Reall Household Income and Affordability Calculator

Methodology Paper: Malawi 2019

This document outlines the methodology used to process the Malawi Fifth Integrated Household Survey 2019-2020 data for use in Reall's Household Income and Affordability Calculator. The calculator is available at www.reall.net/calculator.

Background and Data Source

This study uses the Fifth Integrated Household Survey (IHS5) 2019–2020[1], implemented by Malawi's National Statistical Office (NSO) under the LSMS/LSMS-ISA program. IHS5 is nationally representative (urban/rural, region, and district), fielded April 2019–April 2020, and collected via CAPI. The survey follows a stratified two-stage design based on the 2018 Population and Housing Census frame: enumeration areas (EAs) were selected with PPS within each district/stratum, and 16 households per EA were sampled. In total, 12,288 households were selected from 768 EAs; due to COVID-19 restrictions, 51 EAs could not be visited, yielding a 93% response rate. Household weights (inverse of selection probabilities) must be applied to obtain estimates representative at national, urban/rural, regional, and district levels.

IHS5 comprises a multi-topic household questionnaire, an agriculture questionnaire (for farming/livestock households), a fisheries questionnaire, a community questionnaire, and a market survey, enabling analysis of consumption, income, labor, health, education, housing, assets, food security, shocks, smallholder production, and local prices.

[1] https://microdata.worldbank.org/index.php/catalog/3818/study-description

Dataset Creation

Input files used:
The following data files and mappings were used:

- **ihs5_consumption_aggregate.dta** – main welfare file with total household consumption (rexpagg), housing-consumption item (rexp_cat041), household size (hhsize), survey weight (hh_wgt), urban/rural area (area), and geography (region, district).
- **HH_MOD_E.dta** – individual employment/activity flags used to derive the number of earners per household.
- **region_mapping_mw.xlsx** – mapping from region to **RegionName**.
- **district_mapping_mw.xlsx** – mapping from district to **DistrictName**.

Variables extracted and transformations performed:

- **HH_exp**: Monthly total household expenditure = rexpagg / 12.
- **HH_size**: Household size = hhsize.
- **no_of_earners**: count of members with **any work activity in last 12 months** (any of hh_e06_1a–hh_e06_6 == 1), summed per household.
- **urban_rural**: "Urban" if area == 1, otherwise "Rural".
- **Percent_Income_Spent_on_Housing**: share of monthly expenditure spent on **actual rent** = (rexp_cat041 / 12) / HH_exp * 100 where rexp_cat041 is **Actual rents for housing (annual, real April 2019 prices)**.
- **norm_weight**: Normalized survey weight = round(hh_wgt / 100).
- **Location**: RegionName (via region_mapping_mw.xlsx).
- **City**: DistrictName (via district_mapping_mw.xlsx).
- Records with missing HH_exp, hhsize, or norm_weight were dropped to ensure consistent analysis.

Notes:

- Household IDs were harmonised (HHID → hhid) to join welfare and individual files.
- Region/district codes were aligned with their lookup tables before merging to ensure clean **Location** and **City** labels.

```
# === Malawi 2019 Data Processing ===

# Step 1: Load consumption data

welfare <-
read_dta("C:/Users/Pratyusha/Documents/Reall/Data/Malawi/ihs5_consumption_aggregate.dta")
%>%

 select(HHID, hhsize, rexpagg, rexp_cat041, hh_wgt, area, region, district) %>%

 rename(hhid = HHID)

# Step 2: Load individual-level data and compute earners

indiv <- read_dta("C:/Users/Pratyusha/Documents/Reall/Data/Malawi/HH_MOD_E.dta") %>%

 select(HHID, hh_e06_1a, hh_e06_1b, hh_e06_1c, hh_e06_2,

       hh_e06_3, hh_e06_4, hh_e06_5, hh_e06_6) %>%

 mutate(earner = if_else(

   hh_e06_1a == 1 | hh_e06_1b == 1 | hh_e06_1c == 1 |

   hh_e06_2 == 1  | hh_e06_3 == 1  | hh_e06_4 == 1 | hh_e06_6 == 1, 1, 0, missing = 0)) %>%

 group_by(HHID) %>%

 summarise(no_of_earners = sum(earner, na.rm = TRUE), .groups = "drop") %>%

 rename(hhid = HHID)

# Step 3: Load region and district mapping

region_map <-
read_excel("C:/Users/Pratyusha/Documents/Reall/Data/Malawi/region_mapping_mw.xlsx") %>%

 rename(region = region, RegionName = RegionName)

district_map <-
read_excel("C:/Users/Pratyusha/Documents/Reall/Data/Malawi/district_mapping_mw.xlsx") %>%
```

```
# Step 4: Merge and transform data
merged_data <- welfare %>%
  left_join(indiv, by = "hhid") %>%
  left_join(region_map, by = "region") %>%
  left_join(district_map, by = "district") %>%
  mutate(
    HH_exp = rexpagg / 12,  # Convert annual to monthly
    norm_weight = round(hh_wgt / 100),
    Country = "Malawi",
    Year = 2019,
    City = DistrictName,
    Location = RegionName,
    urban_rural = ifelse(area == 1, "Urban", "Rural"),
    Percent_Income_Spent_on_Housing = (rexp_cat041 / 12) / HH_exp * 100
  ) %>% drop_na(HH_exp, hhsize, norm_weight)
```

Expanding Dataset by Weight

- Each household was replicated by its **normalized weight**: norm_weight = round(hh_wgt / 100).
- Expansion used uncount(weights = norm_weight) to approximate the population distribution for distributional statistics.
- **Note:** Sample sizes are computed on the **unexpanded** data; quantiles are computed on the **expanded** data.

```
# Step 5: Expand based on survey weights
expanded_data <- merged_data %>%
  uncount(weights = norm_weight)
```

Calculating Sample Sizes

We computed the number of valid survey households (i.e., sample size) for each **region-province-urban/rural** group. Two sets of sample sizes were calculated:

- **Urban/Rural disaggregated**
- **Combined (Urban + Rural**
- Counts represent **interviewed households** in each stratum.

```
# Step 6: Compute sample size by area
sample_size_mw <- merged_data %>%
  group_by(Location, City, urban_rural, Year) %>%
  summarise(sample_size = n(), .groups = 'drop') %>%
  mutate(Country = "Malawi")

sample_size_mw_u_r <- merged_data %>%
  group_by(Location, City, urban_rural = "All", Year) %>%
  summarise(sample_size = n(), .groups = 'drop') %>%
  mutate(Country = "Malawi")
```

Calculating Percentiles

Using a defined quantile probability range (1% to 99%), percentile values were computed for:

- HH_exp (household expenditure)
- HH_size (household size)
- Percent_Income_Spent_on_Housing

These were grouped by:

- Location (region)
- City (province)
- urban_rural (Urban / Rural / All)

Each quantile represents a snapshot of affordability for that income segment in a given location and classification.

These were calculated for both urban/rural disaggregated and combined cases.

```
# Step 8: Compute quantiles
mw_quantiles <- expanded_data %>%
  group_by(Location, City, urban_rural) %>%
  group_modify(~ calculate_quantiles(.x, quantile_probs)) %>%
  ungroup() %>%
  mutate(Country = "Malawi", Year = 2019)

mw_quantiles_u_r <- expanded_data %>%
  group_by(Location, City, urban_rural = "All") %>%
  group_modify(~ calculate_quantiles(.x, quantile_probs)) %>%
  ungroup() %>%
  mutate(Country = "Malawi", Year = 2019)
```

Final Dataset Preparation

- **Quantile and sample size datasets were merged:** The disaggregated and combined quantile datasets (mw_quantiles, mw_quantiles_u_r) were merged with their corresponding sample size datasets (sample_size_mw, sample_size_mw_u_r) using left_join() on Country, Location, City, urban_rural, and Year. This ensured that each quantile summary contained the correct sample size information.
- **Columns were aligned using a common schema:** The resulting dataset (mw_final_combined) was passed through the align_columns() function to match the structure of the master schema (common_columns), maintaining consistency across all country datasets.
- **Combine relevant sample size datasets:** The sample size data (both disaggregated and combined) was appended to the unified dataset combined_sample_size using bind_rows(), enabling a consolidated view of sampling strength across countries and years.
- **Join combined sample size to summary dataset:** Finally, summary_dataset was updated by appending mw_final_combined using bind_rows(), ensuring that Malawi 2019 data was incorporated into the master affordability summary.

```
# Step 9: Merge quantiles and sample size
mw_final <- mw_quantiles %>%
  left_join(sample_size_mw, by = c("Country", "Location", "City", "urban_rural", "Year")) %>%
  select(Country, Location, Year, Quantile, urban_rural, sample_size,
         HH_exp, no_of_earners, HH_size, Percent_Income_Spent_on_Housing, City)
mw_final_combined <- bind_rows(
  mw_final,
  mw_quantiles_u_r %>%
    left_join(sample_size_mw_u_r, by = c("Country", "Location", "City", "urban_rural", "Year")) %>%
    select(Country, Location, Year, Quantile, urban_rural, sample_size,
           HH_exp, no_of_earners, HH_size, Percent_Income_Spent_on_Housing, City)
)
# === Malawi 2019 Data Integration into Final Dataset ===
# Align and bind Malawi data to summary dataset
mw_final_combined_aligned <- align_columns(mw_final_combined, common_columns)
summary_dataset <- bind_rows(summary_dataset, mw_final_combined_aligned)
# combined_sample_size <- bind_rows(combined_sample_size, sample_size_mw,
sample_size_mw_u_r)
```

```
combined_sample_size <- bind_rows(
  sample_size_india,  sample_size_india_u_r,  sample_size_uganda,  sample_size_uganda_u_r,
  sample_size_nigeria,  sample_size_nigeria_u_r,   sample_size_pakistan,
sample_size_pakistan_u_r,  sample_size_kenya,   sample_size_kenya_u_r,
  sample_size_kenya_2020,  sample_size_kenya_2020_u_r,  sample_size_india_2022,
  sample_size_india_u_r_2022,  sample_size_ph_2023,  sample_size_ph_u_r_2023,
sample_size_ph_2018,  sample_size_ph_u_r_2018,  sample_size_rw,  sample_size_rw_u_r,
sample_size_mw, sample_size_mw_u_r, sample_size_ci, sample_size_ci_u_r,sample_size_bf,
 sample_size_bf_u_r, sample_size_kenya_2021,  sample_size_all_kenya_2021
)
summary_dataset <- summary_dataset %>%
  left_join(combined_sample_size, by = c("Country","urban_rural","Location","City","Year"))
```

In addition to the quantile-level summaries, we also generated **aggregated metrics at both state and national levels** to support broader comparisons within and across countries:

- **State-Level Aggregation:**
  The dataset was grouped by Country, Location (region), Year, Quantile, and urban_rural. Within each group, we calculated average household expenditure, household size, and percentage of income spent on housing. The sample sizes were summed to reflect total representation. Kenya was excluded from this step as it is handled separately due to unique formatting.
- **National-Level Aggregation:**
  All regions within each country were collapsed into a national summary by grouping on Country, Year, Quantile, and urban_rural. This allowed for country-level benchmarking across time and income segments.

Combined Dataset Creation:
The original quantile-level data, state-level aggregates, and national-level aggregates were combined using bind_rows() into a single unified dataset.

```
state_aggregated <- summary_dataset %>%
  group_by(Country, Location, Year, Quantile,urban_rural) %>%
    summarise(
    sample_size = sum(sample_size),
    HH_exp = mean(HH_exp, na.rm = TRUE),
    no_of_earners = mean(no_of_earners, na.rm = TRUE),
    HH_size = mean(HH_size, na.rm = TRUE),
    Percent_Income_Spent_on_Housing = mean(Percent_Income_Spent_on_Housing, na.rm =
TRUE)
  ) %>%
 ungroup()
%>%
mutate(City = "All")
```

```
# ----------------------------------------
# National-Level Aggregation
# ----------------------------------------
# Aggregate data for the national level by combining all states within each country
national_aggregated <- summary_dataset %>%
  group_by(Country, Year, Quantile,urban_rural) %>%
  summarise(
    sample_size = sum(sample_size),
    HH_exp = mean(HH_exp, na.rm = TRUE),
    no_of_earners = mean(no_of_earners, na.rm = TRUE),
    HH_size = mean(HH_size, na.rm = TRUE)
    #Percent_Income_Spent_on_Housing = mean(Percent_Income_Spent_on_Housing, na.rm =
TRUE)
  ) %>%
  ungroup() %>%
  mutate(Location = "All",
         City = "All")
# Combine State and National Data
# Combine both the state-level and national-level data
final_dataset <- bind_rows(state_aggregated, national_aggregated,summary_dataset)
```

Calculating Income Growth

All household income data for Malawi was projected using **GDP per capita growth rates**.

For the historical period **2016–2024**, actual annual GDP per capita growth rates (sourced from the World Bank)[2] were applied directly to the dataset. For the forward-looking period **2025–2040**, a **median annual growth rate** of 1.603% was used.

Year	GDP per capita growth rate (%)
2000	-0.83083
2001	-7.24641
2002	-0.76007
2003	3.092548
2004	2.72895
2005	0.539881
2006	1.854815
2007	6.551347
2008	4.558556
2009	5.172349
2010	3.769618
2011	1.917989
2012	-0.9836
2013	2.460764
2014	2.703864
2015	0.007925
2016	-0.25637
2017	1.22138
2018	1.602997
2019	2.95151
2020	-1.82835
2021	1.880764
2022	-1.63749
2023	-0.69416
2024	-0.76312
2025	1.602997

[2] https://data.worldbank.org/indicator/NY.GDP.PCAP.KD.ZG