Wherein lies the difference in the estimate of the size of the South African small business sector?

Thesis Presented for the Degree of

Master of Commerce

in the

School of Economics

By

Boingotlo Gasealahwe

Supervisor: Professor Haroon Bhorat

February 2013

Wherein lies the difference in the estimate of the South African Small Business Sector?

Abstract

This paper investigates the difference of over 4 million business owners found in the estimates produced by the QLFS 2010:2 and the Finscope (2010) South Africa Small Business Survey. The paper discusses a number of possible reasons for the discrepancy and finds that the QLFS 2010:2 understates the number of small business owners, with the converse being true for the Finscope (2010) South Africa Small Business Survey. Even after accounting for differences in the inclusion criteria of the two surveys as well as the use of divergent sampling weights, a large part of the difference in the estimates produced is still unaccounted for. The paper also calls into question the reliability of the estimates produced in Finscope survey as well as the validity of the negative binomial sampling methodology used to conduct the survey. Access to additional data from both Statistics South Africa and Finscope, will further unravel this mystery and provide a better understanding of the sources of the differences in the estimates produced.

1. Introduction

Policy makers and researchers alike require accurate information about the size and nature of the small businesses if they are to conduct meaningful research and to effect welfare enhancing policy. Unlike the formal sector, however, South Africa does not have a register of small informal businesses and instead relies on data from a number of household surveys to derive a credible estimate of the size of the small business sector.

In the latest Survey of Employers and the Self Employed (SESE), a nationally representative survey that identifies individuals running businesses through the Quarterly Labour Force Surveys[1] (QLFS), only 1.1m of the possible 17m individuals participating in the labour force are identified as running any kind of business (SESE 2010). The Finscope 2010 Small Business Survey (FSBS) however, places the number of small business owners in South Africa at 5.6m (Finscope 2010). This difference of over 4 million business owners is no marginal figure, and the under or over count deserves a thorough investigation. Using data drawn from the QLFS 2010:2 and the FSBS this paper investigates the most likely sources of this discrepancy.

The paper begins with an outline of the problem and then goes on to review the literature on the size of the South African small business sector. In section 3, the paper discusses some of the probable reasons for the difference observed in the estimates of the small business sector and then proceeds to compare key business and owner characteristics across both surveys. The accuracy of the estimates is also explored followed by a thorough investigation of the distribution of the sampling weights across both surveys. The paper then concludes.

2. Problem Statement

[1] The technique employed for the SESE is also known as the "1-2" methodology (ILO 1993). In the first stage, data on the economic activities of individuals is obtained through the QLFS; a household based labour force survey already in existence with an existing sampling frame. The data collected through the QLFS then forms the sample frame for the second phase, which is a survey of all businesses that are not registered for Value Added Tax (VAT) and are thus excluded from the Stats SA business sampling frame

Small businesses owners in South Africa exist on a continuum, ranging from survivalist street vendors, to backyard manufacturers and service providers, the occasional home based evening job to more formal businesses (DTI 2008, SESE 2010, Herrington,Kew and Kew 2010). The majority of these businesses are not registered for VAT and possibly other taxes such as Pay as You Earn (PAYE) nor are they registered with other official business bodies meaning that there is no readily available data that would enable one to estimate the size of this sector (DTI 2008).

In order to estimate with any accuracy the size and extent of the small business sector in South Africa, one could draw a random sample of small business owners and then make use of raising weights to convert the sample counts into estimates of population totals. However, in the absence of a census for small businesses there does not exist a sampling frame on which to base these inflation weights. As a solution to this problem a number of household based surveys, for which a sampling frame exists, have been conducted to identify those individuals that are running a business at their place of residence or elsewhere. These surveys include the Annual Population Survey (APS), an annual survey of about 2 000 individuals between the ages of 18 and 64 run by the Global Entrepreneurship Monitor (GEMS) to measure the extent and nature of entrepreneurial activity in South Africa; the bi-annual Labour Force Survey, which is the predecessor to the QLFS, the QLFS, the SESE, and lastly the FSBS.

Table 1[2] provides a summary of some of the estimates that have been produced in other studies using the above-mentioned surveys. Taken together, the results strongly suggest that the number of small business owners has been stagnant over the past decade, declining in 2010 following the recent economic downturn. The results produced by the SESE however, suggest a rapid decline in the number of small business owners. Because of changes in the sampling and fieldwork methodology, Stats SA advises against a comparison of the SESE results, necessitating that the results be interpreted with caution.

The Finscope Small Business survey was conducted for the very first time in South Africa in 2010 and estimated the number of small businesses owners to be 5,6m. This

[2] See Appendix

paper, using data from the QLFS 2010:2 finds only 1,7m business owners in the same period, a difference of over 4m business owners. With the survey design of both surveys being almost identical, the vast discrepancy becomes even more puzzling. The sections that follow thoroughly investigate the sources of the discrepancy.

3. Possible sources of discrepancy

The QLFS and the FSBS both identify business owners by going directly to households, employing a multistage stratified sampling design with probability proportional to size (PPS) sampling in the first stage and systematic random sampling (SRS) in the second (see figure 1 and 2). With the exception of the final stage of selection, the survey design of the FSBS bears a strong resemblance to that of the QLFS. Table 2 summarizes the most salient points concerning the survey design and inclusion criteria of both surveys. Both the QLFS 2010:2 and the FSBS were conducted in the second quarter of 2010. The sampling frame for both surveys is based on the census 2001 enumerator areas (EAs), with census EAs being used as Primary Sampling Units (PSUs) in the QLFS. The EAs/PSUs are first stratified[3] along similar geographical dimensions followed by the selection of EAs/PSUs in each stratum using PPS sampling. Then, using (SRS), dwelling units within each EA/PSU are randomly selected into the sample using a fixed sampling interval. The FSBS does not exclude non South Africa citizens in its sample and is twice the size of the QLFS. The paragraphs that follow discuss some of the most likely sources of discrepancy between the two surveys.

3.1.1 Coverage

The sampling frame of any survey defines its potential coverage of the target population. Sampling frames are typically constructed by dividing the country into a number of small area units (enumerator areas) that cover the entire geography of the country. Most countries typically make use of one master sample, from which all household surveys can draw subsamples on which to base their sampling frames.

[3] Dividing the population into homogeneous subgroups (called strata) from which separate independent samples are selected.

In preparation for the 2001 population census the country was divided into 80 787 EAs with each EA consisting of anything from 1 up to more than 500 households. The FSBS benchmarked the census frame to the 2009 population estimates as published by Stats SA and used their in-house Geographical Information system to map out an updated EA list which spatially locates each household in order to ensure complete and current coverage of the country (Finscope 2010). The QLFS however, alongside other Stats SA household surveys uses census EAs as Primary Sampling Units (PSU's) for its master sample. In constructing this sample, EAs with a household count of less than 25 are excluded from the frame while those that contain between 25 and 99 households are combined with other EA's to form PSU's. Also, EAs with a household count of more than 500 are split (Stats SA 2008). Excluded EAs are accounted for with the use of sampling weights.

With the number of households as the measure of size, the FSBS and the QLFS draw a sample of EAs based on PPS in the first stage. This means that larger EAs have a higher probability of appearing in the sample. With large EAs left as is in the FSBS, FSBS EAs are more likely to be larger than QLFS EAs and will also provide greater coverage of the small business sector than the QLFS, since no EAs were excluded from the frame.

3.1.2. Inclusion Criteria

Small businesses can be classified as micro, very small, small or medium enterprises (SMME's) according to their industry, size, turnover and asset value as defined in the National Small Business Act of 1996. The FSBS identified all individuals above the age of 16 who perceive themselves as business owners and employ less than 200 business owners as eligible for inclusion in their sample (Finscope 2010). The QLFS however, uses an identification criterion that is in line with ILO guidelines for the identification of the informal self-employed (ILO 1982). Specifically, the QLFS sample includes individuals aged 15 years or older whose main job, business or economic activity, as identified by where they work the most hours per week is an employer [4] (employing one or more employees) or an own-account worker[5] (not

[4] Employer: (employing one or more employees) a person who operates his/her own economic enterprise or engages independently in a profession or trade, and hires one or more employees

employing any employees). Respondents are asked to report all economic activities undertaken even for only one hour in the past week, ensuring that even those who work for just a few hours a week are more likely to report their activities and as well as provide detailed information about main work activities. Previous surveys failed to correctly categorize this segment of the labour force, capturing them as unemployed, thus leading to an undercount of those who are self-employed (Muller 2003).

The biggest difference in the inclusion criteria between the FSBS and the QLFS is the the treatment of non South African citizens across the two samples. Foreign households are considered ineligible for inclusion in the QLFS, but form part of the count in the FSBS (Stats SA 2008 pg3). Non South African citizens make up for 17% of the Finscope sample. Accounting for just under 1m of the total estimate, these individuals provide some explanation for the observed discrepancy. This point is returned to later in the paper.

Additionally, the FSBS identification criteria is subjective. It is based on the self-reporting of individuals, a methodology which may be subject to some non-random error. On the other end of the spectrum, the QLFS criterion limits the identification of business owners who run a business as a secondary form of employment. Whilst the QLFS establishes whether an individual is engaged in multiple forms of employment, detailed questions are only asked about their main economic activity and not their secondary employment, making it difficult to classify this type of employment. For example, an individual with a typical 9-5 job who runs an informal business on the side and perceives themselves as a business owner will be counted as a business owner in the FSBS and not in the QLFS because the QLFS does not classify secondary employment. This is likely to lead to an underestimation of the informal sector in the QLFS. Muller (2003) asserts that this inability to classify secondary emplyment as a result of the failure to capture detailed information on individuals secondary employment may result in an underestimation of informal sector size. However, with only 6%, and less than one percent of individuals in the FSBS and QLFS respectively reporting additional income from another job, this problem is unlikely to account for a huge fraction of the discrepancy. In fact, the majority of

[5] Own-account worker (not employing any employees): a person who operates his/her own economic enterprise or engages independently in a profession or trade, and hires no employees. (Stats SA 2008)

respondents in the FSBS who reported additional sources of income beyond their businesses derived this income from their spouses, other family members as well as government grants.

Lastly, the QLFS counts individuals from the age of 15 while FSBS counts from the age of 16. With the youngest business owner in the QLFS being 17 years of age this difference cannot cause the discrepancy.

3.2. Sampling Weights

Almost all household surveys make use of weights to compensate for differences between the target population and the chosen sample. The differences are typically the result of unequal probabilities of selection of the sample units, non-coverage of the target population and non-response. In addition weighting is used in poststratification[6] to adjust the weighted sample distribution of certain variables to make it conform to a known distribution (Wittenberg 2009).

Sample weights in their simplest form, are just equal to the inverse of the inclusion probability of the sampling unit. In both the FSBS and the QLFS, the inclusion probability of any unit consists of three parts, namely:

1. The inclusion probability of the **EA**
2. The inclusion probability of the **household**
3. The inclusion probability of the **business owner**.

3.2.1. Inclusion Probability of EA

In theory, the probability of an EA being selected is equal to the probability of the EA appearing on the StatsSA master sample multiplied by the probability of being selected from the master sample. All EAs with less than 25 households had absolutely no chance of making it into the sub-sample of PSU's selected for the QLFS. Having used the original 2001 census master sample, large EA's, which were split for the QLFS and not the FSBS, had a much greater chance of being selected in the FSBS since selection is based on probability proportionate to size. The QLFS intentionally splits these EAs to ensure that these EAs are not selected with a higher probability

6 Calibrating the weights to get the sample to reflect population totals.

(QLFS 2008). This difference in methodology is likely to result in smaller EA weights in the FSBS, and thus a comparatively lower estimate.

3.2.2. Inclusion Probability of household

The probability of including a certain household in a sample depends on it being drawn from the reduced list of qualifying EAs/PSUs and the household consenting to be interviewed. In both the FSBS and the QLFS, households were systematically chosen to appear in the sample. In theory each household has an equal probability of inclusion, but not every possible sample of households has the same probability of inclusion due to the random sampling interval.

3.2.3. Inclusion Probability of Business Owner

The probability of inclusion of a business owner depends on the household being selected for inclusion, and the individual qualifying as a business owner. Once a household has been identified for inclusion in the QLFS, all individuals who are identified as business owners have a 100% probability of inclusion in the sample.

The FSBS however, uses the negative binomial listing approach to identify six business owners in six qualifying households per EA. The negative binomial distribution expresses the number of failures occurring while waiting for a fixed number of successe, where there is a known probability of success with each independent trial (Johnson, Kotz & Kemp 1992). For example, the distribution of the number of coin flips required to achieve 2 heads (which can take on any integer value between 2 and infinity in this example) is said to follow a negative binomial distribution. This distribution can formally be expressed as:

$$p(n,k,p) = \binom{n+k-1}{n} p^k (1-p)^n$$

where n= r-k is the number of failures

r = the number of independent trial

k= the number of successes

p= probability of success with each independent trial and $n \geq 0$

In the context of the FSBS, the number of households visited before six qualifier households are identified per EA forms a negative binomial distribution. This approach is underpinned by the following statistical properties that characterize the negative binomial experiment:

1. The experiment consists of *r* repeated trials.
2. Each trial can result in just two possible outcomes: "success" or "failure".
3. The probability of success, *p*, which is the same on every individual trial.
4. The trials are independent i.e. the outcome on one trial does not affect the outcome on other trials.
5. The experiment continues until *k* successes are observed, where *k* is specified in advance (Johnson, Kotz & Kemp 1992)

The negative listing approach is put to the test below:

1. **Repeated trials:** Recall that within each selected EA, six qualifier households are selected systematically, selecting every k^{th} household. Thus every k^{th} households that was interviewed represents a trial.

2. **Possible Outcomes**: Within selected households one of two outcomes can occur: a business owner is not found (failure) or one or more business owners are found (success).

3. **Constant Probability of Success**. The probability of finding a business owner per EA is equal to the proportion of business owners in each EA. This information can be obtained from the listing exercise conducted by the Finscope team.

4. **Independent trials**: A critical assumption underpinning the negative binomial distribution is that each draw is independent. Sampling without replacement however, automatically means that the draws are not independent. Returning to the earlier example; flipping "heads" on your first toss of a coin has no bearing on the outcome of your second, third and fourth toss and so forth because each toss is independent.

In the context of the FSBS, sampling six business owners per EA with replacement would mean that within each EA you first pick a household from the newly constructed sampling frame, interview it to determine if it is a qualifier household, then put the household back into the draw. You then draw another household, ascertain whether it qualifies and put the household back into the draw. You repeat this exercise until you find six qualifier households per EA. In this case, each household has an equal probability of selection with each draw that is independent from the last draw. Sampling with replacement also means that each household has the possibility of being selected more than once during the sampling process.

However sampling without replacement means that if you sample six households per EA without replacement, you first draw a household from the list, set it aside, and then pick another. If for example there are 10 households per EA, the first selected household has a 1/10 probability of selection and the second a 1/9. Sampling without replacement thus means the draws are not independent, since the first draw affects the second and so on. By the time a FSBS enumerator gets to the last household in an EA with no business owners at that point, they know with certainty which household is next and that they will not reach the target of six business owners per EA.

5. **The experiment continues until a fixed number of successes have occurred.** The goal was to achieve six interviews with small business owners (i.e. six successes) per EA. The assumption is that there are an infinite number of repeated trials in which to do so. The number of trials per EA is constrained by the number of selected households per EA. In the case were 6 business owners could not be found per EA, the FSBS sampled at random and added additional EA's to the sample.

3.2.4. Non-Response

Eligible households in any sample can typically be divided into two response categories: respondents and non-respondents. By the end of the FSBS, data was collected from 1075 EA's yielding 6450 potential respondents. However, only 5676

across 1000 EA's respondents were included in the final sample and the 774 potential respondents were excluded either due to refusal to participate in the interview or to unavailability. FSBS gave no indication of any adjustments for non-response in the calculation of the weights.

The QLFS 2010:2 report provides information on the response rates across the survey by province. Imputation is used for item non-response and sampling weights adjust for non-respondent households (refusal, no contact) and those that were excluded in the PSU master sampling frame.

3.2.5. Postratification

The sampling weights for both the FSBS and the QLFS were benchmarked to known population estimates for demographic variables such as age, gender, race, geography-type and province estimates so that they could be representative of the demographic of the South Africa population.

3.2.6. Sampling Error

The QLFS and the FSBS survey are both sample surveys, and are therefore subject to sampling variability. Even in the absence of all other sources of discrepancy, this would result in differences in the two estimates.

The discussion above however, suggests that differences in the inclusion criteria of both surveys as well as the sampling weights in use are likely to be the major sources of discrepancy between the two samples. The FSBS includes non South African business owners in its sample. If one counts the number of all businesses in South Africa irrespective of nationality, then the QLFS is going to understate the total number of businesses in South Africa by excluding foreigners, especially if these businesses account for as large a fraction of small businesses in the country as found in the FSBS. Although the QLFS excludes EAs with less than 25 households from its sample, it is likely to only marginally understate the number of business owners as excluded EAs are likely to be located in the national parks, desert areas and some rural areas, all of which are largely under populated. In addition, the inability of the QLFS to classify secondary employment is also unlikely to cause the discrepancy since less than 1% of individuals in the sample have more than one job.

The use of the negative binomial sampling approach in the FSBS is likely to have crucial implications for the calculation of the sampling weights that are in use. For the negative binomial formula to be correct, one needs to have an infinite number of independent trials. This however is not the case if you have a finite population that is sampled without replacement. In practice, surveys of this nature are usually sampled without replacement as there is no need to collect information more than once on the same unit and because the sample sizes of these surveys is typically large. In large sample sizes, sampling without replacement is approximately the same as sampling with replacement. The same however, cannot be said for small EA sample sizes. The section that follows discusses the main differences between the two surveys.

4. Main Differences Between the FSBS and the QLFS 2010:2

The inclusion of non-citizens as well as the use of the negative binomial distribution has proven to potentially be the most influential points of departure between the two surveys. In this section, the implication of the inclusion of foreign nationals as well as the use of the negative binomial sampling approach is discussed in further detail.

4.1. The inclusion of non South African Citizens.

Appreciating the fact that there may be a significant number of foreign owned small businesses in South Africa, despite official reports placing the number of foreign nationals at roughly 3-4% of the total national population one may look the number of all business owners in South Africa irrespective of their nationality as a first potential solution to the problem at hand (Schachter 2009, Polzer 2010). The fact that there are 17% of foreign owned businesses in the FSBS sample means that either there are substantially more foreigners than 3-4% in the country or that the probability of owning a business is much higher for foreigners or some combination of these two. Indeed Schachter (2009) and Polzer (2010) concede that owing to the scarcity and quality of migration data, the number of migrants in South Africa may be severely underestimated. If however, it is indeed the case that foreign owned businesses account for a huge fraction of all small businesses, excluding them in the count, as does the QLFS, will severely underestimate the number of small businesses that are in existence in the country.

4.2 Negative Binomial Sampling

4.2.1. Implications of Negative Binomial Sampling on Weights

"The validity of any survey depends on the statistical reliability of the sampling framework" (Finscope 2010). The sampling technique described by the FSBS fails to satisfy two crucial properties of the negative binomial approach, thus calling into question the reliability and validity of the approach. The negative binomial distribution is characterized by an infinite number of independent trials and this does not hold in the case of the FSBS. In calculating the final sampling weights, the number of 'failures' has to be taken into account. If the number of 'failures' is incorrectly calculated, the weights applied will also be incorrect, resulting in misleading weights and biased estimates.

Additionally, the replacement EA's have implications for the inclusion probabilities of the other EA's and households already discussed above and appropriate adjustments would need to be made. Furthermore, business owners across qualifying households have different inclusion probabilities because in the event of more than one business owner per household, the FSBS team selects a respondent at random (the qualifying individual with a birth date closest to the date of the interview was selected to be interviewed). This makes the calculation of the weights an intricate and sensitive exercise.

4.2.2. Implications of Negative Binomial Sampling on Survey Design

The FSBS survey design is almost identical to the QLFS, in that EAs/PSUs are selected based on PPS in the first stage of sampling and households in the second stage based on systematic random sampling. Having selected a sample of households, the FSBS constructs a detailed demographic list of every member of every household in each sampled EA. This allows them to identify all business owners in every household per EA, which is essentially what the QLFS does. Having done so, the team then uses the negative binomial distribution to randomly sample six business owners per EA with the use of the newly constructed sampling frame of business owners. With the validity of this approach in question, it begs the question why the

Finscope team decided to go ahead with it anyway because even if the approach was statistically valid, it is not practical nor is it the most efficient use of their time and resources. Assuming that each of the 1000 selected EAs had 250 households in them each with 5 household members, the Finscope team would have conducted 1250 interviews per EA and 1 250 000 interviews in total in order to list the demographic details of every household member while at the same time identify business owners. They would then construct a universe of small business owners, only to go back "door to door" to these same households and essentially repeat the same exercise, only this time for the sake of the negative binomial distribution. The list in theory eliminates the need for this additional step as the sample counts from this subsample can be used to extrapolate population totals (UN 2009). This approach does not impose the number of business owners per EA and is more likely to be representative than a sample drawn using the negative binomial distribution. An EA with no business owners is just as informative as an EA with only 6 or 5 or any number for that matter, and preserving this information is essential for producing credible results.

5. Exploring the Discrepancy

These sections explores the data contained in the two surveys, and where appropriate the SESE 2009, to determine whether the two surveys generally produce broadly different results, or whether the differences lie in the weights that are used across both surveys. Unfortunately a number of the possible sources of discrepancy that are listed in section 3 cannot be explored by looking at the data. The Stats SA's EA list or the one used by Finscope is not readily available. The data from the Finscope listing exercise and the negative binomial sampling is also unavailable. What is however explored using the available data, is the accuracy of the estimates produced in both surveys, the differences in the inclusion criteria across the surveys as well as the distribution of the final weights that were applied across both surveys.

5.1. Precision of the Estimates

In table 3, the precision of the estimates produced by both datasets is tested. The results show that despite the fact the QLFS is almost half the size of the FSBS, the FSBS estimate has a standard error of more than ten times that of the QLFS, thus calling into question the reliability of the Finscope estimates.

To further probe the efficiency of the estimates produced by both estimates, I look at the design effect (DEFF) on the estimates of the mean age across the sample in Table 4. The design effect provides a measure of the precision gained or lost by the use of the more complex survey design as opposed to a simple random sample (SRS), by comparing the variance obtained under the complex survey design and the variance that would have been obtained through SRS with the same number of observations (Salganik 2006). Stratification often increases the precision while clustering (i.e. aggregating households into EAs/PSUs) does the opposite. Both surveys have a DEFF greater than 1, implying that both suffer from a loss in precision as a result of the choice of survey design. The design effect of 4,3 and 1,6 respectively means that as a result of the complex survey design, the precision of surveys is only as good as a SRS of 1303 and 1587 observations respectively.

5.2. Profile of Business Owners and Businesses

I further probe the two surveys by doing a comparative analysis of some of the key statistics pertaining to small businesses. In table 5 a basic profile of the average business owner is constructed.

Both surveys tend to agree that the majority of business owners are black, have some form of secondary level education and that the probability of business ownership peaks in mid-life. These findings are consistent with those found in other studies, with the GEM 2010 report also finding that Gauteng province houses the most number of business owners and that the 25-44 year age group is the most entrepreneurially active (DTI 2008 Herrington, Kew and Kew 2010). This age cohort typically benefits from having worked in the same industry prior to starting their business (Herrington, Kew and Kew 2010). The two surveys however, differ in the raw estimates as well as the gender profile of business owners. Although the share of female business owners has increased over the last decade, the GEM has consistently found that more men are engaged in entrepreneurial activity than women. However, the share of female business owners increased from 40% in 2009 to 46% in 2010 (Herrington, Kew and Kew 2010). Using data from the LFS, the Department of Trade and Industry (DTI) finds that the share of female business owners averages between 43% and 48% during

March 2005 and March 2007 (DTI 2008). The SESE like the FSBS however, finds that more women operate small businesses than men (SESE 2010).

As a further robustness check, I profile the actual businesses in table 6 and find similar results to those found in table 5 as far as levels are concerned. With many business owners in both surveys citing unemployment as one of the main reasons for starting a business, it is no surprise that business owners only own one business with more than two thirds of these creating employment opportunities only for the owner. Again these results are consistent with those found in other studies. However, the result that more that 59% and 54% of businesses in the FSBS and the SESE respectively are older than 3 years is quite surprising as it differs from other studies. The DTI, Herrington, Kew and Kew (2010) and Mahembe (2011) all find that the majority of small businesses are survivalist enterprises that are still in the start-up phase, with most lasting only for 1.5 to 2.5 years. While the broad trends between the two surveys are consistent, there are more nuanced differences.

5.3. Sampling Weights

Following the evidence in the previous section, I now turn to investigate the nature of the raising weights that are applied in both surveys. A look at the distributions of the sampling weights in the first panel of figure 3 reveals that the distributions of the QLFS and the SESE are both right shifted when compared to the FSBS. The two distributions are also more tightly centered around the mean, unlike the FSBS distribution, which is dispersed over a much wider range. The QLFS and SESE weight distributions are almost identical with the mode of the SESE being the highest. The FSBS weight distribution is distinguished by its long upper tail, which extends far beyond that of the QLFS so that the average FSBS weight is higher than the average QLFS weight. These long FSBS tails suggest that more of the variance observed in the FSBS sampling weights could be the result of infrequent extreme deviations, as opposed to frequent modestly sized deviations.

Indeed we see this in table 7. The top percentile of the FSBS survey contributes roughly 16.7% to the final estimate as opposed to just 5.7% in the QLFS. In fact non-citizens account for three quarters of all observations in the top percentile. The second

panel of figure 3, clearly illustrates how the inclusion of non citizens further extends the long FSBS upper tail. This confirms the earlier hypothesis that it is hard to sample foreigners and to get them to respond, so that when you do, the ones you capture have higher weights. This is worrisome, given that this 1% of the sample contributes more than the bottom half of the sample to the total.

Consistent with previous findings, I find in table 8 that the FSBS weight distribution has a much larger variance with the weights being distributed over a much wider range than the QLFS and disproportionately so, again indicating the low precision of the FSBS estimates. This is seen quite clearly in figure 4, which provides a graphical presentation of the results found above. For every percentile below the 50th percentile, the values for QLFS weights are greater than those for FSBS and the converse is true above the 50th percentile. The increase in weights is far steeper in the FSBS with the gap between the two surveys widening quite steeply (Table 8).

Additionally, in table 9, I calculate a Gini coefficient (a measure of inequality) on the weight distribution to determine the extent to which the weights are unequally distributed across individuals in the FSBS and the QLFS. The QLFS Gini coefficient is 0.39 while that of the FSBS 0.74 for the entire sample and increases marginally to 0.75 after the exclusion of foreign observations. These findings provide strong evidence for the support of the hypothesis that there are quite a number of extreme outliers i.e. a few observations that carry a heavy weight. It is therefore not surprising that the average FSBS weight is larger than the average QLFS weight, nor that the FSBS estimate is far greater than the QLFS estimate. This despite the fact that QLFS weight distribution is more right shifted than the FSBS.

5.4. Impact on the numbers

Taken together, the results above suggest that the inclusion of foreigners, coupled with extreme sample weights may only partly explain the differential between the FSBS and the QLFS. Table 10 summarizes the impact of these differences on the final FSBS estimate. The inclusion of foreigners and deviant weights accounts for just over 1million business owners, leaving a large part of the difference still unaccounted for.

6. Conclusion

This paper makes use of the QLFS 2010:2, the Finscope Small Business Survey survey and the SESE 2009 to investigate the source of the discrepancy in the estimated size of the small business sector. Although the paper finds a number of possible reasons for the observed discrepancy, differences in the sampling weights as well as the inclusion criteria of non South African citizens have proven to be the most likely source of discrepancy between the two surveys. The paper finds that either the QLFS is underestimating the total number of businesses in South Africa by excluding foreigners or that the QLFS and the FSBS are measuring different things- one measuring South African businesses and the other all businesses. The disproportionately large sampling weights in the FSBS also suggest that the final FSBS estimate could be slightly overstated.

However, despite accounting for some of the key differences between the two surveys the final estimates of the FSBS and the QLFS still differ quite significantly. Even if one attempts to include the excluded foreign households in the QLFS, the difference is still not nearly explained. One can thus only conclude that the number of small business owners in South Africa lies somewhere between 2 and 6m. Access to additional data from both Statistics South Africa as well as Finscope, will help to further unravel this mystery and provide a better understanding of the differences in the estimates produced.

7. Reference List

Finscope (2010). "Finscope South Africa Small Business Survey 2010" Johannesburg.[online] Available at: http://www.finscope.co.za/new/pages/Initiatives/SmallBusiness.aspx?randomID=b6af2bb2-5289-4efc-9b14-3f8145241714&linkPath=3&lID=3_3 [accessed 1 June 2012]

International Labour Organisation (1982) "Resolution concerning statistics of the economically active population, employment, unemployment and underemployment" Thirteenth International Conference of Labour Statisticians (ICLS), Geneva.

Herrington, M., J. Kew and P. Kew (2010). *"*Global Entrepreneurship Monitor South African Report 2010*"* The UCT Centre for Innovation and Entrepreneurship, Cape Town.

International Labour Organisation (1993) "Resolution concerning statistics of employment in the informal sector" Fifteenth International Conference of Labour Statisticians, Geneva.

Johnson, N. L., Kotz, S. and Kemp, A. W (1992) "Univariate discrete distributions" Second Edition: John Wiley & Sons, p199-206.

Maas, G and Herrington, M (2007). "Global Entrepreneurship Monitor South African Report 2006*"* Cape Town: GEM. [online] Available at: http://www.gemconsortium.org/docs/602/gem-south-africa-2006-report [accessed 1 June 2012]

Mahembe, E. (2011) "Literature Review on Small and Medium Enterprises' Access to Credit and Support in South Africa". Prepared for the National Credit Regulator (NCR). Compiled by Underhill Corporate Solutions (UCS). Pretoria, South Africa

Muller, C (2002) "Measuring South Africa's Informal Sector: An Analysis of National Household Surveys" University of Natal, Durban.

Polzer, T. (2010) "Population Movements in and to South Africa" Forced Migration

Studies Programme. University of Witswatersrand. [Online] Available ar: http://www.cormsa.org.za/wp-content/uploads/2010/07/fmsp-fact-sheet-migration-in-sa-june-2010doc.pdf [accessed 1 June 2012]

Republic of South Africa. (1996) National Small Business Act, 1996: NO. 102 OF 1996. South Africa

Salganik, M.J. (2006) "Variance Estimation, Design Effects, and Sample Size Calculations for Respondent-Driven Sampling" Journal of Urban Health 83(Suppl 1): 98–112.

Schachter, J.P. (2009) "Data Assessment of Labour Migration Statistics in the SADC Region: South Africa, Zambia and Zimbabwe" International Organisation for Migration, South Africa.

Statistics South Africa (2008) "Guide to the Quarterly Labour Force Survey, August 2008". Report number: 02-11-01. Pretoria. Statistics South Africa

Statistics South Africa (2008) "Concepts and definitions used in the Quarterly Labour Force Survey, August 2008" Report number: 02-11-01. Pretoria. Statistics South Africa

Statistics South Africa (2010) "Quarterly Labour Force Survey: Quarter 2, 2010" Statistical Release P0211. Pretoria. Statistics South Africa

Statistics South Africa (2010) "Survey of Employers and Self Employed 2009" Statistical Release P0276. Pretoria. Statistics South Africa

Statistics South Africa (2006) "Survey of Employers and Self Employed 2005" Statistical Release P0276. Pretoria. Statistics South Africa

Statistics South Africa (2002) "Survey of Employers and Self Employed 2001" Statistical Release P0276. Pretoria. Statistics South Africa

Statistics South Africa (2010) "Survey of Employers and Self Employed 2009" Statistical Release P0276. Pretoria. Statistics South Africa

The Department of Trade and Industry (2008) "An Annual Review of Small Business in South Africa 2005-2007". DTI, South Africa.

Wilson, L.T., and Room, P.M. (1983) "Clumping Patterns of Fruit and Arthropads in Cotton, with implications for binomial sampling" Environmental Entomology (12): 50-54

Wittenberg, M (2009) "Weights: Report on NIDS Wave 1" Technical Paper No. 2, NIDS, University of Cape Town

United Nations (2009) "The "1-2" Survey: A data collection strategy for informal sector and informal employment statistics" Working Paper No. 1. United Nations ESCAP Statistics Division

Appendix

Table 1: Estimates of no of small businesses 2001-2010

'million	2001	2005	2006	2010
SESE	2,3	1,7	N/a	1,1
LFS	1,9	1,8	1,9	1,7
GEM	N/a	1,7	1,8	1.8
FSBS	N/a	N/a	N/a	5,6

Sources:

SESE 2002-2010 (Statistics South Africa), Muller (2003): LFS 2001:1 (Statistics South Africa)

DTI (2008): LFS 2005:2 and LFS 2006:2 (Statistics South Africa), Finscope(2010)

Maas and Herrington (2007): Global Entrepreneurship Monitor South African Report 2006

Herrington Kew and Kew (2010): Global Entrepreneurship Monitor South African Report 2010

Own Calculations: QLFS 2010:2 (Statistics South Africa)

Table 2: Summary of Differences between QLFS and FSBS

	Finscope 2010	**QLFS 2010: 2**
No of business owners	5 579 767	1 656 648
Definition of Small Businesses owner	16 years or older, self perceived business owner running a business with less than 200 employees	*An employer* (employing one or more employees) or an *own-account worker* (not employing any employees) above the age of 15
Survey design	**Stage 1:** EA's stratified by province and Geotype[7] and a sample of 1000 EA's was drawn based on 2001 census and benchmarked against 2009 stats was used, EA's selected based on PPS **Stage 2:** *negative binomial approach* used to systematically identify 6 qualifying households[8] per sampled EA. A qualifying hh, was one with one	**Stage 1:** Use census EA's as PSU's for master frame, which is stratified by province, within province by metro level, and by geotype within metro. A sample of 3080 PSU's was drawn. EA's selected based on PPS **Stage 2:** Sampling of dwelling units [9](DUs) with systematic sampling. Business owners identified within these

[7] The four geography types are: *urban formal, urban informal, farms and tribal*

[8] All people who live/ stay together for more than for nights a week.

[9] A structure, part of a structure or group of structures that can be lived in by a household

	or more business owners.	households.
Time survey conducted	April to May 2010 2nd Quarter 2010	April to June 2nd Quarter 2010
Sample size	5676	2659

Sources: *Finscope (2010), QLFS 2010:2 (Statistics South Africa). Own Calculations*

Table 3: 95% Confidence Interval on Estimates

	Estimate	Std. Error	Lower Limit	Upper Limit	n
Finscope	5579767	434210,2	4727697	6431837	5676
QLFS 2010:2	1632824	32810,5	1568456	1697192	2617

Sources: *Own Calculations: Finscope (2010) and QLFS 2010:2*
Notes: *The final QLFS estimate is in fact 1 656 648. The estimate provided above was run on a smaller sample due to data constraints.*

Table 4: Design Effects on Mean Age Estimate

	Mean	Std. Error	DEFF	n	effective n
QLFS 2010:2	42.58424	0,299	1,6	2617	1587
Finscope	41.15551	0,367	4,3	5659	1304

Sources: *Own Calculations: Finscope (2010) and QLFS 2010:2*

Table 5: Biographical Profile of Business Owners

	Finscope 2010		QLFS 2: 2010	
Average Age	41.15		42.25	
Age range	16-94		17-89	
Age-Groups	**5 579 767**	**100%**	**1 656 648**	**100%**
16-24	615 448	11%	70 242	4%
25-34	1 228 107	22%	412 505	25%
35-44	1 616 458	29%	466 512	28%
45-44	1 189 606	21%	420 292	25%
55-64	626 608	11%	220 169	13%
65 +	303 539	5%	66 929	4%
Gender	**5 579 767**	**100%**	**1 656 648**	**100%**
Male	2 318 951	42%	1 044 517	63%
Female	3 260 816	58%	612 132	37%
Population Group	**5 579 767**	**100%**	**1 656 648**	**100%**
Black	4 661 337	84%	1 176 883	71%
White	427 410	8%	346 074	21%
Coloured	291 264	5%	71 567	4%
Asian/Indian	199 756	4%	62 124	4%

Education	**5 579 767**	**100%**	**1 656 648**	**100%**
No schooling	163 487	3%	92 772	6%
Some primary completed	472 048	8%	209 732	13%
Primary completed	666 224	12%	104 866	6%
Secondary not completed	2 403 764	43%	570 053	34%
Secondary completed	1 369 275	25%	376 059	23%
Tertiary	427 410	8%	275 335	17%
Other	77 001	1%	27 832	2%
Province	**5 579 767**	**100%**	**1 656 648**	**100%**
Eastern Cape	828 595	15%	174 114	11%
Free State	446 939	8%	95 589	6%
Gauteng	1 278 883	23%	498 320	30%
KwaZulu-Natal	770 008	14%	319 070	19%
Limpopo	545 701	10%	182 563	11%
Mpumalanga	386 120	7%	127 396	8%
Northern Cape	154 002	3%	17 560	1%
North West	718 116	13%	70 408	4%
Westen Cape	450 845	8%	159 535	10%
n	5676		2659	

Sources: *Own Calculations: Finscope (2010) and QLFS 2010:2*

Table 6: Distribution of Business Characteristics

	Fisncope 2010	**SESE 2009**
Business size	**100%**	**100%**
0 employees	67%	81%
1 employee	14%	9%
2 employees	8%	4%
3 employees	3%	2%
4+ employees	7%	2%
Business Age	**100%**	**100%**
<12 months	10%	21%
1-3 years	32%	25%
3-5 years	22%	17%
5-10 years	16%	18%
10+ years	21%	20%
No of businesses	100%	100%
1	95%	97%
2	4%	2%
3	1%	1%
4+	0%	

n	5676	1944

Sources: *Own Calculations: Finscope (2010) and SESE (2010)*

Table 7: Distribution of Weights, Outliers

	FSBS	QLFS
50th percentile		
n	2838	1329
Contribution to total estimate	564 002	439 363
Percentage	10,10%	26,52%
100th percentile		
n	56	26
Contribution to total estimate	932 712	93716
Percentage	16,72%	5,70%
N	**5676**	**2659**

Sources: *Own Calculations: Finscope (2010) and QLFS 2010:2*

Table 8: Distribution of Weights

	FSBS	QLFS 2010:2	SESE
Mean	983	623	577
Std. Dev	2583	494	404
Range			
min	4.1	50	54
max	71714.5	7126.5	5577.8
Percentiles:			
10th	79	232	258
25th	180	343	340
50th	444	508	476
75th	1 075	749	685
90th	2 209	1119	983
95th	3 277	1435	1293
99th	6 730	2597	2179

Sources: *Own Calculations: Finscope (2010) and QLFS 2010:2*

Table 9: Gini Coefficients on the Weight Distribution

	n	Gini
QLFS 2010:2	2659	0,40
Finscope (incl. foreigners)	5676	0,75
Finscope (excl. foreigners)	4683	0,74

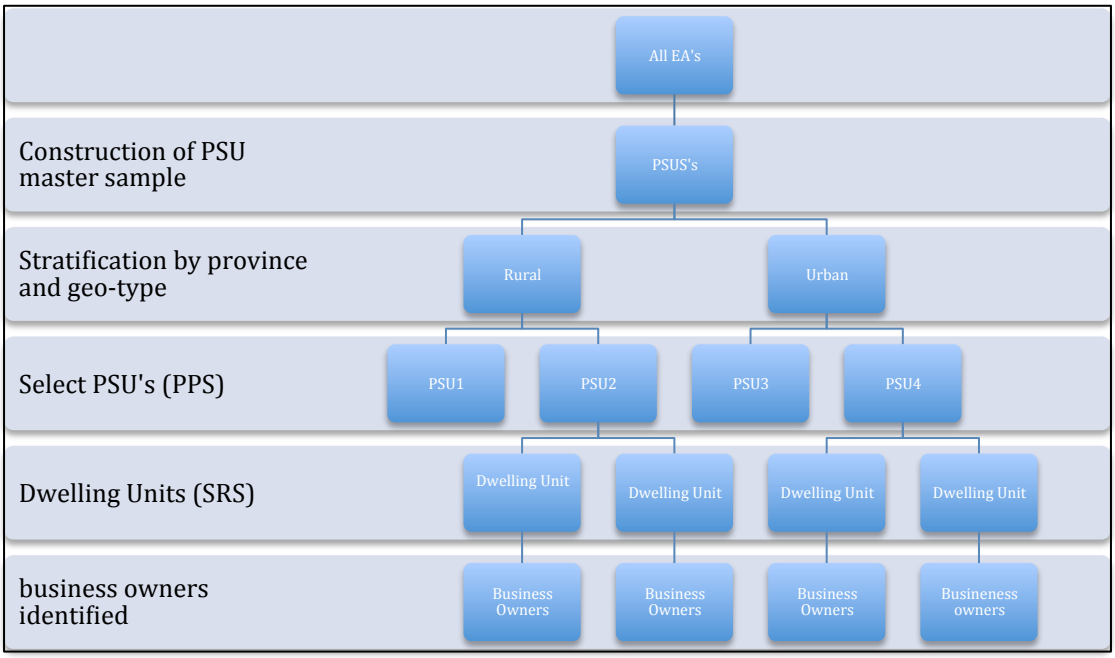
Sources: *Own Calculations: Finscope (2010) and QLFS 2010:2*

Table 10: Impact on the numbers

	n	% of sample	Contribution to total
Foreigners	993	17%	975 901
Extreme Weights			
Fisncope (incl foreigners)	56	1%	932 712
Finscope (excl foreigners)	14	0%	157 871
QLFS	26	6%	93716
Secondary Employment			
Finscope	429	6%	361 011
QLFS 2010:2	17	1%	10 603

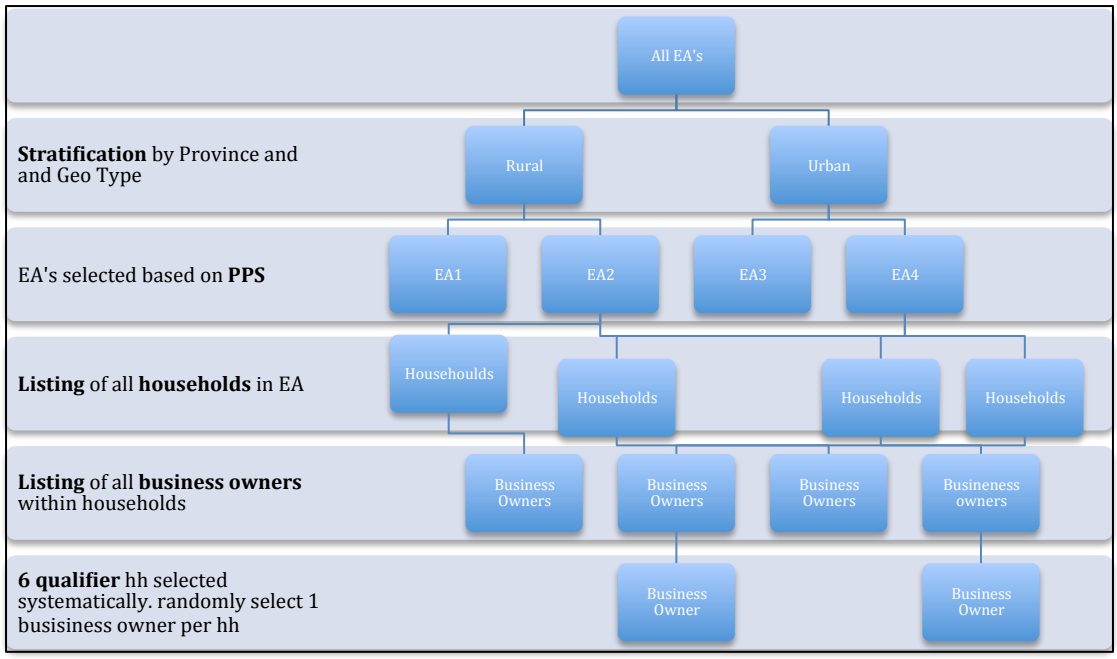
Sources: *Own Calculations: Finscope (2010) and QLFS 2010:2*

Figure 1: QLFS Survey Design



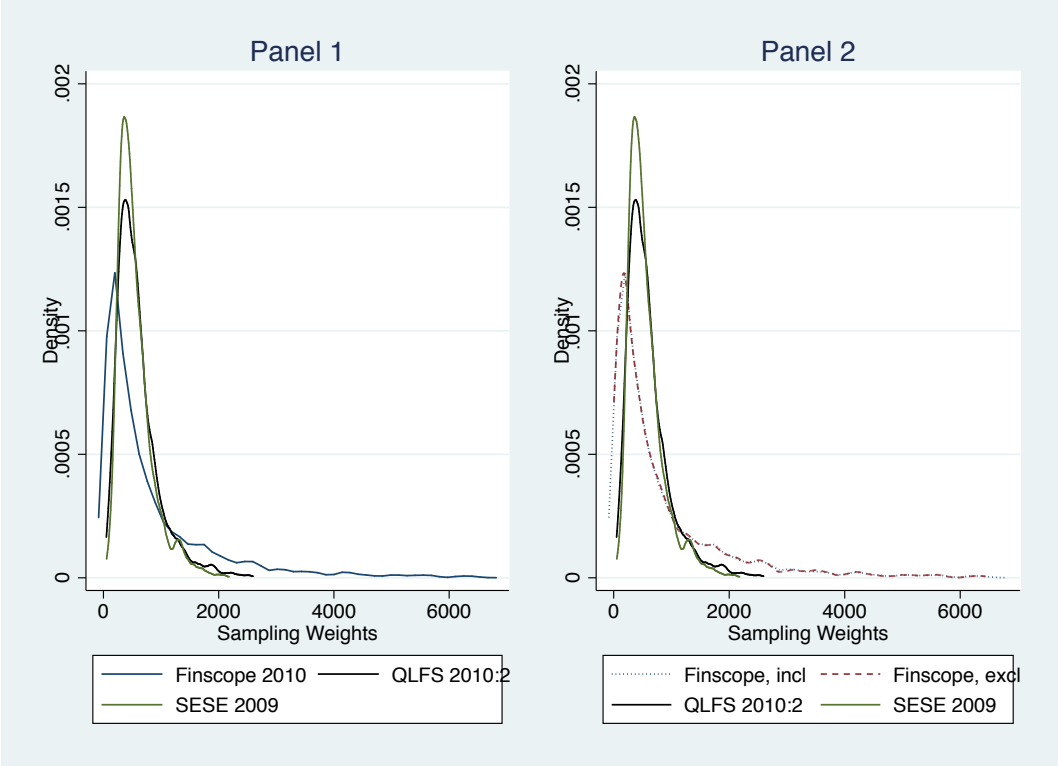
Sources: Statistics South Africa (2008), *QLFS 2010:2*

Figure 2: FSBS Survey Design



Sources: Finscope *(2010)*

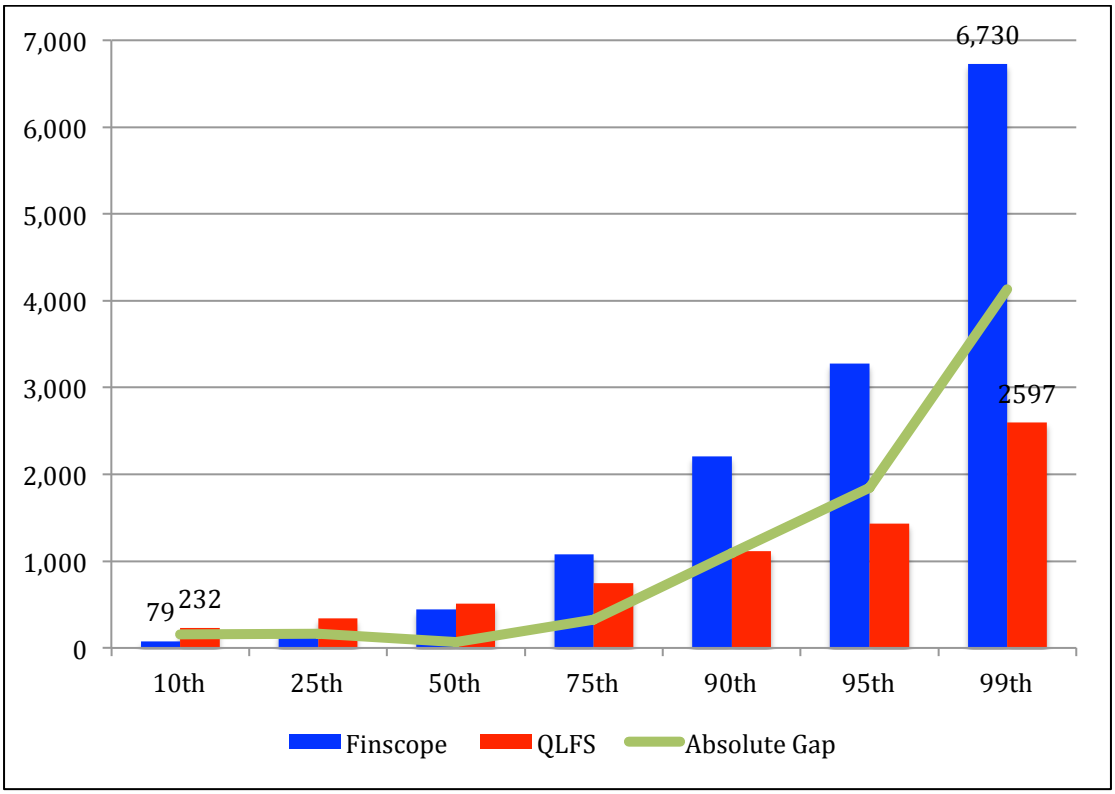
Figure 3: Distribution of Sampling Weights



Sources: *Own Calculations:* Finscope *(2010), QLFS 2010:2, SESE 2010*
Notes: *All graphs plotted from 1st to 99th percentile of each distribution respectively*

99th Percentile, QLFS 2010:2: 2597
SESE 2009: 2179
Finscope (2010) Incl. foreigners: 6730
Finscope (2010) Excl. foreigners: 6438

Figure 4: Comparison of Sampling Weights



Sources: *Own Calculations:* Finscope *(2010), QLFS 2010:2*