A Feasibility Study on Calibration for Selected South-East Asia Household Surveys

Short Term Consultancy Final Report

Diego Zardetto

November 2015

Contents

Executive Summary

Modern large-scale household surveys are generally expected to provide high quality estimates of population parameters. However, strong signals of bias have occasionally been detected for some South-East Asia countries. For instance, the World Bank's Household Survey Development Team found significant discrepancies between survey-based estimates of age-sex and household size distributions and the corresponding Census counts for Vietnam, Thailand and the Philippines.

A feasibility study has been carried out to investigate whether this issue could be solved through a preliminary calibration procedure.

Calibration is a systematic and mathematically rigorous method to achieve higher quality estimates by incorporating auxiliary information on the target population, available from external sources, into the survey estimation infrastructure [1]. From an algorithmic standpoint, a calibration procedure minimally adjusts the survey weights in such a way that the resulting calibration estimates exactly match selected known population totals [4].

Three Proofs of Concept (POC) have been carried out, adopting as empirical test bed the following household surveys:

- 2008 Vietnam Household Living Standards Survey – VHLSS 2008
- 2010 Thailand Household Socio-Economic Survey – SES 2010
- 2006 Philippines Family Income and Expenditure Survey – FIES 2006

For each survey, calibration constraints have been imposed on known population totals derived from the closest Population and Housing Census round:

- 2009 Vietnam Census
- 2010 Thailand Census
- 2007 Philippines Census

To tackle the feasibility study, the ReGenesees system was used: an open source software for design-based and model-assisted analysis of complex sample surveys [10], based on R [3].

All POCs were successful: calibration algorithms were run without noticeable technical problems, and exact convergence was always obtained. Overall, the study showed that *it is technically feasible to integrate a calibration procedure in the production workflow of all the household surveys taken into account*.

Beyond the feasibility study, two possible implementation lines can be envisioned:

1) A calibration procedure could be executed directly by the National Statistical Institute (NSI) in charge of the household survey, as a process step to be routinely performed preliminary to estimation. Of course, enabling NSIs to adopt calibration estimators would require appropriate capacity building actions.
2) A calibration procedure could be executed ex-post for analysis purposes, i.e. after data dissemination and outside the involved NSIs, in order to increase the quality of the estimates derived from the surveys. For instance, the World Bank could manage the calibration procedure on its own, and integrate the obtained calibration weights into its microdata repositories.

In both cases, using calibration weights for estimation would be straightforward, while estimating sampling errors would require specialized software, like the ReGenesees system.

1. Motivation of the Study

The World Bank's Household Survey Development Team noticed that in some countries (mainly in South-East Asia), the distribution of the population by age and sex and the distribution of households by size differ very significantly between large-scale socio-economic sample surveys and the Population Census (see Figure 1 and Table 1 for an example concerning Vietnam). As these surveys are supposed to be (at least) nationally-representative, discrepancies of this magnitude are not expected.

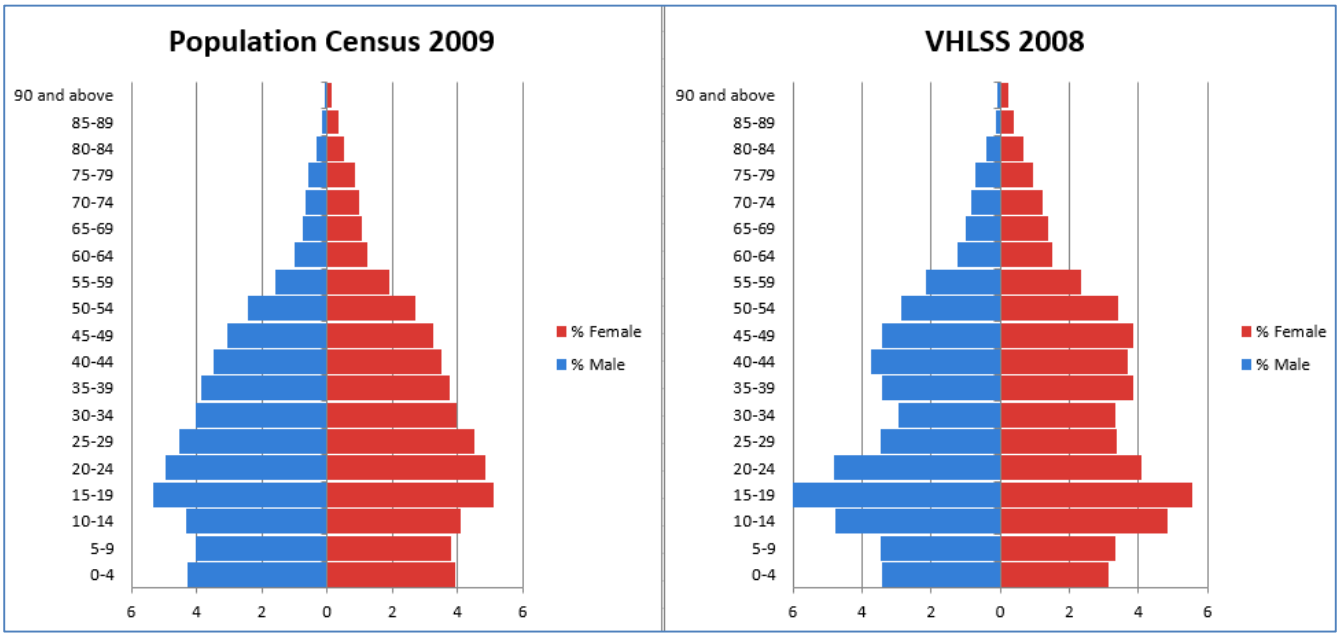

Figure 1: Age-Sex Pyramids of Vietnam Population – Counts from Census 2009 (left) vs Estimates from VHLSS 2008 (right)

	Household Count		Household Percentage (%)	
Household Size	**Census 2009**	**VHLSS 2008**	**Census 2009**	**VHLSS 2008**
1 person	1,625,592	915,340	7.2%	4.4%
2 persons	3,216,733	2,366,551	14.3%	11.3%
3 persons	4,684,820	3,646,577	20.9%	17.4%
4 persons	6,432,702	6,553,118	28.7%	31.3%
5 persons	3,397,237	3,985,423	15.1%	19.0%
6 persons	1,864,916	1,990,573	8.3%	9.5%
7 persons	611,496	879,844	2.7%	4.2%
8 persons	308,380	337,449	1.4%	1.6%
9 persons and above	302,446	282,411	1.3%	1.3%
Total	22,444,322	20,957,286	100.0%	100.0%

Table 1: Absolute and Percentage Frequency Distributions of Households by Size – Vietnam Census 2009 vs VHLSS 2008

Large deviations of survey estimates from the corresponding "true" population parameter (as measured by the Census) are a serious concern, since they likely hint at systematic flaws affecting the survey. Furthermore, there is not any guarantee that such flaws are actually confined to the estimates in which they happen to show up manifestly. Indeed, survey bias tends to "propagate" across estimates, owing to the correlation structure of the involved variables. Therefore, the impact of the aforementioned "large discrepancies" on estimates of poverty and inequality should be carefully investigated.

Calibration algorithms adjust the survey weights in such a way that the resulting calibration estimates exactly match the corresponding known population totals (as derived from reliable sources outside the survey) [4]. Therefore – by construction – a successful calibration procedure automatically *removes* any bias possibly affecting the estimates of the auxiliary variable totals. As a remarkable side effect, it is generally agreed that calibration at least *decreases* the bias affecting interest variables that are correlated to the auxiliary ones [5].

The line of reasoning illustrated above triggered a research project on calibration, which is currently being conducted by the Household Survey Development Team of the World Bank's Development Data Group, with the methodological and technical assistance ensured by a short term consultant. As a first mandatory step, a technical feasibility study was committed to the consultant. The aim of the study, to be carried out on a reasoned selection of South-East Asia household surveys, was threefold:

(1) Investigate whether the observed large discrepancies between survey-based estimates and Census counts are artifacts of random sampling, or rather genuine symptoms of *bias*.

(2) Verify whether a calibration procedure can actually succeed in making both the age-sex pyramids and the distribution of households by size simultaneously *consistent* with the Census data.

(3) In case the task at point (2) is feasible, assess the impact of the aforementioned large discrepancies on key poverty and inequality indicators, by comparing their Horvitz-Thompson (HT) and calibration (CAL) estimates.

As sketched in the Executive Summary, overall the outcome of the feasibility study was positive. This result will very likely pave the way to further actions to be undertaken within the project.

The rest of this document provides a concise description of the feasibility study on calibration. The 'Terms of Reference' (TOR) of the consultancy are reported in Annex 1:. For further information on the work (detailed analyses, plots, tables, and the commented R code describing how the ReGenesees system was used), the interested reader is referred to the World Bank's Box folder dedicated to the consultancy.

2. Description of the Proofs of Concept

As anticipated in the Executive Summary, three Proofs of Concept (POCs) have been set up for the feasibility study on calibration. These POCs involve the following household surveys:

- 2008 Vietnam Household Living Standards Survey – VHLSS 2008
- 2010 Thailand Household Socio-Economic Survey – SES 2010
- 2006 Philippines Family Income and Expenditure Survey – FIES 2006

For the sake of clarity, from now on, we will identify each POC through the name of the corresponding country (e.g. we will refer to the 'Vietnam POC').

All the large-scale surveys listed above suffer the "large discrepancies" issue discussed in Section 1, with very similar manifestations to those testified for Vietnam by Figure 1 and Table 1.

For each survey, calibration constraints have been imposed on population totals derived from the closest round of the Population and Housing Census:

- 2009 Vietnam Census
- 2010 Thailand Census
- 2007 Philippines Census

For each POC, survey microdata and Census aggregates were provided by the World Bank Development Data Group (DECDG). Survey datasets were complemented by sampling design metadata, i.e. information on the way survey samples were drawn from the corresponding list frames.

Concerning the POCs, few additional observations are in order:

(i) With the exception of the Thailand POC, survey and Census reference years do not coincide. Even though such modest time lags cannot jeopardize our feasibility study, they could severely impair estimation in production settings. Real-world calibration procedures of cross-sectional data should always involve population totals referred to the right point in time.

(ii) From a purely computational standpoint, sampling design metadata do not play any material role in calibration, nor in calculating calibrated estimates. Anyway, their knowledge is mandatory for sampling variance estimation, hence for a proper assessment of the uncertainty in survey estimates. This, in turn, is obviously relevant to the first objective of our study (see Section 1).

(iii) To limit statistical disclosure risks, the Thailand NSI usually does not disseminate complete information about the survey design. This explains why identifiers of Primary Sampling Units (PSUs) were not available inside the SES 2010 dataset: simply PSUs variables had not been shared with the World Bank. This circumstance makes the Thailand POC special, in that we had to adopt a "pseudo" sampling design for variance estimation purposes. As a consequence, estimated confidence intervals are not entirely reliable for the Thailand POC.

All the three objectives of the feasibility study outlined in Section 1 have been pursued for all three POCs. This means that each POC was structured into three different tasks. For later convenience, we will identify these tasks as follows:

(1) Bias Analysis

(2) Calibration

(3) Impact on Poverty and Inequality Estimates

The following sections of the report will be devoted to illustrating these tasks. There, we will not try to provide a complete description of how we tackled a given task for each POC. Instead, we will abstract those features that are common to all POCs, and possibly focus on just a single survey for presentation convenience.

3. Bias Analysis

The first task we addressed for each POC was to analyze the large discrepancies between survey-based estimates and Census counts that had been pointed out by the Household Survey Development Team of the World Bank. The goal of this study was to discriminate between two alternative hypotheses:

H_0: The observed discrepancies occurred by chance, i.e. they must be accepted as a mere random sampling effect.

H_1: Survey estimates have been derived from estimators affected by significant bias (whatever the cause could be).

In order to test whether the null hypothesis H_0 must be rejected or not at a given significance level α, we had to:

(i) Estimate confidence intervals for the relevant point estimates at confidence level CL = $(1 - \alpha)$.

(ii) Verify if the estimated confidence intervals cover or not the true population parameters, namely the corresponding Census figures.

Now, computing reliable confidence intervals obviously demands sampling variance estimation, and ReGenesees was used to this end. Indeed, the ReGenesees system can handle a variety of complex sampling designs and can provide estimates and sampling errors for a wide range of estimators, including very complex ones (see Annex 2: ReGenesees in a Nutshell).
Enabling ReGenesees to compute confidence intervals only required us to bind survey data with the appropriate sampling design metadata. Afterwards, we simply asked the system to compute the estimates we were interested in, along with the relative confidence intervals.

In what follows, we will keep using the Vietnam POC as running example.

Table 2 reports HT estimates of the joint percentage distribution of sex and age in five-year classes, together with their 95% confidence intervals and percent coefficient of variation (CV%). HT estimates are also contrasted with their Census counterparts. As highlighted by the surrounding blue rectangles, only 2 estimated confidence intervals out of 32 happen to cover the associated Census percentages. This means that, at a significance level $\alpha = (1 - CL) = 5\%$, we would be compelled to reject H_0. Even repeating the same analysis with a more stringent significance level $\alpha = 1\%$, which entails using a CL = $(1 - \alpha) = 99\%$ confidence level, only 4 Census percentages out of 32 would be covered by the estimated confidence intervals: still a very strong evidence against H_0. As a result, we must conclude that the alternative hypothesis H_1 is the favorite one: even taking into account sampling variability, *the discrepancies observed for age-sex distributions are true symptoms of bias in VHLSS 2008 data*.

The same analysis, this time addressing the distribution of households by size, is graphically illustrated in Figure 2. Here black dots with horizontal error bars represent survey estimates and 95% confidence intervals, whereas Census counts are identified by red squares. Only 4 Census figures out of 9 happen to be covered. If the confidence level is increased to 99%, an identical outcome is obtained. Again, we are led to the conclusion that H_0 cannot hold. With the likely exception of the mode (4 persons) and the right tail of the distribution (8 persons and above), *the discrepancies observed for the distributions of households by size must be understood as true symptoms of bias in VHLSS 2008 data*. Moreover, the VHLSS 2008 seems to be consistently underestimating the amount of small-to-medium sized households, while overestimating the number of medium-to-large ones.

Sex	Age	Census 2009	VHLSS 2008	CI.l(95%)	CI.u(95%)	CV%
Female	[0,5)	3.93%	3.13%	2.93%	3.33%	3.3%
Female	[5,10)	3.79%	3.31%	3.12%	3.51%	3.0%
Female	[10,15)	4.10%	4.82%	4.58%	5.06%	2.5%
Female	[15,20)	5.11%	5.53%	5.29%	5.78%	2.3%
Female	[20,25)	4.87%	4.10%	3.88%	4.32%	2.8%
Female	[25,30)	4.53%	3.36%	3.14%	3.57%	3.3%
Female	[30,35)	3.97%	3.34%	3.14%	3.54%	3.1%
Female	[35,40)	3.77%	3.85%	3.63%	4.06%	2.8%
Female	[40,45)	3.49%	3.70%	3.51%	3.90%	2.7%
Female	[45,50)	3.27%	3.84%	3.64%	4.04%	2.7%
Female	[50,55)	2.71%	3.42%	3.22%	3.63%	3.1%
Female	[55,60)	1.89%	2.33%	2.18%	2.49%	3.4%
Female	[60,65)	1.25%	1.49%	1.35%	1.63%	4.6%
Female	[65,70)	1.05%	1.36%	1.24%	1.49%	4.6%
Female	[70,75)	0.98%	1.22%	1.10%	1.35%	5.3%
Female	[75,Inf]	1.89%	2.22%	2.06%	2.38%	3.7%
Male	[0,5)	4.27%	3.43%	3.24%	3.62%	2.9%
Male	[5,10)	4.03%	3.48%	3.26%	3.70%	3.2%
Male	[10,15)	4.34%	4.79%	4.54%	5.04%	2.7%
Male	[15,20)	5.33%	6.09%	5.82%	6.36%	2.2%
Male	[20,25)	4.95%	4.80%	4.55%	5.05%	2.6%
Male	[25,30)	4.55%	3.45%	3.25%	3.65%	3.0%
Male	[30,35)	4.03%	2.95%	2.77%	3.14%	3.2%
Male	[35,40)	3.84%	3.42%	3.24%	3.61%	2.7%
Male	[40,45)	3.46%	3.72%	3.52%	3.93%	2.8%
Male	[45,50)	3.08%	3.42%	3.23%	3.61%	2.8%
Male	[50,55)	2.43%	2.86%	2.68%	3.03%	3.1%
Male	[55,60)	1.59%	2.13%	1.97%	2.29%	3.9%
Male	[60,65)	1.00%	1.22%	1.10%	1.34%	5.0%
Male	[65,70)	0.76%	0.99%	0.89%	1.10%	5.5%
Male	[70,75)	0.66%	0.83%	0.73%	0.94%	6.1%
Male	[75,Inf]	1.08%	1.36%	1.23%	1.48%	4.7%

Table 2: Joint Sex-Age Percentage Frequency Distribution – Vietnam Census 2009 vs VHLSS 2008 HT Estimates

Bias analysis for the remaining POCs gave analogous results. In particular, evidence of bias for Philippines FIES 2006 turns out to be at least as strong as for the Vietnam POC. Moreover, also Philippines data show a consistent tendency to underestimate the amount of small-to-medium sized households.

On the contrary, our bias analysis cannot be deemed conclusive for Thailand SES 2010. This is because estimated confidence intervals are *not* entirely reliable for the Thailand POC, owing to lacking PSU identifiers (recall observation (iii) in Section 2).

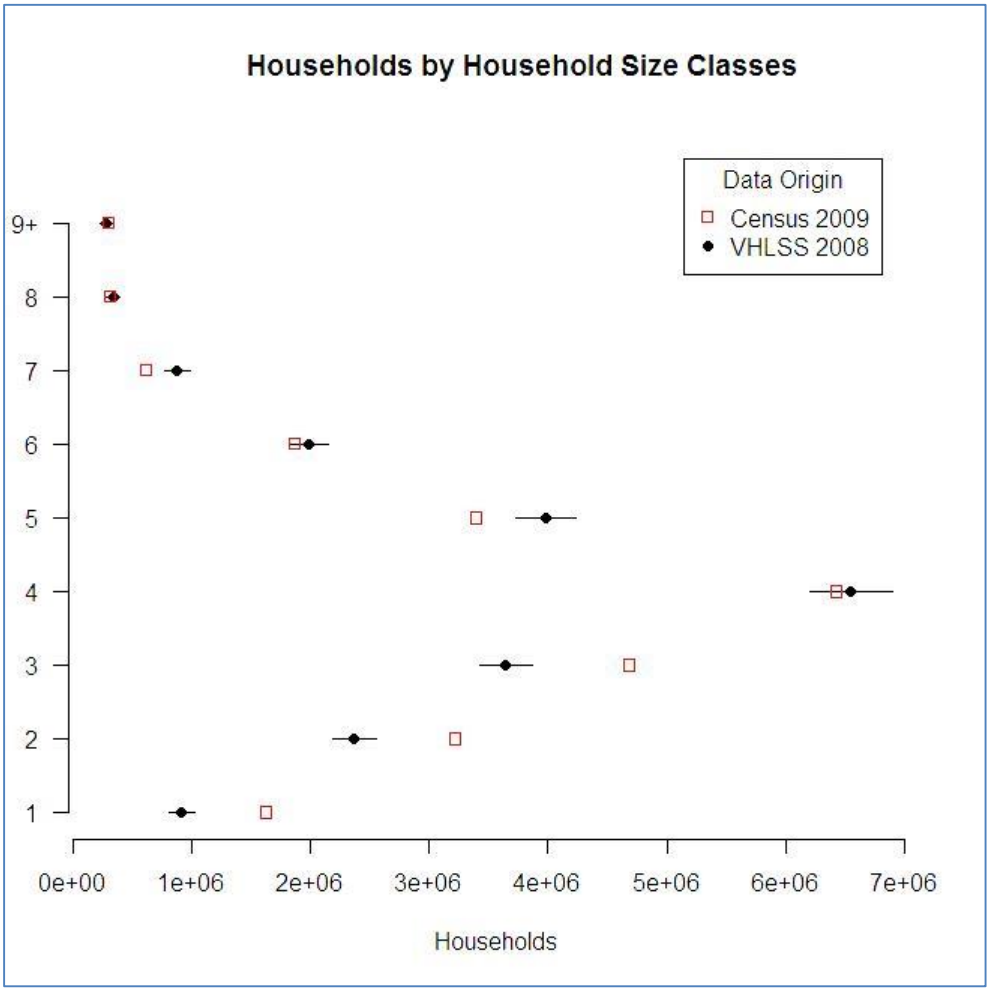

Figure 2: Absolute Frequency Distributions of Households by Size – Vietnam Census 2009 vs VHLSS 2008 HT Estimates

In particular, we cannot exclude that for the Thailand POC our confidence intervals are *tighter* than they should be. This can be understood as follows.

The actual sampling design of Thailand SES 2010 is a stratified two-stage cluster sampling, with municipal areas/villages as PUSs and households as secondary sampling units (SSUs). The "pseudo" sampling design we were forced to adopt in our analysis is, instead, a stratified one-stage sampling design, with real-world SSUs playing the role of first stage clusters. Since the sample variance of estimated totals between *actual* PSUs (i.e. sampled municipal areas/villages, which are unfortunately unknown to us) is expected to be *larger* than the sample variance of estimated totals between "*pseudo"* PSUs (i.e. households), our "pseudo" sampling design can very likely lead to sampling variance *underestimation*, and hence to too tight confidence intervals. Of course, this would surreptitiously decrease the *effective* statistical significance of our tests below the *nominal* level α and potentially undermine our bias study.

We end this Section by stressing that our analysis cannot shed any light on the likely causes of the bias we detected. One can speculate on many possible systematic flaws (e.g. master sample obsolescence or imperfections, selective non-response, influent measurement or processing errors, etc.), but a dedicated study would be required to understand their actual role and relevance.

4. Calibration

The calibration task constitutes the core of our feasibility study. The goal of this task was to investigate whether a calibration procedure could overcome the "large discrepancies" issue discussed in Section 1. More explicitly, our aim was to verify if a calibration algorithm could succeed in adjusting the survey weights so that *both* the estimated age-sex pyramids *and* the estimated distribution of households by size become *simultaneously* consistent with the corresponding Census aggregates.

For each POC, according to the TOR (see Annex 1:), we searched for calibration weights **w** meeting the requirements of being:

(i) as close as possible to the direct weights, **d**

(ii) positive

Moreover, for each POC, we produced two *different* sets of output weights fulfilling both conditions (i) and (ii), with the following distinctive features:

$\mathbf{w}^{CAL1}$: Individuals within each household share a *common* calibration weight, i.e. weights are adjusted at *household level*.

$\mathbf{w}^{CAL2}$: Individuals within the same household may have *different* calibration weights, i.e. weights are adjusted at *individual level*.

For the sake of conciseness, from now on, we will refer to $\mathbf{w}^{CAL1}$ and $\mathbf{w}^{CAL2}$ as *household-level* and *individual-level* calibration weights respectively.

From a mathematical point of view, calibration is a constrained optimization problem. Calibration weights are obtained by minimizing an appropriate distance function from direct weights, subject to *calibration constraints* ensuring that the calibrated estimates of the totals of a set of auxiliary variables exactly match the corresponding known population totals.

This means that condition (i) is inherently satisfied by any calibration algorithm, whereas the positivity condition (ii) is actually an additional constraint to be fulfilled, though not a *calibration* one. The resulting constrained optimization problem is usually known as *range-restricted* calibration. Range-restricted calibration algorithms are often employed to prevent output weights from becoming either negative or exceedingly high. Negative calibration weights might lead to pathological results, such as negative estimates for strictly positive population parameters, even though this risk is likely to materialize only in small domain estimation. Extreme weights, on the other side, might determine unstable estimates and artificially inflate sampling error estimates.

Similarly, ordinary calibration methods do not automatically produce identical calibration weights across members of the same household. Finding household-level calibration weights (like $\mathbf{w}^{CAL1}$) is effectively a harder task than finding individual-level solutions (like $\mathbf{w}^{CAL2}$), and requires specialized algorithms: these are commonly referred to as *cluster-level* calibration algorithms. The rationale for tackling this harder problem is threefold and can be explained as follows:

(1) Practitioners in the official statistics field almost invariably tend to prefer calibration weights that *preserve* notable features of the direct weights. Household surveys typically adopt multi-stage sampling designs with households playing the role of ultimate clusters: all individuals belonging to each sampled household are eventually surveyed. Thus, members of a given household share the same inclusion probability, which – in turn – equals the household inclusion probability. As a consequence, (neglecting

possible non-response effects) *direct* weights are inherently constant within each household, and the same property is perceived as desirable also for calibration weights.

(2) Estimation of parameters concerning the population of households dictates that a *single* weight is attached to each household. If individuals in the same household share a common calibration weight, then calibration estimators of household-level parameters are straightforward. Otherwise, a method must be devised to synthesize individual weights to yield a single household weight.

(3) Household-level calibration can be advocated for *statistical efficiency* considerations. In fact, even if individual-level calibration weights are entirely legal from a methodological standpoint, they may sometimes exhibit higher variability than household-level calibration weights. There is solid empirical evidence that higher weights variability tend to translate into less precise estimates [8], in particular for parameters related to the population of households [2][9].

Coming back to the calibration POC, the TOR left us with two degrees of freedom that are worth mentioning:

(a) The detailed structure of the known totals.

(b) The distance function to be minimized.

Of course, we were in a position to use only population totals that were actually available inside the Census aggregates provided by the World Bank. Moreover, to meet our calibration goal, we were obviously compelled to include as *mandatory benchmarks* both the joint age-sex distribution of individuals and the distribution of households by size. However, we were still free to choose a more fine-grained and larger set of available population totals, provided the mandatory benchmarks could be deduced from that *superset*. We strove to exploit this freedom to add further auxiliary variables that would later benefit poverty estimates. One such variable is *rural/urban* status, thanks to its correlation to poverty and inequality indicators. Recall that the ability of calibration to reduce the sampling variance of estimators (and, at least partially, to soften their bias) crucially depends on how good the auxiliary variables are at "predicting" the interest variables.

As is well known, under mild conditions on the involved distance functions, all calibration estimators are asymptotically equivalent to the generalized regression estimator (GREG) [1][4][6][7]. However, for finite samples, different distance functions will generally determine different calibration weights. Since asymptotic theory does not offer any clue on the distance to be preferred, practitioners often first compare calibration weights arising from different distances, then select those weights that exhibit lowest variability. Again, the justification for this pragmatic rule is that weights with smaller variability usually lead to more efficient calibration estimators.

To tackle the calibration POCs in practice, we used again the ReGenesees system. Indeed, as shown in Annex 2: ReGenesees in a Nutshell, ReGenesees can easily handle all the technical requirements of our calibration task (e.g. benchmarking simultaneously to auxiliary information on individuals and households, range-restricted calibration, cluster-level weights adjustment, different distance functions).
Moreover, ReGenesees makes it very simple to solve even very complex calibration problems, as it does not require any specific data preparation effort. Indeed, the system allows the specification of *calibration models* in symbolic way, using R model formulae. Driven by a calibration model formula, ReGenesees automatically and transparently generates the right values and formats for the auxiliary variables at the sample level, and assists the user in defining and calculating the population totals corresponding to the generated auxiliary variables.

For each POC, we carried out several calibration experiments, exploring different choices of auxiliary variables, distance functions, and range restrictions. For each explored combination, we produced both household-level ($\mathbf{w}^{CAL1}$) and individual-level ($\mathbf{w}^{CAL2}$) calibration weights. Finally, we analyzed all the obtained sets of output

weights and selected the best performing ones: typically those with lowest variance and/or smaller deviations from the direct weights. In what follows we will sketch some of the considerations that led us to identify the optimal calibration weights we eventually delivered to the World Bank.

Table 3 summarizes selected sample variability measures related to different sets of calibration weights that we obtained for each POC. For comparison, the same variability measures are also reported for the direct weights. For each measure, bold figures identify the set of calibration weights with lowest variability.

POC	Weights Type	Distance Function	Standard Deviation	Interquartile Range	Median Absolute Deviation
Vietnam	Calibration *household-level* ($\mathbf{w}^{CAL1}$)	Linear	1932.6	1776.4	1234.3
		Raking	**1930.1**	**1597.3**	1100.5
		Logit	1935.9	1649.1	**1080.4**
	Calibration *individual-level* ($\mathbf{w}^{CAL2}$)	Linear	**1404.8**	1281.9	912.5
		Raking	1408.6	1241.2	882.2
		Logit	1413.5	**1226.0**	**867.6**
	Direct (**d**)	–	1017.5	995.0	732.4
Thailand	Calibration *household-level* ($\mathbf{w}^{CAL1}$)	Linear	**591.4**	457.1	276.2
		Raking	592.4	**453.9**	**274.5**
		Logit	592.1	454.2	276.7
	Calibration *individual-level* ($\mathbf{w}^{CAL2}$)	Linear	562.0	446.3	273.9
		Raking	**541.5**	447.9	**271.0**
		Logit	562.4	**445.8**	273.9
	Direct (**d**)	–	453.3	504.1	294.7
Philippines	Calibration *household-level* ($\mathbf{w}^{CAL1}$)	Linear	**220.9**	276.4	184.8
		Raking	221.0	**272.2**	**182.1**
		Logit	222.1	276.0	182.8
	Calibration *individual-level* ($\mathbf{w}^{CAL2}$)	Linear	**176.6**	214.3	155.8
		Raking	176.7	213.3	155.0
		Logit	177.2	**213.1**	**154.8**
	Direct (**d**)	–	143.3	181.6	136.3

Table 3: Sample Variability Measures for Different Weights – All POCs

For the Thailand POC, taking into account all the variability measures simultaneously, *raking* weights seems to perform (slightly) better than the others, for both household-level and individual-level calibration. The same seems to happen again for Vietnam and Philippines, even though *logit* calibration weights could be an equally defensible choice, especially when individual-level calibration is concerned.

Table 4 summarizes sample distributions of different sets of achieved calibration weights, for each POC. For comparison, the same summaries are also provided for the direct weights. In addition, the obtained range of *g-weights*, i.e. the ratios between calibrated and direct weights ($\mathbf{g}_k = \mathbf{w}_k / \mathbf{d}_k$), is reported.

Overall, the sample distribution of calibration weights is only weakly affected by the adopted distance function, as expected. In particular, for both Vietnam and Philippines POCs, individual-level calibration weights corresponding to *raking* and *logit* distance functions seem to be nearly indistinguishable.

POC	Weights Type	Distance Function	Sample Distribution Summary						Range of g (g = w/d)
			Min	1st Q	Median	Mean	3rd Q	Max	
Vietnam	Calibration *household-level* ($\mathbf{w}^{CAL1}$)	Linear	130	1030	1765	2244	2806	33520	[0.40, 2.60]
		Raking	130	1104	1750	2244	2701	33520	[0.40, 2.60]
		Logit	132	1092	1702	2244	2741	33440	[0.40, 2.60]
	Calibration *individual-level* ($\mathbf{w}^{CAL2}$)	Linear	195	1429	1976	2244	2711	26790	[0.60, 2.40]
		Raking	195	1446	1972	2244	2687	26790	[0.60, 2.40]
		Logit	199	1448	1956	2244	2674	26650	[0.60, 2.40]
	Direct ($\mathbf{d}$)	–	325	1667	2114	2257	2662	12890	–
Thailand	Calibration *household-level* ($\mathbf{w}^{CAL1}$)	Linear	4	133	282	477	590	9262	[0.36, 2.35]
		Raking	4	135	283	477	588	9262	[0.36, 2.35]
		Logit	4	133	285	477	587	9258	[0.36, 2.35]
	Calibration *individual-level* ($\mathbf{w}^{CAL2}$)	Linear	4	149	297	477	596	9262	[0.36, 2.35]
		Raking	5	155	300	477	603	8501	[0.27, 2.70]
		Logit	4	149	297	477	595	9256	[0.36, 2.35]
	Direct ($\mathbf{d}$)	–	11	147	303	462	651	3941	–
Philippines	Calibration *household-level* ($\mathbf{w}^{CAL1}$)	Linear	58	310	414	467	586	3282	[0.70, 1.60]
		Raking	58	310	414	467	583	3282	[0.70, 1.60]
		Logit	58	310	412	467	586	3282	[0.70, 1.60]
	Calibration *individual-level* ($\mathbf{w}^{CAL2}$)	Linear	69	347	444	467	562	3446	[0.80, 1.50]
		Raking	70	348	443	467	561	3514	[0.80, 1.50]
		Logit	68	347	443	467	560	3452	[0.80, 1.50]
	Direct ($\mathbf{d}$)	–	82	356	443	452	538	2428	–

Table 4: Sample Distribution Summaries for Different Weights – All POCs

This close resemblance is even more evident in Figure 3 and Figure 4, where individual-level raking weights are plotted against individual-level logit weights, for Vietnam and Philippines respectively. Therefore, we felt free to deliver to the World Bank *raking* calibration weights as final output for all POCs, for both household-level and individual-level calibration subtasks.

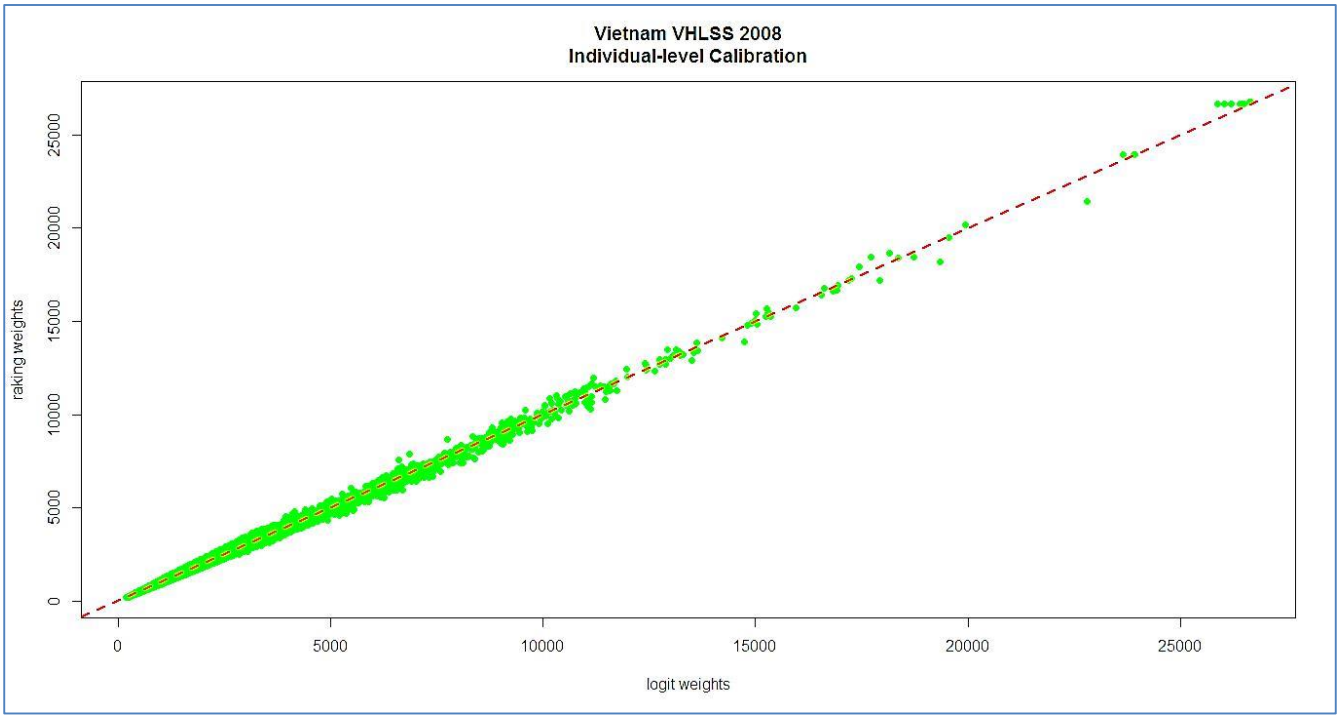

Figure 3: Scatterplot of Raking Weights vs. Logit Weights – Vietnam POC, Individual-level Calibration

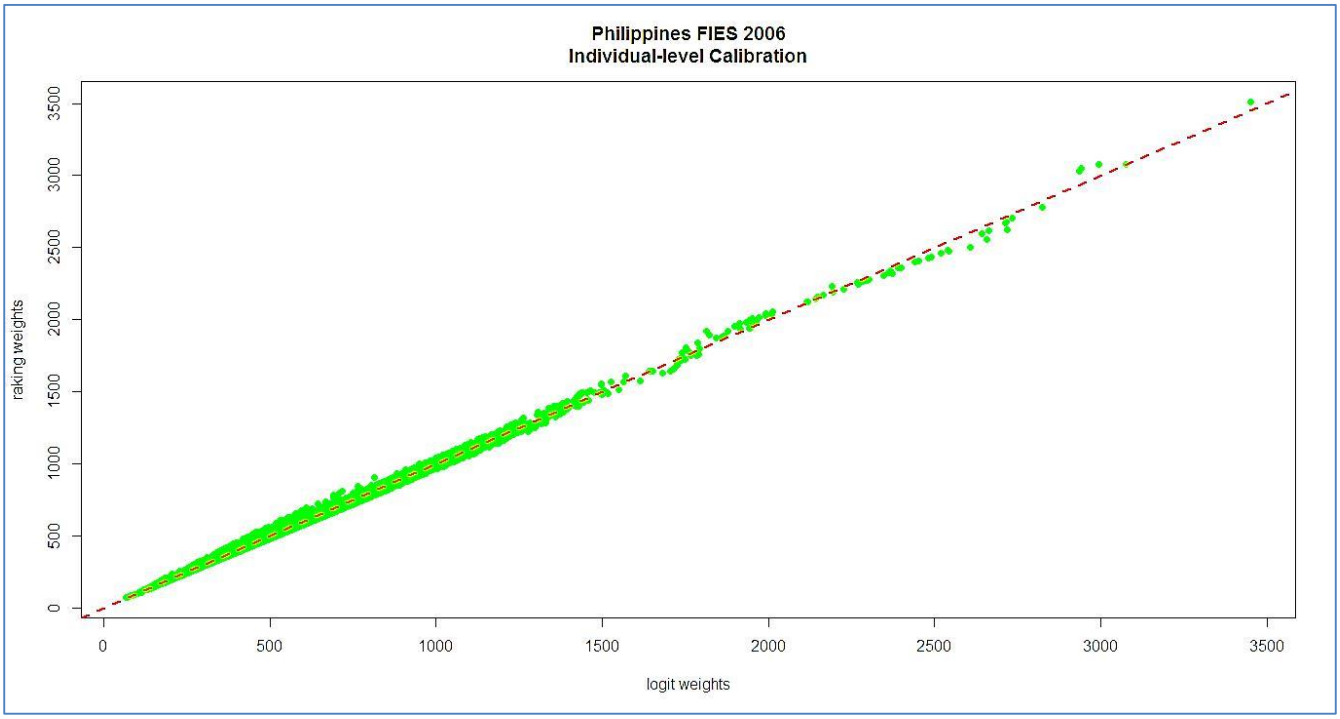

Figure 4: Scatterplot of Raking Weights vs. Logit Weights – Philippines POC, Individual-level Calibration

Table 5 reports notable features of the final calibration weights we delivered for each POC. Note that, unless where expressly stated, these features are common to both household-level ($\mathbf{w}^{CAL1}$) and individual-level ($\mathbf{w}^{CAL2}$) calibration weights. Note also that, for the Philippines POC, rural/urban status was not available, and we could not exploit it as auxiliary variable.

Feature	Vietnam POC	Thailand POC	Philippines POC
Calibration Constraints	• Population counts by age (five-year classes), sex and rural/urban • Household counts by size (9 classes) and rural/urban	• Population counts by age (five-year classes), sex and rural/urban • Household counts by size (5 classes) and rural/urban	• Population counts by age (five-year classes) and sex • Household counts by size (8 classes)
Number of Constraints	82	74	40
Number of Individual Weights	38,253	138,282	189,079
Number of Household Weights	9,189	44,273	38,483
Distance Function	Raking	Raking	Raking

Table 5: Notable Features of the Delivered Calibration Weights w^{CAL1} and w^{CAL2} – All POCs

To give a visual impression of the distribution of the best performing calibration weights we achieved for each POC, we plotted them against the direct weights in Figures Figure 5Figure 6Figure 7, with household-level weights ($\mathbf{w}^{CAL1}$) in red and individual-level ($\mathbf{w}^{CAL2}$) weights in blue.

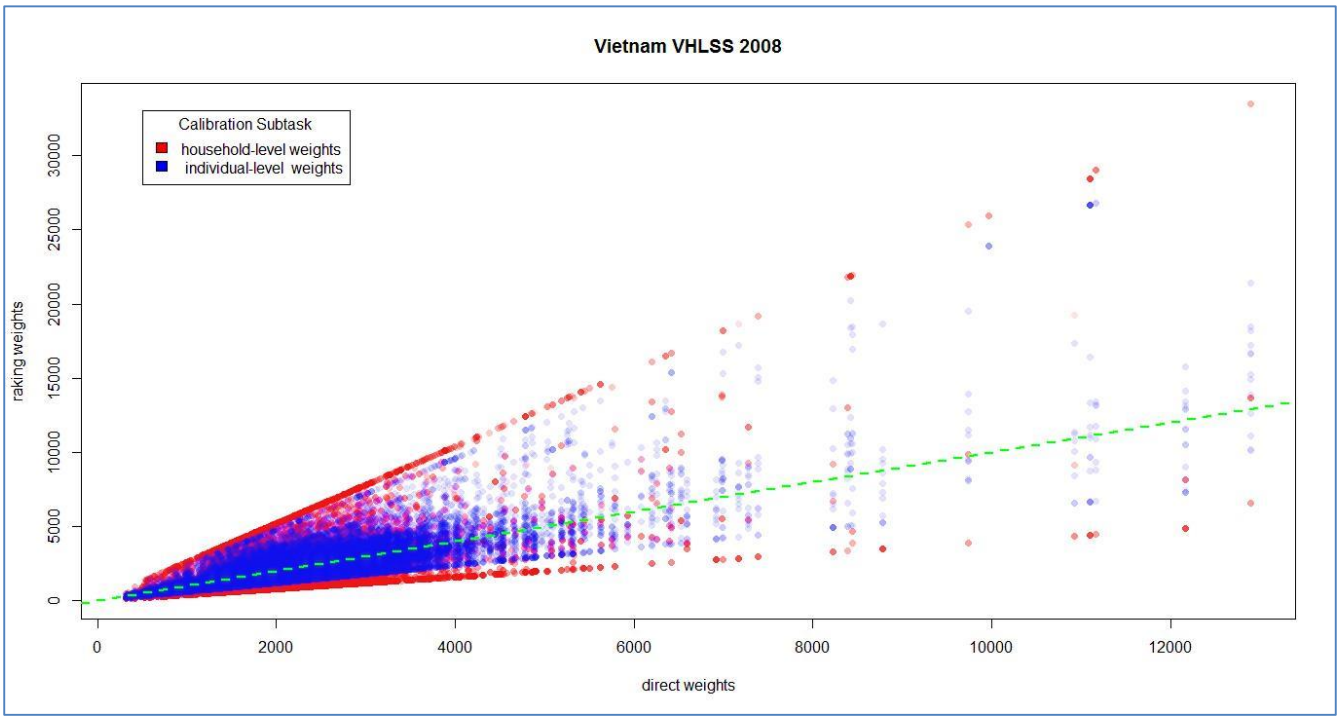

Figure 5: Best Calibration Weights vs. Direct Weights – Vietnam POC, Household-level and Individual-level Weights

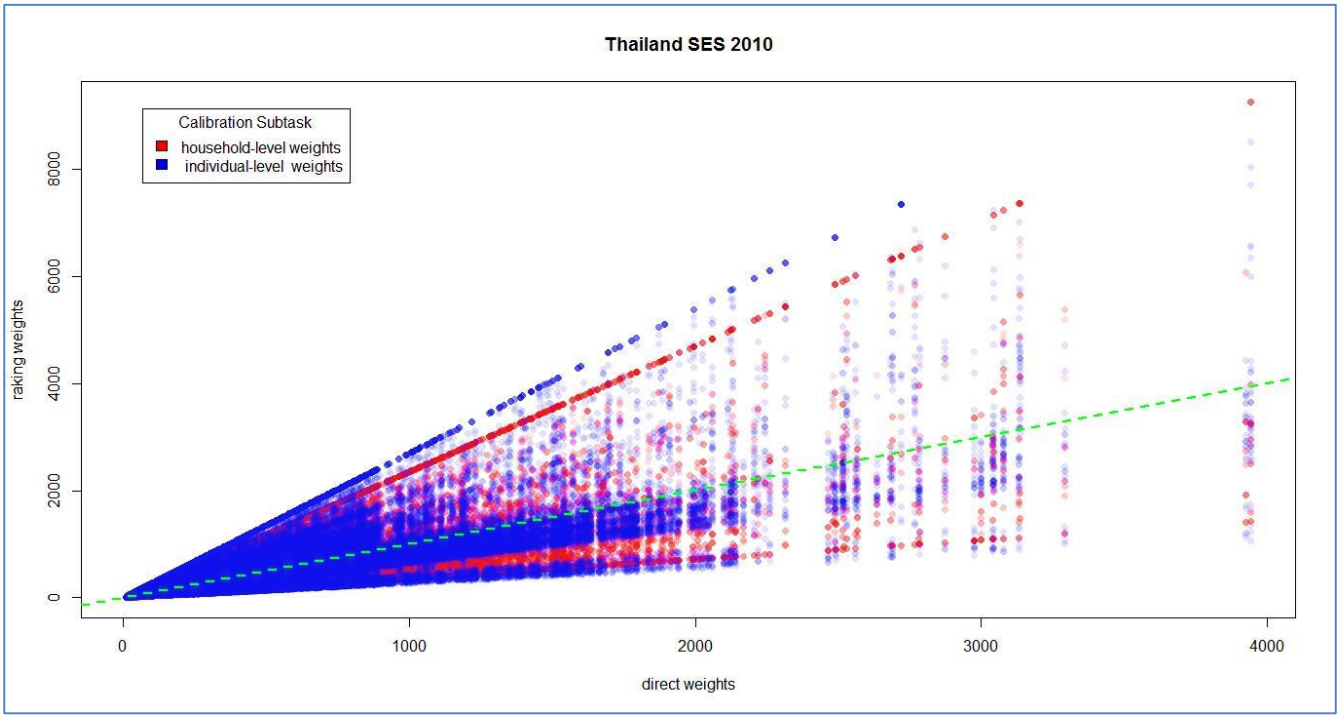

Figure 6 Best Calibration Weights vs. Direct Weights – Thailand POC, Household-level and Individual-level Weights

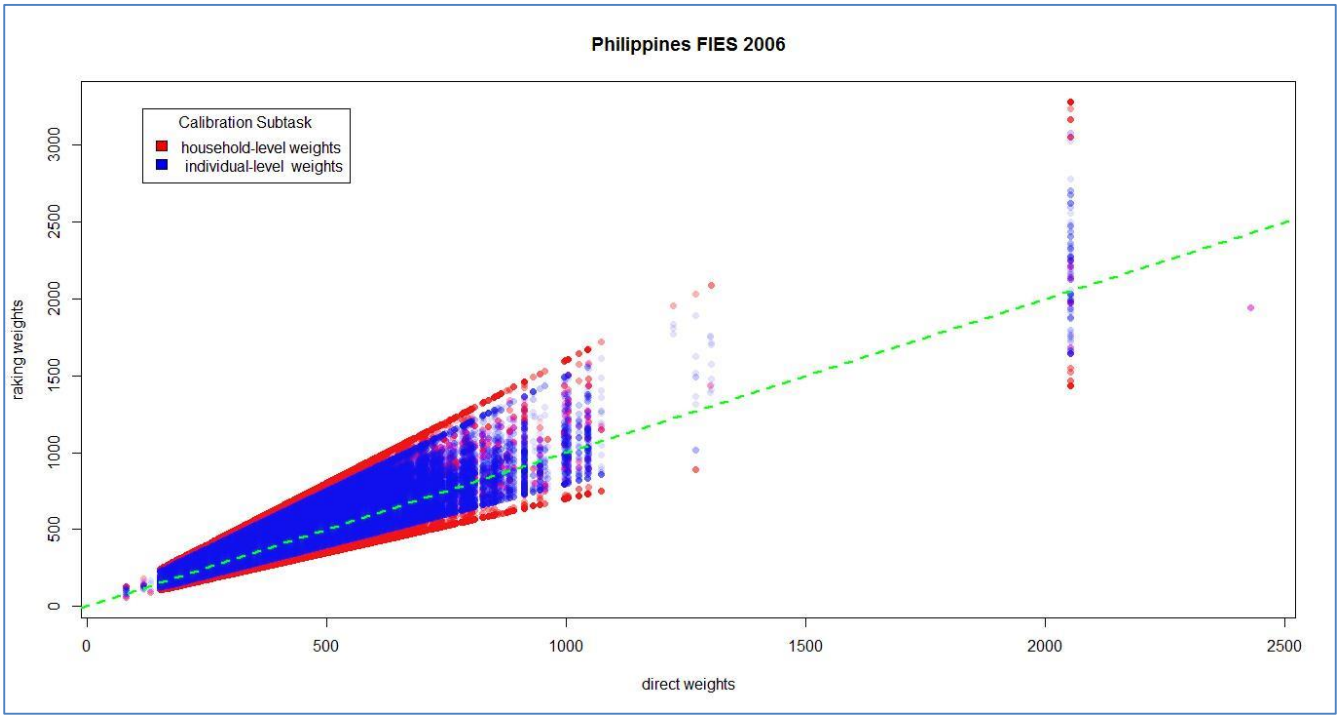

Figure 7: Best Calibration Weights vs. Direct Weights – Philippines POC, Household-level and Individual-level Weights

We end this Section by stressing that our feasibility study on calibration was successful for all three POCs. Exact convergence of numerical optimization routines was always obtained, and calibration weights fulfilling all the requirements laid down in the TOR were delivered. In addition, relying on ReGenesees, we have been able to execute all the calibration tasks in an ordinary PC environment (Windows 7 64-bit OS, 4 GB RAM, dual-core CPU 1.5 GHz + 1.5 GHz), without running into computational troubles.

5. Impact on Poverty and Inequality Estimates

As we already observed, survey bias can propagate across estimates, owing to the correlation structure of the involved variables. Therefore, we cannot take for granted that the bias we diagnosed in Section 3 stays confined to the estimates of age-sex and household size distributions.

In the last task of our study, we investigated possible signals of bias affecting poverty and inequality estimates. For simplicity we addressed only *absolute* poverty measures, relying on *consumption* as pivot variable and adopting as poverty thresholds two *international poverty lines* defined by the World Bank, namely $PPP1.25 and $PPP2.5 per capita per day. The reason for adopting international – rather than *national* – poverty lines is that the household consumption aggregates provided by the World Bank were not exactly the same as the ones used by the NSIs. Even using national poverty lines, we would have not been able to replicate official national poverty rates.

Since no reliable external sources of information were available for poverty, we could not detect potential symptoms of bias by direct inspection of HT estimates. Instead, we assumed calibrated estimates of poverty as a benchmark, and compared HT estimates to that benchmark. Of course, this approach rests on the hypothesis that calibrated estimates provide a better approximation of the unknown poverty parameters than HT.

Needless to say, we delegated again to ReGenesees the computation of all the estimates and sampling errors needed to complete the last task of our feasibility study.

Given the pivotal relevance of the consumption variable, we tried first to assess the impact of calibration on the population distribution of that variable. Table 6 reports selected measures of central tendency for the population distribution of yearly per capita consumption, for all three POCs. HT estimates are contrasted to calibrated estimates obtained through household-level calibration weights (CAL1) and individual-level calibration weights (CAL2). Notably, the observed discrepancies between HT and calibrated estimates are modest, though not entirely negligible. In absolute values, relative differences for Vietnam and the Philippines stay below 5 and 2 percentage points respectively. The effect of calibration seems more significant for Thailand, with discrepancies of up to 10 percentage points.

Interestingly, all the selected measures of central tendency have been shifted toward *higher* values by calibration, with stronger effects for Thailand and with the only exception of individual-level calibration in the Philippines POC. In the light of this finding, one would expect calibration to *decrease* estimated absolute poverty rates, with stronger effects for Thailand. Surprisingly, Thailand will actually contradict such expectation, an outcome that deserves a dedicated explanation. As we will see in the following, the very low incidence of absolute poverty in Thailand will offer a clue to solve this puzzle.

		Estimator (local currency)			Percent Variation (%)	
POC	**Yearly per Capita Consumption**	**HT**	**CAL1**	**CAL2**	**(CAL1 – HT) / HT**	**(CAL2 – HT) / HT**
Vietnam	1st Q	4,443,565	4,491,347	4,540,060	1.1%	2.2%
	Median	6,297,648	6,416,742	6,466,776	1.9%	2.7%
	Mean	7,747,350	7,791,349	7,998,217	0.6%	3.2%
	3rd Q	9,100,144	9,206,175	9,430,316	1.2%	3.6%
Thailand	1st Q	27,909	28,786	28,351	3.1%	1.6%
	Median	40,477	43,701	42,736	8.0%	5.6%
	Mean	52,360	56,577	55,653	8.1%	6.3%
	3rd Q	62,560	69,001	67,577	10.3%	8.0%
Philippines	1st Q	11,936	11,755	11,976	-1.5%	0.3%
	Median	19,566	19,276	19,700	-1.5%	0.7%
	Mean	26,897	26,451	27,065	-1.7%	0.6%
	3rd Q	33,762	33,212	34,012	-1.6%	0.7%

Table 6: Central Tendency of Yearly per Capita Consumption – All POCs, HT and Calibration Estimates

The tendency of calibration to drag the consumption distribution to the right emerges clearly from Figure 8 for the Vietnam POC. Here dots and vertical error bars represent estimated deciles of yearly per capita consumption and 95% confidence intervals. Blue dots and bars stand for HT, green ones for CAL1, and red ones for CAL2. Evidently, the noted effect has a quite modest size: calibrated estimates of deciles are always covered by the corresponding 95% confidence intervals of "uncalibrated" HT estimates.

The same features can be read at a deeper detail in Figure 9. Here CAL1 and CAL2 estimates of percentiles are plotted as red and blue continuous lines respectively. Moreover, to help visualize the impact of calibration while taking into account sampling uncertainty, 95% confidence intervals of HT estimates are drawn as a grey area. Once more, the effect of calibration seems not statistically significant, because calibrated percentiles curves never happen to leave the HT grey area.

As far as the statistical significance analysis underlying Figure 9 is concerned, the Philippines POC gives identical results. On the contrary, the same analysis cannot be deemed sound for the Thailand POC, since the lack of PSU identifiers prevented us from computing reliable confidence intervals (recall the discussion of Section 3).

Even though we were unable to assess the uncertainty affecting consumption estimates for Thailand, we could investigate further the impact of calibration for the Thailand POC by studying the population distribution of consumption at a deeper level. Figure 10 focuses on low quantiles of consumption, more precisely on quantiles below the tenth percentile. Here CAL1 and CAL2 estimates of low quantiles are plotted as red and blue continuous lines respectively, whereas the dashed black lines represent HT estimates in both panels.

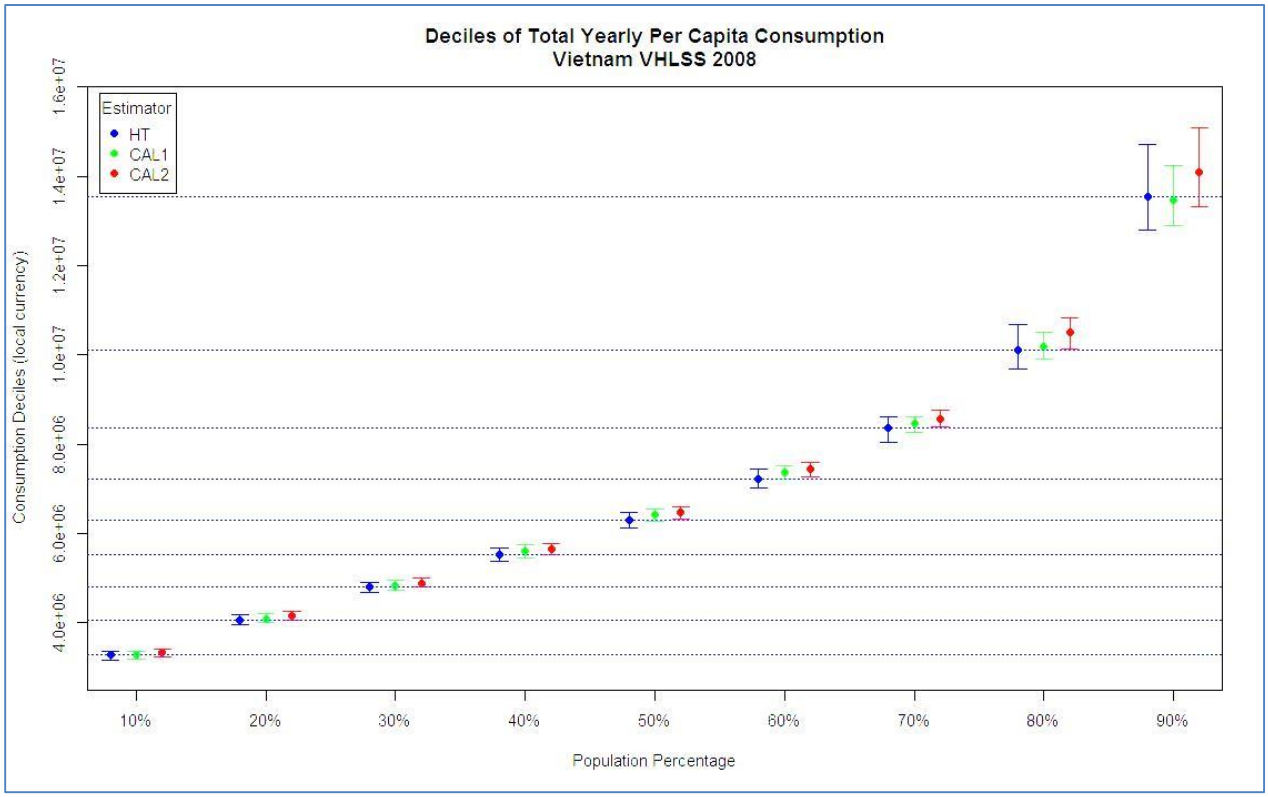

Figure 8: Deciles of Yearly per Capita Consumption – Vietnam POC, HT and Calibration Estimates

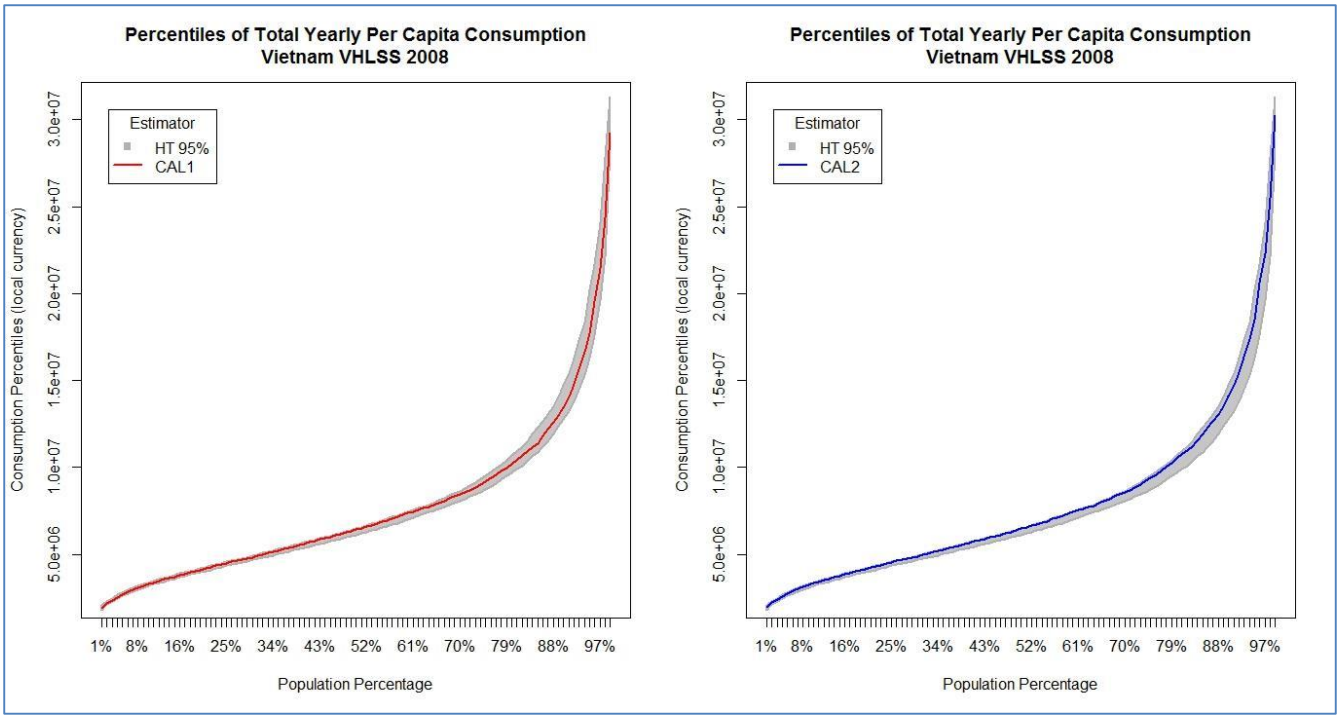

Figure 9: Percentiles of Yearly per Capita Consumption – Vietnam POC, HT 95% Confidence Region and Calibration Estimates

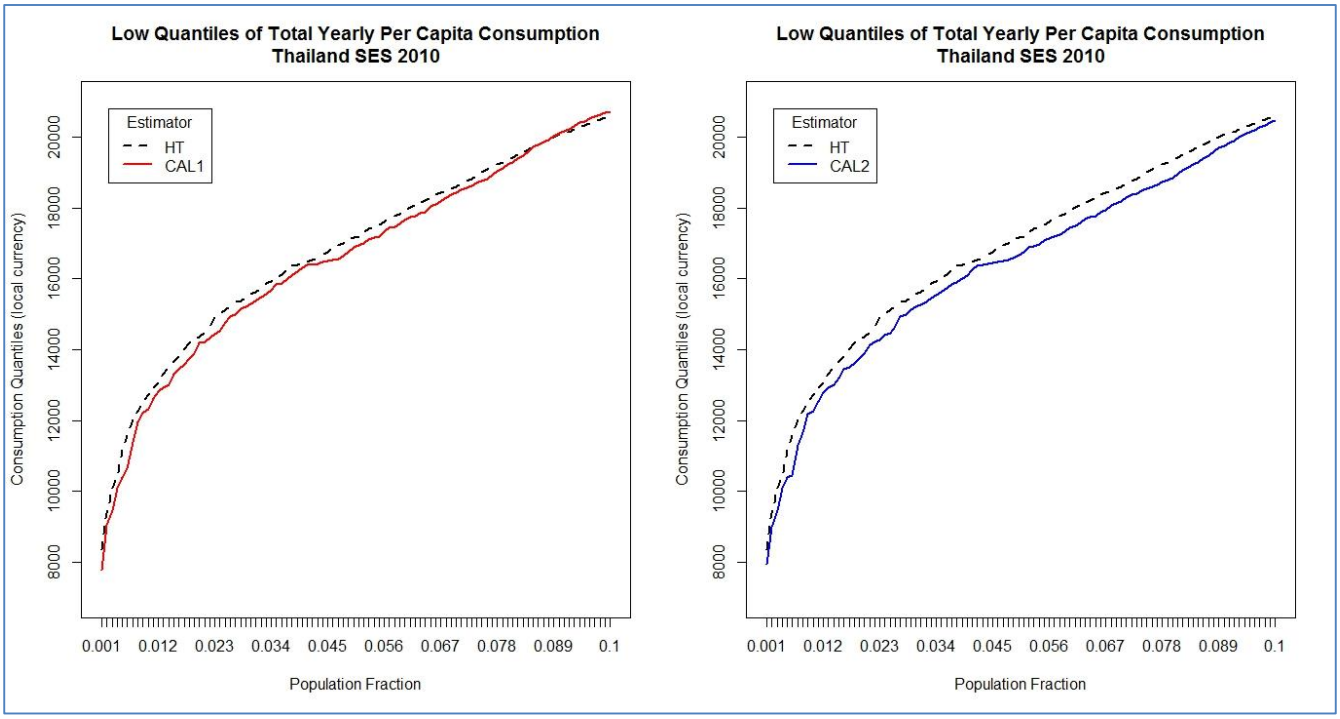

Figure 10: Low Quantiles of Yearly per Capita Consumption – Thailand POC, HT and Calibration Estimates

A clear pattern emerges from Figure 10. Both calibration procedures (CAL1 and CAL2) turn out to *lower* the estimated quantile curve in the low probability region. Indeed, almost all calibration estimates of quantiles below the tenth percentile are *smaller* than the corresponding HT estimates. Therefore, for the Thailand POC, calibration affects low quantiles and quartiles in *opposite* ways: while the center of the consumption distribution is pushed towards higher values (recall Table 6), the left tail is pulled towards lower values. This is a remarkable observation: because absolute poverty has a very low incidence in Thailand (see Table 10), the dynamics of the low quantiles region will turn out to be the decisive one. Eventually, we are led to expect that calibration will *increase* poverty rate estimates.

Summarizing all the findings discussed above, for Vietnam and the Philippines we can state that:

(i) The impact of calibration on investigated consumption statistics is moderate, yet not completely negligible.

(ii) We see no compelling evidence of bias affecting HT estimates of investigated consumption statistics.

Of course, we cannot definitely rule out the possibility that HT consumption statistics are indeed biased, but, if this is the case, then we must conclude that either this hypothetical bias is very small or calibration did not reduce it appreciably.

As far as Thailand is concerned, we have a more controversial situation:

(i) The effect of calibration on investigated consumption statistics is noticeable, but we are unable to evaluate the statistical significance of that effect.

(ii) We cannot draw any conclusion on possible signals of bias affecting HT estimates of investigated consumption statistics.

Switching to absolute poverty estimation, Table 7 converts the international poverty lines of $PPP1.25 and $PPP2.5 per capita per day into local currencies and to annualized values. For conciseness, from now on, we will refer to the transformed international poverty lines – expressed in local currency per capita per year – as IPL_1 and IPL_2.

Country	Year	Purchasing Power Parity (conversion factor)	Days	International Poverty Line (dollars per capita per day)	International Poverty Line (local currency per capita per year)
Vietnam	2008	7688.730729	365	1.25	3,507,983
				2.5	7,015,967
Thailand	2010	18.072579	365	1.25	8,246
				2.5	16,491
Philippines	2006	24.885278	365	1.25	11,354
				2.5	22,708

Table 7: International Poverty Lines of PPP $1.25 and PPP $2.5 per Capita per Day and Local Conversions

It is worth clarifying, at this stage, that only *household* consumption values were available in the survey data files provided by the World Bank. Thus, to be able to analyze per capita consumption, we had to attribute an equal share of household consumption to each household member. This definition of individual consumption implies that if a person has a consumption value below the poverty line, then the same happens to all the members of his household. Hence no household can contain a poor and a non-poor member at the same time. Therefore, in this setting, 'poor households' are univocally defined as households whose members are all poor (and, conversely, 'non-poor households' are univocally defined as households without any poor member).

Table 8 summarizes the impact of calibration on absolute poverty rates for the Vietnam POC, comparing HT estimates to CAL1 and CAL2 estimates. Estimated absolute poverty rates defined with respect to both poverty lines IPL_1 and IPL_2 are reported both for individuals and for households. Moreover, for each unit of analysis, estimates are presented for the whole population, as well as for rural and urban subpopulations.

		Absolute Poverty Rate (%)					
		IPL_1			IPL_2		
Unit of Analysis	Domain	HT	CAL1	CAL2	HT	CAL1	CAL2
Individual	Whole population	12.7%	12.5%	12.0%	57.8%	56.4%	55.8%
Individual	Rural population	16.9%	17.1%	16.3%	71.2%	70.4%	70.2%
Individual	Urban population	1.6%	1.7%	1.7%	22.6%	23.1%	21.9%
Household	Whole population	10.6%	10.5%	9.9%	54.8%	52.9%	**52.5%**
Household	Rural population	14.2%	14.6%	13.7%	68.1%	67.1%	66.9%
Household	Urban population	1.3%	1.3%	1.3%	20.6%	20.5%	19.7%

Table 8: Absolute Poverty Rates for Individuals and Households – Vietnam POC, HT and Calibration Estimates

Apparently, calibration did not affect absolute poverty estimates that much. As expected from the observed tendency of calibration to drag the consumption distribution to the right, the majority of reported calibration estimates of poverty are smaller than the corresponding HT estimates. Anyway, estimates are fairly stable: in absolute values, relative differences between HT and calibration estimates stay below 7 percentage points. In addition, all calibration estimates are covered by the corresponding 95% confidence intervals of HT estimates, with just a single exception signaled by the bold value. We note, incidentally, that this calibration estimate (namely: household, whole population, IPL_2, CAL2) would be covered by the relevant 99% HT confidence interval. Lastly, the calibration effect shows roughly the same magnitude on individual and household poverty rates.

		Absolute Poverty Rate (%)					
		IPL_1			IPL_2		
Unit of Analysis	**Domain**	**HT**	**CAL1**	**CAL2**	**HT**	**CAL1**	**CAL2**
Individual	Whole population	22.7%	23.4%	22.6%	57.2%	57.9%	56.9%
Individual	Rural population	36.8%	37.8%	36.7%	78.1%	78.7%	77.8%
Individual	Urban population	8.4%	8.9%	8.5%	36.1%	37.1%	36.0%
Household	Whole population	18.1%	18.0%	**17.4%**	51.8%	51.3%	**50.6%**
Household	Rural population	29.6%	29.6%	28.6%	72.9%	72.5%	**71.7%**
Household	Urban population	6.4%	6.4%	6.1%	30.4%	30.1%	29.4%

Table 9: Absolute Poverty Rates for Individuals and Households – Philippines POC, HT and Calibration Estimates

As shown in Table 9, which refers to the Philippines POC, calibration effects on poverty estimates exhibit almost identical patterns for the Philippines and Vietnam. Again, calibration does not strongly affect estimates, with relative shifts from HT whose absolute values are below 6 percentage points. Once more, the shifts on absolute poverty estimates induced by calibration are mostly statistically not significant, with only 3 exceptions stressed by bold figures. Unsurprisingly, the statistical significance of two of these three exceptions evaporates if one increases the confidence level of the analysis to 99%. Once again, the magnitude of the calibration effect on individual and household poverty rates is roughly the same.

		Absolute Poverty Rate (%)					
		IPL_1			IPL_2		
Unit of Analysis	**Domain**	**HT**	**CAL1**	**CAL2**	**HT**	**CAL1**	**CAL2**
Individual	Whole population	0.09%	0.15%	0.16%	4.11%	4.42%	4.61%
Individual	Rural population	0.13%	0.26%	0.27%	5.80%	7.24%	7.58%
Individual	Urban population	0.01%	0.02%	0.01%	0.77%	0.86%	0.86%
Household	Whole population	0.05%	0.07%	0.07%	2.70%	2.46%	2.60%
Household	Rural population	0.07%	0.13%	0.13%	3.95%	4.27%	4.53%
Household	Urban population	0.01%	0.01%	0.01%	0.46%	0.46%	0.47%

Table 10: Absolute Poverty Rates for Individuals and Households – Thailand POC, HT and Calibration Estimates

As summarized in Table 10, and consistently with our previous analysis on consumption, Thailand poverty estimates exhibit a quite different behavior, as compared to Vietnam and the Philippines. First, Table 10 shows a marked impact of calibration on poverty estimates, with relative shifts from HT that even exceed 100% in absolute values. Second, calibration systematically increases all the reported estimates of poverty rates. At this stage, this comes as no surprise, given the very low incidence of poverty in Thailand and the already observed tendency of calibration to pull the left tail of the Thailand consumption distribution toward lower values. Third, the magnitude of the calibration effect on poverty seems considerably bigger for individuals than for households, again at odds with what we saw for Vietnam and the Philippines. We end this comment on the Thailand POC with an important general remark: Thailand poverty rates are *so small* that *direct* estimation methods could fail (i.e. provide unreliable estimates), no matter if one adopts HT or calibration estimators. Just to give an impression: out of ~138,000 individuals and ~44,000 households belonging to the Thailand SES 2010 sample dataset, only 122 individuals and 22 households happen to fall below the international poverty line IPL_1. Such small domain sample sizes would claim the adoption of *indirect*, model-based Small Area Estimation (SAE) methods, rather than ordinary design-based and model-assisted ones.

Summarizing all the findings discussed above, for Vietnam and the Philippines we can state that:

(i) The impact of calibration on investigated poverty estimates is moderate, yet not completely negligible.

(ii) We see no compelling evidence of bias affecting HT estimates of investigated poverty rates.

Of course, we cannot definitely rule out the possibility that HT estimates of poverty rates are indeed biased, but, if this is the case, then we must conclude that either this hypothetical bias is very small or calibration did not reduce it appreciably.

As far as Thailand is concerned, we have a more interesting, yet controversial, situation:

(i) The effect of calibration on investigated poverty estimates is marked (and bigger for individuals than for households), but we are unable to evaluate the statistical significance of that effect.

(ii) We cannot draw any conclusion on possible signals of bias affecting HT estimates of investigated poverty rates.

(iii) Thailand absolute poverty rates are so small that neither HT nor calibration estimators are likely to provide estimates of adequate precision. The adoption of Small Area Estimation methods should be considered, instead.

6. Conclusions

The World Bank's Household Survey Development Team noticed that in some South-East Asia countries the distribution of the population by age and sex and the distribution of households by size differ very significantly between large-scale socio-economic sample surveys and the Population Census. As these surveys are supposed to be (at least) nationally-representative, discrepancies of this magnitude are not expected.

A technical feasibility study has been carried out to investigate whether this issue could be solved through a preliminary calibration procedure [4]. Three Proofs of Concept (POC) have been carried out, adopting as empirical test bed the following household surveys: (i) 2008 Vietnam Household Living Standards Survey – VHLSS 2008, (ii) 2010 Thailand Household Socio-Economic Survey – SES 2010, (iii) 2006 Philippines Family Income and Expenditure Survey – FIES 2006. For each survey, calibration constraints have been imposed on

known population totals derived from the closest Population and Housing Census round: (i) 2009 Vietnam Census, (ii) 2010 Thailand Census, (iii) 2007 Philippines Census.

The aim of the study was threefold, and each POC was accordingly structured into three consecutive tasks:

(1) Bias Analysis – *Investigate whether the observed large discrepancies between survey-based estimates and Census counts are artifacts of random sampling, or rather genuine symptoms of bias.*

(2) Calibration – *Verify whether a calibration procedure can actually succeed in making both the age-sex pyramids and the distribution of households by size simultaneously consistent with the Census data.*

(3) Impact on Poverty and Inequality Estimates – *In case the task at point (2) is feasible, assess the impact of the aforementioned large discrepancies on key poverty and inequality indicators, by comparing their Horvitz-Thompson (HT) and calibration (CAL) estimates.*

The main findings of the study can be summarized as follows:

(1) Bias Analysis – *The discrepancies observed for the distributions of individuals by age-sex and of households by size are true symptoms of bias, for Vietnam and the Philippines. For Thailand, a definitive conclusion cannot be drawn, owing to insufficient information on the sampling design.*

(2) Calibration – *The feasibility study on calibration has been successful for all three POCs. Exact convergence of numerical optimization routines has been always obtained, and calibration weights fulfilling all the requirements laid down in the 'Terms of Reference' (TOR) have been delivered.*

(3) Impact on Poverty and Inequality Estimates – *For Vietnam and the Philippines, no compelling evidence has been found of bias affecting HT estimates of investigated consumption statistics and poverty rates. Moreover, the impact of calibration turns out to be moderate, yet not completely negligible. For Thailand – once more – a definitive conclusion cannot be drawn on bias, owing to insufficient information on the sampling design. Nevertheless, calibration markedly increases estimated poverty rates. Anyway, Thailand poverty rates are so small that direct estimation methods are likely to fail (i.e. provide unreliable estimates), no matter if one adopts HT or calibration estimators. Indirect, model-based Small Area Estimation (SAE) methods should be preferred.*

To tackle the feasibility study, the ReGenesees system has been used: an open source software for design-based and model-assisted analysis of complex sample surveys [10], based on R [3]. Overall, the study showed that *it is technically feasible to integrate a calibration procedure in the production workflow of all the household surveys taken into account*.

Beyond the feasibility study, two possible implementation lines can be envisioned:

- A calibration procedure could be executed directly by the National Statistical Institute (NSI) in charge of the household survey, as a process step to be routinely performed preliminary to estimation. Of course, enabling NSIs to adopt calibration estimators would require appropriate capacity building actions.
- A calibration procedure could be executed ex-post for analysis purposes, i.e. after data dissemination and outside the involved NSIs, in order to increase the quality of the estimates derived from the surveys. For instance, the World Bank could manage the calibration procedure on its own, and integrate the obtained calibration weights into its microdata repositories.

In both cases, using calibration weights for estimation would be straightforward, while estimating sampling errors would require specialized software, like the ReGenesees system.

References

[1] Deville, J.C., and Särndal, C.E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 376-382.

[2] Lemaitre, G., and Dufour, J. (1987). An Integrated Method for Weighting Persons and Families. Survey Methodology, 13(2), 199-207.

[3] R Core Team (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org (accessed November 2015).

[4] Särndal, C.E. (2007). The Calibration Approach in Survey Theory and Practice. Survey Methodology, 33 (2), 99-119.

[5] Särndal, C.E., and Lundstrom, S. (2005). Estimation in Surveys with Nonresponse. John Wiley & Sons

[6] Särndal, C.E., Swensson, B., and Wretman, J. (1989). The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator of the Finite Population Total. Biometrika, 76, 527-537.

[7] Särndal, C.E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling, Springer Verlag.

[8] United Nations (1993). Sampling Errors in Household Surveys. NHSCP Technical Study UNFPA/UN/INT-92-P80-15E, New York.

[9] Wu, S., Kennedy, B., and Singh, A.C. (1997). Household-level versus Person-level Regression Weight Calibration for Household Surveys. Annual Meeting of the Statistical Society of Canada, Proceedings of the Survey Methods Section.

[10] Zardetto, D. (2015[a]). ReGenesees: an Advanced R System for Calibration, Estimation and Sampling Error Assessment in Complex Sample Surveys, Journal of Official Statistics, 31 (2): 177-203.

[11] Zardetto, D. (2015[b]) ReGenesees: R Evolved Generalized Software for Sampling Estimates and Errors in Surveys. R package version 1.7, Istat, Italy. Available at: https://joinup.ec.europa.eu/software/regenesees/description (accessed November 2015).

[12] Zardetto, D., and Cianchetta, R. (2015) ReGenesees.GUI: a TclTk Interface for the ReGenesees Package. R package version 1.7, Istat, Italy. Available at: https://joinup.ec.europa.eu/software/regenesees/description (accessed November 2015).

Annex 1: Terms of Reference

BACKGROUND

To better coordinate the activities of international organizations engaged in household surveys and develop joint initiatives, the World Bank and other international organizations established the International Household Survey Network (IHSN – www.ihsn.org).

The mission of the IHSN is to improve the availability, accessibility and quality of survey data to encourage their use by national and international development decision makers, users and stakeholders. Its members agreed to work towards these goals through a program consisting of

- coordination and harmonization of survey activities;
- development of improved tools and guidelines related to the processing, analysis, archiving, and dissemination of survey data and metadata;
- assessment of survey instruments and research work on survey methods.

The IHSN activities are coordinated by the IHSN Secretariat at the World Bank, and implemented in close cooperation with various international partners.

TASKS

The consultant will be recruited for a total of up to 30 working days (between May 8, 2015 and December 31, 2015) to assist the IHSN Secretariat in the testing and documentation of sample calibration techniques.

Survey datasets from Thailand, Vietnam and the Philippines will be provided by the World Bank Data Group to the consultant, together with tabular data from population censuses. The survey datasets will include the following variables (at least): household survey ID, individual ID, sampling weight, stratum and PSU (as information on sampling design), region, urban/rural, sex, age and household consumption. The census tables will provide data on the distribution of the population by age group and sex, as well as the distribution of households by size. The datasets will be organized in a standardized way (consistent variable names across datasets). The datasets will be provided in Stata and/or CSV.

The consultant will apply sample calibration methods, using the ReGenesees R package. The objective will be to produce calibrated sets of sample weights to adjust the survey extrapolated population to the population census tables. Two sets of calibrated weights will be produced:

- one set where the constraints are that the calibrated sample weights are > 0 and as close as possible to the original weights (assuming a solution matching the requirement of positive sample weights can be found; otherwise the condition will be lifted)
- one set with the additional constraint that all members of each household have the same sample weight.

These calibrated weights will then be used to calculate a set of key indicators (poverty headcount using an international poverty line, Gini coefficient and other inequality indicators) and to assess the impact of calibration on these estimates. These calculations will be made using publicly-available R package(s) chosen by the consultant.

DELIVERABLES

The expected deliverables include:

- a dataset with calibrated weights for each of the three countries
- estimates of key poverty and inequality indicators with sampling errors
- the R scripts that would allow full replication of the calculations by the World Bank Data Group or other users
- a technical note on the calibration process.

LOCATION AND REPORTING

The consultant will work from home office (Italy), with regular interaction with the IHSN Secretariat by e-mail or telephone/Skype.
Data and other materials will be shared using a Box folder created by the World Bank Data Group.
The main counterpart for the consultant will be Olivier Dupriez, Lead Statistician at the World Bank Data Group and coordinator of the IHSN Secretariat.

Annex 2: ReGenesees in a Nutshell

What is ReGenesees?

ReGenesees (R Evolved Generalized Software for Sampling Estimates and Errors in Surveys) is a full-fledged R software for design-based and model-assisted analysis of complex sample surveys [10]. This system is the outcome of a long-term research and development project, aimed at defining a new standard for calibration, estimation and sampling error assessment to be adopted in all large-scale sample surveys routinely carried out by Istat (the Italian National Institute of Statistics).

The first public release of ReGenesees for general availability dates back to December 2011. The current version is ReGenesees 1.7. The system is distributed as open source software under the European Union Public License (EUPL). It can be freely downloaded from JOINUP (the collaborative platform for interoperability and open source software of the European Commission) and from the Istat website.

System Architecture

ReGenesees has a clear-cut two-layer architecture: the application layer of the system is embedded into an R package named ReGenesees [11]. A second R package, called ReGenesees.GUI [12], implements the presentation layer of the system (namely a Tcl/Tk GUI). Both packages can be run under Windows as well as under Mac, Linux and most of the Unix like operating systems. While the ReGenesees.GUI package requires the ReGenesees package, the latter can be used also without the GUI on top. Thus the statistical functions of the system will always be accessible by users interacting with R through the traditional command-line interface. On the contrary, less experienced R users will take advantage from the user-friendly mouse-click GUI.

Data Input/Output

The ReGenesees system can import data in a variety of ways. First, it can load R workspace files (with .RData or .rda extensions) storing previously saved data. Second, data can be imported from Text Files (with extensions .txt, .csv, .dat). Third, the system can import data from MS Excel spreadsheets and/or MS Access database tables. Currently, ReGenesees can save output data into R workspace files (.RData, .rda) and/or export them into Text Files (.txt, .csv, .dat). Further extensions are possible.

Main Statistical Functions

- **Complex Sampling Designs**
  - Multistage, stratified, clustered, sampling designs
  - Sampling with equal or unequal probabilities, with or without replacement
  - “Mixed” sampling designs (i.e. with both self-representing and non-self-representing strata)
- **Calibration**
  - Global and partitioned (for factorizable calibration models)
  - Unit-level and cluster-level weights adjustment
  - Homoscedastic and heteroscedastic models
  - Linear, raking and logit distance functions
  - Bounded and unbounded weights adjustment
  - Multi-step calibration
- **Basic Estimators**
  - Horvitz-Thompson
  - Calibration Estimators
- **Variance Estimation**
  - Multistage formulation
  - Ultimate Cluster approximation

- Collapsed strata technique for handling lonely PSUs
- Taylor-linearization of nonlinear “smooth” estimators
- Generalized Variance Functions method

- **Estimates and Sampling Errors (standard error, variance, coefficient of variation, confidence interval, design effect) for:**
  - Totals
  - Means
  - Absolute and relative frequency distributions (marginal, conditional and joint)
  - Ratios between totals
  - Multiple regression coefficients
  - Quantiles
- **Estimates and Sampling Errors for Complex Estimators**
  - Handles arbitrary differentiable functions of Horvitz-Thompson or Calibration estimators
  - Complex Estimators can be freely defined by the user
  - Automated Taylor-linearization
  - Design covariance and correlation between Complex Estimators
- **Estimates and Sampling Errors for Subpopulations (Domains)**
  - All the analyses above can be carried out for arbitrary domains

Sample GUI screenshots