Reall Household Income and Affordability Calculator

Methodology Paper: India 2018

This paper lays out the methodology used to develop the 2018 India data for Reall's Household Income and Affordability Calculator, which can be found at www.reall.net/calculator.

The National Sample Survey (NSS) is produced by the Ministry of Statistics and Programme Implementation in India. NSS 76th Round is a nationally-representative household survey focused on presenting income and consumption expenditure data, with the current round covering 106,838 households. The full published report and tables[1] provide a broad overview of the landscape, with the raw data freely available from the same page for more detailed analysis.

The NSS focuses on consumption expenditure, so this is used as a proxy for income. This is generally seen as good practice, providing more reliable responses from participants. However, it may also result in slightly lower figures than would have been collected for income, particularly for wealthier households. There are various reasons for this, but one of the most significant is that expenditure data generally excludes regular and irregular household savings.

Data is split into separate data files for different sections of the survey. Key sections of the NSS survey used for this work are 'R76120L02 – Demographic and other particulars of household members', which provides employment data for each household members; and 'R76120L03 – Household characteristics', which provides household size and expenditure. Weightings are provided in all files, with location codes in Appendix II (List of NSS Regions and their Composition).[2]

Methodology

The following actions were undertaken on these datasets:

Creating Dataset

- Appendix II
  - Extract all PDF data into Excel
- R76120L02
  - Create new [is_earner] column, in which value is 1 for each household member where [nco2digitcode] (employment type code) is greater than 0.
  - Group all data by [HHID] (household ID), summing number of [is_earner] household members in a new [no_of_earners] column
- R76120L03
  - Convert the following columns to integers: [Sector], [NSS_Region], [District], [HH_size], [Multiplier]

[1] http://microdata.gov.in/nada43/index.php/catalog/153
[2] http://microdata.gov.in/nada43/index.php/catalog/153/download/1956

- Convert [Sector] to new [urban_rural] column, where values of 2 are defined as 'Urban' and others as 'Rural'
- Merge [no_of_earners] from R76120L02 by [HHID]
- Merge [state/ut], [NSS region code], [name of district], [code] from Appendix II, where [NSS_Region] equals [NSS region code]

```
# Calculate the number of earners per household
LO2_dataset <- LO2_dataset %>%
        mutate(is_earner = ifelse(nco2digitcode > 0, 1,0)) %>% #ifelse(is.na(nco2digitcode),
        0, 1)) %>%
        group_by(HHID) %>%
        mutate(no_of_earners = sum(is_earner)) %>%
        ungroup()

# Filter and merge datasets
LO7_dataset <- LO7_dataset %>% filter(monthly_rent != 0)
data <- data %>%
        mutate(urban_rural = ifelse(Sector == 2, "Urban", "Rural")) %>%
        left_join(LO2_dataset[c('HHID', 'no_of_earners')], by = "HHID") %>%
        left_join(LO7_dataset[c("HHID", "monthly_rent")], by = "HHID") %>%
        left_join(Appendix_II[c('state_name', 'district_name', 'district_code','nss_code')], by =
        c('NSS_Region'='nss_code', 'District' = 'district_code'), ) %>%
        mutate(Percent_Income_Spent_on_Housing = (monthly_rent /
        Total_Monthly_expenditure) * 100) %>%
        distinct() %>%
        drop_na(HHID)

# Rename columns for consistency
data <- data %>%
        rename(HH_exp = Total_Monthly_expenditure, HH_size = HH_size, state =
        'state_name', City = 'district_name')
```

Calculating Sample Sizes

- Group dataset by [name of district] and [state/ut] to create sample sizes for each district
- Group dataset by [name of district], [state/ut] and [urban_rural] to create separate urban and rural sample sizes for each district

```
sample_size_india <- data %>%
        group_by(City, state, urban_rural, Year = 2018) %>%
        summarise(sample_size = n())

sample_size_india_u_r <- data %>%
        group_by(City, state, urban_rural = "All", Year = 2018) %>%
        summarise (sample_size = n())
```

Calculating Percentiles

- Multiply each [HHID] record by [Multiplier] to create full weighted dataset
- Group data by [name of district] and [state/ut], and extract data for records at 1% increments, creating figures for each percentile of every district
- Repeat the step above but also grouping by [urban_rural], creating separate urban and rural percentile figures for each district

```
# Function to calculate quantiles and related statistics
calculate_quantiles <- function(data, quantiles) {
        do.call(rbind, lapply(quantiles, function(q) {
                data.frame(
                        Quantile = q * 100,  # Convert quantile probability to percentage
                        HH_exp = quantile(data$HH_exp, probs = q, na.rm = TRUE)
                        )
                }))
        }

all_data_india <- data %>%
uncount(weights = norm_weight)

# Calculate quantiles for each location
quantile_probs <- seq(0.01, 0.99, by = 0.01)
India_location_quantiles <- all_data_india %>%
        group_by(Country = "India", urban_rural, Location = state, City, Year = 2018) %>%
        group_modify(~ calculate_quantiles(.x, quantile_probs)) %>%
        ungroup()

# Additional mixed urban/rural quantiles for each location
quantile_probs <- seq(0.01, 0.99, by = 0.01)
india_location_quantiles_u_r <- all_data_india %>%
        group_by(Country = "India", urban_rural = "All", Location = state, City, Year = 2018)
        %>%
        group_modify(~ calculate_quantiles(.x, quantile_probs)) %>%
        ungroup()
```

Creating Final Dataset

- Define and align common columns to enable merging of quantiles datasets
- Combine relevant ‘quantiles’ and ‘quantiles_u_r’ datasets into a single ‘summary_dataset’
- Define and align common columns to enable merging of sample size datasets
- Combine relevant ‘sample_size’ datasets into a single ‘combined_sample_size’
- Join ‘combined_sample_size’ dataset to ‘summary_dataset’
- Create a ‘state_aggregated’ version of ‘summary_dataset’ by grouping all household data by state
- Create a ‘national_aggregated’ version of ‘summary_dataset’ by grouping all household data by country

- Join ‘state_aggregated’ and ‘national_aggregated’ datasets to ‘summary_dataset’

```
# Define common columns for final summary
common_columns <- c("Country", "urban_rural", "Location", "City", "Year", "Quantile",
"HH_exp", "no_of_earners", "HH_size", "Percent_Income_Spent_on_Housing")

# Function to align columns across datasets
align_columns <- function(df, common_cols) {
        df %>%
        mutate(across(setdiff(common_cols, colnames(df)), ~ NA)) %>%
        select(all_of(common_cols))
        }

# Align columns and combine all datasets
india_location_quantiles <- align_columns(india_location_quantiles, common_columns)
india_location_quantiles_u_r <- align_columns(india_location_quantiles_u_r,
common_columns)

# Combine all country datasets into a single summary dataset
summary_dataset <- bind_rows(india_location_quantiles, india_location_quantiles_u_r)

sample_size_india <- sample_size_india %>%
        rename(Location = state, sample_size = sample_size) %>%
        mutate(Country = "India")

sample_size_india_u_r <- sample_size_india_u_r %>%
        rename(Location = state, sample_size = sample_size) %>%
        mutate(Country = "India")

# Combine the sample size tables into one
combined_sample_size <- bind_rows(
        sample_size_india,
        sample_size_india_u_r,

summary_dataset <- summary_dataset %>%
        left_join(combined_sample_size, by = c ("Country", "urban_rural", "Location", "City",
        "Year"))

# Aggregate data for state level
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),
                ) %>%
        ungroup()%>%
        mutate(City = "All")

# 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)
                ) %>%
        ungroup() %>%
        mutate(Location = "All", City = "All")

# Combine both the state-level and national-level data
final_dataset <- bind_rows(state_aggregated, national_aggregated,summary_dataset)
```

Calculating Inflation

All data is inflated using median annual inflation rates from 2010-23. Median rates were used rather than actual figures to help compensate for large-scale inflation across many economies in 2022 and 2023.

Consumer Price Index inflation rates were taken from the World Bank[3] and consisted of the following figures, creating a final median rate of 6.14%.

CPI	India
2000	4.009436
2001	3.779293
2002	4.297152
2003	3.805859
2004	3.767252
2005	4.246344
2006	5.796523
2007	6.372881
2008	8.349267
2009	10.88235
2010	11.98939
2011	8.911793
2012	9.478997
2013	10.01788
2014	6.665657
2015	4.906973
2016	4.948216
2017	3.328173
2018	3.938826
2019	3.729506
2020	6.623437

[3] https://data.worldbank.org/indicator/FP.CPI.TOTL.ZG

2021	5.131407
2022	6.699034
2023	5.649143