Sample Method:
The Khayelitsha/Mitchell's Plain 2000 Survey[1]

Owen Crankshaw,
Sociology Department
University of Cape Town

Matthew Welch
School of Economics and Centre for Social Science Research
University of Cape Town

1. The Sample Method

The sample was designed to represent all adults (18 years of age and older) in the Mitchell's Plain Magisterial district. As discussed above, the most cost-efficient method of interviewing residents of such a large area is to use a two-stage cluster sample. The first stage of this sample entails selecting clusters of households and the second stage entails the selection of the households themselves. For our clusters of households, we relied on the Enumerator Areas as defined by Statistics South Africa for the 1996 Population Census. These Enumerator Areas are neighbourhoods of roughly 50 to 200 households. They are drawn up by the Chief Directorate of Demography at Statistics South Africa. This directorate is responsible for developing and maintaining a GIS system that provides the maps that are used for conducting the five-yearly national population census (Statistics South Africa, 2001:42-44). Although Enumerator Area boundaries do not cross municipal boundaries, they do not correspond to any other administrative demarcations such as voting wards. Enumerator Areas are designed to be homogeneous with respect to housing type and size. For example, Enumerator Area boundaries within the Mitchell's Plain Magisterial District do not usually cut across different types of settlements such as squatter camps, site and service settlements, hostels, formal council estates or privately built estates. Instead, each Enumerator Area is homogeneous with respect to any one of these housing types.

[1] This survey was funded by The Population and Forced Migration Program of the Andrew W. Mellon Foundation.

The method of selection used was that of Probability Proportional to Size (PPS). The measure of size being the number of households in each Enumerator Area as measured by the 1996 Population Census. This method was chosen as it provides the most efficient way to obtain equal sub-sample sizes across two stages of selection, i.e. we are able to select the Enumerator Areas and then select from each Enumerator Area a constant number of households for all Enumerator Areas in the sample. The sample is implicitly stratified by location and by housing type.

Sample procedure: First Stage: Selecting the clusters

The *first stage* of the sample entailed the selection of the Enumerator Areas. Before selecting the Enumerator Areas, we excluded all non-residential and institutional Enumerator Areas (except for hostels) from the sample frame. Enumerator Areas were selected systematically in such a way as to ensure that their probability of selection was proportionate to their population size. The Mitchell's Plain Magisterial District, as defined in the 1996 Population Census, consists of 1,486 populated Enumerator Areas. The survey population that we were interested in were the adults (eighteen years and older). Using the 1996 Census results, we calculated that the average number of adults per household is 2.66.[2] It was our intention to administer 2,875 questionnaires. Dividing the target number of questionnaires by the average number of adults per household, we determined that we would select 1,081 households.

We aimed to interview at least 10 households from each selected Enumerator Area. The number of Enumerator Areas to be selected in the first stage was calculated by dividing the number of households that we needed to sample to reach 2,875 questionnaires by the number of households to be interviewed per Enumerator Area giving a total of 108 Enumerator Areas. All the Enumerator Areas were listed in geographical order and by housing type. By doing this, the sample was implicitly stratified by location and housing type.

We used the following procedure to select the 108 Enumerator Areas with a probability that was proportional to their population size.[3]

First, the total number of households in the first enumeration area was added to the total number of households of the second enumeration area on the spreadsheet. The sum of these two household totals was then added to the total number of households of the third enumeration area on the spreadsheet. This procedure was carried out for all the following enumeration areas on the list and is commonly referred to as a cumulative total. Table 1 below shows the first few rows of the calculations.

---

2 As calculated from the 1996 Population Census conducted by Statistics South Africa.

3 See Levey, P., *et al* (1999) and Delaine,G., *et al* (1992) for a discussion of this method of selection.

Table 1: Calculating the cumulative household total.

Area	Enumerator Area	Number of Households	Cumulate
Gugulethu-New Rest	1066535	95	95
Gugulethu-New Rest	1066534	105	200
Gugulethu-New Rest	1066538	82	282
Gugulethu-New Rest	1066536	101	383
Gugulethu-New Rest	1066539	76	459
Gugulethu-New Rest	1066547	56	515
Gugulethu-New Rest	1066544	103	618
Gugulethu-New Rest	1066546	141	759

Secondly, we calculated a sampling interval by dividing 169,884 (the total number of households in the Mitchell's Plain Magisterial District) by 108 giving an interval for selection of 1,573. Thirdly, we randomly chose a number between 1 and 1,573 (this was 723) and selected the first Enumerator Area with a cumulated total equal to or greater than 723. The Enumerator Area in the last row of Table 1 has a cumulated total of 759 and is the first Enumerator Area with a cumulated total equal to or greater than 723, it was therefore chosen as our first Enumerator Area and is listed in Table 2 below. The process was repeated by adding the sampling interval of 1,573 to the random number and the Enumerator Area with a cumulated total greater than or equal to this number selected. We repeated this procedure until all of the 108 Enumerator Areas were selected. Table 2 below shows the list of selected Enumerator Areas and Figure 1 shows their geographical distribution.

Table 2: Enumerator Areas Selected for the Mitchell's Plain Survey, 2000

	Area	Enumerator Area	Population	Number of Households	Probability of Selection	Household Weight
1	Gugulethu-New Rest	1066546	399	141	0.008264463	121
2	Gugulethu-Kanana	1066392	222	80	0.008264463	121
3	Gugulethu-Europe	1066415	269	77	0.008264463	121
4	Gugulethu	1066003	366	93	0.008264463	121
5	Gugulethu	1066071	542	90	0.008264463	121
6	Gugulethu	1066006	524	189	0.008264463	121
7	Gugulethu	1066061	600	119	0.008264463	121
8	Gugulethu	1066005	478	129	0.008264463	121
9	Gugulethu	1066023	559	92	0.008264463	121
10	Nyanga-Lusaka	1066525	258	93	0.008264463	121
11	Gugulethu-Tambo Sq	1066052	411	114	0.008264463	121
12	Nyanga-KTC	1066970	371	119	0.008264463	121
13	Nyanga-KTC	1066967	265	78	0.008264463	121
14	Old Crossroads-Gqobhasi	1066459	215	54	0.008264463	121
15	Old Crossroads-Boys Town	1066437	391	107	0.008264463	121
16	Old Crossroads	1066420	414	118	0.008264463	121
17	Old Crossroads	1067152	698	129	0.008264463	121
18	Nyanga-Mpeta Sq	1067027	457	136	0.008264463	121
19	New Crossroads	1066102	628	106	0.008264463	121
20	Gugulethu-Waterfront	1066955	248	103	0.008264463	121

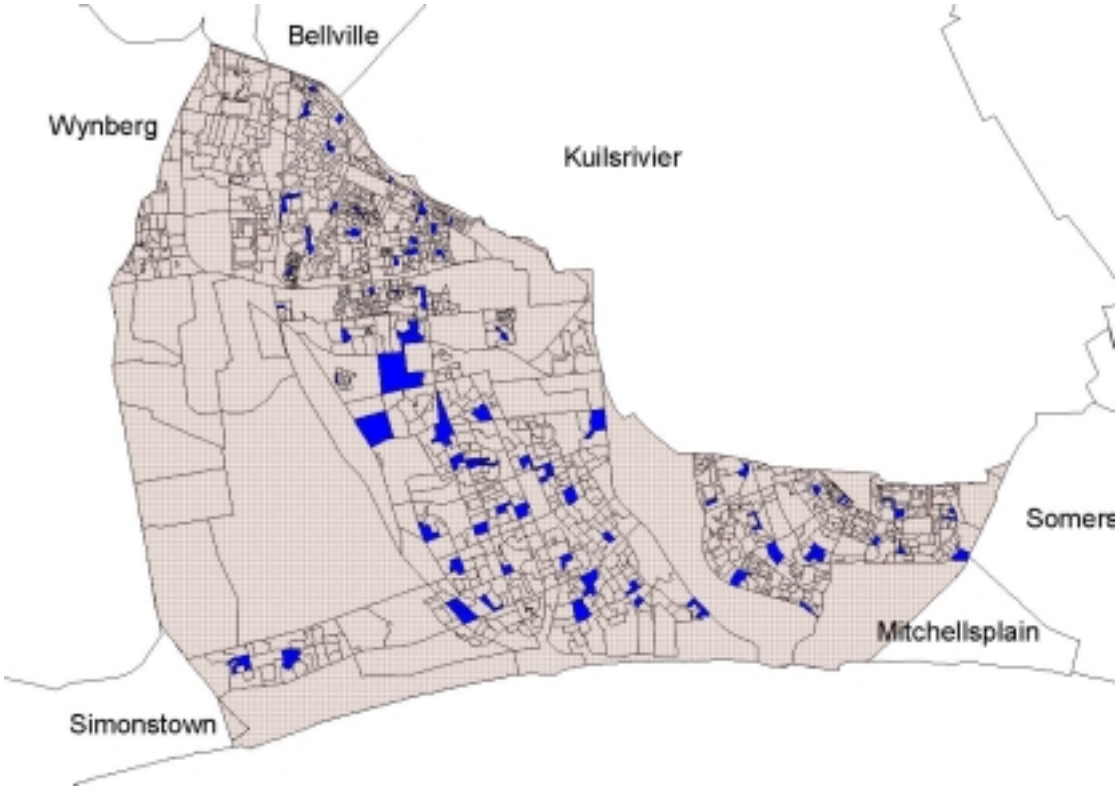
21	Gugulethu-Kick Hostels	1067025	370	96	0.008264463	121
22	Nyanga	1066884	279	94	0.008264463	121
23	Nyanga	1066508	710	110	0.008264463	121
24	Nyanga	1066506	602	92	0.008264463	121
25	Nyanga	1066501	450	83	0.008264463	121
26	Lower Crossroads	1067101	465	114	0.008264463	121
27	Browns Farm	1066579	259	83	0.008264463	121
28	Browns Farm	1066617	291	88	0.008264463	121
29	Browns Farm	1066639	668	141	0.008264463	121
30	Browns Farm	1066563	440	132	0.008264463	121
31	Browns Farm	1066651	742	171	0.008264463	121
32	Browns Farm	1066644	730	156	0.008264463	121
33	Browns Farm	1066656	529	130	0.008264463	121
34	Browns Farm	1066601	261	77	0.008264463	121
35	Browns Farm-S Machel	1066946	312	107	0.008264463	121
36	Weltevreden Valley	1060326	603	132	0.008264463	121
37	Weltevreden Valley	1060333	325	87	0.008264463	121
38	Ikwezi Park	1066116	542	127	0.008264463	121
39	Khayelitsha-Taiwan	1066847	243	87	0.008264463	121
40	Khayelitsha-SiteC	1066764	353	86	0.008264463	121
41	Khayelitsha-SiteC	1066825	421	89	0.008264463	121
42	Khayelitsha-SiteC	1066797	429	95	0.008264463	121
43	Khayelitsha-SiteC	1067086	179	75	0.008264463	121
44	Khayelitsha-SiteC	1066795	464	100	0.008264463	121
45	Khayelitsha-SiteC	1066800	560	130	0.008264463	121
46	Bongweni	1066108	227	51	0.008264463	121

47	Khayelitsha-DM	1066742	248	88	0.008264463	121
48	Khayelitsha-T Vilakazi	1066716	613	131	0.008264463	121
49	Khayelitsha-T Vilakazi	1066740	225	83	0.008264463	121
50	Khayelitsha-T Vilakazi	1066714	504	122	0.008264463	121
51	Khayelitsha-T Vilakazi	1066722	354	112	0.008264463	121
52	Khayelitsha-SiteB	1066910	295	114	0.008264463	121
53	Khayelitsha-SiteB	1066347	530	117	0.008264463	121
54	Khayelitsha-SiteB	1066364	711	154	0.008264463	121
55	Khayelitsha-SiteB	1066375	593	132	0.008264463	121
56	Khayelitsha-SiteB	1067055	373	124	0.008264463	121
57	Khayelitsha-SiteB	1067044	299	108	0.008264463	121
58	Khayelitsha-SiteB	1066371	681	143	0.008264463	121
59	Lentegeur	1060019	659	121	0.008264463	121
60	Lentegeur	1060021	816	159	0.008264463	121
61	Lentegeur	1060055	541	106	0.008264463	121
62	Lentegeur	1060035	748	126	0.008264463	121
63	Woodlands	1060313	944	186	0.008264463	121
64	Woodlands	1060311	700	139	0.008264463	121
65	Woodlands	1060305	878	167	0.008264463	121
66	Beacon Valley	1060079	665	123	0.008264463	121
67	Beacon Valley	1060072	896	168	0.008264463	121
68	Beacon Valley	1060059	692	134	0.008264463	121
69	Khayelitsha-SectD	1066894	411	163	0.008264463	121
70	Khayelitsha-Green Point	1066864	246	89	0.008264463	121
71	Khayelitsha-SectI	1066287	530	115	0.008264463	121
72	Khayelitsha-SectE	1066269	470	108	0.008264463	121

73	Khayelitsha-SectA	1066246	418	82	0.008264463	121
74	Khayelitsha-Ilitha Park	1066485	186	45	0.008264463	121
75	Khayelitsha-Town2	1066858	550	126	0.008264463	121
76	Khayelitsha-G Mxenge	1066687	345	110	0.008264463	121
77	Khayelitsha-G Mxenge	1066679	276	83	0.008264463	121
78	Khayelitsha-Macassar	1066194	405	117	0.008264463	121
79	Khayelitsha-Macassar	1066233	335	93	0.008264463	121
80	Khayelitsha-Macassar	1066156	448	116	0.008264463	121
81	Khayelitsha-Macassar	1066185	373	102	0.008264463	121
82	Khayelitsha-Macassar	1066220	415	105	0.008264463	121
83	Khayelitsha-Macassar	1066165	409	122	0.008264463	121
84	Khayelitsha-Makhaya	1066144	631	124	0.008264463	121
85	Khayelitsha-Makhaya	1066143	269	59	0.008264463	121
86	Khayelitsha-Harari	1066303	445	130	0.008264463	121
87	Khayelitsha-Harari	1066325	493	121	0.008264463	121
88	Khayelitsha-Harari	1066294	486	123	0.008264463	121
89	Eastridge	1060111	860	150	0.008264463	121
90	Eastridge	1060093	1063	194	0.008264463	121
91	Tafelsig	1060340	368	118	0.008264463	121
92	Tafelsig	1060131	1002	184	0.008264463	121
93	Tafelsig	1060163	1065	184	0.008264463	121
94	Tafelsig	1060162	991	178	0.008264463	121
95	Tafelsig	1060147	701	135	0.008264463	121
96	Tafelsig	1060128	1143	222	0.008264463	121
97	Portlands	1060230	724	147	0.008264463	121
98	Portlands	1060232	756	155	0.008264463	121

99	Portlands	1060215	596	127	0.008264463	121
100	Rocklands	1060195	643	135	0.008264463	121
101	Rocklands	1060201	688	140	0.008264463	121
102	Rocklands	1060196	509	103	0.008264463	121
103	Westridge	1060256	948	180	0.008264463	121
104	Westridge	1060243	730	153	0.008264463	121
105	Westridge	1060262	503	109	0.008264463	121
106	Strandfontein	1060286	630	147	0.008264463	121
107	Strandfontein	1060294	702	162	0.008264463	121
108	Strandfontein	1060283	556	128	0.008264463	121

Figure 1: The Geographical Distribution of the Selected Enumerator Areas within the Mitchell's Plain Magisterial District



Second Stage of the Sample: Selecting the households

Using results from previous surveys that we had conducted, we expected a household response rate of around 80 percent. To ensure that we interviewed adults in at least 10 households in every Enumerator Area, we selected 13 households at the second stage of the sample to fit our expected response rate of 80 percent. The households were selected using the systematic sampling method with a random start. Using the results of the 1996 Population Census we calculated the sampling interval by dividing the total number of households in each Enumerator Area by 13.

Where it was possible, we used aerial images (orthophotographs) of each Enumerator Area to draw a sample of dwellings (see Crankshaw *et al*, 2001). We numbered all the dwellings within the Enumerator Area, always starting in the most South-West corner. For each Enumerator Area we generated a random number that fell within the range of the sampling interval for that Enumerator Area. The dwelling corresponding to this number was therefore chosen as the starting point for the systematic sample. Subsequent households were selected according to the sampling interval. To ensure that we counted households that were sharing a dwelling or a stand, we enquired at each dwelling or stand to establish the number of resident households.

The digital orthophotographs were not much help when it came to drawing samples of households in hostels. What was useful, of course, was that the orthophotograph gave us advance information that our Enumerator Area was a hostel. This certainly facilitated our fieldwork organisation because access to hostels is best secured well in advance of the interviewing. Once we were granted access to the hostels, we had to develop a sample frame with a field visit. On the basis of these sample frames, we then drew a systematic sample with a random start from the population of adult hostel residents.

Our fieldworkers aimed to interview every adult in the selected households. To do this, they followed the rule that the household had to be re-visited at least three times on different days. In practice, however, households were revisited more often than this.

2. Calculating the Probability of Household Selection

We can now calculate the probability that each household in the population has of being selected into the sample. This then allows us at the analysis stage to draw conclusions about the population based on the sample drawn.

The overall probability of selecting a household into the sample is the product of the probabilities at each selection stage.[4]

Let p_1= the first stage probability for the i-th Enumerator Area.
Let p_2= the second stage probability for the household.

Then the overall probability is:

$$F_i=p_1.p_2 \qquad (1)$$

[4] The description below follows Delaine, G., *et al* 1992, p.29.

The second stage probability is then:

$p_{2i}=b/N_i$ (2)

Where:

b= the fixed number of households selected for all Enumerator Areas

N_i= the number of households listed in the i-th Enumerator Area

Substituting in (1) results in:

$Fi=p_{1i}b/N_i$ (3)

If Enumerator Areas are selected with probability proportional to size N_i then:

$p_{1i}=kN_i$ (4)

Where k is a constant. In PPS sampling k is the reciprocal of the sampling interval I, the value of I being N/m, where N is the total number of households in the population and m is the number of selected clusters.

Substituting into (3) then results in:

$F_i=bk=constant$ (5)

The overall probability is then constant throughout and this is termed self-weighting.

In practice, however, when a census is used as the sample frame (especially given that we are four years on from the last census) listing the households will result in a different number of households. This may be due to population growth or incorrect listing in the last census. So if we take a fixed number from each Enumerator Area (and the listing of households in the Enumerator Areas is significantly different from the census), then in this case the weights will no longer be constant and we will have to weight the clusters (Enumerator Areas) at the analysis stage.

In such a case weights, for any household selected in Enumerator Area hi, will be applied using the following formula:

$W_h=(N_h/bm_h).(N'_{hi}/N_{hi})$ (6)

Where:

N_h = the number of households in the in the census.
N_{hi} = the Enumerator Area size in the sampling frame.
N'_{hi} = the listed cluster size.
b= households selected per cluster (10)
m = the number of selected clusters.

3. Practical Calculation of Sample Weights for the Mitchell's Plain Survey

The last column in Table 2 shows how the probability of selecting a household into our sample is constant.

For this discussion we assume that our listing of households is the same as those of the census, and that we do not need to make the adjustments as described in formula (6) above.

Our formula (5) shows that the overall probability of a household being selected into the sample should be equal to the number of households selected from a specific Enumerator Area, in our case 13, multiplied by the reciprocal of the sampling interval, in our case, 1/1573. This gives us an overall probability of a household being selected into the sample of 0.00826. This can be checked by applying formula (1) to each of the selected Enumerator Areas. The results of these calculations produce a constant probability (see the last column in Table 2).

Households can then be weighted by the reciprocal of there inclusion probabilities, resulting in a constant weight of 1/0.00826 = 121. This means that each household in the sample represents 121 households in the total population.

4. Adjustment for Non-response

The simplest way of dealing with non-response is to weight the responses with the inverse of the response rate. For example, if the response rate is 80 percent then a suitable weight would be 1/0.80 =1.25, this could be done for each Enumerator Area and applied to each responding household.[5] Further more, if it is found that the population proportions as estimated from the survey differ significantly from a known reliable source, such as the recent population census, a further post stratification adjustment can be made to the weights to adjust the survey proportions to match the census proportions.

References:

Crankshaw, O., Welch, M. and Butcher, S. 2001, 'GIS Technology and Survey Sampling Methods: The Khayelitsha/Mitchell's Plain 2000 Survey', *Social Dynamics* 27(2), pp.156-174

Delaine, G**.,** Demery, L., Dubois, J., Gradjic, B., Grootaert, C., Hill, C., Marchant, T., McKay, A., Round, J., Scott, C. 1992, 'The Social Dimensions of Adjustment Integrated Survey, A Survey to Measure Poverty and Understand the Effects of Policy Change on Households', SDA Working Paper No.14, The World Bank, Washington D.C.

Levey, P., Lemeshow, S. 1999, *Sampling of Populations Methods and Applications*, 3rd edition, John Wiley & Sons, Canada.

Qaba, O., 'Intra-Cluster Homogeneity In A South African Survey', unpublished paper.

Statistics South Africa, 2001, *Annual Report 2000/2001*, Statistics South Africa, Pretoria.

Thanks also to Professor Jim Lepkowski at the University Of Michigan for his expert advice and comments.

[5] Under such a method of dealing with non-response it is assumed that all households selected into the sample have the same probability of responding.