STATISTICS DENMARK

Statistisk sentralbyrå
Statistics Norway

Statistiska centralbyrån
Statistics Sweden

MZ:2015:09

A household data collection system with a continuous multipurpose survey (INCAF)

Report from a seventh short term mission to the National Statistical Institute of Mozambique, Maputo Mozambique

19 – 27 October and 2 – 11 December 2015

within the frame work of the

AGREEMENT ON CONSULTING ON INSTITUTIONAL CAPACITY BUILDING, ECONOMIC STATISTICS AND RELATED AREAS

between

INE and Scanstat

Lars Lundgren

INSTITUTO NACIONAL DE ESTATÍSTICA

Ref: Contract DARH/2008 /004
October, 2012

Lars Lundgren
larsalundgren@hotmail.com

Contents

Summary

INE has modernised its household survey system by introducing a continuous multi-purpose survey (INCAF). The current INCAF has a household budget module IOF (Inquérito ao Orcamento Familiar), with an extended sample. The third quarter was not implemented due to lack of resources. At the visit it was clear that data were not finally cleaned. It was decided to interrupt the mission to come back in the beginning of December when data are expected to be cleaned. The first part of the mission was devoted to check the data and to recommend how to further clean them.

Data are still not fully controlled and cleaned:

- Coding has to be finalised
- Outliers have to be checked and corrected
- Prices for own produced consumption has to be imputed after agreement with main stakeholders.

A random sample of 100 households showed that the basic quality of IOF is high, if the mentioned problems are solved.

The consultancy

The aim of the consultancy was to analyse the results, particularly the expenditure data. This and coming two missions are a continuation of previous missions. The main expected activities are:

1. Support the INE in the verification of the database to identify any potential problems that could hide within the data;
2. Support the INE in making tabulations plans both quarterly, semi-annual as well as yearly,
3. Support the INE with the procedures of data aggregation by quarter, semester and year, given the methodology of a panel survey;
4. Make recommendations for the coming quarters related to the problems encountered in the actual data base.
5. A second line of activities is to review and propose a solution of a possible integration between INCAF and the annual agricultural survey IAI.
6. The consultant will work closely with the Continuous Multi-propose Survey team at INE

The current mission was limited to analyse the quality of the data. The remaining part of this mission will focus on analysis and reporting of the complete IOF data.

Full ToR are found in annex 1.

Recommendations for INCAF/IOF 2014/15

Context

Before using any data for statistical analysis it is important to know the quality of data. Previous IOFs have often been blamed for quality limitations, not always based on evidence or insight. This is partly explained by limited documentation. All statistics need documentation to facilitate transparency and confidence.

Often different errors and problems are known, but the frequencies and impact on the results are unknown. Different from accounting, statistics can be viewed as a tool to manage errors and

problems to generate relevant pictures of the reality, not being exact but avoiding bias. No statics are without errors or problems, so we have to know not only the types but also how common they are and what impact they have or can have.

Methodology

People analysing data easily find errors and problems even after extensive editing and control processes, but they don't see the frequency and possible impact. Central staff at INE have flagged suspicious data as part of the editing process, without objective criteria. About one fourth of the price data for auto-consumption was flagged.

To get a more robust picture of the quality it was decided to take a random sample of questionnaires to guide further editing. When assessing quality, it is important to be more careful than at the original process. A huge sample could risk the quality of the assessment. After a while it becomes very boring.

100 households were sampled in two stages. In the first stage 50 clusters (enumeration areas) were selected systematically[1] to give a good representation of provinces, urban/rural, etc. Then the two first questionnaires in each box were selected[2].

The assessment was limited to the fourth quarter, due to easy access to the questionnaires. It is hard to tell if the quality varies over time. The staff can be more careful in the beginning, but also get more experienced over time.

The assessment was also limited to daily consumption, both expenditure and auto-consumption, due to limited time. More work could risk the quality of the assessment but more important delay the survey results. Those data did not have any built in computer controls, as they were collected by paper and they were perceived as more problematic (flagged problems).

To further limit the work, quantities were checked only for auto-consumption. It is globally known that values have much better quality than quantities. People know and regularly use money, but lack knowledge and experience about quantities as kg, litre, etc. There are also many different local units. The quantities can sometime give supportive information when editing the values, but they have no general statistical use. Any use of the quantities, e.g. to build new poverty lines, should first be specified to see if limited parts can be used before editing them further. Editing all of them now would delay the results.

For auto-consumption, however, it was necessary to assess the quantities as the values are calculated by multiplying the quantities with prices.

The quality was assessed by comparing the data in the questionnaire with the data files. This limits the detailed assessment to the data processing from data entry to the "final" output including coding, manual and computerised editing, concatenation and storing of data and converting them from CSPro to Excel and SPSS. The weighting procedures were not included.

The assessment was made in six steps:

1. Each questionnaire was scrutinised to get a visual overall assessment
2. Relevant data for the sampled households were copied from the database to two tables, one for expenditure and one for auto-consumption
3. Different errors and problems were marked with different colours
4. The impact was calculated when it was possible

[1] The first cluster was sampled within the interval 1-25 and then all other selected with an interval of 25.

[2] The data quality is not expected to be correlated with the order within the cluster.

5. The prices for auto-consumption was compared with expenditure data
6. The findings were summarised in tables.

The assessment took a full working week to perform.

An assessment was also done of the data for monthly consumption (clothing and other semi-durable goods), but only on registered data (no comparisons with the questionnaires). Data registration errors and missing registration are expected to be similar as for daily consumption.

Visual assessment

By scrutinising the handwriting, frequency and patterns (combinations) of data it is possible for a statistician with long experience in data collection for HBS (IOF) to get a feeling of the process quality in the field. Of course a sample with 100 field visits would have been much better.

Based on more than 40 years' experience from more than 30 countries, the 98 IOF questionnaires were found adequate. Two have no data at all (non-response) and none is suspicious in any meaning. Most households had consumed some own produced food (and collected wood) combined with a few expenditures. Some households had only auto-consumption and others only expenditures. Few had many expenditures or many self-produced food items. Two households were buying the monthly food in one day (Friday or Saturday).

Assessment of the data administration

All questionnaires were found including the two without data. A few ID numbers were corrected in the field or at data entry and all questionnaires linked to the files.

Assessment of the data

Two types of error and three types of problems were identified and counted. All those were found in the first 20 questionnaires, telling that even a very small sample will probably fund the most frequent errors and problems[3]. The errors were all related to data entry mistakes. In total 0.8 % of the data were not registered. This type of error leads to underestimation as there are no double data entry (at least not in those 100 questionnaires). If not having double data entry or checking all questionnaires again, these errors have to be accepted.

The other error is when something is registered but something else than in the questionnaires. Sometimes they can be corrected because they are noticed by being too small (including blanks) or too big (including 999999…, which is an old fashioned way to signalise that there is information but the value is not known). The zeros and 9999999 should be corrected or removed. Particularly the 99999999… should be changed or removed as they will have huge impact on the results (depending on the number of 9s). Due to the risk of bad impact, 999999… should not be used. Most of the registration errors cannot be found and corrected without checking all questionnaires. These errors are mainly random, but tend to overestimate as negative numbers would be recognised.

In total 1.4 % of the data have registration errors, which is low. For single data entry, less than 4 % is recognised as good quality. The errors were evenly distributed between expenditure and auto-consumption.

[3] For testing questionnaires in general and questions in particular it can be enough to make well controlled interviews with 10-20 "normal" people to find the most common problems, which often are not observed in big pilots which not have the focus on how the questions are perceived by the respondents.

The main question is what impact it has on the results. Including the errors that could be corrected reveals that the total impact would be the missing of 0.2% of the total daily consumption. Correcting data when possible may increase this figure marginally.

Household Budget Surveys are recognised as OK, if they don't underestimate consumption with more than 20 %. Of course most information is lost when the respondents don't remember what they bought. It is globally recognised that consumption can be remembered for at least one week. IOF had several visits during each week. From the 100 questionnaires it is clear that almost all should be able to remember their consumption, both expenditure and auto-consumption for at least a week. The only exceptions are the two households that bought most of their items at one time. Even visiting the household next day may cause underestimation. These households need to keep diary or save their receipts.

The auto-consumption is very similar from day-to-day making it easier to remember, but also suggesting that a shorter period (a yesterday approach) would do as well.

One common problem (12 % of the records for auto-consumption) is related to fetched firewood. The problem is caused by changed instructions. In the questionnaire the price cell is grey, meaning that no price should be noted. Later the enumerators were instructed to enter the price 5 Mt/unit. The cell was made grey (probably in earlier IOF questionnaires) because it is very difficult to measure this item and it would be better to impute a value. This issue can easily be solved by multiplying the number of units with an imputed value (mostly 5 Mt in the sampled questionnaires). Alternatively, representative values can be generated from the same item in the expenditure file. Fetched wood is frequent so imputation should be implemented carefully also considering what was done in the previous IOF.

Another common problem is missing COICOP codes (14 % of the records in auto-consumption and 3 % of the records in the expenditure file). Like the firewood it's not a big problem, it's just to go back and code. As the items were registered also as text, coding can be done without going back to the questionnaires. First it should be seen if existent codes can be used. Otherwise new codes should be created for e.g. "other fish". It must also be checked if the item should be registered at all. A very common case is own produced water that should be deleted. Auto-consumption is supposed to measure the market value of the own produced ingredients. In the common case of a traditional cake baked with bought oil and home grown beans the cake should not be priced with the market price for the cake. An imputed value for the used beans should apply.

The most common problem is the "flags" that marks suspicious data. No flagged data in the 100 questionnaires are out of sense (outliers) – they are possible. In an HBS, data should not be removed or changed if they only look suspicious. It should be some evidence that they are wrong. Using comparisons with values and quantities are not very efficient either. Luckily only 2.8 % of the expenditures and 1.3% of the quantities in auto-consumption are flagged and as said none out of sense. They are not in any biased direction, some are higher and some are lower than may be expected. The variation is probably random. A good solution may be to just accept them.

A total different story is the prices collected for auto-consumption. Price collection in general is fairly complex and the Consumer Price Index is very much regulated. There are many aspects to take into account:

- Is the quality of the own produced item the same as was priced in the market? Probably the quality and price is higher in the market.
- Where was the price collected? In a normal shop at a coffee-stop on the way to the enumeration area or at the very cheapest outlet?
- What was the price? A price-tag on a shelf or a well negotiated price?

- From what period was the price? The current price or the probably lower price at the time for harvest?
- How aware are sellers and buyers about prices, units and measures like kg and litre? Do they use those measures regularly or do they just have a bunch of beans in plastic case that you will get for your 20 Mt bill and can get some more beans after negotiation not mentioning any price?

The field staff used own weighting scales to try to convert prices of local units to price per kg. In the household they converted the estimated quantity in local units to kg. All together this is not an easy procedure especially for people not very familiar with the measures. A mistake may not be noted.

The collected prices are mostly not out of sense, whether flagged (22 % of the records) or not. But can they be used for valuing own produced food? Maybe not.

The table below is based on all data/questionnaires for the 4th quarter. Item 11141 is maize meal and counts for almost half (45 %) of the own produced items. None of the others have more than 4%. So let's investigate the prices for maize meal.

Own produced food and firewood by COICOP codes for the 4th quarter (all data)

COICOP	%
11111	4%
11121	2%
11133	0%
11141	45%
11144	1%
11146	5%
11272	2%
11279	7%
Other	29%
Firewood	5%

Following table shows the mean prices for maize meal during the 4th quarter based on all data but limited to rural areas. The own production is marginal in urban areas. The X column shows the flagged prices and Blank the non-flagged prices. The flagged prices are lower than the not flagged except for province 5 which have a very high price. As the quality for own produced maize probably is lower than the quality on the market, the flagged prices actually seems more reliable. The next column RURAL shows both the flagged and not flagged in rural areas collected by the field staff. The last column (DD RURAL) shows the mean prices taken from the expenditure data by dividing the bought value with the quantity. There are no such prices in the province 1, 5 and 11, telling that the market is limited in rural areas in those provinces, which should make relevant price collection difficult. But it also tells that prices paid by the responding households are much lower (22 Mt) than the collected prices (31 Mt). The median that is slightly lower than the mean may be more relevant.

Prices for own produced maize meal for the 4th quarter (all data), Mt per kg

	Mean price collected			DD
Province	X	Blank	RURAL	RURAL
1	36	39	37	
2	32	34	33	22
3	18	21	20	22
4	19	20	19	16
5	**44**	35	41	
6	34	35	34	**6**
7	24	28	27	23
8	24	25	25	24
9		26	26	28
10		26	26	31
11				
All rural	32	30	31	22
Median			30	20

There are provincial differences but one could argue for using the same price in all places. Any price will be fictive as no money are involved and the welfare value of a kg of rice is the same wherever it is consumed. For imputed rent the location is included, but in that case the welfare value is different if you live in an own house in Maputo or in a remote village. The harvest was during the 4th quarter and should be lower than in the other quarters and more relevant for the full year.

Other errors

A few less frequent errors have been noticed in full database. Sorting the data from the smallest to the biggest, 0 or very low values and 99999… or very high values still remain. They are not many but should be corrected or removed. The same should be done with "#NULL".

Quarterly participation

Only one household had no consumption in the first quarter, six in the second and two in the fourth quarter. One household had no consumption in both the second and the fourth quarter.

Monthly consumption

COICOP codes were missing for 1 % of the records. Updating this is fairl simple with help of the registered item text. In a few records the values were too high. It is very important to control for this, e.g. by sorting the values from the biggest to the smallest. Though they only counts for 0.6 % of the records in the sample, they count for 44 % of total monthly consumption.

There is no expenditure for communication among the sampled 100 households. In the 4th quarter data there are no communication expenditure among the daily expenditures (DD) and only 91 records for consumption in the monthly consumption file. That is only 0.16% (0.0016) of the consumption measured as monthly consumption. With mobile phones taking a bigger part of total consumption, this must be investigated.

Lesson to learn

The found problems are based on data from the fourth quarter. Early control of the data would have found these problems and could have been communicated back to the field to avoid them in the coming quarters and save time and efforts later.

Annex A
October 2014

TERMS OF REFERENCE

for 3 short-term missions on

INCAF/IOF
the Household Budget module of the Continuous Multi-purpose Survey

23 November – 20 December 2014,
23 February – 20 March 2015 and
12 October – 6 November 2015

within the

AGREEMENT ON CONSULTING IN INSTITUTIONAL CAPACITY BUILDING, ECONOMIC STATISTICS AND RELATED AREAS
between INE and Scanstat.

Consultant: Lars Lundgren through Statistics Sweden

Counterparts*:* Arão Balate and the INCAF team at INE

Background
The National Statistics Institute of Mozambique (INE) has re-designed the household budget survey Inquérito sobre o Orçamento Familiar (IOF) to function as a module of the continuous multi-purpose survey (INCAF) which covers the seasonality in household income and expenditures over a period of 12 months. The International Sampling Consultant will provide technical assistance and training to INE on the INCAF/IOF sampling and estimation procedures. This consulting assignment will take place in three separate missions planned for November 2014, February 2015 and September 2015 and the consultant will coordinate the work with other Scanstat consultants working in the same field.

Main reasons for the mission
The INCAF/IOF has been going on since July 2015 and results are obtained each quarter. There is a need for oversight, assistance and further advice on how to best go forward.

Objective
To secure that the work goes along international standards.

Activities
This is a continuation of previous missions and the Consultant is responsible for the following activities during each of the missions:

7. Support the INE in the verification of the database to identify any potential problems that could hide within the data;
8. Support the INE in making tabulations plans both quarterly, semi-annual as well as yearly
9. Support the INE with the procedures of data aggregation by quarter, semester and year, given the methodology of a panel survey;

10. Make recommendations for the coming quarters related to the problems encountered in the actual data base.
11. A second line of activities is to review and propose a solution of a possible integration between INCAF and the annual agricultural survey IAI.
12. The consultant will work closely with the Continuous Multi-propose Survey team at INE

Expected outputs

The consultant will prepare a written documentation of the above activities;

Beneficiaries of the mission

The mission will benefit INE and the whole National Statistical System of Mozambique including users of statistical information.

Tasks to be done by INE to facilitate the mission

- Elaborate ToR for the missions
- Prepare and supply the consultant with necessary documents and information, like the new visual identity of INE
- Supply good communication conditions for the consultant.

Source of Funding

Project: MPD-2008-0006 – Inquérito Sobre Orçamento Familiar – IOF
PAAO14 – 1.2.2 Inquérito Contínuo aos Agregados Familiares
PAAO15 – 1.2.2 Inquérito Contínuo aos Agregados Familiares

Timing of the mission

Three missions on four weeks each, as written above. Number of workdays for each mission: 18 days in Maputo (6 days x 3 weeks) and a further 6 days of remote work.

Place

The premises of INE in Maputo with possible allocations to the provinces.

Language

English.

Report

The consultant will prepare a short final report to be discussed with INE before ending assignment. Statistics Denmark as Lead Party will publish the final version on www.dst.dk/mozambique within 3+ weeks of the end of the mission. The structure of the report should be according to Scanstat format.

Approved by Arão Balate, INE/DCI

Day / / .....

Confirmed by Leia Gimo Macamo, INE/DARH and Contract Manager for the INE – Scanstat Contract

Day / / .....