Statistisk sentralbyrå
Statistics Norway

Pacific Community
Communauté du Pacifique

PROCESSING FOOD CONSUMPTION DATA FROM HOUSEHOLD CONSUMPTION AND EXPENDITURE SURVEYS (HCES)

Guidelines for countries collecting data in line with the United Nations Statistical Commission-endorsed guidelines on food data collection in HCES

With the technical support of:

Food and Agriculture Organization of the United Nations

THE WORLD BANK

Processing Food Consumption Data from Household Consumption and Expenditure Surveys (HCES)

Guidelines for countries collecting data
in line with the United Nations Statistical Commission-endorsed guidelines
on food data collection in HCES

Prepared by
Members of the United Nations Committee of Experts on Food Security, Agricultural and Rural Statistics (UN-CEAG)
and endorsed by
the Fifty-fifth session of the
United Nations Statistical Commission
New-York March 2024

Original text: English

Pacific Community Cataloguing-in-publication data

Processing food consumption data from household consumption and expenditure surveys (HCES): Guidelines for countries collecting data in line with the United Nations Statistical Commission-endorsed guidelines on food data collection in HCES

1. Food consumption – Data processing – Oceania – Handbooks, manuals, etc.
2. Food security – Oceania – Handbooks, manuals, etc.
3. Food – Composition – Oceania – Handbooks, manuals, etc.
4. Household surveys – Data processing — Oceania – Handbooks, manuals, etc.
5. Consumption (Economics) – Oceania – Handbooks, manuals, etc.

I. Title II. Pacific Community

339.40995 AACR2

ISBN: 978-982-00-1606-4

Cover photo: SPC

Prepared for publication and printed at SPC's headquarters,
B.P. D5, 98848 Noumea Cedex, New Caledonia, 2024
www.spc.int | spc@spc.int

Contents

ABBREVIATIONS

AFH	Away from home
AME	Adult male equivalent
CAPI	Computer-assisted personal interview
COICOP	Classification of individual consumption according to purpose
COMESA	Common Market for Eastern and Southern Africa
CPI	Consumer price index
D4P	Data for Policy initiative
DDI	Data documentation initiative
DEC	Dietary energy consumption
DOS	Domain, obvious and systematic
EA	Enumeration area
FAO	Food and Agriculture Organization (of the United Nations)
FAO ESS	Food and Agriculture Organization (of the United Nations) – Statistics Division
FAO ESN	Food and Agriculture Organization (of the United Nations) – Nutrition Division
FAFH	Food away from home
FCDB	Food composition database
FCT	Food composition table
FPI	Food price index
GIFT	Global individual food consumption data tool
GSDEM	Generic statistical data editing model
HBS	Household budget surveys
HCES	Household consumption and expenditure survey
HIES	Household income and expenditure surveys
IAEG-AG	Inter-Agency Expert Group on Food Security, Agricultural and Rural Statistics (now UN-CEAG)
IFPRI	International Food Policy Research Institute
IHSN	International Household Survey Network
INFOODS	International Network of Food Data Systems (of the FAO)
ISU	International system of unit
ISWGHS	Inter-Secretariat Working Group on Household Surveys
Kcal	Kilocalories
Kg	Kilograms
LCU	Local currency unit
LSMS	Living Standard Measurement Study
MAD	Median absolute deviation
N/A	Not applicable
NCT	Nutrient conversion table
Norad	Norwegian Agency for Development Cooperation
NSO	National statistical office
NSU	Non-standard unit
OECD	Organisation for Economic Co-operation and Development
PAPI	Pen-and-paper interviewing
PSU	Primary sampling unit
SDG	Sustainable Development Goal (of the United Nations)
SSB	Statistics Norway (Statistisk sentralbyrå)
UN	United Nations
UN-CEAG	United Nations Committee of Experts on Food security, Agriculture and Rural Statistics (previously IAEG-AG)
UNHS	Uganda National Household Survey
UNU	United Nations University
USDA	United States Department of Agriculture
WB	World Bank

Boxes

Figures

ACKNOWLEDGEMENTS

These guidelines were drafted under the aegis of the United Nations Committee of Experts on Food Security, Agricultural and Rural Statistics (UN-CEAG, previously known as the Inter-Agency Expert Group on Food Security, Agricultural and Rural Statistics [IAEG-AG]) through its task team on food security and consumption statistics.

The drafting team included Nathalie Troubat, Pacific Community (SPC); Elizabeth Foster, World Bank (WB); Ana Moltedo, Food and Agriculture Organization – Statistics Division (FAO ESS); Astrid Mathiassen, Statistics Norway (SSB); and Ellen Cathrine Kiøsterud, Statistics Norway (SSB).

Experts who provided inputs and comments at several stages of the writing process included Alberto Zezza (WB), Kevin McGee (WB), Josefine Durazo (WB), Michael Sharp (SPC), Andrea Borlizzi (SPC) and Edward Joy (London School of Hygiene & Tropical Medicine), as well as all participants in a workshop held in Rome in October 2022.

The team also relied on inputs from other contributors, including Carlo Cafiero (FAO ESS), Abdul Sattar (FAO ESS), Haoyi Chen (UN Statistical Commission), Owen Siyoto, Common Market for Eastern and Southern Africa (COMESA), and Bridget Holmes, FAO – Nutrition Division (FAO ESN). The team also wishes to acknowledge Valerie Bizier (FAO), who coordinated all the activities on behalf of the UN-CEAG; and Muriel Borderie (SPC) who provided design and layout support of these guidelines.

The team is grateful to the many national statistical offices (NSOs) that provided comments about the guidelines during the global consultation. During this consultation, the NSOs were also requested to provide various information related to food data collected through household consumption and expenditure surveys (HCES) in their countries. This information will be used to inform the dissemination strategy for the guidelines. Seventy NSOs from all over the world provided valuable feedback.

The structure of the guidelines has been tested in practical work by the SPC, WB, FAO and COMESA/SSB during the production process. The team is grateful for feedback from all of those involved.

The team would also like to acknowledge the organisations that provided financial support to this activity. This included the organisations mentioned above, by way of allowing their staff members or consultants to spend time contributing. We would particularly like to acknowledge the Norwegian Agency for Development Cooperation (Norad), through its Agriculture for Development programme; the World Bank, through the Statistical Innovation and Capacity Building in the Pacific Islands (PACTSTAT) project; and the Australian Centre for International Agricultural Research, through its Pacific food system project (project number FIS2018/155), for direct financial contributions.

PREFACE

Household consumption and expenditure surveys (HCES) were primarily undertaken to compile information for important macroeconomic indicators, such as consumer price indices, and to provide input into national accounts. With time, the use of HCES has been further extended to welfare and poverty analysis and more recently to food security and nutrition analysis.

The food data collected in HCES provides core information for these types of analyses. However, the data is comprehensive and complex to process, and users based on their needs or interests quite often tend to follow different approaches when preparing the data for analysis. When data from the same survey is processed independently for different uses, it quite often leads to inconsistent results. Such practice is inefficient and costly.

With this in mind, in 2020, Statistics Norway (SSB) proposed at the 5th meeting of the United Nations Committee of Experts in Food Security, Agricultural and Rural Statistics (UN-CEAG), to develop unified guidelines on how to prepare food consumption data from household consumption and expenditure surveys for "all uses" going from poverty to economic and food security statistics. The proposal was endorsed and a task team was created and these guidelines have been developed through the food security and food consumption measurement task team of the UN-CEAG.

A team of experts from SSB, World Bank, FAO and SPC was put together to draft the guidelines with inputs from representatives of national statistical offices (NSOs), international organisations, survey practitioners, academics and experts in different disciplines (statistics, economics, nutrition and food security). The guidelines draw on materials from existing recommendations, practices and recent research. An initial version of the guidelines was discussed during a workshop in Rome in October 2022 with members of the UN-CEAG group and the Common Market for Eastern and Southern Africa (COMESA) countries. Following the workshop, a revised draft was circulated for comments to the UN-CEAG in July 2023. Based on those comments, a new draft was prepared for global consultation and sent out from mid September to mid-October 2023 to all NSOs from low- to high-income countries.

The result of this process is a set of recommendations in this guideline document. These recommendations are intended to provide countries with standard methods for processing food data from HCES, so as to increase efficiency, quality and consistency when the data is further used for national accounts, consumer price indices, poverty and food security analyses.

INTRODUCTION

The guidelines describe how to process data collected in the food consumption modules of household consumption and expenditure surveys (HCES).[1] These are the modules that collect the monetary value and quantity of each food item a household has acquired and consumed over a given reference period. They also capture which sources the household obtained the food from. Users of these data are diverse (such as price analysts, experts in poverty or in food security) and they may adopt different approaches when they process these data to address their own needs. When these users process the data independently of each other, it often leads to inconsistent results from the same survey. It is also inefficient and costly.

The food consumption module in HCES is part of a larger survey that also captures how much household consumes of other goods and services. The processing of the data described here should, therefore, be integrated into the overall plan for processing the survey.

The primary goal of these guidelines is to assist data owners (mostly NSOs) or organizations or institutions in charge of preparing the data in following one standard process when preparing their data for the main users and proposing a unique dataset with **quantity, dietary energy and monetary value for every food item consumed by the household, from every source of consumption**,[2] **to be used for further analysis.**

Background

HCES were primarily undertaken to compile information for important macroeconomic indicators, such as consumer price indices, and provide input into national accounts.

With time, welfare and poverty measurements have become a core output of the HCES, and over the past decades, these surveys have been further used to inform food security and other socio-economic analyses. There is a long tradition in many countries of implementing HCES, leading to various adaptations of the survey tools to address countries' needs. In the past decade, to provide more reliable and relevant data for food security analysis, both methodological improvements and expanded use of the data have taken place to better assess the amount and types of food consumed in households.

The food module captures the quantity and monetary value of the food acquired and consumed by a household, by individual food items, and the sources the food comes from.

In 2018, the UN Statistical Commission endorsed data collection guidelines (hereafter referred to as the 'IAEG-AG 2018 guidelines')[3] to better capture food consumption in HCES and improve the quality of statistics used to inform poverty, food security and nutrition analyses, while maintaining the traditional purpose to derive the weights for use in the rebase of consumer price indices (CPIs) and in the compilation of national accounts (see Box 1).

The IAEG-AG 2018 guidelines, besides providing recommendations on how to better capture food consumption data, aim to harmonise survey design worldwide to derive global monitoring indicators that can be compared over time and between countries. This also brings up the issue of harmonising survey data processing.

1 Household consumption and expenditure surveys (HCES) are also referred to by a variety of other names including household income and expenditure surveys (HIES), household budget surveys (HBS) or living standards measurement surveys (LSMS). These surveys are conducted on a nationally representative sample to characterise important aspects of household socio-economic conditions.

2 If the food consumption module covers other information than those mentioned in this document, then processing that data must also be integrated. The core idea is that only one round of processing is necessary for the data.

3 The guidelines are available at: https://openknowledge.worldbank.org/server/api/core/bitstreams/327da989-2cda-50a3-8688-1d14110bd554/content

Box 1: Guidelines on food data collection

Many years of research and experience are behind the recommendations endorsed by the UN Statistical Commission in 2018. The guidelines, called *Food Data Collection in Household Consumption and Expenditure Surveys: Guidelines for Low- and Middle-Income Countries*, were prepared by the Inter-Agency Expert Group on Food Security, Agricultural and Rural Statistics (IAEG-AG). IAEG-AG has since become the UN Committee of Experts on Food Security, Agricultural and Rural Statistics (UN-CEAG). These guidelines are recommended as background and for tips on how to improve a survey to better capture food consumption.

About the guidelines for processing food consumption data

The food data processing guidelines presented in this document provide some basic principles to adopt when transforming the food data collected in HCES to data ready for poverty or food security analysis (among other things). The goal is to enable more and more timely, consistent and reliable statistics derived from food consumption data, while also improving the quality and transparency of data processing.

The guidelines and related documentation have been developed under the aegis of the United Nations Committee of Experts on Food Security, Agricultural and Rural Statistics (UN-CEAG). The guidelines bring together information used by different experts working on household food consumption data. The main sources are as follows.

- Guidelines on food data collection developed by the IAEG-AG in 2018.
- Guidelines and training materials from the World Bank teams working on its Living Standard Measurement Study (LSMS) and Data for Policy (D4P) initiative. The World Bank is the custodian of measurement of the indicators on poverty using HCES data under UN Sustainable Development Goal (SDG) 1, *End poverty in all its forms everywhere* (see SDG 1.1.1 and 1.1.2).
- Guidelines and training materials from the UN Food and Agricultural Organization (FAO), which is the custodian of the SDG indicators on food security using HCES data for the prevalence of undernourishment under SDG Goal 2, *End hunger* (see SDG 2.1.1).
- Guidelines and training materials from the Pacific Community (SPC), Inter-secretariat Working Group on Household Surveys (ISWGHS), International Household Survey Network (IHSN), International Food Policy Research Institute, and Statistics Norway.
- Expertise from international consultants and national statistical offices' (NSOs') staff involved in HCES food data processing and analysis,[4] including a five-day seminar in Rome in October 2022 with contributions from experts in 12 different organisations and NSOs.
- Written feedback from experts on the second draft after the seminar in Rome.
- Consultations with members of UN-CEAG, NSOs and other interested stakeholders.

The guidelines suggest a flow of work and various steps on how to approach it, while pointing out issues to be aware of and what to document. The guidelines are built assuming food data is collected or will be collected following the IAEG-AG 2018 guidelines. It is therefore assumed here that the food data collected in the HCES refers to the amount of food *consumed* by the household over

4 See acknowledgements.

the reference period and not the amount of food *acquired to be consumed* during a different period involving building or withdrawing from stocks.[5] This is an important point to keep in mind.[6]

To account for the differences in HCES, advice on how to adjust the process to match different designs is provided. However, there will be variations not covered, and for those, the aim is that the survey owners follow the principles laid out in these guidelines while adjusting the process to fit their data.

The new guidelines do not aim to substitute well-functioning national systems or approaches already established by NSOs for their food data processing. Rather, they provide recommendations on how to improve current systems when appropriate and aim to encourage harmonisation of methods between countries. The guidelines can also be a helpful roadmap when different stakeholders come together to plan data processing that covers the main user needs.

Some other important considerations to keep in mind are as follows.

- The process needs to be transparent and replicable, and it is therefore important that each step is well documented, including each decision taken or adjustment made to the data. Analytical programs and all ad-hoc information should be made available.
- While these guidelines were designed for the use of data after collection, it is recommended that NSOs also refer to the guidelines during the training of enumerators. A basic knowledge of the most important steps to follow when preparing the data will improve data quality when enumerators understand why they are collecting data and why the data should be collected in a certain way.
- Periodical updates of these guidelines may occur to adjust to any improvement in food data collection methods and advance in new technology or knowledge.

Finally, the process discussed in these guidelines have been adopted and implemented successfully by SPC in collaboration with FAO since 2018 to prepare food data collected in the most recent HIES conducted in Pacific Island countries. These surveys have also adopted the harmonized survey design based on the IAEG-AG 2018 guidelines (Kiribati HIES 2019 and 2024, Vanuatu HIES 2019, Marshall Islands HIES 2019/20, Tonga HIES 2021, Palau HIES 2023/24, Tuvalu HIES 2022, Samoa HIES 2023).

Outline of the guidelines

These guidelines are divided into two main parts.

The first part presents food consumption modules and provides some useful principles and general methods to consider before starting work. The analyst needs to assess the data collection tools and other available information before embarking on processing the data. Furthermore, the analyst should decide on the overall approach to cleaning the data.

The second part provides a step-by-step description of food data processing, following 11 steps that describe how to bring the food consumption data from its raw form, as collected in the survey, to transformed data ready to be used for statistical analysis. These steps are referred to as 'food data processing'. At the end of the process, the key variables for each food item reported in the survey, from every source (purchase, own production, etc.) and for each household will be derived as:

- quantity in grams;
- food monetary value (referred to as 'value'); and
- dietary energy (kcal).[7]

See Annex 1 for an example of the final file to be used for future analysis.

5 See the IAEG-AG 2018 guidelines for discussion on acquisition versus consumption.

6 Acquisition surveys need to also collect changes in stocks to calculate consumption. If stocks are not captured, the result may be very high or low average dietary energy consumption at household level. Bulk acquisition could also impact the outlier identification process.

7 These guidelines stop at the dietary energy consumption distribution, needed among other, for both the estimation of poverty and the prevalence of undernourishment. However, once all the quantities of products are converted into grams, it will be possible for analysts interested in other essential nutrients to merge the database with the same food composition tables used for the estimation of dietary energy

These guidelines are intended to be used in the following ways.

1) Make users familiar with the principles and methods outlined here *before* starting processing. If there was a previous survey, check that the methods used to process the food data followed the basic recommendations described. Otherwise, identify the gaps between the method previously used and the recommendations, and plan to assess the difference in the estimates between new and old methods at the end of the process.

2) Check the users food modules versus those described in the next section. This is because the specificities of the module used will determine the choices in the processing plan, so the process will need to be adjusted if the data collected deviates from what is collected in the modules shown.

3) Check the survey data, what is collected and what will be needed as auxiliary data, and make sure to have all the information needed and ready for the process. An overview of what is needed is listed under Step 1 in the process. Then follow the other steps, as illustrated in Figure 1 and further developed in Part 2.

Finally, these guidelines encourage NSOs and other organizations involved in food data collection and processing to further adopt the practices listed in Box 2 below.

Box 2: Practices to adopt for an efficient process

Joint processing of food consumption data

Ensure that the processing of food consumption data from HCES is done in a single process, accommodating main users. In most low- and middle-income countries, these users include welfare/poverty or food security analysts and consumer price index and national accounts' compilers. The standardisation of data processing through a team of coordinated users will allow more control and harmonisation of survey results, avoiding different results for the same indicators from the same survey. It will also reduce the costly and inefficient practice of having data processed in different ways by several users.

Full processing of food consumption data

Data processing must result in a dataset that can be used for all the main analyses, and that the statistics can be disaggregated for all relevant populations. This implies full processing of monetary values and quantities of all foods consumed. The main improvement in most cases, if data collection has made it possible, will be to calculate quantity in grams and/or nutrient content (mainly dietary energy) for all food items reported.

Consistency in data processing between surveys

To allow for comparison over time, successive surveys should, to the extent possible, be processed following the same methods. Food consumption statistics are very sensitive, not only to survey design but also to choice of methods used during data processing. Analysts make different choices according to cost-benefit considerations of their main interests, as well as personal skills and methodological preferences. Food modules and approaches to data processing may be revised to follow international recommendations. After stakeholders have agreed on revisions needed and the approaches to follow, there must be routines in place to ensure that these are replicated for later survey. The process needs to be revised to adjust to any improvement in the food consumption module.

Generate high-quality nutrient conversion tables (NCTs)

NCTs provide information about the nutrient composition of each food item collected in the HCES. They are built using information from national, regional or international food composition tables and databases (FCT/FCDB). The NCT is used to convert quantities collected into nutrient values, such as macronutrients, from which dietary energy is estimated (kilocalories [kcal]). Different analysts may produce different NCTs depending on the FCTs/FCDBs of reference they have used. Such practice usually translates into the generation of different distributions of dietary energy and macronutrients from the same surveys. NSOs and others involved in food consumption data processing are **encouraged to create only one NCT (of good quality) for the survey in cooperation with nutrition experts and ensure consistency in its use.** Of note, it is recommended to invest in the development of a good NCT that can be used in further survey processing and be updated over time to follow changes in dietary patterns in the country.

Documentation

A prerequisite to ensure consistency, transparency, replicability and trust in the processed data is that the methods used are properly documented and available for users. The NSOs/organisations – and in particular the experts involved in the processing – are encouraged to ensure that this takes place. It includes making processing syntax files available with the data, along with the critical auxiliary data, including the NCT and weight in grams (often called 'conversion factors') for non-standard units used to measure quantity. Documentation also includes good metadata to describe each variable in the datasets.

Share practices

Making methods and documentation available is beneficial and efficient for the harmonisation of data collection and statistics production. All actors in this area – members of the national statistical systems, UN organisations, research organisations, universities, consultants, and international and national non-governmental organisations – are encouraged to share research, methods and training materials to contribute to the common knowledge base.

PART 1: REVIEWING MODE OF DATA COLLECTION AND CLEANING

1 FOOD CONSUMPTION DATA COLLECTION MODULES

These guidelines assume that the following data is available.

- The **quantity** of each food item consumed by each household and the time period for which it was consumed, for example, 'six cups of rice consumed over the past seven days'.

  The quantity should be collected directly through the HCES. A choice here is just on how many different units of measurement the household may use when reporting the quantity. The IAEG-AG 2018 guidelines recommend the use of non-standard units (NSUs) when the respondent does not know how to express the quantities in a standard unit (such as grams, kilograms [kg] or pounds [lb]). For all NSUs, the weight in grams of one unit must also be available. For some special foods, such as those prepared and consumed away from home, it may be not possible to collect quantities. In such cases, monetary value may be considered enough to estimate quantity or nutrient values.[8]

- The **monetary value** corresponding to the quantity of the food consumed or a **price** (see Box 3 below on the terminology) that can be used to impute the monetary value of the food consumed from the quantity reported, for example, 'one bag of 250 grams of rice costs 1.5 dollars'.

  This information can be generated directly from the survey or by using external sources.

Quantity and monetary value information is collected in different ways and usually in separate modules for in-house consumption, food away from home (see Box 4) and for the different sources of consumption (see Box 5.) It can also be collected using a diary or recall interview. Users can follow the steps of the process and apply the main principles regardless of which survey design has been used to collect the food data.[9]

Box 3: Terminology: Value and prices

1) For the sake of simplification, the expression 'monetary value' or the word 'value' are used interchangeably to refer to the monetary value in local currency units (LCUs) of the food consumed without distinguishing if the food was purchased or coming from non-market sources. Some users may be more familiar with the expression 'cost' of purchases, which was discarded by the authors and a decision made to stick to only one terminology throughout the document to avoid potential confusion.
2) In the text, the word 'price' is used when referring to prices in general regardless of the source of information.
   We talk about 'market price' when referring to prices collected in a market. For example, prices can be collected at the local market for the unit of measurement for a food.
   The term monetary 'unit values' is used to refer to the price calculated from the household data, and is found by dividing the value by the quantity. Unit values will be missing for records that we need to impute (because both valid quantities and monetary values are needed to calculate it).

8 How to collect reliable data on foods consumed away from home in a cost-efficient way remains a challenge. The IAEG-AG 2018 guidelines recommend that more research work is conducted to reach this goal.

9 For a thorough discussion on differences in survey design, see Smith et al. 2014, available at: http://www.ihsn.org/food

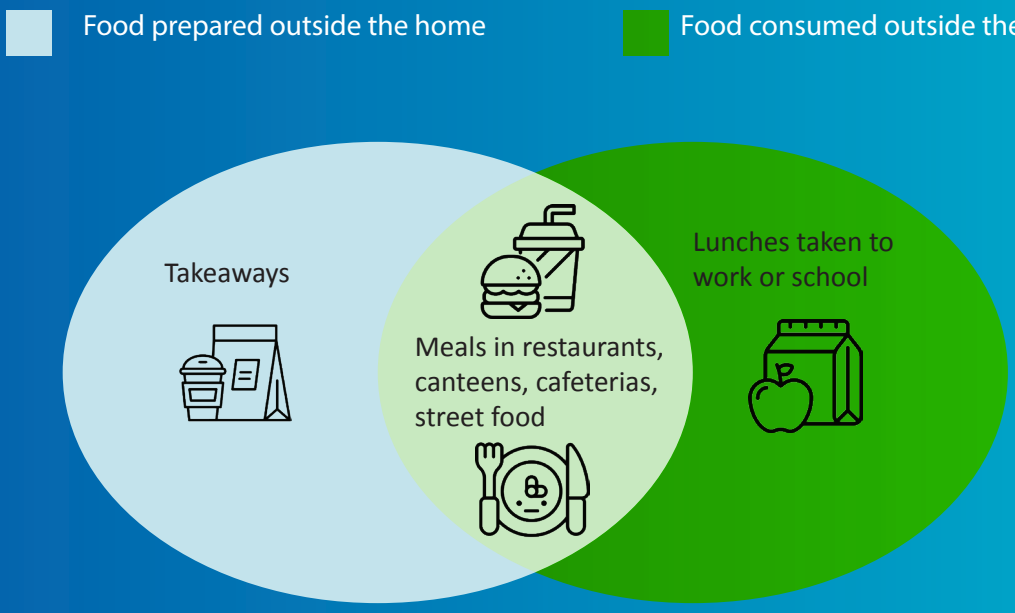
Box 4: In-house food consumption versus away-from-home consumption

Many questionnaires separate in-house (or at-home) food consumption from food away from home.

In-house food consumption generally refers to the food prepared and consumed in the house. It is almost always collected at the household level from one main respondent, ideally the person responsible for food purchase and preparation. Consumption of specific individual food items (rice, oil, salt, etc.) is recorded with quantities. Information on the source of food from purchases, own production and other sources, such as gifts, is also collected for each food item.

Food away from home (FAFH) generally refers to all meals prepared and consumed outside the dwelling. Meals or foods prepared away from home refer to meals prepared outside the dwelling with ingredients that were not reported in the in-house section of the questionnaire. When these meals are prepared and consumed outside the dwelling, they are considered as FAFH. Quite often FAFH consists of prepared meals, but it can also include beverages. FAFH is collected using only a few categories (referring either to a meal event such as breakfast, lunch or dinner or specific categories of foods such as fast food, pizza, snacks or barbecued food). FAFH includes purchased foods and meals received for free as a gift or payment but excludes consumption from own production. FAFH information may be collected at the household or individual level.

Food prepared at home but consumed outside the home (such as packed lunches) and food prepared outside the home but consumed at home (takeaway) are not included in the definition of FAFH. Both are considered as in-house food consumption. All the ingredients needed to prepare packed lunches – as well as takeaway foods consumed inside the house by household members – should be reported in the in-house food consumption module.



It is important to include both in-house and away-from-home consumption while making sure to avoid any double counting.

SOURCE: Authors' elaboration.

Box 5: Sources of food consumption

Households can obtain foods from different sources, and it is important to assess how external shocks may affect sources and consumption. For example, a price hike may have a lower impact on households consuming food from their own production than on households mainly purchasing food. Therefore, it is important to keep information about the sources of consumption throughout the process because it can later be used to inform policy analysis.

The main sources of food consumption are as follows.

- Purchased food:
    - raw food bought in shops and from street vendors, etc., and cooked and/or consumed at home;
    - prepared foods, such as baked goods and processed meats etc., bought in retail establishments; and
    - ready-made meals bought at restaurants, takeaway establishments, from street vendors, etc.
- Own production:
    - from own farm and/or garden;
    - hunted, fished, collected in the wild; and
    - from own retail shops.
- Food received as gifts, in-kind payment or other:
    - in-kind payment for the provision of goods and services;
    - gifts from family, friends or events, such as celebrations or church (includes gifts of food items and meals eaten in another's home or at an event);
    - food aid; and
    - free meals received at school or work.
- Food from stocks:
  Food obtained before the reference period of food data collection, which the household consumed during the reference period. This is only applicable in an acquisition survey where the collection of opening and closing stocks at the respective start and end of the reference period allows consumption to be inferred.
- Food from rations:
  Rations are "the provision of quotas of food items for free or at below-market price".

Note: Food away from home (FAFH) is in this context not a separate source but will be counted in as one of the sources above.

1.1 In-house consumption

Food *quantities* can be reported in standard units of measurement, such as kilograms, grams and litres, or in non-standard units (NSUs), such as heaps and cups. The use of NSUs is recommended in the IAEG-AG 2018 guidelines to simplify food data collection. In cases where all quantities are reported in standard units, the food data processing is simplified but the steps and principles to use remain the same.

Information about the *monetary value* of food can be reported in various ways, which are summarised by the following two questions.

1) **Direct or indirect report of value of purchases?**

   Information on the value of purchases has almost always been included. However, respondents were traditionally asked to give the total value of consumption from purchases or the total value of purchases over the recall period[10] (direct report). The new recommendation from the living standard measurement study (LSMS) team is to collect only quantity consumed from purchases and then quantity and value on the last single purchase of the item made by the household, possibly allowing a longer recall period. This is believed to be easier for the respondent and produce more accurate information (indirect report) than previously. Another form of indirect report is to ask the respondent about the quantity consumed from purchases, as well as the total monetary value and quantity of all purchases over the same recall period. Both methods require imputation, but the 'direct report' requires that each respondent does the imputation while the 'indirect report' provides a consistent imputation from the data team.

2) **Is an estimate of the value of non-market consumption collected?**

   Respondents are not always asked to provide an estimate of the value of their non-market consumption[11] (own production or other sources such as food received as gifts or ration). Values of non-market consumption need to be imputed when they are not collected.

   The following are three example questionnaires that answer these questions in different ways and show how to proceed in each case.

Example 1: 2019/20 Uganda National Panel Survey (UNPS)

Direct report of value of purchases and estimate of value of non-market consumption

In the following questionnaire (Figure 1), both total value of consumption from purchases (column CEB07) and respondent estimates of value of consumption from own production (column CEB11) and other sources (column CEB13) are recorded. In this example, the values can be used directly. In some surveys, however, the respondent's value estimates are not used directly if they are judged to be too noisy, and the non-market quantities consumed are then valued using a unit value constructed from purchases (see Step 5 later in this document on monetary value imputation).

10 Although acquisition and consumption are theoretically different, in practice they are treated the same, assuming that consumption is approximately equal to acquisition. The validity of this assumption depends on the context, for example, it is more valid in contexts with limited storage and where households shop daily. (See the IAEG-AG 2018 guidelines for the discussion on acquisition versus consumption).

11 The IAEG-AG 2018 guidelines regarding non-market consumption include that: "… Such food should be valued at 'basic prices' of similar goods, which can be approximated by the price of similar goods sold on a local market, or the price declared by the household producer if he or she had sold the food rather than consumed it. Information on food and meals acquired through in-kind transfers is also important. Valuation should be based on actual cost if actually purchased by the provider or production cost, both being unknown and difficult to evaluate by the beneficiary".

Figure 1: In-house consumption module from the 2019/20 Uganda National Panel Survey (pen-and-paper interviewing [PAPI] questionnaire)

PART B: FOOD, BEVERAGES, AND TOBACCO (DURING THE LAST 7 DAYS)

Item Description	Code	Did your HH consume [ITEM] In the last 7 days? 1= Yes 2=No (>> NEXT ITEM)	How many days was [ITEM] consumed out of the last 7 days?	In what unit can you best quantify/ describe the amount of [ITEM] consumed? CODES	In what state was [ITEM] mainly consumed in the last 7 days? 1=Baked 2=Boiled 3=Fried 4=Raw/fresh 5=Roasted 6=Steamed 7=N/A	AT HOME		Where did you purchase the [ITEM]? 1= Supermarket 2= Open Market 3= Kiosk 4= General Shop 5= Specialized shop 6= Public or semi-public service providers 7= Private Service Providers 8= Mobile shops and street Vendors 9= Online Local 10=Online International 11= Outside Uganda 12= Other Households 96=Others (specify) 97=N/A	OUT OF HOME PRODUCTION/HH ENTERPRISE STOCK		RECEIVED IN-KIND/FREE		What is the market price per [CEB05] of [ITEM]?	What is the Farm gate /producer price per [CEB05] of [ITEM]
						How many [CEB05] of [ITEM] did your household consume from purchases at home during the last 7 days?	How much did the household spend in total on these [CEB06] [CEB05] of [ITEM] during the last 7 days? for consumption at home? UGX		How many [CEB05of [ITEM] did your household consume out of home production during the last 7 days?	What would be the total value of these [CEB10] [CEB05of [ITEM] if you were to sell them at the farm gate? UGX	How many [CEB05] of [ITEM] did your household consume from received in-kind/free during the last 7 days?	What would be the total value of these CEB12] [CEB05] of [ITEM] if you were to sell them in the market? UGX		
CEB01	CEB02	CEB03	CEB04	CEB05	CEB05A	CEB06	CEB07	CEB16	CEB10	CEB11	CEB12	CEB13	CEB14	CEB15
Matooke (Bunch)	101													
Matooke (Cluster)	102													
Matooke (Heap)	103													
Matooke (Sack)	104_1													
Matooke (Piece)	104_2													
Sweet Potatoes (Fresh)	105													
Sweet Potatoes (Dry)	106													
Cassava (Fresh)	107													
Cassava (Dry)	108_1													
Cassava (Flour)	108_2													
Yams (arrow root)	108_4													
Sugarcane	147_3													
Pancakes(Kabalagala)	108_3													
Irish Potatoes	109													
Rice (white/brown)	110													
Macaroni/Spaghetti	172_1													
Maize (grains) -white/ yellow	111													
Maize (cobs) -white/ yellow	112													
Maize (flour) -white/ yellow	113													
Roasted goat meat	119_3													

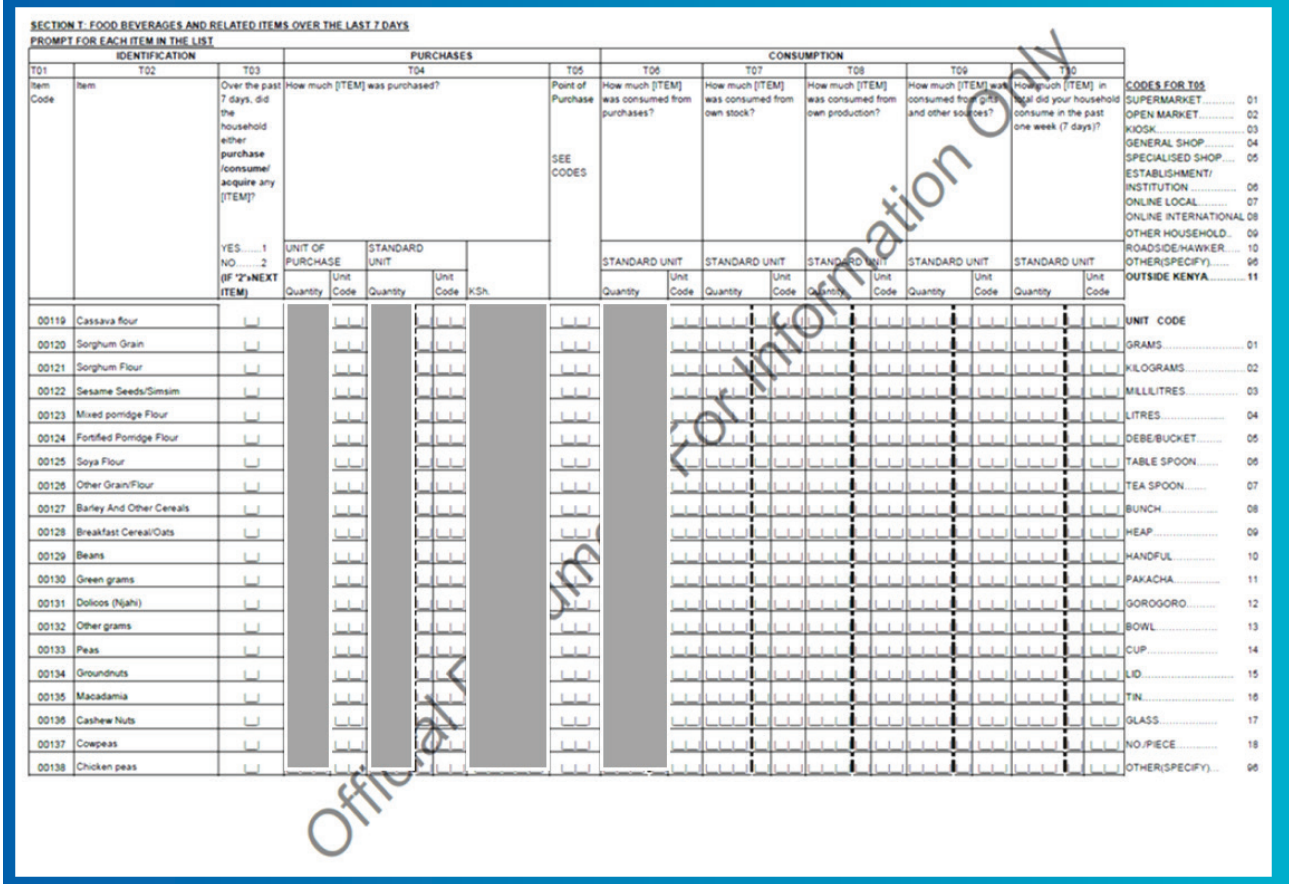
SOURCE: 2019/20 Uganda National Panel Survey questionnaire.

Example 2: 2015/16 Kenya Integrated Household Budget Survey (KIHBS)

Indirect report of value of purchases and no estimate of value of non-market consumption

In this example (see Figure 2, the respondent is asked to report both the quantity and the value of all food purchased over the past seven days (column T04) and the quantity consumed from purchases (column T06). For instance, if the household purchased five kilograms (kg) of rice in the past seven days and paid 500 Kenyan shillings (KSh), but consumed only four kg. This requires an additional step to value the consumption from purchases, using the unit value estimated from total purchase for this household. (In the example, the cost of one kg of rice estimated from the purchases is 100KSh, which means that the cost of rice consumed in the past seven days was 400KSh). In cases where no purchase occurred during the recall period or the unit of measurement of quantities purchased did not correspond to the unit value of quantities consumed from purchases (for example, if the household bought five kg of rice and consumed 10 cups), the monetary value corresponding to the quantities consumed needs to be estimated (see discussion in Step 5 on monetary value imputation). In this example, the value of the quantities consumed from non-market sources would have to be valued using an appropriate price (see discussion in Step 5 on the choice of the price).

Figure 2: In-House Consumption module from the 2015/16 Kenya Integrated Household Budget Survey (PAPI questionnaire)



SECTION T: FOOD BEVERAGES AND RELATED ITEMS OVER THE LAST 7 DAYS
PROMPT FOR EACH ITEM IN THE LIST

IDENTIFICATION		PURCHASES					CONSUMPTION									
T01 Item Code	T02 Item	T03 Over the past 7 days, did the household either purchase /consume/ acquire any [ITEM]? YES.....1 NO.....2 (IF "2"»NEXT ITEM)	T04 How much [ITEM] was purchased? UNIT OF PURCHASE Quantity	Unit Code	STANDARD UNIT Quantity	Unit Code	KSh.	T05 Point of Purchase SEE CODES	T06 How much [ITEM] was consumed from purchases? STANDARD UNIT Quantity	Unit Code	T07 How much [ITEM] was consumed from own stock? STANDARD UNIT Quantity	Unit Code	T08 How much [ITEM] was consumed from own production? STANDARD UNIT Quantity	Unit Code	T09 How much [ITEM] was consumed from gifts and other sources? STANDARD UNIT Quantity	Unit Code	T10 How much [ITEM] in total did your household consume in the past one week (7 days)? STANDARD UNIT Quantity	Unit Code
00119	Cassava flour																	
00120	Sorghum Grain																	
00121	Sorghum Flour																	
00122	Sesame Seeds/Simsim																	
00123	Mixed porridge Flour																	
00124	Fortified Porridge Flour																	
00125	Soya Flour																	
00126	Other Grain/Flour																	
00127	Barley And Other Cereals																	
00128	Breakfast Cereal/Oats																	
00129	Beans																	
00130	Green grams																	
00131	Dolicos (Njahi)																	
00132	Other grams																	
00133	Peas																	
00134	Groundnuts																	
00135	Macadamia																	
00136	Cashew Nuts																	
00137	Cowpeas																	
00138	Chicken peas																	

CODES FOR T05
SUPERMARKET 01
OPEN MARKET 02
KIOSK 03
GENERAL SHOP 04
SPECIALISED SHOP 05
ESTABLISHMENT/ INSTITUTION 06
ONLINE LOCAL 07
ONLINE INTERNATIONAL 08
OTHER HOUSEHOLD 09
ROADSIDE/HAWKER 10
OTHER(SPECIFY) 96
OUTSIDE KENYA 11

UNIT CODE
GRAMS 01
KILOGRAMS 02
MILLILITRES 03
LITRES 04
DEBE/BUCKET 05
TABLE SPOON 06
TEA SPOON 07
BUNCH 08
HEAP 09
HANDFUL 10
PAKACHA 11
GOROGORO 12
BOWL 13
CUP 14
LID 15
TIN 16
GLASS 17
NO./PIECE 18
OTHER(SPECIFY) 96

Example 3: Living Standard Measurement Study (LSMS) survey model

Indirect report of value of purchases and no estimate of value of non-market consumption

The IAEG-AG 2018 guidelines provide recommendations on what to collect in the food consumption module, along with significant background and research on how to collect it, but fall short in providing a template to collect food data. To address the gap, the World Bank LSMS team released a guidebook[12] in 2021 (hereafter, referred to as the LSMS 2021 guidebook) for designing household surveys and collecting food data. The guidebook provides practical applications for the IAEG-AG 2018 guidelines. For example, it recommends collecting only quantities of food consumed the past seven days (from purchases and from non-market sources, columns 3, 4 and 5 in Figure 3) together with the quantity and monetary value of purchases that occurred the past 30 days (columns 6, 7 and 8). The unit value estimated from the most recent purchase is used to estimate the value of the food consumed. For example, if a household paid Kenyan Shillings (KSh) 1500 to purchase 15 kg of rice two weeks ago, then the unit value of one kg is KSh 100. If that household consumed five kg of rice in the past seven days, using the unit value of KSh 100, the household's consumption of rice is valued at KSh 500). If a household consumed a specific food during the seven-day recall period for which no purchase occurred in the past month, or if different measurement units were used to record consumption and last purchases, the consumption would need to be valued differently (see discussion in Step 5 on imputation).

12 See Oseni et al. 2021, available at: https://www.worldbank.org/en/programs/lsms/publication/CapturingWhatMattersEssentialGuidelinesforDesigningHouseholdSurveys

Figure 3: In-home consumption, living standard measurement study (LSMS) model (PAPI questionnaire)

SECTION 9B: WITHIN-HOUSEHOLD FOOD CONSUMPTION & EXPENDITURE

1. PRIMARY RESPONDENT FOR THIS SECTION ID CODE

	REPORT ONLY ITEMS CONSUMED WITHIN THE HOUSEHOLD. FOOD CONSUMED OUTSIDE THE HOUSEHOLD MUST BE REPORTED IN THE PREVIOUS SECTION. ASK Q1 FOR ALL ITEMS, BEFORE MOVING TO FOLLOW UP QUESTIONS.	ITEM CODE	1	2			3	4	5	6	7			8
			During the past 7 days, did members of this household eat/drink any [ITEM] within the household? YES... 1 NO... 2 ▸ NEXT ITEM	In total, how much of [ITEM] did your household consume in the past 7 days? DO NOT INCLUDE FOOD CONSUMED OUTSIDE THE HOUSEHOLD SEE UNIT CODES ON NEXT PAGE			Of the [QUANTITY] [UNIT] of [ITEM] consumed, how much came from purchases? USE UNIT AND SIZE FROM Q2	Of the [QUANTITY] [UNIT] of [ITEM] consumed in the last 7 days, how much came from own production? USE UNIT AND SIZE FROM Q2	Of the [QUANTITY] [UNIT] of [ITEM] consumed, how much came from gifts and other sources? USE UNIT AND SIZE FROM Q2	Did your household purchase any [ITEM] in the past 30 days? YES... 1 NO... 2 ▸ NEXT ITEM	In the most recent purchase, how much [ITEM] did the household buy? SEE UNIT CODES ON NEXT PAGE			How much did your household spend on this [QUANTITY, UNIT, SIZE IN Q7] of [ITEM]?
				QUANTITY	UNIT	SIZE	QUANTITY	QUANTITY	QUANTITY	CODE	QUANTITY	UNIT	SIZE	CURRENCY
Grains & flours	Guinea corn/sorghum	100												
	Millet	101												
	Rice - local	102												
	Rice - imported	103												
	Maize flour	104												
	Cassava flour	105												
	Wheat flour	106												
	Maize (unshelled/on the cob)	107												
	Maize (shelled/off the cob)	108												
	Other grains and flour (specify)	109												
Baked/ processed products	Bread	110												
	Cake	111												
	Buns/pofpof/donuts	112												
	Biscuits	113												
	Meat pie/sausage roll	114												

SOURCE: Oseni et al. 2021, p108.

In examples 2 and 3, the value of food consumed from purchases needs to be imputed with the differences noted – in example 2, we have the value of the total purchases over the recall period, while in example 3, the LSMS model, we have only the value of the last purchase. These differences could slightly affect the value of a unit of product depending on the frequency of purchases.

1.2 Food away from home (FAFH)

The collection of data about food consumed away from home varies even more than in-house food consumption data.

Traditionally, FAFH information has been collected with one or only a few lines asking about the monetary value, for example, 'Food in restaurants (value in local currency)'. Free food consumed away from home, such as free school lunches, has often not been recorded at all. Various studies found that this approach tends to underestimate the amount of food obtained and consumed away from the home or distort the distribution of food between households (see for example Borlizzi et al. 2017 or Sharp et al. 2022). In the cases where some population groups (often urban) eat more food away from home than others, it will also lead to skewed results. To address the issue of misreporting, the IAEG-AG 2018 guidelines suggest that "all surveys collect data at the individual level and all surveys differentiate meal types and make explicit reference to snacks"[13] but also that "more research on this topic is urgently needed".[14]

The following are two variants of food away-from-home modules.

13 See FAO/World Bank (2018), p.24, available at: https://www.worldbank.org/en/programs/lsms/publication/Food-data-collection-in-household-consumption-and-expenditure-surveys-guidelines-for-low-and-middle-income-countries

14 Ibid, p.36.

Example 4: Marshall Islands 2018 Household Income and Expenditure Survey

Individual level, value of purchases and value of free meals

This example (see Figure 4)[15] is implemented at the individual level as recommended by the IAEG-AG 2018. Each household member is asked to report on both the monetary value corresponding to the meals purchased and consumed away from home and the number of meals consumed at each meal event (p23b3 and p23b4). Household members are also asked to provide the number of meals consumed away from home and received for free (p23b5) and an estimate of the value of these meals (p23b6). Note that the module shown is for lunch meals, but similar modules for breakfast and dinner are included in the questionnaire.

Example 5: Living standard measurement study (LSMS) model

Household level, purchased and free, value of purchases and value of free meals

The second variant of the food away-from-home module (see Figure 5) comes from the LSMS 2021 guidebook. In this example, one respondent in the household reports consumption of food away from home on behalf of all household members. Asking at the individual level, as in the previous example, gives more accurate information but is also more time-consuming than asking one person to respond on behalf of the entire household.[16] Moreover, the module in Figure 6 does not allow the analyst to distinguish between FAFH purchased and received for free and does not capture the number of meals taken away from home.

15 See, for example, the recent HIES conducted in Pacific Island countries (information about the surveys is in the Pacific Microdata library at https://microdata.pacificdata.org/index.php/home), and the recent research study from Sharp et al. (2022).

16 Not that individual data can be used to analyse who in the households eat outside the home, but when the data is merged with the at-home consumption module (which does not separate between which household members ate what), then the element of individual consumption is lost.

Figure 4: Module on food away from home (FAFH) section on lunch, collected in the 2018 Marshall Islands HIES (CAPI setup), individual response

4. FOOD AWAY FROM HOME / [23] FAFH

[B] LUNCH

P23b1. In the last 7 days, did %hmName% consume any LUNCH meals outside the house that were purchased OR gifted?	SINGLE-SELECT p23b1 01 Yes 02 No

P23b2. In the last 7 days, how many LUNCH meals did %hmName% consume away from home?
`p23b1 == 1`
`self > 0`
There should be at least 1 lunch meal.
`self.InRange(0, 7)`
The number of lunch meals entered seems unlikely. Please verify.
	NUMERIC: INTEGER p23b2

P23b3: How many of the %p23b2% LUNCH meal(s) were purchased?
`p23b1 == 1`
`p23b3 <= p23b2`
Quantity purchased in cash should not be greater than quantity consum ed.
	NUMERIC: INTEGER p23b3

P23b4: In total how much (in USD) was paid for the %p23b3% LUNCH meal(s)?
Allow for 2-decimal points (e.g 12.50, 1.69, 59)
`p23b3 > 0`
`self > 0`
The amount entered is invalid.
	NUMERIC: DECIMAL p23b4

P23b5: How many of the %p23b2% LUNCH meal(s) were provided for free?
`p23b1 == 1 && (p23b2 > p23b3)`
`p23b3 + p23b5 == p23b2`
Quantity purchased in cash + quantity gotten for free should be equal to quantity consumed.
	NUMERIC: INTEGER p23b5

P23b6: If %hmName% were to pay for the %p23b5% free LUNCH meal(s), what is the amount (in USD) %heShe% could have paid?	NUMERIC: DECIMAL p23b6

Credit: Unsplash - Daria Strategy

Figure 5: Food away from home (FAFH) module from the Living Standard Measurement Study (LSMS) 2021 guidebook (paper version) – household response

SECTION 9A: MEALS AWAY

1. Primary respondent for this section ID code

MEALS PREPARED AND CONSUMED OUTSIDE THE HOME		ITEM CODE	1 In the past 7 days, did any members of this household consume any of the following meals or drinks away from home? READ OUT EACH MEAL TYPE AND RECORD YES/NO. ANSWER Q1 FOR ALL MEAL TYPES BEFORE PROCEEDING. DO NOT COUNT FOOD EATEN FROM THE HOUSEHOLD'S OWN FAMILY BUSINESS. YES... 1 NO... 2 (▸ NEXT ITEM)	2 In the past 7 days, how much IN TOTAL did household members pay for [MEAL]? If it was free, please estimate what it would have cost if you had to pay. CURRENCY
Breakfast	Complete meals [local example]	1		
Lunch	Complete meals [local example]	2		
Dinner	Complete meals [local example]	3		
Snacks such as chips, biscuits, [local examples], etc.		4		
Dairy-based beverages such as milk, yoghurt, [local example], etc.		5		
Vegetable snacks (ex: carrot, pears, roasted corn, roasted plantain, roasted yam, etc.)		6		
Non-alcoholic drinks (Coke, Fanta, [local examples], etc.)		7		
Alcoholic drinks (palm wine, beer, etc.)		8		

Designer Note:
The item list is an example and must be tailored to the country and survey needs in consultation with local experts.

The following table summarises variations in the food consumption modules that align with the IAEG-AG 2018 guidelines.

Table 1. Variants in in-house and away-from-home modules based on the IAEG-AG 2018 guidelines

Source of consumption	In-house food (household level)	Food away from home (household or individual level or both)
Total	Total quantity consumed (i.e. aggregated from all sources*) *[Sometimes included]*	Total quantity consumed *[Rarely included]*
Purchases	Quantity consumed and its respective monetary value over the recall period for food consumption (direct report) OR ONLY quantity consumed over the recall period for food consumption AND quantity and value from purchases (either last single purchase or total over a different recall period) (indirect report)	Quantity consumed (usually in rough, pre-defined units such as number of meals) and its respective value over the recall period for food consumption OR Only value over the recall period for food consumption
Own production	Quantity consumed with respondent-estimated monetary value over the recall period for food consumption OR Only quantity consumed over the recall period for food consumption	N/A
Gifts/other	Quantity consumed with respondent-estimated monetary value over the recall period for food consumption OR Only quantity consumed over the recall period for food consumption	Quantity consumed (usually in rough, pre-defined units such as number of meals) and respondent-estimated value over the recall period for food consumption OR Only respondent estimated value over the recall period for food consumption

* Note that some surveys do not collect purchased quantities, instead the information is constructed as total quantity from all sources, including own production and gifts/other. As this is considered a practice to avoid, how to impute the monetary value of food consumed from purchases in such cases is not discussed in these guidelines.

2 APPROACHES TO CLEANING FOOD DATA

Good and reliable data depends on the checks performed throughout the process as well as the methods used to correct and impute values. This section reviews some basic principles to follow when cleaning food data. These are general and apply no matter which survey design is adopted. They are further described in Part 2 under each step of the process and when they apply. It is also important to note that this chapter is not intended to be exhaustive and does not substitute cleaning systems adopted by the data provider (usually a NSO), which may provide alternative options or complement these systems.

It is important to remind users that these guidelines assume that food data collected refers to consumption and not acquisition, as recommended in the IAEG-AG 2018 guidelines. The identification of outliers can be influenced if this assumption does not hold true.

2.1 Four stages of data cleaning

Data cleaning takes place at different stages within the 11-step process.

The goals and methods used in each stage are summarised below. Details about how to perform each type of cleaning are included later in the appropriate step.

1) First stage: The goal is to identify and deal with any truly impossible values (negative monetary values, quantities of milk reported in heaps, etc.), and to uncover any systematic mistakes in the data (because of questionnaire administration, data processing etc.). All relevant variables in all datasets are inspected. The methods are mostly ad hoc inspection. Systematic issues identified should be corrected if and when possible.

2) Second and third stages: The goal is to systematically address errors that would significantly impact the nutritional or welfare status of the household, and to prevent random errors or clusters of errors from unduly influencing the construction of unit values or other imputations. Specific key variables (such as quantities, monetary or unit values) are inspected and outliers are identified based on assumptions about their distribution (usually that per-capita values should be log-normally distributed) with a conservative definition of what constitutes an outlier. Outliers are replaced by imputed values. If the quantity is only reported in one standard unit of measurement (for example, grams), then the second stage of cleaning is not relevant.

3) Fourth stage: The goal is a final check on the distribution of the main aggregates. Outliers are not generally addressed at this stage, but any irregularity discovered is a reason to return to earlier steps in the process to identify and adjust – when possible – for the causes of these outliers.

2.2 Consistency and traceability

Existing methods

NSOs might have already developed their own methods to detect outliers and these guidelines are not meant to replace systems already in place. Unless there are good reasons for changing the methods, it is important to stick to the choice made over time, as some of the statistics produced from this data are very sensitive to changes in methods. To enhance comparability over time, data must be processed in the same way.

Retain original values

It is recommended to always save the original datasets and work on a copy, which is referred to as the 'working dataset'. Further, make sure to keep the original values in the working dataset. Variables that are foreseen to be altered (mainly quantities and monetary value) will be duplicated and a new name given to the duplicate. Never alter the original variables.

Flagging outliers

Whenever an observation is detected as an outlier, it needs to be flagged. A dummy variable can be created with Value 1 when an observation has been detected as an outlier, and Value 0 otherwise. Flagging outliers is important for further reference to the original data and for documentation.

2.3 Basic overview of outliers

Outliers versus errors

Looking for outliers is one tool to identify potential errors, but not every outlier is an error and not every error is an outlier. The survey might cover a very wealthy household for which consumption appears in a certain distribution as an outlier but is accurately reported. For example, a household holding a party will report huge quantities of food consumed and this will appear as an outlier if the quantities are looked at only at the household level but will not be an outlier if quantities are adjusted to account for the number of people who actually consumed the food.

Systematic versus random errors

Errors in the data can occur for various reasons. Some are random, such as a punching error or a household question interpretation error,

and some are systematic. Systematic errors can take place in a variety of processes, including mistakes in questionnaire design, programming of data entry application, data processing, misunderstanding of the questionnaire and instrument failure. Both systematic and random errors have the potential to be hugely influential on the results, and they need to be detected and corrected at the earliest stage possible of the process using proper methods.

Univariate versus multivariate approach

In a 'univariate' approach, each variable is checked separately. This approach identifies extreme values within a variable's distribution.

A 'multivariate' approach is used when looking simultaneously at the distribution of more than one variable to assess the consistency between, say, quantity and value reported for a food purchase. This allows for more rigorous detection of outliers. For example, where neither a low monetary value nor a related high quantity are detected as outliers within their own distributions, it is then through the abnormally low unit value (monetary value divided by quantity) that either the quantity or the monetary value will be confirmed as outlier.

Automatic versus visual inspection methods

Visual inspection uses graphs, such as histograms and box plots, to explore distributions and tables to look at correspondences between variables. It is used mainly to detect systematic errors. Examples are futher provided in Step 2 (first cleaning – 'domain, obvious and systematic [DOS] editing) of the process (see Part 2, Step 2). However, when dealing with large numbers of food items, visual inspection methods may be considered too time-consuming to implement and/or too subjective in their interpretation, and automatic methods are used. The expression 'automatic method' refers to all the statistical methods that can be embedded in any automatic routines developed to process food data. These methods measure the statistical dispersion of the distribution, assuming that data is symmetrically distributed. They are usually easy to program and are embedded in most analytical software, such as Stata, SPSS, R or Python.

2.4 Basic considerations when cleaning data

Use of subgroups

Detection of outliers is sometimes performed separately for subgroups of households by location or month of data collection. As it is more difficult to detect errors in small groups, it is advised not to subdivide unless there is good evidence that expenditure and consumption patterns are likely to be different from each other. Often, strata or urban/rural as a sub-level is enough. Sometimes a subgroup is by definition very small, for example, when looking for errors in a combination of a particular food and measurement unit. It is recommended to have at least 10 observations in each subgroup (this number is the result of a consensual agreement between experts with large experience working on food data collected in HCES).

Use of sampling weights

The practice varies on whether to use sampling weights or not when detecting outliers in household food consumption data. The International Household Survey Network study, *Detecting outliers in household consumption survey* (Filzmoser, et. al. 2016), hereafter the '2016 IHSN study', refers to Todorov et. al. (2011) on this matter. Todorov recommended that sampling weights be used both for detection and imputation. However, there is no standard recommendation and practices differ.

Per capita versus total

In some cases, it is recommended to detect outliers for quantities and food monetary values expressed in per capita or adult male equivalent. Information on the number of partakers should be used or, if not available, household size, to estimate the number of partakers. This is because a single-person household is likely to have a much lower total consumption level than a large household and there are natural limits on what a person can eat.

Transformation to standard normal distribution

Most robust outlier detection methods assume that data is symmetrically distributed around the mean. However, the distribution of income and consumption variables is commonly skewed to the right (longer tail to the right of the distribution). To address the issue of skewed distribution, the detection methods discussed below require distribution to be normalised. To transform the data to behave like it was generated from a normal distribution, the log-normal or Box–Cox transformations can be used.[17] The basic prerequisite when applying one of these two transformations is that data must be continuous and positive (no negative values). Zeros because of no consumption should not be included.

Choice of the cut-off point

Any distribution has a tail. When defining rules to identify outliers, it is important not to use a rule that would identify an observation as an outlier that is perfectly within the acceptable range of the distribution. For example, if the household consumption follows the normal distribution, the choice of a cut-off point to identify an outlier is crucial. If the cut-off point is based on the distance to the mean, the further the distance, the lower the probability of identifying a consumption as an outlier that indeed belongs to the range of plausible consumption.

Checks for lower outliers

It may also be useful to check for lower outliers in untransformed variables, for instance, looking for values less than the smallest currency unit commonly used, or quantities smaller than the smallest amount of the item that can be purchased. For example, when the lowest currency unit is 5 cents, a reported cost of one cent would be atypical and need to be checked. Likewise, a reported quantity of one gram of rice for example.

2.5 Basic methods for outlier identification

Automatic detection of outliers is a key tool used in Steps 4, 7 and 10 of the process summarized in the flowchart Figure 6. There is a large amount of literature available about outlier detection methods. One particularly relevant for these guidelines is the 2016 IHSN study. This section will limit the discussion to briefly describing approaches and methods commonly used to detect outliers when working with HCES. Methods described are relatively simple and easy to implement, and methods based on more sophisticated algorithms are not necessarily better ones. Regardless of which method is used, it is advised not to follow any algorithm blindly without doing a reality check.

Z-score: The z-score is traditionally defined as the distance away from the mean in standard deviations ($Z=(x-\text{mean})/\text{standard deviation}$). The z-score method assumes that the variable is normally distributed, and the number of observations is large enough to guarantee that both the mean and the standard deviation are good approximations of the real parameters. An observation is often considered an outlier if its z-score is higher than 3 or lower than (−3). At this level, the probability of finding a value with a z-score higher than 3 or lower than (−3) is less than 0.3% (which corresponds to a significant level of 99.7%). Higher or lower conservative values for the threshold can also be used.

Robust z-score: One problem with the z-score is that both the calculation of the mean and the standard deviation are influenced by extreme values that may be outliers. A more robust z-score can be constructed by replacing the mean with the median and the standard deviation with the median absolute deviation (MAD),[18] divided by 0.675 to obtain a consistent estimator of the standard deviation under the assumption of normality. An observation is detected as an outlier when it is outside the range ($\text{median}-c*\text{MAD}/0.675;\ \text{median}+c*\text{MAD}/0.675$). A value

17 The Box–Cox transformation is a parametric family of transformations where each observation xi becomes $(x_i^{\lambda}-1)/\lambda$. The log transformation is a special case of the Box–Cox transformation when $\lambda = 0$.

18 The MAD is defined as the median of the absolute deviations from the data's median (MAD=median(|X-median(X)|). Note that the MAD will be equal to zero if more than 50% of the observations are equal to the same unique value, and it does not work well on some small highly concentrated distributions. For example, in a distribution of 8, 10, 10, 10, 11, both the 8 and the 11 would be identified as outliers.

of c=3[19] is often used. At this threshold, and under the assumption of normality, less than 0.3% of the values are outside the interval. This is a relatively conservative threshold as at c=2.5, the percentage of observations outside the range increases to 1.24%.

Tukey approach: Another robust method to detect outliers is the 'Tukey approach', which measures the distance from the median in reference to the interquartile range (IQR=Q3-Q1) where Q1 and Q3 are respectively the lowest (smallest values) and upper quartile (largest values). Observations that lie outside the range (Q1−c*IQR, Q3+c*IQR), where c=1.5,[20] are often considered as outliers. This is equivalent to a z-score cut-off of +/− (3) (that is a +3 or −3 [21] standard deviation from the mean) under the assumption of normality.

2.6 Imputation to replace outliers

Once an observation has been identified as an outlier, it should not be deleted but rather replaced with estimated values. Two different approaches are usually used to replace outliers:

1) Impute using prices together with information on quantity or monetary value; or
2) Replace using information from the distribution of the variable inspected. That could be using the mean, median or another suitable value, such as the highest or lowest non-outlier value, based on the characteristics of the data. This is known as 'winsorization'.

The choice of the approach depends on the information available.

Approach 1) requires using household-specific information (on either quantity or monetary value) already collected for the food item for which we need to impute and approach 2) uses only the distribution of the variable of interest.

Imputation should be done carefully, as it may introduce bias if not appropriately handled. If the data was normalised using the Box–Cox (or logarithm as special case of Box–Cox) transformation, the reverse of the transformation needs to be applied to go back to the original distribution (for example, using the exponential function in the case of the logarithm).

! Whenever a value is detected as an outlier it should not be included in any aggregate that will be further used for imputation.

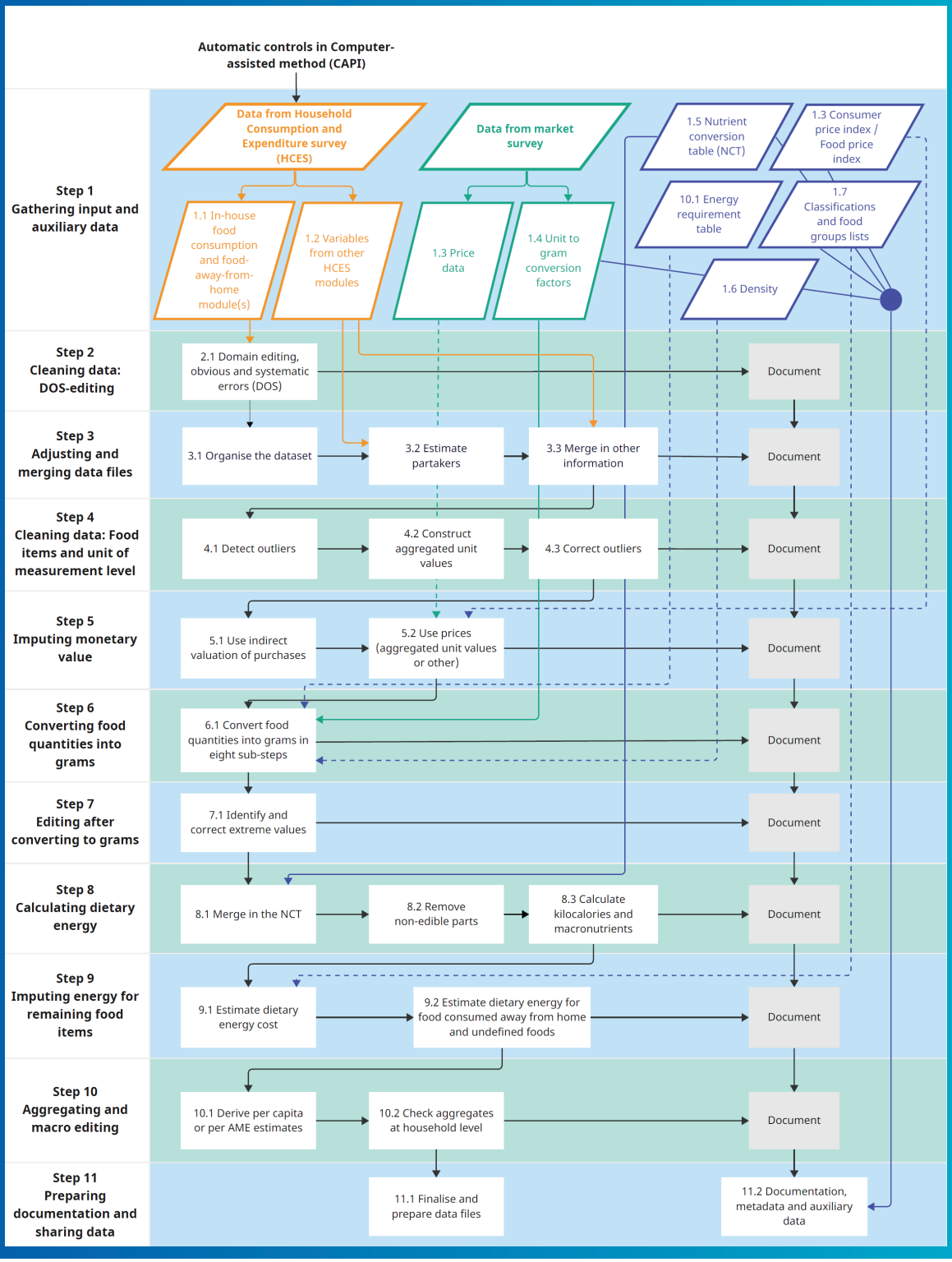

Credit: Unsplash - Will Mayers

19 Which is equivalent to assuming a z-score higher than 3 or lower than −3 under a normal distribution.
20 Note that under the assumption of normality, the IQR is approximatively equal to 1.35 standard deviation, and Q1 and Q3 respectively +0.675 and −0.675 standard deviation.
21 The exact value is 2.7.

PART 2: THE STEP-BY-STEP PROCESS

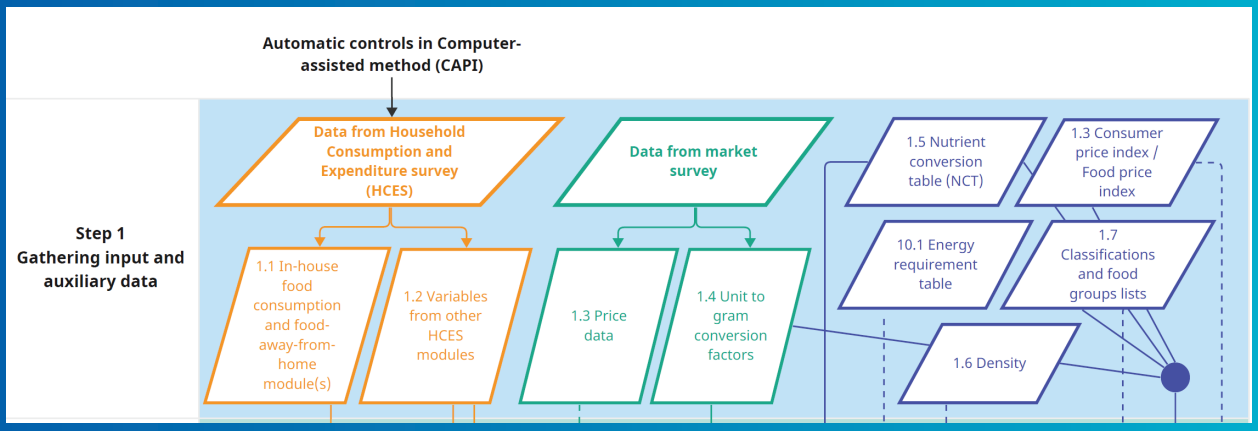
The following flowchart (Figure 6) summarizes all the steps for preparing food data collected in HCES. The numbers refer to sub-chapters in this document, which follow the steps of the process. Dotted lines mean the data may or may not be used, depending on other data and decisions made.

Figure 6. Flow chart on food data processing



STEP 1: GATHERING INPUT AND AUXILIARY DATA

Figure 7: Step 1



The first step represented in the flowchart (Figure 7) is to gather all the relevant data that will be needed at some stage of the process to build the final dataset. This includes the data collected with the household survey food consumption module, some variables from other modules of the survey questionnaire, data collected with the market survey (if available) and auxiliary data from other sources, such as price index or information needed to estimate human energy requirements.[22] It is also recommended to complete the nutrient conversion table (NCT) at this stage. (See further in these guidelines and the companion document[23] for details about how to build the NCT).

It is recommended that the food data be processed by a single team, so that only one file is generated, and the food data is processed only once. This also means that any data collected in the food consumption module that is not mentioned here (if any) should also be processed as part of this work.

Below follows the list of variables needed to process the food data to obtain food monetary values, quantities and dietary energy at the household level.

1.1 Data required from in-house and away-from-home food consumption survey modules

- Unique household identifier (household ID).
- Food item code (each item should be allocated a unique code according to national classification systems, disaggregated enough to allow for food security or nutrition analysis).
- Food item name.
- Quantity and unit of measurement of the food item consumed over the reference period for each source of consumption (purchase, own production, received as a gift, in-kind payment and stocks if collected).
- Number of meals consumed away from home during the same reference period (if this is collected).
- Monetary value of consumption from purchases over the reference period OR quantity, unit of measurement and monetary value over a different/longer reference period OR quantity, unit of measurement and monetary value corresponding to the last purchases (LSMS 2021 guidebook).
- Respondent-estimate of the value of consumption from own production or food received as gift/other sources over the reference period (if collected).

22 This information is needed to estimate adult male equivalencies based on energy requirements.

23 To guide users in the preparation of the survey-specific nutrient conversion table, the authors have drafted a short manual *"Processing food consumption data from HCES - The Nutrient Conversion Table"*. The manual (forthcoming) brings together information about good practices, developed by nutritionists, on building food composition tables/databases and is available upon request.

1.2 Non-food data needed for data processing and collected with other survey modules

- Unique household identifier (household ID).
- Geographic location, all levels (region/province/district; urban/rural area).
- Month and year of the household's interview.
- Household size.
- Number of guests (non-household members) who shared meals with the household during the reference period (if collected).
- Number of household members absent during the reference period of food data collection (if collected).
- Household sampling weight.
- Household non-food expenditure distribution, or any qualitative information collected in the survey related to the welfare level of the household.[24]
- Strata and primary sampling units to build confidence intervals and to be used for imputation.

Additional data needed to estimate the adult male equivalent (AME) based on energy requirements:

  - household member ID;
  - age of each household member;
  - sex of each household member;
  - median height for each sex and age class in the country (if the information is available for the country otherwise the information from a neighbour country can be used); and
  - under-five mortality rates.[25]

1.3 Data on price

Price per one unit of measurement is needed in Step 4 (cleaning data – food item and unit of measurement) and Step 5 (monetary value imputation), while the price of one gram of product is needed in Step 6 (convert food quantities into grams) and Step 7 (editing after converting into grams). The three main sources that inform the price details are:

- market prices collected in a market survey conducted at the same time as the HCES, if they are available and reliable, and if they refer to the price per unit as offered in the market (and not already converted into standard units); and/or
- unit value estimated directly from the HCES; and/or
- ad hoc prices when market prices are not available or unit value cannot be estimated from the survey (ad-hoc prices can be obtained from national experts, the internet or other sources).

See Step 5 (monetary value imputation) for the choice of price and methods of aggregating unit values.

The additional data needed to adjust monetary value for changes in value because of price fluctuation over time is the:[26]

- monthly consumer price index (CPI); and
- monthly food price index (FPI).

24 This information is optional and is mainly needed for some imputation procedures.

25 This information is not from the survey but can be obtained from the United Nations Children's Fund (UNICEF) website: https://data.unicef.org

26 This information is mainly used in the construction of consumption aggregates to ensure data comparability within the country and perhaps also across time periods/survey rounds for the same country. It may also be used to estimate the median price from the survey and further used to impute or correct food monetary value data in cases where inflation is high and the data cannot be cleaned monthly. Therefore, using appropriate CPI/price data is a prerequisite, rather than an option, to construct high-quality food consumption data.

1.4 Data on weight in grams[27] for one non-standard unit of food item

Food quantities are usually collected in both standard and non-standard units (NSUs).

'Standard units' refer to units following the International System of Units (ISUs) and includes all units for which there is an established and agreed international weight equivalent in kilograms (such as grams, pounds, ounces and litres). Other country, market or season-specific units of measurement for weight are considered non-standard units. In these guidelines, the chosen unit of measurement that all other units are transferred into is grams because most food composition tables provide the nutrient value per 100 grams of edible quantity of food item. Information about the gram-equivalent factor is crucial to convert non-standard units of measurement into grams. Common non-standard units are bag, bunch, heap, piece and bucket. Most countries or regions within a country have their own distinctions and the size of a heap in one country or region is not the same as the size of a heap in another. The weight of the same non-standard unit also differs between food items (for example, a heap of spinach has a different weight than a heap of potatoes). Further, the weight in grams for a non-standard unit may fluctuate over time in a given location and should ideally be collected through a market survey in parallel with each HCES survey. The data on weight in grams collected through market surveys should be cleaned and aggregated to an appropriate level. Often the mean or median weight in grams for each item/NSU for each region is used (and season if this is relevant).[28]

As second-best sources, information can also be obtained from other surveys that also use non-standard units, such as data collection for the CPI or an agricultural survey. Weight in grams of one unit of food can also be found on the internet (when the food item and unit refer to a standard product with a standard weight in grams). Some datasets may also provide additional descriptions of the unit used (for example, 'pack of biscuits of 500 grams').

It is important to obtain the weight in grams for as many combinations of food items/NSU as possible. Weight in grams for specific combinations of food items/units may be available for some regions but not others, and in such cases, the national average weight in grams may be used if it can be assured that the weights are likely to be the same. In all cases, it is important to consult local experts to validate the weight in grams for each NSU.

The creation of the database containing the weight in grams of one unit of food item (or conversion factors for NSUs) requires several checks on the quality of these factors (for example, invalid entries that are not possible or do not make sense, such as milk in baskets).

Finding the weight in grams for each combination of food items/units of measurement can be challenging and time-consuming if a thorough market survey has not been undertaken in parallel to the HCES survey. Where this is the case, priority should be given to finding the weight in grams for all the combinations of food items/NSUs widely reported in the survey. As a rule of thumb, weight in grams should be identified for combinations of food items/NSUs used for more than 30 per cent of the quantities collected.[29]

The file containing the information on the weight in grams for all (or as many as possible) combinations of food items/units of measurement should include the following variables.

- Food item code and name of item (same codes and names as those used in the HCES survey).
- Unit of measurement (same code and names as those used in the HCES survey).
- Weight in grams per one NSU for each valid combination food item/NSU reported in the survey.

27 Nutritionists will usually refer to gram-equivalent factors, while other analysts often refer to conversion factors for NSU. For the sake of consistency in these guidelines, we use the term 'weight in grams' to avoid confusion with nutrient-conversion factors that are used to convert quantities into dietary energy.

28 For more detailed information about how to collect and prepare the weight in grams of non-standard units, see the LSMS 2021 guidebook (Oseni et al. 2021) and Oseni et al. 2017.

29 This threshold can be revised upwards based on the number of combinations of food items/units of measurement and the quality of the food data collected.

- Price in local currency units (LCU) of one unit for each valid combination food item/NSU (if collected).
- Geographical level (for example, state/province, urban/rural) to be matched with geographic variables in the HCES data.
- The date when the information was collected, as some units may change size over the year. A 'heap', for example, may be of different size depending on whether the food item is in or out of season.[30] If prices are collected, then information on the date is important to adjust for price fluctuations.

This file will be further merged with the working food data file in Step 6 of the process (convert food quantities into grams).

Figure 8 provides a good example of a market survey collecting information on the weight in grams per unit of measurement of various food items for different geographic areas (extracted from Oseni et al. 2017, Figure 6, p17).

Figure 8: Excerpt from a Conversion Factor Library for Nigeria

NIGERIA GHS-PANEL WAVE 3

CONVERSION FACTORS

ITEM CODE	ITEM NAME	UNIT CODE	UNIT DESCRIPTION	UNIT SIZE	CONVERSION FACTOR (KG)						
					NATIONAL AVERAGE	BY ZONE					
						NORTH CENTRAL	NORTH EAST	NORTH WEST	SOUTH EAST	SOUTH SOUTH	SOUTH WEST
GRAINS AND FLOURS											
10	GUINEA CORN/ SORGHUM	11	Paint rubber		3.612	3.758	3.612	3.612	3.768	3.832	2.828
		12	Milk cup		0.161	0.205	0.125	0.163	0.180	0.161	0.159
		13	Cigarette cup		0.205	0.205	0.205	0.205	0.215	0.198	0.205
		14	Tin		14.738	15.510	14.738	14.738	13.965	14.738	14.738
		20	Congo	Small	1.000	1.280	1.000	1.000	1.000	1.000	.720
		21	Congo	Large	1.978	1.978	1.978	1.978	1.978	1.978	1.978
		30	Mudu	Small	1.073	.978	1.103	1.145	1.073	1.060	1.073
		31	Mudu	Large	1.353	1.368	1.248	1.445	1.353	1.353	1.353
		40	Derica	Small	0.238	0.238	0.238	0.238	0.238	0.138	0.338
		41	Derica	Medium	0.639	0.639	0.612	0.639	0.639	0.639	0.745
		42	Derica	Large	1.587	1.587	1.587	1.587	1.813	1.587	1.361
		43	Derica	Very large	1.889	1.889	1.889	1.880	1.890	1.870	1.925
		51	Tiya	Medium	1.825	1.825	1.825	1.825	1.825	1.825	1.825
		52	Tiya	Large	2.650	2.650	2.650	2.650	2.650	2.650	2.650
		60	Kobiowu	Small	0.595	0.595	0.595	0.595	0.595	0.595	0.595
		61	Kobiowu	Medium	1.110	1.110	1.110	1.110	1.110	1.110	1.110
		62	Kobiowu	Large	1.210	1.210	1.210	1.210	1.210	1.210	1.210
11	MILLET	11	Paint rubber		3.765	3.672	3.765	3.765	3.767	3.805	3.840
		12	Milk cup		0.153	0.153	0.145	0.165	0.150	0.153	0.155
		13	Cigarette cup		0.210	0.210	0.210	0.210	0.215	0.205	0.210
		14	Tin		15.060	15.685	15.060	15.060	14.435	15.060	15.060
		20	Congo	Small	0.924	1.160	0.924	0.924	0.924	0.924	0.688
		21	Congo	Large	1.437	1.437	1.437	1.437	1.437	1.437	1.437
		30	Mudu	Small	0,988	0,893	1,058	1,135	0,988	0,988	0,988
		31	Mudu	Large	1,260	1,260	1,210	1,323	1,260	1,170	1,260
		40	Derica	Small	0.243	0.243	0.243	0.243	0.243	0.145	0.340

SOURCE: World Bank, LSMS Team (Oseni et al 2017, Figure 6, p17).

! Check that all weights are expressed in grams per unit and that they do not refer to a mix of measures, such as litre, kilograms and pounds. If the quantity is given in a volumetric unit, such as litres, then density factors should be applied. This is explained in Step 6 of the process.

30 Some surveys avoid this problem by showing images of the units during the interviews.

1.5 Data from food composition table (FCT) needed to build the survey-specific nutrient conversion table (NCT)

The data needed from external food composition databases or food composition tables (FCDB/FCT) are the nutrient values per 100 edible grams of food for each food item collected in the survey, along with the 'refuse' factors[31] to convert whole foods into edible portions. The information is then used to build nutrient conversion tables (NCTs) that should contain, as a minimum:

- food item code (corresponding to the code used in the survey data);
- refuse factors or edible portions of the food item;
- macronutrient values per 100 edible grams of food – proteins, fats, available carbohydrates (total carbohydrates minus total fibre), total fibre and alcohol – to compute the corresponding number of calories that will be used in the analysis OR the kcal per 100 edible grams if already available in the NCT and estimated following nutritional standards, as explained in the manual on NCT;
- other essential micronutrients per 100 edible grams of foods (vitamins and minerals) – if that is part of the analysis plan – and the survey limitations are well understood.[32]

Users are invited to consult the manual on the creation of NTCs (forthcoming) or visit the FAO International Network of Food Data Systems (INFOODS) website https://www.fao.org/infoods/infoods/en/ before starting with the food data processing. The match between the food item collected in the survey with the reference food composition table is a time-consuming and meticulous exercise that can be better performed with the assistance of experts in nutrition, who have a good knowledge of the types of food consumed in the country.

1.6 Information on density

Density is the gram-equivalent factor of one millilitre of product for all products measured in volume. In this context, it is mainly relevant for liquids or semi-liquids but can be of help to identify non-standard units when the volume of the container is known. This information may already be integrated into the gram-equivalent factor from the market survey or may be part of the FCT. If neither of these is a source, a density database can be found on the FAO / INFOODS website.[33]

31 The 'refuse' factor corresponds to the part of the food that is not edible, such as peel, bone and seeds.
32 Not all HCES collect food data relevant and reliable enough to conduct nutrition analysis. Before using a survey to conduct a nutrition analysis, it is important to ensure that the food data collected meets the minimum of the prerequisite for the analysis. Users can refer to the survey assessment conducted by Smith et al. 2014.
33 The FAO / INFOODS density database version 2.0 (2012) can be found at: https://www.fao.org/3/ap815e/ap815e.pdf

1.7 Food group list

Each food item in the HCES module belongs to a pre-defined food group, such as cereals, fruits or dairy products. Food groups are used in Step 6.2 (convert into grams the quantities not yet converted) and Step 9 of the process (impute energy from remaining food items) and are also helpful when inspecting aggregated results. When it comes to analysis, the allocation of food items in relevant food groups is an important task that should be handled by experts to avoid wrong coding. At this point, the users for economic statistics may have other preferences than the users for nutrition analysis. The international reference classification of household expenditure 'Classification of individual consumption according to purpose' (COICOP)[34] is widely used for economic statistics, while nutritionists may use the system of food classification developed by FAO, the Global Individual Food Consumption Data Tool (GIFT).[35]

There are a few different diet diversity indicators which classify the food items differently. At the processing stage described in these guidelines, it is not critically important which of the standard food groupings are used, but a good NCT should also contain relevant food groups. Finally, it is important to note that the classification of food items into groups widely depends on the number of food items collected in the HCES, their description and their coding. Classification of each food item into different food groups needs to be well documented and quite often it is recommended to include this information in the NCT to keep consistency in the food groups when analysing future surveys.

STEP 2: DATA CLEANING: DOS- EDITING

Figure 9: Step 2

34 COICOP 2018 contains an annex with an optional high-detail structure for food items. https://unstats.un.org/unsd/classifications/unsdclassifications/COICOP_2018_-_pre-edited_white_cover_version_-_2018-12-26.pdf

35 The food items can be grouped, for example, following the classification proposed by FAO nutrition experts who developed the GIFT platform (https://www.fao.org/gift-individual-food-consumption/en), adapted from FoodEx2 classification. FoodEx2 is a comprehensive food classification and description system aimed at covering the need to describe food in data collections across different food safety domains (https://efsa.onlinelibrary.wiley.com/doi/pdf/10.2903/sp.efsa.2015.EN-804).

Domain editing consists of reviewing the data to see if it contains only eligible units and classification variables. Obvious errors are those that are easy to detect and quite often easy to correct, such as an entry error in the unit of measurement of a specific food item (for example, one litre of bread or one meter of rice) or the use of numerical code for missing values (99999). Systematic errors, such as wrong coding (for example, the code corresponding to kilograms is entered instead of grams for a product) that can be detected at this stage should be identified and, if possible, edited. In the glossary of the generic statistical data editing model (GSDEM), this step brings the data from 'raw' to 'edited DOS'.[36]

Whatever method is used to collect data in the field (paper or computer-assisted interviews), checks can be performed on the food data as follows.

2.1 Check for negative or zero values for variables, such as quantities or monetary values

If a zero value does not explicitly mean a '0' consumption (that is, when it is associated with a non-zero quantity or non-zero food expenditure), then it should be replaced by missing values to be imputed at a later stage if appropriate.

2.2. Check the filter questions ('yes' or 'no' to having consumed a food item or anything from the food group), if used

- If a household reports not having consumed a food item ('no' in the filter question) but at the same time reports a quantity or food monetary value, then this observation cannot be dropped. It should be further investigated if this is a true or false report. If the enumerator wrongly entered 'no' instead of 'yes', the observation should be kept. If there was a likely error in reporting, the observation should be dropped, for example, if the enumerator realised the information was entered for the wrong food item and afterwards changed the 'yes' to 'no'.
- Similarly, if a household reports a consumption event but both quantities and food monetary values are left empty, the decision on whether to drop this observation depends on the food item. If it is widely consumed, then it is likely that it corresponds with a true consumption event for which household did not report the information. In such a case, dropping the observation could bias the information on the food consumption pattern of this food item (such as percentage of households consuming the food). It can then be decided to impute the quantities and food monetary values using the approaches further described. If on the other hand, this case happens only once, then it is likely that the household or the enumerator made an error when answering the filter question. In such a case, it can be decided to drop the observation.

2.3 Identify duplicates

Duplicates are more likely to be found when data is collected through a diary. They obviously correspond to a data-entry problem if it has been ascertained that the data does not correspond to a true quantity of the same food consumed several times on the same day. However, duplicates can also be observed if the same questionnaire is scanned twice or if a response is uploaded twice when using a computer-assisted method (CAPI). In all cases, it is recommended to go back to the questionnaire whenever possible.

To identify duplicates, it is advised to look at the number of observations for which there is the same quantity and food monetary value for each combination household/food item/unit/day of the week/source of consumption. If there is more than one observation for this combination, then it is likely it is a duplicate and has to be deleted.

36 See: https://statswiki.unece.org/display/sde/5+SDE+Flow+Models

Credit: Nicole Pankalla - Pixabay

2.4 Check the consistency between total quantity consumed and quantities reported by food source

In case the food consumption module collects information on 'total quantity consumed', which is sometimes reported in addition to the quantities from all sources, it is recommended to check the consistency between the total and the sum of each quantity reported ('total calculated from source') from different sources. If inconsistencies are found, it is important to further investigate their causes and correct information accordingly, if and when possible. Hopefully there should not be many cases, and these could be inspected manually to identify obvious errors.[37]

2.5 Check the validity of the codes applied for the food items and the units

This check is important because, at this stage, issues in the data because of wrong coding of a product or unit of measurement can be detected and corrected. There are two main types of errors.

1. Combinations of food items and units that are not possible (for example, an entry for rice in metres), or other types of errors that might not get caught when checking for outliers at a later stage. These will need to be flagged. It is probably not worth the effort, or in many cases not even possible, to try to 'correct' all these one-offs, which can be treated as missing and their values be imputed at a later stage.
2. Systematic mistakes in coding that affect a significant number of observations, such as mistakes from one team in coding a unit of measurement differently from other teams.

! These mistakes need to be corrected with reference to raw data. This step is crucial in the process. Any systematic or random error not corrected at this stage will be carried through the process and will impact all the estimates, if not detected later in the process.

37 Of note, it can happen that the quantities of the same product purchased, own-produced or received as gift were collected in different units of measurement. In such a case, it is difficult to reconcile all the quantities consumed from each source with the total quantity. Such practice should be avoided, and more restrictions should be imposed on the units of measurement when designing the survey instrument.

Box 6 provides some examples of the types of errors that can be found at this stage of the process.

Box 6: Examples of data entry issues

Below are some examples of errors that can occur during data entry or coding. These cases may refer to errors or implausible consumption that need to be further confirmed with national experts.

- A food item is reported in an invalid unit, such as one litre of potato or one meter of rice.
- Many observations corresponding to a food that is rarely consumed in a country or a region, for example, raspberry in a Sahel country, such as Niger.
- The unit of measurement of a fresh food refers to a processed food, for example, the unit 'can' is associated with 'ocean fish, fresh'. In such cases, this could either correspond to a misallocation of the code ('canned fish' and not 'fresh fish') or a misallocation of a unit ('kg' instead of 'can'). But before correcting, a further check with experts from the country is needed to confirm it is a mistake, not a local unit. For example, tiny tomatoes in Sierra Leone are quite often reported in old tomato-paste tin units.
- Only one type of product exists in a country, but it is coded as a different type in one region or by one enumerator. For example, only one type of rice is consumed in a country ('white rice') and the code allocated to rice by all enumerators is the same except for one that coded it as a different type ('brown rice').
- The quantities reported for a food item in one region are well above the quantities reported in other regions, for example, if the median quantity of a food item is equal to one kg in one region and is one gram in other regions. Again, these extreme cases need to be further checked with experts from the country as they can reflect true disparities in the consumption of specific products in a country.
- A food item usually consumed in very low quantities over a period of seven days is reported many times in an implausibly high unit of measurement. For example, salt is reported in kg for a too-high number of observations. This may not be an error but requires further investigation.
- Inconsistency between the unit of measurement of the total quantities consumed and the unit of measurement of quantities consumed by sources of consumption. For example, a household reports a total quantity of rice consumed in the past seven days of five kg, but three kg of rice is reported as purchased and two cups of rice is reported as received in-kind. In such a case, chances are that the code for 'cup' was not correctly entered and might be referring to kg.
- The quantity and unit of measurement are inverted, for example, if the quantity of cooking oil reported is 500 and the unit code is 'bottle of 500ml'. In such a case, this is analysed as 500 times 500ml, which is 250,000ml, when the correct entry should be either quantity is one and unit is 'bottle of 500ml' or quantity is 500 and unit is 'ml'.

At this stage, automatic methods to detect data entry or coding errors are not efficient, and visual inspection of the data with the use of graphs can help.[38]

Box plots (or whisker plots)[39] and histograms are two types of graphs that can be used to detect errors and outliers.

- Box plots visualise five summary statistics (the median, two hinges that usually correspond to the first and third quartiles of the distribution, and two whiskers that correspond to the thresholds above or below which a point is considered as an outlier) and all 'outlying' points individually. Basically, they are graphs that give a good indication of how the values in the data are spread out. In comparison with other graphical representations, box plots have the advantage of taking up less space, and therefore it is possible to compare distributions of quantities of a food item by unit of measurement or regions (if the number of observations is large enough).
- A histogram or bar chart is a one-dimensional bar plot that provides information about the distribution of the variable. It is used for continuous data, where the bins represent ranges of data. Histograms represent another good way to look at the distribution. Any outstanding value will appear at the upper or lower tail of the distribution. It is beneficial to identify a realistic upper bound on per capita consumption and expenditure of individual food items based on knowledge of consumption habits in the country.

Exploration of the data through tables can also help detect errors. For example, a table can show the correspondence between food items and the unit codes used, and may help identify wrong associations of code with product. Such tables can also be built by area or region, for instance, and allow for identifying a food item that does not match with a unit of measurement.

Box 7 provides two examples of how a box plot can be used to identify issues in the data due to data entry or coding errors.

Documentation

It is important to keep track as best as possible of all the editing performed at this stage and adopt all the following basic principles.

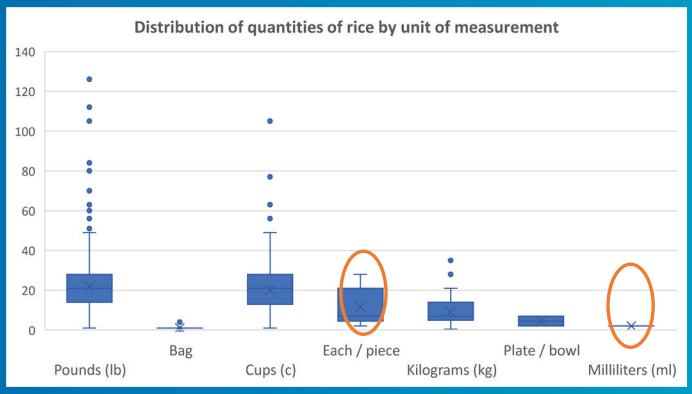
- Do not change the original variables – all edits/revisions should be done on a separate variable.
- Describe the mode of data collection (CAPI, PAPI, self-administered, etc.).
- Describe the method used to check and correct the data.
- Document all the programs developed (if any) to improve data entry.
- Document how many entries were changed.

38 Some of these errors can still be avoided and detected during field work if tables with all valid combinations of food items/units of measurement are developed and part of the computer-assisted method (CAPI) programming.

39 For more information about box plots, see McGill et al. 1978: https://www.jstor.org/stable/2683468

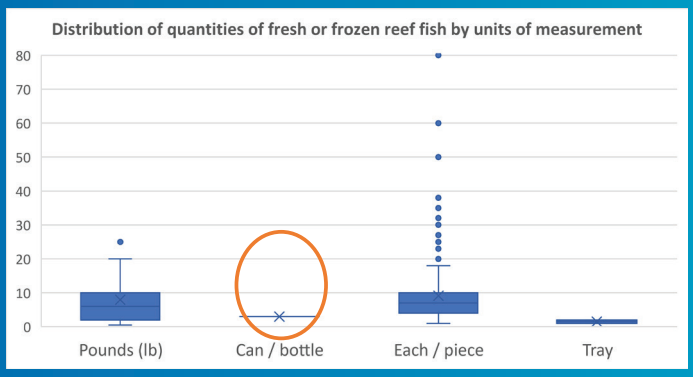
Box 7: Example of the use of the box plot to detect data coding issue

In the first graph, which plots quantity of rice, it can be shown that in urban areas some quantities were reported in millilitres. It is likely that the unit of measurement by which the rice was reported by these households is wrong and needs to be checked. Also, the unit corresponding to 'each/piece' seems suspicious and it may well be that the unit was not properly coded (maybe referring to a bag of rice).



SOURCE: Authors' elaboration from food data collected in the Kiribati 2019/20 HIES. (The survey data can be requested from the Pacific Data Hub, https://pacificdata.org, 'Data Catalogue'.)

In the second graph, which plots the quantity of fresh or frozen reef fish, it seems that the code corresponding with 'can' was entered instead of that corresponding to 'pound' or 'piece'. It is unlikely that 'fresh fish' is available in cans.



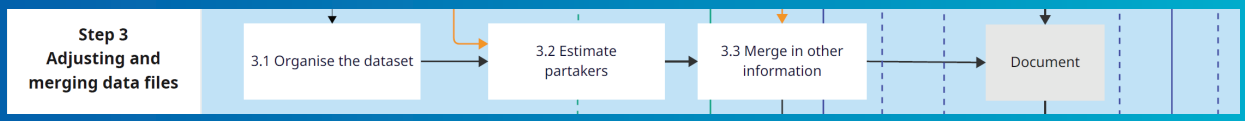
SOURCE: Authors' elaboration from food data collected in the Marshall Islands 2019/20 HIES. (The survey data can be requested from the Pacific Data Hub, https://pacificdata.org, 'Data Catalogue'.)

Of note, box plots were used to illustrate the case but for this specific example, a simple tabulation of units by food item is enough to detect invalid combinations of product/unit.

Unit of measurement	Frequency	Per cent	Cumulative percentage
Pounds (lb)	75	23.22	23.22
Can / bottle	1	0.31	23.53
Each / piece	245	75.85	99.38
Tray	2	0.62	100
Total	323	100	

STEP 3: ADJUSTING AND MERGING DATA FILES

Figure 10: Step 3



The information needed for processing food consumption data is collected and stored in various ways across surveys. It is important that all information needed is identified and prepared for processing, and often it is convenient to merge all into one data file. Step 3 describes ways to do this.

The approach to organising the data files depends on the information collected, the software used and the preferences of the statisticians working on it. Step 3 describes ways to organise the food consumption datasets and then merge them with the other data files that contain variables to be used throughout the process.

3.1 Organise the datasets

Restructuring the dataset

The working dataset can take a wide form (when there is one line/record for each household/food item and as many columns as there are combinations of sources of consumption and variables collected for each source) or a long form (as many rows as there are combinations household/food item/source of consumption). The advantage of the long form is that the number of variables decreases and work can be undertaken on all sources of consumption at once rather than replicating the process for each source of consumption, which is the case with the wide form.

The long form may be more convenient than the wide form when the households report directly on the value of purchases and on an estimate of the value of non-market consumption (see Section 1, example 1). The wide form may be more appropriate than the long form when there is no estimate of the value of non-market consumption (see Section 1, examples 2 and 3). A long form of the working dataset in the latter case would require the creation of additional temporary files that would be used for imputation, while the wide form allows all the calculations to be performed in the same dataset.

Whether long or wide forms are adopted, the process is the same, but it is important to ensure that all information needed for the food data processing is present in the working dataset for each in-house or away-from-home food consumption event:

- household ID;
- food item code;
- food item name;
- quantity consumed;
- unit of measurement;
- source of consumption (purchased and consumed in the house, own production, received as a gift and consumed in the house, food consumed away from home from purchases, food consumed away from home received as a gift, etc.);
- monetary value if an estimate was provided by the respondent, otherwise this will be estimated later; and
- any other variable related to the food item (origin, type, place of acquisition, etc.).

If the long form is adopted for surveys for which the value of non-market consumption needs to be imputed from purchases, it is recommended to create a second dataset that will be further used during the process. The second dataset will undergo Steps 4, 6, 7 and 8 of the process. This dataset does not need to be reshaped but will need to contain information about:

- household ID;
- food item code;
- food item name;
- quantity acquired from purchases;
- unit of measurement; and
- monetary value of the quantities purchased.

If the wide form is adopted, there is no need to reshape the working dataset. Users can go directly to Step 4 related to the first data cleaning at the level of food item/unit of measurement.

Creating the working dataset if long form is adopted

In-house food consumption data

It is important to keep all the consumption information disaggregated by source of food consumption, without aggregation, as the source is an important variable in many analyses. In case the food consumption module collects information on 'total quantity consumed', which is sometimes reported in addition to the quantities from all sources, it is recommended to drop this variable. Before dropping the variable corresponding to total consumption check the consistency between the total quantity consumed and quantities consumed by sources (see point 2.4 above).

The file with in-house consumption can finally be reshaped from wide to long form in such a way that for each household and each food item consumed, and each source of consumption and day of consumption (in the case of a diary), there will be one row. A new variable should be created to identify the source of food consumption and whether food was consumed in the house or outside the house. For example, the new variable 'source' can take the value of 1 for 'purchased and consumed in the house', 2 for 'own production', 3 for 'received as gift and consumed in the house and so forth.

The next example, Figure 11, shows a simplified food consumption dataset before and after it was restructured. In this example, both quantities and values were collected for each source of consumption, in addition to information on units of measurement. There is one line for each food item consumed by a household. The household in the example consumed rice from purchases and gifts. After restructuring, there is only one variable for quantity, unit and value, respectively, and a new variable denoting source has been created. There are now two records for rice, one for each source, that the household consumed from.

Figure 11: Examples of a simplified food consumption dataset

Before restructuring

House-hold ID	Food item name	Food item code	Quanti-ty, pur-chase	Unit, pur-chase	Value, pur-chase	Quanti-ty, own prod.	Unit, own prod.	Value, own prod.	Quanti-ty, gifts	Unit, gifts	Value, gifts
2345	Rice	1	0.5	kg	64				2	Cups	47

After restructuring

Household ID	Food item name	Food item code	Quantity	Unit	Value	Source
2345	Rice	1	0.5	kg	64	Purchase
2345	Rice	1	2	Cups	47	Gifts

SOURCE: Authors' elaboration.

Away-from-home food consumption data

In some surveys, all food data is covered in only one module, for example, food away from home is collected in the in-house section of the questionnaire as 'amount spent in restaurants/ bars' or a listing of prepared foods consumed away from home. In such a case, the information is reshaped together with the other data on in-house food consumption.[40] However, in most recent surveys that follow the IAEG-AG 2018 guidelines, in-house food consumption is collected in one section of the questionnaire and food consumed away from home is collected in a different section, which leads to two different datasets.

In the following, we assume the module on food away-from-home consumption asks each household member to report on the value of meals consumed away from home (and in some cases also the number of meals consumed) by type of meal or meal event, and for the two main sources of consumption (purchases and gifts). The data can be aggregated by meal events and sources of consumption to obtain the total number of meals consumed and respective amount spent by all household members. Depending on the structure of the module (meal type/event in line or in column), the file may need to be reshaped to have one row for each meal event and the number of meals and corresponding monetary value in columns. To keep the information on the meal type or meal event, it is recommended to allocate a code product to each meal following the same structure of the coding system that is being used for in-house food consumption. The file is then reshaped into the same structure as the one for in-house food consumption. The variables 'item code' and 'item name' correspond to the meal type/event (breakfast, lunch, diner, snack, etc.); the variable 'quantity' corresponds to the total number of meals consumed by all household members over the reference period (if this information is collected in the module); the variable 'unit' corresponds to 'meal' and the variable 'monetary value' corresponds to the amount spent on (or estimated value of) the meals consumed by all household members away from home. Before merging the data with the in-house food consumption file, make sure that the reference period is consistent between both files, otherwise, include a variable for the recall period and convert all in-house and away-from-home food consumption on a per day basis.

Once the file with information on food consumed away from home is restructured, a new variable to distinguish between food consumed in the house and food consumed outside the house can be created. This variable can take the same name as the variable created in the in-house dataset if you want to have only one variable for the source of consumption. In such a case, new values can be added, such as 4 for 'food consumed away from home from purchases' and 5 for 'food consumed away from home received as gift'.

Annex 2 provides an example of working datasets adopting the long form.

! Sometimes other food information, such as meals provided at school may be collected in other places in the questionnaire. Make sure to include these in the food consumption data, renaming the variables to be consistent with that in the food data file. When doing this, be very careful not to double count the monetary value of the food. For example, if payment for school meals is specifically recorded in the education module, while the meal is also reported as free food in the food away from home module, then the payment for school meal needs to be dropped from the education module because there is risk of double counting when the value of this meal is imputed.

40 Of note, food away from home, even if collected at individual level, is brought back to the level of the household, and the total amount of energy consumed away from home is divided by all the household members present during the reference period. This is because HCES do not inform on the intra-household distribution of food consumed in the house.

3.2 Estimate partakers

The group of people who consume food in the household during the recall period (known as 'partakers') may not correspond to the household size. For example, one member of the household might be absent during the whole reference period[41] or the household had visitors who ate with them during the reference period. This matters for two main reasons.

1) The identification of outliers and imputation of missing or erroneous values will be more accurate if the quantities and values are expressed in per capita terms based on the number of partakers rather than the number of household members. If a one-person household hosted visitors over the recall period, their consumption might be incorrectly flagged as an outlier if the number of partakers, including visitors, is not accounted for.

2) Food data collected in HCES reflects consumption that occurs over a short period of time while, by construction, some indicators refer to a situation that occurs over a long period of time. This is the case of some of the Sustainable Development Goal (SDG) indicators, such as SDG 2.1.1 (prevalence of undernourishment). To inform these indicators, what is needed is an estimate of the habitual consumption in the population rather than occasional consumption in households. To obtain such an estimate, it is important to account for the exact number of people in the house that consumed the food during the reference period. For example, if one member of the household was absent the whole reference week, then the quantity of food consumed per person in that week will be underestimated if the total amount of food is divided by the size of the whole household. Similarly, if the household had a visitor, the average amount of food consumed by the household will be overestimated if that visitor is not considered.

Of note, in the construction of the welfare aggregate for poverty measurement,[42] food consumed by visitors is considered as a transfer to another household, and NOT included in the hosting household's welfare aggregate. This adjustment is usually made at the last stage of processing food data, once the total value of food consumed has been constructed and before combining it with non-food consumption. Total household consumption is then rescaled by multiplying the total value of food consumption by the number of partakers and divided by the number of household members.

While the importance of accounting for the number of partakers has been widely acknowledged, templates or recommendations on how this information should be collected do not exist. Box 8 attempts to provide some guidance on how to estimate the number of partakers based on information currently collected.

41 If food away from home (FAFH) is collected at the individual level, check if this person's food while away was reported. If not, or FAFH was collected only at the household level, it is safest to assume that their consumption was not captured, and not include them as a partaker. If they are a usual household member, they are included in the count of household members.

42 SDG 1.1.1: Proportion of the population living below the international poverty line by sex, age, employment status and geographical location (urban/rural).

Box 8: Estimation of the number of meal partakers based on information collected in survey design

The approach to count meal partakers differs depending on whether the survey collects information on visitors.

1. If the survey does not collect information about the number of visitors or number of meals consumed by visitors, the number of partakers can be proxied by the number of household members present in the household during the recall or reference period for the food consumption. (If the recall period is seven days, then you need to have the number of household members present or absent in the past seven days). This information is usually collected in the roster on household members. The true amount of food consumed per capita in the household during the reference period may be over-estimated if the number of household members absent is not captured.

2. If the survey collects information about the number of visitors present in the house during the recall or reference period for food consumption, the total number of partakers is proxied by the sum of household members present plus the visitors. It is quite often assumed that visitors consumed the same number of meals as household members and stayed for the entire reference period. If this later assumption does not hold, the per capita food consumption will be underestimated.

3. If the survey collects the number of people who visited the household during the reference period and the number of days they stayed with the household, the number of partakers is then estimated as follows (assuming the visitors shared all the daily meals with the household):

   Number of household members present in the household during the reference period + number of visitors(number of days they stayed with the household/number of days of the reference period)*

   For example, if three visitors stayed for two days with a household composed of four members, over a reference period for food consumption of seven days, the total number of partakers is 4+(3*2)/7=4.85.

4. If the survey collects information on the number of meals consumed by visitors during the reference period (see the 2019 Kiribati HIES), the number of partakers is estimated as follows, assuming the household consumes three meals a day:

   *Household members present in the household during the reference period + number of meals consumed by visitors during the reference period /(3*number of days of the reference period)*

For cases 2, 3 and 4, if the information on visitors is collected by age and gender, the same calculation is performed disaggregating by age class and gender, and what is obtained is then the number of partakers for each combination of age class and gender. This level of disaggregation can be useful in obtaining the number of partakers as 'adult equivalent'.

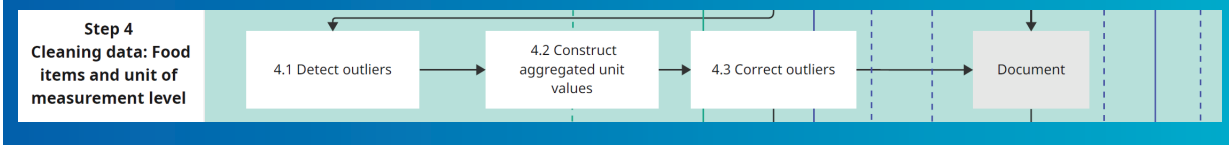
3.3 Merge in other information

External data files can be merged or added to the food consumption dataset when needed or at the beginning of the processing. For ease of illustration, in addition to the food consumption data, the working data file will contain all the variables listed for each household and each food item:

- household size;
- number of partakers;
- size of the household in adult equivalent parameters (described later in this chapter);
- geographic identifiers, from lowest available level (enumeration area (EA)) to highest (region and urban/rural area);
- timestamp or day/week/month and year for data collection (whatever is collected);
- household sampling weights;
- indicator of the welfare level of the household (if available);
- monthly food price index (if relevant);
- dataset with the market price information (if available and of good quality);
- dataset with the weight in grams (or gram-equivalent factors) per one unit of food item from the market survey or ad hoc sources.

STEP 4: CLEANING DATA: FOOD ITEMS AND UNIT OF MEASUREMENT LEVEL

Figure 12: Step 4



The editing in Step 4 is performed for each combination of food item and unit of measurement before data is converted into grams and after data was corrected for errors because of data entry or coding (Step 2).

There are some advantages of editing the data at the level of the unit of measurement prior to its conversion into grams, as follows.

- Distribution at the unit-code level does not suffer from systematic errors that can be introduced when transforming a variable. For example, when a wrong conversion factor is used to convert the quantities from NSU to grams, this introduces an error that would not appear on the non-transformed quantities.
- Extreme quantities can be detected before they are converted into grams and covered up by the other quantities. For example, a quantity corresponding to 21 cups of rice will be detected as an outlier within the distribution of rice reported in cups, but this quantity might not be detected as an outlier once converted into grams and merged with the other quantities in grams.
- Prices are likely to differ across units. For example, food items bought in large quantities are likely to be cheaper than the same foods bought in smaller quantities. Thus, prices estimated at unit-code level may provide better estimates than those at gram level.
- Unit values per unit of measurement can be estimated and used to impute monetary values before food quantities are translated into grams. These monetary values can later be used to impute quantities in grams when conversion factors are missing for some units (which is a common situation). This is further explained in Step 6 of the process.

A challenge is that there may not be enough observations to detect outliers and to impute values for some combinations food item/unit of measurement.

4.1 Detect outliers

Two approaches are used at this stage: the 'multivariate' approach, to check the consistency between the quantities and values when both are given, and the 'univariate' approach to detect abnormal values where only quantity or monetary value are given.

1. **The multivariate approach** should be used where both a quantity and its corresponding value are reported for purchases, or when the respondent provided estimates of the value of own production or food from other sources.

In the multivariate approach, the distribution of quantity, monetary value and unit value are looked at simultaneously. The unit value is constructed as monetary value divided by quantity, and outliers in value, quantity and unit value are identified. Although some quantities, values or unit values will be detected as outliers when looked at independently, the decision on whether to correct a quantity and a monetary value will depend on the position of the other variables within their distribution. This process can be described as a 'consistency check'. For example, a quantity detected as an outlier for a combination food item/unit of measurement will not be corrected if it is associated with a unit value that is not an outlier within its distribution for that combination. In such a case where no inconsistency between quantity and monetary value is evidenced, there is no need to correct the quantity. The decision to correct, or not to correct, a quantity or monetary value is further explained in the decision matrix shown (Table 2).

Table 2: Outlier correction decision matrix

Quantity (total OR quantity per capita)*	Monetary value (total OR per capita)*	Unit value (construct-ed)	Probable cause	Action
Not detected as an outlier	Not detected as an outlier	Not detected as an outlier		No action is needed
Upper outlier	Not detected as an outlier	Lower outlier	Quantity is wrong	Flag and correct the quantity and correct the error
Lower outlier	Not detected as an outlier	Upper outlier		
Not detected as an outlier	Upper outlier	Upper outlier	Monetary value is wrong	Flag and correct the monetary value
Not detected as an outlier	Lower outlier	Lower outlier		
Upper outlier	Upper outlier	Not detected as an outlier	There is consistency between the quantity and monetary value reported – respondent might have reported bulk quantities of food to be consumed over the reference period, for example, to host a party	No action is needed at this stage**
Lower outlier	Lower outlier	Not detected as an outlier	Probably not a problem	
Not detected as an outlier	Not detected as an outlier	Upper outlier	The quantity and monetary value are not outliers with respect to their overall distribution but compared with the quantity reported, the monetary value is too high	Flag, check and correct for obvious errors if possible

Quantity (total OR quantity per capita/AME)*	Monetary value (total OR per capita)*	Unit value (construct-ed)	Probable cause	Action
Not detected as an outlier	Not detected as an outlier	Lower outlier	The quantity and monetary value are not outliers with respect to their overall distribution but compared with the monetary value reported, the quantity is too high	If it is not possible to correct the error, then the decision to impute, or not, the quantity or monetary value will need to be further assessed***
Other combinations (less likely)			Flag	

SOURCE: Authors' elaboration.

* When the multivariate approach is used, the outlier detection can be performed either on the total quantity and monetary-value distribution or on the per-capita quantity and per capita monetary-value distributions. In both cases, the unit value calculated as monetary value, divided by quantity, is the same, and a quantity or a monetary value will be confirmed as an outlier only if the unit value is also detected as an outlier.

** No action is needed at this stage, but quantities and monetary value need to be monitored because they may appear as outliers at a later stage.

*** Decision to correct, or not, the quantity or monetary value needs to be further assessed. This decision can be made, for example, using the distance of the unit value from the median unit value. If, for instance, unit value=4*median unit value, then it may be decided to correct the monetary value. If unit value=0.25*median unit value, then it may be decided to correct the quantity.

2. **The univariate approach** will be used, when only the quantity or the monetary value is given.[43] This covers cases where only the quantity of consumption from purchases is given and indirect valuation of purchases is used, and where only the quantity of consumption from own production and other sources is given if the respondent was not asked for an estimate of value. In the FAFH section, when only the monetary value is recorded, and not the quantity, the univariate approach also applies.

For the variables inspected using the univariate approach, outliers are identified using methods discussed in Part 1. In the case of the univariate approach, outlier detection is performed on the per capita quantity and per capita monetary value distributions.

In both multivariate and univariate approaches, the analysis is done at the level of food item and unit of measurement, which can be further disaggregated by other subgroups if that is deemed appropriate.

If the outlier detection is performed on the quantity and monetary value per capita, it is important to bring back the quantity and monetary value at the level of the household using the same per capita number.

4.2 Construct aggregated unit values

There are a number of factors to consider when aggregating unit values to construct prices to value consumption, as follows.

1) If central tendency is used, instead of winsorizing or using other imputation methods, then which concept of the central tendency should be used? Although the median is often the more robust measure, in some cases the mean of the trimmed distribution is more appropriate.

2) What level of disaggregation should be used? Should prices be constructed at primary sampling unit (PSU) or national level or somewhere in between? Should proximity in time as well as space be considered? Should other household characteristics be considered to reflect that different types of households

43 One exception to this refers to the LSMS 2021 guidebook template that collects only the last purchases. It can be enough to use the univariate approach on unit value for each combination of food item and unit of measurement, if this information was collected only to impute the value corresponding to the quantities consumed.

buy different varieties and qualities of the same food, or face different prices for other reasons?

3) Should other adjustments be made, such as, deflating the unit values temporally using a measure of inflation, such as, the CPI for aggregating?

a) Central tendency

The three main measures of the central tendency are the mean, median and mode. The median is considered a robust estimate of the distribution central tendency because it is not affected by outliers (already flagged observations should be excluded anyway from the calculation). The median is often used in estimation and imputation.

In some cases, the mean can be considered after discarding the outliers or tails of the distribution. The example in the Box 9 illustrates that, if in doubt, it can be wise to check the distribution before deciding on which central tendency measure to use. In the example, the mean proves to be a better estimate than the median unit value.

Box 9: Example of a case where the choice of the mean prevails over the median

There can be instances where the mean is better to use than the median. The example here shows the distribution of unit value in two different regions. In both regions, most of the units are sold at either 50 or 100 values. In this case, the mean better reflects the different prices that are somewhat higher in region 1 than region 2, but not twice as high.

Region	Unit value	Median	Mean after dropping bottom 10% and top 10%
Region 1	2, 50, 50, 50, 50, 50, 100, 100, 100, 150, 1000	50	72.2
Region 2	0.001, 50, 50, 50, 50, 100, 100, 100, 100, 100, 100, 5000	100	83.3

SOURCE: Authors' elaboration.

These types of cases are common in many contexts of cash economies with limited ability to adjust prices by small amounts. For example, in a cash economy, if the smallest bill or coin in circulation is 50 cents, even if the optimal price to charge for an item is 75 cents, workarounds may be found that mean optimal price may be reached by selling the item in groups of three for 200 cents. Prices are often clustered around a few multiples of the smallest denomination of the currency in circulation.

b) Survey weights

Survey weights are often used when calculating the median or mean used for imputation. However, there is no standard practice here (see Part 1, Section 2.4). If the level of aggregation is at or below the survey sampling strata, the use of weights will make little or no difference.

c) Level of disaggregation

The level of disaggregation refers to the layer at which the variable used in the imputation will be estimated. In the previous example, the level of disaggregation is region.

Various levels of disaggregation include geographic levels (PSU/region/area of residence), month of the interview, welfare level (by income quintile, for example) or source of consumption (purchases, own production or received as gift). Which level to choose depends on the expected variation in the variable of interest. The goal is to identify groups of households that are expected to have a similar consumption pattern for the particular food item.

For example, if you need to calculate the median quantities per capita, you may want to select households at a similar welfare level in a similar place as the households you need to impute.

When calculating the median unit value, it is important to particularly consider temporal and/or spatial price fluctuations, as well as expected variability in the quality of the product, to establish a reliable median unit value. The following list shows factors affecting price variations. The presence of any of these factors may decide the level of disaggregation.

- **Temporal and seasonal variations**. Prices change over time, particularly for seasonal food items and in times when inflation is high. To address this issue, median unit values should be based on unit values for households interviewed at the same time.
- **Spatial variations**. Prices are likely to vary across urban/rural areas and geographical locations, mainly because of differences in transportation costs. Imported foods are often cheaper in urban areas than in rural areas where, in turn, locally produced foods are cheaper. Thus, this is an argument for calculating median unit values at disaggregated regional areas.
- **Quality of the food**. Foods that are seemingly the same, such as butter, may be of different quality and thus have different prices. Different prices for the same item in the questionnaire may reflect different quality or just that the item was bought in different shops, charging different prices for the same item. Some surveys ask about where the food item was bought, and this can then be used as a layer for disaggregation when calculating median unit values. Another way to handle this is to disaggregate based on household welfare levels, if that information is available. The hypothesis is that households at similar economic levels shop in similar shops or buy the same quality of food items.

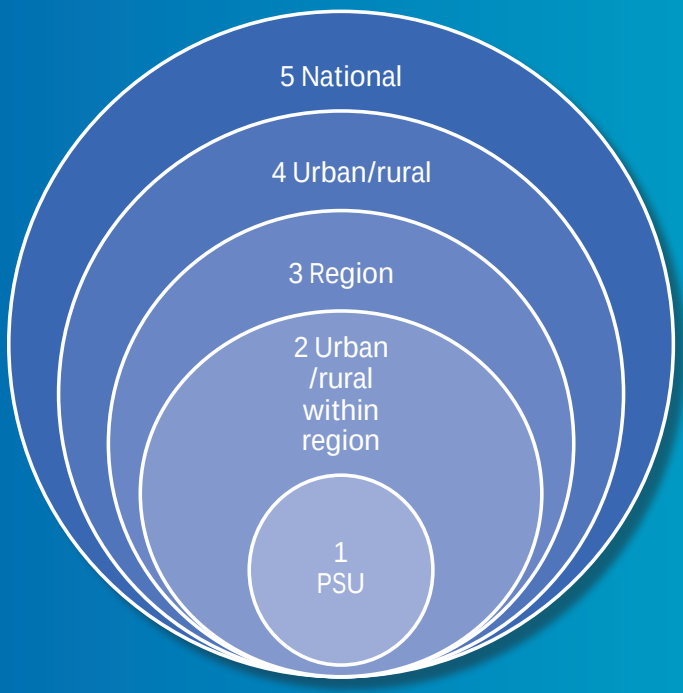
Thus, the layer to choose may depend on the variable (quantity, monetary value or unit value), the geographical, temporal or environmental contexts, and the particular food item of interest. There is no single recommendation on how to balance the tradeoffs, but general practice is to choose the lowest possible level that reaches a minimum number of observations (at least 10 observations).

Note also that any flagged or imputed observations should be excluded from the calculation of the central tendency. It is also acceptable to systematically exclude the top and bottom 1–5% of unit values to be more conservative.

d) Geography-based hierarchy

This section provides an example of how to choose the geographical level at which the variable used in the imputation will be estimated. First, you set the minimum number of observations required. If the minimum is met at the PSU level, use the observations at that level to estimate the median. If it is not enough observations, move one geographical level up and so forth until reaching national level.

Figure 13. Example of geography-based hierarchy



SOURCE: Authors' elaboration.

In the Figure 13 example, PSU is the first level of aggregation (1). Second is urban/rural areas within regions (2). If there are not enough observations at the second level of disaggregation, you move to third. Third is region (3) chosen ahead of urban/rural (4). However, for some countries or some items, urban/rural location has a higher impact on prices than region and it is better to pick urban/rural before regional level of aggregation. The last level to use is all households in the national sample (5) to find the median.

Introducing the 'aspect of time', usually because of expected seasonality or price fluctuation, will add another criterion. In this case that the first level after PSU (households interviewed at the same time) could be region+urban/rural+month, and so forth.

e) Deflating using price indices

To the extent possible, editing takes place on nominal, unadjusted data. However, if the inflation at the time of the survey was particularly high in the country, and there were not enough observations to consider only households interviewed at the same time, food price indices (FPI) may be used to adjust for the variation that might occur in the reported food monetary values. A ratio of the monthly FPI over the annual FPI is applied to the monetary value.

f) Inspect constructed prices

Before using prices (constructed as explained) in imputation, they should be checked for implausible values. The box plot is a good tool to look at the range of prices, especially for the most common item/units. Box plots can also be used to look at the difference in the distribution of prices between rural and urban areas, and across regions of the country. When checking the prices, it is important to take into account elements that may cause real and substantial price differences within a country. These elements include the cost of transport from production to sales points, whether the food item is mainly imported or grown locally, whether different regions are dispersed within the country,[44] whether different regions within a country are trading with different

44 Many small countries in the Pacific are composed of many islands or atolls spread over thousands of km.

countries, and whether there are regional quality differences in products that read the same in the survey. If any implausible values for prices are detected, the analyst should check the underlying data and possibly change the level of disaggregation, as explained earlier.

4.3 Correct outliers

Before correcting any outliers, create a new variable with the original values. Once a quantity or value is detected as an outlier it is replaced as follows.

1. If the univariate approach is used, outliers can be replaced using, for example, an estimate of the central tendency of the distribution, often the median, as this is a robust estimate of the mean and is also easy to calculate (see 4.2 for more details about the use of the central tendency). Outlier value is replaced as:

$$\textit{Corrected quantity per capita (unit)} = \textit{median quantity per capita (unit)}$$

$$\textit{Corrected monetary value per capita (LCU)} = \textit{median monetary value per capita (LCU)}$$

2. If the multivariate approach is used, then quantity and/or monetary value[45] detected as outliers are corrected as follows:

(i) monetary value is detected as an outlier:

$$\textit{Corrected monetary value (LCU)} = \textit{median unit value (LCU per unit)} * \textit{original quantity (unit)}$$

(ii) the quantity is detected as an outlier:

$$\textit{Corrected quantity} = \frac{\textit{original monetary value (LCU)}}{\textit{median unit value (LCU per unit)}}$$

(iii) Both the quantity and the monetary value are detected as an outlier:

In such a case, one variable will be corrected first and the other one will be corrected using the median unit value. For example, the quantity can be corrected using the median quantity and the monetary value corrected using the median quantity, times the median of unit value for that food item reported in that unit of measurement (in that region).

See discussion above on the choice of the level of disaggregation for the median unit value used to replace outlier.

Note that records with corrected quantities or monetary values should not be used in the estimation of unit values that will be further used for imputation.

! When the outlier detection is done on per capita distributions, corrected per capita quantities and per capita monetary values need to be multiplied by the same number used to convert distributions in per capita (partakers, if available or proxy household size).

Documentation

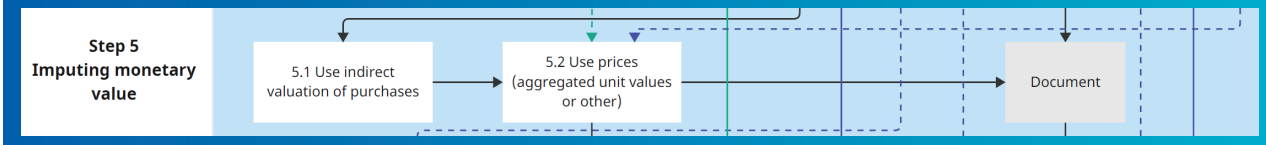
It is important to report on:

- the outlier detection and correction procedure adopted;
- the percentage of quantities or food monetary values corrected at this stage; and
- the number of observations used to set the threshold for the estimation of the median and the level of disaggregation.

45 Referring to per capita or the total household distributions.

STEP 5: IMPUTING MONETARY VALUE

Figure 14: Step 5



Step 5 refers to cases where monetary values for consumption of food were not reported for some sources of consumption. If the household reported directly the value of food consumed from purchases (or the total value of purchases and no information on quantity consumed versus quantity purchased) and an estimate of the value of food from own production and other sources, then this step can be skipped.

In general, it is preferable to impute a monetary value corresponding to the quantities reported in the unit of measurement and, to the extent possible, before quantities are converted into grams. There are two major reasons for this: foods purchased in small quantities are normally more expensive per gram than the same product bought in larger quantities; and the food monetary value can be used at a later stage to convert the quantities into grams when the gram-equivalent factor of one unit of food item is not available. However, if a weight in grams for each combination of food item/unit of measurement is available, and it has been asserted that there is no major difference between the prices of one small and one large unit of product, the imputation of monetary value can be performed after all quantities are converted into grams. In such a case, Step 5 (monetary value imputation) will follow after Step 6 (convert food quantities into grams).

The monetary value is imputed using the quantity consumed and a price as follows:

Monetary value of quantity consumed (LCU) =

*quantity consumed (unit)*price (LCU per unit)*

where the price used for the imputation depends on the information collected.

1. In case the quantity purchased and its corresponding monetary value (from last single purchase or total over a fixed period) are collected, the price refers to the unit value estimated from purchases.

$$\textit{Unit value of purchases (LCU per unit)} = \frac{\textit{Monetary value of purchases (LCU)}}{\textit{Quantity purchased (unit)}}$$

The household unit value can be used when the household reports both consumption and purchase of a food item, the quantities are reported in the same unit, and the unit value for this household was not previously flagged as an outlier (see step 4 above), otherwise an aggregate unit value constructed as explained in section 4.2 can be used. In some (rare) cases, good quality market prices are used as first priority for the imputation (see 1 below). The aggregate unit value is then constructed from observations with both monetary values and quantities for each household and each food item (ignoring any observations flagged as outliers in Step 4). The basic approach is to consider the distribution of unit values for households facing similar markets and use the central tendency of that distribution (see section 4.2)

2. In cases where no purchases occur for a product or information on values is missing or deemed erroneous or unreliable, additional price information will have to be used to value consumption. In such cases there are two main choices for the prices used to value consumption. Which one gives the best-quality result must be assessed for each country.

- Prices from well-undertaken market surveys from the same area and time of household data collection are the first choice. The prices should be available at unit-code level, and should cover, if possible, the same markets used by households in each PSU. In addition, they should be coded the same way as

the food data collected in the HCES. Good market surveys are very costly to implement. They are not always done and their quality is often questionable. When prices from a local market survey are collected, they first need to be edited for potential data entry issues, coding, presence of outliers, too much variability within the distribution or too much heterogeneity in the prices.
Some checks can consist of looking at the standard deviation of the distribution of the price per one unit of product; comparing the central tendency from one region to the other (the same order of magnitude should be expected in the price if the country is not too large or dispersed); and comparing the price of one unit collected in the market survey with that estimated from the HCES data. If the prices from the market survey do not prove to be reliable enough, it is better not to use them as a priority source but rather as a last-resort option when imputing or estimating a value.

- When there are no unit values for a food only consumed from own production or gifts, and market prices were not collected in parallel with the survey. In such cases, you should search for other sources. Sources for 'last-resort prices' could be prices from ad-hoc surveys, such as those collected to estimate CPI and FPI (however, the use of ad-hoc surveys should be limited because they quite often include data collected during a different period than the HCES survey or collected only in central/urban areas).

Box 10: Estimation of monetary value corresponding to rations

In some countries or for some population sub-groups, rations may be an important part of food consumption. If the survey collects information about rations, this is a special case to value. Rations are "the provision of quotas of food items for free or at below-market price". These items need to be included, but it may be difficult to find appropriate prices to use. Options are (in order of preference):

1. if a secondary market for buying and selling these items exists, and there is data on it, then use this value; or
2. use data on items similar to those in the rations but be aware that quality may be systematically different; or
3. ask respondents to self-report what they would pay for such items in a market; or
4. use expert judgement (ask local informants, ration agents); see more in Mancini and Vecchi 2022, p33.

! If the monetary value corresponding to the quantity of non-purchased foods is collected in the survey, it can happen that these values need to be re-estimated. This can be the case if the self-assessment of monetary values is not considered reliable, whereas the food quantities reported are considered as more reliable. In such a case the monetary value needs to be estimated following same process as described earlier in this section.

Documentation

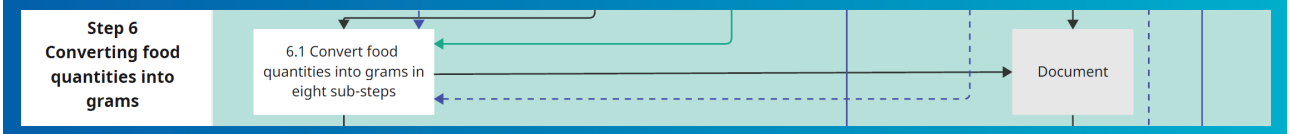
It is important to report on:

- the choice of price or monetary unit value, including sources used as second option when necessary;
- the level of disaggregation used to estimate the median prices, as well as the number of observations on which price estimates were based;
- methods used if there are special cases, such as, valuations of rations or food aid.

The price dataset can be saved separately for documentation purposes (and possibly used to construct price indices later).

STEP 6: CONVERTING FOOD QUANTITIES INTO GRAMS

Figure 15: Step 6



The next step is to transform all food quantities into one comparable standardised unit. This guide uses grams since nutrient values are usually given per 100 grams of food.

There are three possible scenarios for how food quantities can be reported.

1) All quantities are reported in standard units (grams, kilograms, pounds or litres).

2) Quantities are reported in both standard and non-standard units, and weight in grams for each combination of food item and non-standard units is available.

3) Quantities are reported in both standard and non-standard units, and weight in grams is missing for many combinations of food item and non-standard units.

In the first two cases, the conversion of quantities into grams is straightforward. The third case is the most common, and the conversion of quantities into grams involves additional steps that are described in this section.

The information needed for the conversion into grams includes the following.

- Weight in grams for each combination of product and unit of measurement.
- Weight in grams for each food item reported in a volumetric unit (density factor).
- Prices per gram of each food item (estimated from the survey or from a market survey, if available).

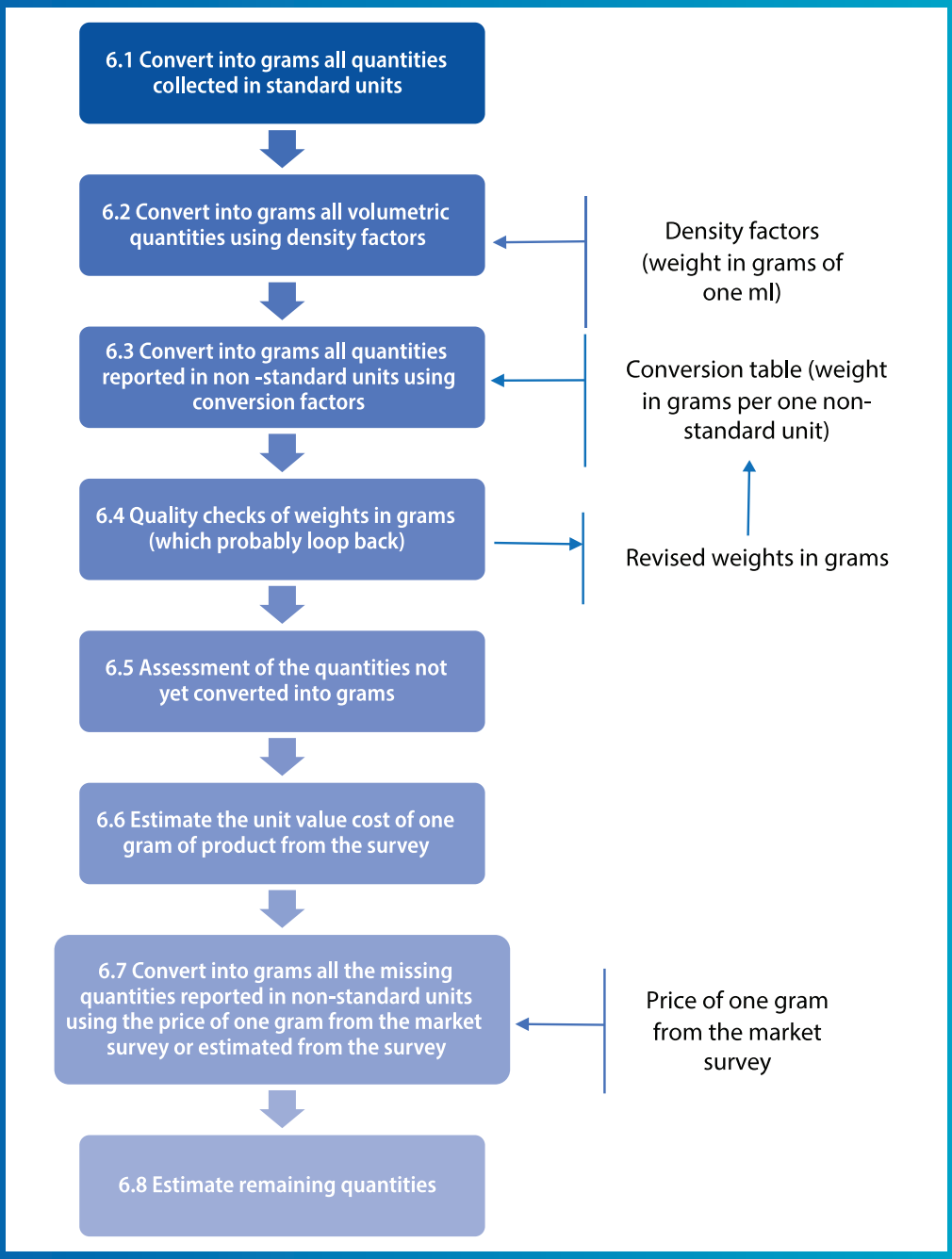
The dataset, with the information on weight per gram of unit of measurement and market price per gram, will be further merged with the food dataset. Merging variables are food item, unit of measurement, geographic location[46] and date,[47] if information is available at these levels.

Figure 16 shows the main steps that can be followed to convert quantities into grams.

46 'Geographic location' means the geographic variable(s) matching one or more location from the survey. It can be the enumeration area (if the market survey was done in connection with the interviews), or (more commonly) a combination of region (state, province) and urban/rural.

47 Date can be quite relevant for some items, if available, when the size of the unit changes between seasons.

Figure 16: Main steps to convert quantities into grams



6.1 Convert into grams all quantities collected in standard units

This conversion is straightforward, and consists of directly applying the universal conversion factor. For example, when a quantity is reported in pounds the conversion is:

*Food quantity (grams) = food quantity (pounds)*453.6 (grams per pound)*

6.2 Convert into grams all quantity measured in volume (for example, expressed in litres)

The density factors are given in grams per millilitres, therefore the volumetric quantity needs to be converted first into milliliters. For example, when a quantity is reported in litres, the conversion is:

*Food quantity (millilitres) = food quantity (litres)*1000 (millilitres per litres)*

Then, convert the millilitres to grams using the density factor (grams/ml) corresponding to the food item.

For example, the quantity in grams corresponding to one litre of product is:

*Food quantity (grams)= food quantity (millilitres)*Density (grams per millilitres)*

It can happen that the same food item is measured in both volume and unit of mass (for example, yogurt can be reported in volumetric units, such as, 6oz or 250ml, and in grams). In such a case, the density applies to quantities reported in volumetric unit while density do not apply for the mass unit.

6.3 Convert non-standard units into grams

If a market survey has been conducted, directly apply the weight in grams per one unit of food item (or conversion factors for NSU) collected in the market survey. If a market survey has not been conducted or conversion factors have not been collected for some combinations item/NSU, use, if possible, information from other secondary sources.

For example, the market survey has concluded that one heap of a specific food item weighs 270 grams. The quantity into grams of one heap of the food item becomes:

*Food quantity (grams)= food quantity (number of heaps)*270 (grams per heap)*

6.4 Check the quality of the weight in grams per NSU

Using a wrong weight in grams can create systematic bias in the data, and it is therefore important to choose reliable values. The survey data may be used to do additional checks on the conversion factors by assessing whether there are significant discrepancies between the results from different areas, regions and markets. This test can be performed in two ways (see examples in Box 11):

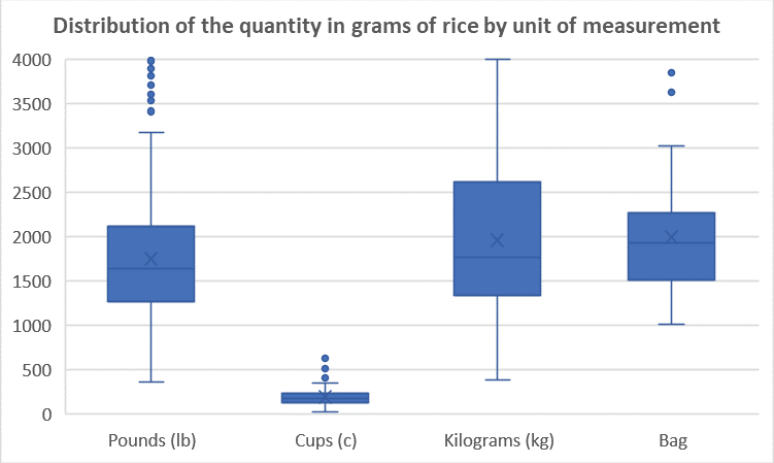
- looking at the distributions of quantity converted into grams, as per Steps 6.1, 6.2 and 6.3, by unit of measurement; or
- looking at the distribution of the unit value of one gram of food item by unit of measurement.

If an error in the gram-equivalent factor is detected, it needs to be flagged, corrected and Step 6.3 re-executed.

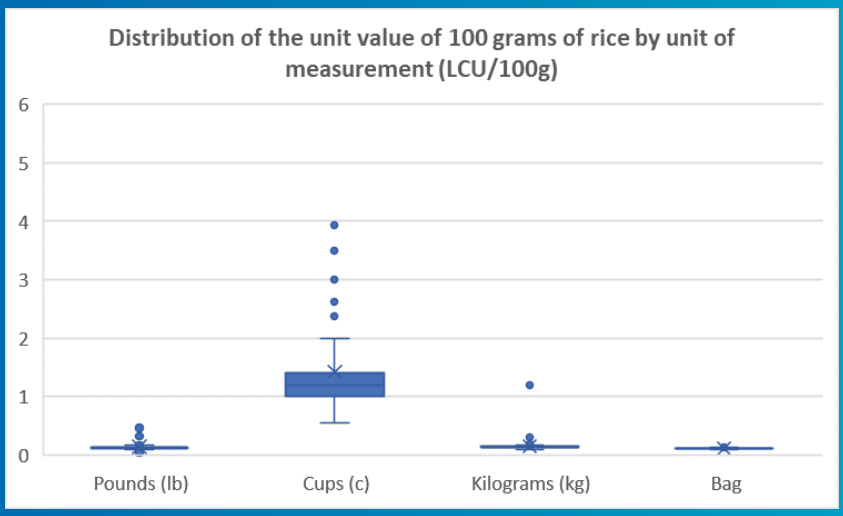

Credit: Firefly

Box 11: Example of detection of wrong weight in grams per NSU

The box plot below represents the distribution of the quantities of rice converted into grams. As can be seen, the distribution of the quantities corresponding to cups is well below that of quantities reported in kg and pounds after converting them into grams. This graph suggests that the gram-equivalent factor used for cups is not correct and should be revised.



The box plot below represents the unit value of 100 grams of rice, estimated as food monetary value divided by the quantity of rice converted into grams. The graph shows that the price per 100 grams of rice, estimated from quantities reported in cups and converted into grams, is systematically much higher than the price in grams of the quantities reported in other units and converted into grams. This box plot is another way of looking at the gram-equivalent issue.



SOURCE: Authors' elaboration using food data collected in the Kiribati 2019/20 HIES. (The survey data can be requested from the Pacific Data Hub, https://pacificdata.org, 'Data Catalogue'.)

6.5 Assess the number of quantities that have not yet been converted into grams

After Steps 6.1 to 6.4. in Figure 16 have been completed, it is important to take stock of the:

- number and percentage of missing quantities in grams due to missing conversion factors;
- number of records with standard units; and
- quantities reported in non-standard units that were converted into grams.

This information will be considered, as follows, when remaining quantities will be converted into grams using prices of one gram of food item.

6.6 Estimate the unit value of one gram of food item

This step is not necessary if good and reliable market prices per gram for all food items reported in the HCES are available from a market survey. If such prices are not available, the aggregated unit value of one gram of food estimated from the survey needs to be used.

$$Household\ unit\ value\ (LCU\ per\ one\ gram) = \frac{Value\ (LCU)}{Quantity\ (gram)}$$

The unit value of one gram of food item is estimated for each household using all the quantities of the product converted into grams as outline in the previous three steps.

The underlying data used to estimate this price differs according to the survey design adopted.

- If the survey does not collect an estimate of value of non-market consumption, the unit value per gram is calculated using the monetary value and the quantities, converted into grams (following Steps 6.1 to 6.5 in Figure 16 from the module collecting information on purchases (for example, food purchased in the past month or the last purchase of food the past month).
- If the survey collects the monetary value of purchases and estimates of value of non-market consumption, then the unit value per gram is estimated using the quantities in grams and the corresponding monetary values from each source of consumption provided there is no major difference in the price of one unit per source of consumption.

The next step is to calculate an estimate of the central tendency of the monetary unit value at the lowest possible level (see earlier discussion under 4.2).

Note that if a large number of quantities could not be converted into grams, the price estimate based on unit values may not be reliable. There is no defined standard on the number of observations, but if the food item is reported by more than 10 households and at least 60 to 70 per cent of the quantities of that food item could be converted into grams, then the estimated unit value of one gram can be considered as relatively reliable. If these conditions are not met, it is important to obtain more weights in grams to increase the number of quantities into grams to be used to estimate the unit value of one gram (quick ad hoc surveys can be conducted by the NSO to obtain the weight in grams of the missing combinations food items/NSUs).

6.7 Convert missing quantities using the price of one gram of product

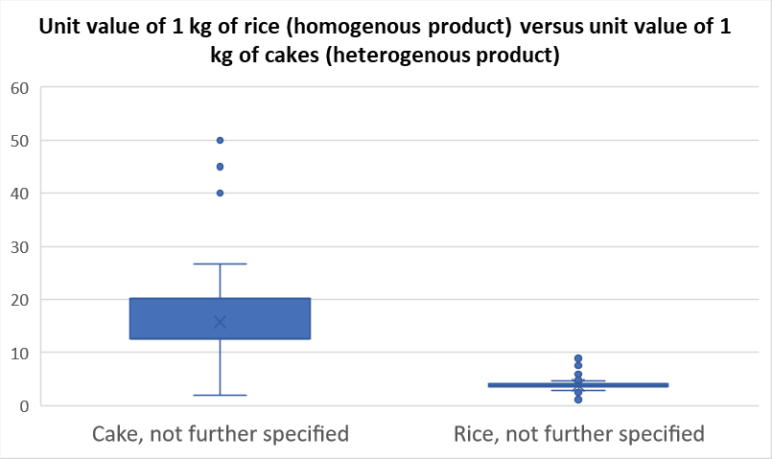
The price per one gram of product is used to convert into grams those quantities that could not be converted because a gram-equivalent factor of one unit was not available. The choice of price to use depends on the quality of the price information at hand, as noted earlier. If market price per gram is available and is assessed as the best source, then it should be used (see discussion on the choice of price in Step 5). If not, the aggregated unit value should be used. The missing quantities in grams can be imputed as follows:

$$Quantity\ (gram) = \frac{Value\ (LCU)}{Median\ price\ (LCU\ per\ one\ gram)}$$

Note that there can be significant variation in the price of one gram of food item when the food item refers to a category rather than a well-defined food (such as 'cakes' or 'cheese'). In such a case, it is recommended to consult with local experts to better understand the type of food it refers to and the unit of measurement in which this food is mainly acquired. (For example, 'one piece' of cake can refer either to a birthday cake or a 'muffin', and the price of one gram of the two can be very different). Box 12 shows an example of differences in price within a heterogenous type of food items (cakes) and a more homogenous food item (rice).

Box 12: Difference in the unit value of heterogenous versus homogenous products

The graph below shows the distribution of the unit value of 1 kg of cake, not further specified (as coded in the HCES) and 1 kg of rice. As can be seen, the unit value of 1 kg of cake is widely spread. This is because this food item refers to different types of foods, of different quality and gram equivalence per one unit of food item.



SOURCE: Authors elaboration using food data collected in the Tonga 2021 HIES. (The survey data can be found in the Pacific Data Hub Microdata Library at: https://microdata.pacificdata.org/index.php/catalog/865)

6.8 Convert into grams the quantities not yet converted

At this stage of the process, there can still be some cases where monetary value and/or quantities in grams are missing.

i) Cases with missing monetary values

This may happen for two reasons. First, in cases where for some combinations food items and unit of measurement we had very few observations and we were not able to clean and impute monetary values for that food in step 4 and 5. If quantities in grams are available, we can impute the monetary values for these items as follows:

Monetary value of the quantity consumed (LCU) = quantity consumed (gram) price (LCU per gram)*

Second, in cases where no price and no weight in grams per unit are available for a given combination of food item and unit of measurement. Without a price the monetary value could not be imputed in Step 5 and without a weight in grams per one unit the quantity could not be converted into grams in Step.[48]

Thus, both monetary value and quantities in grams will be missing, and the consumption event needs to be flagged. There are two options: not to consider this record in the final aggregate; or to impute the monetary value and quantities in grams, based on consumption of the same product by households with similar characteristics. However, it is recommended not to drop the entire observation because it can be used to derive other statistics, such as, the percentage of the household having consumed that food item. These cases are likely to be rare.

48 For example, if five households in a HECS report tomatoes consumed from their own production in 'heap' quantities – and only the price of one gram of tomato is available (not per heap) – it is not possible to estimate the monetary values corresponding to the quantities of tomatoes. Further, if the weight in grams of one heap of tomatoes is not available, it is not possible to convert the quantities reported in heaps to grams and apply the price per gram to impute the monetary values.

ii) Cases with missing quantities in grams

1) Food items referring to a category[49] of homogenous foods (such as, 'other vegetables', which may include items such as cabbage and tomatoes) and for which there is no unit of measurement.

2) Foods corresponding to mixed foods (such as 'other food' or 'a plate of food') and for which estimating the unit value of one gram of product is not possible, or foods belonging to a broad category (such as 'other dairy products') and for which there is no entry in the NCT.

3) Foods referring to 'take away foods consumed in the house' or 'prepared meals procured and consumed away from home' for which the composition is not known and that cannot be quantified.

In the second and third cases, it will not be possible to impute quantities in grams, and dietary energy will be directly imputed in Step 9. In the first case, quantities in grams may be imputed when the food item refers to a broad category of relatively homogenous foods.[50] In those cases, the quantities can be estimated as follows.[51]

- Assign food category to the food item (see Step 1).
- Calculate the household unit value of one gram for that food category by dividing the total monetary value by total quantity. For example, all the foods that belong to the category 'vegetables' will be used to estimate the missing quantities in grams from the foods reported as 'other vegetables'.

$$Household\ unit\ value\ (LCU\ per\ gram)_{hj} = \frac{\sum_{i=1}^{I} Value(LCU)_{ihj}}{\sum_{i=1}^{I} Quantity_{ihj}\ (grams)}$$

where h refers to the household, j refers to the food category (for example, 'vegetables') and i refers to the food item belonging to category j (for example, 'carrots', 'onions', etc.)

- Calculate the median household's unit value per gram by food category using the most appropriate level of disaggregation (as described in section 4.2).
- Impute missing quantities in grams by dividing the monetary value for the food with the missing quantity by the median unit value per gram for the food category the item belongs to.

$$Quantity\ (grams)_{hk} = \frac{Value(LCU)_{hk}}{Median\ unit\ value\ (LCU\ per\ gram)_{j}}$$

where k refers to food item (for which quantities in grams are missing, for example, 'other vegetables'), h refers to household and j refers to food category.

49 Category does not necessary correspond to the food groups defined earlier.

50 For some food consumption, trade analysts may want to present consumption statistics on food quantities in a standard unit, such as grams, for food groups (for example, to assess whether the WHO-recommended level of 400 grams of fruit and vegetable consumption per capita per day is met). For this purpose, it is important to obtain the quantities in grams for the food items corresponding to a broad category made up of relatively homogenous foods such as 'other vegetables' or 'other fruits'.

51 We acknowledge this two-step imputation may induce some slight bias, but this approach may be used to address the need to have the total quantities of the food consumed. In such a case, the imputed quantity of this food item can be added to the total aggregate.

Documentation

The conversion of quantities into grams is an important step in the process, and each decision made, and piece of information used, needs to be well documented. Information to gather and/or prepare is as follows.

- Source of information on the weight in grams of one NSU.
- Source of information on prices used in imputation.
- Summary table showing the price of one gram of food item and the weight in grams for each combination food item/unit of measurement. See Annex 3 for an example of such a table.
- Percentage of quantities collected in non-standard units.
- Information about the combination of food item/unit of measurement for which the unit value of one gram was estimated (number of times the food item is reported, number of observations collected by unit of measurement for this food items, etc.).
- Method developed to convert missing quantities into grams (quantities for which there was no unit of measurement) if different from the one described previously.

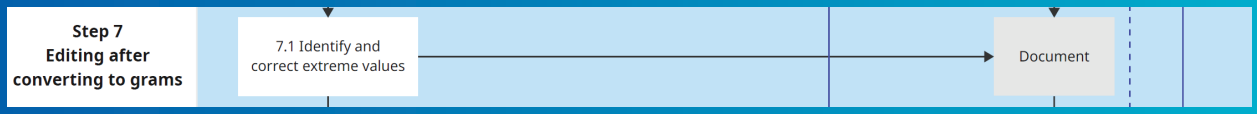
It may also be useful to report the percentage or number of quantities converted at each level of the conversion process. A large share of the quantities converted using the direct method (conversion factors) indicate better quality data.

Remember to keep the flags for all cases where imputation was needed.

Credit: Steve Buissinne-Pixabay

STEP 7: EDITING AFTER CONVERTING TO GRAMS

Figure 17: Step 7



At this stage the data will be cleaned for extreme values. The cleaning is performed on the overall distribution for each food item. The approach adopted differs if an estimated value of non-market consumption is reported, or not, in the survey and for cases where quantities in grams are not available (for example, meals consumed away from home).

A. The survey does not collect the estimated value of non-market consumption

In this case **only**, the distribution of quantities per capita in grams is checked for outliers. If the quantity in grams is detected as an outlier, then both quantities and monetary value need to be corrected. The reason why the monetary value for outliers in this case is not checked, is that it has already been imputed using the quantity variable and unit values in Step 5. If it now turns out that the quantity in grams is an extreme value, you subsequently correct the monetary value using the median unit value per gram and the imputed quantity, so that the quantity and monetary value remain consistent. The prices used in the imputation are the same prices used to impute missing quantities in grams in Step 6, and quantities and monetary values are then corrected as follows.

$$\textit{Corrected quantity per capita (grams)} = \textit{median quantity per capita (grams)}$$

$$\textit{Corrected food monetary value per capita (LCU)} = \textit{corrected quantity per capita (grams)}^{*} \textit{price (LCU per gram)}$$

B. The survey reports on an estimated value of non-market consumption

In this case the distribution of quantities per capita in grams, as well as the distribution of monetary value per capita, are checked for outliers. (Note that the multivariate approach should **not** be used because you are not checking the consistency between quantities and monetary values, rather you want to identify extreme values within each distribution for each food item.)

1. If the quantity in grams is detected as an outlier and the monetary value is not detected as an outlier, then the quantity in grams is corrected using the monetary value, as follows.

$$\textit{Corrected quantity per capita (gram)} = \frac{\textit{monetary value per capita (LCU)}}{\textit{price (LCU per gram)}}$$

2. If the quantity in grams per capita is detected as an outlier and the monetary value is also detected as an outlier, then both quantities and monetary value need to be corrected because they are both confirmed as outliers. In such a case the correction is performed as follows.

$$\textit{Corrected quantity per capita (grams)} = \textit{median quantity per capita (grams)}$$

$$\textit{Corrected food monetary value (LCU)} = \textit{corrected quantity per capita (grams)}^{*} \textit{price (LCU per gram)}$$

3. If the quantity in grams is not detected as an outlier but the monetary value is detected as an outlier, then it is corrected as follows.

$$\textit{Corrected food monetary value (LCU)} = \textit{quantity (grams)}^{*}\textit{price (LCU per gram)}$$

C. Quantities in grams of a food item are not available at all

In this case, the distribution of monetary value of the food item is checked for outliers using the univariate approach.

If a monetary value per capita is detected as an outlier it is replaced with the median of the monetary value of that food item.

Corrected monetary value (LCU) = median monetary value (LCU)

! Data is edited at per capita level, therefore, corrected food monetary value per capita and quantities per capita need to be multiplied by the number of partakers (or the household size as proxy) to also correct the household consumption variables.

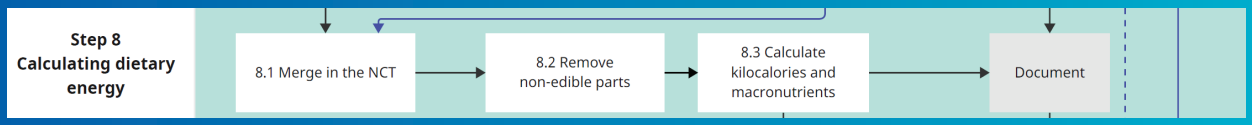
Documentation

As for Steps 2 and 4, it is important to flag all values detected as outliers and corrected. The method used to detect and correct outliers needs to be well documented (in case of the interquartile range specify the multiplier used), as well as the level of disaggregation at which the outlier detection is performed.

The number and percentage of outliers need to be further indicated.

STEP 8: CALCULATING DIETARY ENERGY

Figure 18: Step 8



The dataset now includes cleaned values for quantities in grams and for monetary values, for all households and each food item. Imputations of originally missing or flagged variables have been done to the extent possible. This step describes how to convert the quantities in grams into dietary energy for food items that have information in the nutrient conversion table (NCT). See Annex 4 for an example of of NCTs.


Credit: SPC

8.1 Merge in the nutrient conversion table (NCT)

This is the time to merge the survey-specific NCT file that contains the information on the nutrient content of 100 edible grams of food and the edible portions. This step uses the prepared NCT mentioned in Step 1. and further described in the guidebook about building the NCT.

The data from the NCT is merged by food item (food item code) with the working dataset.

Box 13 below discusses some errors that may occur because the match between the food data collected in the HCES and the NCT was not perfect or correct.

Box 13: Potential errors because of poor food matching

Many errors can occur because of poor food matching or poor preparation of the NCT. Most of them need to be sorted out in the preparation of the NCT. Some issues to watch out for when doing the matching are as follows.

- The refuse factor was not applied to the quantity. For example, the food item 'chicken' reported in the HCES was associated with the food item 'chicken, breast'. This leads to an important difference in the edible quantity of chicken consumed, as almost all of 'chicken breast' is edible while less than 60 per cent of a whole chicken is edible, and an over estimation of the quantity consumed of 40 grams for a quantity reported as 100 grams.
- The food matching was not correctly performed. For example, the food item 'milk powder, whole' reported in the HCES was associated with the food item 'milk, cow powder, skim', and brings a difference of 150 kcal per 100 grams of product. If this product is widely consumed in the country, then the impact on the overall dietary energy consumed can be important.
- The item code in the NCT does not represent the same food item as the code in the survey data. This happens more often when the NCT is originally produced for a different (previous) survey and item codes in the new survey do not match exactly with the item codes used previously. If the data from the previous survey is available, then comparison of kilocalories per food item for both surveys should make these kinds of errors stand out. Manual checks of whether the labels describe the same item are also helpful.

Treating NCT information with care is important because a mistake here will systematically affect the data for every household that consumed the specific food item.

8.2 Remove non-edible parts

The quantity of food in the dataset is reported the way it was bought, taken from own production or received as a gift. However, not all the food in the state acquired is necessarily edible. Some parts of it, such as bones, peels and seeds, need to be removed before being consumed. To obtain the amount of dietary energy available for consumption by the household, the part of the food that is not edible (refuse factor) needs to be removed from the total quantity.

The edible portion or the refuse factor is included in the NCT to adjust the food quantities for non-edible parts. Many foods have a refuse factor equal to zero, which implies that everything is edible (for example, wheat flour, butter or white rice). The refuse factor is given as a proportion ranging from zero to 100 per cent or a number between zero and one. Occasionally, a NCT will provide the reverse of the refuse factor, which is called 'edible portion' (the part of the food that *can* be consumed). Either way, one of the calculations that follows needs to be used to generate the new quantities in edible grams per food item. (In the formulas, the values for the refuse factor and the edible portion are given in per cent).

Edible quantity (grams)= reported quantity (grams)(100-refuse factor/100)*

or

*Edible quantity (grams)= reported quantity (grams) *(edible portion/100)*

8.3 Calculate kilocalories and macronutrients

The NCT provides the calories per 100 edible grams of food item. The dietary energy consumed, in kilocalories (kcal), is estimated as follows:[52]

Dietary energy (kcal)=Edible quantity (grams) kcal per 100 edible grams/100* [53]

52 It is important to ensure that the dietary energy content provided in the NCT has been estimated from the essential macronutrients, and that the method of estimation is well documented. For more information, refer to the manual on the creation of the NCT.

53 The need to divide by 100 is justified by the fact that nutrient content is provided for 100 edible grams of food item.

Similar formulas apply to macronutrients. For example, quantity of fats is estimated as follows:

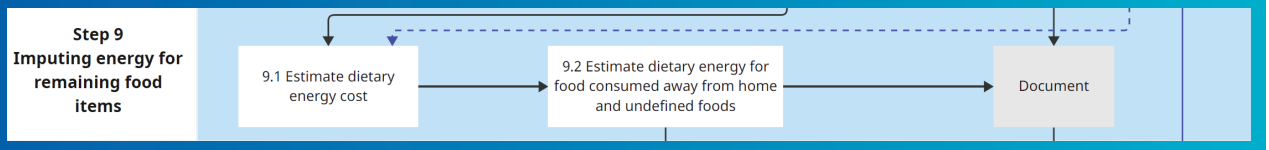
Quantity of fats (grams)=Edible quantity (grams) fat content per 100 edible grams/100*

Documentation

- The NTC, with its relevant documentation (see Annex 4 for an example of an NCT with documentation).
- The number of food items in the survey that did not have corresponding information in the NCT.
- The number of matches, and cases not matched, by item code. The purpose is to see the share that needs to be imputed with other methods than matching with the NCT (see next step).

STEP 9: IMPUTING DIETARY ENERGY FOR REMAINING FOOD ITEMS

Figure 19: Step 9



The remaining records without dietary energy estimates are food items for which quantities could not be converted into grams (see discussion in Step 6 about converting food quantities into grams), and for which there is no nutrient value in the NCT.

For these cases, the dietary energy will be imputed using the monetary value and an estimate of the price per calorie. As mentioned in Step 6, these cases cover two types of food items, which will be further referred to 'unspecified' foods for ease of reading.

1) Food items belonging to a specific food category and for which there are no nutrient values in the NCT. The most common example will be a category such as 'other dairy products' to cover for all the food items in that group that have not been mentioned already. Food items reported here might have very different nutrient content per gram (for example, cheese and liquid milk).

2) Meals cooked outside the home that are difficult to quantify, for which the composition is not known and there are no nutrient values in the NCT. These foods quite often refer to food consumed away from home.

In both cases, the dietary energy consumption will be estimated by applying the cost of one calorie (referred to as 'dietary energy unit cost') to the monetary value of the unspecified foods. The dietary energy unit cost of unspecified foods is proxied by a dietary energy unit cost of similar products reported in the survey, and for which a dietary energy value is available (referred to later as 'well-defined foods'). This approach assumes that the cost of one calorie from unspecified foods is similar to the cost of one calorie from a basket of well-defined foods consumed.

9.1 Estimate dietary energy unit cost

The dietary energy unit cost is calculated by dividing the total monetary values from well-defined foods by the total calories consumed from the same foods. In other words, it is the weighted average of the dietary energy cost of each food consumed, weighted by its respective share of dietary energy in the total dietary energy consumed.

When estimating the dietary energy cost needed to impute the dietary energy from unspecified foods of Type 2 (meals cooked outside the home that are difficult to quantify), it may be advisable to consider only foods purchased. This is because purchased foods may be more representative of the composition of meals consumed in food establishments than foods consumed from own production. It is also recommended **not** to include entries that might have been imputed at some step of the process, for example, observations detected as an outlier or missing monetary values.

Tips: Before using the dietary energy cost to impute the missing dietary energy, you may wish to perform an outlier detection on the dietary energy cost distribution using the univariate approach. This will allow you to identify some potential issues in the data that outlier detection methods described earlier failed to detect.

Adjustments

The use of the dietary energy cost approach to estimate the dietary energy of composite meals, and food away from home, assumes that the cost of one calorie consumed in the house is equal to the cost of one calorie consumed away from home. To account for the additional costs to run a food business, a multiplier can be used to adjust the median dietary energy unit cost. The FAO does not adjust the cost, but accounts for differences in the cost of a food basket by income group. (See Box 14 on the rationale to account for the welfare of the household when building the median dietary energy cost).

Box 14: Accounting for the welfare level of the household when estimating the median dietary energy unit cost

The dietary energy unit cost used to impute missing dietary energy is estimated from the values reported by a group of households similar to the one under observation (including the household itself).

The most similar households are usually considered those living in the same areas. Using the same stages for choosing households, this implies: PSU, region, urban/rural and time of year (season). In addition, the dietary energy unit cost is often affected by the household's economic status, for example, rich households may eat in more expensive restaurants than poor people. The assumption is usually that poor households buy cheap, energy-dense foods, while rich households often buy more expensive (and sometimes healthier) foods. However, this may vary between food items because rich households may also buy cheap food in bulk quantities using credit/debit cards and keep it in big freezers or other storage areas.

To impute the dietary energy for poorly defined food items, the FAO uses the 'expenditure quintiles' method. The rationale behind this choice is to identify the 'typical' cost per calorie, based on the assumption that households belonging to the same expenditure quintiles face similar costs (i.e. poor households have a lower dietary energy unit cost than wealthier households). A typical way to rank households according to wealth is in quintiles in line with the distribution of total consumption expenditure per capita (consumption expenditure aggregate). At this stage, the consumption expenditure aggregate is probably not available. Thus, for the sake of classifying households as poor or rich, two approaches can be adopted:

1) use a proxy of total consumption expenditure estimated as the sum of preliminary estimates of non-food and food expenditures; or
2) use a welfare indicator derived from the survey, if available (for example, the multidimensional poverty indicator).

Once households are categorised based on their welfare level, the median dietary energy unit cost can be estimated for each welfare level and the most appropriate level of disaggregation applied.

9.2 Estimate dietary energy for food consumed away from home and undefined foods

The following describes the approaches to handle the imputation for the two types of food items for which calories are missing.

Type 1: If the food item belongs to a specific category, such as 'dairy products', apply the median dietary energy cost of the relevant food category to impute missing dietary energy,[54] and do as follows.

- Assign a food category to the 'unspecified' foods (for example 'other dairy products' belongs to 'dairy products') (see Step 1).
- Calculate the household dietary energy unit cost for that food category by dividing the total monetary value by total calories (using only observations with non-missing on monetary values and calories).

$$Dietary\ energy\ unit\ cost\ (LCU\ per\ one\ kcal)_{hj} = \frac{\sum_{i=1}^{I} Value\ (LCU)_{ihj}}{\sum_{i=1}^{I} DEC_{ihj}\ (kcal)}$$

where h refers to the household, j refers to the food category, i refers to the foods belonging to category j (for example, milk, cheddar, etc.) and DEC refers to the dietary energy of food i expressed in kcal.

- Calculate the median dietary energy unit cost for that food category at the most appropriate level of disaggregation.
- Impute the missing dietary energy by dividing the food monetary values for the unspecified food (for example, 'other dairy products') by the median dietary energy unit cost for the corresponding food category (for example, 'dairy products').
-

$$DEC\ (kcal)_{hk} = \frac{Value(LCU)_{hk}}{Median\ dietary\ energy\ unit\ cost\ (LCU\ per\ kcal)_{j}}$$

where k refers to the unspecified foods (for which calories are missing), h refers to household, and j refers to the food category to which the unspecified foods belongs to.

Type 2: If the food item is a mix of several food groups (composite foods or meals consumed away from home), apply the median dietary energy unit cost estimated on all food items and not just a category of foods. In such a case, the same steps as for Type 1 are followed with the exception that all the foods purchased by the household, and for which the dietary energy is available, are used in the estimation of the household average dietary energy cost.

Meals consumed away from home by meal event or type of meals

It is worth noting that if the survey collects information about meals consumed away from home by meal event, such as snack or type of meal and including non-alcoholic drinks, the dietary energy consumed from these meals can be estimated using the same approach as that used in the case of foods belonging to a food group category. In the case of non-alcoholic beverages, the dietary energy unit cost is estimated using information about dietary energy from all non-alcoholic beverages reported in the survey.

Inclusion, or not, of drinks in meals consumed away from home

It may not be known if drinks are included in the meals consumed away from home, particularly in restaurants. Some drinks that are more expensive than others, such as alcoholic beverages, may or may not be included in the meals consumed away from home. In such cases, it may be best to estimate the in-house dietary energy cost including and excluding expensive drinks, and to use these two costs to estimate the dietary energy consumed away from home. If there is a major difference between the estimated dietary energy using one cost and the other, then further investigation on whether or not it is common in the country (or for specific population groups) to consume expensive drinks during meals.

! The dietary energy unit cost approach cannot be used to estimate the dietary energy of foods with no energy (such as 'bottle of water consumed away from home'), and for these foods the nutrient value should be set to zero.

54 The same approach can also be used to estimate quantities of macronutrients.

Free food

If there is a system of free (school) meals, the provider will often know the nutritional value and the cost of an average meal. In such cases, it is recommended to use this information directly when estimating the dietary energy from school meals and the associated cost. This is done by multiplying the number of school meals received by the nutrient content and the cost of an average meal.

Outliers

At this point, it can save time to check for extreme outliers in per-capita consumption, measured in dietary energy of individual food items and distinguishing between in-house and away-from-home food consumption. It makes the most sense to look for outliers in dietary energy consumption when aggregating all food items, which will come in the next step, but correction happens on food-item level and some may be picked up here. If a household, for example, dines at the most expensive restaurant in town, then imputing the dietary energy based on what they paid will generate a very high amount of dietary energy.

Limitations

There is no standard approach to estimate the amount of dietary energy consumed away from home, and the approach mainly depends on the information collected.[55] To date, the use of the dietary energy cost remains the most widely used approach in the absence of a better option. It is widely acknowledged that further research in this area is needed.

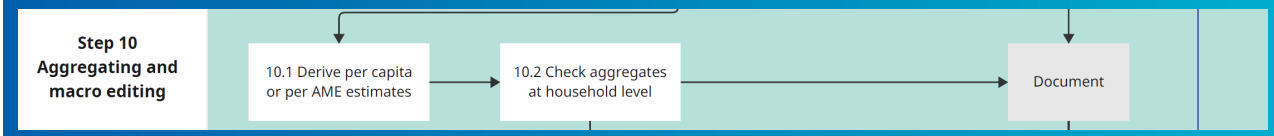
Documentation

- The source and method used to estimate the dietary energy unit cost.
- The choice of level of disaggregation for imputation.
- Information about whether different approaches were used to estimate the dietary energy from specific food-away-from-home components (school meals, snacks, non-alcoholic beverages, etc.).
- Information about whether an adjustment factor was applied to the dietary energy cost, what the rationale was behind this choice, and how it was estimated.
- Information about whether alcoholic beverages were included, or not, when estimating the in-house dietary energy unit cost.
- Information about whether food purchases only, or all foods consumed by the households, were used to impute food away from home.

55 The calculation of dietary energy from food away from home is linked with the challenges on how to collect the data, as mentioned in the IAEG-AG 2018 guidelines. The Samoa Bureau of Statistics, in collaboration with the Pacific Community, has conducted a survey experiment in Samoa to assess the difference in the cost of one kcal consumed in the house and one kcal consumed away from home. The benchmark being used for information about meals consumed away from home is collected through a highly monitored individual diary and also visual aids.

STEP 10: AGGREGATING AND MACRO EDITING

Figure 20: Step 10



At this stage, for each household and for each food item and source, the dataset contains the dietary energy and corresponding food monetary value. These values can be aggregated to obtain the total amount of dietary energy consumed and the corresponding total food monetary value for each household.

10.1 Derive per capita or per adult equivalent estimates

In order to derive estimates of dietary energy consumption per capita, total calories consumed at the household level have to be divided by the number of people who consumed them. Such a number can be one of four different types.

1) Simple counts (number of person);

2) Adult equivalent or adult male equivalent (if age and sex is provided for all, including guests);[56]

3) Person-meals (assuming, for example, three meals per person per day); or

4) Adult–male–equivalent meals (AME–meals).

The last option, if feasible, is the recommended one.

The estimation in 'adult equivalent' differs on whether the adult equivalent concept is based on dietary energy requirement (in such case we refer to adult male equivalent) or on food and non-foods requirements (economic concept).

Box 15 explains the difference between the two concepts and outlines the information needed.

The use of an adult equivalency scale requires information on age and gender for all the people who consumed the food during the reference period.


Credit: SPC

56 To estimate the adult male equivalent (AME), the normative average dietary energy requirements are estimated for each household member following the WHO/FAO/UNU 2004 recommendations. The average requirement of a male adult is then used as reference, and the AME is then estimated as the ratio of the average requirements of each household member divided by the average requirements of an adult male.

Box 15: Conversion of the number of partakers in adult equivalent

Some analysts may be interested in expressing the dietary energy or food monetary values in per adult equivalent (AE) rather than per capita. Adult equivalency scales are based on the human dietary energy requirements or human economic costs.

Which adult equivalent scale to use depends on the analysis to be conducted.

To conduct **food consumption analysis**: the adult male equivalent (AME) is used, based on the different energy needs of people. To estimate the dietary energy requirements, it is recommended to use age, sex and height of the individuals, along with some specific reference tables on height or height-for-age for children under five years.[57]

For welfare or **poverty** analysis, the equivalence scales used have a different logic. All adult men are assumed to need the same amount of food, so if a man moves into a household that has just one other person, the dietary energy needs of the household increase[58] – and so would the AME. Clothing budget and personal hygiene may also double, but the expenses for rent, water, furniture, household equipment and so forth may stay almost the same. This is economies of scale. The modified Organisation for Economic Co-operation and Development (OECD) scale is, for example, used by Eurostat (the statistical office of the European Union) to measure living conditions in Europe.[59] (See the Mancini and Vecchi 2022[60] chapter about adult equivalent scales, where they state there is no scientific consensus on which equivalence scale is best to use.)

Finally, to obtain the per-capita daily consumption, the per-household, per-capita values are divided by the number of days of the survey reference period. Attention should be paid to diaries, for which each entry usually corresponds to a daily consumption, but the sum of the consumption per household should be divided by the number of days the diary was undertaken. For example, if a household received two seven-day diaries but filled out only one, then the total consumption reported should be divided by seven.

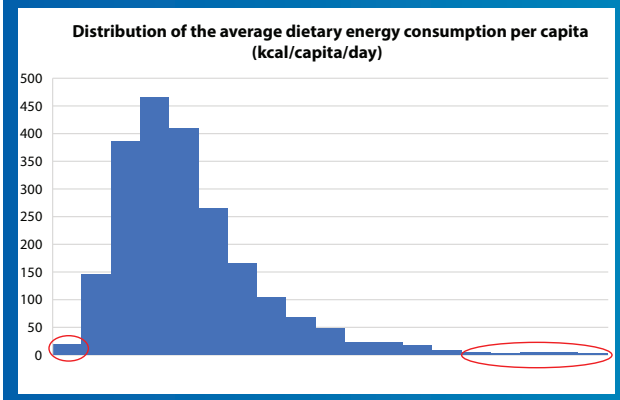
10.2 Check aggregates at household level

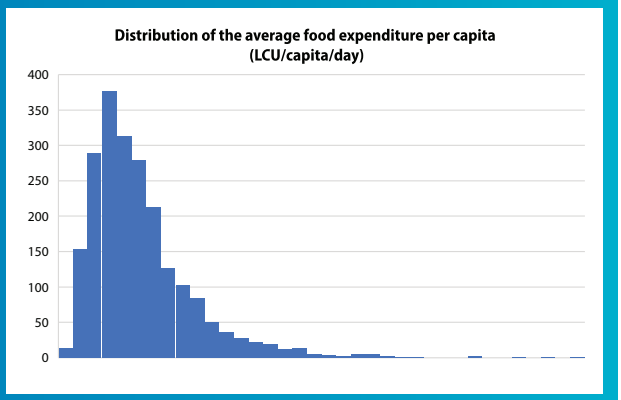
At this level, the data is aggregated, and for each household there is only one record corresponding to the average daily dietary energy consumption (DEC) and food monetary value. Both distributions are expected to be log-normal, but the tail of the distribution in the case of the DEC is expected to be shorter than that of the food monetary value (see the next graph showing the distributions of DEC and food monetary value). Even if the average per-capita dietary energy consumption estimate reflects the amount of dietary energy available for consumption by the household over a certain reference period, there is a realistic limit to the amount of dietary energy that can be absorbed by the body. Therefore, the average dietary energy consumption cannot be as skewed as the food monetary value.

A final check needs to be made on the DEC and the food monetary value distributions using either expert judgment or a robust outlier-detection method, such as those described in section 2.3. Abnormal values will need to be flagged and underlying data leading to the aggregate will need to be checked back and corrected according to the methodology described earlier.

57 See FAO 2004. https://www.fao.org/3/y5686e/y5686e00.htm
58 If they have different height and weight, and that data is available, this is not exactly the same.
59 See: https://www.oecd.org/els/soc/2.2b%20Eurostat-EUSILC-Comparability.pdf
60 Mancini and Vecchi 2022 is a follow-up/update of the Deaton and Zaidi 2002 World Bank report, which has been a core reference point for constructing consumption aggregates over the past two decades, and why it is being quoted in this guideline.

Figure 21: Examples of distribution of average dietary energy consumption per capita and average food monetary value per capita





SOURCE: Authors' elaborations.

Some reasons why low- or high-DEC might be observed are listed as follows. In most, cases they are the consequence of a poor, earlier cleaning of the data.

There are many reasons why too high values of DEC may be observed.

- The main reason why high-DEC would not be discovered earlier in the process is if a household reported many food items consumed in quantities slightly lower than the upper fence of the box plot. Therefore, they are not detected as outliers, but once aggregated lead to an implausibly high value of DEC. For example, one household consumed 1500 kcal of rice per day, per capita and 1500 kcal of taro per day, per capita, which aggregated the sum to 3000 kcal per capita, per day. When taken individually, these values were not detected as outliers because the upper fence of the box plot was 1550 kcal per capita per day, for instance.
- The estimation of dietary energy consumed away from home, unless checked in the previous step, might have been over-estimated because of the use of a too-low cost for one kcal consumed in the household or high expenditures reported when consuming in high-standing restaurants.
- None of the outlier detection procedures could be applied because the number of observations were too few. For example, if two households in the entire sample reported consumption of caviar and one of the quantities reported was 10 kg.
- Systematic errors in the data. For example, wrong coding, wrong unit of measurement, wrong gram-equivalent factors in grams of one unit. If good work was performed during Step 2 of the process ('first cleaning – DOS editing'), then such errors should not appear at this stage. If, however, such errors appear, then the prior editing of the data was probably not efficient and it is important to go back to the raw data.
- A wrong price was used when converting some quantities into grams. For example, a food item referred to a very expensive imported food while the unit value used to convert the quantity reported in non-standard units into grams was the local price.
- Some households might have reported consuming quantities of food items that were not meant to be consumed by the household. For example, a proportion of the quantities of coconuts reported was meant to be fed to pigs or a proportion of quantities of rice was used to feed livestock. If this is known as a common practice in the country, the over reporting of the quantities can be corrected after consulting experts from the country (for instance, if it is known that on average 25 per cent of the rice acquired is given to livestock).
- A significant quantity of food was wasted during storage or cooking.

- Some visitors were not accounted for in the total of partakers.

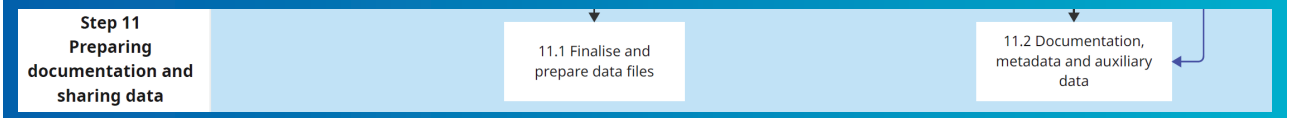
Low DEC values may be because of the following reasons.

- Under reporting. For example, in a diary collecting food consumption, not food acquisition, a household reported only two or three food items that were very low in dietary energy content, (such as salt, water, fresh skimmed milk or spices) over the entire reference period.
- Households did not fill in the in-house and/or away-from-home food consumption sections of the questionnaire.
- Households reported only the consumption of the first food item in the recall list and stopped reporting on the consumption of the other food items after realising that saying "yes" to the filter question was leading to too many follow up questions. These households in theory should have been dropped from the raw data from the beginning.
- Fatigue of diary respondents. For example, respondents were asked to fill in a diary each day for two weeks but they filled it in only for the first week. In such cases, it is recommended to consider the first week of the diary for these households.
- Wrong number of partakers. The total amount of food was divided by a number of people higher than the number of people who actually consumed the food. In such cases, you may need to go back to the questionnaire to check if all household members were truly present in the household during the reference period.

! As a rule, adjustment to aggregates should not be performed, and any gross outliers need to be traced back to determine the source and adjusted at the item level. It is not recommended to drop values, but to go back to the data and identify the food quantities or monetary values that led to excessive consumption and correct them if the earlier outlier detection procedure did not prove efficient enough. However, in the case of low-DEC, it may be needed to drop some households and to re-calculate the entire household sampling weights. A decision to drop households needs to be part of a consultation process and cannot be taken unilaterally.

STEP 11: PREPARING DOCUMENTATION AND SHARING DATA

Figure 22: Step 11



Documenting all the steps of the food data processing is key to establishing trust in the data and explain decisions made during the process. Further, documentation allows future users to decide whether or not to apply additional steps to the process, or, for example, to use a different approach when cleaning the data, depending on the quality of the data, survey design and analysis to be conducted.

Data cleaning processes can, for example, strongly affect the overall distribution of original quantities and monetary values reported. A non-transparent cleaning process can lead to uncertainty about the data and lack of trust, even if the process followed strict and robust approaches.

An important motivation for this guideline point is to ensure that each step in the process is well documented.

Syntax files, metadata and basic information

The most detailed insight into the methods used will be the syntax files developed, which should include proper narrative documentation of decisions made throughout the process. Users should not need to go back to syntax files to check how the data was processed. Therefore, it is important to document beyond syntax files.

Suggestions on what to document are included in each step and briefly summarised as follows.

- Document the survey modules, mode of data collection and limitation in the data. For example, what was missing to conduct proper analysis?
- Document the method(s) used to identify the outliers and relevant thresholds applied, and the method(s) used to correct outliers and order of imputation (quantity first, food monetary value after, and both quantities and food monetary value, etc.).
- Document the choice of monetary value for imputation (including the data file, if it comes from an outside the survey).
- Whether price indexes have been used in the process, and if they were, how they were produced and for which geographical areas.

In addition to the above, it is recommended to attach all the files with auxiliary data such as the nutrition conversion table and weight in grams per one (non-standard) unit of food item, including the metadata on how these were produced. This needs to be made available to analysts of the survey and to also process future surveys.

It is also important to document the results of the process to reflect the quality of the raw data. For example, indicators that provide information about the quality of the data can be added, such as how many cases that needed corrections (rate of actual errors and imputation rate), missed matches, the percentage of observations detected as outliers, and the frequencies of selected responses, as mentioned earlier in this document.

Metadata and documentation are considered an overarching process in the United Nations Economic Commission for Europe (UNECE) 'Generic Statistical Business Process Model'. Information is derived from every step, and the information from this process must be included in the overall documentation of the survey.

Likewise, after processing the food consumption data, the microdata should be systematically organised and integrated in the production of public-use files and licensed files from the whole survey.

How to document

These guidelines have provided suggestions for *what* to keep and document, based on what is relevant for future users. It is not in the scope of these guidelines to suggest *how* to document the survey. It is assumed that the survey owner has appropriate systems in place for data and metadata storage, documentation and transparent sharing. It also falls under the responsibility of other expert groups to provide guidance.

Data owners are encouraged to make public-use microdata files, metadata and supporting data documentation widely available. The goal is to ensure transparency for both the users of microdata and for those using statistical products produced from the data. This must be done responsibly, under legal and ethical rules and principles, and ensuring that confidentiality is secured.

Work continues to be done to improve standards in this area. A note was presented by the Inter-Secretariat Working Group on Household Surveys (ISWGHS) to the United Nations Statistical Commission in 2023, called 'Standards and good practice for survey data documentation' (Dupriez and Asghar 2022), and is a recommended starting point. The note presents the benefits of rich and structured metadata, along with standards and tools.

Other recommended reading is the IHSN 'Quick Reference Guide for Data Archivists',[61] which provides information about how to document microdata in compliance with the Data Documentation Initiative (DDI),[62] the Pacific Community's documentation on Microdata dissemination,[63] Eurostat's 'European Statistical

61 Dupriez, O., Castro, D. M. S., & Welch, M. (2019). *Quick Reference Guide for Data Archivists—Guide for Data Archivists documentation*. International Household Survey Network. https://guide-for-data-archivists.readthedocs.io/en/latest/

62 DDI Alliance. (n.d.). *Data Documentation Initiative.* Retrieved 31 July 2024, from https://ddialliance.org/

63 Pacific Community. (2024). *Methodology Knowledge Base.* Statistics for Development Division. https://sdd.spc.int/methodology-knowledge-base

System Handbook for Quality and Metadata Reports',[64] and the UNECE 'Generic Statistical Business Process Model'.[65]

The use of food consumption data in analyses

In the feedback received to these guidelines during the drafting process, there were several questions about further processing of data. The two issues most-often mentioned were the use of price indices or deflators to adjust monetary values and preparing DEC information for 'Prevalence of undernourishment' (SDG indicator 2.1.1). The following responds to this.

- The monetary values in the survey will most often be adjusted with price indices before being used in analysis. The nominal monetary value of food consumption in the beginning of the year will, in a year with inflation, not reflect the same amount of food at the end of a year. Price differences between geographical areas, for example, between rural and urban areas or between provinces, may also be relevant to adjust for when the goal of the analysis is to reflect living standards. Different users of the data may need to do this in different ways – or at least want to do it across different time periods, areas or with different deflators. These guidelines are not moving into this area because the end goal for this process is only to get the data ready for food consumption analysis.
- If the team preparing the data already knows what value adjustments they will use for analysis, this can then be integrated into the process. There is one step where it is relevant to use adjusted monetary values, that is, if there is a combination of relevant price change between periods and too few observations in one period for items/units/values used to estimate median monetary values for imputation. This is mentioned in Step 4.2.
- Another element is that the FAO methodology to estimate the prevalence of undernourishment does not use the household distribution of DEC directly from following Steps 1 to 11. The FAO method adds two additional steps at the end of the process (1) to adjust the variability in the DEC distribution to estimate the variability due to income, and (2) to add the variability of DEC due to differences in the population body weights and physical activity levels. These two steps are needed to estimate the coefficient of variation of the DEC distribution that will be further used to estimate the prevalence of undernourishment. This is a complex procedure, which can be found in the FAO documentation related to the estimation of the prevalence of undernourishment.[66]

64 European Commission. Statistical Office of the European Union. (2021). European Statistical System handbook for quality and metadata reporting: 2021 re edition. Publications Office. https://data.europa.eu/doi/10.2785/616374

65 Generic Statistical Business Process Model. (n.d.). UNECE Statswiki. Retrieved 31 July 2024, from https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model

66 See the series of technical papers on the prevalence of undernourishment available at https://www.fao.org/statistics/methods-and-standards/en/; or refer to the SDG 2.1.1 e-learning course available at https://elearning.fao.org/course/view.php?id=386; or the methodological annex of past and recent reports on the 'State of Food Security and Nutrition in the World' available at: https://www.fao.org/publications/sofi

ANNEXES

Annex 1: Structure of final data file

Table 3: Example of final data file

Household ID	Area	Strata	Sampling weight	House-hold size	Parta-kers	Code product	Description product	Source	Quan-tity original	Unit	Amount original	Quantity grams	Amount new	Total household calories	Average dietary energy consumption per capita	Average food mon-etary value per capita	Year	Month	Expenditure quintiles	COICOP class
00-01-09-31	rural	strata1	7.3	4	4	01.1.8.2.0_99	Sugar, not further specified	cash	2	(kg) Kilo-grams	3.8	1500	3.8	5910	4185	96.7	2021	June	4	Other sugar and sugar substitutes (ND)
00-01-09-31	rural	strata1	7.3	4	4	01.1.9.1.2_93	Lamb, cooked, not further specified	gifts_received	2	(kg) Kilo-gramss	20.0	1750	20.0	3768	4185	96.7	2021	June	4	Ready-made food (ND)
00-01-09-31	rural	strata1	7.3	4	4	01.1.9.1.3_19	Cabbage, bush, slippery, boiled	gifts_received	3	Bundle / Bunch / Pack	9.0	3000	9.0	870	4185	96.7	2021	June	4	Ready-made food (ND)
00-01-09-31	rural	strata1	7.3	4	4	01.1.2.5.1_99	Sausage, not further specified	gifts_received	340	(g) Grams	4.5	340	4.5	746	4185	96.7	2021	June	4	Meat, offal, blood and other parts of slaughtered animals' preparations (ND)
00-01-09-31	rural	strata1	7.3	4	4	01.1.7.2.9_99	Capsicum, not further specified	gifts_received	3	Unit/ each	5.0	1500	5.0	460	4185	96.7	2021	June	4	Fruit-bearing vegetables, fresh or chilled (ND)
00-01-09-31	rural	strata1	7.3	4	4	01.1.6.1.8_03	Coconut, brown	gifts_received	11	Unit/ each	11.0	6300	5.7	12235	4185	96.7	2021	June	4	Dates, figs and tropical fruits, fresh (ND)
00-01-09-31	rural	strata1	7.3	4	4	01.1.2.2.4_96	Chicken, not further specified	gifts_received	9	(kg) Kilo-grams	45.0	9000	45.0	13487	4185	96.7	2021	June	4	Meat, fresh, chilled or frozen (ND)
00-01-09-31	rural	strata1	7.3	4	4	01.1.1.3.1_97	Bread, loaf, not further specified	cash	6	(kg) Kilo-grams	28.0	5810	28.0	13688	4185	96.7	2021	June	4	Bread and bakery products (ND)
00-01-09-31	rural	strata1	7.3	4	4	110.1.1.1.5	Non-alcoholic drinks away from home	cash	5	..	10.0		10.0	408	4185	96.7	2021	June	4	Restaurants, cafés and the like – with full service (S)
00-01-09-31	rural	strata1	7.3	4	4	01.1.7.5.5_01	Taro, common	gifts_received	7	Unit/ each	35.0	3500	35.0	3474	4185	96.7	2021	June	4	Tubers, plantains and cooking bananas (ND)
00-01-09-31	rural	strata1	7.3	4	4	01.2.6.0.0_99	Soft drink, not further specified	cash	1	Can / Bottle	7.5	221	7.5	52	4185	96.7	2021	June	4	Soft drinks (ND)
00-01-09-31	rural	strata1	7.3	4	4	110.1.1.1.8	Snacks away from home	cash	5	..	25.0		25.0	11222	4185	96.7	2021	June	4	Restaurants, cafés and the like – with full service (S)
00-01-09-31	rural	strata1	7.3	4	4	110.1.1.1.3	Lunch away from home	gifts_received		..	3.8		3.8	736	4185	96.7	2021	June	4	Restaurants, cafés and the like – with full service (S)
00-01-09-31	rural	strata1	7.3	4	4	01.1.1.2.9_99	Flour, not further specified	cash	2	(kg) Kilo-grams	4.0	2000	4.0	7056	4185	96.7	2021	June	4	Flour of cereals (ND)
00-01-09-31	rural	strata1	7.3	4	4	110.1.1.1.8	Snacks away from home	cash	5		25.0		25.0	11222	4185	96.7	2021	June	4	Restaurants, cafés and the like – with full service (S)
00-01-09-31	rural	strata1	7.3	4	4	01.1.7.1.9_03	Leaves, taro	gifts_received	2	Bundle / Bunch / Pack	6.0	1200	6.0	419	4185	96.7	2021	June	4	Leafy or stem vegetables, fresh or chilled (ND)
00-01-09-31	rural	strata1	7.3	4	4	01.2.4.0.0_02	Beverage, chocolate flavour, from base (Milo)	gifts_received	200	(g) Grams	5.3	200	5.3	877	4185	96.7	2021	June	4	Cocoa drinks (ND)
00-01-09-31	rural	strata1	7.3	4	4	01.1.2.2.3_03	Lamb and mutton, regular, cuts not specified	cash	1	(kg) Kilo-grams	17.5	1000	17.5	1957	4185	96.7	2021	June	4	Meat, fresh, chilled or frozen (ND)
00-01-09-31	rural	strata1	7.3	4	4	01.1.7.4.3_03	Onion, spring	cash	4	Unit/ each	2.0	505	2.0	121	4185	96.7	2021	June	4	Other vegetables, fresh or chilled (ND)
00-01-09-31	rural	strata1	7.3	4	4	01.2.3.0.3_01	Tea, iced, commerical	home_ production	28	Unit/ each	1.4	140	1.4	55	4185.376	96.7	2021	June	4	Tea, maté and other plant products for infusion (ND)
00-01-09-31	rural	strata2	7.3	4	4	01.1.4.3.2_99	Milk, powdered, not further specified	cash	450	(g) Grams	9.8	450	9.8	1807	4185.376	96.7	2021	June	4	Other milk and cream (ND)
00-01-09-31	rural	strata2	7.3	4	4	01.1.5.2.1_99	Butter, not further specified	cash	454	(g) Grams	13.0	454	13.0	3342	4185.376	96.7	2021	June	4	Butter and other fats and oils derived from milk (ND)
00-01-09-31	rural	strata2	7.3	4	4	01.1.1.5.0_01	Noodles, instant (Maggi-type), dry	cash	340	(g) Grams	4.8	340	4.8	1524	4185.376	96.7	2021	June	4	Macaroni, noodles, couscous and similar pasta products (ND)
00-01-09-31	rural	strata2	7.3	4	4	01.1.9.4.0_99	Spices, not further specified	cash	25	(g) Grams	0.5	25	0.5	87	4185.376	96.7	2021	June	4	Spices, culinary herbs and seeds (ND)
00-01-09-31	rural	strata2	7.3	4	4	01.1.1.3.1_99	Crackers, not further specified	cash	1	(kg) Kilo-grams	11.2	1440	11.2	6231	4185.376	96.7	2021	June	4	Bread and bakery products (ND)
00-01-09-31	rural	strata2	7.3	4	4	01.1.7.2.4_01	Tomato, common	gifts_received	6	Unit/ each	5.0	300	5.0	49	4185.376	96.7	2021	June	4	Fruit-bearing vegetables, fresh or chilled (ND)
00-01-09-31	rural	strata2	7.3	4	4	01.1.2.2.1_04	Beef, regular, cut not specified	cash	1	(kg) Kilo-grams	15.5	1000	15.5	1688	4185.376	96.7	2021	June	4	Meat, fresh, chilled or frozen (ND)
00-01-09-31	rural	strata2	7.3	4	4	110.1.1.1.5	Non-alcoholic drinks away from home	cash	5	..	10.0		10.0	408	4185.376	96.7	2021	June	4	Restaurants, cafés and the like – with full service (S)
00-01-09-31	rural	strata2	7.3	4	4	11.1.1.1.1_49	Takeaway, fish, fried, bbq'd	gifts_received	740	(g) Grams	16.0	740	16.0	1362	4185.376	96.7	2021	June	4	Restaurants, cafés and the like – with full service (S)
00-01-09-31	rural	strata2	7.3	4	4	01.1.6.5.4_01	Watermelon	gifts_received	1	Unit/ each	10.0	6000	10.0	759	4185.376	96.7	2021	June	4	Other fruits, fresh (ND)
00-01-09-31	rural	strata2	7.3	4	4	01.1.9.3.1_01	Salt, iodised	cash	59	(g) Grams	0.5	59	0.5	0	4185.376	96.7	2021	June	4	Salt, condiments and sauces (ND)
00-01-09-31	rural	strata2	7.3	4	4	01.1.5.1.9_01	Oil, cooking	cash	1	(ltr) Litres	5.0	920	5.0	8280	4185.376	96.7	2021	June	4	Vegetable oils (ND)
00-01-09-31	rural	strata2	7.3	4	4	01.1.8.6.0_05	Ice cream, vanilla	gifts_received	2	(ltr) Litres	18.0	1120	18.0	2153	4185.376	96.7	2021	June	4	Ice, ice cream and sorbet (ND)
00-01-09-31	rural	strata2	7.3	4	4	110.1.1.1.3	Lunch away from home	gifts_received			3.8		3.8	736	4185.376	96.7	2021	June	4	Restaurants, cafés and the like – with full service (S)
00-01-26-36	rural	strata3	17.8	3	3	01.1.4.2.0_99	Milk, fresh, not further specified	cash	2	(ltr) Litres	7.0	2060	7.0	1013.52	2616.779	96.7	2021	Sep-tember	4	Skimmed milk (ND)
00-01-26-36	rural	strata3	17.8	3	3	02.3.1.8.7_01	Kava	gifts_received	85	(g) Grams	10.0	200	10.0	0	2616.779	96.7	2021	Sep-tember	4	2400

This is an example of the structure with the core variables. The processing should provide an overview of all the data collected, so a final data file will also contain other information from each specific survey. It is further recommended to add as many variables as possible on household characteristics or food groupings to meet the needs of different users.

Annex 2: Examples of working datasets – using the long form

Table 4: Structure of the working file when the survey adopts the same design as that from recent HIES conducted in Pacific Islands countries

Household ID	Code product	Product description	Quantity	Unit	Monetary value (local currency)	Source
00-01-09-31	01.1.7.2.9_99	Capsicum, not further specified	3	Piece	5	Gift
00-01-09-31	01.1.9.4.0_99	Spices, not further specified	25	Grams	0.5	Purchase
00-01-09-31	110.1.1.1.5	Non-alcoholic drinks away from home	5		10	Purchase
00-01-09-31	01.2.4.0.0_02	Beverage, chocolate flavour, from base (Milo)	200	Grams	5.3	Gift
00-01-09-31	01.1.1.3.1_99	Crackers, not further specified	1.44	Kg	11.2	Purchase
00-01-09-31	01.1.9.3.1_01	Salt, iodised	59.2	Grams	0.5	Purchase
00-01-09-31	01.1.7.5.5_01	Taro, common	7	Piece	35	Gift
00-01-09-31	01.2.3.0.3_01	Tea, iced, commercial	28	Piece	1.4	Home produce
00-01-09-31	01.1.2.2.3_03	Lamb and mutton, regular, cuts not specified	1	Kg	17.5	Purchase
00-01-09-31	01.1.7.1.9_03	Leaves, taro	2	Bundle/bunch	6	Gift
00-01-09-31	01.1.6.1.8_03	Coconut, brown	11	Piece	11	Gift
00-01-09-31	01.1.7.2.4_01	Tomato, common	6	Piece	5	Gift
00-01-09-31	01.1.5.1.9_01	Oil, cooking	1	Litres	5	Purchase
00-01-09-31	11.1.1.1.1_49	Takeaway, fish, fried, barbequed	740	Grams	16	Gift
00-01-09-31	01.1.9.1.2_93	Lamb, cooked, not further specified	1.75	Kg	20	Gift
00-01-09-31	110.1.1.1.8	Snacks away from home	5		25	Purchase
00-01-09-31	01.1.1.5.0_01	Noodles, instant (Maggi type), dry	340	Grams	4.8	Purchase
00-01-09-31	01.1.2.2.1_04	Beef, regular, cut not specified	1	Kg	15.5	Purchase
00-01-09-31	110.1.1.1.8	Snacks away from home	5		25	Purchase
00-01-09-31	01.1.1.3.1_97	Bread, loaf, not further specified	5.81	Kg	28	Purchase

Annex 3: Example of summary table combining information on weight in grams and price per unit

Table 5: Summary example combining weight and price

Description	COICOP	Unit	Region	Number of combinations product/unit/region	Ad hoc conversion in grams from NSO	Regional median weight in grams from market survey	Number of observations on which regional median weight is estimated	National median weight in grams from market survey	Regional median price of one gram from market survey	National median price of one gram from market survey	Median unit value from purchase section of recall	Number of observations on which median unit value is estimated	Median price in grams from purchase section of recall	Ad hoc price in grams from NSO/CPI
Rice	111001099	Bag	Region 1	10	5000	18000	7	18000	0.20	0.19	2800	303	0.2	0.18
Rice	111001099	Bag	Region 2	6	5000	18150	13	18000	0.18	0.19	2800	303	0.2	0.18
Rice	111001099	Bag	Region 3	5	5000	18000	3	18000	0.20	0.19	2800	303	0.2	0.18
Rice	111001099	Bag	Region 4	21	5000	10000	5	18000	0.19	0.19	2800	303	0.2	0.18
Rice	111001099	Bag	Region 5	18	5000	25000	14	18000	0.20	0.19	2800	303	0.2	0.18
Rice	111001099	Bag	Region 6	77	5000	18000	25	18000	0.17	0.19	2800	303	0.2	0.18
Rice	111001099	Bundle/bunch/packet	Region 1	1		1000	2	1000	0.20	0.19	200	5	0.2	0.18
Rice	111001099	Bundle/bunch/packet	Region 3	1		1000	12	1000	0.20	0.19	200	5	0.2	0.18
Rice	111001099	Bundle/bunch/packet	Region 4	5		1000	35	1000	0.19	0.19	200	5	0.2	0.18
Rice	111001099	Bundle/bunch/packet	Region 5	1		1000	2	1000	0.20	0.19	200	5	0.2	0.18
Rice	111001099	Bundle/bunch/packet	Region 6	1		1000	1	1000	0.17	0.19	200	5	0.2	0.18
Rice	111001099	Cups	Region 1	263		399	10	399	0.20	0.19	77.5	164	0.2	0.18
Rice	111001099	Cups	Region 2	76				399	0.18	0.19	77.5	164	0.2	0.18
Rice	111001099	Cups	Region 3	576				399	0.20	0.19	77.5	164	0.2	0.18
Rice	111001099	Cups	Region 4	926				399	0.19	0.19	77.5	164	0.2	0.18
Rice	111001099	Cups	Region 5	9				399	0.20	0.19	77.5	164	0.2	0.18
Rice	111001099	Cups	Region 6	77		316.5	6	399	0.17	0.19	77.5	164	0.2	0.18
Rice	111001099	grams	Region 2	2					0.18	0.19	200	1	0.2	0.18
Rice	111001099	grams	Region 3	1					0.20	0.19	200	1	0.2	0.18
Rice	111001099	grams	Region 4	1					0.19	0.19	200	1	0.2	0.18
Rice	111001099	grams	Region 6	3					0.17	0.19	200	1	0.2	0.18

SOURCE: Authors' elaboration.

NOTES:
COICOP – Classification of individual consumption according to purpose
NSO – National statistics office
CPI – Consumer price index

Annex 4: Example of an NCT table with documentation

The following list shows the minimum number of variables needed in the NCT food item code in household surveys.

- Food item code in data file – if changed or added (e.g. from chronological to COICOP)
- Food item description in household survey
- Reference food composition table (FCT)
- Food code in FCT
- Food description in FCT
- Food item index matching (note the text in the comment cells)
- Refuse factor
- Food group to which the item belongs to
- Weighting factor (used when the nutrient value of a food item is estimated as the weighted average of the nutrient values of other foods)
- Macronutrients in grams (fats, proteins, carbohydrates, fibre and alcohol)
- Kcal from FCT
- Kcal calculated from macronutrients.

Table 6: Example of an NCT showing some nutrients

Food item code in household survey	Food item description in household survey	Reference FCT	Food code in FCT	Food description in FCT	Food item matching	Refuse factor	Item group	Water (g)	Ash (g)	Protein (g)	Total fat (g)	Fibre (g)	Carbo-hydrates including fibre (total)
12	Farine de maïs	From worksheet average		Weighted average	A2	0	Cereals	10.80	0.65	7.35	2.20	3.85	75.10
13	Farine de mil	West African FCT	01_063	Mil chandelle, farine (sans son)	A	0	Cereals	12.00	0.79	7.40	3.16	4.60	72.10
14	Farine de blé local ou importé	West African FCT	01-043	Blé, farine, blanc	A	0	Cereals	11.85	0.64	10.36	1.45	3.17	72.54

SOURCE: Authors' elaboration.

Table 7: Example instructions for how to build an NCT

Available carbohydrates = 100 - Water - Ash - Protein - Fats - Fibers - Alcohol

Total carbohydrates = Available carbohydrates + Fibers

Computed calories = Protein * 4 + Fats * 9 + Available carbohydrates * 4 Fiber * 2 + Alcohol * 7

Food description in FCT	Food Item Index Matching (Read the text in the comment of this cell)	Refuse factor (refuse)	Item group (Item_grp)	Item group (diversity_grp)	Water (grams)	Ash (grams)	Protein (grams) (fd_pro)	Fat (grams) (fd_fat)	Fiber (grams) (fd_fib)	Alcohol (grams) (fd_alc)	Carbohydrates including fiber (Total) (grams)	Available carbohydrates by difference (grams) (fd_car)	Calories (Kcal)	Computed calories (Kcal) (fd_kcal)
From weighted average	A2	0	1	1	11,83	0,84	7,08	1,32	2,53	0	78,93	76,39	364,17	351
Cornstarch	B	0	1	1	8,32	0,09	0,26	0,05	0,90	0	91,27	90,38	381,00	365
Bananas, raw	A	36	8	7	74,91	0,86	1,09	0,33	2,6	0	22,84	20,21	89	93
From weighted average	A2	9	7	5	94,40	0,55	1,06	0,21	0,98	0	3,79	2,815	18,00	19
Egg, whole, raw, fresh	B	12	13	10	76,15	1,06	12,56	9,51	0	0	0,72	0,72	143	139
Chicken, broilers or fryers, breast, meat and skin, raw	A	20	12	9	69,46	1,01	20,85	9,25	0	0	0	0	172	167
Tea, leaves	A	95	9	16	9,3	7	19,6	2	55,8	0	62,1	6,3	229,53	233
Alcoholic beverage, beer, regular, all	A	0	11	16	91,96	0,16	0,46	0	0	3,9	3,55	3,52	43	43
	D	0	19	17										

Estimation based on the assumption that only 1/20 of nutrients is going to the liquid tea/coffee

Example: Missing value for ash in USDA for item code 09040

Ash value from another FCT has to be adjusted by total solids

	USDA FCT	DANISH FCT
Ash value	0.86	0.8
Water content	74.91	76.6

Ash value = (value in FCT) x (100-water content of food item with missing value)/ (100-water content of food item in FCT)

Ash value = 0.8 * (100 - 74.91) / (100 - 76.6)

Differences due to:
Chemical analysis,

Other calorie conversion factors used in the formula, Others

SOURCE: Authors' elaboration.

NOTES:
* A – Single, perfect match, no modifications required (apart from edible portion, if indicated). A2 – Exact match, but multiple selections, need weighting. B – Similar, single match. B2 – Similar match, multiple selections, need weighting. C – Poor, single match.
FCT – Food composition table
USDA – United States Department of Agriculture

REFERENCES

Borlizzi A., Delgrossi M. and Cafiero C. 2017. National food security assessment through the analysis of food consumption data from Household Consumption and Expenditure Surveys: The case of Brazil's *Pesquisa de Orçamento Familiares* 2008/09. Food Policy 72:20–26. Available at: https://www.sciencedirect.com/journal/food-policy/vol/72/suppl/C

COICOP (Classification of Individual Consumption According to Purpose) 2018. United Nations Department of Economic and Social Affairs, Statistics Division. Available at: https://unstats.un.org/unsd/classifications/unsdclassifications/COICOP_2018_-_pre-edited_white_cover_version_-_2018-12-26.pdf

Deaton A. and Zaidi S. 2002. Guidelines for Constructing Consumption Aggregates for Welfare Analysis. LSMS working paper no. 135. World Bank. Available at: http://hdl.handle.net/10986/14101

DDI Alliance. (n.d.). *Data Documentation Initiative*. Retrieved 31 July 2024, from https://ddialliance.org/

Dodge Y. (ed.) 2003. The Oxford Dictionary of Statistical Terms. Oxford University Press. DOI: doi.org/10.1002/sim.1812

Dupriez O. and Asghar M. 2022. Standards and good practice for survey data documentation. Inter-Secretariat Working Group on Household Surveys. Available at: https://ecosoc.un.org/sites/default/files/documents/2023/BG-3c-MetadataStandard-E.pdf

Dupriez O., Castro D. and Welch M. 2019. Quick Reference Guide for Data Archivists. International Household Survey Network. Available at: https://guide-for-data-archivists.readthedocs.io/en/latest/

Eurostat 2020. European Statistical System handbook for quality and metadata reports. Available at: https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-gq-19-006

European Commission. Statistical Office of the European Union. (2021). *European Statistical System handbook for quality and metadata reporting: 2021 re edition.* Publications Office. https://data.europa.eu/doi/10.2785/616374

Human energy requirements 2004. Report of a Joint FAO/WHO/UNU Expert Consultation, Rome 2001. Available at: https://www.fao.org/3/y5686e/y5686e00.htm

FAO International Network on Food Data Systems (INFOODS) 2022. Available at: https://www.fao.org/infoods/infoods/en/ (accessed 5 November 2023).

FAO / INFOODS 2012. Databases: Density Database (version 2.0). Available at: https://www.fao.org/3/ap815e/ap815e.pdf

FAO/World Bank 2018. Food Data Collection in Household Consumption and Expenditure Surveys: Guidelines for Low- and Middle-Income Countries. [Abbreviated in this publication to the 'IAEG-AG 2018 guidelines'.] Available at: https://www.worldbank.org/en/programs/lsms/publication/Food-data-collection-in-household-consumption-and-expenditure-surveys-guidelines-for-low-and-middle-income-countries

Filzmoser P., Gussenbauer J. and Templ, M. 2016. Detecting outliers in household consumption survey. Vienna University of Technology/World Bank. Available at: http://www.ihsn.org/node/695

Generic Statistical Business Process Model. (n.d.). UNECE Statswiki. Retrieved 31 July 2024, from https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model

Kenya Integrated Household Budget Survey 2015–2016. Kenya National Data Archive (KeNADA). Available at: https://statistics.knbs.or.ke/nada/index.php/catalog/13/related-materials

McGill R., Tukey J. and Larsen W 1978. Variations of Box Plots. The American Statistician 32.1:12–16. Available at: https://www.jstor.org/stable/2683468

Mancini G. and Vecchi G. 2022. On the Construction of a Consumption Aggregate for Inequality and Poverty Analysis. Washington, DC: World Bank Group. Available at: https://documents.worldbank.org/en/publication/documents-reports/documentdetail/099225003092220001/p1694340e80f9a00a09b20042de5a9cd47e

National Panel Survey: Uganda, 2019/20. Uganda Bureau of Statistics. DOI: https://doi.org/10.48529/nqzx-f196

OECD (Organisation for Economic Co-operation and Development). Adjusting household incomes: Equivalence scales. Available at: https://www.oecd.org/els/soc/OECD-Note-EquivalenceScales.pdf

Oseni, G., Palacios-Lopez, A., Mugera, H.K. and Durazo, J. (2021). Capturing What Matters: Essential Guidelines for Designing Household Surveys. Washington DC: World Bank. Available at: https://documents1.worldbank.org/curated/en/381751639456530686/pdf/Capturing-What-Matters-Essential-Guidelines-for-Designing-Household-Surveys.pdf

Oseni, G., Durazo, J., & McGee, K. 2017. The Use of Non-Standard Units for the Collection of Food Quantity: A Guidebook for Improving the Measurement of Food Consumption and Agricultural Production in Living Standards Surveys. Washington DC: World Bank. Available at: https://documents1.worldbank.org/curated/en/981711587040434398/pdf/The-Use-of-Non-Standard-Units-for-the-Collection-of-Food-Quantity-A-Guidebook-for-Improving-the-Measurement-of-Food-Consumption-and-Agricultural-Production-in-Living-Standards-Surveys.pdf

Pacific Community. (2024). *Methodology Knowledge Base*. Statistics for Development Division. https://sdd.spc.int/methodology-knowledge-base

Sharp M., Buffière B., Himelein K., Troubat., N. and Gibson J. 2022. Effects of Data Collection Methods on Estimated Household Consumption and Survey Costs: Evidence from an Experiment in the Marshall Islands (English). Policy Research working paper; no. WPS 10029. Washington, DC: World Bank Group. Available at: http://documents.worldbank.org/curated/en/099834004282216747/IDU014657f420d00e042a30a68903563423f5b37

Smith L.C., Dupriez O. and Troubat N. 2014. Assessment of the Reliability and Relevance of the Food Data Collected in National Household Consumption and Expenditure Surveys. IHSN working paper no. 008. Available at: http://www.ihsn.org/sites/default/files/resources/IHSN_WP008_EN.pdf

Smith L.C. and Subandoro A. 2007. Measuring food security using household expenditure surveys. International Food Policy Research Institute (IFPRI): Food security in practice technical guide series 3. Available at: https://ideas.repec.org/p/fpr/fsprac/3.html

Todorov V., Templ M. and Filzmoser P. 2011. Detection of multivariate outliers in business survey data with incomplete information. Advances in Data Analysis and Classification 5:37–56. Available at: https://link.springer.com/article/10.1007/s11634-010-0075-2

UNECE (United Nations Economic Commission for Europe) Statswiki 2019a. Generic Statistical Data Editing Model (version 2.0): SDE Flow Models. Available at: https://statswiki.unece.org/display/sde/5+SDE+Flow+Models

UNECE (United Nations Economic Commission for Europe) Statswiki 2019b. Generic Statistical Business Process Model (version 5.1). Available at: https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model

GLOSSARY

This is an overview of some central concepts used in this publication. Some are directly copied from the IAEG-AG 2018 guidelines to align vocabulary with what is used in that publication.

Adult male equivalent (AME) – expresses energy requirements on the basis of gender, age and physiological status as a proportion of the energy requirements of an average adult male (Weisell and Dop, 2012; World Health Organization, 2004).

Coefficient of Variation (CV) – a measure of dispersion of a (frequency or probability) of distribution and is defined as the ratio of the distribution's standard deviation to the mean.

Consumer price index (CPI) – a measure of the aggregate price level in an economy. The CPI measures the price level of goods and services and the changes in the purchasing power of a country's currency over time.

Density – the factor used to convert quantities in volume (for example, litres) to quantities in mass (kilograms). One litre of water is 1000 grams, but other liquids may be heavier or lighter than water.

Diary – a method of self-reported data collection. In a food diary, one or more individuals in a household are asked to record, at a daily level, the household's or individual's food acquired, consumed or both during the reference period of food data collection.

Dietary energy – the energy needed by the human body to function and maintain body temperature and the continuous action of the heart and lungs. In children, energy is essential for growth. Energy is needed for the breakdown, repair and building of tissues. It is usually measured in kilojoules (kJ) or kilocalories (kcal). Kcal is quite often used synonymously with dietary energy.

Dietary energy consumption (DEC) – a measure of calories consumed by a population group. It is expressed in kilocalories per capita per day, and is estimated from households' food quantities collected in surveys, after being adjusted for non-edible portions (e.g. bones and peels). DEC is calculated by converting consumed quantities into calories using a food composition table.

Edible portion – the factor describing the proportion of a food item that could be ingested. In HCES, quantities of foods are normally reported by households as acquired (i.e. with peel, bones or other parts that are not consumed). Dietary energy is estimated from only the part that is edible.

Estimate – the value yielded by an estimator in a given set of circumstances. The expression is widely used to denote the rule by which such particular values are calculated. It seems preferable to use the words 'estimator' for the rule of procedure and 'estimate' for the values to which it leads in particular cases.[67]

Food acquisition – food (quantity or monetary values or both) acquired by households for the purpose of human consumption. It includes food from purchases, own production and from other sources, such as food received as a gift, aid or as a payment.

67 Source: Dodge Y. (ed.) 2003. See also: estimator, trend estimates, non-sampling error, and error of estimation.

Food away from home (FAFH) – food prepared and consumed outside of the dwelling. Foods prepared at home and consumed outside the dwelling (such as lunch boxes) and foods prepared outside the dwelling and consumed in the dwelling are not considered as foods consumed away from home. Food away from home includes all meals, snacks and nonalcoholic beverages, as well as fast food, take aways and deliveries, meals consumed at concession stands, buffets and cafeterias, full-service restaurants, and purchased at vending machines or from mobile vendors that are consumed outside the dwelling. Also included are board (including at school) meals as pay, special catered affairs (such as weddings, bar mitzvahs and confirmations), school lunches and meals away from home on trips.

Food composition table – a table with information about nutritional properties of foods usually consumed in a country (national) or a group of countries (regional), which includes information about the content of selected macro- and micro-nutrients and the fraction of edible portions. It is generated according to international guidelines to be comparable and reliable.

Food consumption – food (quantity or in monetary values or both) consumed by a household. It refers to apparent consumption or household-level food consumption and should not be confused with individual food consumption. The apparent food consumed can come from food acquired during the same reference period or from household stocks.

Food from gift, in-kind payment and other – food acquired or consumed at home or outside the house from sources different from purchases and own production (e.g. gift, charity, as part of payment, government programmes, food received as part of a school feeding programme, etc.).

Food from own production – food acquired or consumed by one or more household member from the household's own production for the household's at-home consumption.

Food from purchases – food acquired or consumed by one or more household members that was paid for. It includes food consumed at home or away from home.

Food matching – the process of pairing a food item reported in a HCES with a food item belonging to a referenced Food Composition Table (FCT).

Food price index (FPI) – a measure of the aggregate price of food in an economy. The FPI consists of a bundle of commonly purchased foods.

Food recall period – the period over which respondents were asked to recall the consumption of food items.

Food reference period – the time period for which respondents were asked to report on their food acquisitions and/or food consumption. 'Recall' and 'reference' periods are quite often used interchangeably, though they are different terms. The recall period differs from the reference period in regards to when households were interviewed, multiple times, during multiple visits to the household. For example, if households were interviewed about their food consumption over the past seven days in four, weekly visits, the recall period is seven days and the reference period is 28 days.

Household – usually an individual or group of individuals, related or unrelated, who live together in the same dwelling unit; share the same living arrangements; pool some or all of their income and wealth; consume certain types of goods and services collectively, mainly housing and food; and are considered as one unit.

Imputation – the procedure for entering a value for a specific data item where the response is missing or unusable. Context: Imputation is the process used to determine and assign replacement values for missing, invalid or inconsistent data that have failed edits. This is done by changing some of the responses or assigning values when they are missing on the record being edited to ensure that estimates are of high quality and that a plausible, internally consistent record is created.[68]

List of food items – food items that are pre-filled in a food consumption module.

Macronutrients – includes fat, protein, carbohydrate, dietary fibre and alcohol. Macronutrients are eaten in large amounts and include the primary building blocks of the diet. They provide the body with energy and are used to calculate the kilocalories in a food item.

Micronutrients – essential nutrients such as vitamins (for example, vitamin A, C, D, B1, B2, B6, etc.) and minerals (for example, calcium, iron, zinc, etc.) that are needed by the body in small doses.

Multivariate approach – used when looking simultaneously at the distribution of more than one variable to assess the consistency between, say, quantity and value reported for a food purchase.

Outlier – an observation that lies an abnormal distance from other values in a random sample from a population.

Partakers – number of individuals who actually consumed the total amount of food reported by the household during the reference period of food data collection. The household size may be different from the number of food partakers because foods may be shared with non-household members, such as guests, employees and relatives, and household members may have been absent from the household during the reference period.

Percentile – each of the 100 equal groups into which a population can be divided according to the distribution of values of a particular variable.

Prevalence of undernourishment – an estimate of the proportion of the population whose habitual food consumption is insufficient to provide the dietary energy levels that are required to maintain a normal active and healthy life. It is expressed as a percentage. This indicator measures progress towards SDG target 2.1 (end hunger and ensure access by all people, in particular the poor and people in vulnerable situations, including infants, to safe, nutritious and sufficient food all year round).

Recall interviews – a method of data collection, in which one or all individuals in the household are asked to recall information that the interviewer records. For food data collection, the focus of recall is the household's or individual's food acquired and consumed during the reference period.

Refuse factor – corresponds to the part of the food that is thrown away because not edible. A refuse factor is different from waste, which refers to the food that is thrown away because it is in bad form or left on the plates.

Sampling weights (or survey weights) – positive values associated with the observations (rows) in the dataset (sample). They correspond to the inverse of the probability for a household to be selected. They are used to ensure that metrics derived from a dataset are representative of the population (the set of observations).

68 Sources: *Statistics Canada Quality Guidelines*, 4th edition, October 2003, p41, Statistics Canada; *Glossary of Terms on Statistical Data Editing*, United Nations Economic Commission for Europe (UNECE), https://unece.org/DAM/stats/publications/editing/SDEGlossary.pdf; and Conference of European Statisticians methodological material, Geneva, 2000. See also: Missing data French equivalent: Imputation.

Seasonality – the variations in food acquired or consumed or both and the related expenditure over a long period (for example, six months or one year). Seasonality is usually linked to the agricultural production season. Other cyclical events, such as floods and droughts, may also cause variations affecting both food availability and prices.

Univariate approach – where each variable is checked separately to identify extreme values within a variable's distribution.

Visitors – non-household members who join the household to share meals prepared at home.

TERMINOLOGY

This is an overview of the terminology used in this publication and what it refers to.

Data cleaning – the process of detecting and correcting errors (systematic errors and outliers) in the dataset.

Dietary energy unit cost – the cost of 1 kcal, estimated from the food consumption module data.

Editing – the process involving the review and adjustment of collected survey data.

Food item – a food or beverage product reported in the survey as being consumed (or acquired).

IAEG-AG 2018 guidelines – the World Bank and FAO guidelines on food consumption, designed under the aigis of the UN-CEAG (at the time called IAEG-AG).

In-house consumption (or at-home consumption) – refers to the food consumed by household members and household guests. Food prepared outside the house and consumed in the house is considered as in-house food consumption.

Living Standards Measurement Study (LSMS) – the World Bank's flagship household survey programme focused on strengthening household survey systems in client countries and on improving the quality of microdata to better inform development policies.

LSMS 2021 guidebook – the 2021 World Bank essential guidelines for designing household surveys and collecting food data.

Measurement unit – the unit of measurement in which the quantity is reported. We separate between standard units of measurement and non-standard units of measurement.

Market price – the price in local currency of one unit from a market survey, as opposed to the unit value cost.

Market survey – the survey of foods in the local market, conducted in parallel with the HCES to collect prices and/or weights of non-standard units for conversion to grams.

Monetary value – the food expenditure for purchases and the respondent estimated value of the food consumed from own production or received in kind. The monetary value is expressed in local currency unit.

Non-standard units – the measurement units in which quantities for food consumption (or acquired for consumption) are reported, and that are different from the measurement units included in the metric system (e.g., kilogram and litre). Examples of non-standard units are heaps, cups, bags and baskets.

Price – the word 'price' is used when we refer to the amount of money a consumer must spend per unit to acquire a product, disregarding where the price comes from (market survey, HCES or other sources).

Raw data – the non-altered form of the data, as received from the NSOs, after being collected and available in electronic form.

Standard unit of measurement – all the units of measurement in which quantities of food consumed are reported in the HCES and that are part of the International System of Units (SI), which is commonly known as the metric system. Kilogram is one of the base units of the SI but all units of measurement that have a standard and accepted conversion in kilogram (such as pound, once and litre) are also considered in these guidelines as standard units.

Transformed data – any data that has been modified from its original form during the process.

Unit value – the price in local curency of one unit of a food item. It is calculted by dividing total cost by total quantity for purchases reported by the household in the food consumption module data of the HCES, as opposed to the prices collected in the market surveys.

Weight in grams – (or conversion factors for NSU or gram-equivalent factor) – the weight in grams of one non-standard measurement unit.

Produced by the Pacific Community (SPC)
Pacific Community
B.P D5 - 98848 Noumea Cedex
New Caledonia

Telephone: +687 262000
Email: spc@spc.int
Website: www.spc.int

ISBN 978-982-00-1606-4