Poverty and Environment Network (PEN): A brief introduction to the global data set

By Ronnie Babigumira, with Arild Angelsen and Sven Wunder

December 2013

Contents

1 Background to the project 2

2 Questionnaire 3

3 Technical guidelines 4

4 Codes 5

5 Data entry, checking, and files 6

6 Modification of the data set 8

6.1 Moving and reclassifying products 8

6.2 Missing prices and quantities 9

6.3 Capping fuelwood prices and quantities 9

7 References 11

1 Background to the project

The PEN network was launched in September 2004 by the Center for International Forestry Research (CIFOR) with the aim of collecting uniform socio-economic and environmental data at household and village levels in rural areas of developing countries. The data presented here were collected by 33 PEN partners (mainly PhD students) and comprise 8,301 households in 334 villages located in 24 countries in Asia, Africa and Latin America.[1]

Three types of quantitative surveys were conducted:

1. Village surveys (V1, V2)
2. Annual household surveys (A1, A2)
3. Quarterly household surveys (Q1, Q2, Q3, Q4)

The village surveys (V1-V2) collected data that were common to all or showed little variation among households. The first village survey, **V1**, was conducted at the beginning of the fieldwork to get background information on the villages while the second survey, **V2** was conducted the end of the fieldwork period to get information for the 12 months period covered by the surveys.

The household surveys were grouped into two categories: quarterly surveys (Q1, Q2, Q3, Q4) to collect income information, and, household surveys (A1, A2) to collect all other household information.

A critical feature of the PEN research project was to collect detailed, high-quality data on forest use. This was done through quarterly income household surveys, for two reasons: first, short recall periods increase accuracy and reliability and, second, quarterly data would allow us to document seasonal variation in (forest) income and thus, *inter alia*, help us understand to what extent forests act as seasonal "gap

[1] The number 8,301 is the total household number included in the first household survey A1, but we had dropouts (attrition) and some households missed one out on of the quarterly income surveys.

fillers". There are, however, three partners (10101, 10203, and 10301[2]) who, because of various particular circumstances, only conducted three of the four income surveys.

In addition, 598 of the households missed out on one of the quarterly surveys, e.g., due to temporal absence or sickness, or insecurity in the area. These are still included in the database, while households missing *more* than one quarter were excluded.

Two other household surveys were conducted. The first annual household survey (A1) collected basic household information (demographics, assets, forest-related information) and was done at the beginning of the survey period while the second (A2) collected information for the 12-month period covered by the surveys (e.g., on risk management) and was done at the end of the survey period. Note, however, that we did not collect any systematic data on the time allocation of households: while highly relevant for many analyses, we believed that it would be too time-consuming a component to add to our standard survey questions.[3]

The project is further described and discussed in two edited volumes by Angelsen *et al.* (2011) (describes particular the methods used) and Wunder *et al.* (2014) (includes six articles based on the PEN project).

2 Questionnaire

Questionnaire design and development among the core of PEN researchers began in 2004. Following extensive pretesting and revisions, the first set of partners (10101 and 30402) began data collection using V3 of the questionnaire. Partly based on

---

[2] The partners adjusted their recall periods to account for this which makes their surveys comparable with the rest, i.e. the sum the income collected in the three surveys is the household's estimated annual income. Users should take this into consideration when computing annual income.

[3] We did include a few questions related to time use, e.g., "who in the household collected the forest product?"

their feedback and partly from requests from new partners, some changes and additions were made resulting in V4, which was to become the final version of the questionnaire. As such, we lack some global-comparative data for the two partners who had already used V3 of the questionnaire. For these partners we entered -8 (not applicable) in the relevant tables.

All survey tools are available (in multiple languages) on http://www.cifor.org/pen/. Because our goal was to have the different datasets aggregated into a common database, partners were asked to keep the exact meaning of the questions. However, in cases where they also needed to capture site-specific information or had special lines of own research, they had the flexibility to add supplementary questions, corresponding to their particular research interests. These added questions are not part of the database presented here.

3 Technical guidelines

Detailed and extensive technical guidelines were developed to ensure a uniform and common understanding among all partners. The technical guidelines (in the documents folder) describe the project design and the research format, and outline definitions used in the data collection. These guidelines should be consulted for details on how PEN dealt with contentious issues such as what constitutes a household, a forest, and forest income, among others. In addition, the methods book mentioned above (Angelsen *et al.,* 2011) is an in-depth going supplement.

While the technical guidelines address site selection and sampling issues, we did not *impose* a site or village sampling frame. We did, however, *encourage* partners to select representative samples and achieve a desirable degree of variation along pre-defined gradients in their village sampling choices (see Cavendish, 2003). Within the selected villages, partners used random sampling to select participant households e.g., households drawn randomly from village census lists.

4 Codes

The PEN database exclusively uses numeric data for at least three reasons: numbers are less prone to coding errors than text, it is easier to restrict the range of values with numeric data, and, given the global coverage of partners, differences in character encoding formats would have created unnecessary problems when creating the global database. Three types of codes can be distinguished:

Identifier codes: The PEN database has three observation unit identifiers; household id (*ghousecode*) a 7-digit code, village (*gvillcode*) a 6-digit code , and site (*siteid*), a 6-digit code. These identifier codes uniquely identify the observation units in the database. All data tables have both *gvillcode* and *siteid* while *ghousecode* is found only in the household surveys.

Missing data codes: We use -8 and -9 to denote missing data. These negative codes were purposefully selected, as they were less likely to overlap with or be confused with real data values. -8 is used whenever the question *does not apply*. For example, in A1 section B, question 4 only applies to respondents who answered question 3 in the affirmative -- "yes" (1) -- meaning that -8 was entered for all respondents who answered "no" (0) to question 3. The other code for missing data, -9, was used to denote "real" missings, that is, cases where the respondent did not know the answer to a question (e.g. prices for non-traded commodities). In general, assigning -9 was the last choice done after all efforts to get an answer had failed. For example, in the case of prices, enumerators would assist respondents by use of the different valuation methods discussed in the technical guidelines, however, if in spite of this, the respondent was unable to value the product/service then -9 (*respondents don't know*) was recorded.

Categorical data codes: All categorical data were pre-coded before entry, these codes can be found in the codebook. The largest component is code-product, which has more than 1 000 products. These are very disaggregated and, while we have tried to clean them out, many still often overlap in part because some of the codes reflect local product names rather than distinct products.

Our view regarding codes was that it was easier and better to aggregate codes post data collection than at the time of data collection. The later would have resulted in non-reversible data loss.

For purposes of our global analysis (see articles of special issue of *World Development*), we have done some of this aggregation and added, to each of the relevant tables, an aggregate product code that classifies the original disaggregated products into one of the 20 aggregate product codes below.

```
11 Food: Plant Products
12 Food: Animal Products
13 Food: Mushroom
21 Fuel: Firewood
22 Fuel: Charcoal
31 Medicine, Resins & Dyes: Plant Products
32 Medicine, Resins & Dyes: Animal Products
33 Medicine, Resins & Dyes: Mineral & Other Products
41 Structural & Fibre: Sawn Wood
42 Structural & Fibre: Poles And Construction Materials
43 Structural & Fibre: Other Wooden Products
44 Structural & Fibre: Pulp And Paper
45 Structural & Fibre: Non-Wood Products
51 Fodder: Browse & Graze
52 Fodder: Fed Products
61 Non-Animal Manure
62 Animal Manure
91 Other: Mineral & Metals
92 Aquarium - Ornamental
99 Unclassified
```

As an example, the table *qtr_b_fup*, which contains data on unprocessed forest products, has a variable *fup_pdt_agg*, which is the aggregate product code for the variable *fup_pdt*, the variable with disaggregated products for that section.

5 Data entry, checking, and files

Data quality and related issues are addressed in several chapters in Angelsen *et al.* (2011). In this section we present a summary of what we did with to respect to data entry and checking as well as an introduction to the data files structure.

Data were entered using Microsoft Access databases[4] in part because this would allow us to restrict data values for all pre-coded categorical variables, simplify data entry by entering each section as its own table, and, ensure uniformity and compatibility across the different study sites.

All datasets were subjected to four main checks, namely:

Logical checks: For example, children not older than their biological parents, reasonable travel time (walking minutes).

Mathematical checks: For example, the sum of quantity consumed and sold should equal to the total quantity produced or collected.[5]

Conditional checks: For example, if a household head's marital status is married, then we should have demographic data on at least one spouse.

Potential outliers: Extreme values were identified and shared with partners so as to verify whether these were due to, for example, misunderstandings or data entry errors, or whether these very just due to natural variation.

The checked and aggregated data files are made available as plain-text comma-separated values (CSV) files a platform- and program-independent format. The data are found in the data folder and are further filed under sub-folders that corresponds to their respective modules. Each section in the various questionnaires has its own data file. The data shared in this folder has be de-identified, i.e., individual/household/village names and GPS coordinates have been removed.

The file naming system follows the following convention: **module_section_content**. Using the example of direct forest income in the quarterly survey, the data are in the table **qtr_b_fup** (where fup is "forest unprocessed")

---

[4] Designed by Ronnie Babigumira and Betty Abang, with input from Arild Angelsen.

[5] We did not include storage as a third option, but asked respondents (in cases when they had stored the product) to give the likely end use of the product (consumed or sold).

Variable names and definitions are in the "Variable Names and Labels" folder in the documentation folder.

6 Modification of the data set

After the initial checks of the raw data submitted, we identified a series of errors and inconsistencies, as could be expected in a large, multi-country data set. We thus undertook a number of modifications to make the data set ready for analyses. These included a reclassification of some products, filling in missing prices and quantities, and capping and ceiling the prices and quantities of the most important but often ill-measured forest product, namely fuelwood (see below).

6.1 Moving and reclassifying products

Moving products across sections: Some products that PEN partners originally (mis-) classified as forest products were moved to the environmental and crop sections. For example, we moved some environmental and forest products between forest, non-forest environmental and agricultural income to comply with the used definition of these incomes, i.e., the main criteria for where they are classified is the land they are taken from (see guidelines). Further, we classified (wild) fish income into forest and non-forest environmental income, based on the surrounding land category of the river or lake.

Land: As can be seen in the codebook (code-land), forestland had three classes, namely: natural (10), managed (20), and plantations (30). These could be further classified depending on whether they were closed or open (see code-forest). However, because most partners only entered the disaggregated forest-land classes (10, 20, 30), we re-coded the few disaggregated entries into their corresponding aggregate classes.

6.2 Missing prices and quantities

We replaced missing product prices with average prices of location, quarter, product, and unit combinations, starting at the most disaggregated level (1) and working our way up until we found a replacement value. The four levels were:

1. village & quarter (village-quarter-product-unit)
2. village & year (village-product-unit)
3. site & quarter (site-quarter-product-unit)
4. site & year (site-product-unit)

To further explain this, consider that in the first quarter, household 1011 in village 101 harvested Y units (U) of product 323, however, the unit price was not recorded. To assign the missing price, we started by calculating the average village price of product 323 in Unit U for the first quarter. If there were no other households in the village that harvested the product by that unit in the first quarter, we would try to calculate the annual average village price. If that too was missing, we would move on the site level.

Having recovered missing prices, we used the total value and computed prices to calculate the missing quantities (very few cases). If prices/quantities have been imputed, there is an indicator variable (e.g., *fup_unpx_imp* in the direct forest products table *qtr_b_fup*) to show which level the imputation was done at.

6.3 Capping fuelwood prices and quantities

Fuelwood emerged as the most important direct forest product across most sites, with the only exceptions being sites where a majority of households were involved in harvesting high-value forest products, e.g., Brazil nuts. Given some known ranges for fuelwood values and household income shares in the literature, there was a possibility that prices and/or quantity units in our sample were overestimated, or in a minority of cases, underestimated.

A potential source of error/exaggerated importance of fuelwood was conversion factors. For each product collected, enumerators recorded the quantity collected, unit and unit price. In addition, the conversion factors from local units to kilograms were recorded in the narratives submitted with each dataset. Using these conversion factors, we calculated annual fuelwood consumption per adult equivalent (ae) with the objective of assessing whether or not the units reported were reasonable. This required that, as in the case of prices, we come up with a reasonable reference against which to check potential errors. Based on literature estimates, 5 kg per capita is the maximum quantity of fuelwood an average person would use for daily fuel needs, or just over 1,800 kg per year. On this basis, we converted all fuelwood quantities into kilograms per adult equivalent and, using a 10% margin 2,000 kg/ae as the cut off, flagged the suspect cases. Our discussions with the partners revealed that this was often a problem of imprecise conversion factors. However, there were also some cases where consumption levels were just unreasonable (for example, a widow living alone reported to have consumed the equivalent of 8,000 kg/year). For these cases, as we did with prices, we introduced a ceiling of 2,000 kg annually, which we used to calculate the quantity adjustment factor we applied on the reported quantities.

Regarding prices, the technical guidelines (section 6.7, page 51) outlined a number of methods of valuing products -- typically products produced/collected for home consumption – for which farmers may have had trouble assigning price. Distant market prices were one, but they would have to be adjusted by transport costs. Willingness to pay was another and seemed like the obvious method for fuelwood, however, as we looked at the data, we suspected that some respondents reported distant market prices for fuelwood, prices we didn't think they were willing to pay and yet these were not adjusted for transport costs.

To give an accurate picture of the importance of fuelwood, it was important that these prices were checked and corrected. To identify suspect cases we converted the unit fuelwood prices into USD/kg/ae using purchasing power parity (PPP) rates from the PENN world tables (version 7.0

http://pwt.econ.upenn.edu/php_site/pwt_index.php) and the reported conversion factors. The resulting prices were compared with a "reasonable" maximum estimate of USD (PPP) 0.15/kg from the literature and cases where prices were greater than this value were flagged for clarification with the concerned partner. In a few cases we found that these were data entry errors which were then corrected by the partner, however, the majority had indeed been distant market prices. These were then top and bottom coded these cases using USD 0.15/kg as the ceiling and USD 0.01/kg as the floor. The direct forest product table (*qtr_b_fup*) includes three variables (*an_sub_g_2k*, *px_g_15cents*, and *px_l_1cent*), which indicate which observations have been capped.

7 References

Angelsen, A., Larsen, H. O., Lund, J. F., Smith-Hall, C. & Wunder, S. (eds.) 2011. *Measuring Livelihoods and Environmental Dependence: Methods for Research and Fieldwork,* Earthscan, London.

Cavendish, W. (2003). *How do forests support, insure and improve the livelihoods of the rural poor: A research note.* Bogor, Indonesia: Center for International Forestry Research (CIFOR).

Wunder, S. Angelsen, A. & Belcher, B. (eds). 2014 (forthcoming): Special issue of *World Development* on "Forests, livelihoods and conservation"