GCRO 2015/16 Quality of Life Survey: Sample Design and Weighting

1. Sampling frame

An updated database of population and demographic information is key to successful research, studies and surveys. Since Statistics South Africa did not release an EA (enumerator area) sampling frame based on the 2011 population census, a new 2011 EA sampling frame was constructed by a team of specialists consisting of GTI (GEOTerraImage (Pty) Ltd) and Dr Ariane Neethling.

StatsSA's Census 2011 information on Small Area Layer, main and sub place was superimposed on the 2011 set of EAs through GIS techniques and statistical modelling. This information was simultaneously integrated with the latest GTI Building Based Land Use (BBLU) and dwelling unit counts dataset.

BBLU data is constantly updated by the GTI New Development dataset. This information as well as information from fieldwork reports, and from other sources are used to update the sampling frame annually. This updated sampling frame includes inter alia new developments, changes of EA type, and changes in administrative boundaries such as municipality and ward boundaries. Benchmarking and statistical modelling techniques are applied to the updated information annually to ensure that the demographic variables align to the latest mid-year estimates as released by StatsSA. For the QoL 2015/16 survey the 2014 StatsSA midyear estimates at District Municipality level were used.

The EA sampling frame consists, for each EA, of its demographic information and estimated population, counts of number of households, number of people as well as numbers per population group, gender and per five-year age interval, etc.

Since EAs do not roll-up to form a ward, some EAs are split over different wards. Therefore, the sampling frame and GTI's dwelling counts per EA (or section of an EA) in a specific ward are used in the design of the sample.

The main benefit of using this sampling frame was that it formed the basis of the sampling process from the design of the sample, draw of dwelling units to the weighting of the final survey data.

2. Sampling methodology

A stratified multistage sample design was designed for the 2015/16 GCRO QoL survey. A sample of 30 000 households was drawn among the 508 wards in Gauteng. This sample was designed by selecting 6000 EAs with 5 visiting points in each EA.

In stratification a distinction is made between 'explicit' and 'implicit' stratification. 'Explicit stratification' refers to when the population of sampling units are explicitly divided into strata and a separate sample is selected per stratum. The sample sizes of the explicit strata are determined beforehand. 'Implicit stratification' is where the population of sampling units is sorted by some characteristic(s) and then the sample is selected from the sorted list.

For this survey, wards were considered as the primary explicit stratification variable, with dominant population groups of EAs as the secondary explicit stratification variable to ensure good coverage per ward. The variables main-place, sub-place and EA-Type (including formal residential, informal residential, etc) were used as implicit stratification variable to improve the representativeness in the sample.

The EAs were considered as the primary sampling units (psus) and households as secondary sampling units (ssus). The number of persons 18 years and older per EA was considered as the measure of size (MOS).

2.1. Allocation

The allocation of the sample was done using a probability proportion to size (pps) approach.

Sampling units in proportion to some MOS can be extremely efficient in complex sampling. According to Valliant et. al (2013), when clusters are selected with probability proportional to population totals which are fairly accurate, the negative effect of clustering on the variance is lessened for a design that selects clusters with unequal selection probability.

Another advantage of pps sampling is that if the psus (e.g. EAs) are drawn pps, and the ssus (e.g. households) are drawn with equal probability, the sample could be self-weighting, which mean that all ssus (households) in the stratum will have the same probability (change) to be selected. This approach will also decrease the variance and thus increate the precision of the estimates.

A pps sample of EA's per ward was determined based on the number of persons aged 18 years and older in the EA (or section of an EA) in a specific ward. Based on the specification of GCRO, the allocation was done in the following manner:

1. The number of EAs/visiting points per ward, using proportional allocation was determined.
2. After the allocation was done, all wards in local municipalities with less than 30 visiting points were increased to 30 and all the wards in metropolitan municipalities with less than 60 visiting points were increased to 60.
3. Visiting points in local municipalities greater than 30 and in metropolitan municipalities greater than 60 were proportionally decreased to compensate for the increased size of the smaller wards that needed to be supplemented with additional interviews (step 2 above).

After the number of EAs per ward was determined (e.g. 6 EAs), the EAs were proportionally divided among the different race groups in that ward (e.g. 3 Black, 2 White, 1 Indian dominant EA).

2.2. Selection of EAs

The EAs in each of the above explicit strata were ordered according to main-place, sub-place, EA-Type and EA number upon which the predetermined numbers of EAs were drawn using pps (i.e. probability proportional to size) systematic sampling with the number of persons aged 18 years and older per EA as measure of size. Vacant, recreational and industrial EAs are excluded from the survey design.

In instances where wards consisted of fewer EAs than was required by the sample design, some EAs were drawn more than once. Note that every sample in an EA (called hits) was drawn independently.

EAs were only substituted in selective cases, for example if the area was inaccessible or dangerous to visit.

2.3. Selection of households and respondents

GTI supplied a list of all dwelling units per EA, with their GIS coordinates. The dwelling units in each EA were sorted according to its GIS coordinate whereafter 5 visiting points were selected using systematic sampling. An additional 5 visiting points were selected per EA as oversampling points. These visiting points were used when the original visiting point resulted in a substitute because of refusal to participate, vacant homes, or when nobody was at home after three independent visits.

In order to assist the fieldworkers to find the correct visiting points, maps were printed for each of the EAs. These maps clearly indicated the EA boundaries and street names, as well as the GPS coordinates of the selected visiting points. This technique is used to limit fieldwork bias.

At the selected household, all adults 18 years and older are listed and one respondent is randomly selected, using a Kish grid.

3. Weighting process

Weights are assigned to make weighted sample records represent target population as closely as possible. A weight indicates number of population elements "represented" by a single sample element. Therefore, the sum of the weights should be equal to the population total of elements.

According to sampling theory, weights have to be calculated as follows:

1. Calculate the design weights for the realised sample to compensate for the design which deviates from a simple random sample.
2. Hereafter the design weights are compensated for non-response, if present.
3. In the final step, the design weights are benchmarked (using rim weighting, calibration or another benchmark technique) to resemble the 18+ population total in the population, according to the latest released midyear estimates of StatsSA.

Since the sample design could not fully be applied due to fieldwork problems and other reasons, the design weights (base weight) could not be calculated completely according to the initial design. The best option was to calculated the design weights (under assumptions) as follows, based on Census 2011 numbers per ward.

First the selection probability and weight of a household in a ward were determined by

$$W_{HH} = \left(\frac{n_{HH}}{N_{HH}} \right)^{-1}$$

where n_{HH} is the number of households interviewed during the survey in a specific ward and N_{HH} is the population number of households in that ward.

In the final stage, a person aged 18 years or older was selected from the drawn household. The respondent weight was calculated by

$$W_{PP} = W_{HH} * n_{18+},$$

where n_{18+} is the average number of persons aged 18 years and older in the selected households in that ward. The average is used, instead of the observed number of persons 18+ in a household, to obtain more smooth design weights with less variation.

After the calculation of the design weights, these weights were benchmarked to the Census 2011 gender-by-race adult population per ward as desired by GCRO. The benchmarking was executed on the gender-race cells per ward as far as possible. If no males or females of a certain race group in a ward was interviewed, the two gender groups of that race group were combined. If a race group was not present in the realised sample, its population size was proportionally divided among the other race groups in the ward so that the population total of that ward was still correct.

The weights calculated by this method sum to the population number of persons aged 18 years and older per ward.

Lastly, since GCRO also wanted the weights to sum to the sample size of 30 000, the weights were proportionally down-scaled.

References

- Cochran, W.G. (1977), Sampling Techniques, John Wiley & Sons, Inc.
- Kish, L. (1965), *Survey Sampling*, John Wiley & Sons, Inc.
- Lohr, S.L. (2010) *Sampling: Design and Analysis*, 2nd Edition, Brooks/Cole, Cengage Learning.
- Valliant, R., Dever, J.A, and Kreuter, F. (2013), *Practical Tools for Designing and Weighting Survey Samples*, 2013, Springer.

Dr A. Neethling (Pr.Nat.Sci)
(PhD Statistics (Sampling), Wits University)