Basic Documentation[1]

Albanian panel survey description Waves 1 and 2 (2002/2003)

December 2004

1. Panel design

The Albanian panel survey sample was selected from households interviewed on the 2002 LSMS conducted by INSTAT with support from the World Bank. The sample size for the panel took approximately half the LSMS households and has re-interviewed these households annually in each of 2003 and 2004. The LSMS data collected in 2002 therefore constitute 'Wave 1' of the panel survey and giving three waves of panel data altogether.

Two waves of data, LSMS Wave 1 (2002) and Wave 2 (2003), are currently available for analysis. The fieldwork for Wave 3 was carried out in the spring of 2004 and the data will be available in the future.

The sample selected from the LSMS for the panel was designed to provide a nationally representative sample of households and individuals within Albania (see Appendix B for full description of the sample design and selection procedure). This differs from the LSMS where the sample was designed to be representative of each strata which broadly represented the main regions in Albania so that regional level statistics could be generated (Mountain, Central, Coastal, Tirana). Appendix B contains a description of the sample design for the panel survey.

The panel also has no over-sampling as in the LSMS. This design was adopted as the smaller sample size for the panel would have made it more difficult to produce regionally representative samples and increased sampling error while over-sampling can introduce additional complications for analysis in the context of a panel. The panel data can be used for analysis broken down by strata to assess any differences between areas but should not be used to produce cross-sectional estimates at the regional level. The relatively small sample size for the panel must always be considered as cell sizes which are small have higher levels of error and can produce estimates which are less reliable.

Panel surveys have a number of elements of which data users need to be aware when carrying out their analysis. The main features of the panel design are as follows:

- All members of Wave 1 households were designated as original sample members (OSMs) including children aged under 15 years.
- New members living with an OSM become eligible for inclusion in the sample
- All sample members are followed as they move address and any new members found to be living in their household included
- Sample members moving out of Albania are considered to be out of scope for that year of the survey (note that they remain potentially eligible for interview and it is possible they may return to a sample household at a future wave)
- From Wave 2, only household members aged 15 years and over are eligible for interview. As children turn 15, they become eligible for interview (This differs from the LSMS where the individual questionnaire collected some data on children under 15 from the mother or main carer)

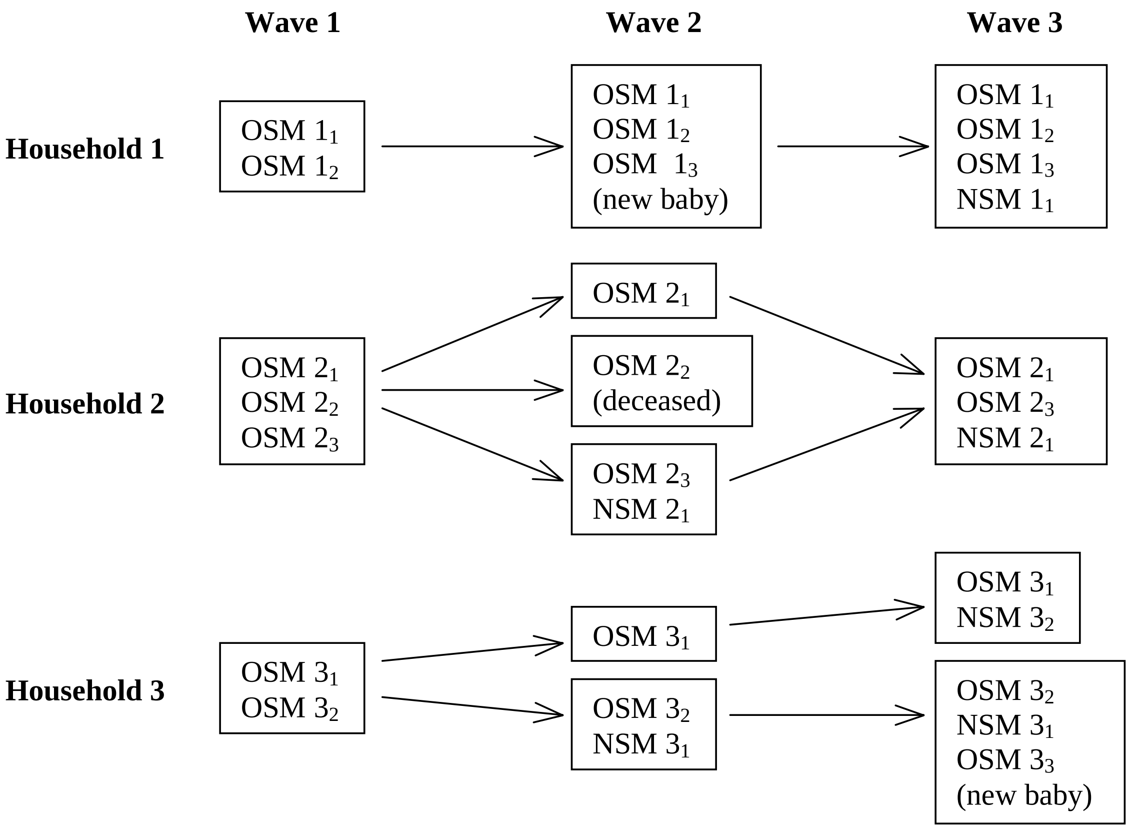
The diagram below gives a schematic outline of the panel design. The panel is essentially an individual level survey as individuals are followed over time regardless of the household they are

[1] This documentation was prepared by Institute for Social and Economic Research, University of Essex.

living in at a given interview point. This is the key element of the panel design. Households change in composition over time as members move in and out, children are born and others die. New households are formed as people marry or children leave the parental home and households can disappear if all members die or all members move in different directions. The fact that households do not remain constant over time means that it is only possible to follow individuals over time, observing them in their household context at each interview point.

It should also be noted that a ‘household’ is not equivalent to a current address. A household may move to a new address but maintain the same composition. Similarly, an individual sample member may move between several addresses during the life of the survey. In this design, there is no substitution or recruitment of new households moving into addresses vacated by sample members.

Outline of panel design



2. Panel questionnaire content

The data for Wave 1 of the panel survey are the LSMS data so contains all the modules carried for the LSMS. To minimise respondent burden and help maintain response rates in the panel survey it was necessary to reduce the length and complexity of the LSMS questionnaire. However, it was also important to maintain comparability in question wording and response categories wherever possible as only variables which are comparable over time can be used for longitudinal analysis. The Wave 2 questionnaire is therefore a reduced version of the LSMS questionnaire with some additional elements that were required for the panel e.g. collecting details of people moving into and out of the household, and some new elements that had not been included on the LSMS. A cross-wave list of variables for

Waves 1 and 2 shows which variables have been carried at both waves, which were carried at Wave 1 only and which at Wave 2 only (see 'Variable Reconciliation LSMS_PANEL_final). The most notable changes were that the LSMS detailed consumption module was not collected at Wave 2 and the agriculture module was a reduced form compared to the LSMS.

The Wave 2 individual questionnaire contains some routing depending on whether or not the person is an original sample member interviewed on the LSMS or a new person who had joined the household since Wave 1. This is because some information only needs to be collected once e.g. place of birth and other information only needs to be updated on an annual basis. For example all qualifications were collected on the LSMS so for original members we only need to know if they have gained any new qualifications in the past year but for new members we need to ask about all qualifications. Users of the data need to be aware of this routing and in some cases may need to get information from an earlier wave if it was not collected at the current wave. Users are recommended to use the data in conjunction with the questionnaires so they are aware of the routing for different sample members.

The Wave 2 questionnaire contained the following modules:

Front cover: Pre-printed household identifier and address details fed forward from Wave 1.
Collected details of new address in case of moves.
Details of calls, interviewer and supervisor, interview outcome.

Module1: Control Form (Original and split-off households)

This module contained the details of all household members from Wave 1 as the starting point for Wave 2. Details of person number within the household, name, sex, date of birth of all Wave 1 household members were pre-printed on the form.

Prior to feeding forward the sample data each sample member was assigned a unique personal identifier (PID). Each sample member carries their PID with them through the life of the survey regardless of the household they are living in. This was also pre-printed on the form. The PID is the key linking variable for cases over time so is critical for the correct identification of sample members.

Household roster including details of members who had moved out of the household since Wave 1 and new members who had joined the household, including births.

Module 2: Dwelling, utilities and durable goods

Household questionnaire asked of one person only.
Details of housing conditions, tenure, utilities and cost of utilities,
durable goods owned by household

Module 3: Education

Part A collects all details of education and qualifications in the past year for original sample members.
Part B collects details of all education and qualifications for new sample members.

Module 4: Communication

This is a new module at Wave 2 not carried at W1.
Collects details of internet and mobile phone use

Module5: Health

Reduced version of W1 health module. Collects details of health conditions, usage of health services, smoking, hospital stays, subjective health status

Module 6: Labour

Part A Labour force participation as at W1, job search

Part B Overview past 7 days as at W1
Part C Main and secondary job details as at W1
Part D Employment Grid – new section collects details of all spells in and out of employment/non-employment in past 12 months.

Module 7: Migration
Part A Full migration history past 10 years and in last year
– much extended from W1 section
Part B Details of children living away in Albania and abroad
Part C Details of extended family members living abroad – asked of Ho Hans spouse only.

Module 8: Agriculture
Part A Land - reduced version of W1 module
Part B Livestock, access to land - reduced version of W1

Module 9: Credit
Details of loans and why taken out

Module 10: Subjective
Part A Family situation as per W1 – asked of one person only
Part B Rating of local services – asked of HoH and spouse only

Module 11: Interview outcomes

Module 12: Social assistance
Part C as per W1 – receipt of social assistance payments, pensions and benefits

3. Panel files description

There are household level and individual level files for each year of the panel 2002/2003 as below. These files have been produced by combining the numbers of files produced following data entry of the questionnaires. Appendix A1 gives a description of how these files have been combined, giving details of which have been included on the household level file and which the individual level file. The files listed below are those resulting from this process of combining various data files for data collected at either the household level or the individual level. Where questions on the individual data file were answered by only one person, these modules have been attached to the household level data files.

W1_hh_all / W2_hh_all
These are the household level sample files.

Contain household questionnaire data including data for consumer durables owned by household and modules where one person only responded including subjective questions, and the agriculture modules. All repeating loops flattened so is one record per household. For example, at Wave 2, modules M2, M2c, M8a, M8b, M9, M10a, M10b, and M12 are on this household level file as only one person in the household responded to these modules on behalf of the whole household.

Note that at wave 1 all households were interviewed.
At wave 2 this file includes interviewed and non-interviewed households. For analysis, interviewed households only should be selected.

Wave 1 N= 1782 interviewed households (917 urban/863 rural/2 uncoded)

Wave 2 N= 2155 households (1780 interviewed/375 not interviewed). The majority of the non-interviewed households were due to split-off moves out of the country (N=348). A further 23 households had moved out of scope within Albania e.g. had moved into an institution of some kind or prison or were non-contacts or refusals. Only 4 households had moved and could not be traced. 83 households had moved and were traced to their new address.

W1_hh_basic / W2_hh_basic

Content as per WX_hh_all but without the agriculture modules

Wave 1 N= 1782 interviewed households (917 urban/863 rural/2 uncoded)
Wave 2 N= 2155 households (1780 interviewed/375 not interviewed).

W1_hh_farm / W2_hh_farm

Household level file with agricultural modules only (flattened)

Wave 1 N= 1782 interviewed households (917 urban/863 rural/2 uncoded)
Wave 2 N= 2155 households (1780 interviewed/375 not interviewed).

W2_roster_all

Household roster file for all individuals enumerated at Wave 2.
This file includes some sample members who appear twice i.e. in their issued household and in the household they were finally located in, even if this household was not interviewed.
The derived variable BFINLOC allows you to select cases at their final location (bfinloc eq 1).
The reasons for people leaving the household and the date they left are on their issued household record not their final location record.

W1_ind_all / W2_ind_all

Individual level file for all household members including children under 15 years.
Includes household roster module and individual level questionnaire Modules.

Wave 1 N= 7973 household members including children aged under 15.

Wave 2 N= 8110 household members including children under 15. There are 5431 members aged 15 or over with individual questionnaire data.

This file contains one record per interviewed individual in the sample and does not include individuals who were not interviewed (including refusals, out of scope, untraced or deceased).

Note that at Wave 2 the individual questionnaire did not collect data on children under 15 years. The questionnaire was for all present household members aged 15 or over.

The Wave 2 individual file contains weights for cross-sectional analysis and longitudinal analysis. The weighting variable "wt_nr" deals with non-response at Wave 2 and is used for cross-sectional analysis. Then there is the weighting variable "wt_w2" which combines the design weight and the non-response weight to produce an overall weight. This is the weight used for Wave 1 to Wave 2 longitudinal analysis and is defined for those people who participated at both waves of the survey.

The Wave 2 individual file includes variable called “outcome”. This takes the following values; 1 “interviewed” 2 “other non-response” 3 “absent from household” 4 “child under 15” 5 “moved away” 6 “deceased”.

Variable naming conventions

On the LSMS files all substantive variables started with the letter ‘m’ to denote ‘module’ followed by the module number and question number. For the panel data the redundant ‘m’ has been removed and replaced with ‘a’ or ‘b’ to denote the wave of the panel followed by the module number and question number as on the questionnaire.

As the questionnaire for Wave 2 is not the same as for Wave 1 (LSMS), the question numbers did not remain constant across both waves. This means that the variable names are also not constant across the two waves. The cross-wave variable reconciliation table gives the variable names for each wave. However we strongly suggest that you work with a questionnaire when doing the analysis to be sure of selecting and using the correct variables.

Key linking variables

There are three key linking variables that must be used to match records. These are
The HID (household identifier), PNO (person number within the household), and
PID (unique personal identifier)

For matching within a wave e.g. matching the household and individual level files you should use the HID (AHID or BHID depending on the wave) and PNO (APNO or BPNO as required).

For matching cases across waves you must use the PID. Matching cross-wave must be done at the individual level and cannot be done at the household level. This is because there is no necessary relationship between HIDs and PNOs over time as these can change as individuals move in and out of households and new households are created while others disappear. The PID is the identifier which remains constant over time and is attached to the same individual throughout the life of the panel regardless of the household they are found in or their PNO within a given household at the point of interview.

In SPSS the basic syntax for matching files is

```
match files file= NAME/file=NAME/by KEY VARS
```

Where the match is not a one to one match e.g. matching household to individual level records you need to use the TABLE subcommand

```
match files file=IND/table=HHOLD/by KEY VAR
```

Appendix A

Albanian LSMS and Wave 2 files.

Original Files – restructuring carried out

File	ahid	apno	Pid	Questions	Level
WAVE 1					
W1 filters	Yes	No	No		household
W1_agric_a_1	Yes	No	No	mca1_q0a- mca1_q13	plot
W1_agric_a_2	Yes	No	No	mca2_q0a- mca2_q11	plot
W1_agric_a_3	Yes	No	No	mca3_q0a- mca3_q12	plot
W1_agric_b	Yes	No	No	mcb_q00- mcb_q06	equipment
W1_agric_c	Yes	No	No	mcc_q00- mcc_q05	crop
W1_agric_d	Yes	No	No	mcd_q00- mcd_q05	input
W1_agric_e	Yes	No	No	mce_q0a- mce_q10	animal
W1_agric_f	Yes	No	No	mcf_q00- mcf_q03	product
W1_anthro1ch	Yes	Yes	No	mf_q3a-mf_q10b	individual
W1_anthro2	Yes	Yes	No	mf2_q3a-mf2_q7b	individual
W1_bmetadata	Yes	No	No		household
W1_dwelling	Yes	No	No	m3a_q1-m3b_q46	household
W1_health_b	Yes	No	No	m5b_q01-m5b_q10	household
W1_hholds	Yes	No	No	psu, hh	household
W1_hhroster_pids	Yes	Yes	Yes	m1-m5	Individual
W1_labor_a	Yes	Yes	No	m7aq1-ma7q14	Individual
W1_labor_b	Yes	Yes	No	m7bq1-m7bq8	Individual
W1_labor_c	Yes	Yes	No	m7cq1-m7cq27	Individual
W1_labor_d	Yes	Yes	No	m7dq1-m7dq19	Individual
W1_matern	Yes	Yes	No	m6b_q03	individual
W1_materna	Yes	Yes	No	m6a_q00-m6a_q15b	individual
W1_maternb	Yes	Yes	No	m6b_q03	individual
W1_metadata	Yes	No	No	m0q1-m0q8	household
W1_ndurables	Yes	No	No	m3c1b.1-m3c1c.24	household
W1_nonagr_a-c	Yes	No	No	mda_q2b-mdc_q17	enterprise
W1_nonagr_e	Yes	No	No	mde_q00-mde_q06	asset
W1_NonFood_1	Yes	No	No	mba_0.1-mba_3.15	household
W1_NonFood_2	Yes	No	No	mbb_0.1-mbb_3.18	household
W1_NonFood_3	Yes	No	No	mbc_0.1-mbc_3.15	household
W1npids	Yes	Yes	Yes	pid, apno	individual
W1_OtherIncome	Yes	No	No	me_0.1-me_3.11	household
W1_subjpov	Yes	No	No	m09_q0a-m09_q14	household
W1_transfer_a	Yes	No	No	m8aq2-m8aq13	Transfer
W1_transfer_b	Yes	No	No	m8b_q02-m8b_q08	livestock
W1_transfer_c	Yes	No	No	m8cq1-m8cq22	Transfer
W1_weights	Yes	No	No	weights	household
WAVE 2	bhid	bpno	pid	Questions	Level
agriculturea1_cl	Yes	No	No	m8a_q02-m8a_q12	Plot
agriculturea2	Yes	No	No	m8a_q14-m8a_q19	Plot
agricultureb1_cl	Yes	No	No	m8b_q00-m8b_q08	Livestock type
agricultureb2_cl	Yes	No	No	m8b_q09-m8b_q15	Land access
communication	Yes	Yes	No	m4q1-m4q11	Individual

Credit	Yes	No	No	m9_q02-m9_q04	Lender
dwelling_cl2	Yes	No	No	m2a_q01-m2b_q42	Household
educationnew	Yes	Yes	No	m3bq1-m3bq15	Individual
educationorig	Yes	Yes	No	m3aq1-m3aq10	Individual
Filters	Yes	No	No	various	household
health_cl2	Yes	Yes	No	m5q1-m5q35	Individual
hhroster_cl2	Yes	Yes	Yes	m1q1-m1q32	Individual
Labora	Yes	Yes	No	m6aq1-m6aq16	Individual
laborb_cl	Yes	Yes	No	m6bq1-m6bq6	Spells
laborc_cl	Yes	Yes	No	m6cq1-m6cq32	Individual
labord_cl2	Yes	Yes	No	m6dq1-m6dq6	Spells
metadata_cl	Yes	No	No	m0q1-m0q19	household
Migrationa_cl	Yes	Yes	No	m7a_q01-m7a_q33b	Individual
Migrationb_cl	Yes	Yes	No	m7b_q01-m7b_q16	Individual
Migrationc	Yes	No	No	m7c_q01-m7c_q12	household
Numdurables_cl	Yes	No	No	m2c_q1a- m2c_q1c	item
Outcome	Yes	No	No	m11_q01-m11_q05	household
socialassistance_cl	Yes	No	No	m12q1-m12q22	Source
Subjectivea_cl	Yes	No	No	m10a_q00-m10a_q12af	household
Subjectiveb	Yes	No	No	m10b_q00-m10b_q3e	household
Subjectivec	Yes	No	No	m10c_q01-m10c_q3e	household

Many of the files had to be “flattened out” so that each row was a household and each household had just one row. Variables were renamed to include the loop number so that all the information was on the row. For example, W1_transfer_c had 13 sources for each household. Rather than 13 rows of 22 variables for each household the file was flattened so that each row was one household with 13x22 variables. For the sixth source the variable m8c_q02 was renamed m8c6_q02.

Files W1_bmetadata, W1_metadata and W2_metadata were also not at the individual level, but at the household level. Each row contained the data from one household.

The aim was to end up with an individual-level file, with each row containing data on an individual. To accomplish this the loop-files had to be flattened out (see above) and these and the household-level variables had to be matched onto the individual-level files using the household identifier.

The result is a data-set with each row representing an individual. That row contains the individual’s own data (from the individual-level files) plus a copy of their household’s data (from the household-level files) and a flattened-out version of their household-level looped data. When analysing household-level data it is necessary to just select one person per household.

For some types of analysis you may need to aggregate individual level data to the household level e.g. when computing total household income from individual level data.

Appendix B

LSMS Sample Design

The LSMS design consisted of an **equal-probability sample of housing units** (HUs) within each of 16 explicit strata[2]. These were selected **in two stages**. The first was to select – within strata **-** an agreed number of enumeration units (EAs) with **probability proportional** to number of HUs in the EA (according to 2001 Census data). The second stage was to select 8 HUs systematically from each selected EA. (Substitutes were used where necessary to ensure that 8 households were successfully interviewed in each EA, but I shall ignore that for current purposes.)

Although probabilities within strata were (approximately) equal, probabilities varied greatly between the strata. Notably, the mountain region was heavily over-represented and the Central Rural region was under-represented in the sample.

Table 1 compares the population and sample distributions of households. The numbers of households for the first 13 strata come directly from the sampling spreadsheets (Census data), whereas for Tirana the spreadsheets showed only numbers of persons. For Tirana, using the total number of households in Tirana, the numbers were estimated in each of the 3 Tirana 'strata' by assuming an equal mean household size in each stratum. It can be seen that the main difference in the distributions is in the two 'mountain' strata, which contain only 9.4% of households in Albania, but 27.8% of LSMS sample households. Conversely, only 14.4% of LSMS households are in 'Central, Rural' stratum, compared with 27.9% of all Albanian households.

Table 1: Population and LSMS sample distributions over strata

Stratum	Primary	Secondary	Tertiary	Households in Population		Households in LSMS sample	
				No.	%	No.	%
1	Coastal	City	Durres	24323	3.3	80	2.2
2			Fier	14098	1.9	80	2.2
3			Vlore	19393	2.7	80	2.2
4		Other urban		41199	5.7	240	6.7
5		Rural		122747	16.9	520	14.4
6	Central	City	Shkoder	20331	2.8	80	2.2
7			Elbasan	20604	2.8	80	2.2
8			Berat	9841	1.4	80	2.2
9			Korce	13879	1.9	80	2.2
10		Other urban		46724	6.4	160	4.4
11		Rural		203013	27.9	520	14.4

[2] In fact, probabilities were not exactly equal within strata, for two reasons:

- Where more than one household was found in a HU, one was selected at random. This means that households in multi-household HUs had a smaller probability of selection than other households;

- In Tirana, households were re-enumerated (listed) in the 75 selected EAs and the sample selected from the new list. Thus, the number of households on the list may have differed from the number on the Census. Households in areas where the population had increased will have had a smaller probability of selection than others, and *vice versa*. In all other strata, households were selected from the Census list.

12	Mountain	Other urban		15044	2.1	400	11.1
13		Rural		53061	7.3	600	16.7
14	Tirana	Blue: low poverty		34483	4.7	136	3.8
15		Red: Medium poverty		50934	7.0	248	6.9
16		Green: High poverty		37221	5.1	216	6.0
Total				726895		3600	

1. Panel Survey Sample Design

The LSMS was so-designed, partly to enable separate analysis by broad strata (e.g. separate estimates for the mountain region). Regional analysis is much less important for the panel. The sample size will in any case be considerably smaller, so some regional sample sizes would inevitably be too small to permit robust estimation. The prime objective for the panel is to enable national-level estimates with the highest possible precision.

To achieve this, the sample was structured in a way that minimises the overall variation in households' selection probabilities. In other words, the sample distribution over strata matched as closely as possible the population distribution. To achieve this, the sampling fractions shown in **Table 2** were applied to the LSMS sample. The resultant panel sample distribution is also shown in **Table 2** and can be seen to be broadly similar to the population distribution in **Table 1**.

Table 2: Sampling fractions for the panel survey

Stratum	Primary	Secondary	Tertiary	Sampling fraction	Households in panel sample	
					No.	%
1	Coastal	City	Durres	5/8	50	2.8
2			Fier	4/8	40	2.2
3			Vlore	4/8	40	2.2
4		Other urban		4/8	120	6.7
5		Rural		4/8	260	14.6
6	Central	City	Shkoder	5/8	50	2.8
7			Elbasan	5/8	50	2.8
8			Berat	4/8	40	2.2
9			Korce	4/8	40	2.2
10		Other urban		6/8	120	6.7
11		Rural		7/8	455	25.5
12	Mountain	Other urban		1/8	50	2.8
13		Rural		2/8	150	8.4
14	Tirana	Blue: low poverty		5/8	85	4.8
15		Red: Medium poverty		4/8	124	7.0
16		Green: High poverty		4/8	108	6.1
Total					1782	

The rationale for this sample distribution was three-fold:

- Statistical precision for national estimates is greatly improved, compared with the LSMS design. Design effects (under the assumption of equal stratum population variances) can be expected to be around 1.02 for the panel sample, compared with 1.28 for the LSMS sample. In other words, a panel sample of 1500 interviews would give precision equivalent to an equal-probability sample of 1172 households if it followed the LSMS distribution of households over strata, but gives precision equivalent to an equal-probability sample of 1471 households with the panel design. Precision is also further improved by retaining all 450 EAs in the sample, thus reducing the design effect due to the clustering (as mean responding sample size per cluster will reduce from 8.0 to around 3.3);

- The design was simple to implement as, within each stratum, the number of households to select was the same in each EA. (Note that sampling fractions have been expressed as a fraction of 8 for this reason);

- The sample size was set so as to make it likely that the number of achieved interviews would be between 1600 and 1700. Substitute households were not be used in the case of non-response. Rather, all attempts were made to maximise the response rate. This enables the use of potentially powerful non-response weighting using the LSMS data.

Appendix C

Summary of principles of longitudinal data analysis and basic SPSS syntax

Introduction to Longitudinal Data Analysis

1. Why longitudinal research?

- Net vs gross change: gross change visible only in longitudinal data
- Causal inferences from temporal sequence (e.g. become employed > income increases)
- *Inherently* longitudinal phenomena (eg unstable employment) requires longitudinal evidence.
- Controlling for unobserved heterogeneity: change over time cannot be result of a fixed "time invariant" characteristic so models explaining first differences in effect "control" for the effects of unmeasured respondent characteristics

2. Types of longitudinal design:

Administrative data (collected for bureaucratic/management purposes) an alternative to the costly process of collecting sample surveys

Advantages of admin data:

- comprehensive coverage of clientele of admin agency
- provide authoritative statement of behaviour/circumstances related to the agency's activities (eg benefit records)

Problems:

- limitations on access and restrictive confidentiality requirements
- Limited scope of data (since limited to the agency's own purposes)

- ….but the Scandinavian model in which admin records are used as the basis or sampling frame for further sample survey activity may provide promise for the future.

"Prospective" (ie "repeated measures) vs retrospective data collection

Advantages of retrospective:

- Quick, in the sense that all the data arrives at the same time,
- and therefore cheap, requiring only a single measurement cycle
- provides "noise reduction": respondents' narrative have internal consistency

Problems with retrospective studies

- Slow, since repeated measures take time to accumulate into a longitudinal narrative
- And the internal consistency of retro studies is achieved at the cost of recall error

- Subject to “survivor bias” since some longitudinal processes mean that particular sorts of respondents (eg weak, sick, poor, those subject to an environmental stressor) are disproportionately unlikely to survive long enough to be interviewed.

3. Prospective survey designs

Cohort studies

- Birth cohorts: eg National Child Development Study – all British children born in a particular 1958 week, reinterviewed repeatedly, followed in admin (death) records
- Also youth cohorts (eg National Longitudinal Survey of Youth (USA)), ageing cohorts (eg English Longitudinal Study of Ageing, UK)

Panel studies

- Revolving: eg UK Labour Force Survey -- designed with a regular, fixed and limited cycle of interviews, and with recruitment of new samples
- Perpetual: designed with an indefinitely long horizon of regular repeated measurements e.g. Albanian Panel Survey

Household panel studies

- Particular case of a perpetual panel study
- Designed to maintain representativeness of sampled population over an extended period.

4. The Albanian Panel Survey

Is a household panel survey with a perpetual design.

Each wave has data about:

- Households,
- respondent individuals
- and children
- other non-respondent individuals

Hence two main files for each wave with substantive data:

- W1_HH_ALL
- W1_IND_ALL
- W2_HH_ALL
- W2_IND_ALL

The key index variables for matching and identifying cases on these files when using them as cross-sectional data are AHID and APNO/ BHID and BPNO.

5. Albanian Panel Following Rules

The Albanian Panel Survey is an indefinite life panel study, without replacement by drawing new samples.

‘Following rules’ are required to maintain representativeness of original population and their descendants – these specify who should be eligible to be interviewed at each wave.

- The Longitudinal Sample consists of: members of original households, and their natural descendants born since the start of the panel
- The above are eligible for interview each wave so long as they remain in scope (i.e. in Albania)
- Sample members are followed as they move.
- At each wave the interviewed sample also contains co-residents of longitudinal sample members. These are also followed if they no longer live with a sample member.

The panel sample will be reduced by:

- Attrition – refusal and non-contact
- Members becoming ineligible - Deaths and moves out of scope

6. Choosing a sample for analysis

Researcher needs to decide which patterns of response are usable for particular needs, e.g.:

- Balanced panel – requires a longitudinal sample of people present throughout the historical period covered.

- Cross-section – requires merely a single instance of interview: cross-sections may be “pooled” across years

- Pooled sample of transitions – in this case, what is analysed represents a sequence of years, and the sample describes an averaged view of the sequences over the period of the panel.

7. Households and Individuals across Time

- The concept of a longitudinal household is highly problematic – households change over time, and it is not helpful to seek to identify consistent units, and treat them as continuous households.
- on the other hand the study of household composition change is a most important role for panel studies
- We match individuals across time, and treat household data (and data about other household members) as important contextual information – e.g. household poverty level – this has been the main rationale for the household panel design, rather than the simpler individual panel design.
- Household contextual information is main reason for collecting data from new entrants
- Therefore we have a second system of identifiers at the individual level to match data about individuals across waves – the PID.

8. Database operations

4 straightforward operations to enable data management:

“matching”, “distribution”, “aggregation”, “disaggregation” (and “pooling”).

Matching

- Joining two (or more) files at the same level of observation (eg person files) where both (all) have the same “index” or “key” variables to enable matches.

- Files “joined” putting all info from both into a new file with cases defined by the index variable.

- Make an “inner match” in new file only where both old files have a value for the index variable.

- An “outer match” where one (some) of the files do not have particular cases (= particular index values) represented.

Distribution

- Joining two files at different levels of observation, in hierarchical relationship (eg household and individual) where the files share a common index.

- The distributing operation treats the higher level (eg household) file as a “table”, from which values for cases in the relevent lower level files can be read. (Eg information from a household file is “distributed” to each of its members.

- Again inner and outer matches (but outer-matched files have missing data only from the higher-level “table” file).

Aggregation

- Calculating values for a new file from old files with index variables that can be grouped hierarchically (eg calculating household incomes from an individual level file with a household index variable, then saving results as a new household-level file).

Disaggregation

- The reverse of aggregation: splitting information in cases of a higher-level file into multiple cases of a lower-level file (eg a household-level file listing all household members plus characteristics, split into an individual level file).

Pooling

- (This device is less fundamental than the other four)
- Treating *successive* measurements of the same case as if they are *independent* observations, by adding files with no index-based matching (eg adding files with occupation and wage data from successive waves to estimate mean wage rates).

SPSS file management syntax

To join two (or more) files at the same level of observation (eg person files) where both (all) have the same “index” or “key” variables to enable matches the syntax is:

```
MATCH FILES FILE=file1 /FILE = file2 /by =indexvar.
```

To join two files at different levels of observation, in hierarchical relationship (eg household and individual) where the files share a common index the syntax is:

```
MATCH FILES FILE=file1 /TABLE = file2 /by =indexvar.
```

To aggregate data i.e. calculating values for a new file from old files with index variables that can be grouped hierarchically (eg calculating household incomes from an individual level file with a household index variable, then saving results as a new household-level file) the syntax is:

```
AGGREGATE OUTFILE=filename /BREAK=indexvar
                /aggvar1 = function( vars  )
```

Pooling - Treating *successive* measurements of the same case as if they are *independent* observations, by adding files with no index-based matching use

ADD CASES or ADD FILES

Matching individuals across waves

PID is the basic cross-wave identifier

wHID and wPNO only for within-wave matches

Use the MATCH FILE ……/ FILE = instruction