Technical Appendix: Wave 1 Somali HFS

This technical appendix describes sample design, cleaning and construction of consumption aggregates for the Wave 1 Somali High Frequency Survey data.

Introduction

Estimating monetary poverty rates requires a sound, reproducible methodology. The methodology starts with the sample design, continues with questionnaire design and the construction of food and non-food consumption aggregates, selection of spatial price deflators and how to determine the consumption value derived from assets, and what process to use to construct the poverty lines. This appendix describes the methodology used to estimate poverty for the Wave 1 Somali High Frequency Survey.

The chosen methodology balances a trade-off between feasibility and accuracy. Somalia is a fragile country with severe security constraints for field work and wide spread displacement. The sampling methodology was adapted to the context by excluding several inaccessible areas. The questionnaire design utilized the Rapid Consumption methodology that can be easily and quickly implemented. The choice of deflators and the poverty line were driven by data quality.

A household is defined as poor if the per-capita household consumption does not exceed a given threshold

$$y_i \leq z \tag{1}$$

where y_i is the nominal per-capita household expenditure and z is the poverty line at the nominal level. In the following, we discuss the selection of households i as part of the sample design and present the construction of the consumption aggregate y_i before discussing the choice of the poverty line z and standard poverty measures.

Sample

The Population Estimation Survey of Somalia (PESS) was used as sample frame alongside a list of settlements from three different sources (UNDP 1997, UNDP 2006 and FSAU 2003) to complement missing rural and semi-urban settlements. The combined sample frame was cleaned and preprocessed before the number of enumeration areas per strata was calculated and enumeration areas selected proportional to size. Depending on the strata, different multi-stage clustering approaches were used to select households.

Sample Frame

Due to the combination of the different data sources, the resulting sample frame included enumeration areas as well as settlements. While enumeration areas are defined as geographical areas with about 50 to 200 households, settlements often are larger areas with a larger population. In fact, all rural and a large fraction of semi-urban enumeration areas and settlements did not have boundaries available but were only defined by a GPS position.

Since PESS is also partially based on the same data sources (especially UNDP 1997 and UNDP 2006) and since some PESS enumeration areas had the same GPS location, several GPS positions were very close of each other and, thus, considered duplicates (Figure 1). Technically, duplicates are defined where the distance between the GPS position is below 75m. In groups with multiple duplicates, the additional criteria was introduced that all GPS positions must have pair-wise distances below 200m to prevent large sequential areas of GPS positions. Duplicates were merged into one 'hypothetical' enumeration area with a tag of the number of duplicates. Those duplicate counts were used to position manually midpoints for new enumeration areas around the main duplicate GPS position to ensure that larger settlements have the appropriate number of surrounding enumeration areas.[1]

In a second step, boundaries of enumeration areas without corresponding shape files were drawn automatically. First, the GPS positions were used as midpoints of circles with a radius of 200m. Overlapping circles were transformed to Thiessen polygons where the line connecting the overlapping points becomes the new boundary. The algorithm was tested for areas where PESS shapefiles were available (Figure 2).

[1] Note that this was only done for selected duplicate enumeration areas to reduce manual processing.

Figure 1: Example of duplicate GPS positions.

Figure 2: Test of Thiessen polygons with bold boundaries representing the known enumeration area boundaries.

Sample Stratification and Size

The sample is designed based on predicted statistical precision of consumption as well as cost considerations. Without political implications, the survey stratifies the sample into four zones, A including Mogadishu, B including Garowe, C including Hergeiza and D for Sanaag, Sool and Togdheer. The sample is stratified for each zone into economic/political centers, urban centers, other urban settlements, rural settlements and – if existent – IDP camps. The result are 16 strata (star marks areas where a micro-listing approach was utilized; see below):

- A: Mogadishu*; IDPs*
- B: Garowe; Urban Centers; Other Urban; Rural; IDPs*
- C: Hergeiza; Urban Centers; Other Urban; Rural; IDPs*
- D: Sanaag Urban; Sanaag Rural; Sool Urban; Sool Rural; Togdheer Urban; Togdheer Rural

The sample employs a clustered design with the Primary Sampling Unit (PSU) being the enumeration area. Within each enumeration area, 12 households will be selected for interviews. A larger number of households per enumeration area would only marginally benefit the statistical estimation of indicators. A smaller number of households would result in less than 3 observations for each of the four optional modules capturing consumption data.

A total sample of about 3,800 households is sufficient to obtain consumption estimators with a relative standard error below 1 percent. After rounding the number of enumeration areas ensuring that 12 households per enumeration area, 324 enumeration areas were initially selected. The 324 enumeration areas are first distributed into the 16 strata. The number of enumeration areas per strata is determined by (i) the population of the strata, (ii) the variability of consumption within the strata, and (iii) the requirement of at least two enumeration areas per strata. Strata with larger population and larger variability will need a larger sample to retrieve the same relative standard error as a strata with smaller population and consumption variability (Table 12). Variability is estimated based on previous surveys and a pilot in Mogadishu. The strata for Mogadishu was later amended by an additional 20 enumeration areas to correct against a faulty optional module assignment in the first days of data collection.

Household Selection

Depending on the strata, different clustering approaches were used. In strata with more volatile security as well as for IDP camps, a multi-stage cluster design was employed called micro-listing. Each selected enumeration area was divided into multiple segments and each segment was further divided into blocks. A block is defined as a geographical area where an enumerator can see (and list) all households from one location in the center of the block. Within each enumeration area, one segment was randomly selected and within the segment 12 blocks were chosen. In each block, all structures were listed before selecting randomly one structure. Within the selected structure, all households were listed and one household randomly selected for interview. This multi-stage clustering approach reduces the time in the field substantially and contributes to a lower profile of enumerators, which is paramount in fragile areas. In strata less volatile, the complete enumeration area was listed before 12 households were randomly selected for interviews (called full-listing).

Data Collection and Replacements

The survey was implemented using tablets as survey devices (CAPI). The data collection system consisted of Samsung Smartphones equipped with SIM cards, mobile data plans, microSD cards (16 GB capacity), and external battery packs. The phones were secured with Android's native encryption and protected by a password. GPS tracker helped to track all devices using a web interface (www.gps-server.net), Barcode Scanner allowed to use barcodes for the identification of enumerators and a parental control application provided a safe contained working environment for enumerators. Interviews were conducted using SurveyCTO Collect on the tablet with data transmitted to a secure SurveyCTO server in a cloud computing environment.

EAs were replaced if security rendered field work unfeasible (Table 12). Replacements were approved by the project manager. Replacement of households were approved by the supervisor after a total of three unsuccessful visits of the household.

Incoming data is processed to create a raw consistent data set. Interviews with wrongly entered EAs were manually corrected. Interviews conducted outside sampled EAs were discarded. For duplicate submissions, only one record is kept.[2] Sampling weights are added to the final dataset and subsequently anonymized at the strata level. Missing values are recoded into four different types of missing values: (i) genuinely missing values coded as "."; (ii) respondent indicated "don't know" coded as ".a"; (iii) respondent refused to respond to the question coded as ".b"; and (iv) missing values due to the questionnaire skipping pattern because the question does not apply to the respondent coded as ".z".

Cleaning Process of Submissions

The total number of interviews submitted through SurveyCTO was 4,590, and the breakdown by zone the following:

- A: 1,06
- B: 1,035
- C: 2,366

[2] Two types of duplicate households are identified. Technical duplicates are defined as duplicate submission of the same interview. They are identified as households with identical GPS data (latitude, longitude and altitude coordinates). Manual duplicates are defined as two interviews conducted with the same household. They are identified by almost identical household rosters. The interview with more information is kept based on manual inspection.

- D: 120

The first step corresponds to a cleaning process identifying general issues and inconsistencies with submissions.

- B: 1 empty household record dropped
- C
    - 3 household records deleted as they were submitted through the web and they were part of a test to monitor scripts before fieldwork
    - 1 submission dropped as it corresponds to a test that a team leader made to check if the GPS of one of his enumerator's phone was working
    - 1 additional household record dropped as it corresponds to an interview completed by the enumerator to check he had the latest version of the questionnaire

Therefore, after making the described adjustment, the number of correct submissions became 4,584, with the following breakdown by region:

- A: 1,069
- B: 1,034
- C: 2,361
- D: 120

The second step excludes submissions from EAs and blocks that were not included in the final sample.

- A: 3 submissions were dropped as they belong to a block that was not included in the final sample
- B: 12 submissions dropped, as they correspond to an EA that was not included in the final sample, since it was a replacement EA that was never executed
- C: 3 interviews dropped because the enumerators selected a wrong EA that had been replaced

Therefore, after making the described adjustment, the number of correct submissions became 4,566, with the following breakdown by region:

- A: 1,066
- B: 1,022
- C: 2,358
- D: 120:

The next step was to validate the acceptance of submissions, for which six criteria were defined and interviews were dropped that failed to meet at least one of them:

- The duration of the interview had to exceed a threshold of 30 minutes
    - 26 submissions were excluded because they were completed in 30 minutes or less
- Random sound bites check, including respondent and enumerator voices. This criterion will be assumed to hold if a specific interview was not checked on this criterion.
    - No interview was removed for this reason
- The interview has GPS coordinates and it was conducted within a buffer area of the correspondent EA
    - 5 interviews did not have GPS coordinates; and
    - 5 were also excluded as the GPS coordinates indicate the interview did not took place within the boundaries of the EA
- If the interview was not completed in the first visit, then the household record for the first visit must be valid using the previous criterions (except for the duration), and both household records must contain a matching GPS positions, with a margin of +/- 10 meters
    - 34 interviews were dropped as they corresponded to a second visit, and the record from the previous visit did not exist or was not valid
    - 26 additional submissions were not considered, as the GPS coordinates of the first visit did not match with those of the subsequent visit

- If the interview corresponds to a replacement household, the record of the original household must be valid, except for the duration of the interview: 67 submissions were not considered as the interview corresponded to a replacement household with an inexistent or invalid record for the original household
- Finally, unsuccessful interviews were discarded; the ones where no one answered the door, there was not a knowledgeable adult present or the respondent did not give permission to continue: 282 submissions were not successful and thus were also excluded

Therefore, at this point, the dataset had a total number of 4,121 submissions, with the following breakdown by region:

- A: 1,031
- B: 929
- C: 2,045
- D: 116

The final step excludes interviews that were incomplete, and thus have several sections without any single response. 4 households did not have any record in the sections corresponding to food consumption, assets and livestock, and thus they were excluded. Therefore, the final dataset includes a total number of 4,117 complete, valid and successful submissions from valid EA and blocks, with the following breakdown by region:

- A: 1,031
- B: 929
- C: 2,041
- D: 116

Sampling weights

This section describes calculation of sample weights for households in the dataset. The sample design was different for some strata due to security volatility. Thus, the methods differ between micro-listing and full-listing. After the sample weights were calculated as described below, they were scaled to the number of households accessible with GPS from the sample frame.

- Full listing: The sample was drawn in a two-stage process for strata 201-204, 301-304 and 1103-1304. Therefore, the weights were calculated based on the sampling probabilities for each sampling stage and for each cluster in the following way:

$$P_{hij} = P_1 P_2 = \frac{EA_j \, H_i}{H_j} \frac{HS_i}{HL_i}$$

such that

P_{hij}: Probability of selecting household *h* in EA *i* of strata *j*
P1: Probability of selecting the EA in stage 1
P2: Probability of selecting the household in stage 2
EA_j: Number of EAs selected in strata *j*
H_i: Number of households estimated in the sample frame for EA *i*
H_j: Number of households estimated in the sample frame in strata *j*
HS_i: Number of households selected in EA *i*
HL_i: Number of households listed in EA *i*

Therefore, the sample weight for each household corresponds to

$$w = 1/P_{hij}$$

- Micro-listing: In strata 101, 105, 205 and 305, the sample was segmented in blocks within EAs, in addition to the two-stage, stratified cluster sampling, design.[3] Therefore, the weights were calculated based on the sampling probabilities for each sampling stage and for each cluster in the following way:

$$P_{hij} = P_1 P_2 P_3 = \frac{EA_j\, H_i}{H_j} \frac{BS_i}{B_i} \frac{HS_i}{H_i}$$

such that

P_{hij}: Probability of selecting household *h* in EA *i* of strata *j*
P_1: Probability of selecting the EA
P_2: Probability of selecting the Block
P_3: Probability of selecting the household
EA_j: Number of EAs selected in strata *j*
H_i: Number of households estimated in the sample frame for EA *i*
H_j: Number of households estimated in the sample frame in strata *j*
BS_i: Number of blocks selected in EA *i*
B_i: Number of blocks in EA *i*
HS_i: Number of households selected in EA *i*
HL_i: Number of households in EA *i*

Therefore, the sample weight for each household corresponds to

$$w = 1/P_{hij}$$

Finally, three types of sampling weights were estimated:

1) Unadjusted weights: Considers all submissions (4,117) and scales the weights so that the sum of the sampling weights by analytical strata matches the total number of accessible households with GPS according to sample frame.

2) Adjusted weights: Considers all submissions (4,117) and scales the weights uniformly so that the sum of the weights by analytical strata matches the total number of households according to the PESS (Table 3).[4]

3) Adjusted weights for consumption and poverty variables: Considers only submissions with consumption data (excludes 53 submissions with missing values in the consumption of food, non-food and durables) and adjusts the weights of the remaining 4,064 submissions according to the following scenarios:

   - If the number of accessible households with GPS (i.e. the sum of weights) is larger than the total number of households according to PESS by analytical strata, then the weights were scaled downwards uniformly to match the total number of households from PESS, which already reflects the re-allocation of the weights from the 53 submissions excluded

   - If the number of accessible households with GPS (i.e. the sum of weights) is smaller than the total number of households according to PESS, then the weights were scaled upwards in two steps: i) re-allocating uniformly the weights from the 53 households excluded across the 4,064 submissions; and then ii) assigning the additional weights needed to match the figures from PESS only to those households or submissions in the bottom 25 percent of the total consumption distribution for the respective analytical strata. The bottom 25 percent were taking up the weight of the additional households to reflect the fact that excluded enumeration areas were not randomly chosen but differed from other enumeration areas by

[3] The segmentation step cancels out as exactly one segment is chosen.
[4] Usually, the household number from the sample frame should reflect the number of households from the last Census. However, the incomplete sample frame necessitated using different (overlapping) data sources for the sample frame. While the probabilities for selection for duplicates are adjusted for already in the EA selection step, the total number of households did not automatically sum up to the number of households from PESS.

inaccessibility due to security and/or infrastructure. As those enumeration areas are expected to be more deprived than the average enumeration area, they were assumed to be similar to the bottom 25 percent.

Table 1: Total number of households by PESS region and analytical strata

PESS Region	Type	Analytical Strata	Number of households
All	IDP	All IDPs	201,963
Banadir	Urban	Mogadishu	187,246
Nugaal	Urban	Garowe	23,119
Bari and Mudug	Urban	Urban Bari and Mudug	140,334
Woqooyi Galbeed	Urban	Hergeiza	123,390
Awdal, Sanaag, Sool and Togdheer	Urban	Urban Awdal, Sanaag, Sool and Togdheer	158,279
Bari, Mudug and Nugaal	Rural	Rural Bari, Mudug and Nugaal	27,684
Awdal, Sanaag, Sool, Togdheer and Woqooyi Galbeed	Rural	Rural Awdal, Sanaag, Sool, Togdheer and Woqooyi Galbeed	61,086

Consumption Aggregate

The nominal household consumption aggregate is the sum of three components, namely 1) expenditures on food items, 2) expenditures on non-food items, and 3) the value of the consumption flow from durable goods:

$$y_i = y_i^f + y_i^n + y_i^d \quad (2)$$

This section describes in detail the cleaning of the recorded data for each of three components. Subsequently, the construction of the consumption aggregate using the Rapid Consumption Methodology is explained as well as the estimation of the consumption flow for durables and the details on the deflator used to calculate spatial price indices.

Moreover, 53 households were assigned a missing value in consumption since 52 of them reported not consuming any food items, and 1 household only reported consuming a non-core food item.

Cleaning

Food

Food expenditure data is cleaned in a four-step process. First, units for reported quantities of consumption and purchase are corrected. Typical mistakes include recorded consumption of 100 kg of a product (like salt) where the correct quantity is grams. These mistakes are corrected using generic rules (Table 5). Then, we introduce a conversion factor to kg for some specific items and units. For example, we recognize that a small piece of bread must have a different weight than a small piece of garlic (Table 6:). The third step consists of correcting issues with the exchange rate selected (Table 7). Finally, outliers are detected using the six cleaning rules below to correct quantities and prices.

- Rule 1
    - Consumption quantities with missing values for items reported as consumed were replaced with item-specific median consumption quantities.
    - Missing purchase quantities and missing prices for items consumed were replaced with item-specific median purchase quantity and item-specific median purchase price.
- Rule 2: Records where the respondent did not know or refused to respond if the household had consumed the item, were replaced with the mean value, including non-consumed records.
- Rule 3: Records with the same value for quantity consumed or quantity purchased and price are assumed to have a data entry error in the price or quantity and are replaced with the item-specific medians.
- Rule 4: Records that have the same value in quantity consumed and quantity purchased but different units are assumed to have a wrong unit either for consumption or purchase. For both quantities, the item-specific distribution of quantities in

kg is calculated to determine the deviation of the entered figure from the median of the distribution. The unit of the quantity that is further away from the median is corrected with the unit of the quantity closer to the median.

- Rule 5:
    - Missing and zero prices are replaced with item-specific medians
    - Outliers for unit prices were identified and replaced with the item-specific median. This includes unit prices in the top 10 percent of the overall cumulative distribution (considering all items), and unit prices below 0.07 USD.
- Rule 6: the consumption value in USD was truncated to the mean plus 3 times the standard deviation of the cumulative distribution for each item, if the record exceeded this threshold.

All medians are estimated at the EA level if a minimum of 5 observations are available excluding previously tagged records. If the minimum number of observations is not met, medians are estimated at the strata-level before proceeding to the survey level. In addition, medians greater than 20 kg and smaller than 0.02 kg were not considered for quantities, while medians greater than 20 USD and smaller than 0.005 USD were also excluded for unit prices.

Non-Food

The non-food dataset only contains values without quantities and units. First, we apply the same cleaning rules for currencies (Table 7), and then the following cleaning rules:

- Rule 1: Zero, missing prices and missing currency for purchased items are replaced with item-specific medians.
- Rule 2: Records where the respondent did not know or refused to respond if the household had purchased the item, were replaced with the mean value, including non-consumed records.
- Rule 3: Prices that are beyond a specific threshold for each recall period (Table 8) are replaced with item-specific medians.
- Rule 4: Prices below the 1 percent and above the 95 percent of the cumulative distribution for each item are replaced with item-specific medians
- Rule 5: the purchase value in USD was truncated to the mean plus 3 times the standard deviation of the cumulative distribution for each item, if the record exceeded this threshold.

The item-specific medians were applied at the EA, strata and survey level as described above.

Durables

For durables, we also apply the same cleaning rules for currencies (Table 7), and then the following cleaning rules:

- Rule 1: Vintages with missing values and greater than 10 years are replaced with item-specific medians.
- Rule 2: Current and purchase prices equal to zero are replaced with item-specific medians.
- Rule 3: Records that have the same figure in current value and purchase price are incorrect. For both, the item-vintage-specific distribution is calculated to determine the deviation of the entered figure from the median. The one that is further away from that median is corrected with the item-year-specific median value.
- Rule 4: Depreciation rates are replaced by the item-specific medians in the following cases:
    - Negative records
    - Depreciation rates in the top 10 percent and vintage of one year
    - Depreciation rates in the bottom 10 percent and a vintage greater or equal to 3 years

- Rule 5: Records with 100 items or more, and those that reported to own a durable good but did not report the number were replaced with the item-specific medians of consumption in USD
- Rule 6: Consumption in the top and bottom 1 percent of the overall distribution were replaced with item-specific medians
- Rule 7: Records where the respondent did not know or refused to respond if the household owned the asset, were replaced with the mean consumption value, including non-consumed records.
- Rule 8: the consumption value in USD was truncated to the mean plus 3 times the standard deviation of the cumulative distribution for each item, if the record exceeded this threshold.

All medians are estimated at the EA level if a minimum of 3 observations are available excluding previously tagged records. If the minimum number of observations is not met, medians are estimated at the strata-level before proceeding to the survey level. Table 9 contains a general overview of consumption of durables, while Table 10 presents the details by item. Table 11 shows the median depreciation rate by durable good.

Rapid Consumption Methodology: Food and Non-Food Aggregates

The survey used the new Rapid Consumption methodology to estimate consumption. A detailed description including an ex post assessment of the methodology is available in a separate document.[5] The rapid survey consumption methodology consists of five main steps. First, core items are selected based on their importance for consumption. Second, the remaining items are partitioned into optional modules. Third, optional modules are assigned to groups of households. After data collection, fourth, consumption of optional modules is imputed for all households. Fifth, the resulting consumption aggregate is used to estimate poverty indicators.

First, core consumption items are selected. Consumption in a country bears some variability but usually a small number of a few dozen items captures the majority of consumption. These items are assigned to the core module, which will be administered to all households. Important items can be identified by its average food share per household or across households. Previous consumption surveys in the same country or consumption shares of neighboring / similar countries can be used to estimate food shares.[6] In the worst case, a random assignment results in a larger standard error but does not introduce a bias.

Second, non-core items are partitioned into optional modules. Different methods can be used for the partitioning into optional modules. In the simplest case, the remaining items are ordered according to their food share and assigned one-by-one while iterating over the optional module in each step. A more sophisticated method takes into account correlation between items and partitions them into orthogonal sets per module. This leads to high correlation between modules supporting the total consumption estimation. Conceptual division into core and optional items is not reflected in the layout of the questionnaire. Rather, all items per household will be grouped into categories of consumption items (like cereals) and different recall periods. Using CAPI, it is straight-forward to hide the modular structure from the enumerator.

Third, optional modules will be assigned to groups of households. Assignment of optional modules will be performed randomly stratified by enumeration areas to ensure appropriate representation of optional modules in each enumeration area. This step is followed by the actual data collection.

Fourth, household consumption will be estimated by imputation. The average consumption of each optional module can be estimated based on the sub-sample of households assigned to the optional module. In the simplest case, a simple average can be estimated. More sophisticated techniques can employ a welfare model based on household characteristics and consumption of the core items. The results presented in this note uses a multiple imputation technique based on a multi-variate normal approximation.

Next, the methodology is formalized and assessed using an *ex post* simulation based on the consumption data from Hergeiza using the Somaliland 2012 household survey (SHS12). Food and non-food consumption for household *i* are estimated by the sum of expenditures for a set of items

[5] Pape & Mistiaen (2015), "Measuring Household Consumption and Poverty in 60 Minutes: The Mogadishu High Frequency Survey", World Bank (2015).
[6] As shown later, the assignment of items to modules is very robust and, thus, even rough estimates of consumption shares are sufficient to inform the assignment without requiring a baseline survey.

$$y_i^f = \sum_{j=1}^{m} y_{ij}^f \text{ and } y_i^n = \sum_{j=1}^{m} y_{ij}^n$$

where y_i^f and y_i^f denote the food and non-food consumption of item *j* in household *i*. As the estimation for food and non-food consumption follows the same principles, we neglect the upper index *f* and *n* in the remainder of this section. The list of items can be partitioned into M+1 modules each with m_k items:

$$y_i = \sum_{k=0}^{M} y_i^{(k)} \text{ with } y_i^{(k)} = \sum_{j=1}^{m_k} y_{ikj}$$

For each household, only the core module $y_i^{(0)}$ and one additional optional module $y_i^{(k^*)}$ are collected.

The item assignment to the modules are based on the SHS12 survey with manual modifications especially to treat 'other' items correctly.[7] The core module was designed to maximize its consumption share resulting in 91 percent and 76 percent of food respectively non-food consumption captures in the core modules (based on SHS12 consumption; Table 2). Optional modules are constructed using an algorithm to assign items iteratively to optional modules so that items are orthogonal within modules and correlated between modules. In each step, an unassigned item with highest consumption share is selected. For each module, total per capita consumption is regressed on household size, the consumption of all assigned items to this module as well as the new unassigned item. The item will be assigned to the module with the highest increase in the R2 relative to the regression excluding the new unassigned item. The sequenced assignment of items based on their consumption share can lead to considerable differences in the captured consumption share across optional modules. Therefore, a parameter is introduced ensuring that in each step of the assignment procedure the difference in the number of assigned items per module does not exceed *d*. Using *d=1* assigns items to modules (almost) maximizing equal consumption share across modules.[8] Increasing *d* puts increasing weight on orthogonality within and correlation between modules. The parameter was set to *d=3* balancing the two objectives.

In each enumeration area, 12 households were interviewed with an ideal partition of three items per optional module.[9] The assignment of optional modules must ensure that a sufficient number of households are assigned to each optional module. Household consumption was then estimated using the core module, the assigned module and estimates for the remaining optional modules

$$\hat{y}_i = y_i^{(0)} + y_i^{(k^*)} + \sum_{k \in K^*} \hat{y}_i^{(k)}$$

where $K^* := \{1, \dots, k^* - 1, k^* + 1, \dots, M\}$ denotes the set of non-assigned optional modules. Consumption of non-assigned optional modules is estimated using multiple imputation techniques taking into account the variation absorbed in the residual term.

Multiple imputation was implemented using multi-variate normal regression based on an EM-like algorithm to iteratively estimate model parameters and missing data. This technique is guaranteed to converge in distribution to the optimal values. An EM algorithm draws missing data from a prior (often non-informative) distribution and runs an OLS to estimate the coefficients. Iteratively, the coefficients are updated based on re-estimation using imputed values for missing data drawn from the posterior distribution of the model. The implemented technique employs a Data-Augmentation (DA) algorithm, which is similar to an EM algorithm but updates parameters in a non-deterministic fashion unlike the EM algorithm. Thus, coefficients are drawn from the parameter posterior distribution rather than chosen by likelihood maximization. Hence, the iterative process is a Monte-Carlo Markov –Chain (MCMC) in the parameter space with convergence to the stationary distribution that averages over the missing data. The distribution for the

[7] Items 'other' are often found to capture remaining items for a food category. Using the Rapid Consumption Methodology, this creates problems as 'other' will include different items depending on which optional module is administered. This can lead to double-counting after the imputation. Therefore, 'other' items are re-formulated and carefully assigned so that double counting cannot occur.

[8] Even with *d=1*, equal consumption share across modules is not maximized because among the modules with the same number of assigned items, the new item will be assigned to the module it's most orthogonal to; rather than to the module with lowest consumption share.

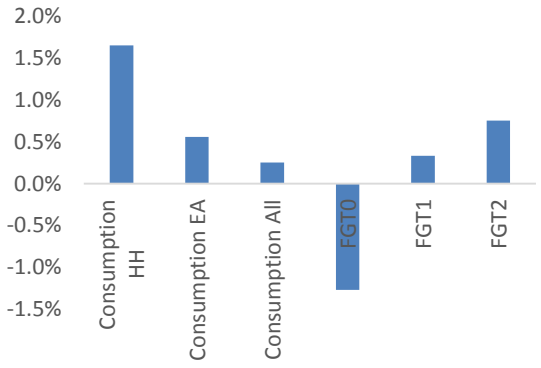
[9] Field work implementation aimed to achieve a balanced partition among optional modules but due to challenges in following the protocol exactly some enumeration areas are not completely balanced.

missing data stabilizes at the exact distribution to be drawn from to retrieve model estimates averaging over the missing value distribution. The DA algorithm usually converges considerably faster than using standard EM algorithms:

$$\hat{y}_i^{(k)} = \beta_0^{(k)} y_i^{(0)} + x_i^T \beta^{(k)} + u_i^{(k)}$$

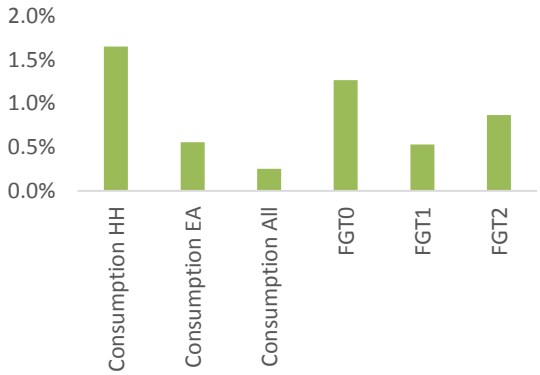
The performance of the estimation technique was assessed based on an *ex post* simulation using the Hergeiza data from SHS12 and mimicking the Rapid Consumption methodology by masking consumption of items that were not administered to households. The results of the simulation were compared with the estimates using the full consumption from SHS12 as reference. The simulation results distinguish between different levels of aggregation to estimate consumption.[10] The methodology generally does not perform well at the household level (HH) but improves considerably already at the enumeration area level (EA) where the average of 12 households is estimated. At the national aggregation level, the Rapid Consumption methodology slightly over-estimates consumption by 0.3 percent. Assessing the standard poverty measures including poverty headcount (FGT0), poverty depth (FGT1) and poverty severity (FGT2), the simulation results show that the Rapid Consumption methodology retrieves estimates within 1.5 percent of the reference measure (Figure 3). Generally, the estimates are robust as suggested by the low standard errors (Figure 4).

Figure 3: Relative bias of simulation results using Rapid Consumption estimation.



Source: Authors' own calculations based on SHS12 data.

Figure 4: Relative standard error of simulation results using Rapid Consumption estimation.



Source: Authors' own calculations based on SHS12 data.

Table 2: Item partitions based on SHS12 and pilot in Mogadishu.

	Food Items				Non-food Items			
	Number of items	Share Hergeiza	Share Mogadishu	Share Mogadishu Imputed	Number of items	Share Hergeiza	Share Mogadishu	Share Mogadishu Imputed
Core	33	91%	64%	54%	26	76%	62%	52%
Module 1	19	3%	9%	16%	15	7%	9%	12%
Module 2	20	2%	14%	14%	15	5%	9%	12%
Module 3	15	2%	5%	6%	15	6%	8%	9%
Module 4	15	2%	8%	9%	15	6%	11%	15%

Source: Authors' own calculations based on SHS12 and Mogadishu Pilot data.

Durable consumption flow

The consumption aggregate includes the consumption flow of durables calculated based on the user-cost approach. The consumption flow distributes the consumption value of the durable over multiple years. The user-cost principle defines the consumption flow of an item as the difference of selling the asset at the beginning and the end of the year as this is the opportunity

[10] The performance of the estimation techniques is presented using the relative bias (mean of the error distribution) and the relative standard error. The relative error is defined as the percentage difference of the estimated consumption and the reference consumption (based on the full consumption module, averaged over all imputations). The relative bias is the average of the relative error. The relative standard error is the standard deviation of the relative error. The simulation is run over different household-module assignments while ensuring that each optional module is assigned equally often to a household per enumeration. The relative bias and the relative standard error are reported across all simulations.

cost of the household for keeping the item. The opportunity cost is composed of the difference in the sales price and the forgone earnings on interest if the asset is sold at the beginning of the year.

If the durable item is sold at the beginning of the year, the household would receive the market price p_t for the item and the interest on the revenue for one year. With i_t denoting the interest rate, the value of the item thus is $p_t(1 + i_t)$. If the item is sold at the end of the year, the household will receive the depreciated value of the item while considering inflation. With π_t being the inflation rate during the year t, the household would obtain $p_t(1 + \pi_t)(1 - \delta)$ with the annual physical or technological depreciation rate denoted as δ assumed constant over time.[11] The difference between these two values is the cost that the household is willing to pay for using the durable good for one year. Hence, the consumption flow is:

$$\mathrm{y}^d = p_t(1 + i_t) - p_t(1 + \pi_t)(1 - \delta) \tag{3}$$

By assuming that $\delta \times \pi_t \cong 0$, the equation simplifies to

$$\mathrm{y}^d = p_t(i_t - \pi_t + \delta) = p_t(r_t + \delta) \tag{4}$$

where r_t is the real market interest rate in period t. Therefore, the consumption flow of an item can be estimated by the current market value p_t, the current real interest rate r_t, and the depreciation rate δ. Assuming an average annual inflation rate π, the depreciation rates δ can be estimated utilizing its relationship to the market price[12]:

$$p_t = p_{t-k}(1 + \pi)^k(1 - \delta)^k \tag{5}$$

The equation can be solved for δ obtaining:

$$\delta = 1 - \left(\frac{p_t}{p_{t-k}}\right)^{\frac{1}{k}} \frac{1}{(1 + \pi)} \tag{6}$$

Based on this equation, item-specific median depreciation rates are estimated assuming an inflation rate of 0.5 percent, a nominal interest rate of 2.0 percent and, thus, a real interest rate of 1.5 percent (Table 11).

For all households owning a durable but did not report the current value of the durable, the item-specific median consumption flow is used. For households that own more than one of the durable, the consumption flow of the newest item is added to the item-specific median of the consumption flow times the number of those items without counting the newest item.[13]

Deflator

Prices fluctuate considerably between regions, thus we calculated spatial price indices using a common food basket and spatial prices to make consumption comparable across regions. The Laspeyres index is chosen as a deflator due to its moderate data requirements. The deflator is calculated by analytical strata areas based on the price data collected by the HFS.

The Laspeyres index reflects the item-weighted relative price differences across products. Item weights are estimated as household-weighted average consumption share across all households before imputation. Based on the democratic approach, consumption shares are calculated at the household level. Core items use total household core consumption as reference while items from optional modules use the total assigned optional module household consumption as reference. The shares are aggregated at the national level (using household weights) and then calibrated by average consumption per module to arrive at item-weights summing to 1. The item-weights are applied to the relative differences of median item prices for each analytical strata. Missing prices are replaced by the item-specific median over all households. A large Laspeyres indicates a high price level deflating consumption stronger than a lower Laspeyres index. The resulting indices show the fluctuation of prices across regions (Table 4).

Table 3: Laspeyres Deflators by Analytical Strata

Analytical Strata	Deflator

[11] Assuming a constant depreciation rate is equivalent to assuming a "radioactive decay" of durable goods (see Deaton and Zaidi, 2002).
[12] In particular π solves the equation $\prod_{i=t-k}^{t}(1 + \pi_i) = (1 + \pi)^k$

[13] The 2016 HFS questionnaire provides information on a) the year of purchase and b) the purchasing price only for the most recent durable owned by the household.

All IDPs	0.923
Mogadishu	0.964
Garowe	0.862
Urban Bari and Mudug	1.107
Hergeiza	1.133
Urban Awdal, Sanaag, Sool and Togdheer	0.922
Rural Bari, Mudug and Nugaal	1.013
Rural Awdal, Sanaag, Sool, Togdheer and Woqooyi Galbeed	1.075

Tables for Cleaning Rules

Table 4: Summary of Unit Cleaning Rules for Food Items

Unit	Condition	Correction	Affected Records[14]
250 ml tin	<=0.03	Multiply by 4	2; 39
Animal back, ribs, shoulder, thigh, head or leg	>=7	Divide by 10	4; 35
Basket or Dengu (2 kg)	>=10	Divide by 10	1,004; 20
Bottle (1 kg)	>=10	Divide by 10	473; 281
Cup (200 g)	>200	Divide by 2	447; 24
Faraasilad (12kg)	>12	Divide by 12	544; 60
Gram (if item corresponds to a spice)	<1	Multiply by 100	115; 5
Gram (if item does not corresponds to a spice)	<1	Multiply by 1,000	69; 19
Haaf (25 kg)	>=25	Divide by 25	357; 921
Heap (700g)	>=0.69	Divide by 7	182; 11
Kilogram	>=100	Divide by 1,000	68; 4
Large bag (50 kg)	>=50	Divide by 50	1; 27
Liter	>=10	Divide by 10	3; 32
Madal/Nus kilo ruba (0.75kg)	>=7.5	Divide by 10	849; 20
Meals (300 g)	>2.1	Divide by 10	366; 208
Packet sealed box/container (500 g)	>=5	Divide by 10	340; 16
Piece (large - 300g)	>=3	Divide by 10	397; 43
Piece (small - 150g)	>=1.5	Divide by 10	95; 5
Rufuc/Jodha (12.5kg)	>=12.5	Divide by 10	37; 15
Saxarad (20kg)	>=20	Divide by 10	312; 793
Small bag (1 kg)	>=10	Divide by 10	110; 8
Teaspoon (10 g)	<0.009	Multiply by 10	45; 4

[14] The first number indicates the number of affected records reported for consumption while the second number states the number of affected records for purchases.

Table 5: Conversion factor to Kg for specific units and items

Item	Unit	Conversion to Kg
Biscuits	Piece – large	0.030
	Piece - small	0.010
Bread	Piece – large	0.400
	Piece - small	0.100
Eggs	Piece – large	0.070
	Piece - small	0.050
Canned fish/shellfish	Piece – large	0.420
	Piece - small	0.140
Grapefruits, lemons, guavas, limes	Piece – large	0.350
	Piece - small	0.100
Milk	Piece – large	0.750
	Piece - small	0.250
Milk powder	Piece – large	0.450
	Piece - small	0.100
	Small bag	1.00
Garlic	Piece – large	0.065
	Piece - small	0.040
Onion	Piece – large	0.150
	Piece - small	0.095
Tomatoes	Piece – large	0.200
	Piece - small	0.110
Bell-pepper	Piece – large	0.150
	Piece - small	0.080
Sweet/ripe bananas	Piece – large	0.110
	Piece - small	0.070
Canned vegetables	Piece – large	0.400
	Piece – small	0.200
Sorghum, flour	Cup	0.200
Cooking oats, corn flakes	Cup	0.200
Other cooked foods from vendors	Small bag	1.00
Purchased/prepared tea/coffee consumed at home	Small bag	0.400
Other spices	Small bag	0.400

Table 6: Summary Cleaning Rules for Currency

Currency	Condition	Correction
Somaliland shillings	Entry in Somaliland shilling	Replace currency to Somali shillings
	Price <=500	Replace currency to Somaliland shillings (Thousands)
	Price>=500,000	Divide by 10
Somali shillings	Entry in Somali shilling	Replace currency to Somaliland shillings
	Price <=500	Replace currency to Somali shillings (Thousands)
	Price>=500,000	Divide by 10
USD	Price >1,000	Replace currency to Somali(land) shillings

Table 7: Threshold for Non-Food Item Expenditure (USD)

Recall period	Min	Max
1 Week	0.05	30
1 Month	0.20	95
3 Months	0.45	200
1 Year	0.80	1,200

Table 8: Consumption of durable goods (per week in current USD)

	SOM Wave 1 All regions	SOM Wave 1 Mogadishu	Pilot Mogadishu
Median	0.74	1.17	1.01
Mean	1.24	1.52	1.91
Sd	1.51	1.49	2.62

Table 9: Median consumption of durable goods (per week in current USD)

Item	SOM Wave 1 All regions	SOM Wave 1 Mogadishu	Pilot Mogadishu
Air conditioner	0.005	0.005	0.041
Bed	N/A	N/A	0.861
Bed with mattress	0.700	0.746	N/A
Car	0.001	0.001	0.001
Cell phone	0.361	0.413	0.430
Chair	0.073	0.072	0.253
Clock	0.028	0.003	0.046
Coffee table (for sitting room)	0.005	0.005	0.106
Computer equipment & accessories	0.020	0.020	2.837
Cupboard, drawers, bureau	0.240	0.240	1.099
Desk	0.047	0.005	0.429
Electric stove or hot plate	0.001	0.001	N/A
Electric or gas stove; hot plate	N/A	N/A	0.012
Electric stove	N/A	N/A	0.004
Fan	0.069	0.064	0.101
Gas stove	0.007	0.007	0.275
Generator	0.000	0.000	0.000
Iron	0.043	0.035	N/A
Kerosene/paraffin stove	0.024	0.007	0.009
Kitchen furniture	0.023	0.015	1.112
Lantern (paraffin)	0.000	0.000	0.002
Lorry	0.000	0.000	0.000
Mattress without bed	0.217	0.212	N/A
Mini-bus	0.000	0.000	0.001
Mortar/pestle	0.016	0.009	0.112
Motorcycle/scooter	0.002	0.002	0.006
Photo camera	0.001	0.001	0.595
Radio ('wireless')	0.021	0.001	0.016
Refrigerator	0.282	0.018	0.267
Satellite dish	0.117	0.008	0.265
Sewing machine	0.002	0.002	0.732
Small solar light	0.003	0.003	N/A
Solar panel	0.000	0.000	0.018
Stove for charcoal	0.032	0.023	0.020
Table	0.042	0.042	0.092
Tape or CD/DVD player; HiFi	0.001	0.001	0.092
Television	0.330	0.278	0.417
Upholstered chair, sofa set	0.019	0.019	2.657
VCR	0.000	0.000	0.000
Washing machine	0.405	0.368	0.557

Table 10: Median depreciation rate of durables goods

Item	SOM Wave 1 All	SOM Wave 1 Mogadishu	Pilot Mogadishu	Wave 1: Awdal, Sanaag, Sool, Togdheer and Woqooyi Galbeed	SHS12
Air conditioner	0.278	0.241	0.210	0.134	0.145
Bed	N/A	N/A	0.364	N/A	0.088
Bed with mattress	0.172	0.172	N/A	0.172	N/A
Car	0.118	0.118	0.111	0.118	0.066
Cell phone	0.188	0.188	0.296	0.188	0.169
Chair	0.149	0.149	0.371	0.149	0.114
Clock	0.204	0.204	0.228	0.204	0.110
Coffee table (for sitting room)	0.279	0.279	0.329	0.279	0.114
Computer equipment & accessories	0.182	0.240	0.364	0.150	0.204
Cupboard, drawers, bureau	0.150	0.150	0.296	0.150	0.098
Desk	0.134	0.134	0.502	0.134	0.108
Electric stove or hot plate	0.262	0.257	0.005	0.252	N/A
Electric stove	N/A	N/A	0.296	N/A	0.138
Fan	0.131	0.131	0.235	0.131	0.134
Gas stove	0.174	0.135	0.296	0.174	0.333
Generator	N/A	N/A	0.296	N/A	0.127
Iron	0.161	0.161	N/A	0.161	0.110
Kerosene/paraffin stove	0.224	0.224	0.296	0.224	0.210
Kitchen furniture	0.188	0.188	0.393	0.188	0.101
Lantern (paraffin)	0.064	N/A	0.067	0.064	0.114
Lorry	0.154	N/A	0.296	0.154	0.052
Mattress without bed	0.185	0.185	N/A	0.185	N/A
Mini-bus	0.153	0.172	0.296	0.153	0.039
Mortar/pestle	0.210	0.210	0.254	0.210	0.114
Motorcycle/scooter	0.172	0.172	0.138	N/A	N/A
Photo camera	0.134	0.134	0.296	0.122	0.171
Radio ('wireless')	0.210	0.210	0.337	0.210	0.134
Refrigerator	0.133	0.133	0.065	0.133	0.096
Satellite dish	0.110	0.110	0.303	0.110	0.097
Sewing machine	0.138	0.114	0.296	0.138	0.134
Small solar light	0.296	N/A	N/A	0.471	N/A
Solar panel	0.005	0.038	0.296	0.005	0.110
Stove for charcoal	0.226	0.226	0.337	0.254	0.188
Table	0.157	0.157	0.296	0.160	0.114
Tape or CD/DVD player; HiFi	0.172	N/A	0.138	0.172	0.092
Television	0.131	0.131	0.240	0.131	0.099
Upholstered chair, sofa set	0.168	0.168	0.289	0.168	0.101
VCR	0.166	0.488	0.296	0.130	0.092
Washing machine	0.138	0.138	0.171	0.138	0.114

Table 11: Sample size calculation, number of replacement and final sample.[15]

Strata	Strata Code New	Number of EAs	Number of accessible EAs	Number of accessible EAs with GPS	Number of house-holds	Number of house-holds in accessible EAs with GPS	Percent in Sample	N	"weights"	cons-umptio n	standard deviation	n	Design Effect	(NhSh)	"optimum" allocation	rounded optimum	clusters	est. standard error	relative standard error	Re-place-ments	Post Sample
A - Mogadishu	101	1347	1299	136592	136592	131914	97%	131,914	0.146	512.9	38.4	676	3.327	5,062,022	558.88	564	47	5.4	0.01	27	67
A - IDPs	105							33,333	0.037	97.9	54.0	122	2.858	1,801,172	198.86	204	17	10.8	0.11	10	17
B - Garowe	201	149	149	149	16351	16351	100%	16,351	0.018	484.6	59.4	573	1.091	971,660	107.28	108	9	6.2	0.01	1	9
B - Urban Centers	202	111	111	111	12534	12534	100%	12,534	0.014	484.6	59.4	573	1.091	744,834	82.23	84	7	7.1	0.01	1	7
B - Other Urban	203	1426	1399	1230	160891	142339	88%	142,339	0.157	386.0	27.8	279	2.130	3,962,628	437.50	432	36	2.9	0.01	6	36
B - Rural	204	1230	1067	475	101226	46235	46%	46,235	0.051	347.1	30.9	873	2.406	1,428,335	157.70	156	13	6.0	0.02	6	13
B - IDPs	205							21,500	0.024	97.9	54.0	122	2.858	1,161,756	128.27	132	11	13.4	0.14	3	11
C - Hergeiza	301	1617	1617	1617	139345	139345	100%	139,345	0.154	484.6	59.4	573	1.091	8,280,591	914.24	912	76	2.1	0.00	5	76
C - Urban Centers	302	1071	1071	1071	114435	114435	100%	114,435	0.126	386.0	27.8	279	2.130	3,185,798	351.73	348	29	3.2	0.01	1	29
C - Other Urban	303	268	250	237	26294	23224	88%	23,224	0.026	386.0	27.8	279	2.130	646,542	71.38	72	6	7.0	0.02	0	6
C - Rural	304	1296	1218	1013	241531	185700	77%	185,700	0.205	347.1	30.9	873	2.406	5,736,817	633.38	636	53	2.9	0.01	98	53
C - IDPs	305							14,167	0.016	97.9	54.0	122	2.858	765,498	84.52	84	7	16.8	0.17	0	7
D - Sanaag Urban	1103	57	57	57	6088	6088	100%	6,088	0.007	386.0	27.8	279	2.130	169,486	18.71	24	2	12.1	0.03	0	2
D - Sanaag Rural	1104	43	43	43	5131	5131	100%	5,131	0.006	347.1	30.9	873	2.406	158,512	17.50	12	2	15.2	0.04	4	0
D - Sool Urban	1203	10	10	10	1352	1352	100%	1,352	0.001	386.0	27.8	279	2.130	37,639	4.16	12	2	12.1	0.03	4	2
D - Sool Rural	1204	29	29	10	2387	919	39%	919	0.001	347.1	30.9	873	2.406	28,391	3.13	12	2	15.2	0.04	4	3
D - Togdheer Urban	1303	128	128	128	11064	11064	100%	11,064	0.012	386.0	27.8	279	2.130	308,015	34.01	36	3	9.9	0.03	2	3
D - Togdheer Rural	1304	8	8	2	541	150	28%	150	0.000	347.1	30.9	873	2.406	4,634	0.51	12	2	15.2	0.04	2	0
Total		**8790**	**8456**	**142745**	**975762**	**836781**		905,781		369.8	288.1	8124	36.65526	34,454,327	3,804	3,840	324	1.4	0.004	174	341

[15] Note that the number of (accessible) households does not resemble necessarily the number of PESS households due to the merging of multiple data sources. Therefore, sample weights were adjusted accordingly to scale with PESS household estimates.