DATAWORLD

think innovate create

Data Report:
Second Quality of Life Survey

conducted on behalf of the

Gauteng City Region Observatory (GCRO)

DOCUMENT NO:	1
VERSION:	0.1
CONTACT:	Adhir Nursayhe
EMAIL:	adhir@dataworld.co.za
DATE:	09 March 2012

Table of Contents

1. Introduction

This data report should be read in conjunction with the completed SPSS dataset, which contains the data for 16729 successful surveys. The purpose of this report is to outline changes which were made to the data during the data validation stage.

While the general consensus from fieldworkers was that the questionnaire was fairly straightforward in terms of flow, it does appear that some of the skip patterns were not followed correctly in certain cases Based on these cases, changes were made to the questions which would lead to either continued flow or skip patterns being applied to subsequent questions.

All variables were checked by Ross Jennings on behalf of GCRO, and cleaned by Data World accordingly. Among the checks conducted was to verify that all variables were within the correct ranges, and that the value labels per variable were correct as per the survey instrument.

Fieldwork was conducted between 15 August 2011 and 15 December 2011, which included work on public holidays and certain Sundays. Lost field time due to inclement weather was minimal, with fieldwork being halted on only one day due to severe rain.

Taking the size of the sample as well as number of variables into account, the number of missing data per question was very minimal. Checks were done against missing data in all variables, which entailed the rechecking of hardcopy survey forms, as well as making telephonic contact with respondents for verification.

2. Universal codes in the data

-1 – No response (indicates that the respondent has not provided an answer)
-2 – No response due to skip scenario (indicates non responsiveness due to the question being skipped based on answers to previous questions)

3. Data checks and updates

General checks

Monetary questions (1_19 and 4_16) were checked for format used, to ensure that fieldworkers did not add extra zero's to denote cents as well. The rule which was given to fieldworkers in training was that figures should be rounded to at least the nearest 50, and this was followed in most cases.

Place code / place name questions (2_4,2_8,4_2) were checked for validity of place codes used, as well as place names which were written even though codes could be assigned. Where place codes could be assigned, they were, and the place name captured was blanked out as a result. Where invalid place codes were used, these were removed.

Municipal code questions (2_8,4_2) were checked for correctness of municipal code, as the place code booklets had different sets of codes for the different questions and this may lead to come confusion with the field workers. Where the wrong set was used, these were updated accordingly.

Time questions which required HH:MM formatting (4_3,4_14,4_15) were checked to confirm correctness, with reformatting being done on all the answers to ensure all were in the correct format.

Other/specify questions (6_3 , 8_7) were checked to confirm uniformity in case and spelling of respondent specified answers.

Country and place names (2_5,2_8,4_2) were checked to confirm uniformity in case and spelling, and updated accordingly.

Question 6_23, while ideally meant to be a single response variable, was recorded as multiple response in a few cases. This is thought to have arisen from the phrasing of the question, which says “Preferred, single mention only”.

Age question (8_2) was checked to confirm format used, as although the question pertains to age in years, the format in the instrument is “YYYY”. The age in years was calculated for those records where age was listed as an actual year (which denoted when the business was started), as opposed to the age in years (which should denote how old the business is).

All questions which required a number input from the respondent (such as how many refuse bags are used, how many people live in this household) were checked against hardcopy surveys as well as by telephonic contact with respondents. In cases where values are high for number of people in household, these were verified as well by confirming the number of respondents listed for application of the birthday rule.

Respondent age (12_2) was checked against the birth year given when the birthday rule was applied to select respondents, with the only noted variance being 1 year (this is assumed to stem from respondents either counting or not counting the current year as their birthday would not have yet occurred based on application of the birthday rule).

Filling in of missing data for skip questions

The following changes were made to questions where there was no response (-1), however by following the flow of the questionnaire it was apparent that the answer to the question under consideration could be ascertained. This was done in consultation with and approved by GCRO.

1.4 Updated 127 records to “Other” based on 1.5 and 1.6 being answered.
1.14 Updated 235 records to “No” based on 1.15 being answered.
4.1 Updated 70 records to “Other Purpose” based on 4.2 until 4.7 being answered.
4.9 Updated 387 records to “Yes” and 79 to “No” based on answers to 4.10/ 4.11
4.12 Updated 162 records to “Yes” based on answers to 4.13
5.24 Updated 57 records to “Yes” based on answers to 5.25
6.2 Updated 41 records to “No” based on answers to 6.3
6.15 updated 62 records to “Yes” based on answers to 6.16
6.16 Updated 16 records to “Yes” based on answers to 6.17
6.20 Updated 50 records to “Yes” based on answers to 6.21
8.1 Updated 2 records to “Yes” based on answers to 8.2
8.4 Updated 4 records to “Yes” based on answers to 8.5.

- For questions which were meant to be skipped, responses were set to “-2” to denote reason for no response.

4. GIS and co-ordinates

During the fieldwork process, GPS co-ordinates were taken at every household surveyed. The GPS co-ordinates were attached from the cellular phone when forms were sent via digital pen to the cellular phone, and then onto the database via GPRS/3G connection.

This minimized the risks associated with manually capturing co-ordinates, as no co-ordinates were hand captured - the field worker would only need to ensure that the cellular phone returned a success message when the forms were sent.

The survey instrument also required capture of fields such as municipal code, ward number, as well as street address. The format for street addresses was trained in detail with the fieldworkers, and examples of how to record addresses for the various dwelling types (houses, flats , informal settlements) are provided in the training and user manuals for field workers.

For the 16729 surveys completed, a spatial analysis was done to verify that surveys were located within the correct wards as recorded on field. It was found that 1008 surveys fell in the correct municipality, however not within the correct ward. These 1008 records were not removed from the dataset, however a decision was taken by the GCRO to weight these specific records at municipal (as opposed to ward) level.

The balance of 15721 surveys were within the correct wards, and were weighted at ward level.

There were a few reasons as to why co-ordinates did not come through directly from the field, mainly:

Collectors were in unsafe areas and reluctant to remove devices from their bags / pockets, hence GPS co-ordinates were not sent.

There were a few wards where GPS and cellular signal were unavailable.

Field workers were not waiting to get a GPS signal lock on their device prior to sending forms.

Forms were sent in manually, that is not via the digital pen and cellular phone. Reasons for this include devices running out of battery life, as well as devices being damaged while on the field.

For surveys which were sent in manually, each survey was manually geocoded utilizing a variety of data sources which included:

- Google geocoding
- Bing geocoding
- Yahoo geocoding
- National Address database (NAD Layer)
- Streets Database

Prior to being run through the above processes, addresses were corrected by adjusting spellings on street names, suburbs and towns.

Checks were also done when verifying ward accuracy, which entailed comparing teams and collector's daily routes and work done. As an example, if a collector had done 5 surveys within a ward for the day, but survey number 3 had issues with co-ordinates yet had a proper street address, it was assumed the survey was done within the correct ward, and it was geocoded accordingly.

When geocoding, it was found that 2212 addresses were not geocodable to address or street level. These were mainly due to unknown street names being used, as well as surveys being done within informal settlements or blocks of flats. These 2212 surveys were rechecked to confirm that they were within the correct wards, and once confirmed the centre of the ward was used as the co-ordinate for the survey. It was noted that especially in township areas, the street names provided by the various geocoding sources are not that same as the street names on the ground, which makes geocoding to this street level impossible.

Where addresses could not be found but either the street or suburb was located within the ward recorded, the addresses were geocoded to these levels. These were 1048 at street level and 1117 at suburb level.

For the surveys with municipal level accuracy, the main reasons noted for the accuracy issues was that the field workers had crossed into neighbouring wards, or that the suburbs which were used for geocoding crossed over ward boundaries, or that streets which were used for geocoding crossed over ward boundaries.

The table below shows the total successful surveys, as well as the source of their GPS co-ordinates, against the precision level (either ward or municipal level).

	GPSSource				
	GPS Device	**Geocoded to Street**	**Geocoded to Suburb**	**Geocoded to Ward Centroid**	**Total**
Ward Level	11386	1048	1117	2170	**15721**
Munic Level	160	88	718	42	**1008**
					16729

5. Conclusion

The overall number of wards sampled was 507, out of the possible 508 wards in the study area under consideration, with the ward excluded being one made up of holiday homes and non residents. As covered in the weighting document, two levels of weighting were applied to the surveys, which were based on either ward level (15721 surveys) or municipal level (1008 surveys).