GCRO TECHNICAL REPORT
QoL 7 SURVEY #NO. 3

SAMPLING REPORT

GCRO QUALITY OF LIFE SURVEY 7 (2023/24)

OCTOBER 2024

Authors:

Laven Naidoo, Christian Hamann and Yashena Naidoo

GCRO | Gauteng
City-Region
Observatory

The GCRO comprises a partnership of:

SAMPLING REPORT

GCRO QUALITY OF LIFE SURVEY 7 (2023/24)

OCTOBER 2024
DOI: https://doi.org/10.36634/EOSA3096
Written by: Laven Naidoo, Christian Hamann and Yashena Naidoo
Design: Breinstorm Architects
Cover typesetting: Lumina Datamatics
Cover photographs: Tshepiso Seleka

Published by the Gauteng City-Region Observatory (GCRO), a partnership of the University of Johannesburg, the University of the Witwatersrand, Johannesburg, the Gauteng Provincial Government and organised local government in Gauteng (SALGA).

SAMPLING REPORT

GCRO QUALITY OF LIFE SURVEY 7 (2023/24)

Contents

Figures and tables

PREFACE

The Gauteng City-Region Observatory (GCRO) is a partnership between the University of Johannesburg, the University of the Witwatersrand, Johannesburg, the Gauteng Provincial Government (GPG) and organised local government in Gauteng (SALGA-Gauteng).

The Quality of Life (QoL) Survey has become the flagship project of the GCRO. The QoL Survey is designed to provide a regular understanding of the quality of life, socio-economic circumstances, satisfaction with service delivery, psycho-social attitudes, values and other characteristics of residents in Gauteng. It serves as a tracking and diagnostic tool, affording a rich information resource for those people in policy-making, business, civil society and the public wanting to see where progress is being made, and where concerns remain.

The QoL Survey is a household-based survey with randomly selected adults (18+ years of age) as respondents. The GCRO has conducted seven QoL surveys since its inception in 2009:

- QoL I (2009) with 5 836 respondents in Gauteng and a total of 6 636 across the wider Gauteng City-Region (GCR).
- QoL II (2011) with 16 729 respondents in Gauteng.
- QoL III (2013/14) with 27 490 respondents in Gauteng.
- QoL IV (2015/16) with 30 002 respondents in Gauteng.
- QoL V (2017/18) with 24 889 respondents in Gauteng.
- QoL 6 (2020/21) with 13 616 respondents in Gauteng.
- QoL 7 (2023/24) with 13 795 respondents in Gauteng.

This publication is one of a series of technical reports about QoL 7 (2023/24). The reports include the Questionnaire, Fieldwork Report, Data Report, Sampling Report and the Weighting Report, as well as a generic guide to weighted analysis. These reports go hand in hand with the public dataset and should be consulted when analysing the QoL 7 (2023/24) data.

Additional information on the QoL Survey can be found on the GCRO website.

Photograph by Tshepiso Seleka

1. INTRODUCTION

The GCRO Quality of Life (QoL) Survey is a longstanding survey of randomly selected adults in households across the entire Gauteng province. The rich and longitudinally comparable data generated by our QoL surveys enables analysis that is both wide and deep, allowing a detailed, regularly updated understanding of the overall quality of life in the province, along with much more nuanced insights into trends, shifts and variations over time and space.

The survey is designed to be representative at ward level.[1] Since the first survey iteration in 2009, the details of the sampling process have varied (Orkin, 2020), but four general stages have remained constant. Stage 1 is the selection of clusters within wards where visiting points of residential dwelling units should occur, although the nature of clusters have varied over time. Stage 2 is the selection of visiting points in the clusters. Stage 3 is the selection of a household at the visiting point and stage 4 is the selection of an adult respondent in the selected household. Stages 1 and 2 are conducted before going into the field, and stages 3 and 4 are carried out in the field by fieldworkers. This report pertains to the first two stages – the selection of survey clusters within wards and the selection of visiting points. Stages 3 and 4 are documented in the Fieldwork Report (de Fortier and Loots, 2024).

Nonparametric bootstrap sampling was used to select clusters within each ward (stage 1). Simple random selection was then used for the selection of residential dwelling units as visiting points (stage 2). This multistage stratified cluster sampling strategy, using the updated ward layer as the stratification variable, was used because it brings substantial advantages over pure random sampling in terms of the logistical feasibility and cost efficiency of the data collection process (Orkin, 2020). This document outlines the rationale and guidelines for the QoL 7 (2023/24) sample design and then presents the detailed design and process that was followed to draw the QoL 7 (2023/24) sample.

[1] Wards are geopolitical subdivisions of municipalities, delimited by the Municipal Demarcation Board and used for electoral purposes.

2. RATIONALE FOR THE SAMPLE DESIGN

2.1 Key parameters

The QoL 7 (2023/24) sample design was guided by the need for a simple, automated and statistically robust approach. It was inspired by machine learning modelling sampling protocols. The available project budget was also a key factor. The key parameters and guidelines for the QoL 7 (2023/24) sample design included the following:

1. A total project budget allowed for a minimum of 13 500 interviews.
2. Retention of wards as the basis for sampling, both for reasons of comparability with previous QoL iterations and due to their ongoing salience to key constituencies.
3. Interviews in all 529 wards.
4. A consistent sample size for wards within each municipality to ensure consistent ward-level precision.
5. The minimum interviews per ward in metropolitan municipalities is set at 24. The minimum interviews per ward in local municipalities is set at 20.
6. A high as possible minimum floor per municipality to ensure adequate precision of estimates in smaller municipalities, set at a minimum of 600 interviews per municipality.
7. A minimum of five clusters per ward.
8. A minimum of four visiting points per cluster.

An overview of the sample distribution implemented based on these guidelines is provided in section 3 below.

2.2 Data sources

The QoL 7 (2023/24) sample design drew on various data sources. The backbone of the sample design is administrative boundaries, of which the GCRO holds various spatial layers, including wards (as demarcated in 2020) and enumerator area (EA) boundaries. In addition, the sample design drew on data for the point location of residential dwelling units to serve as visiting points to sample (GeoTerraImage, 2022).

Scripts were developed in PostgreSQL and R to execute the various methodological steps. These scripts were developed to efficiently handle the number of points that represent residential units (~6 million points) and the need to perform 'big data wrangling' in both a database and GIS environment in order to achieve the desired outputs.

2.3 Nonparametric bootstrap sampling for cluster ranking

Traditionally, nonparametric bootstrap sampling is used in machine learning modelling exercises to allow for the 'equal' chance of any particular sample point or cluster to be selected. In the QoL 7 (2023/24) sampling process, nonparametric bootstrap sampling was only used for stage 1 – the selection of clusters within wards – via a ranking process. Stage 2 – the selection of the specific residential dwelling units for the fieldworkers to visit – was done through a simple random sampling approach (see section 3.4). A method for ranking clusters and specific residential dwelling units is required to determine the order in which these dwellings were selected during the sampling process. This ranking also dictated the order in which dwellings were visited during fieldwork.

A bootstrapping random selection approach was chosen to give every cluster in any given ward an equal chance of being selected. This approach is formally known in the literature as nonparametric bootstrap sampling, in which randomly selected data is generated by resampling with replacement from the original dataset a certain multiple of times. One thousand bootstrapping iterations were used in this study. Nonparametric refers to the fact that the approach makes no assumptions on the distribution of parameters of a dataset. In comparison to the previously used approach for QoL 6 (2020/21) (i.e. the probability proportional to size, or PPS), this bootstrapping random selection approach might have a reduced field sampling efficiency (PPS prioritises clusters with higher residential dwelling unit counts), but it makes up for statistical unbiasedness with a high degree of replicability due to its flexible, robust and simple nature no matter the dataset type. This approach has been supported in various demographic surveying studies (Buil-Gil et al., 2020; Berrar et al., 2018; McCarthy and Snowden, 1985).

This approach was scripted as an R-script, which is included in Annexure A. The chosen methodology ensured that the cluster rankings approach could be easily replicated for future QoL surveys.

2.4 Substitution of primary visiting points

Clusters and primary visiting points were drawn before fieldwork commenced. However, given that response rates are usually low in Gauteng and, because it is a significant challenge in some areas to gain access to the sampled residential dwelling units (due to security concerns of residents), it was anticipated that substitutions and secondary visiting points would be required to complete the desired sample. If an entire cluster proved to be inaccessible (e.g. when access to a security estate or farm was not granted), the substitution was done on a case-by-case basis where the next highly ranked cluster was recommended. The inaccessible cluster was substituted with another cluster that was drawn on the same principles (described in section 3.3). If the entire primary cluster list of a ward was exhausted, the secondary cluster list was utilised. The secondary cluster list was simply an extension of the primary cluster list, where all other available cluster options are ranked for use.

The substitution of visiting points occurred when a particular primary visiting point proved inaccessible. For each primary visiting point, three substitution visiting points were drawn before fieldwork commenced and were provided to the fieldwork service provider. Substitution visiting points were also drawn from the data for the location of dwelling units (GeoTerraImage, 2022) after the selection of the primary visiting points and their exclusion from the dataset. Substitution visiting points were drawn on the same principles described in section 3.5. A high-level overview of the extent of substitutions at the EA and visiting point levels is provided in the QoL 7 Fieldwork Report (de Fortier and Loots, 2024).

3. DRAWING THE SAMPLE

3.1 Sub-ward geography

The development of polygons that encompass clusters of interview locations is beneficial for the efficiency of fieldwork logistics. Fieldworkers can navigate quickly between sample points when the sample points are clustered in a relatively contained area within a ward.

The starting point for drawing clusters is the enumerator area (EA) layer, as demarcated and used by Statistics South Africa for the 2011 census. The EA layer has been chosen as the starting point because it is the smallest official geographical unit that is used for enumeration while it also aligns with residential developments and makes it easier to sample distinct parts of settlements (e.g. complexes, estates, housing suburbs and the like). Statistics South Africa updated the 2011 EA boundaries for the 2022 census, but the updates were not shared with the public by the time the QoL 7 (2023/24) sample was drawn. The rationale behind the selection of the EA geographical unit (and subsequent clusters) over others (such as voting districts, small area layers (SALs), etc.) was elaborated upon in the QoL 6 (2020/21) sample design report (Hamann and de Kadt, 2021).

The EA layer did not neatly fit the 2020 ward boundaries and needed to be adjusted to ensure that clusters of interviews remained contained in the designated wards. Clusters that did not have the minimum number of residential units required to reach the desired sample (the required number of interviews per EA X 3, i.e. 12) were excluded. This improved the efficiency of fieldwork in that the sampled EAs will be more likely to yield the required number of interviews. A unique numeric cluster ID was given to each cluster in order to connect the clusters to the sampling of residential dwelling units. The cluster layer was further refined by the contracted fieldwork company (GeoSpace International) where topological boundary errors between the 2020 wards and 2011 EA boundaries (i.e. boundary slivers) were removed. Additionally, particular clusters bordering some municipalities (e.g. the boundary between Emfuleni and Midvaal) had to be manually reallocated to the correct ward to ensure fair sampling and to follow the sampling statistics illustrated in the primary sampling distribution (Table 1).

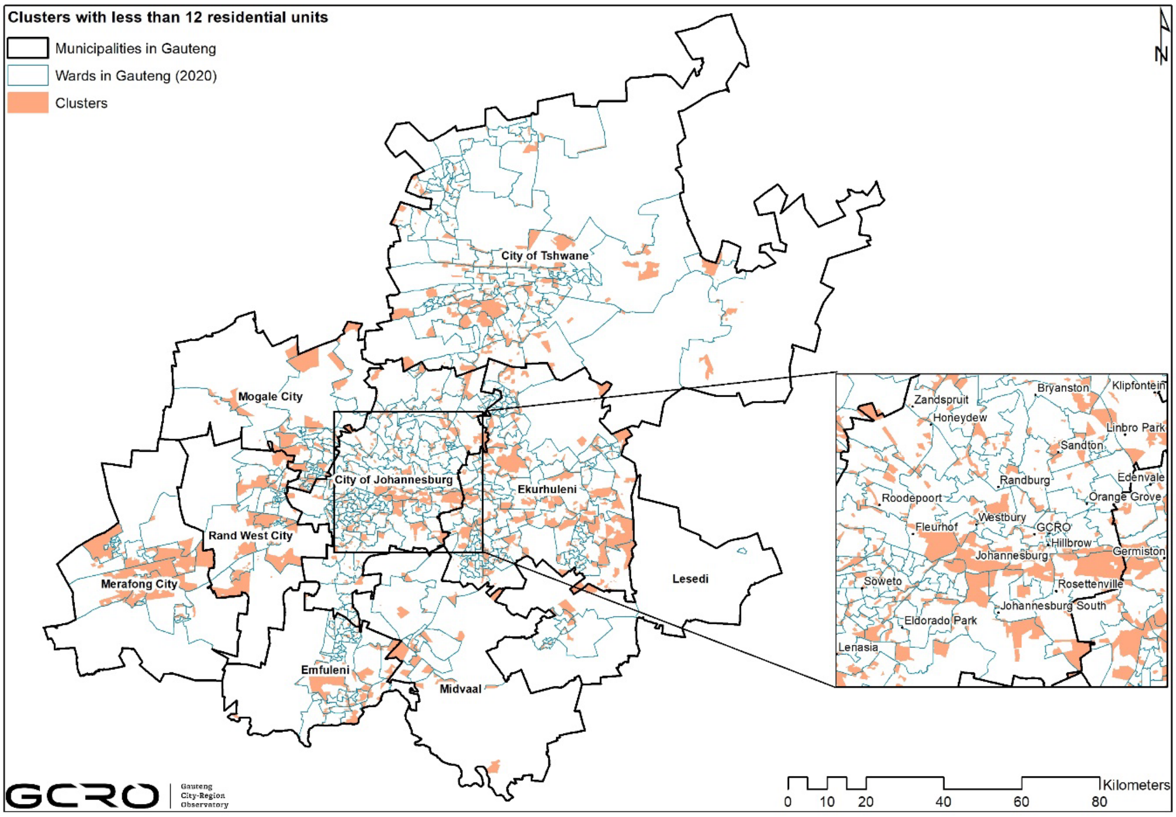
Table 1: The sample distribution of QoL 7 (2023/24)

Municipality	Number of wards	Total number of clusters	Sample structure	Visiting points per ward	Visiting points per municipality
City of Johannesburg	135	6 490	7 EAs per Ward; 4 interviews per EA	28	3 780
City of Tshwane	107	5 226	6 EAs per Ward; 4 interviews per EA	24	2 568
City of Ekurhuleni	112	5 168	6 EAs per Ward; 4 interviews per EA	24	2 688
Emfuleni	45	1 358	5 EAs per Ward; 4 interviews per EA	20	900
Lesedi	13	212	8 EAs per Ward; 6 interviews per EA	48	624
Merafong City	28	393	5 EAs per Ward; 4 interviews per EA	24	672
Midvaal	15	231	8 EAs per Ward; 5 interviews per EA	48	720
Mogale City	39	678	5 EAs per Ward; 4 interviews per EA	20	780
Rand West City	35	626	5 EAs per Ward; 4 interviews per EA	20	700
GAUTENG	**529**	**20 382**			**13 432**

3.2 The cluster sample frame

The first stage of the sampling process was to draw the sub-ward geographies in which interviews were clustered. The starting point was all the clusters in Gauteng (n = 31 454). All the clusters with fewer than 12 residential dwelling units,[2] as per the data for the location of dwelling units (GeoTerraImage, 2022), were excluded from the sample frame (n = 11 072; see Figure 1). This included some clusters which are sparsely populated, but mainly included clusters drawn around nature reserves, mining areas, industrial parks, and retail or school precincts. The nonparametric bootstrap sampling of clusters was drawn from the remaining 20 382 clusters.

Figure 1: Clusters with fewer than 12 residential dwelling points and therefore not eligible for the sample



[2] If we draw four primary visiting points per EA and three potential replacement points, we require a minimum of 12 potential visiting points. It is too costly to risk going to EAs that do not have enough potential visiting points.

3.3 Drawing the cluster sample

The nonparametric bootstrapping approach to sample clusters randomly selected one cluster per ward in each municipality. This process was repeated or 'bootstrapped' a thousand times in each municipality, and the number of times that a cluster was selected was tallied up using a pivot table function in Excel and became the ranking variable for the clusters per wards and municipality. The cluster that was randomly selected the most often out of the thousand times of random selection was given the highest cluster rank (i.e. its frequency number of selection) while the opposite was true for the cluster that was selected least often. Occasionally, some clusters could be selected the same number of times and yield the same rank. When this occurred, the clusters were followed sequentially. This bootstrapping random cluster selection process was repeated for all nine municipalities in Gauteng to ascertain the cluster ranks within all the wards (i.e. via sorting by ward ID and then by bootstrap count/rank) and municipalities.

The ward and cluster distribution per municipality are displayed in Table 1 above. Figure 2 illustrates the primary clusters that were selected via this bootstrapping approach. An example below (Figure 3) illustrates the cluster rank per cluster per ward within the municipality of Emfuleni.

Figure 2: The random selection of primary clusters for the QoL 7 (2023/24) sample

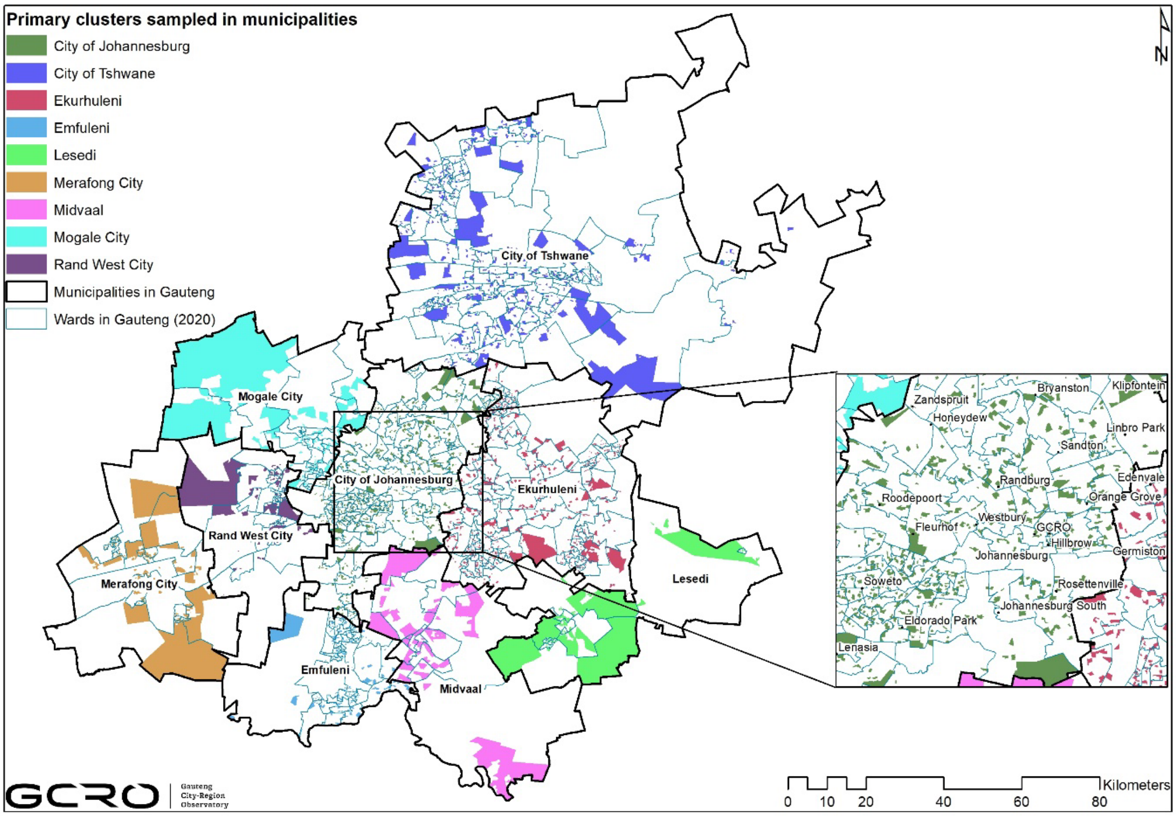
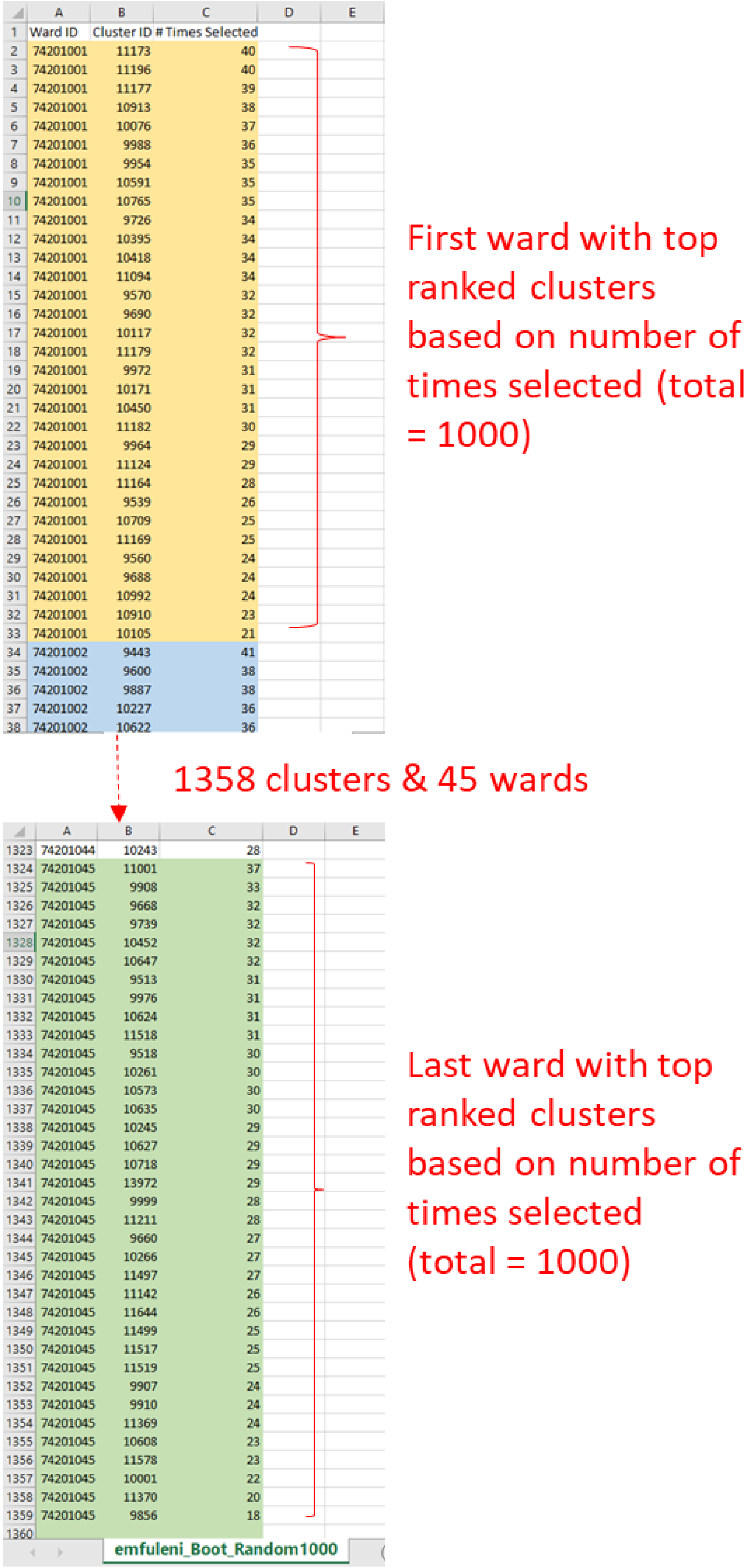


Figure 3: Bootstrap sampling rank results for Emfuleni per cluster and per ward (different colours outline the different cluster per ward ranked from highest to lowest; the higher the bootstrap count, the higher the rank as selected from the thousand bootstrap iterations)



	A	B	C	D	E
1	Ward ID	Cluster ID	# Times Selected		
2	74201001	11173	40		
3	74201001	11196	40		
4	74201001	11177	39		
5	74201001	10913	38		
6	74201001	10076	37		
7	74201001	9988	36		
8	74201001	9954	35		
9	74201001	10591	35		
10	74201001	10765	35		
11	74201001	9726	34		
12	74201001	10395	34		
13	74201001	10418	34		
14	74201001	11094	34		
15	74201001	9570	32		
16	74201001	9690	32		
17	74201001	10117	32		
18	74201001	11179	32		
19	74201001	9972	31		
20	74201001	10171	31		
21	74201001	10450	31		
22	74201001	11182	30		
23	74201001	9964	29		
24	74201001	11124	29		
25	74201001	11164	28		
26	74201001	9539	26		
27	74201001	10709	25		
28	74201001	11169	25		
29	74201001	9560	24		
30	74201001	9688	24		
31	74201001	10992	24		
32	74201001	10910	23		
33	74201001	10105	21		
34	74201002	9443	41		
35	74201002	9600	38		
36	74201002	9887	38		
37	74201002	10227	36		
38	74201002	10622	36		

First ward with top ranked clusters based on number of times selected (total = 1000)

1358 clusters & 45 wards

	A	B	C	D	E
1323	74201044	10243	28		
1324	74201045	11001	37		
1325	74201045	9908	33		
1326	74201045	9668	32		
1327	74201045	9739	32		
1328	74201045	10452	32		
1329	74201045	10647	32		
1330	74201045	9513	31		
1331	74201045	9976	31		
1332	74201045	10624	31		
1333	74201045	11518	31		
1334	74201045	9518	30		
1335	74201045	10261	30		
1336	74201045	10573	30		
1337	74201045	10635	30		
1338	74201045	10245	29		
1339	74201045	10627	29		
1340	74201045	10718	29		
1341	74201045	13972	29		
1342	74201045	9999	28		
1343	74201045	11211	28		
1344	74201045	9660	27		
1345	74201045	10266	27		
1346	74201045	11497	27		
1347	74201045	11142	26		
1348	74201045	11644	26		
1349	74201045	11499	25		
1350	74201045	11517	25		
1351	74201045	11519	25		
1352	74201045	9907	24		
1353	74201045	9910	24		
1354	74201045	11369	24		
1355	74201045	10608	23		
1356	74201045	11578	23		
1357	74201045	10001	22		
1358	74201045	11370	20		
1359	74201045	9856	18		
1360					

Last ward with top ranked clusters based on number of times selected (total = 1000)

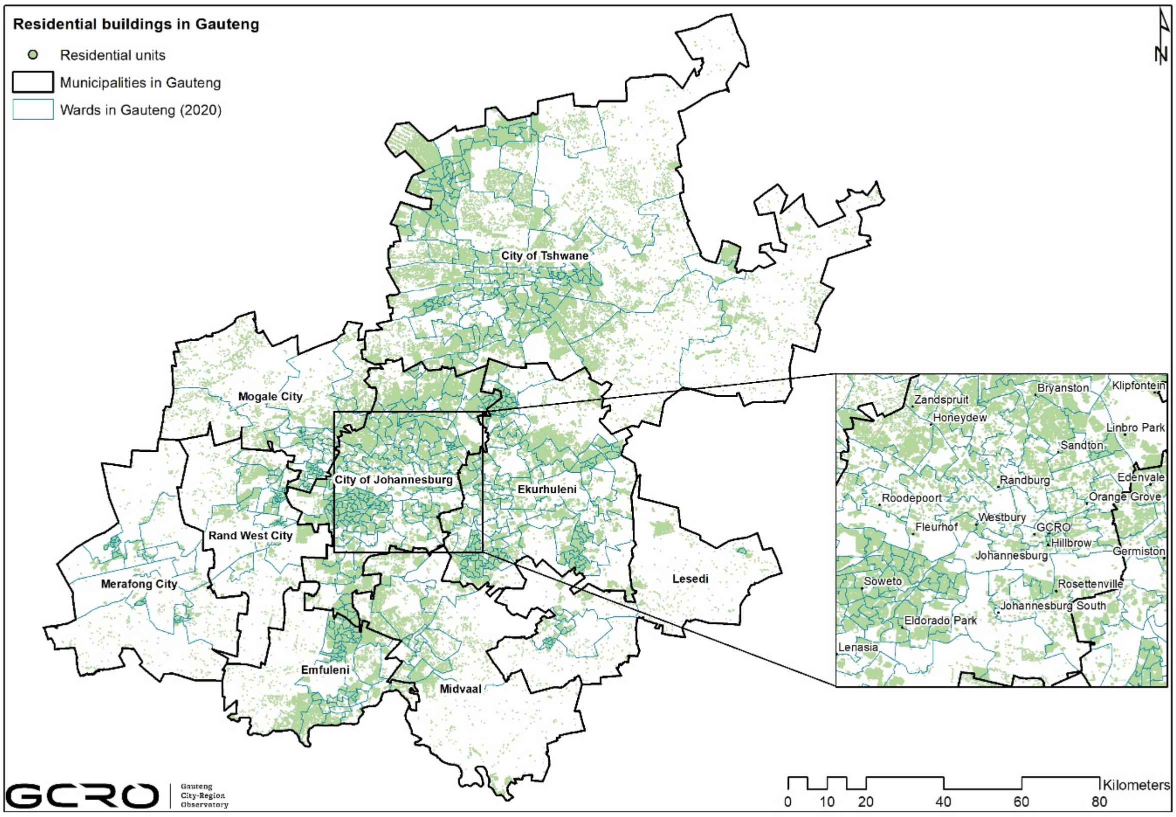
3.4 Dwelling unit sample frame

Following the selection of the clusters, the next step was the selection of visiting points in each sampled cluster. Visiting points were drawn from the GeoTerraImage Building Based Land Use (BBLU) 2022 dataset. GeoTerraImage provided the data as a BBLU point dataset in which the land use of each building is classified as one of 70 land-use classes (Figure 4). The result is a spatial layer that marks each building in Gauteng with a point according to a set of comprehensive land-use definitions (GeoTerraImage, 2022).

In addition to land-use categorisation, each point in the GeoTerraImage BBLU layer also included attributes for administrative layer association (ward and EA) and the number of residential units per building. Before the building unit selection and ranking could begin, the BBLU layer required pre-processing. The residential land-use categorisation (i.e. primary land-use code 7) was used to filter out all non-residential buildings from the BBLU dataset. Additional exclusion was made of children's homes, correctional facilities, security/estate gate points and garage units based on the secondary (code 7.9) and tertiary (codes 7.3.4, 7.3.7, 7.3.8, 7.4.1 and 7.5.1) land-use codes. Additional pre-processing steps involved duplicating the number of building points by the number of dwelling units associated with each building in order to produce a dataset in which each row represents a dwelling unit and potential sampling point. Geographic coordinates were also generated as additional fields for each dwelling unit.

Figure 4: Distribution of dwelling units in Gauteng

Data source: GeoTerraImage (2022)



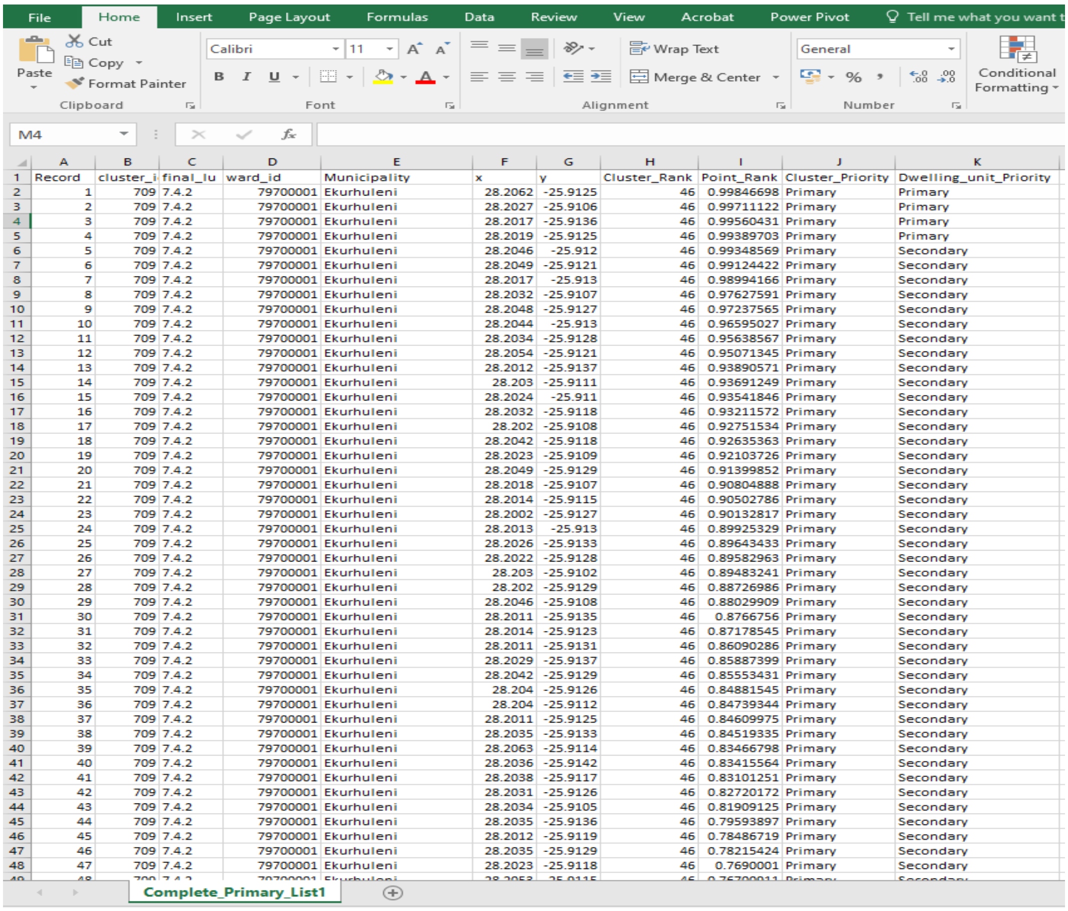
The finalised BBLU layer was spatially joined with the cluster layer for the ranking process so that the number of dwelling units per cluster could be determined. In order to handle this large dataset, the processed BBLU layer was further divided into the nine Gauteng municipalities in which the cluster and building unit ranking was conducted. This was done in accordance with the primary sample distribution requirements given in Table 1 above.

3.5 Drawing the dwelling units for visiting points

Stage 2 of the sample selection involved simple random selection of visiting points from the dwelling unit sample frame described above. For the dwelling unit ranking within clusters, a simple random number (a multi-decimal number between 0 and 1) was generated for every building unit per cluster and sorted from highest to lowest to assign the dwelling unit ranking in each cluster. This was conducted using PostGreSQL and R scripts for automation purposes (see Annexure A).

Once the primary sample distribution requirements (Table 1) were fulfilled based on the cluster and dwelling unit ranking (which populated the primary cluster and dwelling unit list), the remaining ranked data was allocated as the secondary cluster and dwelling unit list for the fieldwork team to use when the primary list was exhausted at the ward or municipality level. Figure 5 illustrates the resultant output of the cluster and dwelling unit rankings (including primary and secondary list designation) for the primary cluster and primary-and-secondary dwelling unit lists.

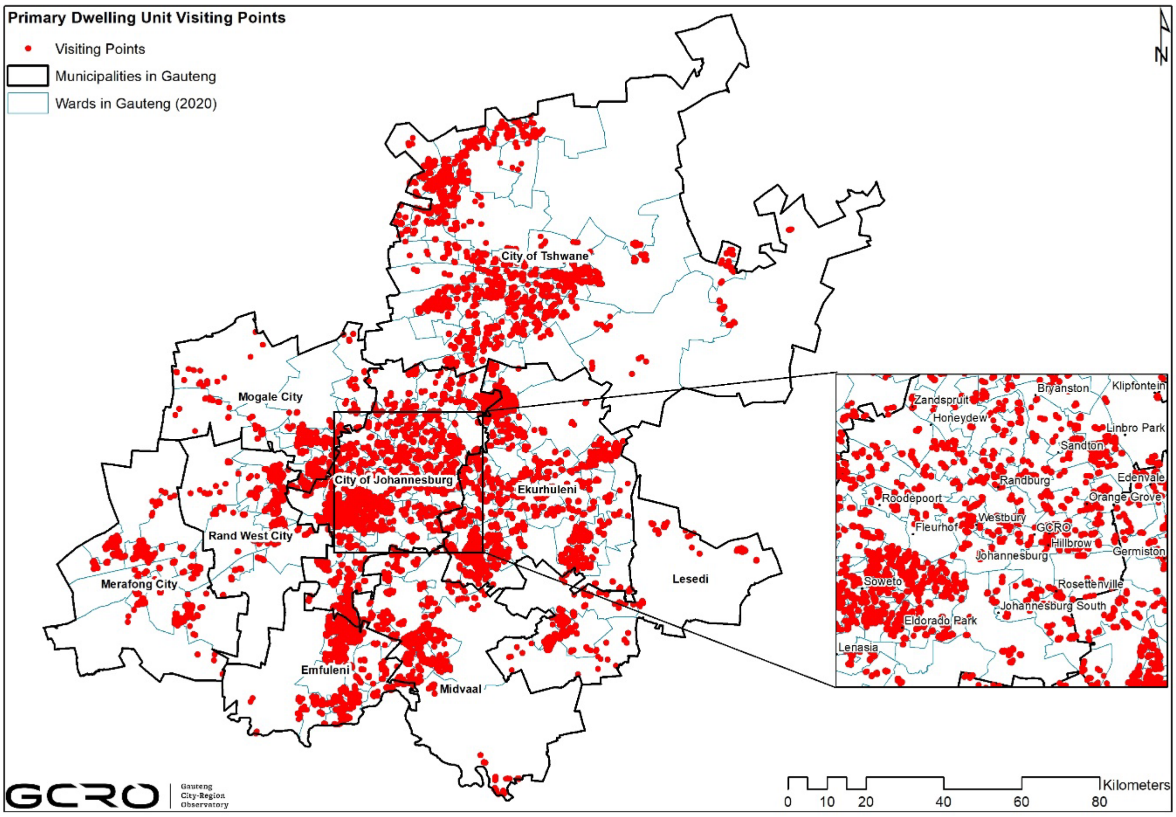
Figure 5: Primary cluster list of ranked clusters and dwelling unit points (insert from Ekurhuleni)



	A	B	C	D	E	F	G	H	I	J	K
1	Record	cluster_i	final_lu	ward_id	Municipality	x	y	Cluster_Rank	Point_Rank	Cluster_Priority	Dwelling_unit_Priority
2	1	709	7.4.2	79700001	Ekurhuleni	28.2062	-25.9125	46	0.99846698	Primary	Primary
3	2	709	7.4.2	79700001	Ekurhuleni	28.2027	-25.9106	46	0.99711122	Primary	Primary
4	3	709	7.4.2	79700001	Ekurhuleni	28.2017	-25.9136	46	0.99560431	Primary	Primary
5	4	709	7.4.2	79700001	Ekurhuleni	28.2019	-25.9125	46	0.99389703	Primary	Primary
6	5	709	7.4.2	79700001	Ekurhuleni	28.2046	-25.912	46	0.99348569	Primary	Secondary
7	6	709	7.4.2	79700001	Ekurhuleni	28.2049	-25.9121	46	0.99124422	Primary	Secondary
8	7	709	7.4.2	79700001	Ekurhuleni	28.2017	-25.913	46	0.98994166	Primary	Secondary
9	8	709	7.4.2	79700001	Ekurhuleni	28.2032	-25.9107	46	0.97627591	Primary	Secondary
10	9	709	7.4.2	79700001	Ekurhuleni	28.2048	-25.9127	46	0.97237565	Primary	Secondary
11	10	709	7.4.2	79700001	Ekurhuleni	28.2044	-25.913	46	0.96595027	Primary	Secondary
12	11	709	7.4.2	79700001	Ekurhuleni	28.2034	-25.9128	46	0.95638567	Primary	Secondary
13	12	709	7.4.2	79700001	Ekurhuleni	28.2054	-25.9121	46	0.95071345	Primary	Secondary
14	13	709	7.4.2	79700001	Ekurhuleni	28.2012	-25.9137	46	0.93890571	Primary	Secondary
15	14	709	7.4.2	79700001	Ekurhuleni	28.203	-25.9111	46	0.93691249	Primary	Secondary
16	15	709	7.4.2	79700001	Ekurhuleni	28.2024	-25.911	46	0.93541846	Primary	Secondary
17	16	709	7.4.2	79700001	Ekurhuleni	28.2032	-25.9118	46	0.93211572	Primary	Secondary
18	17	709	7.4.2	79700001	Ekurhuleni	28.202	-25.9108	46	0.92751534	Primary	Secondary
19	18	709	7.4.2	79700001	Ekurhuleni	28.2042	-25.9118	46	0.92635363	Primary	Secondary
20	19	709	7.4.2	79700001	Ekurhuleni	28.2023	-25.9109	46	0.92103726	Primary	Secondary
21	20	709	7.4.2	79700001	Ekurhuleni	28.2049	-25.9129	46	0.91399852	Primary	Secondary
22	21	709	7.4.2	79700001	Ekurhuleni	28.2018	-25.9107	46	0.90804888	Primary	Secondary
23	22	709	7.4.2	79700001	Ekurhuleni	28.2014	-25.9115	46	0.90502786	Primary	Secondary
24	23	709	7.4.2	79700001	Ekurhuleni	28.2002	-25.9127	46	0.90132817	Primary	Secondary
25	24	709	7.4.2	79700001	Ekurhuleni	28.2013	-25.913	46	0.89925329	Primary	Secondary
26	25	709	7.4.2	79700001	Ekurhuleni	28.2026	-25.9133	46	0.89643433	Primary	Secondary
27	26	709	7.4.2	79700001	Ekurhuleni	28.2022	-25.9128	46	0.89582963	Primary	Secondary
28	27	709	7.4.2	79700001	Ekurhuleni	28.203	-25.9102	46	0.89483241	Primary	Secondary
29	28	709	7.4.2	79700001	Ekurhuleni	28.202	-25.9129	46	0.88726986	Primary	Secondary
30	29	709	7.4.2	79700001	Ekurhuleni	28.2046	-25.9108	46	0.88029909	Primary	Secondary
31	30	709	7.4.2	79700001	Ekurhuleni	28.2011	-25.9135	46	0.8766756	Primary	Secondary
32	31	709	7.4.2	79700001	Ekurhuleni	28.2014	-25.9123	46	0.87178545	Primary	Secondary
33	32	709	7.4.2	79700001	Ekurhuleni	28.2011	-25.9131	46	0.86090286	Primary	Secondary
34	33	709	7.4.2	79700001	Ekurhuleni	28.2029	-25.9137	46	0.85887399	Primary	Secondary
35	34	709	7.4.2	79700001	Ekurhuleni	28.2042	-25.9129	46	0.85553431	Primary	Secondary
36	35	709	7.4.2	79700001	Ekurhuleni	28.204	-25.9126	46	0.84881545	Primary	Secondary
37	36	709	7.4.2	79700001	Ekurhuleni	28.204	-25.9112	46	0.84739344	Primary	Secondary
38	37	709	7.4.2	79700001	Ekurhuleni	28.2011	-25.9125	46	0.84609975	Primary	Secondary
39	38	709	7.4.2	79700001	Ekurhuleni	28.2035	-25.9133	46	0.84519335	Primary	Secondary
40	39	709	7.4.2	79700001	Ekurhuleni	28.2063	-25.9114	46	0.83466798	Primary	Secondary
41	40	709	7.4.2	79700001	Ekurhuleni	28.2036	-25.9142	46	0.83415564	Primary	Secondary
42	41	709	7.4.2	79700001	Ekurhuleni	28.2038	-25.9117	46	0.83101251	Primary	Secondary
43	42	709	7.4.2	79700001	Ekurhuleni	28.2031	-25.9126	46	0.82720172	Primary	Secondary
44	43	709	7.4.2	79700001	Ekurhuleni	28.2034	-25.9105	46	0.81909125	Primary	Secondary
45	44	709	7.4.2	79700001	Ekurhuleni	28.2035	-25.9136	46	0.79593897	Primary	Secondary
46	45	709	7.4.2	79700001	Ekurhuleni	28.2012	-25.9119	46	0.78486719	Primary	Secondary
47	46	709	7.4.2	79700001	Ekurhuleni	28.2035	-25.9129	46	0.78215424	Primary	Secondary
48	47	709	7.4.2	79700001	Ekurhuleni	28.2023	-25.9118	46	0.7690001	Primary	Secondary
49	48	709	7.4.2	79700001	Ekurhuleni	28.2053	-25.9115	46	0.76700011	Primary	Secondary

The final sample of primary visiting points in primary clusters (n = 13 428) is shown in Figure 6. The selected primary visiting points were removed from the sample frame and the remaining dwelling units became eligible for selection as substitution points (i.e. secondary cluster and dwelling unit lists). Substitution points within each EA were drawn in the same way as the primary visiting points.

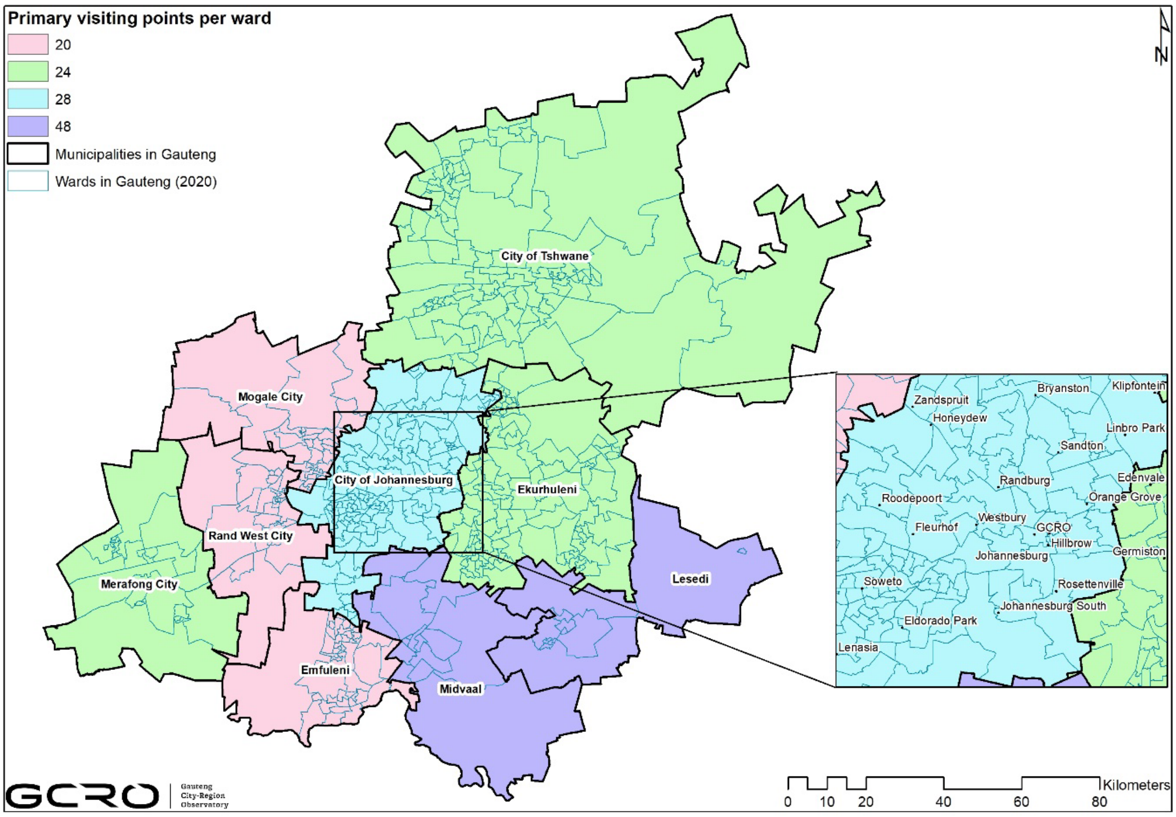
Figure 6: The random selection of primary visiting points for the QoL 7 (2023/24) sample



The number of primary visiting points per ward and municipality is shown in Figure 7.[3] In some instances, selected wards within some municipalities have a lower number of primary visiting points than other wards in the same municipality. This is the result of the sample distribution (shown in Table 1) and for the need to balance the sample and reach the minimum required visiting points for each municipality.

[3] Note that the final sample is slightly different from this sampled distribution due to the nature of fieldwork. Additionally, in a number of wards, slightly more interviews than planned were completed.

Figure 7: The number of primary visiting points per ward for the QoL 7 (2023/24) sample



3.6 Drawing the substitution points

Since the entire collection of clusters and dwelling units have been ranked, substitution points could be pulled from the primary and secondary cluster lists as long as the rank order was maintained throughout the field sample process. The sample frame for the substitution points excluded multi-unit buildings that had already been sampled. This prevented substitution points from being inside a building where access had already been denied, and also meant that fieldwork teams did not need to negotiate access to a building more than once.

Instead, once access to a multi-unit building had been negotiated, if refusals were encountered in the building, teams would identify substitute points in the same building. In-field substitution across multi-unit buildings also took place, but only after all options in the building that had been first selected were exhausted. At that point, the 'substitution buildings' were used to gather the remaining required interviews in the cluster by following the dwelling unit ranking order.

4. CHANGES DURING FIELDWORK

During fieldwork, it was necessary to adapt to unforeseen challenges and delays. A high rate of refusals in high-wall security estates and hostels proved to be problematic and slowed down fieldwork progress. In many cases in these areas, primary sampled and secondary substitute visiting points had been exhausted through outcomes such as access refusals or 'no one at home'. To overcome this, the complete ranked primary and secondary cluster lists were shared with the fieldwork team so that all available interview options could be considered for the solution. In some instances, it even meant repeating gatekeeper meetings in order to work in the same previously visited estate.

In peripheral areas, the dwelling unit point ranking was not always followed by the fieldwork team due to the long travel distances between ranked points. As a result, nearby points were grouped together for sampling before resuming the dwelling unit point rankings in the affected clusters. This change in protocol affected less than 1% (12) of the total number of clusters. During the sampling process, oversampling was conducted in a few wards where Indian and Coloured respondents were believed to be present. This was done to ensure these population groups were not underrepresented. The final list of interviews conducted in each ward in Gauteng is provided in Annexure B of the Fieldwork Report (de Fortier and Loots, 2024).

5. QUALITY CONTROL AND SUBMITTED DATA TO THE FIELDWORK TEAM

The primary and secondary lists for all nine municipalities were combined into a single dataset of complete primary and secondary lists in CSV format. The XY coordinates were mapped in a shapefile format. Quality control of random samples of the primary and secondary lists was performed in a Google Earth and GIS software environment in which the building unit sample points were interrogated (i.e. Is the building/dwelling unit there? Is it the correct land use? Is it abandoned? etc.) and to ascertain whether the primary sampling distribution had been adhered to.

In terms of disseminating the relevant dataset to GeoSpace International, the team had to limit the amount of primary and secondary datasets shared to avoid clashing with GeoTerraImage's data-sharing agreements. Using the R Studio command 'slice_head', the top 25 ranked sample points of the highest ranked clusters as per the number of required clusters per ward (see Table 1) were extracted from the master dataset and shared with GeoSpace International as the primary cluster and point list. The sample process was repeated for the secondary clusters and points (classified during the last step of the final script; see Annexure A of this report), but this was limited to the top ten ranked sample points of the remaining clusters per ward due to the large number of remaining clusters. This dataset was extracted and shared with GeoSpace International as the secondary cluster and point list.

References

Berrar, D. (2018). 'Introduction to the non-parametric bootstrap'. *Encyclopedia of Bioinformatics and Computational Biology*, *1*, 766–773. https://doi.org/10.1016/B978-0-12-809633-8.20350-6

Buil-Gil, D., Solymosi, R. and Moretti, A. (2020). 'Nonparametric bootstrap and small area estimation to mitigate bias in crowdsourced data'. In C.A. Hill, P.P. Biemer, T.D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov and L.E. Lyberg (Eds.), *Big data meets survey science*: *A collection of innovative methods*. https://doi.org/10.1002/9781118976357.ch16

de Fortier, E. and Loots, H. (2024). *Fieldwork report: Quality of Life Survey 7 (2023/24)*. Johannesburg: Gauteng City-Region Observatory. https://doi.org/10.36634/EOSA3094

GeoTerraImage. (2017). *Building Based Land Use point dataset* [Spatial layer, points represent buildings].

GeoTerraImage. (2018). *Demographics estimates: EA summary population* [Spatial layer, polygons are Statistics South Africa 2011 EA boundaries].

GeoTerraImage. (2020a). *Demographics estimates*. https://geoterraimage.com/tag/building-based-land-use/

GeoTerraImage. (2020b). *Land use data products: Building Base Land Use*. https://geoterraimage.com/tag/building-based-land-use/

GeoTerraImage. (2022). *Land use data products: Building Base Land Use*. https://geoterraimage.com/tag/building-based-land-use/

Hamann, C. and de Kadt, J. (2021). *GCRO Quality of Life Survey 6 (2020/21): Sample design*. Johannesburg: Gauteng City-Region Observatory (GCRO). https://d3iy114y1zl2zq.cloudfront.net/media/documents/2021.11.17_QoL_6_Sample_design.pdf

McCarthy, P.J. and Snowden, C.B. (1985). 'The bootstrap and finite population sampling'. *Vital and Health Statistics*, *2*(95), 1–23. PMID: 3992938.

Neethling, A. (2024). *Weighting report: Quality of Life 7 Survey (2023/24)*. Johannesburg: Gauteng City-Region Observatory. https://doi.org/10.36634/EOSA3095

Orkin, M. (2020). *Technical review of the GCRO QoL surveys: Synthesis report*. Johannesburg: Gauteng City-Region Observatory (GCRO). chrome-https://d3iy114y1zl2zq.cloudfront.net/media/documents/2020.11.20_Synthesis_report_formatted.pdf

About the authors

Laven Naidoo | https://orcid.org/0000-0002-7091-0566

Dr Laven Naidoo is a Senior Researcher at the GCRO and is the lead of the 'Data analytics, informatics and visualisation' research theme. He completed his PhD in Geoinformatics at the University of Pretoria in 2017 and specialised in using earth observation technologies to monitor savannah, forest and wetland vegetation. He has expanded his areas of interest into savannah tree species mapping using hyperspectral and Light Detection and Ranging (LiDAR) airborne sensors and assessing soil moisture and grass biomass in Afromontane wetlands using SAR technologies. Within the GCRO, his research is focused on the development of advanced AI/machine learning techniques, geospatial data science approaches and novel remote sensing applications within the urban environment and UN Sustainable Development Goal space. He is also currently a Y-rated scientist by the South African National Research Foundation and an extraordinary lecturer at the University of Pretoria's Geography, Geoinformatics and Meteorology Department.

Christian Hamann | https://orcid.org/0000-0002-2129-8550

Christian Hamann is a Researcher at the GCRO. He completed his undergraduate studies in Town and Regional Planning at the University of Pretoria before embarking on an Honours degree in Geography (BSocSci Hons), also at the University of Pretoria. He then enrolled for a Master's degree in Geography at the University of South Africa, which he completed at the beginning 2016. His research interests primarily relate to the GCRO's 'Social change' and 'Spatial transformation' research themes, but he enjoys engaging in a variety of projects related to analytics, cartographies and visualisations. His most recent work focused on socio-spatial change, specifically racial-residential segregation and socio-economic inequality. Christian is currently a registered PhD student at the University of Pretoria, researching the 'Spatial mechanisms that enable equality of opportunity and social mobility in Gauteng'.

Yashena Naidoo | https://orcid.org/0000-0003-3171-448X

Yashena Naidoo is a Junior Researcher at the GCRO, where her work focuses on data science and the use of data to understand urban issues. She obtained her Master's in Geoinformatics with distinction from the University of Pretoria. Yashena's current research focuses on the use of spatial analysis and remote sensing techniques to broaden the understanding of the socio-economic realities in urban environments, and she is interested in the use of data-driven approaches that can inform policy. She also has a keen interest in the use of machine learning and analytics to leverage open data within the geospatial domain to inform better decision making.

Annexure A: PostGreSQL and R automation scripts

Workflow tasks related to the data extraction and preparation were executed in a PostGreSQL script for each municipality and is outlined below.

```
--Extract residential buildings from overall dataset
SELECT *
INTO bblu_2022_residential_lesedi
FROM “GP_BBLU”
WHERE p_lu_code = ′7′ AND t_lu_code != ′7.3.7′ AND
t_lu_code != ′7.3.8′ AND
t_lu_code != ′7.4.1′ AND
t_lu_code != ′7.5.1′ AND
s_lu_code != ′7.9′
AND mn_name_16 = ′Lesedi′;


--generate duplicate points based on final units
SELECT * ,
generate_series(1, final_units) as duplicate_id
INTO bblu_2022_residential_duplicates_lesedi
FROM bblu_2022_residential_lesedi;

--add lat and long
SELECT *, ST_X(geom) AS “x”, ST_Y(geom) as “y”
INTO bblu_2022_residential_duplicates_lat_long_lesedi
FROM bblu_2022_residential_duplicates_lesedi;

--clip cluster
SELECT *
INTO cluster_lesedi
FROM clusters_4326
WHERE mn_name = ′Lesedi′

-- join points and cluster data
SELECT a.*, b.objectid_1 AS “cluster_id”, b.wardid
INTO bblu_cluster_join_lesedi
FROM bblu_2022_residential_duplicates_lat_long_lesedi a, cluster_lesedi b
WHERE ST_Contains(b.geom, a.geom)

-- count per cluster
SELECT cluster_id, ea_code, wardid, COUNT(*)
INTO lesedi_cluster_count
FROM  bblu_cluster_join_lesedi
```

```
GROUP BY cluster_id, ea_code, wardid

--remove anything with count less than 12
SELECT *
FROM lesedi_cluster_count
WHERE count >= 12

SELECT * FROM bblu_cluster_join_lesedi
```

Workflow tasks related to the bootstrapping and cluster selection were executed in an R script for each municipality and is outlined below. For the following R scripts to work, the R packages indicated within the library() and require() commands need to be installed within the local instance/installation of R and/or R-studio.

```
library(dplyr)
library(tidyr)
library(tidyverse)
require(data.table)

# Define the function for nested random sampling
nested_random_sampling <- function(data, group_col, sample_size_per_group) {
 unique_groups <- unique(data[[group_col]])
 sampled_data <- NULL

 for (group in unique_groups) {
  group_data <- filter(data, !!sym(group_col) == group)

  if (nrow(group_data) < sample_size_per_group) {
   warning(paste(“Group”, group, “has fewer observations than the sample size per group.”))
  }

  if (nrow(group_data) >= sample_size_per_group) {
   sampled_group_data <- group_data %>% sample_n(sample_size_per_group, replace =
TRUE)
   sampled_data <- rbind(sampled_data, sampled_group_data)
  }
 }

 return(sampled_data)
}

# Read the CSV file
df <- read.csv('ekurhuleni_cluster_count_final.csv', header = TRUE)
```

```
# Number of iterations
num_iterations <- 1000

# Sample size per group
sample_size_per_group <- 1

# Create an empty list to store sampled data from each iteration
all_samples <- list()

# Iterate the sampling process num_iterations times
for (i in 1:num_iterations) {
 site_sample <- df %>%
  nest_by(ward_id_20) %>%
  summarise(data = list(slice_sample(data, n = sample_size_per_group, replace = TRUE)))

 # Store the sampled data in the all_samples list
 all_samples[[i]] <- site_sample %>% unnest(data)
}

# Combine all the sampled data into a single data frame
output <- bind_rows(all_samples)

write.csv(output,”ekurhuleni_cluster_count_final_Random1000.csv”)
```

The frequency at which each cluster was selected within the bootstrapped sampling result (from the above script) was summarised using a pivot table in Excel (‘cluster ID’ and ‘Count.of.count’ or frequency) and saved as “prefix_Boot_Random1000.csv” for the following process. Related workflow tasks were executed in an R script for each municipality and are outlined below:

```
library(tidyverse)
library(readxl)
library(dplyr)
library(data.table)

#Read in two tables for VLOOKUP
Boot <- read.csv(“eku_Boot_Random1000.csv”)
GPS <- read.csv(“ekurhuleni_bblu_cluster_join1.csv”)

#Additional Cleaning of BBLU layer
GPS1 <- GPS[!(GPS$final_lu==“7.3.4”),]

#VLOOKUP to add Cluster ranking (i.e. Count.of.count) from Bootstrapping Script
VLUP <- merge(GPS1, Boot, by.x = “cluster_id”, by.y = “Row.Labels”, all.x = TRUE)
```

```
#Generate random number column for randomization of building unit points
n <- nrow(VLUP)
VLUP1 <- VLUP %>%
    mutate(RANDOM = runif(n, min = 0, max = 1))

#Sort data and column trimming
Final_Table <- arrange(VLUP1, ward_id_20, desc(Count.of.count), cluster_id,
desc(RANDOM))
Final_Table1 <- na.omit(select(Final_Table, cluster_id, id, gti_id, final_lu, ward_id_20,
munic_name, x, y, Count.of.count, RANDOM))

#Primary and secondary labeling for priority clusters and building unit points
Final_Table2 <- Final_Table1 %>%
 group_by(ward_id_20) %>%
 mutate(subgroup_index = cumsum(c(1, diff(cluster_id) != 0))) %>%
 mutate(Cluster_Priority = ifelse(subgroup_index <= 6, “Primary”, “Secondary”)) %>%
 ungroup() %>%
 group_by(cluster_id) %>%
 mutate(Point_Priority = ifelse(row_number() <= 4, “Primary”, “Secondary”)) %>%
 ungroup()

#Filtering out primary and secondary clusters
Table_P <- filter(Final_Table2,Cluster_Priority == “Primary”)
Table_S <- filter(Final_Table2,Cluster_Priority == “Secondary”)

#Output
write.csv(Table_P,”CoE_Primary_List.csv”)
write.csv(Table_S,”CoE_Secondary_List.csv”)
```

GAUTENG CITY-REGION OBSERVATORY

6th Floor University Corner
11 Jorissen St (Cnr Jorissen and Jan Smuts)
Braamfontein
Johannesburg
Gauteng, South Africa

tel +27 11 717 7280
email info@gcro.ac.za
www.gcro.ac.za