
DRAFT

BASIC INFORMATION

PAPUA NEW GUINEA HOUSEHOLD SURVEY 1996

GUIDE TO THE DATA AND CONSTRUCTED VARIABLES

**JOHN GIBSON
DEPARTMENT OF ECONOMICS
UNIVERSITY OF WAIKATO**

AUGUST 1997

Introduction

The Papua New Guinea Household Survey of 1996 (PNGHS96) is a comprehensive and multipurpose survey. In addition to estimating the value of household consumption, the survey also collected data on demographics, education levels and costs, income sources, the use of health facilities, birth details for young children, diet, housing conditions, agricultural assets, agricultural input spending, and participants' perception of their quality of life. Young children and their parents were weighed and measured. This document describes the data and background material that is supplied on the tape "PNGHS96". This document should be read in conjunction with "Results of the Household Survey Component of the 1996 Poverty Assessment for Papua New Guinea", especially Part I "Measuring the Standard of Living".

The files on the tape "PNGHS96" are arranged into six folders, which are listed in Table 1. These folders contain raw data, data dictionaries, constructed variables, analysis files and text files with original copies of the survey documentation. The raw data are in batch files (*.bch) created by the *IMPS* program, which is used by the PNG National Statistical Office, and is supplied free by the US Census Bureau. The matching data dictionaries are in *.dd files. Analysis files are usually in ASCII format, and are named *.out. For every *.out file there is a matching *.cn file, which contains the instructions that were used by *IMPS* in creating the output file. These *.cn files are vital for understanding the structure of the output file, and for understanding what manipulations were carried out. Additional files with constructed variables are saved as *.wk3, and can be read by any spreadsheet program.

Table 1: Structure of the Data

| Folder Name | Description |
|-------------|---|
| Impsdata | Raw data files and data dictionaries |
| Foodprgs | Program and output files for each food item |
| Othprgs | Program and output files for all data items excluding foods |
| Weights | Sampling weights and survey design files |
| Prices | Raw data, analysis files and summary files from price surveys |
| Text | Copies of all survey documentation |

A note on identification

In PNGHS96 households are uniquely identified by the combination of the codes for:

- Province
- Census Division
- Census Unit
- Dwelling
- Month.

It is particularly important that all five of these variables be used when identifying a household because the survey included a longitudinal component where some of the households were

revisited several months after the initial interviews. For these households, the codes for the first four identifying variables are the same for both sets of observations, and it is only the month of the interview that distinguishes between the initial interviews and the revisit. When calculating summary statistics users will almost always want each household to occur only once. The identification of which observation on a household in the longitudinal sub-sample should be used can be found in the files in the *Weights* folder. Paragraph 14 of “Measuring the Standard of Living” has more details on the choice of which observation to use on a household with two sets of observations available.

1. IMPSDATA

There are three batch files in this folder:

1. *Full.bch* contains all of the raw data except for the data on adult anthropometrics and the data collected at community level. Note that *Full.bch* is a concatenated file. The data dictionary for this batch file is *househld.dd*.
2. *Anthropo.bch* contains the data on adult anthropometrics (repeated observations of heights and weights of parents who had children aged five years and below). The data dictionary for this batch file is *anthropo.dd*.
3. *Commqst.bch* contains the data from the community questionnaires, with the exception of the data collected by the price surveys, which can be found in the folder *Prices*. The data dictionary for this file is *commqst.dd*.

An example of the data dictionaries is printed below. These data dictionaries contain all codes used throughout the survey, and are essential when attempting to interpret any of the **.cn* program files.

| Record Name: | HCF-EDUCATION | | | | Record Type: | 01 | |
|------------------|---------------|----------|------|------|--------------|--------|--|
| Item (occurs) | Data | Item | | | | | |
| Subitem (occurs) | Type | Position | Len. | Dec. | Value Name | Values | |
| P00-PERSON-NO | N | 17-18 | 2 | 0 | | | |
| P02-SEX | N | 19 | 1 | 0 | Male | 1 | |
| | | | | | Female | 2 | |
| P03-AGE | N | 20-21 | 2 | 0 | Age | 00:99 | |
| P04-RELATIONSHIP | N | 22-23 | 2 | 0 | Relationship | 01:10 | |
| P05-PROV-BORN | N | 24-25 | 2 | 0 | Western | 01 | |
| | | | | | Gulf | 02 | |
| | | | | | Central | 03 | |
| | | | | | NCD | 04 | |

2. FOODPRGS

This folder contains files that calculate the quantities and values of household production, sales, purchases, gifts out and in, and stock changes for the 36 food items distinguished by the survey, plus firewood and tobacco (which can be self-produced by households). There is a *.cn file and a matching *.out file for each food (e.g., *Aibika.cn* and *Aibika.out*).

The *.cn files contain hard coding of the median of the production unit values for each food, at the level of each Census Unit. The reasons for using these CU median unit values are discussed in paragraphs 44, 48, and 49 of “Measuring the Standard of Living”. Analysts unhappy with this use of the CU medians would need to modify the *.cn files as they saw fit.

The amount of information collected on each food depended on the characteristics of that food. For example, it was not sensible to ask about household production of tinned fish. Another distinction is whether beginning and ending stocks of the food were measured. To save interviewer and respondent fatigue, only the most important foods had stock changes measured.

Therefore the files for a particular food fall into one of five groups, depending on what information was collected. These groups are:

1. Foods with quantity and value data on production, purchases, gifts, and stocks
kaukau, cassava, taro, yam, banana, sago, coconut, rice, sheepmeat, pork, chicken, other meat, fish
2. Foods with quantity and value data on production, purchases, gifts but not stocks
potato, sugarcane, other fruit, peanuts, aibika, other greens, eggs, betelnut,
3. Foods with value data on production, purchases, and gifts but not stocks and not quantities
firewood, tobacco
4. Foods with quantity and value data on purchases, gifts and stocks but not production
flour, sugar, tinned fish, tinned meat, milk
5. Foods with quantity and value data on purchases and gifts but not stocks and not production
bread, biscuits, butter, other dairy, snacks, spices, tea/coffee, beer, other alcohol, soft drink

All of the *.out files within each of the five groups have the same structure. However the structure differs between groups because some information is not available for the foods in certain groups. The structure of each of the *.out files can always be discovered from the “Write” statement in the matching *.cn file, in conjunction with field length information from the data dictionaries.

3. OTHRPRGS

There are 20 *.out and matching *.cn files in this folder. The bulk of the survey analysis is carried out in these files, excepting for the analysis of food production and consumption. Table 2 summarises the contents of each file.

Table 2: Contents of *othrprgs* Folder

| File | Description |
|-----------------|--|
| <i>agricult</i> | Agricultural input spending by crop and by type of input; livestock numbers; ownership of agricultural assets; receipt of extension advice |
| <i>annpurch</i> | Consumption of infrequently purchased items (described in para. 57-59) |
| <i>child_ap</i> | Anthropometric measurements on children from the first visit to the household |
| <i>chlv2_ap</i> | Repeat of anthropometric measurements on the second (consumption recall) visit to the household |
| <i>days</i> | Daily attendance; days between interviews; summary of demographic structure (all inputs into the extrapolation factor to 12 months – see para 33-36) |
| <i>diet</i> | Foods in the diet (24 hour recall) |
| <i>dropout</i> | Details on students not completing the school year in 1995 |
| <i>durables</i> | Ownership details of 16 types of household durable and calculation of straight line annual depreciation |
| <i>educat</i> | Average educational attainment by gender for adults |
| <i>enrolmnt</i> | Male and female school enrolment rates |
| <i>facility</i> | Household facilities (toilet, water, cooking, electricity) |
| <i>frqpurch</i> | Consumption details for 19 non-food frequently purchased items |
| <i>health</i> | Utilisation of health facilities in the previous 30 days; delivery details for children aged 0-5 years |
| <i>hh_head</i> | Demographic, educational and employment details on household head |
| <i>housing</i> | Construction details for dwelling; rents (where paid) |
| <i>income</i> | Income sources for each adult in the household |
| <i>person</i> | Demographic and educational details on each resident of the household |
| <i>qltylife</i> | Quality of life perceptions; trend perception; articulation of priorities |
| <i>sch_cost</i> | Education expenses in 1995 |
| <i>water</i> | Time spent fetching water, by person |

In addition, the folder contains two constructed files that are not directly linked to the raw data. These files are:

1. *rent.wk3* gives the value of either annual depreciation or actual rent paid per year for each dwelling, constructed according to the calculations reported in para. 62-70.
2. *durables.wk3* gives the total value of annual depreciation of all durables (excluding the dwelling) based on details provided in para. 60-61 and Table 9 of “Measuring the Standard of Living”.

4. WEIGHTS

PNGHS96 was not a self-weighting survey. The sampling frame (the 1990 Census) was sufficiently old to mean that the selection of Primary Sampling Units (CUs) according to the probability proportional to estimated size rule would not give equal probabilities for the selection of each household if that selection had been based on 1996 household counts. Therefore, survey teams listed all households in the selected CU and this, along with other information (described in para. 24-29 of “Measuring the Standard of Living”) was used to form sampling weights, so that statistics estimated from the survey are representative of 1996.

Another reason for the use of sampling weights is that non-respondent households were replaced only at the first interview stage. If a household did not respond to the consumption recall interview (e.g., they had left the village for an extended period) a replacement household was not used. This differs from the practice used in some surveys. The reason for not seeking replacement households for non-respondents at the consumption recall stage was that the rate of progress of survey teams would be impeded if they had to stay behind in a village to re-do the survey (with a 14 day bounded recall period) for one replacement household. The other alternative would be to just use an unbounded recall and just ask the replacement household about their activities in the last two weeks but this would then defeat the efforts made to accurately measure consumption (particularly the recording of own-food production). Non-replacement means that there are more first visit interviews than second visit interviews, and hence there are also differing weights.

A final reason for the use of weights is to deal with the households in the longitudinal sub-sample. For most purposes, one of the two observations on these households should be given zero weight when calculating summary statistics (these weights are referred to as the N1200 rule, with this name based on the approximate sample size under this rule). The alternative would be to give weights of 0.5 to each of the two observations (these weights are referred to as the N1400 sample, with this name based on the approximate sample size under this rule).

There are just two files in this folder. The file *cu_wgths.wk3* calculates the sampling weights for each Census Unit. These weights are discussed in para. 24-29 of “Measuring the Standard of Living”.

The file *surv_des.wk3* reports the stratum number, the cluster number, and the household weight for each household in the N1200 sample (i.e., putting a zero weight on one of the two observations on the households in the longitudinal sub-sample). The details on the stratification are in para. 8-11 of “Measuring the Standard of Living”. The data in this file are essential to the correct calculation of the sampling errors for any statistics of the survey data (see para. 30-31). Any user who does not understand this point probably should not be working with the data until they have familiarised themselves with the characteristics of clustered, weighted samples.

Note that if population weights are required, they can be calculated by multiplying the household weight by the number of people in the household.

5. PRICES

Price surveys were carried out as part of the community questionnaire. These were broken into two parts: surveys of the prices of store-bought goods, typically in village trade stores, and surveys of the prices of fresh produce in local markets. The store survey collected prices for up to 23 items from two stores within each Census Unit. The survey of fresh produce prices in local markets was carried out twice, coinciding with the first and second interview of the household. Up to six lots of each of 11 types of produce were priced during each of these two surveys. Therefore, potentially there are 12 observations on, for example, the sago price in each Census Unit and these can be used to form a CU average (or a median).

The National Capital District was treated as an integrated market so a single (NCD-wide) average was taken (by item) of the prices collected from stores and markets in the 40 CUs that were part of the survey in the NCD.

Data from these price surveys were not entered into *IMPS*, instead being entered and manipulated in spreadsheets. All eight of the files in this folder are named *.wk3. Table 3 describes the files.

Table 3: Contents of Prices Files

| File | Description |
|-----------------|--|
| <i>mktprice</i> | Raw data on food market prices (except in NCD) |
| <i>fp_work</i> | Analysis file to calculate CU average prices for each type of fresh produce |
| <i>no_ncdts</i> | Raw data on trade store prices (except in NCD) |
| <i>ts_work</i> | Analysis file to calculate CU average prices for each type of trade store item |
| <i>fp_by_cu</i> | Summary file with CU averages of fresh produce prices |
| <i>ts_by_cu</i> | Summary file with CU averages of trade store prices |
| <i>ncd_mkt</i> | Food market prices in the NCD |
| <i>ncd_ts</i> | Prices of store-bought items in the NCD |

6. TEXT

There are eight WordPerfect files in this folder: *form_a.wp* (the first visit questionnaire), *form_b.wp* (the consumption recall questionnaire), *form_c.wp* (part of the community questionnaire), *manual.wp* (the interviewer manual), *diary.wp* (the expenditure diary used for one-half of the sampled households in the NCD; see para. 20), *anthr_ad.wp* (the adult anthropometrics questionnaire), *list.wp* (the household listing form), and *split_cu.wp* (the form to use when partitioning large CUs; see para. 22, 24, 25).

There are also six *.wk4 files that contain sections of the community questionnaire.