Reall Household Income and Affordability Calculator

Methodology Paper: Tanzania 2021

This document outlines the methodology used to process the 2020/21 Tanzania National Panel Survey (NPS-R5) data for use in Reall's Household Income and Affordability Calculator. The calculator is available at www.reall.net/calculator

Background and Data Source

The Tanzania National Panel Survey (NPS)[1] is a multi-topic household survey implemented by the Tanzania National Bureau of Statistics (NBS)[2] in collaboration with the World Bank's LSMS–ISA program. The 2020/21 Round 5 survey is part of a long-running panel series (Rounds 1–5, spanning 2008–2021) that collects data on household demographics, consumption, labour, agriculture, and living conditions.

The dataset is nationally representative, based on a stratified multi-stage sampling design, with stratification by rural/urban and region. Data are provided at the household level and include survey weights (y5_crossweight) and location identifiers such as region (hh_a01_1) and district (hh_a02_1).

Two primary files were used in this study:

- **Household Roster (hh_sec_a.dta)** – provided location identifiers (region and district codes), survey weights, and urban/rural classification.
- **Consumption Aggregate (consumption_real_y5.dta)** – contained estimates of household consumption and household size, derived from detailed expenditure modules.

A region lookup table (tanzania_region_lookup.xlsx) was created externally to harmonise region codes with official names and ensure consistency across datasets.

Key Limitation

While the Tanzania NPS contains district-level codes, no official concordance exists to map these codes to administrative units across survey waves. The World Bank's 2022 public release of the data did not provide documentation linking Round 5 district codes with earlier rounds or to official administrative boundaries. This constraint made district-level disaggregation unreliable and forced us to restrict analysis to the regional level only.

[1] https://microdata.worldbank.org/index.php/catalog/5639/data-dictionary
[2] https://www.nbs.go.tz/

Dataset Creation

Input files used:

- hh_sec_a.dta – household characteristics and location identifiers
- consumption_real_y5.dta – household expenditure and size aggregates
- tanzania_region_lookup.xlsx – mapping of region codes to region names

Variables extracted and transformations performed:

- HH_exp: Total household expenditure (expm), converted from annual to monthly by dividing by 12
- HH_size: Household size (hhsize)
- urban_rural: Derived from y5_rural (1 = Urban, 2 = Rural)
- Weight: Cross-sectional survey weight (y5_crossweight), normalised as norm_weight = round(Weight / 100)
- no_of_earners: Not available in this dataset, set to NA
- Percent_Income_Spent_on_Housing: Not available in this dataset, set to NA
- RegionCode: cleaned as integers, merged to region names using lookup

Location variable was created using region names, with City coded as “All” for this dataset.

```
# Step 3: Clean location dataset
hh_tz_clean <- hh_tz %>%
  select(
    y5_hhid, hh_a01_1, hh_a02_1, y5_rural, y5_crossweight
  ) %>%
  rename(
    RegionCode = hh_a01_1,
    DistrictCode = hh_a02_1,
    Weight = y5_crossweight,
    urban_rural = y5_rural
  ) %>%
  mutate(
    urban_rural = ifelse(urban_rural == 1, "Urban", "Rural"))
# Step 4: Clean consumption dataset
tz_cons_clean <- cons_tz %>%
  transmute(
    y5_hhid,
    HH_size = hhsize,
    HH_exp = expm / 12,   # annual → monthly
    Percent_Income_Spent_on_Housing = NA_real_)
# Step 5: Merge datasets + region lookup
tz_merged <- hh_tz_clean %>%
  left_join(tz_cons_clean, by = "y5_hhid") %>%
  left_join(region_lookup_tz, by = "RegionCode") %>%
  mutate(
    norm_weight = round(Weight / 100),
    Country = "Tanzania",
    Year = 2021,
    Location = stringr::str_to_title(RegionName),  # use region names
    City = "All",
    no_of_earners = NA_real_
  ) %>%
  drop_na(HH_exp, HH_size, norm_weight)
```

Expanding Dataset by Weight

To reflect actual population distributions, the dataset was expanded by household weights (weight). Each household record was replicated according to its normalized weight (scaled by 1/100). This ensured that quantile calculations represented Tanzania's weighted population, not just the survey sample.

```
# Step 7: Expand data by survey weights
tz_expanded <- tz_merged %>%
  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:

1. **Urban/Rural disaggregated**
2. **Combined Urban & Rural (All)**

```
# Step 6: Sample size summaries
sample_size_tz <- tz_merged %>%
  group_by(Location, City, urban_rural, Year) %>%
  summarise(sample_size = n(), .groups = "drop") %>%
  mutate(Country = "Tanzania")


sample_size_tz_u_r <- tz_merged %>%
  group_by(Location, City, urban_rural = "All", Year) %>%
  summarise(sample_size = n(), .groups = "drop") %>%
  mutate(Country = "Tanzania")
```

Calculating Percentiles

Using quantile probabilities (1% to 99%), percentile values were computed for:

- Household expenditure (HH_exp)

- Household size (HH_size)

Grouped by:

- Location (Region)
- City = “All”
- Urban/Rural (Urban / Rural / All)

Percentiles provide the basis for affordability comparisons across the income distribution.

```
# Step 8: Quantile calculations
quantile_probs <- seq(0.01, 0.99, by = 0.01)

tz_quantiles <- tz_expanded %>%
  group_by(Location, City, urban_rural) %>%
  group_modify(~ calculate_quantiles(.x, quantile_probs)) %>%
  ungroup() %>%
  mutate(Country = "Tanzania", Year = 2021)

tz_quantiles_u_r <- tz_expanded %>%
  group_by(Location, City, urban_rural = "All") %>%
  group_modify(~ calculate_quantiles(.x, quantile_probs)) %>%
  ungroup() %>%
  mutate(Country = "Tanzania", Year = 2021)
```

Final Dataset Preparation

The final Tanzania dataset was prepared as follows:

1. **Merge quantiles with sample sizes** – Urban/Rural and combined quantile datasets were merged with their corresponding sample size summaries using left_join().
2. **Combine disaggregated and all records** – Quantile datasets were appended together to provide both disaggregated and pooled views of household expenditure.
3. **Schema alignment** – The dataset was passed through the align_columns() function to match the global Reall schema (common_columns), ensuring consistency with other countries.
4. **Integration into master dataset** – The final Tanzania dataset (tz_final_aligned) was appended into the global affordability summary (summary_dataset) using bind_rows(). Sample size

tables (sample_size_tz, sample_size_tz_u_r) were similarly added to the consolidated sample size tracker (combined_sample_size).

```
# Step 9: Combine quantiles with sample sizes
tz_final <- tz_quantiles %>%
  left_join(sample_size_tz, 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)
tz_final_combined <- bind_rows(
  tz_final,
  tz_quantiles_u_r %>%
    left_join(sample_size_tz_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)
)
# Step 10: Align with global schema & integrate into master datasets
tz_final_aligned <- align_columns(tz_final_combined, common_columns)
summary_dataset <- bind_rows(summary_dataset, tz_final_aligned)
```

```
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 Tanzania was projected using GDP per capita growth rates.

For the historical period 2016–2023, actual annual GDP per capita growth rates (sourced from the World Bank) [3] were applied directly to the dataset. For the forward-looking period 2024–2040, a median annual growth rate of 2.145% was used.

Year	GDP per capita growth rate (%)
2000	1.599427
2001	3.250579
2002	4.363687
2003	3.893443
2004	4.631631
2005	4.540186
2006	3.595549
2007	3.834933
2008	2.873538
2009	2.709366
2010	3.670415
2011	4.704850
2012	1.538920
2013	3.627228
2014	3.511479
2015	2.770382
2016	3.304007
2017	3.270743
2018	2.208610
2019	2.710059
2020	-0.998988
2021	1.254018
2022	1.542998
2023	2.080992
2024	2.144801
2025	2.144801

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