Tracing down real socio-economic trends from household data with erratic sampling frames : the case of the Democratic Republic of the Congo

Tracing down real socio-economic trends from household data with erratic sampling frames. The case of the Democratic Republic of the Congo

Abstract

By means of the latest seven national household surveys of the Democratic Republic of the Congo (DRC), this article uncovers the very volatile sampling frame used underneath all survey designs. As a result, the reliability of much associated survey information as well as any corresponding temporal analysis are seriously jeopardized. Relying on recent vaccination, school enrolment and election data, the article proposes a post-stratification technique to retroactively control for these erratic variations in sampling frame in an attempt to identify real socio-economic trends. Although the proposed technique did not restore full comparability of survey data in all respects, it has been able to eliminate an essential part of the spuriousness as illustrated by assessing trends in asset ownership under both the biased and stabilized sampling frames.

Keywords: household surveys, sampling, population estimates, assets, the Democratic Republic of the Congo

1. Introduction

Depending on the nature and resources available, different sampling methods are used to draw inferences about a population from a sample of observations. Although simple random sampling is theoretically the gold standard to generate unbiased and representative[1] samples, in many cases this method is neither feasible (as sampling frames at individual/household level are often simply not available) nor cost- or time-effective (given the geographical dispersion of the population studied). As a result, a combination of (other) sampling techniques is often used to strike a better balance between sample reliability and resource constraints (Deaton, 1997). As a matter of fact, in the Democratic Republic of the Congo (DRC), like for most developing countries, national household surveys involve a combination of stratified, cluster, random and systematic sampling techniques, which, apart from the necessary capacity to administer, also require additional data on the country's demography (Yansaneh, 2005). Unfortunately, for the DRC the latter data can only be labelled as 'guesstimates' (Trefon, 2012) since the core and primary census data on which they are based, goes back to 1984—with irregular and dubious revisions ever since.

After a description of the most common method used in the DRC to assure national survey results to be representative to the whole population, this article points to the highly questionable nature of population data fed into the national sampling frame (section two). In a following section, a three-step post-stratification technique will be proposed to retroactively stabilize the sampling frame[2] in an attempt to isolate real socio-economic changes from methodological fluctuations. In section four, this particular technique will then be applied to the national survey data of the DRC and its impact assessed using a number of socio-economic indicators. Meanwhile, this section will also provide an account of the respective merits and shortcomings of the proposed technique. Section five concludes.

2. Rich sampling menu, poor feeding with population data

Since the beginning of this millennium, the number of representative household surveys executed in the DRC has significantly increased. In total, there were not less than seven national surveys, capturing information on a wide range of socio-economic and demographical topics from more than half a million individuals. Among these surveys, one can find two Multiple Indicators Cluster Surveys (2001 and 2010), two standard budget surveys, called 123 Survey (2004–05 and 2012–13), two Demographic Health Surveys (2007 and 2013–14), and one nationwide survey on the number of Out-Of-School Children (OOSC) executed in 2012[3]. Before the turn of the millennium, national data collection was much less frequent with another MICS survey in 1995 (MICS1) and a national census in 1984. This variable pace in data gathering largely follows the political history of the country, as the end of the Cold War (1989) marked the beginning of a protracted period of instability culminating in overt violence during the second Congo War (1997–2003). After the signing of peace treaties, a period of relative stability began which facilitated the execution of surveys. An overview of the major political events together with the various national survey initiatives is presented in Figure 1.

<< Figure 1 about here >>

For all surveys, a fairly similar sampling design was adopted (as summarized in Table 1), which combines various sampling techniques at different stages. In general, each province's population is first divided in three strata: statutory cities, smaller cities and the rural sector. Within each of these strata, a number of primary sampling units (PSU) are selected using systematic sampling with probabilities proportional to the population size. These PSU concern city quarters (for statutory cities), cities (for the stratum of smaller cities) and chiefdoms/sectors (for the rural sector). For the latter two strata, the secondary sampling stage involves the

selection of city quarters and villages, using respectively simple random sampling and systematic sampling with equal probabilities. In a third stage, a segmentation procedure is first adopted when the previously selected units are still too populated, after which one segment is randomly chosen and all households are enumerated within that segment/unit. Finally, a fixed number of households is selected using systematic sampling with equal probabilities.

<< Table 1 about here >>

Although considerable similarity in sample design exists between all seven national household surveys, some small differences do occur. For example, the sampling procedure of the smaller cities has become identical to that of the statutory cities for the latest 123 Survey (2012–13) and DHS (2013–14), and both these surveys now apply their stratification to each of the 26 new provinces as opposed to the 11 former ones. Further, both 123 Surveys and the MICS2 (2001) make use of systematic sampling with probabilities proportional to the population size to select the secondary sampling units within the rural sector. And finally, the treatment of Kinshasa (and Lubumbashi, for the latest surveys) follows a slightly distinct procedure.

In order for this rich sampling design to be able to generate unbiased and representative samples, it is crucial to have both exhaustive and mutually exclusive lists of PSU and SSU (the latter which only counts within selected PSU) as well as reliable population estimates linked to each (Yansaneh, 2005). For the DRC, both aspects seem to be highly problematic. With respect to the quality of PSU and SSU listings, there seems to be an important lag and disconnect between the administrative records and the reality on the ground (INS, n.d.). This disconnect not only finds its origin in the poor institutional capacity of the country to follow up on local demographic evolutions, but also in local village histories, name claiming being part and parcel of local politics when villages split up or relocate. This phenomenon is exacerbated in the

context of protracted hostilities, characterizing the eastern part of the country (IDMC, 2014). These events, together with other (softer) forms of local contestations, have resulted in an inflation of Congolese villages carrying the same name but differ in number (e.g. Mboma I-IV, Bwisegha I-IV, among others).

However, the most problematic issue by far concerns the demographic data associated with the PSU and SSU frames which largely determine the magnitude of the design weights to assure all survey data and results being representative to the population studied (République Démocratique du Congo, 2014c: 447–460). Indeed, apart from non-response adjustments, these weights are typically constructed based on inverse probabilities of selection, and thus reflect the number of population units that each sample unit 'represents' (Little, 2004; Gelman 2007). In their qualitative assessment of national household surveys up to 2010, Marivoet and De Herdt (2014) already underscored the highly speculative nature of all population data of the DRC—by referring to (i) the outdated and anyway incompletely processed 1984 census, (ii) the variation observed between different sources of information, and (iii) the often mechanical and spatially indifferent yearly updates employed using a standard 3 percent population growth rate.

Indeed, although every single national survey report makes explicit reference to the 1984 census and implicit reference to a 3 percent annual growth rate, reality is more complex—at least when analyzing demographic estimates derived at survey level. Consider Table 2, which provides an overview of provincial population estimates underlying each of the seven national surveys. From this table, one could immediately observe annual growth rates being both different from 3 percent and different among regions. Only the OOSC survey overall seems to adhere to the aforementioned reference, but growth rates do vary between provinces. For the other surveys, the growth rates underlying overall and provincial population updates are oscillating around 3

percent, though with some important deviations through time and place—ranging from 0.1 percent (Nord-Kivu in 2007) to 5.2 percent (Kinshasa in 2007). Although many of these deviations from 3 percent may look rather minimal, the impact over an extended period of time can be quite dramatic: for example, estimating the total population in 2012 by applying either a 3.0 percent (OOSC) or 3.4 percent (123 Survey) growth rate on the census data of 1984, may yield a difference of almost ten million people.

To be sure, making use of varying population growth rates to estimate the current size of the Congolese population in itself should not be worrisome at all—quite the contrary: it shows the willingness of (local) sampling experts and statisticians to give due attention to the various local events which have impacted on the country's demography. On the other hand, the information used for an estimation and its outcome should be credible and justifiable. In the remaining part of this section, several findings will be presented to put serious doubt on the population estimates used to assure national surveys being representative to the Congolese population.

<< Table 2 about here >>

First, many changes in population size are difficult to explain or are counterintuitive. For example, and still making use of the data presented in Table 2, how to explain that the Congolese population would have grown from 55.3 million to 65.8 million individuals over only two years of time (2005–07)? There was not massive immigration into the country after the elections in 2006, nor was the end of the war accompanied by a baby boom. Furthermore, as already highlighted above, the population estimates for both national surveys executed in 2012 (123 survey and OOSC survey) differ by almost ten million people, or 14 percent of the total population. Additionally, it is difficult to explain the demographic declines which would have occurred in certain provinces during certain episodes. As a matter of fact, the population would

have decreased in Kinshasa within the interval 2007–10, and again in 2013–14; in Bas-Congo, between 2005 and 2007, and also in 2012 (OOSC) and 2013–14 ; in the provinces of Bandundu, Equateur and both Kasai between 2007 and 2010 ; and for both Kasai also in 2013–14; in both Kivu provinces and in Katanga within the interval 2005–07 and again in 2013–14 ; and also for the Kivu provinces in 2012 (OOSC). As most of these population declines are lagging quite far behind the episode characterized by the biggest hostilities (i.e. 1997–2003), the presence of conflict cannot be invoked as an explanation. This counts all the more for the regions largely unaffected by conflict. Indeed, why would the ultimate growth pole, Kinshasa, have experienced a decline of 0.6 million people between 2007 and 2010, and another of almost three million in 2013–14?

Second, as a result of these varying changes in population estimates across regions, the demographic weight of each of the country's provinces over time has been fluctuating in quite an erratic and irregular fashion, as can be observed from Figure 2. Apart from Maniema, whose share in overall population has largely remained stable around 3 percent, all other provinces have seen their demographic weight considerably fluctuate between surveys. Whereas these fluctuations are still limited to 2–3 percent for Bas-Congo and Orientale, they become more important for the provinces of Equateur, both Kasai, Kinshasa and Sud-Kivu with variations around 4 percent, and are really striking for Nord-Kivu, Katanga and Bandundu. For the two latter provinces, the share in overall population is oscillating around 13 percent with an amplitude close to 3 percent in both directions.

<< Figure 2 about here >>

And third, a similar question props up when observing Figure 3, which presents the estimated change in urbanization rate in the DRC and in each of the individual provinces[4]. Again,

substantial and irregular variations in urbanization rates can be observed between different surveys administered within a short time interval: for the country as a whole and for Orientale, these fluctuations did not exceed 13 percent, but for Bas-Congo, Bandundu, Equateur, Sud-Kivu and both Kasai provinces, the estimated urbanization rate varied with 20 percent or more; and in Katanga, Nord-Kivu and Maniema, one could even observe changes of around 30 percent. For the latter regions, it then seems that three out of ten Congolese people are floating somewhere in-between the urban and rural sector of their province, without settling in a definitive manner. Another striking observation in this respect is the comparatively high urbanization rate assumed by the DHS survey in 2007. Indeed, this rate has been estimated at 43 percent for the whole country, being on average 10 percent higher than most other national surveys—an observation which largely counts as well across each of the individual provinces.

<< Figure 3 about here >>

In sum, nobody really knows how many Congolese today populate the DRC and how this population is distributed over its vast territory. This is understandable, it is a reflection of both the weak institutional environment and of the difficulty to do standardized survey research in a country like the DRC, but this should not be a reason for sampling experts to just create their own reality. Because what sampling experts and demographers assume to know, based on fragmented or unreliable data fed into projections, not only differs a lot in absolute terms but also follows a very erratic pattern over time for which hardly any explanation can be provided. These population approximations can at best be labelled what they are—very approximate.

3. Technique to stabilize sampling frames

In order to illustrate its impact and to be able to isolate sample fluctuations from real socio-economic changes, this section will now propose a particular technique to stabilize a country's sampling frame over time. Given its core objective of modifying initial design weights, this procedure can be easily classified as a post-stratification technique (Little, 2004). Although this technique will be applied to the national surveys of the DRC, its application can be extended to other survey settings suffering from similar biases. In short, the technique proposed entails the computation of a series of correction factors to adjust the initial design weights comprised in each survey as to get rid of its erratic nature over time. More in particular, the technique involves the following three consecutive steps: (i) identifying an historical and recent population reference which allows for the calculation of spatially refined growth rates; (ii) based on these region-specific growth rates, deriving alternative population figures for each survey year; and (iii) correcting design weights by confronting the initial population estimates with their alternative.

3.1.Finding demographic yardsticks to compute average growth rates

First, in order to minimize the impact of unreliable sampling frames, one cannot—unfortunately—avoid the necessity to come up with a current set of regional population estimates. As a matter of fact, the origin of the problem lies precisely in the application of a too volatile and thus unreliable account on how the Congolese population is distributed over its territory since the census of 1984. As such, the main objective of this alternative set is to provide a steady yardstick, which, in combination with the census data, will allow for the derivation of average population growth rates to level out the erratic pattern of the sampling frame over time. In order to reduce the arbitrariness of this re-estimation exercise and to align with the level of representativeness of all national surveys of the DRC, the data used to execute this first step, as a

principle, should both involve some real counting and allow to derive population estimates by province and sector.

Fortunately, such 'counted' information does exist in the form of enrolled pupils at school, participants to a national school test and vaccinated children. Of course, this type of data is far from perfectly reliable or ready-for-use either, as they only consider a certain age group[5]. Yet, by performing the necessary mark-ups to represent the whole population, these data have been employed to construct an alternative and more recent set of population estimates[6].

More specifically, we first construct a series of population estimates (PE) *by province* for the year 2012 using the equation below. Basically, these PEs are obtained by averaging the population estimates derived from two different sources. The first source represents the accumulated total number of vaccinated children under 5 years of age (VAC) following the immunization campaign executed in 2012 by the Ministry of Health and UNICEF[7]. These numbers have then been adjusted using a province-specific coverage rate (COV) for the same target population (DHS, 2014: 143) in order to obtain an estimate of the under-five population in each province. These estimates have further been scaled up to obtain overall population estimates per province, by multiplying them with the population share of under-five children per province (U5/TOT) surveyed during the 123 Survey (2012). The second source of data used entails the number of children enrolled at primary school (PRIM) during the school year 2012–13 as registered by the Ministry of Education (République Démocratique du Congo, 2014b: 26). These figures are again scaled up using the 123 Survey (2012), this time by the population share of pupils declaring their enrolment in primary school (PUP/TOT). The result of these computations can be read in column (3) of Table 3.

$$PE_{p,2012} = \left[\left(VAC_{p,2012}\ x\ \frac{1}{COV_{p,2013}} x\ \frac{TOT_{p,2012}}{U5_{p,2012}}\right) + \left(PRIM_{p,2012}\ x\ \frac{TOT_{p,2012}}{PUP_{p,2012}}\right)\right]/2$$

$$for\ all\ provinces\ p = 1, \ldots 11$$

To obtain a 2012 estimate for the urbanization rate (UR) in each province (see equation below), we again make use of the urbanization rates comprised in the same vaccination data, complemented this time with data from the national End-Of-Studies test (EOS), sanctioning the cycle of primary education (République Démocratique du Congo, 2013b). The reason why another source of schooling data has been employed, relates to the detail it provides on the number of participants to the test within each major city in the country. As a matter of fact, in order to assure consistency over time, a more limited but common list of major cities has first been identified across all data sources (including the 1984 census) to first derive a preliminary urbanization rate (PUR) per province. Using the population share of these major cities within the urban sector (MCITIES/URBAN) as known in 1984, these preliminary urbanization rates have then been scaled up to represent the overall size of the urban sector in each province. Then, an average of both 2012 estimates has been computed to finally represent the degree of urbanization in each province, which can be read from column (4) of Table 3.

$$UR_{p,2012} = \left[\left(PUR\_VAC_{p,2012}\ x\ \frac{URBAN_{p,1984}}{MCITIES_{p,1984}}\right) + \left(PUR\_EOS_{p,2012}\ x\ \frac{URBAN_{p,1984}}{MCITIES_{p,1984}}\right)\right]/2$$

$$for\ all\ provinces\ p = 1, \ldots 11$$

By applying a simple accounting formula on the first four columns of Table 3, one obtains a set of average annual growth rates for each province and its rate of urbanization (see columns (5) and (6)).

For the interval 1984–2012, the Congolese population would have yearly grown, on average, with 3.5 percent while the share of the urban sector has annually increased with 0.7 percent. Compared to this national trend, some regional variation exists: both Kivu provinces and Orientale seem to have grown more slowly than the national average (between 2.8 and 3.2 percent), but have urbanized at a much faster pace (especially Nord-Kivu at 4.1 percent). This particular demographic evolution could be traced back to the war and continuing hostilities characterizing this part of the country, making survivors flee to urban areas in search for security[8]. An opposite tendency could be observed in Katanga and both Kasai provinces: higher-than-average population growth (around 4 percent) combined with an urban sector losing weight over the years (around -0.3 percent). This pattern can be attributed to the collapse of the formal (mining) economy within each of these (highly) urbanized provinces. A similar argument of industrial decline could be put forward with respect to Bas-Congo, housing the only two international sea harbors of the country and whose importance in attracting people to the city seems to have been negatively affected by the economic crisis as well.

While these region-specific growth rates by and large seem to make sense, further triangulation has been pursued by looking at election data from the national referendum in 2005. Contrary to the (partly boycotted) 2006 and the (fraudulent) 2011 elections, the national referendum was fairly well organized and reliable (MOEUE-RDC, 2006). More precisely, by comparing the number of eligible voters (assuming the population growth rates in column (5) to be correct) with the actual figures of enrolled voters as officially announced, one obtains another series of data to assess the quality of the 2012 population yardstick. As a matter of fact, the ratio of enrolled over eligible voters gives an idea of the spatial coverage of Congolese adults having obtained their election card. A political and infrastructural reading of this information then

allows for an additional check before moving to the next step of the technique. Following the last column of Table 3, not less than 83 percent of eligible people have been able to acquire a voting card for the referendum in 2005. Given the heavily dilapidated infrastructure of the country at that time, this percentage underscores the overall success generally attributed to this enrolment operation (Reyntjens, 2007: 311). It is exactly within this logic that most Congolese tried at all costs to obtain their voting cards, which were considered as identity cards. Despite the overall success, some marked variation in coverage rates can be observed across provinces.

<< Table 3 about here >>

To begin with, in Nord- and Sud-Kivu, coverage rates are substantially higher[9], which could be read as an act of appreciation on behalf of the population to those having ended the war. A similar reasoning could be made for Katanga and Orientale, though their coverage rates are much less pronounced compared to those of the Kivus. This observation could then perhaps be linked to the fact their involvement in the conflict has been spatially less widespread as well as to the remoteness and thus poorly accessible nature of these two provinces, which surely made voter registration all the more challenging. This final argument could be put forward as well when observing the lower coverage rates within the rural provinces of Bandundu, Equateur and Maniema: each of these provinces only registered a bit more than 70 percent of their eligible voting population, and all sharing the same challenges in terms of proximity. The contrary of this argument may be true for Kinshasa, where road infrastructure and agglomeration effects have probably contributed a lot to its successful enrolment rate of 91 percent. With respect to both Kasai provinces, and especially in the eastern part, the boycott issued by a prominent figure with a political stronghold in these particular provinces did not seem to have missed its effect (De

Saint Moulin, 2009: 54) as only 72 percent of eligible voters in East Kasai have actually obtained their voting card.

Even though these region-specific average growth rates derived for the period 1984–2012 seem to be acceptable, both in terms of qualitative triangulation as well as of how they have been quantitatively derived, the underlying data used together with the assumptions adopted are far from perfectly reliable either. As a result, the 2012 yardstick should not be taken as the single best reference; its sole objective is to provide for a more stable benchmark to derive alternative population estimates for each survey year.

3.2.Deriving corrected population estimates for each survey year

Given the average population and urbanization growth rates derived in the previous section, the next step entails the computation of corrected population estimates for each survey year by province and sector. This exercise is trivial and involves nothing more than an application of the region-specific growth rates to the corresponding population counts of 1984 over the exact period that separates each survey date from the census[10]. The result of these computations can be observed in percentage terms in Figure 4 which combines the provincial weight (panel a) and the rate of urbanization of each province (panel b) over time. Overall, this figure by construction largely reflects Table 3 with provinces displaying a higher-than-average population growth to see their share increase and vice versa (panel a)[11], and provinces characterized by a positive urbanization growth rate to see the weight of its urban sector accrue and vice versa (panel b). As such, the two panels of Figure 4 represent the stabilized versions of the two previous figures above, and comprise all information needed to correct design weights in the next step.

<< Figure 4 about here >>

3.3.Correcting household design weights

The last step in the proposed technique to reduce the impact of a volatile sampling frame is to correct the design weights attached to each observation within the studied household surveys. Following the previous section, corrected population estimates (CPE) have been derived by province and sector for each of the survey years. Dividing these estimates with the initial population estimates (IPE) as reflected by the survey data, one obtains the sampling correction factors (SCF) wherewith each initial design weight should be multiplied in order to stabilize the sampling frame over time according to the region-specific growth rates derived in section 3.1. The formula below is a more formal representation of the computation of these correction factors.

$$SCF_{psy} = \frac{CPE_{psy}}{IPE_{psy}}$$

$$for\ all\ provinces\ p = 1, \dots 11; all\ sectors\ s = urban, rural; all\ survey\ years\ y = 1, \dots 7$$

In total, 147 correction factors have been subsequently computed and imputed to the respective surveys[12]. To be sure, these factors only provide a correction at the level of each province's sector so that their overall population size aligns with the demographic alternative derived above: they cannot however correct for possible biases in the design weight proportions *within* each province's sector.

4. Sampling fluctuations versus real socio-economic changes

By means of the correction factors derived in the previous section, we can now retroactively apply the adjusted design weights to the data and seize its impact. To do so, we will focus on a set of socio-economic variables and assess how they have evolved over time, both under the initial and stabilized design weights. More specifically, we will try to estimate the level of noise

generated by the biased population frame and discuss a few cases where the introduction of corrected design weights is salient. Given the variation in survey methodologies across all national surveys, the number of common variables is rather limited. For the purpose of this analysis, we select seven variables on household durables and housing characteristics. This particular selection is not only driven by data availability across all surveys, capturing this information is also considered to be less affected by measurement error[13]. Moreover, in the absence of income or expenditures, this type of data has been extensively used for the construction of asset indices being viable alternatives to explain inequalities in all sorts of development outcomes as well as to serve as longer-term welfare proxies in their own right (Filmer and Scott, 2008; Harttgen and Klasen, 2012). In order to go beyond a mere methodological discussion and to shed light on the evolution in living standards in the DRC over the past 15 years, an asset index based on these seven ownership and housing variables is constructed.

The seven asset components retained for this exercise are household ownership of radio, car and television; type of drinking water; material used for roof and floor; and the number of sleeping rooms per person. After re-categorizing the housing characteristics to assure consistency across surveys, we opted for the polychoric version of principal components analysis (PCA) developed by Kolenikov and Angeles (2009). Compared to standard PCA, this technique relies on polychoric instead of Pearson correlations, and allows for the introduction of all sorts of data without introducing spurious negative associations for ordinal data while at the same time respecting their logical order. Applying this polychoric PCA technique to the data using alternatively the initial and stabilized design weights yields two sets of principal components. The first component of each set then represents the asset index, and captures 60.9 percent of

overall variation when using the initial sampling frame and 61.3 percent for the corrected version.

To assess how much noise is removed by applying the technique presented in this article, Table 4 displays the relative change in standard deviation of the asset index and each of its underlying components when shifting from the initial to the corrected design weights. As can be observed, most changes are negative (i.e. grey cells), pointing to an overall reduction in fluctuations. Especially, the housing characteristics (apart from the number of sleeping rooms) seem to be sensitive to the erratic sampling frame in most provinces. On the contrary, fluctuations in car ownership are only reduced in four provinces when design weights are stabilized, and these reductions are relatively more modest too. By and large, the reduction in volatility is reinforcing across each of the individual components as the overall asset index itself shows the largest reductions in standard deviation. This should not come as a surprise, knowing that urban households in general perform better in terms of household durables and housing quality compared to their rural counterparts and given the fact that the corrective measures introduced exactly intended to stabilize urbanization rates over time.

<< Table 4 about here >>

The effects of stabilizing urbanization rates across surveys can also be captured when looking at the evolution in urbanization and asset ownership (the latter as approached by the asset index), according to both the initial and stabilized sampling frames (Figures 5–6).

For the whole country (Figure 5), one could notice at least two important deviations. On the basis of the initial design weights (dashed line), asset ownership peaked in 2007 and again in 2012 as measured through the 123 survey. Yet, controlling for this biased sampling frame (solid

line), the first peak probably occurred already in 2005 after which asset ownership gradually increased between 2007 and 2013, without any culmination in 2012 but with an acceleration during the last year. By inspecting the variation in urbanization rates in detail, the two peaks in asset ownership under the initial sampling regime are probably much driven by an overestimation of the urban population for these same years. Indeed, both in 2007 as for the second 123 Survey (2012), the urbanization rate has been estimated markedly higher than for the previous and subsequent surveys (dashed bars). By consequence, instead of researching the drivers behind the peaks in 2007 and 2012, an analyst should rather examine what may have caused the particular asset drift in 2005 and its fallback in 2007; as most of the remaining trend is in line with the country's economic reconstruction after the conflict years (1997–2003). Although this type of examination goes beyond the scope of this article, the implications of Congo's biased sampling frame is evident and should be of much concern for policymaking. Take for instance the last data point. Depending on whether design weights have been corrected or not, those concerned with household wealth might get very contradictory information: while uncorrected sampling frames point to severe asset depletion between 2012 and 2013, many Congolese households seem to have actually experienced a drift in asset ownership over the same time period.

<< Figure 5 about here >>

Many similar, and even more pronounced, implications of this biased sampling frame could be observed when looking at the provincial level. Consider Figure 6 which displays trends in asset ownership and urbanization rates for the provinces of Bandundu (panel a) and Katanga (panel b). Again, when taking the initial sampling frame for granted, it is difficult to ignore any association between asset ownership and estimated urbanization: each time the urban sector was

projected to house more people compared to previous and subsequent surveys, each time the asset index peaked. Whereas for Bandundu the timing of both peaks coincides with the one for the DRC discussed above, the second peak in Katanga rather occurred in 2013, when urbanization would have increased from 31 percent to more than 45 percent. When removing the volatility in presumed urbanization, Bandundu's evolution in economic wealth is much less erratic. After a decline between 2001 and 2005, the asset index has remained more or less constant till 2010 when it increased again above its 2001 level, to remain stable ever since. The trend in Katanga on the contrary largely follows the one observed for the DRC, except for 2013 when no steep increase could be noticed. Again, a more elaborated analysis of these 'real' trends falls outside the scope of this article.

<< Figure 6 about here >>

To be sure, the extent to which the technique presented in this article was able to neatly isolate 'real' from methodological changes, of course depends on the validity of the underlying assumptions used. As a matter of fact, for some provinces or asset components the stabilized trends look as erratic as the initial ones—which may either point to a number of shortcomings inherent to the technique proposed or to other reasons and sources of spuriousness. To end this section, we put forward a number of hypotheses which may still complicate a straightforward comparison of the Congolese national survey data over time.

A first series of hypotheses relates to the quality and accuracy of the sources used to derive corrected population estimates. Given their construction, this entails three separate assumptions. First, it has been implicitly assumed that an intervention at the level of each province's sector would be sufficient to stabilize the country's sampling frame. However, if the core problem lies at a lower geographical level, for instance at the stage of selecting the primary sampling units

(city quarters, cities and chiefdoms) which is executed with probabilities proportional to population size, then our technique falls short in addressing any potential bias at this level. Second, no matter how meticulous the demographic yardstick for 2012 has been constructed to derive average growth rates per province and sector, one could not avoid a number of arbitrary choices (see section 3.1.). And third, the application of *average* rates in itself could be read as a violation of reality as populations do not necessarily grow at a flat rate. Here, it might be useful to refer to a debate on the exact death toll resulting from the war in eastern Congo (1997–2003). Whereas the assumption of average growth might still be acceptable when the death toll did not exceed 200,000 people, as estimated by Lambert and Lohlé-Tart (2008), it becomes more difficult to defend when the number of deaths would have amounted to more than 4.5 million people, as defended by Coghlan et al. (2008).

Another series of hypotheses which can explain the moderate effects prompted by the revised sampling frame may be linked to the empirical data on which they are applied. More specifically, the final impact of a biased sampling frame may be largely attenuated by effects of compensation and substitution. In theory, if less weight is assigned to a poor performing province while more weight is given to the poorest performing sector within that province, the overall impact of this particular province on the country's average might be completely offset. Similarly, if the performance of both sectors with respect to a certain indicator is more-or-less equal, then any change in the urbanization rate will not yield much of an impact. That these effects might be at play for the survey data on the DRC is evident from the asset index. Asset ownership in Kinshasa has increased from 1.7 to 2.1 between 2001 and 2013, thus leaving other provinces far behind. As a result of this, the capital's performance has an extensive impact on the overall trend in asset ownership compared to many other provinces, whose change in provincial weight will then only

have a minor effect on the aggregate total. Moreover, Kinshasa's urbanization rate has been kept constant at 100 percent, both under the initial and stabilized sampling regimes, an important issue which thus does not affect the trend in household wealth at all.

A final series of hypotheses jeopardizing any straightforward comparability of survey data entails all sorts of methodological differences beyond sampling. At various stages in the execution of a survey, seemingly trivial choices or slightly different approaches may have important consequences. For instance, the simple yes/no-query 'do you have a car?' should be accompanied with clear guidelines for poll-takers as people may falsely provide an affirmative answer when their neighbor or uncle owns a car which the interviewee may use at all time or when he actually possesses one that is urgently waiting for that essential spare part that is never to arrive. In these circumstances, to be on the safe side, it might be advisable to only compare surveys with similar methodologies. Unfortunately, even for surveys following the same methodology, questions have been slightly altered, making it nearly impossible to assure consistency over time. For example, with respect to assets, the second round of the 123 Survey followed a stricter definition by adding that household durables should be 'properly functioning', resulting in an underestimation of ownership compared to the previous round.

In sum, the technique elaborated in this article has certainly been able to downscale some of the erratic fluctuations produced by Congo's unreliable sampling frame, while it has been unable to control for other methodological biases. Identifying some of the other biases is certainly a first step in improving the current version of the technique and assure more meaningful analyses over time.

5. Conclusion

In this article, a lot of attention has been devoted to question the accuracy of the sampling frame used to make surveys on the DRC representative to the size and distribution of the Congolese population. In fact, this should not surprise as the last census dates back to 1984, giving ample room to each survey's sampling expert to come up with an alternative set of regional population estimates. Putting the estimates of each of the last seven national surveys together, it is nearly impossible to make real sense of the demographical fluctuations observed in both absolute and relative terms. Most salient in this respect are perhaps the changes in provincial weights and urbanization rates, which have proven to follow a highly volatile and erratic pattern over time. As a result, much information coming from these surveys runs the risk of being flawed.

Subsequently, in order to reduce the impact of Congo's biased sampling frame, this article has proposed a three-step post-stratification technique to restore comparability over time. Whereas the first step involved the computation of average growth rates per province and sector, the other two consecutively derived alternative population estimates for each survey year, on the basis of which design weights have then been retroactively corrected. Inevitably, this exercise required the construction of another and more recent demographic reference, for which data from vaccination campaigns, school enrolment and election rounds have been used. To be clear, given the imperfect data used and the assumptions adopted, this alternative series of population figures remains by and large an estimation and therefore should not be considered as a recount of the Congolese population.

Based on an asset index and its underlying components, the technique's impact to disentangle real socio-economic trends from spurious sampling variation has been illustrated using all seven surveys. Overall, the extent of fluctuations, as measured by a reduction is standard deviation, has been noticeably reduced after stabilizing the survey design weights. Moreover, in many instances

trends in asset ownership under the initial sampling frame have proven to be clearly driven by unreliable changes in estimated urbanization. In short, notwithstanding many other sources of spuriousness may still complicate any straightforward comparison of Congo's survey data, the technique proposed in this article has certainly removed an essential part of the erratic volatility introduced by the country's biased sampling frame. Among the other sources of spuriousness, one can either refer to the shortcomings or assumptions underneath the proposed technique or to other methodological challenges beyond sampling.

In policy terms, the importance of this sampling issue cannot be overestimated as policymakers are typically concerned with absolute numbers (of poor, ill, undereducated people, etc.), where they are located and how their situation has evolved over time—all aspects being directly influenced by the sampling frame. Therefore, in order to avoid that the next survey round would be biased from its very conception, it is crucial to finally get a more firm hold on these shaky population estimates, or at least to know exactly what has driven their volatility. Ultimately, the best way to deal with this issue would be the execution of a new census, which of course is an endeavor of quite a different kind but one currently making more sense than any other effort to shed light on the Congolese's living standard.

Endnotes

[1] A sample is unbiased and representative, respectively, if each unit of analysis is equally likely to be included in the sample and if it has the same composition as the population from which it is drawn.

[2] Here, the sampling frame is understood as the list of all sampling units together with their respective population estimates. As elaborated in section three, the technique proposed in this article involves the stabilization of population data without affecting the list of sampling units.

[3] Of course, other large-scale household surveys have been executed over the past 15 years, though their coverage is not always representative at the national level; like the Global Financial Inclusion Database 2011 (which excludes many areas in the east of the country) and the Comprehensive Food Security and Vulnerability Analyses of 2007–8 and 2011–12 (which only covers the rural sector).

[4] Even though the rural municipalities of Nsele and Maluku cannot really claim this label, the Kinshasa province as a whole is considered to be urban—which explains why Kinshasa has been dropped from this figure.

[5] Another important concern for the DRC entails the accuracy of data coming from the many conflict affected areas in the country.

[6] Inevitably, the information used to perform these mark-ups is based on available survey data (executed around the same time as the counted data), which in turn involves the risk of importing some of the imprecision associated with the erratic sampling frame. Yet, as the proposed mark-up procedure will employ province-specific data from variables of which most of the variation is captured by inter-provincial differences, the effect of applying a distorted weighting scheme is somewhat reduced. Apart from this empirical convenience, it should be stressed that the benchmark obtained should mainly be valued for its capacity to downscale the observed sampling fluctuations, rather than for being a correct proxy of the level and distribution of the Congolese population.

[7] As a matter of fact, the data issued by UNICEF and the Ministry of Health already converted immunization data into estimates of overall population, though without taking into account provincial variations in coverage rates and population shares of children under 5 years of age. As a result, the accumulated total number of children vaccinated by 2012 has first been obtained for each province before the formula above could be applied.

[8] Similar urban growth rates could be observed in other countries characterized by violent conflict, like Uganda (see for example Potts (2009)).

[9] Of course, coverage rates above 100 percent are theoretically impossible and may either point to an underestimation of true regional population growth between 1984 and 2005 or an over-registration of voters in these respective provinces.

[10] Given the accelerated sequence of surveys executed over the last years, population counts have been accrued on a monthly basis by considering the timing when most households have been surveyed during each survey round.

[11] For Nord-Kivu, given the marked difference in growth rates between its urban and rural sector, this association does not hold over the complete period: despite its lower-than-average growth rate (3.2 percent versus 3.5 percent), the provincial weight of Nord-Kivu in fact has *increased* after 2005.

[12] Seven less than 11x2x7 since Kinshasa has no rural sector.

[13] Although problems of interpretation can never be fully excluded, yes/no-queries on ownership or housing quality tend to be less problematic than say, for example, questions of self-assessment or household consumption.

References

Coghlan B, Brennan R, Ngoy P, et al. (2008) *Mortality in the Democratic Republic of Congo: An Ongoing Crisis*. New York: International Rescue Committee.

Deaton A (1997) *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Washington DC: The World Bank.

De Saint Moulin L (2009) Analyse du paysage sociopolitique à partir du résultat des élections de 2006. In: Trefon T (ed) *Réforme au Congo (RDC): Attentes et Désillusions*. Paris: L'Harmattan, pp. 49–65.

Filmer D and Scott K (2008) *Assessing Asset Indices*. Policy Research Working Paper, No. 4605. Washington DC: The World Bank.

Gelman A (2007) Struggles with Survey Weighting and Regression Modeling. *Statistical Science* 22(2): 153–164.

Harttgen K and Klasen S (2012) A household-based Human Development Index. *World Development* 40(5): 878–899.

IDMC (2014) *Democratic Republic of the Congo; Multiple crises hamper prospects for durable solutions*. Geneva: Internal Displacement Monitoring Centre, Norwegian Refugee Council.

INS (n.d.) Recensement Général de la Population et de l'Habitat (RGPH2): Questions/Réponses. Available at: www.ins-rdc.org/?q=content/questionsreponses (accessed 15 May 2016).

Kolenikov S and Angeles G (2009) Socioeconomic status measurement with discrete proxy variables: Is principal component analysis a reliable answer? *Review of Income and Wealth* 55(1): 128–165.

Lambert A and Lohlé-Tart L (2008) La Surmortalité au Congo (RDC) durant les Troubles de 1998–2004: Une Estimation des Décès en Surnombre, scientifiquement Fondée à partir des Méthodes de la Démographie. Available at: www.uclouvain.be/cps/ucl/doc/demo/documents/Lambert.pdf (accessed 15 May 2016).

Little RJ (2004) To Model or Not To Model? Competing Modes of Inference for Finite Population Sampling. *Journal of the American Statistical Association* 99(466): 546–556.

Marivoet W and De Herdt T (2014) Reliable, challenging or misleading? A qualitative account of the most recent national surveys and country statistics in the DRC. *Canadian Journal of Development Studies* 35(1): 97–119.

MOEUE-RDC (2006) *Referendum Constitutionnel 2005, Rapport Final*. Kinshasa: Mission d'Observation Electorale de l'Union Européenne en RDC.

Nunley AC (n.d.) African Elections Database. Available at: http://africanelections.tripod.com (accessed 15 May 2016).

Potts D (2009) The slowing of sub-Saharan Africa's urbanization: evidence and implications for urban livelihoods. *Environment & Urbanization* 21(1): 253–259.

République Démocratique du Congo (2002) *Enquête Nationale sur la Situation des Enfants et des Femmes, MICS2/2001*. Kinshasa: Institut National de la Statistique, UNICEF, USAID.

République Démocratique du Congo (2008a) *Enquête 1-2-3 (Phase I: Emploi, Phase II: Secteur Informel et Phase III: Consommation des Ménages) 2004-5*. Kinshasa: Institut National de la Statistique.

République Démocratique du Congo (2008b) *Enquête Démographique et de Santé, République Démocratique du Congo 2007*. Calverton: Macro International, Ministère du Plan.

République Démocratique du Congo (2011) *Enquête par Grappes à Indicateurs Multiples MICS-2010, Rapport Final*. Kinshasa: Institut National de la Statistique, UNICEF.

République Démocratique du Congo (2013a) *Rapport de l'Enquête Nationale sur les Enfants et Adolescents en dehors de l'Ecole (2012)*. Kinshasa: UNESCO, UKAID, UNICEF.

République Démocratique du Congo (2013b) *Rapport Général des Résultats du TENAFEP 2013*. Kinshasa: Ministère de l'Enseignement Primaire, Secondaire et Professionnel.

République Démocratique du Congo (2014a) *Enquête 1-2-3 (Phase I: Emploi, Phase II: Secteur Informel et Phase III: Consommation des Ménages) 2012-13*. Kinshasa: Institut National de la Statistique.

République Démocratique du Congo (2014b) *Annuaire Statistique de l'Enseignement Primaire, Secondaire et Professionnel, Année Scolaire 2012-2013*. Kinshasa: Ministère de l'Enseignement Primaire, Secondaire et Professionnel.

République Démocratique du Congo (2014c) *Deuxième Enquête Démographique et de Santé (EDS-RDC II 2013-2014)*. Rockville: Ministère du Plan et Suivi de la Mise en œuvre de la Révolution de la Modernité, Ministère de la Santé Publique, ICF International.

République du Zaïre (1991) *Zaïre, Recensement Scientifique de la Population, Juillet 1984, Totaux Définitifs*. Kinshasa: Ministère du Plan et Aménagement du Territoire, Institut National de la Statistique.

Reyntjens F (2007) Briefing, Democratic Republic of Congo: Political transition and beyond. *African Affairs* 106(423): 307–317.

Trefon T (2012) Population Census DRC. In: Congo Masquerade. Available at: http://congomasquerade.blogspot.sn/2012/05/population-census-drc.html (accessed 15 May 2016).

Yansaneh IS (2005) Overview of sample design issues for household surveys in developing and transition countries. In: Department of Economic and Social Affairs Statistics Division (ed) *Household Sample Surveys in Developing and Transition Countries*. New York: United Nations, pp. 11–34.