Accounting for the Diversity of Rural Income Sources in Developing Countries: The Experience of the Rural Income Generating Activities Project[1]

L'analisi della diversità delle fonti di reddito rurali nei paesi in via di sviluppo: L'esperienza del progetto RIGA (Attività Produttrici di Reddito Rurale)

Katia Covarrubias[2]
Agriculture and Economic Development Division
Food and Agriculture Organization
Viale Terme di Caracalla
00153 Rome, Italy
Email: katia.covarrubias@fao.org

Ana Paula de la O Campos
FAO-ESA

Alberto Zezza
FAO-ESA

Based on joint work with:

Gero Carletto (World Bank), Benjamin Davis (UNICEF) and Paul Winters (American University, Washington, DC)

May, 2009

Abstract: The RIGA project of the Food and Agriculture Organization created a growing database of 33 household living standards surveys from which a set of income aggregates and other measures of well-being were constructed in a methodologically consistent manner. Through this elaborate task a host of definitional and methodological issues arose that confirmed the need to reflect on the different stages leading to the construction of income aggregates for developing countries. These issues relate to topics such as the defining agricultural households, identifying rural areas, defining reference periods and frequencies, among other topics. We summarize both the RIGA methodology for income aggregate construction and the obstacles faced in their construction and offer a consolidated list of methodological recommendations for the measurement of household income levels.

JEL Keywords: C80, C81, C83, D13, J00, J30, Q10, R20.

[1] Paper prepared for presentation at the Wye City Group Meeting on Rural Development and Agricultural Household Income, 11-12 June, 2009, Rome, Italy.
[2] Corresponding author.

1. Introduction

A number of efforts have been made in recent years to systematise the work on the collection of income data at the household level. Available sources have emphasised the importance of collecting and analyzing income data mainly as a measure of "the economic well-being of individuals and households" (ILO, 2003; Canberra Group, 2001), as well as a tool for looking at the distribution of income in society. The latter focus is, for instance, strongly reflected in most of the recommendations of the Canberra Group (2001). In the Wye City Group Handbook (2007), the basic motivation for looking at income at the household level also seems mostly related to measuring (farm) household well-being as well as distributional issues, including comparing low-income to other households (p. 17) and "farm households to (...) other socio-professional groups" (p. 15), as these income differentials are seen as key in driving the exit from agriculture, a major policy concern in relatively high income countries where farming still absorbs a sizeable share of the workforce.

In developing countries, consumption expenditure is usually preferred to income as a measure of household well-being for a series of both practical as well as theoretical reasons (Lipton and Ravallion, 1995; Deaton, 1997; McKay, 2000). Even though measuring household well being is still considered one of the key reasons to collect income data, other purposes are often more important, such as utilizing income data as an input into the analysis of the determinants of welfare and poverty, to check the accuracy of consumption data, to estimate household savings, and to assess the relative importance of the various activities that contribute to total household income (McKay, 2000).

Much of the focus of welfare analysis in developing countries in the last twenty years or so has focused on the assessment of poverty and the monitoring of its trends. Since consumption expenditure is the preferred metric for poverty measurement, the collection of good consumption expenditure data has received considerably more attention than the collection of income data. In some countries (Integrated) Household Budget Surveys (I-HBS), Living Standard Measurement Study (LSMS) surveys and other similar surveys have collected very little, if any, income data. Practical guidelines have been developed to assist researchers and analysts computing broadly comparable and theoretically consistent consumption aggregates and poverty measures from household surveys (Deaton and Zeidi, 2002; Ravallion, 1998), but much less information is available *for low-income countries* in terms of looking at income data. The Luxembourg Income Study, the Canberra Group and the Wye Group Handbook, the three major efforts in systematising work on household income data, all share a bias towards working with high- and middle-income countries.

On the other hand, during the 1980s and 1990s development economists started devoting increasing attention to issues related to rural non-farm income and employment and the diversification of the rural economy (FAO, 1998; Haggblade et al., 2007). Serious concerns soon began to emerge concerning the comprehensiveness, comparability and coverage of the available data. Much of the literature was based on country case studies, lacking statistical representativeness at the national level. Those based on Census data were strong on coverage, but often collected limited information on employment (e.g. only the primary occupations) and in consequence very little, if any, information on income. Studies based on nationally representative household surveys often used data coming from very different survey instruments and lacking a comparable definition of income and its components, as well as a standardised way of treating the data. Lanjouw and Feder (2001) identified data comparability and coverage issues as major shortcomings of this strand of literature.

The Rural Income Generating Activities (RIGA) project started in 2005 as a collaboration between FAO, the World Bank and American University in Washington, DC[3] with the aim of overcoming some of these issues with income data comparability, furthering our understanding of the sources of income in rural areas, and generating lessons for improving the collection of rural income data. The project has since created a database of 34 household living standards surveys from which a set of income aggregates and other measures of well-being were constructed in a methodologically consistent manner.

Through this elaborate task a host of issues arose that confirmed the need to reflect on the different stages leading to the construction of indicators of well-being, namely the construction of income aggregates. This paper summarizes the RIGA methodology for income aggregate construction and the obstacles faced in their construction to ultimately generate a consolidated list of recommendations for the measurement of household income. These are viewed in the context of both some of the possible research and analytical needs of data users, as well as from the practical point of view of the people engaged in the collection and analysis of the primary data.

The paper is organized as follows. The next section discusses the definitions and components of the RIGA income aggregates in the context of the existing literature on measuring household income. Section 3 elaborates on the survey design and methodological considerations for income aggregate construction. Section 4 reports on some key results of the RIGA work, focusing on how differences in definitions can generate very different analytical results. Section 5 offers conclusions and recommendations. The Annexes included at the end of the paper also provide more detail regarding specific methodological issues faced in the development of the RIGA database that escape the scope of the main body of this paper, but still merit mention due to their linkages of the overall subject matter.

2. Riga Income Aggregate Methodology

2.1 Preliminary considerations

In explaining the RIGA approach to the measurement of household income it is important to frame the discussion in the context of the objectives of the project and the constraints in which it operated. Concerning the former, the project stated goals put a strong emphasis on international comparability of the income measures being computed as well as on the definitions of the different components of income across countries and surveys. The project also has a strong emphasis on comparative research and analysis, particularly concerning the composition of rural household incomes. Regarding the latter, the project did not engage in the collection of new data, but worked with existing, and mostly publicly available surveys.

These considerations clearly drove the choice of the surveys with which the project, as well as the emphasis in the data work that was undertaken by the project. First, the project chose to work with multitopic surveys, such as Living Standards Measurement Study (LSMS) surveys and other, similarly structured surveys. These surveys tend have the desirable quality (from a researcher's point of view) of collecting data on a number of individual, household, and

[3] For more information on the RIGA project and access to the RIGA data see www.fao.org/es/esa/riga .

community characteristics that are essential when the purpose is not only to characterise the level and composition of income but also to investigate its correlates and determinants.

From the pool of possible surveys, the choice of particular countries was guided by the desire to ensure geographic coverage across the four principal developing regions – Asia, Africa, Eastern Europe and Latin America – as well as adequate quality and sufficient comparability in codification and nomenclatures (see Table 1 for the full list of RIGA database surveys). Surveys that did not provide adequate detail on income, or where information on income was collected via very synthetic survey instruments, were not included in the RIGA database because of concerns with their quality and with ensuring a good degree of comparability within the RIGA database. Furthermore, an effort was made to include a number of International Development Association (IDA)countries as these represent developing countries with higher levels of poverty and are therefore of particular interest to the development and poverty reduction debate.

The urban/rural definition adopted in RIGA is an immediate consequence of the choice of surveys. Countries have their own unique mechanisms for defining what constitutes rural. Thus, government definitions tend not to be comparable across countries and this may play some part in explaining cross country differences in comparisons of rural incomes. On the other hand, it may make sense to use government definitions since presumably they reflect local information about what constitutes rural and are used to administer government programs. While recognizing the potential problem with using country-specific definitions of rural, the available survey data do not allow for a straightforward alternative definition and therefore the government definition of what constitutes rurality is used. One additional caveat regarding rurality is that with the information available RIGA identifies rurality via the domicile of the household, and not the location of the job. It is probable that a number of labour activities identified as rural in RIGA are in fact located in nearby urban areas.

A host of issues that are sometimes discussed in the statistical literature on income measurement, including in the Wye City Handbook, concern the differentiation between total and disposable income, the latter being income after certain deductions take place (taxes, social security payments). Often, and namely for wage employment, such deductions are not reported or collection due to the reality of tax collection in developing countries. Nonetheless, and as explained below, income in the RIGA data is defined as 'net income', which is deducting from gross income the cost of any inputs that went into the generation of specific sources of revenues.

2.2 General Principles for Estimating Income Aggregates

Issues related to the definition and classification of income and its components, of the concept of (agricultural and farm) household, and of what constitutes rural have been explored in considerable depth in previous reports (Canberra 2001; ILO, 2003; Wye Group Handbook, 2007). The RIGA definition of income closely follows the definition given by the International Labour Organization (ILO) (Box 1).[4] An income aggregate is a measure of household welfare that is based on the different sources of income – wage and non-wage, dependent and independent – that a given household can earn over a well-defined reference

[4] Source: ILO, Resolution I "Resolution concerning household income and expenditure statistics" Available from: http://www.ilo.org/public/english/bureau/stat/download/res/hiestat.pdf

period. Set up as a monthly or annual indicator, the income aggregate is reported as an average net income figure.

As per the definition of household, the RIGA project applies the definition utilised by the corresponding survey. Generally, LSMS-type surveys define the household based on some variation of the concept that household members share a dwelling and the means of living (e.g. "eating from the same pot"). Each survey provides precise instructions as per which individuals should be considered household members (usually based on a minimum number of months they were present during the 12 months preceding the interview).

No systematic effort was made in RIGA to come up with a consistent definition of what constitutes an agricultural household, as in Chapter IX of the Wye Group Handbook. In some of the analytical papers produced by the RIGA project, agricultural households have been defined as those who had any agricultural production. Recently the RIGA database has been used in a comparative paper that looks at how different definitions can yield very different characterisation of what constitutes an agricultural household (Aksoy et al., 2009). We summarize some of the main results of that study in Section 4.

Box 1: *ILO Definition of Income*

The ILO *Resolution concerning household income and expenditure statistics* defines income as follows: "Household income consists of all receipts whether monetary or in kind (goods and services) that are received by the household or by individual members of the household at annual or more frequent intervals, but excludes windfall gains and other such irregular and typically onetime receipts. Household income receipts are available for current consumption and do not reduce the net worth of the household through a reduction of its cash, the disposal of its other financial or non-financial assets or an increase in its liabilities" (ILO 2003).

Based on the definition proposed by the ILO, we therefore consider as income receipts those that (i) recur regularly; (ii) contribute to current economic well-being; and (iii) do not arise from a reduction in net worth. These three criteria are embodied in each of the components of income; as such, irregular payments such as lottery earnings or inheritances; investments and savings and the value of durables are not included in our estimation of income.

In order to create income aggregates that are comparable across countries and over time, we apply the following criteria in the generation of our income measures:

i. All total income aggregates are estimated at the *household level*. Although income data is reported at individual, household, business and farm levels, depending on the survey module, to facilitate any analysis it is necessary to aggregate income to a common level. Since income strategies and consumption patterns are often jointly determined among household members, the household is an appropriate level of aggregation for the income aggregate.

ii. All income and expenditures are *annualized*. Income is also reported for different time periods ranging from days to weeks, months and the full year since households may earn income from different activities and to different degrees over the course of the year. In order to generate a clear picture of household-level income, it is therefore preferable to establish a broad enough time frame that captures the full extent of

activities undertaken by the household. The straightforward approach to annualization involves multiplying the amount of income received (or expenditure incurred) by the number of times it was received (or spent) such that the total revenues and costs over the course of one year are captured, accounting for the frequency in which they were earned or spent. Often complete information on frequencies is not available in which case some assumptions are drawn to enable the annualization of income and expenditures. Specifically, *when data on frequencies is not available*, the RIGA project assumes 313 working days per year (6 days per week; applied to daily earnings or costs); 52 weeks per year (to annualize weekly data) and 12 months per year (when values are on a monthly basis).

iii. All income components are *net of costs* in order to obtain an estimate of income which is readily available for household consumption, investment and/or savings. Arriving at a net income aggregate takes into consideration expenditures made by the household that are essential to its income activities but only if they are incurred on a regular basis. Two exceptions to this point are for rental income and transfer income, both of which are kept at a gross level. These exceptions are elaborated upon in the following section.

iv. *Purchases and sales of durable goods, investments* and *windfall gains are excluded* from household income and expenditure calculations since these are not transactions undertaken regularly by rural households and can result in the significant over- or under-stating of permanent income.

v. All aggregates are reported in *local currency units* of the year in which the survey took place. Income shares are estimated in order to make cross-country and over-time comparisons. If the comparison of income levels is necessary, PPP US dollars are applied, using exchange rates obtained from the most recent version of the World Bank World Development Indicators.

vi. Although the RIGA project focuses on rural households, income aggregates and household-level variables are constructed for all *rural and urban* households in the sample of each survey. As discussed earlier, defining rural and urban areas in a comparable manner is a difficult task since countries have their own methodologies by which they differentiate these areas.[5] Despite the limitations already described on this matter, the available survey data do not allow for a straightforward alternative definition of rural and urban as that would require data on population density from census data, geo-referencing of households, and information on the location of the employment, each of which is generally unavailable in most household surveys; therefore the survey-specific definitions are used.

2.3 Components of Aggregate Income

Although the construction of the income aggregates takes into consideration all sources of income reported by the household in the survey, the aggregation of the different sources is necessary and practical for analytical purposes. At the least disaggregated level, we can define two categories of income: wage and non-wage. Wage income includes all activities undertaken by persons in which the income received is in the form of a salary paid out by an

[5] See de Ferranti et al (2005) for examples of and discussion on the variability of definitions of rural areas.

employer; in other words, wage income includes earnings from dependent activities. Non-wage income is a broader category referring to (1) independent income, which includes crop and livestock production and self-employment (enterprise) earnings, and (2) non-labour income, containing transfer and other miscellaneous income sources. We therefore disaggregate our income measures into the following principal categories that follow standard practices and best reflect the analytical objectives of the RIGA study: (agricultural and non-agricultural) wages, self-employment, crop production, livestock production, transfers, and other income (see Box 2). In the remainder of this section, we expound further on each of these categories. The full classification is summarized in Table 2.[6]

Box 2: *Components of Total Household Income*

$$TOTY_I = Agwage_i + Nonagwage_i + Crop_i + Livestock_i + Selfemp_i + Transfer_i + Other_i$$

2.3.a. Dependent Income Sources

Wage Income. Wage income consists of all income received in the form of employee compensation either in cash or in kind. Since it is common for household members to simultaneously hold more than one job or change jobs throughout the survey reference period, all income from primary, secondary and any additional jobs held in a 12-month period is considered. Since wage employment information is obtained from the "economic activities" module of the living standards surveys within which all individuals generally report all their dependent and independent activities, it is necessary to only keep reported income from individuals that are not employers, own-account workers, or unpaid workers to ensure that only dependent labour income is captured and to avoid double-counting income that is reported in other survey modules. Income from individuals that identify themselves as employers or own-account workers is considered self-employment income and is accounted for in that category. Individuals identifying themselves as unpaid workers do not report income and are therefore excluded from the wage income estimation.

Wage employment income is first disaggregated by industry. The classification is based on the United Nations International Standards Industrial Classification of All Economic Activities (ISIC). As the classification of industries changes over time, the most appropriate revision of the ISIC classification standards is chosen based on the year the survey was undertaken.[7] As presented in Table 2, industries are grouped into ten principal categories: (1) Agriculture, Forestry and Fishing; (2) Mining; (3) Manufacturing; (4) Utilities; (5) Construction; (6) Commerce; (7) Transportation, Communications and Storage; (8) Finance and Real Estate; (9) Services; and (10) Miscellaneous. Using this industrial classification, total wage employment income is separated into two aggregate categories: agricultural wages (industry 1) and non-agricultural wages (industries 2 through 10).

The wage component is further disaggregated into skilled and unskilled labour and when insufficient information is provided by the respondent in a specific survey, some observations are forcibly classified into an unknown skill level. The distinction among these three sub-

[6] Further country-specific disaggregation and classifications are fully described in the survey-specific methodology, available from the RIGA website: www.fao.org/es/esa/riga.
[7] See: http://unstats.un.org/unsd/cr/family1.asp.

categories is based on the ILO International Standard Classification of Occupations (ISCO-88)[8] and, sometimes also on country-specific documentation. Whereas the skilled labour classification is given by ISCO-88 major groups 1 through 4, unskilled labour corresponds to groups 5 through 9, and the unknown skill level is assigned for major group "0" and the observations for which this information is missing.

2.3.b. Independent Income Sources

Labour income that is not earned in wage activities is accounted for by household enterprise income which is either "on-farm" or "non-farm" in nature. On-farm enterprises are represented by crop and livestock activities, which are agricultural production activities taking place on the household's own, rented in, borrowed, or sharecropped land. Non-farm enterprises represent the household's business(es) that are not directly connected to the household's agricultural production, if it has any. All of these independent activities are accounted for in the income aggregate through the categories of crop, livestock and self-employment income.

Crop. The estimation of crop income accounts for the sale of crop production, crop by-product production, sharecropping, the consumption of household crop production, net of all expenditures incurred in realizing these activities, such as agricultural inputs (seeds, pesticides and fertilizers) and the hire of farm labour.

For the valuation of own crop consumption, two different estimates are generated depending on the availability and quality of the data in each country. In the first approach (the "*crop1"* component in RIGA total income), own crop consumption is calculated based on the quantities consumed of own-produced crops as reported in the agricultural module of the household questionnaire. In cases where the quantities of own consumption are not specifically asked for in the questionnaire, this magnitude is estimated for each crop as a residual by subtracting the total amount sold, bartered, lost or used as an input (such as for seed or fodder) from the total amount harvested. The second approach (the *crop2* component) relies on the food expenditure section of the questionnaire to estimate the quantities of food consumed from own production.

In both approaches, to estimate the value of the reported quantities of own-consumption, the value of consumption is obtained using a set of imputed median prices (the specific procedure is described in greater detail below in Annex 3). For most surveys the second estimate was used in the total income calculation, among other reasons, to improve the comparability with consumption-base welfare measures; in a few cases, however, quantities from the production side were used as they were deemed more accurate or due to survey data limitations.[9]

Livestock. The livestock income category includes income from the sale and barter of livestock, livestock by-product production (i.e. milk, eggs, honey etc.), net of expenses related to livestock production (e.g. fodder, medicines) and livestock purchases, plus the value of household consumption of own livestock and livestock by-product production. The values of own consumption are estimated based on the food consumption/expenditure section of the questionnaire. In cases where this information is not available in that module, the consumption amount is obtained from the agricultural module. The approach for valuing own

[8] See: http://www.ilo.org/public/english/bureau/stat/isco/isco88/major.htm.
[9] These cases include Indonesia (both years), Vietnam (1992), Pakistan (2001), and Cambodia (2004).

consumption is the same as for the valuation of crop own-consumption (see above and Annex 1).

Self-employment. Income earned from all non-farm household enterprises is captured in this category. In most surveys this income is reported in an autonomous module; however, in a few cases, it is identified as "independent income" in the "economic activities" module of the survey and is thus necessary to differentiate from wage employment jobs (see "Wage Income" above). Self employment income includes all in cash and in kind earnings and non-durable, recurrent expenditures for all non-farm businesses operated by any member of the household over a 12-month period. All expenditures for equipment and machinery purchases and other investment expenses are not included in the aggregate. It must be noted that the purchase of raw materials, although clearly not an investment, are often procured in bulk; consequently, in some instances, they had to be treated differently from other expenses. Annex 3 elaborates on the issues encountered with this expenditure category and the treatment applied for some surveys.

Income from household enterprises can also be decomposed into industries according to the ISIC nomenclature. It should be noted that the first industry category for self-employment represents agricultural processing activities, which is differentiated from agricultural production activities. Further, when the information is available, the total income is weighted by the share of the enterprise owned by the household since non-farm enterprises may be owned by more than one household.

2.3.c. Non-labour Income Sources

Transfers. This category refers to both private and public transfers received by the household, both in cash or in-kind. *Private transfers* primarily refer to remittance income, but they can also include benefits obtained from private organizations and/or associations as well as forms of gifts and contributions not associated with the performance of a job or the provision of a service. *Public transfers* are divided into state-funded pensions and social benefits, which include welfare support, maternity benefits, and educational transfers. Pensions and social benefits reported in this section do not include benefits received from employers, as those are included under the wage employment component. Further, transfer income is overall estimated as a gross, rather than net, figure; this is elaborated upon in Section 3.2.b.

Other Sources. All other non-labour income components that do not fall into the previous five categories are accounted for in this last grouping. Other income is separated into (1) gross income from farm land rental, (2) gross income from the rental of non-farm real estate and/or of owned assets, and (3) other miscellaneous non-labour sources not specified in the questionnaire. Some caution should be exercised when comparing this income category across countries because not all surveys contain all three categories and due to the ambiguous nature of the possible sources comprising the third category. Nonetheless, in the vast majority of cases, it only represents an insignificant portion of total income.

2.4. Higher Levels of Aggregation

As shown in Box 2, total income is the sum of the seven categories described above, and is calculated twice, using the two estimates of crop income described in section 2.3.b, above. Although these seven income categories form the basis of the RIGA analyses, they are also aggregated into higher groupings in several cases. For the first grouping, total *agricultural*

activities are composed of crop, livestock and agricultural wage labour. Its complement, *non-agricultural* activities, is made up of non-agricultural wage labour, self employment, transfers and other income activities. *On-farm* income is the sum of crop and livestock production activities whereas *off-farm* activities include all activities performed off of the household's own land (agricultural and non-agricultural wage labour, self employment, transfers and other income activities). Finally, *non-farm* activity is comprised of non-agricultural wage labour and self employment.

2.5. Individual and Job Level Aggregates

Recently, the RIGA project started complementing household level with individual level sources of income data. In this case the focus is limited to agricultural and non-agricultural wage labour income sources, as independent and non-labour income sources are not easily attributed to any one household member in LSMS-type surveys. The focus of this analysis is on individuals of working age, defined as those between the ages of 15 and 60 for sixteen RIGA surveys. The dataset (referred to as RIGA-L) provides comparable data on the rural labour markets, such as individuals' participation in wage employment, income and number of jobs, as well as the frequency and duration of these jobs.

Labour market participants are defined as any individual in the household in the 15-60 age category that responded to labour time and earnings questions in the wage employment modules of the corresponding survey. Along with the data on labour market activities, individual-level and household-level variables are also available in these data sets. This allows for an investigation of how labour market participation and remuneration varies based on individual and household factors. The final dataset includes data on individual labour participation, time participation categories, daily wages, individual characteristics and household level characteristics (Winters et al., 2008).

It should be noted that wage income aggregates forming part of the RIGA-L component of the database are estimated at the *individual and job levels*. This subcomponent of the RIGA study focused on wage employment data and the pluriactivities of individuals within each household; therefore, these lower levels of aggregation were fundamental to the RIGA-L analysis.

Income is aggregated at the job level, providing information on the duration, frequency, participation in and income earned from each job held by each individual. As these are wage jobs, these variables are also disaggregated by industry and also re-aggregated, into agricultural and non-agricultural wages, though always at the job level.

At the individual and job levels in RIGA-L, wage employment is also disaggregated by labour time categories. All employment is categorized into one of the following four classifications: a) Full Year-Full Time (FYFT), b) Full Year-Part Time (FYPT), c) Part Year-Full Time (PYFT), and d) Part Year-Part Time (PYPT). These groups are intended to capture the labour time characteristics of individual employment and reflect the predominant types of jobs that exist. It can be assumed that the FYFT category represents full-time employment while the FYPT category represents part-time jobs. In addition, the PYFT category represents seasonal jobs and the PYPT category represents causal employment (see Table 3 for a synthesis of the methodology).

One major limitation of this individual level dataset from an analytical point of view is that of being limited to wage employment. To fully analyze important policy issues, such as the relation between demographic and human capital characteristics of individual workers and the transformation of the rural economy as well as to different pathways out of poverty, requires filling individual-level information gaps in future surveys. This is in part already happening where time-use modules are being integrated into living standard surveys. This is one of the avenues along which income data collection may progress in the near future.

3. Considerations for Income measurement and Aggregate Construction

Having defined the components of and the basic RIGA methodology for the income aggregate and the wage-labour market activities, we now elaborate on some key methodological points and on the main obstacles faced when creating income aggregates with household income and expenditure data. Following a review of the RIGA database surveys, we identify the principal issues that were characterized by similar, recurring problems across countries and years that must be considered in the construction of an income aggregate. We grouped these issues into those related to questionnaire design and survey implementation (which impact the quality of the raw data obtained) and those related to income aggregate construction and other data-use points, providing a list of survey-module-specific considerations based on the RIGA project experience in working with those surveys.[10] Each of the issues discussed represents a challenge in the creation of an accurate income aggregate. While these obstacles can be overcome with a proper understanding of the surveys and the data, it is necessary to take these issues into account in all stages of income aggregate creation especially the survey design stage.

3.1. Questionnaire Design and Survey Implementation[11]

The way in which the household survey collects information drives the reporting of income and expenditures in the survey. It is therefore critical to structure the questionnaire and implement the survey in a way that encourages reporting that accurately represents all sources of household income. The key issues surrounding questionnaire design relate to the way in which questions are asked, the scope of information asked for, the time frame to which the questions refer and the target respondents. Below we summarize the key points related to these issues.

3.1.a. Reference Periods and Frequencies

The choice of reference period has a large impact on the accuracy, reliability and quality of the data collected and thus should be chosen very carefully. A large body of literature exists debating the appropriate choice of reference period when conducting living standards surveys, specifically those of the LSMS project (see Saunders and Grootaert 1980; Deaton and Grosh 1998; Pettersson 2005; Scott, et al. 1980; Glewwe 2005). A reference period can span from a short reporting period (daily, weekly or biweekly) to a long (seasonal or annual) time frame. Trade-offs exist between selecting a short versus long reference period: proponents of shorter reference periods argue that annual reference periods lead to inaccurate reporting due to recall

[10] Additional issues are covered in Annexes 1-3.

[11] For a much more comprehensive review of issues related to LSMS survey design and implementation the 3 volumes edited by Grosh and Glewwe (2000) are mandatory reading. This section only reviews a few specific issues that are particularly relevant for the RIGA work with income aggregates.

errors. The reference period that the respondent is asked to recall must not be too long, as this would increase the probability and magnitude of recall errors (Pettersson, 2005).

On the other hand, due to seasonal fluctuations, short-term income and expenditures may be too variable to present an accurate picture of annual income and expenditures such that the reference period should be kept at twelve months (Saunders and Grooteart 1980; Deaton and Grosh 1998). A full-year reference period can be regarded as essential for the survey as a whole– unless enough data is collected to show how a reasonably representative shorter period can be selected for future rounds (Scott, et al. 1980). One of the ways to address the problem of seasonal variability is to undertake multiple visits to the household throughout the year, but this solution is costly and may not be feasible due to difficulties in coordinating revisits with households, changes in household structure over time, as well as logistical issues such as the problem of reaching households in unfavourable periods, for example the rainy season. Ultimately, due to these numerous challenges and the financial cost of following households throughout a year, the possibility of obtaining a sufficiently large sample would be greatly limited (Saunders and Grootaert 1980)

In the end, the reference period should coincide with respondent's ability to recall reported information as accurately as possible. The key consideration in dealing with reference periods and frequencies should be ensuring that the information provided from the questions answered in the survey is sufficient to create a reliable annual estimate of the income or expenditure. The income and expenditures of a household range from frequent (food expenses; monthly wages) to infrequent purchases or income receipts (i.e. durable goods purchases; annual interest from savings). As a result, the reference period should reflect the frequency of the incurred income and expenditures. As noted by Pettersson (2005), high-frequency items such as food usually have relatively short reference periods, at most a one-month period. The recall of expenditures on low-frequency items, such as household durables, must cover a relatively longer period since a short reference period could result in large variations in the final estimates. The length of a suitable reference period must consequently differ across item groups and income activities (Pettersson 2005).

In consequence, depending on the survey module at hand, the most appropriate reference period will vary. Even within sections it may differ, such as for wage employment, in which households can report income from current jobs and those undertaken over the previous year. Income that is earned infrequently is better off reported with a broader reference period (the previous 12 months) whereas consistent, regular or frequent sources of income are better off being reported on a shorter time frame. When reference periods are shorter, households must be probed for more information about the frequency with which the income is received (or the expenditure is incurred) to generate a picture of the annual importance of the income or expense. Under a broader reference period, fewer questions need to be asked regarding the frequency of income, but enough should be made to enable cross-checks of the reported values. It is inefficient for surveys to have many questions regarding the frequency of an income source when the value earned is a total reported for the previous twelve months.

The method by which information is collected is also relevant. For example, the use of a 14-day consumption diary, as in the case of the Albania LSMS, is one approach used to obtain a clear picture of daily consumption of food and non-food items among household members. For generating estimates of consumption, this could arguably be the best approach instead of asking households regarding their consumption patterns based on a pre-defined list of items; with a food diary, the scope of the data collected is created by the household, minimizing the

possibility of not accounting for items a household may regularly consume. A similar approach could be applied for measuring income from household enterprises to obtain accurate estimates of regular income and expenses and minimize the recall error of the responses. Of course, the reference period for such an approach must be chosen carefully: it should be broad enough so that one can assume with confidence that household income and expenditures over the course of *N* days/weeks are representative of a longer time period, but not so great so that households fail to complete the booklet. [12] The time frame should also be chosen so to not include "abnormal" periods such as holidays in which consumption behaviour may deviate from the norm (though the questionnaire can also account for this kind of consumption separately).

Collecting complete and representative information on frequency and duration of work is fundamental also for wage employment data, specifically for calculating variables on full-time and part-time jobs; as well as full-year and part-year jobs. Data on frequency and duration should be available for all types of jobs (main, secondary, third, etc.), to improve the accuracy of the estimations of labour-time and ultimately income. One inherent challenge faced is the differing ways in which individual surveys ask labour time questions. For some surveys, labour time questions vary depending on whether the first, second, or third job is being referred to while in other cases all labour time queries are consistent. In addition, in some surveys the first job is designated as the primary or full-time job whereas the second job is considered a casual, other, or default employment, a problematic approach when, for example, a person has two full-time jobs over a 12 month period, or when a person has no full-time job but two or more part-time jobs. In such cases, designating one particular employment as the primary or secondary job is difficult and requires additional criteria on labour time or earnings to be applied, which can introduce further complications when labour time or income questions are inconsistent throughout employment modules. In consequence, in addition to the importance of selecting adequate reference periods for reporting income and expenditures, when dealing with with-in module reporting of similar income sources, the consistency of the questions asked is also critical for accurately estimating income, notably when reference periods may mask the true nature or importance of a given employment.

The experience of the RIGA project allows for the conclusion to be drawn, for this matter, that certain reference periods are more appropriate for estimating income from each of the income survey modules and that linked to those recommended reference periods are other relevant considerations. Table 4 highlights what are identified as "recommended" reference periods for each component of income. It should be emphasized though that these are just guidelines, and therefore should be taken as such.

For crop income to be accurately estimated, it is most practical for questions regarding production to refer to production seasons, such as the previous season, the wet season and/or dry season. Collecting information in a way that follows the actual practices of the producer is logical and encourages better recall for the survey. The number of harvests or production seasons may vary across countries so these should be specified according to the local context. Information on frequency of sales, consumption, and other categories reported should also be

[12] de Mel, et al (2007) find that when household enterprises were requested to complete a booklet of income and expenditures over an extended period of time, accuracy dropped after one month as households were less regular in their accounting.

collected with the goal of accurately annualizing each component of crop income.[13] Similarly, despite that inputs are used on a regular basis, the reference period for these expenditures should be annual since their purchase is not necessarily undertaken on a periodic basis (e.g. seed inputs are likely purchased only prior to the cropping season). Similarly, questions on livestock income and expenditures should also be framed around the previous 12 months as livestock sales and purchases are generally infrequent. Livestock products such as eggs, milk and milk products, however, could be structured on a monthly reference period since their production is more regular in nature. Conversely, by-products related to the slaughter of animals are probably infrequent and therefore should also adhere to the annual reference period.

Regarding reference periods for own consumption, since the objective is to obtain an accurate representation of the regular consumption patterns of the household that reduced recall error, a recommended reference period would be the previous 14 days complemented with information on the frequency of consumption. As emphasized above, it is of utmost importance in this case to undertake the survey in a period that best reflects regular consumption periods. That is to say, avoiding periods where consumption may substantially increase or drop such as the Tet holiday in Vietnam or during Ramadan in countries with a Muslim population.

Multiple reference periods should exist for wage employment activities in order to capture individuals' pluriactivities throughout the year. Whereas the overall reference period for any wage income reporting should be the previous year, it is preferable to have wages reported in terms of the frequency in which they are received; that is, the reference period should be specified by the respondent. Further, the questionnaire should ask questions regarding the duration of each job (in hours, days, weeks and/or months of the previous year), the wage received, the frequency the wage was received, the time worked to receive the wage, and the amount earned over the past year. In-cash or in-kind income from wage jobs obtained as benefits granted by the employer should be divided into those received monthly (e.g. transportation; food) and those received bi-annually or annually (e.g. uniforms). The frequency of their receipt should always be reported.

For self employment activities, RIGA work demonstrates that the preferred reference period would be the previous month. If a household owned an enterprise in the previous year but no longer operated it in the month preceding the survey, then the average monthly income from the enterprise should be reported. It is therefore essential for the survey to also ask about the number of months the enterprise operated in the previous year, as well as the share of household ownership in the enterprise (since businesses are often owned by more than one household). Business costs should be asked with respect to the previous month, the average month, and the frequency with which the household incurs such costs must be reported.

The reference period for transfer income must also vary depending on the type of transfer. Public transfers such as pensions are generally received regularly therefore a monthly reference period is appropriate. Social transfers, such as educational scholarships, may have a lower frequency such that an annual reference period is preferable. For private transfers, households should be asked regarding their receipt during the previous year, the average

[13] For example, the Ghana 1998 Living Standards Survey asks about crop sales income in the previous two weeks but the survey does not contain information that allows the data user to know if those two weeks are representative of crop sales over the course of the year or if crop sales are irregularly distributed over the year.

amount received and the frequency with which that amount is received so to construct a more specific and accurate annual estimation of this sort of income.

Finally, other income sources are comprised of rental income and miscellaneous non-labour sources. Rental income is generally regularly received therefore a monthly reference period, accompanied with the frequency of the income is appropriate whereas the latter category is best reported with respect to the previous year given the unspecified and potentially irregular nature of the income.

An additional important note on reference periods is the way in which they are phrased. To ask an individual about the previous week's consumption is different than to ask about the consumption of the previous seven days. The former specifically refers to the previous Sunday through Saturday (or Monday through Sunday) period whereas the latter refers to the previous seven days, irrespective of the day of the week. The same concept holds for previous year versus the previous 12 months; the previous 14 days versus last 2 weeks; etc.

3.1.b. Units and Coding

The use of appropriate units and coding is an important factor determining the accuracy of constructed consumption and income aggregates. One of the key problems encountered in the generation of the RIGA income aggregates is inconsistency in units and coding across survey modules, across years for one country in different time periods, and across surveys for different countries. These issues engendered problems in the interpretation and processing of household data with the primary objective being the creation of interpretable values that are intra-module, intra-year or internationally comparable.

Although internationally recognized units for measurement exist, it is inevitable that different countries have local units by which measurements are recorded. For example, *cuadras* are common in some Latin American countries, whereas ropes and poles may only be found in Ghana. And though different countries may refer to a unit by the same name, the equivalence to internationally recognized units can vary across or even within countries: one *cuadra*, for example, is equal to 4.17 acres in Argentina, 3.89 acres in Chile and 1.85 acres in Paraguay and anywhere from 1.58 to 4.45 acres in Venezuela. When dealing with weight or mass quantities, some countries will account for units sold in kilos, whereas other countries, when referring to the same items, may report items in sacks, carts, or boxes, depending on the unit of the consumption or transaction. Further, household surveys generally include an "other unit" category which is often undefined, thus complicating the task of placing values in common units.

It is not to say that only standard units should be used in the survey design. As noted by Saunders and Grootaert (1980), the appropriate classifications of groups must be related to the local context. The accuracy of data collection is enhanced by reporting in units with which households are most familiar; nevertheless, it is clear that in all cases, regardless of the type of measurement at hand, equivalencies to international standard units are necessary to provide with each survey. Not doing so will lead to the persistence of the measurement inaccuracies prevalent in many surveys and impede accurate and meaningful cross-country comparisons, and can complicate the comprehensibility of analyses using unit-level data. These inaccuracies may also carry over to analyses which are often the basis for policy-making and program evaluations (Chander et al. 1980; Pettersson 2005).

Table 5 demonstrates the variability in unit reporting across a selection of the RIGA surveys. For some surveys, only one unit is reported, while for others, upwards of seventy are reported. Further, as an added difficulty for the analyst, equivalence scales for converting units often are not included in the surveys where many local or non-standard units are reported. In other cases, the reported unit is identified as "other" but without any additional information to inform the analyst as to what is represented by "other". This is particularly problematic in certain cases, such as in identifying land units in the surveys for Ghana and Nigeria, in which one of the four potential units in which land area can be reported is "other", and in which no extra information is given.

Also of great importance is internal consistency in the codification of items within a survey. This point is notable for living standards surveys containing agricultural production and food expenditure modules. From the perspective of the construction of an income aggregate which requires data from multiple survey modules to be utilized, it is both practical and logical to facilitate analytical work by codifying items from these modules consistently, as well as to ensure consistent lists of production and consumption items, particularly those which are consumed by the household. Within a survey module it is beneficial for the analysis if the unit in which a specific item is reported is always consistent. For example, if a respondent reports harvesting 100 kilos of rice for a given reference period, subsequent questions on the quantity of rice sold, consumed, paid out to labour, and stored, etc, would preferably be reported in the same unit (kilos, in this case). Doing so facilitates the valuation of items for which no real value can effectively be reported (e.g. the value of crops consumed or stored).

Another issue related to units and coding is the application of existing nomenclature in the collection and analysis of wage data. Some surveys do not collect data on the industrial classification of labour activities and only collect occupational sectors. It is important for surveys to collect data on both given that industrial and occupational classifications represent different concepts and are not substitutes for descriptive information about labour activities. Further, given the existence of ISIC and ISCO coding schemes, it is practical for surveys to adhere to such nomenclature rather than incorporating local industrial or occupational classifications in the survey. When a survey does adopt the existing ISIC or ISCO codes, it is also important to follow the current revision of such coding schemes, which are updated when necessary by their respective institutions. Although these updates may challenge consistency in codification schemes over time, it is also important to keep data collection and analysis in line with current definitions.

3.1.c. Data Validation

Structuring surveys in a consistent manner can alleviate many of the problems encountered by data users in creating estimates of income, expenditures and other indicators. Resolving issues of questionnaire design must also be complemented by granting attention to the process by which a survey is carried out. In order to establish confidence in the quality of the data collected by a survey, it is necessary to cross-check the answers provided by respondents.

Data validation can be undertaken through various approaches both during and after the survey. In the questionnaire, surveys can ask additional questions to cross-check the responses to other questions for ensuring accurate reporting. For example, the quantity of crop harvest can be validated by asking follow-up questions about the crop production such as the quantity sold, given away, stored, used as inputs, used to pay the landlord or labour,

consumed and lost;[14] total harvest should equal the sum of these components. Another approach is to include, as with the agricultural module of the 1996 Nepal Living Standards Survey, an accounting table which can serve to cross-check and summarize reported income and expenditures. Finally, data entry ideally takes place concurrent to the survey fieldwork so that field validation can be undertaken, if necessary, to verify questionable responses by revisiting households to reconfirm answers.

3.2. Dealing with Specific Income Estimation Issues[15]

3.2.a. Investments

The estimation of a net income aggregate is based on sources of regular and/or recurring income and expenditure receipts for a specified reference period that contribute to current economic wellbeing (ILO 2003). Such transactions occur on some recurring basis- daily, weekly, monthly, etc., and exclude anything that is beyond the scope of the previous year. The examples of non-regular income are few and generally fall into the category of lottery earnings and inheritances. These kinds of earnings would not be accounted for since generally they are not regular or recurrent earnings and since such earnings would be a source of income beyond the range of normal income for that household.[16] On the expenditure side, the cases that should be identified and dropped are numerous, since large expenditures can be recorded in multiple sections of the survey (namely, where income earned is from an independent activity: crop, livestock and self employment) which places the risk of under-estimating or driving negative household income if the cases are overlooked or misidentified.

The misclassification of expenditures into investments or durables is always a risk if a survey is not properly structured. Chander et al. (1980) emphasizes the importance of clearly grouping questions and sections in a logical structure to avoid such misclassifications. The result of not doing so can produce reduced estimates of net income in a given section, such as in the self employment module of the 1998 Vietnam Living Standards Survey in which, the distribution and magnitude of annualized raw material expenses appear to be more like investment expenditures than regular, non-investment expenditures undertaken by the households. For more detail on this issue, Annex 3 provides a broader discussion.

3.2.b. Gross versus Net (Disposable) Income

In the estimation of a net income aggregate, the identification of revenues and expenditures is highly important. The wage employment module of many questionnaires contain inconsistencies in the reporting of wage income: some surveys ask for net income, others for gross income and in those surveys with gross income reported, taxes and other deductions are not always consistently or clearly accounted for in the reporting of expenditures. The lack of

[14] An aside related to this point is the importance for keeping the reporting of sales and own consumption separate in the agricultural production module; some surveys aggregate these in the question, placing obstacles to accurate price estimations and disaggregated agricultural income analyses. Ideally, households are asked about their total harvest, harvest sold, consumed, given away, used to pay labour, use to pay rent, stored, used as an agricultural input, and lost.

[15] For the RIGA protocol for dealing with outliers and impute missing values see the Annexes.

[16] It should be noted that if the payment is received as a lump-sum it should be unquestionably excluded but if such payment is spread over 30 years though, as could be the case for large lottery earnings, then the treatment of the income would be different. However, the occurrence of this is so rare; such income would be treated as an outlier, and be dropped or imputed.

detail on taxes and deductions is of course a reflection of the reality of the large scale of the informal sector in low income countries in which taxes are often not collected and deductions not applied. This "reality" generates the need to differentiate income estimation methodology for low-income countries from what is the standard for OECD countries and what is described in the Wye City Group Handbook.

Table 6 summarizes this tendency for a selection of the countries in the RIGA study and demonstrates that most surveys do not ask for net income in the reporting of wage income. Many surveys allow the respondents to specify whether their reported income is gross or net, a useful feature of those surveys since doing so allows analysts to structure their work accordingly. Further, this kind of reporting permits, to an extent, the consideration of the informal sector since adherence to tax laws in developing countries may not necessarily be common (Tanzi and Zee 2000). However, the flexibility in reporting of wages may complicate the estimation of wage income if appropriate follow-up questions are not asked in conjunction with the reporting of the earnings. In only one survey of those reviewed for this paper did the questionnaire follow up the question "were taxes deducted from the wage?" (2001 Bulgaria Integrated Household Survey) with a question on the value of taxes deducted. Most often in the surveys for which respondents can choose to report gross or net earnings, no question is included regarding the amount of taxes paid or if they were paid at all. Other surveys may include a question regarding the payment of taxes in wage employment; however it is not clear if these were subtracted from the reported wages. Two examples of these are the Guatemala 2000 *Encuesta de Condiciones de Vida* and the Nicaragua 2001 *Encuesta de Medición de Niveles de Vida*, which both include a question on payments to social security (such as "did you pay quota to social security for the work that you do as…?"), but do not state if the tax amount has been subtracted from the reported wages. Although this lack of information conveys the reality of tax collection in developing countries, it places an obstacle for the analyst who wishes to undertake livelihoods analyses, particularly comparative ones. If international comparisons of income strategies are to be realized, the over-arching principle is to establish a common set of standards and guidelines by which to measure and aggregate income sources This objective is undermined when some surveys report net earnings and others gross without the possibility of estimating taxes and other withholdings.

At the analytical level, the consideration of gross versus net earnings from certain activities is also a subject of debate. Whereas some income sources should clearly be estimated as net figures (agricultural income, self employment earnings and wages, whenever possible), others should be gross earnings. The two cases in which earnings are always estimated as gross figures in the RIGA data work are rental income and transfer income. Given that the purpose of discounting expenditures from revenues in the income aggregate is centred on accounting for regular current costs that contribute to the realization of the labour activities under consideration and, overall, available for current consumption, it is not appropriate to account for the cost of renting in property or sending out transfers.

In the case of transfer income, it is important to first note that expenditures can only be incurred on private transfers since the public and social transfers are generally received at no financial cost to the household. Private transfers, though, can be incoming and outgoing. The outgoing transfers are not accounted for since it is not possible in any of the surveys to differentiate between those that are permanent and those which will be repaid and therefore classified as loans. The RIGA income aggregate does not consider loans in its calculations since they represent a reduction in net worth (see Box 1). This reporting ambiguity obliges

the exclusion of outgoing transfers and the necessary estimation of gross, rather than net, transfer income. Moreover, the payment of transfers to other individuals does not contribute to the household's current economic well-being (these instead contribute to other households' well-being) in the sense that making an outgoing transfer does not enable the household to earn income from a given activity; conversely, expenditures made for agricultural and non-agricultural businesses do contribute to the household's economic well-being since they allow the household to realize its income-generating potential. It follows from this last point that rental expenditures are also not considered in the income aggregate. They do not embody an incoming-outgoing transaction and are not necessarily associated with the generation of income through labour activities.

To summarize some of the main points in this section, for dependent labour employment, regardless of net or gross wages being reported, deductions and taxes should also be reported to enable the estimation of both values. Further, the inclusion or exclusion of in cash and in kind benefits received in the reporting of the wage must be specified. With respect to the agricultural production and non-farm enterprise modules, regular expenditures and durable expenditures should be clearly distinguished in different sections of the survey module so to avoid misinterpreting the kind of expenditures considered in the income aggregate. It thus follows that survey questionnaires should avoid grouping together expenditures for several different items into one code; if expenditures are to be disaggregated by item, they should be reported for each item, not groups of items. Finally, "other" expenditures should be identified or described to determine if they are investments or not.

4. An Overview of Key RIGA Results

4.1 Diversification of Rural Household Income in RIGA Countries[17]

Much of the literature on rural non-farm activities focuses on the diversification of income sources over the rural space, or over groups of households within the rural space. To examine that, the RIGA dataset is able to describe the share of income from, and household participation in, rural income generating activities. Some of this work is summarised in this section based on 16 country datasets from the RIGA database. These results also show how household level income data can be a useful starting point to look beyond the microeconomic level, to better describe and understand the structure of the rural economy.

Overall, as would be expected, the share of rural non-agricultural income increases, with increasing levels of GDP per capita (Figure 1) and as such, its complement, agricultural income, declines with higher GDP levels. Off-farm sources of income account for 50 percent of total income in almost two-thirds of the countries of the dataset (Table 4). This is true of all of the countries from Eastern Europe and Latin America and for all but Vietnam among the Asian countries in the sample. On-farm sources of income tend to be more important for the African countries, where the shares range from 59 to 78 percent of total income. Joining together income from agricultural wage labour with crop and livestock production, around half (9 of 16) of the countries in this dataset had a majority of income from agricultural sources.

These results speak of a highly diversified rural economy and suggest that rural households employ a wide range of activities. The question remains, however, over whether households

[17] This section is based on Davis et al. (2009).

tend to specialize in activities with diversity in activities across households or, alternatively, whether households themselves tend to diversify their activities thereby obtaining income from a range of activities. To answer this question, we examine the degree of specialization and diversification by defining a household as specialized if it receives more than 75 percent of its income from a single source and diversified if no single source is greater than that amount.[18] This provides a sense of the degree of specialization and of the activities through which households specialize, though we are limited from delving into greater details about this specialization by the way in which household survey data are typically collected. The apparent diversification shown in the data may be due to aggregation across seasons (with households specializing seasonally) or across individuals, with specific household members specializing in different activities.

In any case, household diversification, not specialization, is the norm, as can be seen in the data presented in Table 8. Not only are most rural economies highly diversified, but rural households are as well. With the exception of the African countries where it is still most common to specialize in on-farm activities, the largest share of rural households is diversified. When households do specialize, in a majority of cases this specialization is in on-farm activities, although the shares decline with higher per capita GDP. At higher GDP levels specialization in non-agricultural wages becomes more important, whereas no distinct association between GDP levels and specialization in agricultural wage or self-employment is suggested by the data.

This is illustrated in Figure 2 with the average country shares of specialization and diversification identified by the country data points in the figure and the trend lines for the first, third and fifth expenditure quintile. The share of diversified rural households increases only at the higher levels of per capita GDP for low and high income households alike. Clearer patterns linked to the level of development emerge for specialization in farming (declining with GDP), and in non-agricultural wage labour (increasing with GDP for all but the poorest households in the first quintile). In the former case two countries appear to be significantly distant from the pattern set by the others: Nigeria (high share of farm specializers for its GDP level) and Bangladesh (low share). In the latter case, the only significant 'outlier' is Pakistan with a relatively high share of non-agricultural wage specializers for its GDP level.

4.2 Sensitivity of the Characterisation of Agricultural Households to the Criteria for Defining Such Households[19]

Given the diversity of income generating activities in which rural households tend to be involved, defining what constitutes an agricultural household becomes a difficult task. Policy interventions and price changes reallocate income from agricultural households to others and vice versa. Precisely defining the agricultural household is important, particularly when measuring the poverty impact of changes in policy and in the overall economic environment.

[18] Other definitions of diversification and specialization are possible. We also looked at using 100 percent and 50 percent of income from a single source as alternative thresholds to define specialization, in order to ascertain the robustness of our results. The extent of diversification is clearly affected by the choice of the threshold, which drops to around 10 percent or less in all cases when using the 50 percent definition and climbs to around 90 percent when using the 100 percent definition. The broad patterns by country and by level of welfare discussed in this section, however, do not change with the choice of the threshold. Similarly, alternative groupings of income categories are also possible, such as joining together agricultural and non agricultural wage labour, or non agricultural wage labour and non agricultural self employment, which would increase the share of household specializing.

[19] This section is based on Aksoy et al. (2009).

One definition widely used in the literature equates rural with agriculture, thus treating all rural households as agricultural or farming households. Although this may be an appropriate definition in some country contexts, it is subject to some limitations. Firstly, it discounts the fact that farming is often widely practiced in urban and peri-urban areas. Another important shortcoming is the inconsistent definition of rural and urban across countries, which this paper elaborates upon in Section 2.

A second definition uses sources of income to classify households as agricultural or non-agricultural. By this measure, households that earn any income from crop and livestock production are defined as farming households. This is similar to the broad definition that is used in the OECD country studies and also used in the Luxemburg Income Study (LIS) and can thus be used for comparative purposes. Although it is more specific than equating rural with agricultural, this definition also has the limitation of identifying as agriculturally-oriented households with only small shares of agricultural income, namely those that maintain some subsistence production, but are actually more fully employed in other sectors. In the sample of surveys used in Table 9 (which combines RIGA and non-RIGA data), almost 93 percent of households in Vietnam, and 97 percent in Bolivia have some positive agricultural income. Similarly 41 percent of all urban households have some income from agricultural activities. The results indicate that this appears to be too loose a definition to satisfactorily identify farm households in developing countries.

A variation of this definition uses a cut-off point of the level of income originating from agricultural activities. Households that earn more than a given level of income from agriculture are classified as agricultural households. Clearly, there are no "a priori" correct cut-off points. In Table 9, two thresholds are used: 10 and 30 percent. In other studies, thresholds of up to 50 percent are also used[20] and, as Table 9 demonstrates, results change according to the selection of the cut-off points.

Finally, a third approach consists in defining farming or agricultural households based solely on the occupation of the household head. As with the other definitions, caveats exist, the main one in this case is assuming that the household head's main activity is representative of that of all household members. This definition also ignores the tendency, as described in Section 4.1 for households to diversify their income sources.

Table 9 presents the share of households under the various definitions. The share of the rural population varies between 88 percent in Malawi, to 36 percent for Peru. Latin American countries have much lower share of rural populations. Countries in the Asia sample of surveys have the greatest share of rural households and the Africa surveys fall in between. With respect to the income cut-off point definitions, the table also shows that using a threshold of 10 percent of income from agriculture eliminates most of the marginal agricultural producers, and reduces the share of households classified as farmers from 72 percent of all households to 57 percent. Increasing the cut-off point to 30 percent reduces this share to only 46 percent suggesting that most of the marginal agricultural producers have agricultural incomes at levels below 10 percent. This change is more dramatic in urban areas where about 41 percent of urban households have some positive income and when the cut-off point is increased to 10 percent, the share drops to 21 percent.

[20] These cut-off points for identifying agricultural households should not be confused with the thresholds for specialization in agriculture, described in the previous section.

With the 10 percent definition, almost 23 percent of rural households would not be classified as agricultural households, and 21 percent of urban households earn enough agricultural income to be classified as agricultural households. Even at 30 percent, 13 percent of urban households would be classified as agricultural households, and 37 percent of rural households would not be classified as agricultural households. Using the 30 percent cut-off point, two countries, Cambodia and Madagascar have much higher shares of urban households who can be classified as farmers. To a lesser extent, Malawi and Ghana also have high urban agricultural income shares. This might be due to a different rural-urban classification system for these countries, to more extensive urban and peri-urban agricultural production, or to a combination of these factors. Similarly the Bangladesh data report a much lower share of rural households earning more than 30 percent of their income from direct agricultural production.

Table 9 shows the share of households that earn more than 30 percent of their income from agricultural production and the share of households where the household head classifies its sector of operation as agriculture. For the 14 surveys for which this information was estimated, the share of agricultural households who classify their occupation as agriculture is about 46 percent. For the same countries, 46 percent of the households have income above that the 30 percent threshold. Of course the share of households classified as agricultural under the two definitions varies significantly for individual countries. In both urban and rural areas, the unweighted averages point to more households defining their sector of occupation as agriculture than the households earning more than 30 percent of their income from agriculture; but, this result is not robust across countries and thus no generalization should be made.

One more definitional measure is the share of households that classify their occupations as agriculture and also have agricultural incomes which constitute more than 30 percent of their total income. Only about 57 percent of rural households (32 percent of all households) meet each of the three definitions of being an agricultural household. That is, they are located in rural areas, the head of the household lists agriculture as his main occupation, and they earn more than 30 percent of total household income directly from agricultural activities. It is worthwhile to highlight the heterogeneity of the households under each definition. The physical and social characteristics, needs and capabilities of households vary depending on the definition applied; therefore, the interpretation of data about agricultural households should consider those differences.

Another important lesson from this data is the misleading nature of identifying rural with agricultural and urban with non agricultural activities. An important proportion of households with significant agricultural incomes reside in urban areas, as demonstrated for Madagascar, Cambodia, Malawi, Ghana, and Nepal. Overall, more than 12 percent of urban households identify agriculture as their main occupation. Similarly, a significant number of households in rural areas, ranging from 15 to 65 percent, do not earn more than 30 percent of their income from agriculture. While some share of these households might be agricultural labourers etc, the numbers are still large. On average, almost 30 percent of rural households do not list agriculture as their main occupation.

The main message of this discussion is that even if no universal definition of an agricultural household exists, analysts of data from low-income countries need to be especially aware of the sensitivity of their results to small changes in the definition of an agricultural household. Not only do the number and proportion of households defined as agricultural vary, but their

profile and overall characteristics may also change in several important dimensions having far-reaching implications on an analysis if not adequately considered (see Aksoy et al. (2009) for a more comprehensive discussion).

5. Summary of Recommendations and Conclusions

In this paper we have tried to contribute an analyst and practitioner perspective to both the conceptual and analytical motivations, as well as the practical concerns related to measuring household level income in developing countries, focusing in particular on multi-topic household survey data. It is worth emphasising how from the perspective of the analyst there are several additional uses of household income information beyond those most widely cited in the statistical literature (income as a measure of well-being, and as a source of data on economic inequality). The income data from the RIGA database, for instance, are being used for a number of analytical application as diverse as looking at the importance of specific subsectors within the rural economy (e.g. the livestock sector, or the staple vs. commercial crop sector), the extent to which farm rural households are oriented towards commercialization rather than subsistence production, or the impact of economic shocks on household well-being (for instance in the case of the recent food price crisis). None of these analyses are possible at a reasonable level of accuracy if reliable information on income and its components is not available.

We have also tried to emphasise how in low-income countries the characterisation of the agricultural household is particularly sensitive to issues of definition. Not only do the number and proportion of households defined as agricultural/farm change, but their profile and overall characteristics change in several important dimensions as different criteria and thresholds are applied.

On a more specific level, this paper focused on the process of constructing income aggregates with the aim of addressing the most common obstacles encountered when undertaking this endeavour. Accurately estimated income aggregates serve as important indicators of household well-being and provide a wealth of information about income strategies, inequality and the structure of the overall economy that can complement welfare and poverty analyses which are conducted with consumption data. We identify several problems that analysts encounter in the process of creating income aggregates and provide recommendations for how to surmount those obstacles. However, it is important to note that the accuracy of the constructed income aggregates and analysts' ability to deal with these issues are largely dependent on data availability and data quality which ultimately depend on the proper design and implementation of household surveys. A well-designed household survey that aims to provide accurate data for both consumption and income aggregations enables analysts to create more precise income estimates with fewer assumptions tied to them, ultimately increasing the usability of the constructed indicator as a complementary measure of household welfare and improving the quality of livelihoods analyses based on survey data.

We identified several common categories of issues across various surveys that consistently hinder the estimation of precise income measures. Our main conclusions from these issues are summarized below.

Reference Periods and Frequencies. The key consideration in dealing with reference periods and frequencies should be to ensure that the information provided from the questions asked in

the survey is sufficient to create reliable annual estimates and that the chosen reference period coincides with respondent's ability to report as accurate figures as possible. It is important that the chosen reference period is appropriate for the type of survey module such that reference periods will vary across survey modules. In general, infrequently earned income is better off reported with a broader reference period whereas consistent, frequent and regular sources of income should be reported on a shorter time frame. It is vital for surveys to include questions regarding frequency with which income is received in order to enable the annualization of income indicators.

Units and Coding. Inconsistency in units and coding across modules and questionnaires and the lack of appropriate conversion factors can result in problems with data processing. The usage of country-specific units, with which survey respondents are more familiar, allows survey respondents to more accurately report their earnings and expenditures and thus should not be eliminated for the sake of standardized ones. Nevertheless, it should be ensured that whatever kinds of units are chosen, that they are used consistently across modules to enable intra-module comparison. It is absolutely necessary that equivalence scales and/or tables for the conversion of local units to standard units be provided with the survey. With respect to coding, crop, occupational and other codes should be structured in a manner that enables intra-module, intra-year and international comparison.

Dealing with Costs. Income aggregates are based on sources of regular and/or recurring income and expenditures. Household surveys should be structured in a manner that clearly separate questions regarding investments and durables. The misclassification of expenditures can result in understated income figures and a bias the final income aggregate estimates. Clarification on whether income reported is net or gross should also be included followed up by the value of all possible withholdings.

The principal considerations presented in this paper are a valuable set of points that we recommend be reflected upon by future analysts as well as those involved in the design and drafting of household surveys. The resolution of these issues in future household surveys will not only benefit the generation of cross-survey comparable income estimates, but also the estimation of income for individual countries, particularly when an analysis spans more than one year. This paper is certainly not an exhaustive list of the issues for dealing with household surveys but it does establish a core set of "best practices" based on the first-hand experience of creating comparable and consistent income aggregates for 34 household surveys, which should serve as a set of useful considerations in the future design of household surveys and the analysis of data obtained from those surveys.

References

Aksoy, A., Beverinotti, J., Covarrubias, K. & Zezza, A. (2009) Household Income Structures in Low Income Countries, Forthcoming.

Canberra Group (2001) *Expert Group on Household Income Statistics: Final Report and Recommendations*. Ottowa.

Chander, R., Grootaert, C., & Pyatt, G. (1980) Living Standards Surveys in Developing Countries, The World Bank, Living Standards Measurement Study Working Paper No.1.

Coudouel, A., Hentschel, J.S., & Wodon, Q.T. (2002) Poverty Measurement and Analysis, in: *PRSP Sourcebook*, The World Bank, Washington D.C.

Davis, B., Winters, P., Carletto, G., Covarrubias, K., Quiñones, E.J., Zezza, A., Stamoulis, K., Azzarri, C., & DiGiuseppe, S. (2009) A Cross Country Comparison of Rural Income Generating Activities, *World Development*, Forthcoming.

Deaton, A. (1997). *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy.* Johns Hopkins University Press.

Deaton, A. & Grosh, M. (Eds.) (1998) Consumption, in: *Designing Household Survey Questionnaires for Developing Countries: Lessons from Ten Years of LSMS Experience*, Grosh, Margaret & Paul Glewwe, eds., The World Bank.

Deaton, A. & Zaidi, S. (2002) Guidelines for Creating Consumption Aggregates for Welfare Analysis, The World Bank, Living Standards Measurement Study Working Paper 135.

de Ferranti, D., Perry, G.E., Foster, W., Lederman, D., & Valdés, A. (2005) *Beyond the City: The Rural Contribution to Development*. The World Bank: Washington, D.C.

De Mel, S., McKenzie, D.J., & Woodruff, C. (2007) Measuring Microenterprise Profits: Don't ask how the sausage is made, World Bank Policy Research Working Paper 4229.

FAO (1998) *The State of Food and Agriculture: Rural Non-Farm Income in Developing Countries*, Rome: Food and Agriculture Organization.

Glewwe, P. (2005) An overview of questionnaire design for household surveys in developing countries, in: *Household Sample Surveys in Developing and Transition Countries,* United Nations, Department of Social and Economic Affairs, Statistics Division, Studies in Methods, Series F, No 96.

Gradín, C., Cantó, O., & del Río, C. (2004) Inequality, Poverty and Mobility: Choosing Income or Consumption as Welfare Indicators, Prepared for the 28th General Conference of The International Association for Research in Income and Wealth, Cork, Ireland, 22-28 August.

Grosh, M. & Glewwe, P. (2000) *Designing Household Survey Questionnaires for Developing Countries: Lessons from Fifteen Years of Living Standard Measurement Study*, The World Bank.

Grootaert, C. (1982) The Conceptual Basis of Measures of Household Welfare and their Implied Survey Data Requirements, The World Bank, Living Standards Measurement Study Working Paper No.19.

Haggblade, S., Hazell, P.B.R. & Reardon, T. (2007) *Transforming the Rural Nonfarm Economy: Opportunities and Threats in the Developing World*.

Hentschel, J., & Lanjouw, P. (1996) *Constructing and Indicator of Consumption for the Analysis of Poverty: Principles and Illustrations with Reference to Ecuador*, The World Bank, Living Standards Measurement Study Working Paper No.124.

ILO (2003) Report II: Household income and Expenditure Statistics, Prepared for the Seventeenth International Conference of Labour Statisticians, Geneva, 24 November.

Kamanou, G., Ward, M., & Havinga, I. (2005) Statistical Issues in Measuring Poverty from Non-Survey Sources, Chapter 5 of Draft Resolution.

Kozel, V. (1990) The Composition and Distribution of Income in Cote d'Ivoire, The World Bank, Living Standards Measurement Study Working Paper No.68.

Lanjouw, P. and Feder G. (2001) Rural Nonfarm Activities and Rural Development: From Experience towards Strategy, Rural Development Strategy Background Paper #4, The World Bank.

Lipton, M., & Ravallion, M. (1995) Poverty and policy, in: *Handbook of development economics, Vol. 3*, Behrman, J. & Srinivasan, T.N. (Eds). Amsterdam: North Holland.

Johnson, M., McKay, A.D. & Round, J.I. (1990) "Income and Expenditure in a System of Household Accounts: Concepts and Estimation" The World Bank, Social Dimensions of Adjustment in Sub-Saharan Africa Working Paper Series, No. 10.

McKay, A. (2000) Should the Survey Measure Total Household Income? in: *Designing Household Survey Questionnaires for Developing Countries: Lessons from Fifteen Years of Living Standard Measurement Study*, Grosh, M., & Glewwe, P. (Eds.), The World Bank, 83-104.

Mohadjer, L., & Choudhry, G.H. (2001) Adjusting for Missing Data in Low-Income Surveys, in: Citro, C., Moffitt, R., & Ver Ploeg, M. (Eds.) *Data Collection and Research Issues for Studies of Welfare Populations*. National Academy Press, Washington, DC.

Nestic, D. (2003) Inequality in Croatia in the Period from 1973 to 1998, Institute of Public Finance, Occasional Paper No. 17.

Pettersson, H. (2005) Survey design and sample design in household budget surveys, in: *Household Sample Surveys in Developing and Transition Countries,* United Nations, Department of Social and Economic Affairs, Statistics Division, Studies in Methods, Series F, No 96.

Ravallion, M. (1992) Poverty Comparisons: A Guide to Concepts and Methods, The World Bank, Living Standards Measurement Study Working Paper No.88.

Richard R. (1994) Issues relating to the UN system of national accounts and developing countries, *Journal of Development Economics*, 44, 77-85.

Saunders, C. & Grootaert, C. (1980) Reflections on the LSMS Group Meeting, The World Bank, Living Standards Measurement Study Working Paper No.10.

Scott, C., de Andrew, P.T.A., & Chander, R. (1980) Conducting Surveys in Developing Countries: Practical Problems and Experience in Brazil, Malaysia, and the Philippines, The World Bank, Living Standards Measurement Study Working Paper No.5.

Tanzi, V. & Howell H. Zee (2000) Tax Policy for Emerging Markets: Developing Countries, *National Tax Journal.* 53(2): 299-322. June.

Wahab, M. & Wahab, A. (1980) Income and Expenditure Surveys in Developing Countries: Sample Design and Execution, The World Bank, Living Standards Measurement Study Working Paper No. 9.

Winters, P., De la O, A.P., Quiñones, E.J., Hertz, T., Davis B., Zezza, A., Covarrubias, K., & Stamoulis, K. (2008) Rural Wage Employment in Developing Countries, Under review at *Oxford Development Studies*.

The Wye City Group Handbook, Rural Households' Livelihood and Well-Being: Statistics on Rural Development and Agriculture Household Income. (2007) United Nations: New York and Geneva.

Table 1: *RIGA Surveys*

Country	Name of Survey	Year of Survey
Africa		
Ghana	Ghana Living Standards Survey- Round Two	1992
Ghana*	Ghana Living Standards Survey- Round Three	1998
Kenya	Kenya Integrated Household Budget Survey	2004-2005
Madagascar	Enquête Permanente Auprès des Ménages	1993-1994
Madagascar	Enquête Permanente Auprès des Ménages	2001
Malawi*	Integrated Household Survey-2	2004-2005
Nigeria*	Living Standards Survey	2004
Asia		
Cambodia	Socio-Economic Survey	2004
Bangladesh*	Household Income-Expenditure Survey	2000
Bangladesh	Household Income-Expenditure Survey	2005
Indonesia	Family Life Survey- Wave 1	1992
Indonesia*	Family Life Survey- Wave 3	2000
Nepal	Living Standards Survey I	1995-1996
Nepal*	Living Standards Survey I	2003
Pakistan	Integrated Household Survey	1991
Pakistan	Integrated Household Survey	2001
Vietnam	Living Standards Survey	1992-1993
Vietnam*	Living Standards Survey	1997-1998
Vietnam*	Living Standards Survey	2000
Eastern Europe		
Albania	Living Standards Measurement Survey	2002
Albania*	Living Standards Measurement Survey	2005
Bulgaria	Integrated Household Survey	1995
Bulgaria*	Integrated Household Survey	2001
Tajikistan*	Living Standards Survey	2003
Tajikistan	Living Standards Survey	2007
Latin America		
Bolivia	Encuesta de Hogares	2005
Ecuador*	Estudio de Condiciones de Vida	1995
Ecuador	Estudio de Condiciones de Vida	1998
Guatemala*	Encuesta de Condiciones de Vida	2000
Guatemala	Encuesta de Condiciones de Vida	2006
Nicaragua*	Encuesta de Medición de Niveles de Vida	1998
Nicaragua*	Encuesta de Medición de Niveles de Vida	2001
Panama	Encuesta de Condiciones de Vida	1997
Panama*	Encuesta de Condiciones de Vida	2003

* Denotes the surveys also forming part of the RIGA-L dataset.

Table 2: *RIGA Income Aggregate Components*

Wage employment	Agricultural	Agriculture & fishing, Skilled
		Agriculture & fishing, Unskilled
		Agriculture & fishing, Unknown
	Non-agricultural	Mining- Skilled
		Mining- Unskilled
		Mining- Unknown
		Manufacturing- Skilled
		Manufacturing- Unskilled
		Manufacturing- Unknown
		Electricity & Utilities, Skilled
		Electricity & Utilities, Unskilled
		Electricity & Utilities, Unknown
		Construction, Skilled
		Construction, Unskilled
		Construction, Unknown
		Commerce, Skilled
		Commerce, Unskilled
		Commerce, Unknown
		Transport, Storage, Skilled
		Transport, Storage, Unskilled
		Transport, Storage, Unknown
		Finance, Insurance, Skilled
		Finance, Insurance, Unskilled
		Finance, Insurance, Unknown
		Services, Skilled
		Services, Unskilled
		Services, Unknown
		Other, Skilled
		Other, Unskilled
		Other, Unknown
Self Employment	Agricultural	Agricultural Processing
	Non-agricultural	Mining
		Manufacturing
		Electricity & Utilities
		Construction
		Commerce
		Transport, Storage and Communications
		Finance, Insurance and Real Estate
		Services
		Other
Crop Production	Crop income (1)	
	Crop income (2)	
Livestock Production	Livestock income	
Transfers	Private Transfers	
	Public Transfers	Pensions
		Social Transfers
Other Income Sources	Nonfarm Rental Income	
	Farm Rental Income	
	Other	
Total Income	Total income	Wage + Self employment + Crop (*1 or 2*) + Livestock + Transfers + Other

Table 3: *Methods to Determine Duration and Frequency Classifications*

Methods	
Duration	- Months: If the number of months is not available, the number of days per year is divided by days per month to estimate the number of months per year worked.
	- Weeks: In the absence of months per year, weeks per year are used to designate full year or part year employment. It is estimated that 44 weeks are equivalent to 10 months.
	- Another way to determine the number of weeks worked per year is multiplying weeks per month by the number of months worked.
Frequency	- Hours per week: If the number of hours per week is unavailable, we divide hours per month by 4.35 (the estimated number of weeks per month) to get hours per week.
	- Days per week: In the absence of hours per week, days per week are used. Five days or more per week are assumed to designate full time status and less than five days per week as part time status.
	- When days per week are not available but hours per day are available, hours per day are multiplied by the number of days worked in a week.[21]

[21] In general, the median days per week worked for job 1 is six in most RIGA countries/surveys (the means and modes also hover directly around six). As a result, when days per week are not available six days per week is relied on for the purpose of facilitating analysis.

Table 4: *Recommended Reference Periods*

Activity	Recommended Reference Period
Food consumption	Weekly or fortnightly
Crop production	Seasonal (previous season; or previous wet season and previous dry season, depending on the number of harvests/production seasons per year in the given country)
Crop expenditures	Annual
Livestock production	Annual
Livestock expenditures	Monthly
Wage employment	Weekly/monthly earnings for current or most recent work; annual earnings for infrequent or past jobs.
Self employment	Monthly (with data on frequency, as well as seasonality of earnings)
Transfers	Monthly (with frequency of receipt) or Annual
Other income	Annual

Table 5: *Unit Reporting in Agricultural and Consumption Modules*

Survey	Number of Units	Comments	Equivalence Incl.?
Albania 2002	3 (crop; livestock) 1 (land)	Only standard units (gram, kilo, ton, litre, etc)	N/A
Albania 2005	3 (crop; livestock) 1 (land)	Only standard units.	N/A
Bosnia 2001	9 (consumption; crop; livestock) 5 (land)	Only standard units.	Yes
Bulgaria 1995	3 (consumption; crop; livestock) 1 (land)	"Unit" is an unspecified unit.	N/A
Bulgaria 2001	3 (crop; livestock) 1 (land)	"Unit" is an unspecified unit.	N/A
Ecuador 1995	22 (consumption, crop)	Includes non-standard units.	No
Ecuador 1998	23 (consumption, crop)	Includes non-standard units.	Yes
Ghana 1992	25 (crop)	Includes non-standard units	No
Ghana 1998	25 (crop)	Includes non-standard units	
Guatemala 2000	66 (total possible units)	Includes non-standard units	No
Malawi 2004	20 (crop) 4 (land)	Consistent across sections. Includes non-standard units.	No
Nepal 1996	10 (crop) 4 (land)	Includes non-standard units.	Yes
Nicaragua 2001	57 (total possible units) 27 (crop) 3 (land)	Includes non-standard units.	Yes, but incomplete.
Panama 1997	77 (crop, consumption)	Includes non-standard units.	No
Panama 2003	45 (crop, consumption) 2 (land)	Includes non-standard units.	No
Vietnam 1998	4 (crop; consumption) 1 (land)	Only standard units.	N/A

Table 6: *Reporting of Gross vs. Net Wage Income*

Survey	Gross or Net	Are Tax Expenditures & Other Deductions Reported?
Albania 2002	Net	No
Albania 2005	Net	No
Bulgaria 1995	Both	Yes
Bulgaria 2001	Both	Yes
Ecuador 1995	Gross	Yes: Reported in Annual Non-Food Expenditures Module
Ecuador 1998	Gross	Yes: Reported in Annual Non-Food Expenditures Module[22]
Ghana 1992	Unclear/Both	Yes
Ghana 1998	Unclear/Both	No
Guatemala 2000	Gross	No
Malawi 2004	Not Specified	No
Nepal 1996	Unclear/Both[23]	Yes: Reported in Non-Food Expenditures Module
Nicaragua 2001	Net	Yes: Reported in Non-Food Expenditures Module.
Panama 1997	Net	Yes: Reported in Annual Other Household Income and Expenditures Module.
Panama 2003	Gross	Yes: Reported in Annual Other Household Income and Expenditures Module.
Vietnam 1998	Unclear[24]	Unclear: Reported in Annual Non-Food Expenditures Module

[22] Same as Ecuador 1995.

[23] Household can report gross or net income, but if it reports gross (distinguished by a question asking "are taxes already deducted?") it is not asked to report how much in taxes it will pay. Income taxes are accounted for in the Non-Food Expenditures module; but, it is unclear if households that report net pay in the employment module report taxes in the Expenditures module. To make a clear distinction in either case is impossible and it is likely that either (1) wages reported as gross would be over-stated if taxes from the Non-Food Expenditures module are not accounted for or (2) wages reported as net would be understated if taxes are reported in the Expenditures module.

[24] The questionnaire states that in the reporting of the wage, the respondent should "exclude contributions to pensions, health insurance, etc" yet this does not specify if taxes should also be excluded.

Table 7: *Share of rural income generating activities in total income*

	Income-generating activity												
	Group I							Group II		Group III			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(1) + (2) + (3)	(4) + (5) + (6) + (7)	(1) + (2)	(4) + (5)	(6) + (7)	(3) + (4) + (5) + (6) + (7)
Country and year	Agriculture-Crops	Agriculture - Livestock	Agricultural wage employment	Non-farm wage employment	Non-farm self-employment	Transfers	Other	*Agricultural total*	*Non-Agricultural Total*	*On-Farm Total*	*Non-farm total*	*Transfers & Other*	*Off-farm Total*
Malawi 2004	56.1%	9.4%	11.4%	7.4%	8.7%	6.6%	0.3%	*77.0%*	*23.0%*	*65.5%*	*16.1%*	*7.0%*	*34.5%*
Madagascar 1993	57.3%	13.2%	6.5%	6.1%	8.5%	6.2%	2.2%	*77.0%*	*23.0%*	*70.5%*	*14.6%*	*8.4%*	*29.5%*
Bangladesh 2000	15.5%	1.2%	20.2%	19.9%	16.4%	13.4%	13.4%	*36.9%*	*63.1%*	*16.6%*	*36.4%*	*26.8%*	*83.4%*
Nepal 2003	20.3%	17.7%	12.6%	21.1%	9.2%	16.8%	2.4%	*50.6%*	*49.4%*	*38.0%*	*30.2%*	*19.2%*	*62.0%*
Ghana 1998	55.0%	4.4%	1.4%	9.6%	20.5%	8.5%	0.5%	*60.9%*	*39.1%*	*59.4%*	*30.1%*	*9.0%*	*40.6%*
Tajikistan 2003	37.3%	17.4%	16.9%	11.5%	1.1%	15.5%	0.3%	*71.6%*	*28.4%*	*54.7%*	*12.6%*	*15.7%*	*45.3%*
Vietnam 1998	41.5%	14.8%	5.9%	9.2%	21.2%	7.0%	0.3%	*62.2%*	*37.8%*	*56.3%*	*30.5%*	*7.3%*	*43.7%*
Nigeria 2004	73.5%	4.3%	2.0%	7.1%	10.8%	1.7%	0.6%	*79.8%*	*20.2%*	*77.8%*	*17.8%*	*2.4%*	*22.2%*
Pakistan 2001	21.2%	11.4%	8.9%	28.8%	10.7%	14.5%	4.6%	*41.4%*	*58.6%*	*32.6%*	*39.5%*	*19.1%*	*67.4%*
Nicaragua 2001	21.1%	14.3%	21.4%	21.3%	11.1%	6.1%	4.6%	*56.9%*	*43.1%*	*35.4%*	*32.5%*	*10.7%*	*64.6%*
Indonesia 2000	23.8%	2.1%	9.7%	20.3%	17.6%	22.9%	3.6%	*35.5%*	*64.5%*	*25.8%*	*37.9%*	*26.5%*	*74.2%*
Guatemala 2000	27.6%	2.6%	19.9%	20.2%	12.4%	16.9%	0.5%	*50.1%*	*49.9%*	*30.2%*	*32.6%*	*17.3%*	*69.8%*
Albania 2005	17.2%	23.3%	2.8%	18.1%	7.4%	28.0%	3.2%	*43.3%*	*56.7%*	*40.5%*	*25.5%*	*31.2%*	*59.5%*
Ecuador 1995	9.0%	3.4%	10.3%	39.1%	23.2%	8.9%	6.0%	*22.8%*	*77.2%*	*12.5%*	*62.3%*	*14.9%*	*87.5%*
Bulgaria 2001	3.9%	12.0%	4.6%	16.5%	1.3%	60.5%	1.2%	*20.5%*	*79.5%*	*15.9%*	*17.8%*	*61.7%*	*84.1%*
Panama 2003	15.8%	2.0%	16.7%	27.1%	22.6%	14.6%	1.2%	*34.6%*	*65.4%*	*17.8%*	*49.7%*	*15.7%*	*82.2%*

Source: Davis et al. (2009).

Table 8: *Percent of rural households with diversified and specialized income generating activities*

		Principal Household Income Source (>= 75% of Total Income)					
	Diverse Income Portfolio	**Ag Wage**	**Nonag wge**	**Self Emp**	**Transfers**	**Other**	**Farm**
Malawi 2004	39.3%	5.5%	5.6%	5.0%	2.5%	0.0%	42.0%
Madagascar 1993	30.6%	1.3%	2.8%	4.0%	1.4%	0.4%	59.4%
Bangladesh 2000	52.4%	11.4%	12.2%	10.5%	5.5%	2.2%	5.9%
Nepal 2003	52.5%	4.3%	11.7%	4.9%	6.9%	0.3%	19.4%
Ghana 1998	24.0%	0.6%	6.2%	15.4%	3.4%	0.2%	50.1%
Tajikistan 2003	54.3%	4.5%	3.7%	0.6%	4.8%	0.0%	32.0%
Vietnam 1998	44.3%	2.1%	1.9%	12.8%	1.2%	0.1%	37.7%
Nigeria 2004	14.7%	1.0%	5.5%	7.8%	0.9%	0.2%	69.9%
Pakistan 2001	36.1%	5.4%	19.3%	6.6%	9.1%	1.6%	21.9%
Nicaragua 2001	43.8%	12.7%	14.1%	6.2%	0.7%	0.4%	22.1%
Indonesia 2000	41.5%	5.9%	14.0%	10.5%	11.5%	1.1%	15.6%
Guatemala 2000	54.6%	8.7%	12.8%	5.6%	5.0%	0.1%	13.2%
Albania 2005	54.8%	1.4%	9.1%	5.0%	9.8%	0.5%	19.4%
Ecuador 1995	45.5%	13.2%	11.7%	8.9%	2.3%	1.1%	17.4%
Bulgaria 2001	41.1%	1.8%	9.3%	1.4%	43.1%	0.1%	3.4%
Panama 2003	48.8%	9.6%	20.0%	10.0%	6.6%	0.1%	4.8%

Outlined cells represented the greatest share of households for a given country dataset; shaded cells represent the highest among specializing households.

Source: Davis et al. (2009)

Table 9: *Shares of Rural and Agricultural Income Households*

	GNI per capita (PPP) 2005	Percent Rural	Shares of Households with Positive Agricultural Income			Shares of Households with Agricultural Income greater than 10%			Shares of Households with Agricultural Income Greater than 30%			Share of Agricultural Households by Occupation		
			Urban	Rural	Total	Urban	Rural	Total	Urban	Rural	Total	Urban	Rural	Total
Ethiopia (2000)	630	50.7	13.7	91.4	71.3	11.0	84.5	65.4	7.9	76.1	58.4	6.7	89.4	67.9
Malawi (2004)	640	88.1	39.0	95.8	89.0	31.9	92.5	85.2	21.5	85.1	77.5	21.8	72.3	66.3
Zambia (1998)	1090	47.8	33.9	93.2	70.1	20.2	87.7	61.4	11.0	78.5	52.2	5.6	75.6	48.4
Nepal (2003)	960	87.4	56.0	95.6	90.6	32.8	85.6	78.8	18.2	65.8	59.8	20.8	53.7	49.5
Ghana (1998)	1140	63.3	32.1	84.9	65.5	28.2	77.9	59.7	22.7	69.4	52.3	22.7	67.2	49.7
Cambodia (1999)	1380	60.0	49.1	92.6	86.1	42.2	87.3	80.6	35.4	78.1	71.7	n.a.	n.a.	n.a.
Bangladesh (2000)	1120	79.7	19.1	71.9	61.3	8.3	51.7	43.0	4.3	32.3	26.7	10.5	57.6	48.2
Vietnam (1998)	2100	71.2	74.5	98.5	92.7	20.8	90.8	73.9	10.9	78.8	62.3	11.2	76.8	60.8
Madagascar (2001)	820	75.8	34.0	83.1	71.2	31.8	79.8	68.2	29.6	74.4	63.5	24.1	74.6	62.3
Nicaragua (2001)	2250	43.9	72.8	92.7	80.5	25.6	74.2	44.4	12.2	54.9	28.7	11.2	59.2	29.8
Pakistan (2001)	2280	71.0	18.8	71.1	55.9	9.9	59.3	44.9	6.3	49.9	37.3	4.4	42.5	31.5
Bolivia (2002)	3610	42.0	95.7	96.5	96.0	18.4	79.5	41.3	4.2	64.3	26.7	5.7	78.6	33.0
Guatemala (2000)	4030	56.7	43.3	90.8	70.2	18.1	69.1	47.0	9.0	48.4	31.4	12.3	58.9	38.7
Peru (2003)	6030	35.9	14.0	89.8	37.5	8.0	76.3	29.2	4.2	55.9	20.3	14.6	84.9	36.4
Ecuador (1995)	6390	37.4	21.8	80.7	43.9	9.8	64.1	30.1	4.6	45.5	19.9	6.4	80.0	26.4
Unweighted average	2295	60.72	41.19	88.57	72.12	21.14	77.36	56.87	13.49	63.83	45.92	12.7	69.4	46.4

Source: Aksoy, et al. (2009)

Figure 1: *Share of rural non-agricultural income by per capita GDP*

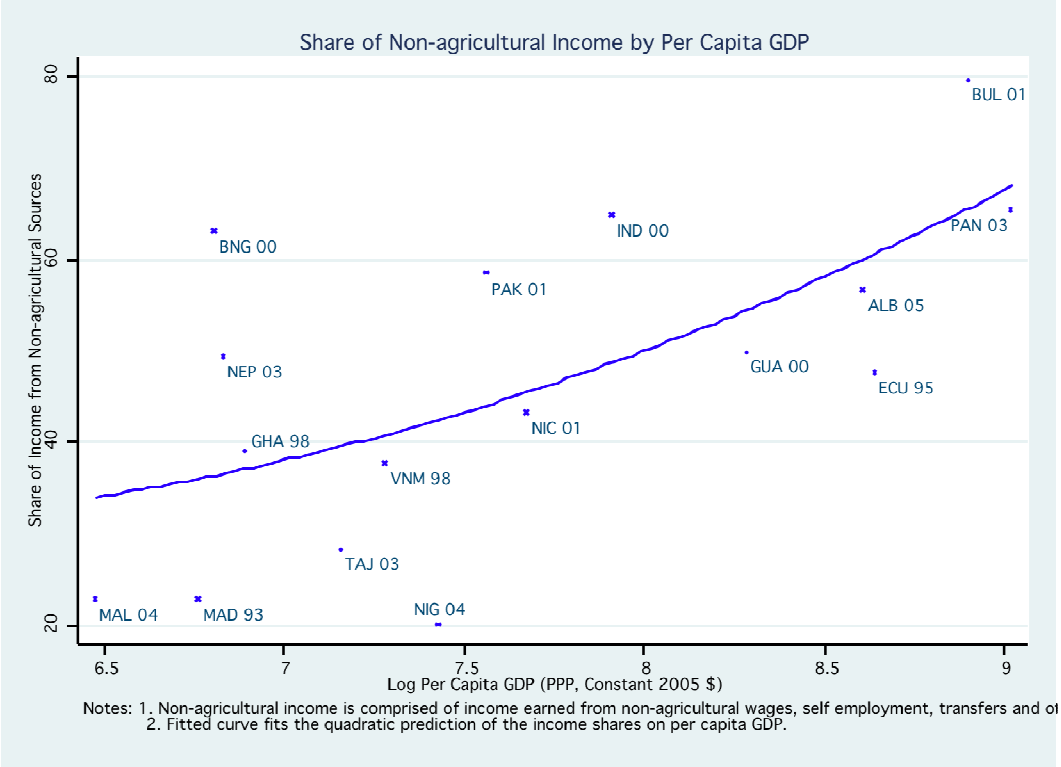
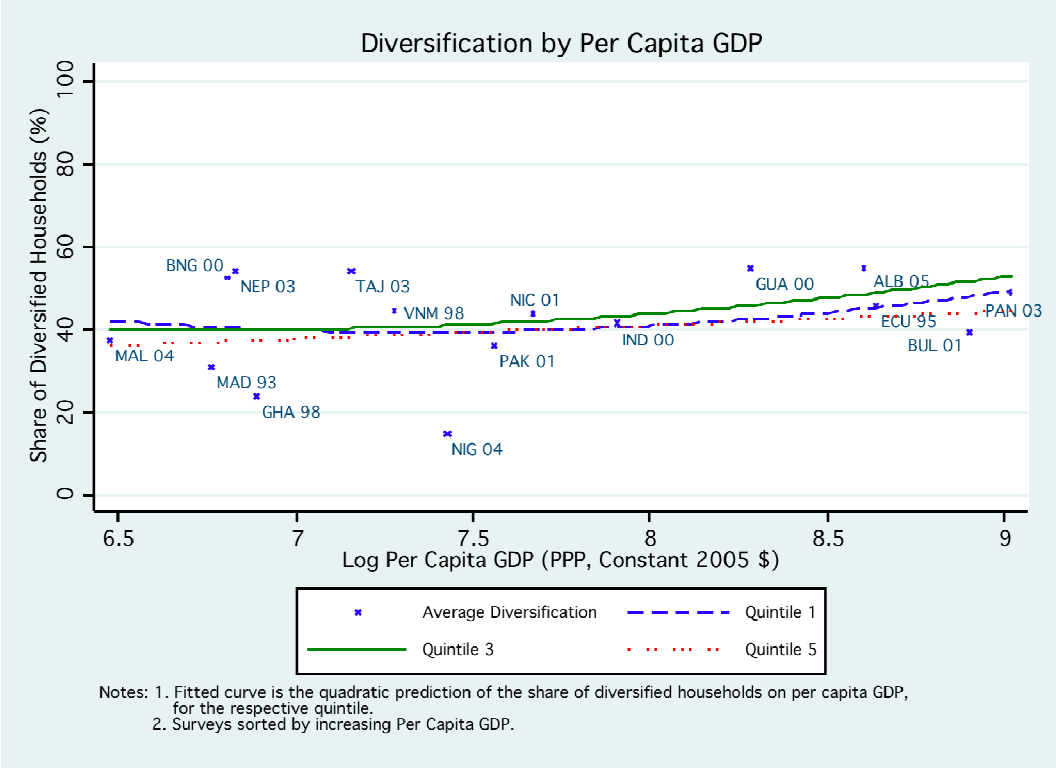
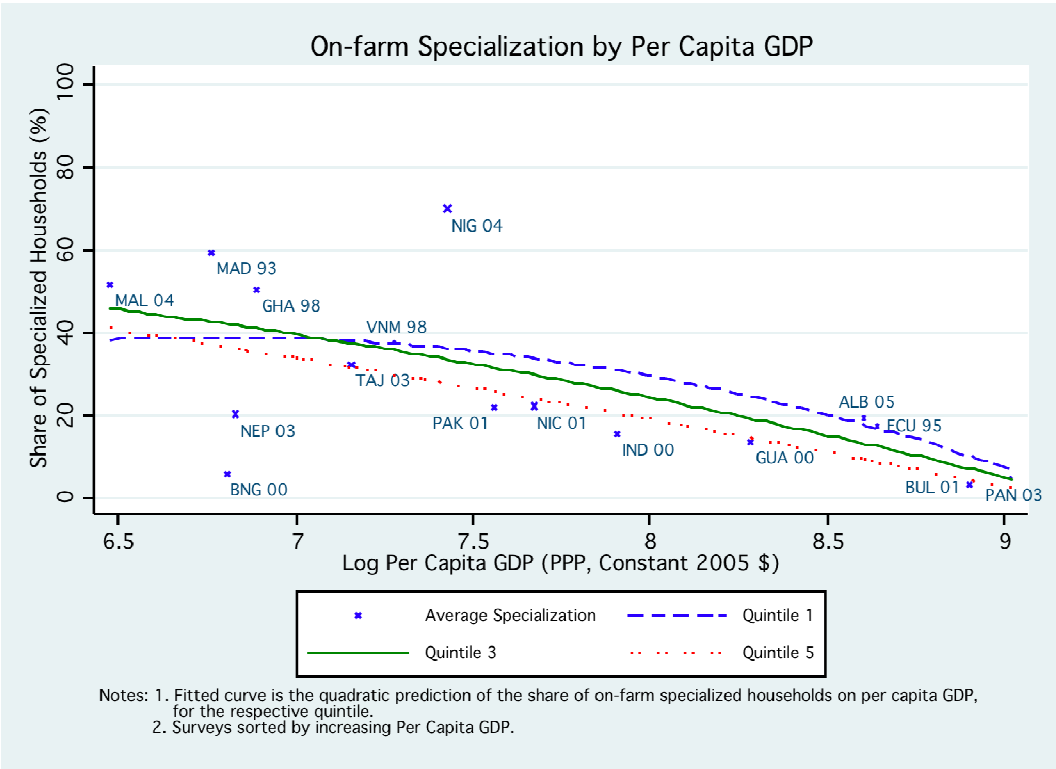


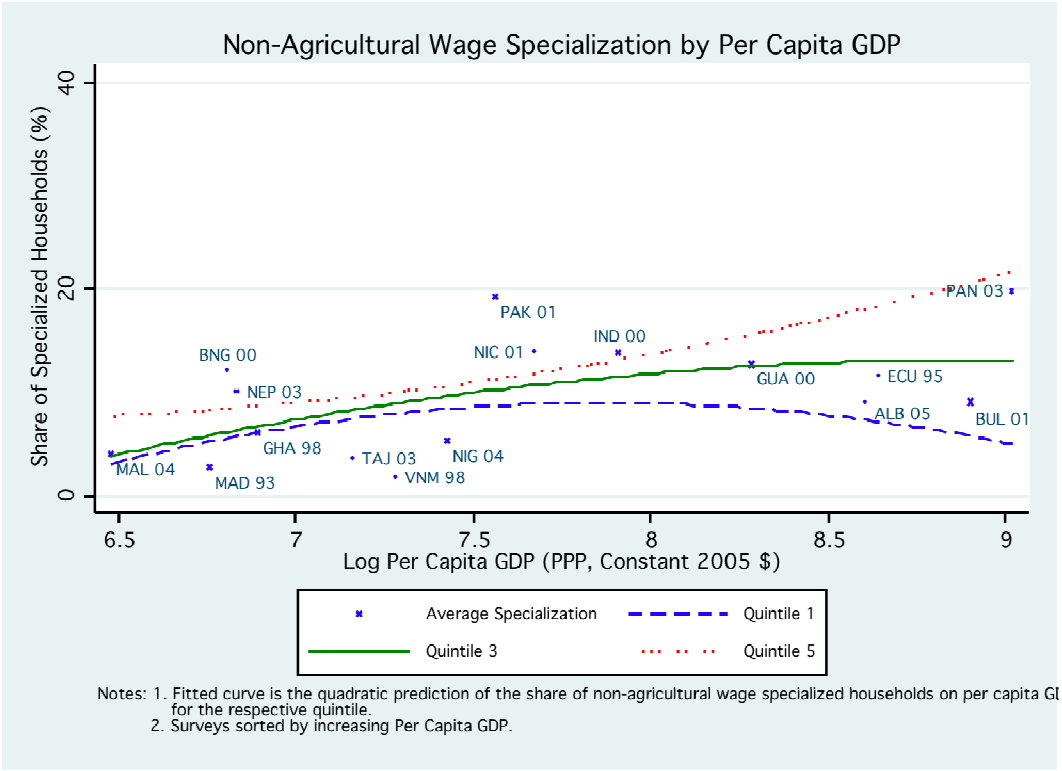
Figure 2: *Share of diversified, on-farm and non-agricultural wage specializing households, by per capita GDP*



Notes: 1. Fitted curve is the quadratic prediction of the share of diversified households on per capita GDP, for the respective quintile.
2. Surveys sorted by increasing Per Capita GDP.



Notes: 1. Fitted curve is the quadratic prediction of the share of on-farm specialized households on per capita GDP, for the respective quintile.
2. Surveys sorted by increasing Per Capita GDP.



Notes: 1. Fitted curve is the quadratic prediction of the share of non-agricultural wage specialized households on per capita GI for the respective quintile.
2. Surveys sorted by increasing Per Capita GDP.

Annex 1: *Detailed issues concerning outlier checks and imputation procedures in RIGA*

A major concern in the entire RIGA project has been that of avoiding that differences in definitions and data manipulation procedures could affect the results of the analysis. Standard protocols have therefore been developed to treat the data as consistently as possible across surveys. This Annex provides a discussion of the procedures adopted in dealing with outliers and data imputation procedures.

A.1. Outlier Checks.

Raw data are assumed to have been already checked for consistency and erroneous values; therefore, in creating the RIGA income aggregate components, checks for outliers are performed only on constructed variables. Such transformations include, but are not limited to: annualizing the original variable; collapsing the data from individual to household level; and multiplying the price received by the quantity sold. In order to obtain comparable estimates of the income aggregate components, there must be consistency as well in the methodology for outlier checks. Different approaches exist by which extreme values can be identified and imputed; this section describes the RIGA approach for undertaking this task.

The outlier check is performed twice, following the same procedure each time. For a given variable, the values are divided by one to two relevant subgroups, for example, crop code for checking the values of crop sales, or industry sector and skill level for wage employment income. For variables without a logical sorting variable and for variables in which the logical sorting variable does not include a sufficient number of observations in the sorting category, an administrative unit variable, such as district or region, is used as the sorting variable. The choice of the sorting variable for the outlier check must be carefully undertaken so to avoid potential endogeneity when performing empirical analyses. Outliers are identified as values greater or less than three standard deviations from the median value of the variable for the specific sorting group. The outlier values are flagged and then replaced with the median value of the variable within the corresponding sorting category. This procedure is described in more detail, along with its corresponding Stata syntax, in Annex 2. After completing the first check for outliers, the data is summed at the household level (or transformed in some other way, if already at the household level) and final aggregate variables are computed. The outlier check is repeated on these variables using an administrative or geographic unit as the categorical sorting variable.

For all outlier checks, zeroes and missing values in the income and expenditure variables in question are excluded from the calculation of the median and standard deviation as well as from the identification of outlier values in order to achieve accurate imputations. Since some income activities may have a significant share of non-participants, a notable share of the values in income and expenditures may be zero or missing. The exclusion of these values ensures the estimation of the median and standard deviation is not skewed and that households with missing values are not erroneously assigned income or expenditures. Further, in order to avoid mis-identifying and mis-imputing outliers, the standard outlier check is performed on groups of minimum 50 observations whenever possible. Smaller groups have greater variance and as such increase the possibility of flagging values as outliers, even when none may be present. [25]

[25] This criterion is adhered to in most cases; however, in some instances when there are few observations participating in labour employment, outlier check sub-groups may be slightly smaller.

One final outlier check is imposed after the construction of the income aggregate to deal with extreme income shares that arise following the aggregation of all income components. After estimating the shares of the seven principal income categories (agricultural wages, non-agricultural wages, crop, livestock, self employment, transfers and other) in total income, observations whose share is greater than or less than 3 (indicating a percentage share of +/- 300%) are dropped from the aggregate, a procedure that leads to the loss of a small yet distortionary share of observations in each survey, usually representing less than 1 percent of the sample. The fact that such extreme shares exist are a product of two factors. The first is the estimation a net income figure: taking into consideration costs related to income generating activities reveals that some households incur net losses in certain activities. Most often, negative income is obtained in self-employment activities in which households incurred substantial costs (e.g. wholesale purchases or high labour costs) but did not report returns that offset those expenditures. However, since the nature of household pluriactivity is such that the losses in one activity can be offset by gains in another activity, in most cases, only a small share of households have negative total household income (see Table A2). Recognizing this possibility, the second factor driving extreme income shares is the persistence of problematic data. The magnitude of certain losses (or gains) are undoubtedly implausible and are attributed to erroneous values introduced either at the data collection stage (due to recall error, enumerator error, or flawed survey design) or at the data entry stage.

Two amendments for flagging outliers in the original procedure apply in the construction of the wage income aggregates in RIGA-L. First, outlier checks are weighted according to the weights provided by each survey. Secondly, outlier checks use the logged wages of monthly income variables instead of the "raw" wages. Although this approach is diverges from the one taken in the construction of the RIGA household level aggregates, the difference in the number of the cases affected is trivial (see Annex 2 for more detail on this procedure).

Here is it worthwhile to note that the approaches selected by RIGA for identifying and imputing outliers are one of multiple possibilities. Whereas other methodologies exist for dealing with outliers- such as regression imputations, the elimination of all negative values, nearest neighbour matching- the above-described methodology is one that is possible to replicate in a systematic manner across surveys applying limited assumptions and reducing the possibility of confounding empirical work with potential endogeneity.

A.2. Imputations.

One major obstacle in working with survey data is the presence of missing and misreported values. Missing values can be identified in three categories:[26] under-coverage, unit non-response and item non-response. Of these three, it is item non-response which is of concern for the RIGA analysis since it is the failure of the respondent to respond accurately to a question posed by the interviewer or due to the interviewer's failure to ask or record a response to a question. Misreported/miscoded values arise either when the answer provided by the respondent does not logically correspond to the question, or in the data entry stage of the survey. In these instances, recoding and imputation procedures may be necessary.

A.2.a. Recoding. As it is always the case, some respondents choose not to answer or are unable to answer some questions in the questionnaire and as a result the answer is coded as missing. In other cases, the missing value is a natural result of the skip pattern in the questionnaire: for example, if a respondent answers "zero" when asked for the quantity of

[26] For more detailed explanation on these types of missing values, see: Mohadjer, L. & Choudry, G.H. (2001).

milk sold, a subsequent question regarding the price at which the milk was sold would be skipped and coded as missing. Missing values of this sort are preserved and not recoded. However, if the amount of milk sold is reported but the price is reported as missing, the price must be recoded to a logical, imputed value, a procedure described below. For the purpose of recoding, it is thus important to first distinguish between missing values resulting from intentional or unintentional omission and missing values caused by the skip pattern of the questionnaire

In the employment module, recoding time variables when these are either missing or erroneously reported is fundamental for calculating the duration of jobs (i.e. full-time jobs) and earnings at different time levels (i.e. daily wage or monthly income). When time variables are missing but income is reported, we procedure to replace those missing values with the median of the values reported. However, recoding also occurs when missing values are outside the range of possibility, i.e. when the number of reported months worked is 13 or when the reported of days worked in a week is 8. These instances exclusively refer to values that are too high, not those below a certain range. In these instances, values are recoded with the maximum possible value; as in the following example:

Box 3: *Recoding example*

Original (implausible) value	Maximum possible value	Recoded Value
Number of months worked = 13	12	12
Number of days per week worked = 8	7	7

Although the existing values are erroneous, it is more appropriate to replace them with their maximum possible value rather than the median, because it is assumed that the true value of these observations is at or closer to the maximum than to the median of the distribution. Table A3 is a list of maximums used, followed by a brief explanation wherever necessary.

Note that in order to calculate the maximum days of work in a year, *365.25* days are used instead of *365* days, to account for the extra day in the calendar every 4 years (leap year). Also, in the case of hours worked per day, *16* hours is a rather generous assumption, intended to minimize the number of observations that are changed and to allow for the instances when individuals work extraordinary numbers of hours in short periods. Both hours per day and hours per week assumptions (*84* hours per week) allow so that no more than a handful of values are replaced.

A.2.b. Imputation Procedures. Following the recoding procedure, imputations can be used (1) to correct values that are erroneously coded or reported; (2) to value a transaction that has a bearing on poverty or livelihoods but goes unreported or minimally reported in the survey; and (3) to appropriately assign values where they have been intentionally or unintentionally omitted. For the first case, outlier checks, as described above, are one form of imputations undertaken to deal with extreme values. Also corresponding to the first objective is the reassigning of miscoded values. For example, if the answer to a filter question (e.g. "Did you sell...?") is coded "No" yet all subsequent questions (which would correspond to quantity sold, price received, etc.) were answered, indicating the answer to the filter question should actually be "Yes", then the response to the filter question would be modified. Similarly, in the self employment module, if there is a missing value for a response to a question regarding the share of ownership in the household enterprise yet the response to another question states

that the household is the sole owner of the enterprise, it would be necessary to recode the missing value to 100% so to be able to accurately weight self employment income by the share of ownership in the enterprise.

In the second case an imputation can be undertaken when it is necessary to capture the value of a transaction that is not adequately captured by data collection efforts. The most common example of this is the estimation of *imputed rent* in an income aggregate. This value of well-being is described by Deaton and Zaidi (2001) as "a measure in monetary terms of the flow of services that the household receives from occupying its dwelling." It is specifically used, in the consumption aggregate estimation in which current rent expenditure is considered, to assign a value to the use of the dwelling for households who own their dwelling so to avoid placing an upward bias on households that rent in their dwelling with respect to those who do not.

Whereas some methodologies include imputed rent in the income aggregate as a measure of household wealth that augments household disposable income,[27] viewing it as a productive asset from which households can derive consumption value (Johnson et al 1990), the RIGA project does not account for imputed rent in its aggregates. Also, a key principle followed in the RIGA project methodology is to not create income or costs where they are not explicitly reported. Assigning rental expenditures (or any other cost or income) to households that have not reported them in the survey counters the methodological approach of RIGA. Finally, even if the previous reasons were not concerns, a major caveat to the estimation and inclusion of imputed rent in the income aggregate computation when working with developing countries' household surveys is that the approach used is based on information for which there is usually limited data. Imputed rent is generated using a hedonic regression which uses reported data on actual rent paid out, along with other household characteristics, to obtain rental values for households that own their dwelling and therefore do not report rental income. As noted by Johnson et al (1990) housing rental markets for rural areas in developing countries are shallow since most households own their dwelling (see Table A1). Few households rent out an owned dwelling (since ownership is usually limited to one dwelling) and few rent-in their homes. Furthermore, many living standards surveys do not ask households to value the potential rental value of their owned dwelling, a fact that, when taken together with the limited data on actual dwelling rental income, implies that an imputation regression for this purpose would be based on a limited number of observations with little variability, and as such would likely generate poor estimates of dwelling rental values.

The third case in which imputations are used to replace missing values corresponds to estimating prices for variables that represent income or expenditures but are not given a monetary value or do not contain any information on price per unit. To value these income sources, we apply the price imputation approach proposed in Deaton and Zaidi (2002) using prices obtained from (1) sales data from the production modules of the survey; (2) the price questionnaire, if it exists; and finally (3) the consumption module of the survey.[28] Specifically, we use the data available in each of the three sources to obtain an overall unit price for each item. The median unit price is then estimated at different levels of the sample going from the narrowest level (item-unit prices in each cluster) to the broadest level (item-unit prices across the full sample). These imputed prices are then merged with the data on

[27] For example, the Luxembourg Wealth Study and Chile's Encuesta de Caracterización Económica Nacional, CASEN.

[28] The consumption module prices are used only when imputing own consumption. Only in the rare cases when no price can be assigned to an own consumption quantity is the household-reported valuation used.

quantities (e.g. quantity consumed; quantity received as gift) and values are estimated multiplying the quantities with the imputed sales prices, then with the imputed price questionnaire prices, and finally with the imputed prices from the consumption module, in each case applying the cluster-level prices first and following up with the sample-level item-unit prices. This procedure is demonstrated step-by-step in Annex 2 along with the Stata syntax used for the imputation.

The imputation of prices is also particularly relevant for obtaining estimates of household "own consumption," the consumption of crop and livestock items produced by the household. Surveys generally collect information on the quantity consumed but also ask households to place a value on their own consumption, an estimate that is not considered reliable since the household's estimate is unlikely to reflect the market price of the good. In this case it useful to impute, as described above, a set of market prices for valuing own consumption which is then incorporated in the estimation of total net agricultural income.

The RIGA income aggregates contain two estimates of crop income, "*crop1*" and "*crop2*". The former applies the own consumption estimated from the survey's agricultural production module whereas the latter uses own consumption estimated from the survey's food expenditure module.[29] The calculation of both figures leads to interesting findings. Firstly, the results demonstrate that the module from which own consumption is estimated matters greatly in terms of capturing the full extent of own consumption income. Participation rates for *crop2* are almost always greater than those for in *crop1* (see Table A4) indicating that there are a number of true subsistence producers (households that produce exclusively for consumption) that go unaccounted for in the *crop1* variable. This finding conveys the importance of using the reported own consumption from the food expenditure module, and therefore using *crop2*, to estimate total household income, whenever possible.

Taking this as given, it is then worthwhile to compare crop income levels and shares in total income across *crop1* and *crop2*, and thus to note how the contribution of crop income to total income varies significantly depending on the variable used. Although for both cases imputed prices are applied to estimate own consumption, in general more confidence can be placed on the results from *crop2*, since a broader range of items is considered (as *crop2* is generated from the expenditure module) and therefore a broader set of prices is applied[30] for the valuation of consumption. To summarize the differences in results, Table A4 reveals that in 21 of the 23 cases, participation in *crop2* is greater than for *crop1*. Meanwhile, in 12 of the 23 cases highlighted, the income level for *crop2* is greater than *crop1*; however, with respect to the share in total income, for 13 of the surveys is the share of *crop2* income greater than that of *crop1*. Although in theory one would anticipate *crop1* and *crop2* to produce comparable results, in practice, the two estimates are subject to great variability.

Many of the issues discussed in this paper are relevant in the generation of the imputed value of own consumption. Pulling own consumption data from different modules of the survey questionnaire to generate two versions of the same variable highlights the importance of

[29] In a small number of surveys, only one of the two crop variables was generated due to the non-existence of a food expenditure module, or the non-separability of own consumption as an independent variable in the expenditure module, or, rarely, since the agricultural module did not ask households to report their own consumption of production. The surveys affected include Indonesia 1993, Indonesia 2000, Tajikistan 2003, and Cambodia 2004.

[30] Although both *crop1* and *crop2* are generated using imputed prices, *crop1* only uses agricultural module prices which essentially represent farm-gate prices. Market prices would be inappropriate for valuing agricultural production.

reference periods, recall error, frequencies, and units and coding. By not adequately treating these issues in a given survey, it is likely that inconsistencies like those displayed in Table A4 will likely to reappear in other analyses.

One particular point of concern is how differences in the reporting of units and consistency may influence the gap between *crop1* and *crop2* estimates. Assigning prices from the *community price questionnaire* to the quantities reported in the *food expenditure module* in order to obtain the household consumption of own production for the *crop module* raises the issue of comparability of values because the units in which items are reported in each section are not necessarily consistent and obtaining a common unit for all items in each section is often impossible. This problem is most complicated when households are given the flexibility of reporting their crop production in the crop module in any unit. In the case of the Guatemala 2000 ENCOVI survey, households report crop sales and consumption in nearly 66 different units for which conversions are not readily available; not surprisingly, the difference between *crop1* and *crop2* in Table A4 is notable for this survey.

Of course, the disparity in unit-reporting is for the obvious and unavoidable reason that in practice, households do not necessarily consume their food items in the same unit in which they produce. For example, a household may harvest 10 kilos of rice but instead records consumption of rice on in terms of grams or cups (i.e. *n* cups of rice are consumed for one meal on day *X*). Further, in the price questionnaire (when it exists for a given survey), items are reported on a per-unit basis, the unit corresponding to that which is most commonly found in the market: for apples it may be per kilo whereas for potatoes it may be per sack and for milk it may be per carton or per litre. Nonetheless, as discussed above, if survey data and/or documentation do not include equivalence factors or tables, finding an accurate unit price to assign to the quantity of household own consumption becomes complex and the estimation of the value of consumption will remain a rough rather than precise one. The estimation of precise values also can be further assisted by structuring questionnaires in a way to make the units comparable across sections.

The principal recommendations stemming from the process of price imputations reflect the same ideas summarized in the previous sections. The process of price imputations underscores the importance of dealing with the obstacles emphasized and the need to address these issues in all stages of income aggregate creation.

A.2.c. Additional considerations

Some methodologies for the estimation of aggregate income consider further adjustments and corrections to the data than those described within the RIGA methodology. Two cases worth highlighting are those of the rescaling of self employment income and that of accounting for price variability in the income estimation.

Several analyses of the quality of household income data have demonstrated that misreporting of self employment income is common and that survey instruments should be designed to minimize this misreporting (De Mel et al 2007). Misreporting may be unintentional due to recall error and the lack of accounting systems in the business, but it may be deliberately understated due to household concerns with taxation or also to convey an image of greater poverty in hopes of receiving social benefits that respondents perceive may be linked to the household survey. Attaining reliable self employment income data can be approached through proper questionnaire design (see De Mel et al 2007 for details on various methods), through the coordination and/or integration of different survey instruments, as well as through

alternative or indirect estimation methods such as Germany's summation of consumption expenditure and savings to yield an estimate of income (Wye City Group Handbook 2007). Despite these possibilities, to date the RIGA methodology for calculating self employment income does not present new innovations for addressing underreporting. Household survey data will inevitably be subject to various kinds of errors and therefore income aggregates should be recognized as providing estimates rather than precise measures of household income levels and shares, a point that should be taken into consideration in the interpretation of any analyses and statistics produced from survey data.

Accounting for price variability is another relevant issue that is not always possible to address. Seasonal price fluctuations are not easily controlled for given that data collection for household surveys often is undertaken in a limited period of the year, on which is normally constrained to the months in which households are most accessible (usually outside the rainy season). Although this does not present an issue for the valuation of income and expenditures for which reference periods are the previous twelve months, some bias may be introduced for income sources with a shorter reference period and those for which imputed prices are applied in order to value income, specifically the valuation of own consumption. In this case, prices will be based upon the reference period of sold production and purchased food items. Overcoming this limitation in a clean and straightforward manner would require data collection to be spread throughout the year so that the household survey would naturally account for changes in prices, or perhaps for a price questionnaire to be administered at the community level several times per year in order to be able to track the evolution of market prices over a longer time frame.

Alternatively, though less accurately, questions could be incorporated into survey instruments that ask households about how the price received for their sold production (or on the price paid for purchased food consumption) compares at the time of the survey to different points in time in the twelve months preceding the survey interview. This approach is likely to be more practical with a low marginal cost, but it is subject to recall error and does not effectively collect price data, but a subjective perception of price changes, which could introduce other biases.

Taking these factors into consideration, the RIGA income aggregate estimation does not introduce adjustments for seasonal price variability. Without the adequate data, making such adjustments would require placing many assumptions for each survey's data construction which would possibly be idiosyncratic across surveys and thus limit the methodological consistency of each RIGA income aggregate. The RIGA project does account for regional price differences through its price imputation approach (described in A.2.b. above) in which prices are imputed and assigned at different administrative levels. Other, more conventional adjustments for regional price differences (e.g. regional price deflator) are not applied in the data construction process. They are relevant and recommended for analyses of income levels. To date, however, RIGA income analyses focus on income shares and participation rates and as such have not applied these sorts of additional adjustments.

Table A1: *Share of households owning dwelling (full sample, select surveys[31])*

Country	Survey Year	Share Owning Home
Ecuador	1995	65.3%
Ghana	1992	37.0%
Ghana	1998	41.4%
Guatemala	2000	57.4%
Indonesia	2000	82.7%
Malawi	2004	80.6%
Nepal	1996	93.8%
Nicaragua	2001	77.6%
Nicaragua	1998	78.7%
Nigeria	2004	71.5%
Panama	1997	78.9%
Vietnam	1992	93.6%
Vietnam	1998	94.7%
Mean		*73.3%*

[31] Table reports means for surveys for which this variable was constructed. No other discretion was applied in the selection of surveys for this table.

Table A2: *Dropped households and negative income*

Country	Year of Survey	Share of Households Dropped, Final Outlier Check	Share Households with Negative Income	
			Totincome1	Totincome2
Africa				
Ghana	1992	1.1%	12.9%	6.8%
Ghana	1998	1.8%	11.2%	7.6%
Madagascar	1993-1994	0.8%	2.2%	1.7%
Madagascar	2001	0.6%	0.9%	1.0%
Malawi	2004-2005	0.5%	3.4%	0.6%
Nigeria	2004	2.7%	4.6%	4.1%
Asia				
Bangladesh	2000	0.3%	0.5%	0.6%
Indonesia	1992	0.7%	1.9%	n/a
Indonesia	2000	0.3%	1.1%	n/a
Nepal	1995-1996	1.5%	0.6%	0.7%
Nepal	2003	1.1%	0.4%	0.7%
Pakistan	1991	3.3%	11.9%	11.2%
Pakistan	2001	0.9%	1.3%	5.5%
Vietnam	1992-1993	2.6%	2.1%	2.5%
Vietnam	1997-1998	0.6%	0.7%	0.9%
Vietnam	2002	1.3%	5.9%	6.1%
Eastern Europe				
Albania	2002	1.1%	2.6%	2.3%
Albania	2005	0.1%	0.3%	0.3%
Bulgaria	1995	0.1%	0.8%	0.7%
Bulgaria	2001	0.3%	1.4%	1.3%
Tajikistan	2003	0.3%	0.7%	n/a
Latin America				
Ecuador	1995	1.3%	3.4%	3.3%
Ecuador	1998	1.2%	4.2%	3.8%
Guatemala	2000	0.1%	0.3%	0.1%
Guatemala	2006	0.0%	0.1%	0.1%
Nicaragua	1998	1.2%	3.5%	3.2%
Nicaragua	2001	0.8%	2.3%	2.4%
Panama	1997	0.2%	0.3%	0.1%
Panama	2003	0.2%	0.3%	0.2%

Table A3: *RIGA-L Recoding Guidelines*

Time Reference Variable	Recoded Value
Months per year	12 (the maximum per year).
Weeks per year	52 (the maximum per year).
Weeks per month	4.35 (365.25 days per year divided by 12 months, all of which is divided by 7 days per week – ((365/12)/7) – which rounds to 4.35).
Days per year	365 days per year (the maximum per year) or 312 working days per year (52 weeks multiplied by 6 working days per week. This is used if it is more appropriate for a select survey).
Days per month	31 (the maximum per longest month).
Days per week	7 (the maximum per week or 6 working days per week, if more appropriate for a specific survey).
Hours per day	16 (assuming that an individual can work a maximum of 16 hours in a single day).
Hours per week	84 (assuming that an individual can work a maximum of 12 hours per day for 7 days or 14 hours per days for 6 days, etc.). Note: this implies that it is not possible for an individual to work for the maximum number of hours per day, 16, for more than 6 days.

Table A4: *"Crop1" and "Crop2" Levels and Shares Comparison*

	Participation		Levels (USD, Survey Year)		Shares	
	Crop1	*Crop2*	*Crop1*	*Crop2*	*Crop1*	*Crop2*
			Africa			
Ghana 1992	87.0%	87.1%	70	632	44.5%	66.2%
Ghana 1998	87.2%	87.8%	142	409	41.7%	55.0%
Madagascar 1993	92.8%	93.4%	228	183	37.0%	57.3%
Madagascar 2001	81.5%	83.6%	211	210	56.9%	59.5%
Malawi 2004	83.2%	96.3%	28	194	17.5%	56.1%
Nigeria 2004	87.5%	88.7%	424	634	71.7%	73.5%
			Asia			
Bangladesh 2000	61.5%	81.6%	97	94	15.8%	15.5%
Nepal 1996	88.8%	92.6%	176	114	37.7%	32.1%
Nepal 2003	89.4%	93.4%	188	93	29.4%	20.3%
Pakistan 1991	44.4%	60.3%	1,196	1,352	17.4%	31.5%
Vietnam 1992*	92.1%	95.1%	145	123	48.3%	50.9%
Vietnam 1998	92.7%	97.8%	319	300	42.4%	41.5%
			Eastern Europe			
Albania 2002	79.6%	91.8%	135	327	15.5%	15.1%
Albania 2005	93.4%	94.7%	888	600	24.5%	17.2%
Bulgaria 1995	45.6%	89.3%	28	275	2.6%	19.8%
Bulgaria 2001	62.5%	68.3%	685	57	19.5%	3.9%
			Latin America			
Ecuador 1995	68.8%	73.5%	677	583	18.4%	9.0%
Ecuador 1998	63.5%	68.3%	240	430	35.8%	22.1%
Guatemala 2000	69.9%	87.8%	189	367	19.2%	27.6%
Nicaragua 1998	73.7%	71.1%	155	231	19.0%	22.9%
Nicaragua 2001	83.6%	84.8%	216	267	-9.0%	20.6%
Panama 1997	45.8%	87.5%	402	232	16.4%	15.5%
Panama 2003	48.6%	78.4%	80	263	5.5%	15.8%

Annex 2: Stata Syntax

Outlier Checks.

The RIGA project defines outliers as values greater than or less than three standard deviations from the median when checking for extreme values by some relevant subgroup. The RIGA-CLSP[32] team developed a Stata command (ado-type) to facilitate the checking of outliers. The command, called "imputeout", has a standard syntax structured as follows:

```
imputeout varlist1 [if] [weight] , bylist(varlist2) range(#)
```

where *varlist1* is the list of variables to check for outliers, *varlist2* is the sorting category by which *varlist1* will be checked for outliers (not an option), and "range(#)" identifies the number of standard deviations according to which an outlier will be defined (the default is 3). Missing values and zero values are automatically excluded from the outlier check when using this command. The command and corresponding help file are available from the team[33].

For a variable "exvar" and a bylist variable of "byvar", the above syntax is equivalent to:

```
(1) g exvarm   = .
(2) g exvarimp = exvar

(3) bys byvar: egen exvarme = median(exvar) if exvar != 0
(4) by  byvar: egen exvarsd = sd(exvar)     if exvar != 0
(5) replace exvarsd = 0 if exvarsd == .
(6) replace  exvarm  =  exvarme  if  !(exvar  >=  (exvarme  -  3*exvarsd)  &  exvar  <=
           (exvarme + 3*exvarsd)) & exvarm == . & exvar != . & exvar != 0
(7) replace exvarimp = exvarm if exvarm != .
```

The outlier check procedure can be summarized as follows. When checking for outliers, each variable that is identified as an outlier is flagged. A flag variable (*exvarm*, above) is constructed, named the same as the variable being checked for outliers, plus an "m"(e.g. the flag for "*exvar*" is "*exvarm*"). In our sample Stata syntax (above), the variable **exvarme** contains the values to be used as replacements for outlier values when outliers are identified. These replacement values are the median value of **exvar** for each subgroup identified by **byvar**. The flag variable **exvarm** is either a missing value or the value identified by **exvarme** if **exvar** was an outlier. The dataset also includes the imputed variable, **exvarimp** in this example. The variable **exvarimp** contains the original values when no imputations are necessary and the imputed values when an outlier is identified.

In order to flag and impute outliers in the variable **exvar,** the variables **exvarm** and the variable **exvarimp** are created in lines (1) and (2). At this stage, the variable **exvarm** consists only of missing values and the variable **exvarimp** is equal to the original variable, **exvar**. Lines (3) and (4) calculate the median and standard deviation of **exvar** by the relevant subgroup category (**byvar**). Line (6) flags any outliers in **exvar** by replacing in **exvarm** the

[32] CLSP stands for Comparative Living Standards Project. This project is based at the World Bank with the objective of creating an online database of comparative living standards indicators. The RIGA project collaborated with CLSP in the construction of the income aggregates, which will be incorporated into the online database. More information on CLSP can be found on http://go.worldbank.org/YIOLNP2T40.

[33] A version of the routine that applies the logarithmic transformation of the variable before imputing the outliers is also available ("imputeoutlog"), but is only used in the construction of the RIGA-L database.

missing value with the median of **exp**. In the last line of the program (7), the original value (in **exvarimp**) is replaced by the median value (in **exvarm**) if an outlier is identified.

As mentioned earlier in this paper, there are two amendments to the outlier check procedure in the RIGA-L database: the use of logs and weights to flag outliers. These amendments are explained below:

Using the coding above for the RIGA database, step (3) and step (4) are replaced by the following codes:

```
(*)  gen lnexvar = ln(exvar)
(**) sum lnexvar if byvar==`i' [aw=weight],detail
(3)         scalar lnexvarmed=r(p50)
(4)         scalar lnexvarsd=r(sd)
```

The new step (*) creates the log of **exvar**. Step (**) summarizes the weighted log of the **exvar** variable by the sorting variable **byvar**. This procedure is later repeated for every **byvar** category "**i**". Then, similarly as the original coding, steps (3) and (4) calculate the median and the standard deviations but of **lnexvar** by the relevant subgroup category **byvar**.
We include a new code (5) of the original coding to create an additional flag variable sorted by the **byvar** categories ***i*** (**outlier_exvar*i***), following the rule of 3 standard deviations lower or higher than the median:

```
(5) gen outlier_exvari =(abs(lnexvar - lnexvarmed)>(3*lnexvarsd)) & byvar ==`i'
```

Then, in step (6) we proceed to replace the flag variable **exvarm** with the exponent of ln**exvar**. The procedure is repeated for all the categories of **byvar**.

```
(6) replace exvarm = exp(lnexvar) if outlier_exvari ==1
```

Finally, in step (7) we replace **exvarimp** with the exponent of **lnexvarmed**. The operation is repeated for all the categories of **byvar.**

```
(7) replace exvarimp = exp(lnexvarmed) if outlier_exvari ==1 & exvar !=. & exvar
           !=0
```

Imputeout in RIGA-L was not used for outlier checks. This is because the procedure in RIGA-L is different and also because at the time when this project was initiated, the RIGA-CLSP team had not yet created the command.

Price Imputations.

The following lines of Stata syntax demonstrate the approach used to estimate a standard set of prices by which to value income and expenditures that were reported in quantities rather than monetary units. In the example, line 1 demonstrates the generation of the price variable when the dataset is at the most disaggregated level. Lines 2 through 21 indicate how the median prices are obtained at different geographic/administrative levels by saving datasets at the cluster, province, district, urban and unit levels. The prices are then merged into the dataset containing the variables for which values will be imputed and assigned prices as demonstrated in lines 22 through 38. The same procedure is used for the estimation of prices from the food expenditures module.

```
* Generate overall price variable
(1) gen price = salesvalue / soldquantity

* Save median prices at different levels
(2) preserve
(3) collapse (median) pricedata1=price , by(crop unit urban district cluster)
(4) sort crop unit urban district cluster
(5) save $OUT\prodprice_clust.dta, replace
(6) restore
(7) preserve
(8) collapse (median) pricedata2=price , by(crop unit urban district)
(9) sort crop unit urban district
(10) save $OUT\prodprice_district.dta, replace
(11) restore
(12) preserve
(13) collapse (median) pricedata3=price , by(crop unit urban)
(14) sort crop unit urban
(15) save $OUT\prodprice_urban.dta, replace
(16) restore
(17) preserve
(18) collapse (median) pricedata4=price , by(crop unit)
(19) sort crop unit
(20) save $OUT\prodprice_unit.dta, replace
(21) restore

***
* Merge in estimated prices
(22) sort crop unit urban district cluster
(23) merge crop unit urban district cluster using $OUT\prodprice_clust.dta
(24) drop if _merge==2
(25) sort crop unit urban district
(26) merge crop unit urban district using $OUT\prodprice_ district.dta
(27) drop if _merge==2
(28) sort crop unit urban
(29) merge crop unit urban using $OUT\prodprice_ urban.dta
(30) drop if _merge==2
(31) sort crop unit
(32) merge crop unit using $OUT\prodprice_ unit.dta
(33) drop if _merge==2

* Assign prices
(34) generate value = .
(35) replace value = quantity * pricedata1 if value==.
(36) replace value = quantity * pricedata2 if value==.
(37) replace value = quantity * pricedata3 if value==.
(38) replace value = quantity * pricedata4 if value==.
```

Annex 3: Raw Material Expenditures in Non-farm Enterprises

The cost of raw materials reported among the expenses of non-farm enterprises, as collected in most surveys, generally refers to the previous month. However, the initial inclusion of these expenditures, when annualized using our standard approach for non-farm enterprises (multiplying the monthly expenditure by the number of months the business was in operation and then weighting by the share of the business owned by the household[34]), was found to lead, in some cases, to over-stated costs and overall negative net self employment income.

Studying this category of expenditures in greater detail yielded the conclusion that due to the large magnitude of these expenditures in certain surveys, the values reported for raw materials expenditures may not necessarily represent regular, monthly spending, but rather bulky purchases that generally can be assumed to take place less frequently. In some cases, raw material expenditures are naturally excluded from the estimation of self employment income since the survey questionnaire identifies the items which characterize this expenditure as non-frequent/non-recurrent costs. For example, the Guatemala ENCOVI 2000 and Panama ECV 2003 surveys group raw materials with capital investment purchases rather than with the set of regular monthly costs while the Pakistan 2001 survey classifies raw materials as a "Special Operating Expense." Some other surveys, such as Panama 1997, do not attach a special label but do separate the expenditure from the list of regular monthly expenditures, indicating that raw materials are characterized differently from other recurrent costs. As survey questionnaires are generally written with consideration to the local-context, clearly there may be a degree of country-specificity about the nature of raw material costs: the Pakistan 1991 survey highlights the ambiguity of this expenditure category by allowing raw material expenditures to be reported as both frequent and infrequent expenses while the Ecuador 1995 and Nicaragua 2001 survey lists them with capital investments (infrequent expense) and also as an autonomous question regarding raw materials purchases (frequent expense).

For the following surveys we were obligated to impose a different approach, or to exclude, raw materials from the estimation of net self employment income: Albania 2002, Vietnam 1992 and Vietnam 1998. For Albania 2002 we implemented a modified outlier check, using a stricter definition of outliers and taking into consideration the expenditure distribution of each cost category. For the Vietnam 1992 and Vietnam 1998 surveys we excluded raw material costs after determining that even modifications to the standard outlier check approach were insufficient to deal with the extreme values, leading us to assume that raw material expenditures may encompass investment-type purchases.

It should also be noted that even with the more explicit reporting in the Guatemala 2000 and Panama 2003 surveys, due to data limitations on the non-farm enterprise module of the survey (income information was not asked for alongside the expenditure questions), raw material expenses were not explicitly considered since net self employment income was instead estimated from the employment module of the questionnaire. In consequence, so to preserve over-time comparability, the Panama 1997 income aggregate also estimates self employment income from the wage employment module of the questionnaire.

[34] The frequency with which the expense was made is also factored in, if that information was included in the survey.