PAUL FATTI CONSULTING cc

GCRO 2017 QUALITY OF LIFE SURVEY
REPORT ON THE STATISTICAL SAMPLING FOR THE SURVEY AND WEIGHTING OF THE RESULTS

The previous four surveys

The GCRO survey is a household survey that has been conducted by GCRO every second year since 2009. Apart from the 2009 survey, it was conducted only in Gauteng. The main stratification used in the Survey is Ward, with random sampling of households carried out within wards.
The previous surveys all used a multistage sample design and generally used PPS (Probability Proportional to Size) for the sample sizes within wards.

The sample sizes and sampling methods used in the previous four surveys were as follows:

2009: 596 Wards sampled in 4 provinces (Gauteng 442, Mpumalanga 72, North-West 70 and Free State 12. A total of 6,636 household interviews were performed – i.e. an average of 11 per ward. Sampling: Random starting points within wards, with approximately 5 households interviewed from each starting point, with every 20th household being selected for interviewing, until the required number of households was achieved.

2011: 508 Wards sampled in Gauteng. A total of 16,729 household interviews were performed– i.e. an average of 33 per ward. Sampling: Random starting points within wards, with approximately 4 households interviewed from each starting point, with every 5th household being selected for interviewing, in the specified direction, until the required number of households was achieved.

2013: 508 Wards sampled in Gauteng. A total of 27,490 household interviews were performed– i.e. an average of 54 per ward – PPS (based on the 2011 Census) with a minimum of 30 per ward in the 172 District Municipality wards and a minimum of 60 per ward in the 336 Metropolitan Municipality wards. Sampling: Random PPS sampling of SALs (Small Area Levels) was conducted amongst 16,400 SALs out of a total of 17,840 across Gauteng. In each of the selected SALs a random number of interviews was conducted, the first house being randomly selected and then every 4th house, until the required number had been selected for interviewing.

2015: 508 Wards sampled in Gauteng. A total of 30,000 household interviews were performed– i.e. an average of 59 per ward – PPS (based on the 2011 Census) with a minimum of 30 per ward in the 172 District Municipality wards and a minimum of 60 per ward in the 336 Metropolitan Municipality wards. Sampling: Within wards the households were sorted by main-place, sub-place and EA. Random PPS sampling of EAs yielded a total of 5,860 EAs across Gauteng, from which to select households to interview. Within each selected EA each dwelling

unit was identified according to its GIS coordinates, and 5 of them were randomly selected for interviewing.

The Current (2017) Survey

1. **Optimising the sample size**
   The smallest stratum of interest in the GCRO survey is the ward and therefore the critical sample size is that of the ward. While PPS sampling within wards provides estimates over municipalities and over Gauteng that are "self-weighting" and therefore representative without requiring weighting, this is not the optimal approach as regards sample size and precision. The precision of a random sample is determined by the sample size, not by the population size. The table below gives the 95% precision for a binary (yes/no) question in which approximately half of the population in ward responds "yes" and the other half responds "no".

Ward sample Size	95% Precision
10	31%
20	22%
30	18%
40	15%
50	14%
60	13%
70	12%
80	11%
90	10%
100	10%

So, for example, if 40% of a random sample of 30 respondents answers "yes" to the question, then you can be 95% confident that the true percentage answering "yes" in the population lies within 18% of 40%, i.e. between 22% and 58%. However, if the sample size is 50, then the 95% confidence interval lies between 26% and 54%.

In view of the wish to bring down the sample size a little this year, and the fact that the number of wards in Gauteng was increased from 508 to 529 in 2016, it is recommended that a sample size of 50 interviews be used in each ward this year. This will require a total sample size of 26,450 and will appreciably improve the precision of the samples in the District Municipality wards while minimally reducing it in the Metropolitan Municipality wards. (Using a sample size of 60 for the metropolitan municipal wards and 50 in the district municipality wards, and assuming that all of the 21 new wards are district municipality wards, will require a total sample size of 29,810.)

2. **Sampling Methodology**
   The **2009 and 2011 surveys** both used random starting points within a ward for identifying the first property on which to conduct interview, and if there was more than one household on the property a random procedure was used to select the household to survey. Subsequent households (typically 3 or 4) in the neighbourhood of the starting point were selected according to a systematic rule. A Kish Grid or random rule was used to identify the adult (18 + years) to interview.

The **2013 survey** used random (PPS) sampling to select small area layers (SALs) within wards as primary sampling units and households as secondary sampling units. For each SAL selected in a ward, the stand with one or more dwellings on it that was nearest the centroid of the SAL was selected, and one of the dwellings was randomly (using the toss of a die) selected for an interview. If more than one interview had to be made in the SAL (or the first one was unsuccessful), then the next stand was selected by moving down the street in a random direction and selecting the fourth stand, and proceeding as before to select a dwelling for an interview.

If a selected dwelling contained more than one household, then the toss of a die was used to select the household for the interview. If the household contained more than one eligible (18+) member then the person with the next birthday was selected for the interview. If this person was not available, then an appointment was made for an interview. If the person was not available after three attempts were made to make an appointment, then another stand was selected as before, and the same methodology was followed to obtain an interview.

The **2015 survey** used an EA (Enumerator Area) sampling frame based on the 2011 population Census which was constructed by Dr Ariane Neethling and GeoTerraImage (Pty) Ltd, as the Census data is not released at EA level. EAs were considered as primary sampling units and households as secondary sampling units. Sample sizes within wards were determined by PPS according to the 2011 census, EAs were randomly selected within wards and 5 households were chosen randomly within each of the selected EAs.

For the **2017 survey** it is recommended that, since information is not available to GCRO at EA level, the procedure adopted in 2013, using SALs as primary sampling units and households as secondary sampling units. (In the writer's opinion, using SALs as primary sampling units is equally as effective as using EAs.)

As discussed earlier, using a sample size of 50 households in each of the 529 wards to interview will result in a total sample of 26,450 interviews. Using two interviews per SAL will require 25 SALs to be randomly selected in each ward. Ideally, from a statistical viewpoint, the two households to interview should be selected randomly and independently, but this will require more travelling for the fieldwork team than if first household is randomly selected in the SAL and the second is selected in the neighbourhood of the first household, according to some rule. This will require a total of 13,225 starting points for the fieldwork team

The table below shows the tradeoff between the number of interviews conducted in a SAL, the number of SALs selected per ward, the number of interviews conducted in a ward and the total number of interviews performed in the survey.

No.Interviews / SAL	No.SALs / Ward	No.Interviews / ward	Total Interviews
1	50	50	26,450
2	25	50	26,450
3	17	51	26,979

4	13	52	27,508
5	10	50	26,450

It is recommended that for consistency the same number of interviews be performed in all SALs across all wards, which explains the slight increase in the number of interviews performed when three or four interviews are conducted in a SAL.

As mentioned, ideally each household to be interviewed in a SAL should be selected randomly and independently. The reason for this is that then the sample within the SAL and across the SALs in a ward will constitute a random sample of the households in the ward, ensuring that it is representative and unbiased and that its precision can be calculated. Otherwise it will be a cluster sample in the SAL and its properties will be unknown. In practice, if some randomness can be used in the selection of the households within SALs, and they are not too close together, then the interviews could hopefully be considered as being "reasonably" random.

The reason for wanting to interview more households within a SAL is the reduction of travelling cost and time required to survey all the households in a ward. However, if too many households are interviewed within a SAL, then the number of SALs included in the sample from a ward will be reduced and may possibly not be representative of the ward as a whole. In the writer's opinion, a reasonable compromise would be to use between two and four interviews per SAL.

3. **Volatility of the survey results in smaller wards**
   This was a problem with the previous surveys which used PPS sampling, so that smaller wards had smaller numbers of interviews, with a lower limit of 30 interviews. The proposal that a fixed sample size of 50 interviews be used per ward, is specifically aimed at solving this problem.

4. **Drawing a sample from an unknown population**
   This refers to the fact that since the last survey there has been a realignment of the Gauteng wards and municipalities and there are now 529 Wards in the province, compared to 508 previously and 9 municipalities, instead of 10 previously. Also, it is now 6 years since the 2011 Census and the population in Gauteng has changed considerably since then.

   There are three potential issues that arise from this:
   a. **Sampling of Wards**
      Under the proposal of using a fixed sample size for all wards, compared to PPS used in previous surveys, this is not an issue as the population size is not required for determining sample size.
   b. **Inclusion of the realigned wards in the sample design**
      A colleague in Statistics South Africa, Mr Arulsivanathan ("Sathie") Naidoo has kindly supplied me with the Gauteng ward population sizes for the 529 wards based on the 2011 Census figures, divided according to race, gender and age (less than 18 and 18+).

c. Updated population numbers required for post-survey weighting

Mr Naidoo has also provided me with Statistics South Africa's 2016 Community Survey population updates, which however only go down to local municipal level, not to ward level. These estimates are divided by race, so they will allow us to calculate updated 2016 population estimates, based on the assumptions that:

i. The relative ward population sizes within municipalities are approximately the same as in 2011
ii. The relative male/female split within race, ward and municipality is approximately the same as in 2011
iii. The relative 0-17/18+ split within gender, race, ward and municipality is approximately the same as in 2011

The updated 2016 population estimates can then be used in the usual manner for the post-survey weighting exercise.

References

- Cochran, W.G. (1963), *Sampling Techniques*, 2nd Edition, Wiley International
- Moser, C.A. and Kalton, G. (1985), *Survey Methods in Social Investigation*, 2nd Edition, Gower

Professor L.P. Fatti
17 February, 2018
(Initial draft, 16 March 2017)