Presentation to

"Seminar for innovative approaches to turn statistics into knowledge"

Using OLAP cubes as a method of information delivery in Statistics South Africa household surveys

Thanyani Maremba

Statistics South Africa

Methodology and Evaluation

Cape Town

09 December 2010

Overview

❶ **Introduction**

❷ **Organisational problem**

❸ **Proposed solution**
- Generic view of multi-dimensional data
- Creation of data marts
  - Response mart;
  - Slippage rate mart;
  - Unemployment rate mart;
  - Sample account mart and;
  - In-coming and Out-going Quality

❹ **Implementation and results**

❻ **Target audience**

❼ **Potential areas of application**

❺ **Future improvements**

❽ **Conclusions**

❾ **Acknowledgements**

Thanyani A Maremba

Introduction

- Lessons learnt from Business Intelligence (BI) to use internal data in building intelligence

- BI tools are commonly used in profit driven organisations such as banks, insurance companies

- These techniques are used to give companies competitive edge by using the data from within to improve profits

- Statistics South Africa is quality driven organisation

- BI techniques can be used to enhances quality of our products by providing relevant quality indicators

- Decision makers will put measures that will enhance quality

Thanyani A Maremba

Introduction (Cont)

- We apply this principle in one of Stats SA survey that is Quarterly Labour Force Survey (QLFS)

- We identify data needs from the regular requests by the management such as slippage rates, response rates, imputation rates etc).

- We create data marts from the series of surveys or by quarters in the case of QLFS with specific topics such as slippage rates

- The data marts are created with imbedded multi-dimensional structures (they are updated by appending every new quarter' data)

- To view this multi-dimensional data sets we implement the view using Online Analytical Processing (OLAP).

Organisational problem

- Emphasis in Stats SA survey is on the publication targeted to the public

- Survey quality indicators such as response rates, slippage rates as well as measures of precision are produced with particular publication

- The challenge is to study how surveys are performing overtime

- We currently do not have quality indicators available to share with the relevant users in a single view for previous surveys

Thanyani A Maremba

- To meet the request for the required indicators overtime we generate adhoc reports using SAS to extract data from various sources on the server and tabulate it according to the request

- The process of generating those reports is time consuming

- Survey methodologists can save time spent in generating reports by implementing the proposed OLAP system

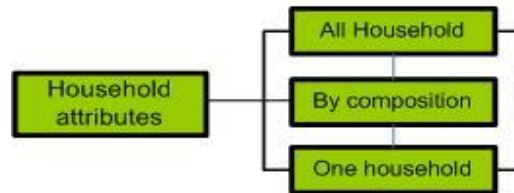Generic view of multi-dimensional hierarchical data structure

Thanyani A Maremba
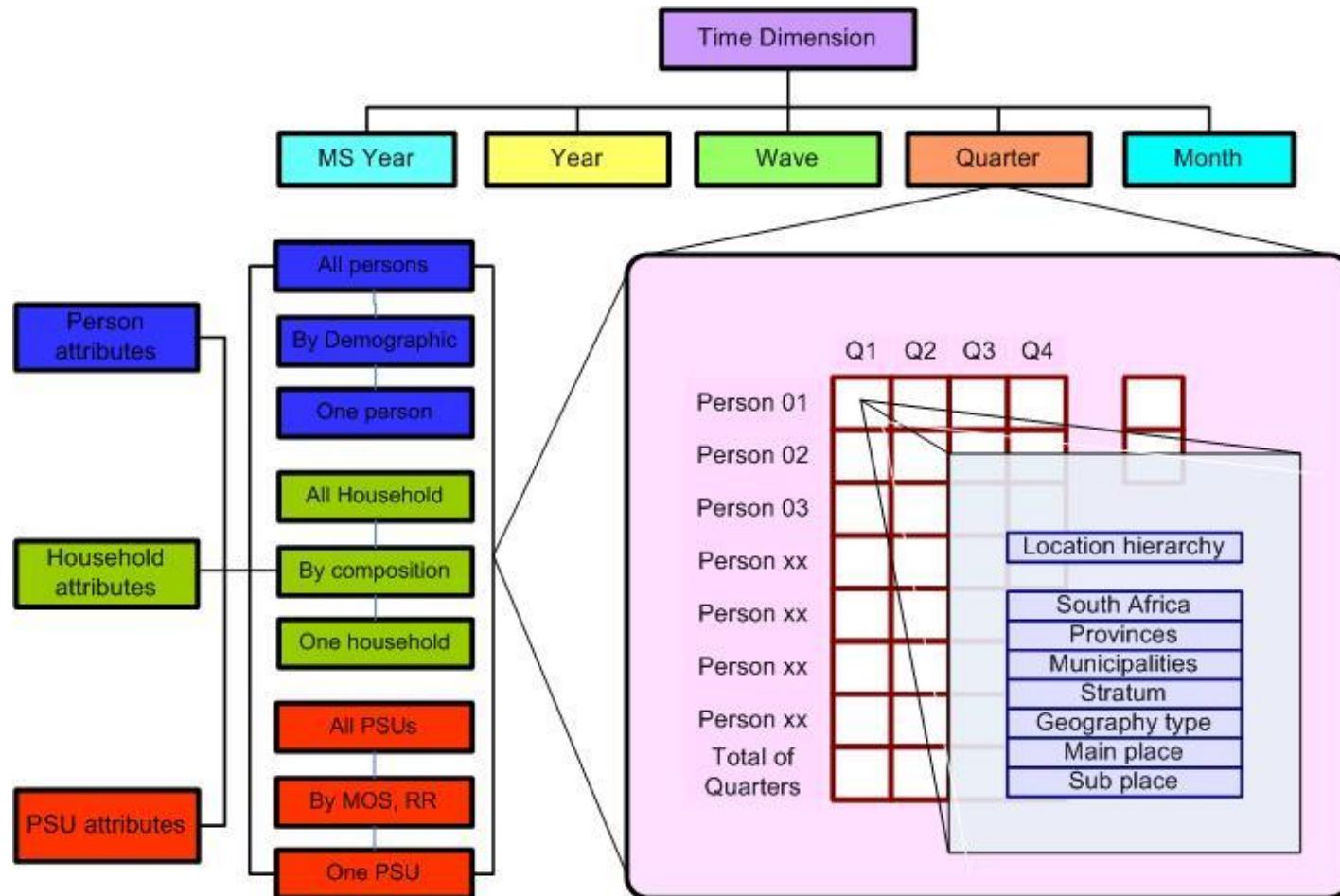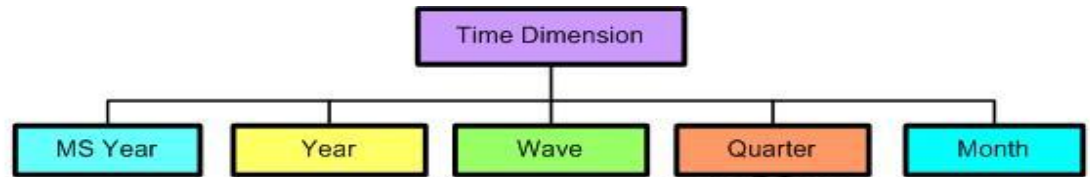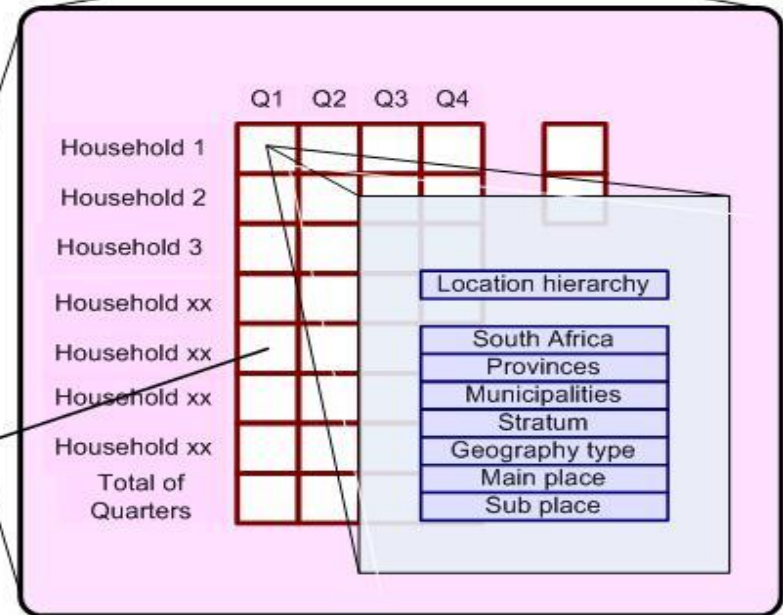
Creation of data marts (cont)

– Response mart;

$$Weighted\_RR = 100 * \left(\frac{\sum w_i response}{\sum w_i \left(response + nonresponse \right)} \right)$$

$$Unweighted\_RR = 100 * \left(\frac{response}{response + nonresponse} \right)$$

Time Dimension

MS Year | Year | Wave | Quarter | Month

Household attributes

All Household

By composition

One household

Q1 Q2 Q3 Q4

Household 1
Household 2
Household 3
Household xx
Household xx
Household xx
Household xx
Total of Quarters

Location hierarchy

South Africa
Provinces
Municipalities
Stratum
Geography type
Main place
Sub place

Measures:
1. Response
2. Non-responses
3. Weighted response rate
4. Unweighted response rates

Thanyani A Maremba
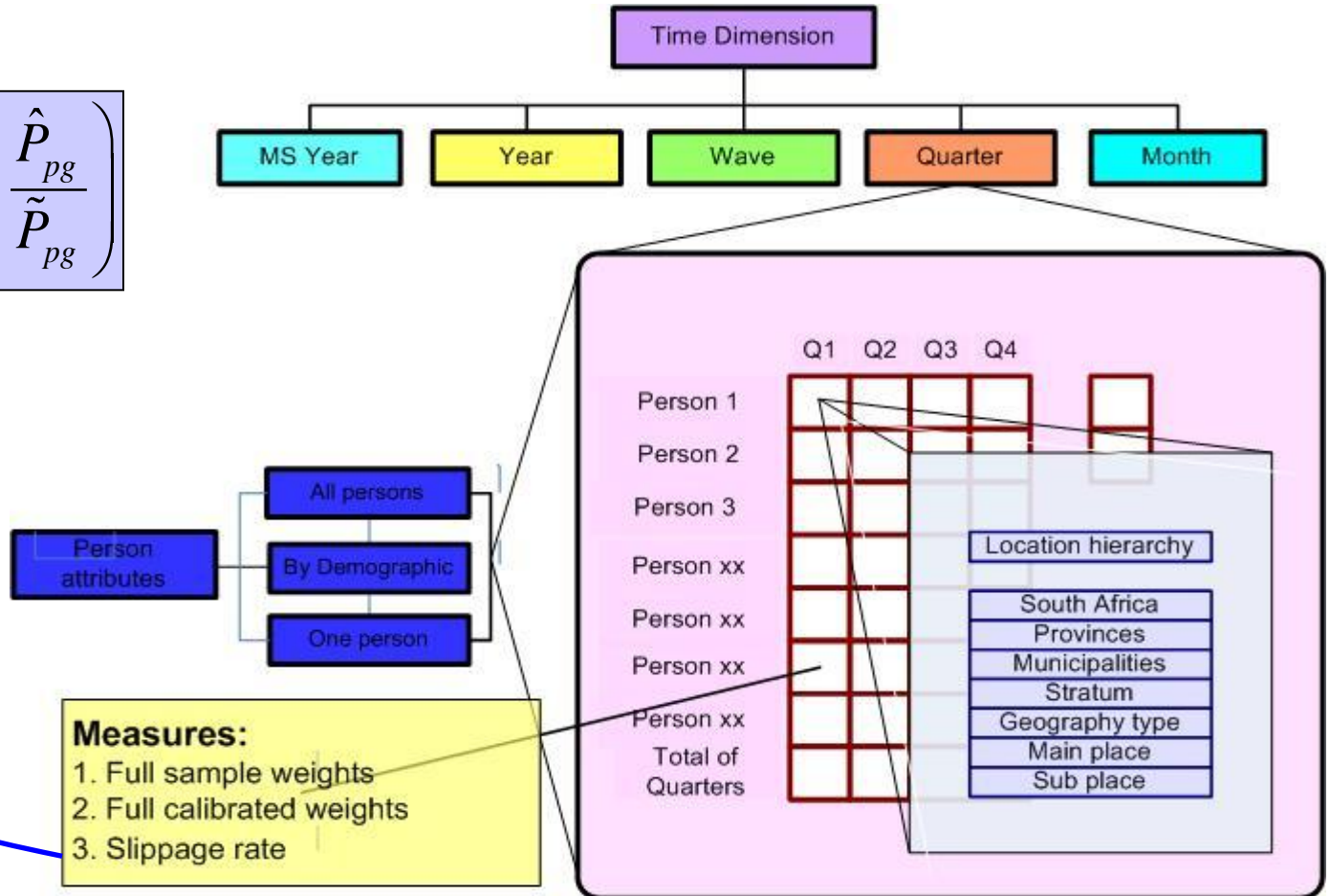
Creation of data marts (cont)
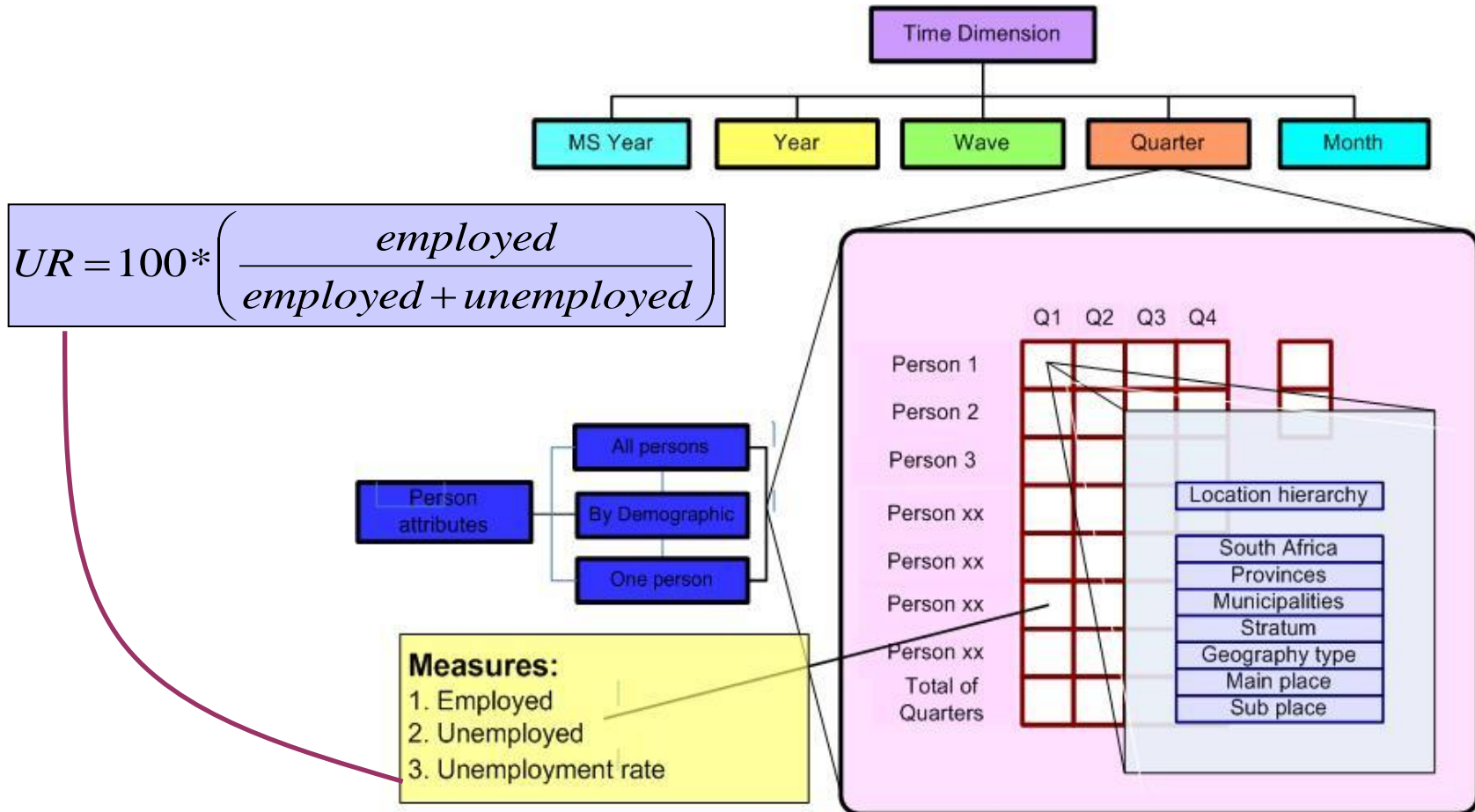
– Slippage rate mart;

$$SR_{pg} = 100 \times \left(1 - \frac{\hat{P}_{pg}}{\tilde{P}_{pg}} \right)$$

Time Dimension

MS Year	Year	Wave	Quarter	Month

Q1 Q2 Q3 Q4

Person 1
Person 2
Person 3
Person xx
Person xx
Person xx
Person xx
Total of Quarters

Person attributes
- All persons
- By Demographic
- One person

Location hierarchy
- South Africa
- Provinces
- Municipalities
- Stratum
- Geography type
- Main place
- Sub place

Measures:
1. Full sample weights
2. Full calibrated weights
3. Slippage rate

Thanyani A Maremba

Creation of data marts (cont)

– Unemployment rate mart *(with measures of precision)*

$$UR = 100*\left(\frac{employed}{employed+unemployed}\right)$$

Thanyani A Maremba

Creation of data marts (cont)

– Sample account mart;

This is an Example of a data mart which is not presummarised

Person level files comes from the following sources:
1. Data processing center
2. Editing and imputation system
3. Weighting and estimation system

Person attributes — All persons, By Demographic, One person

Time Dimension — MS Year, Year, Wave, Quarter, Month

Q1 Q2 Q3 Q4 — Person 1, Person 2, Person 3, Person xx, Person xx, Person xx, Person xx, Total of Quarters

Location hierarchy
South Africa
Provinces
Municipalities
Stratum
Geography type
Main place
Sub place

Measures:
1. Number of PSUs
2. Number of dwelling units
3. Number of households
4. Number of persons

Thanyani A Maremba
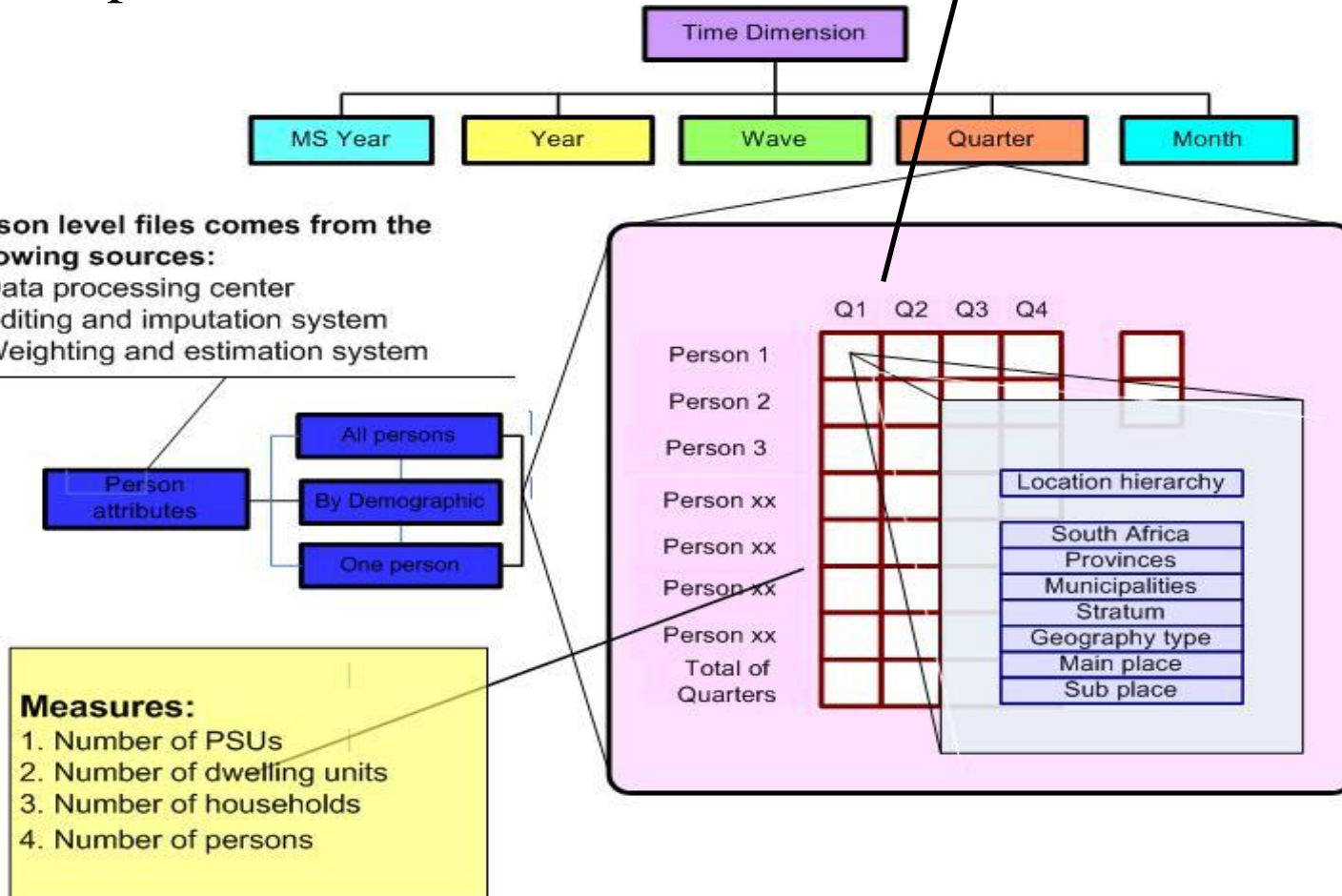
Creation of data marts (cont)

– In-coming and Out-going Quality

$$Verification - Rate = 100 \times \frac{\sum_{b=1}^{B} n_b}{\sum_{b=1}^{B} N_b}$$

$$In - Quality = 100 \times \frac{\sum_{b=1}^{B} w_b d_b}{\sum_{b=1}^{B} w_b c_b}$$

$$Out - Quality = 100 \times \frac{\sum_{b=1}^{B} \left(w_b^* - 1\right) d_b^*}{\sum_{b=1}^{B} w_b^* c_b^*}$$

Time Dimension

MS Year | Year | Wave | Quarter | Month

This process is conducted at the data processing center during capturing

Sample of questionnaires — All questionnaires / By batches / One questionnaire
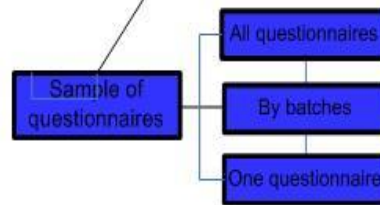
Measures:
1. Verification rate
2. In-Quality
3. Out quality
4. Number of sampled questionnaires

Q1 Q2 Q3 Q4

Questionnaire 1
Questionnaire 2
Questionnaire 3
Questionnaire x
Questionnaire x
Questionnaire x
Questionnaire x
Total of Quarters

Location hierarchy

South Africa
Provinces
Municipalities
Stratum
Geography type
Main place
Sub place

Thanyani A Maremba
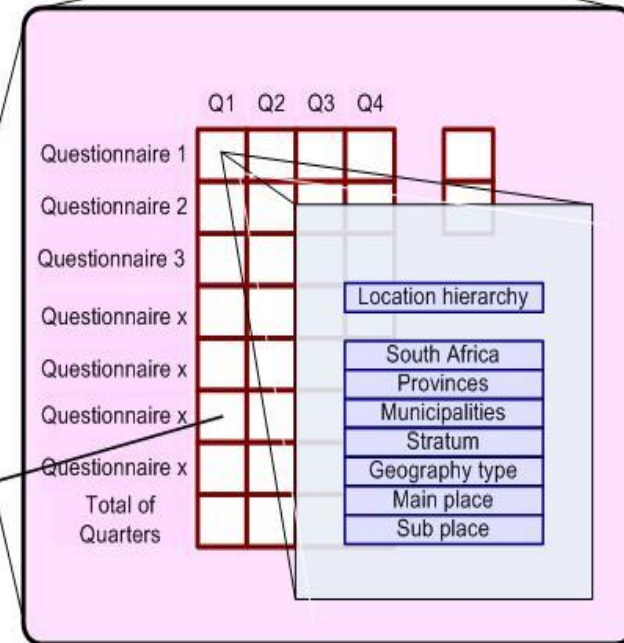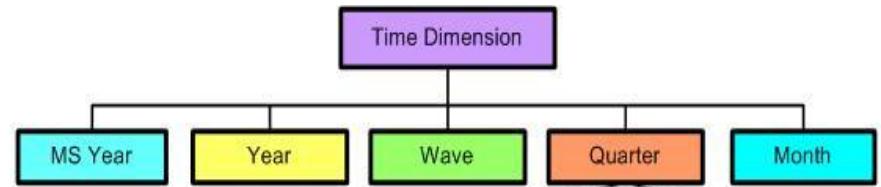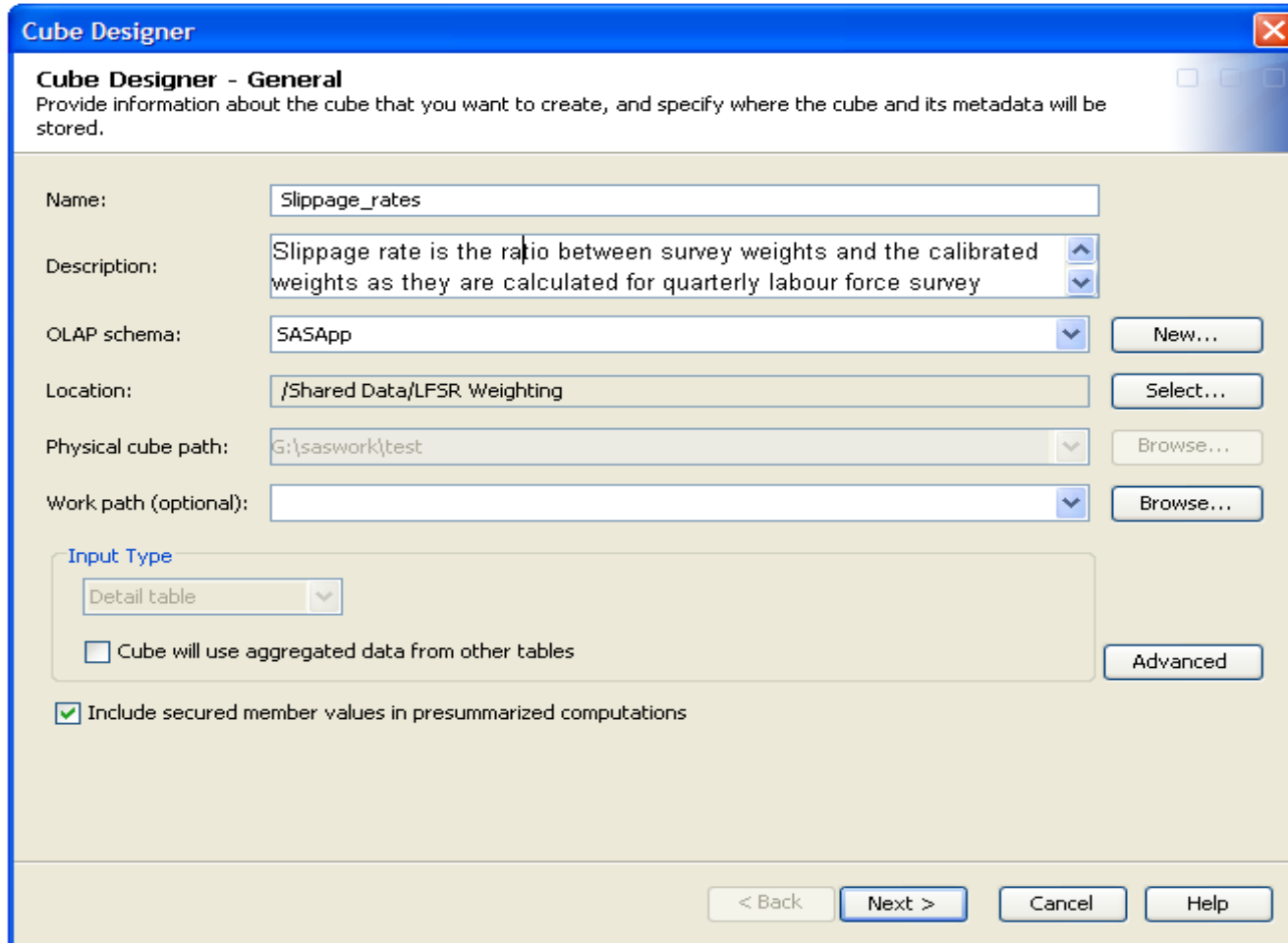
Creation of Cube: SAS OLAP Cube studio

- Cube designer wizard

- Viewing a cube in:

    - SAS Enterprise guide

    - MS Excel

    - SAS Web OLAP Viewer for Java

    - SAS Web OLAP Viewer for .NET

Thanyani A Maremba

Cube creation: SAS OLAP Cube studio

– Naming and describing the cube

Thanyani A Maremba

Cube creation: SAS OLAP Cube studio

– Selecting input dataset

Thanyani A Maremba
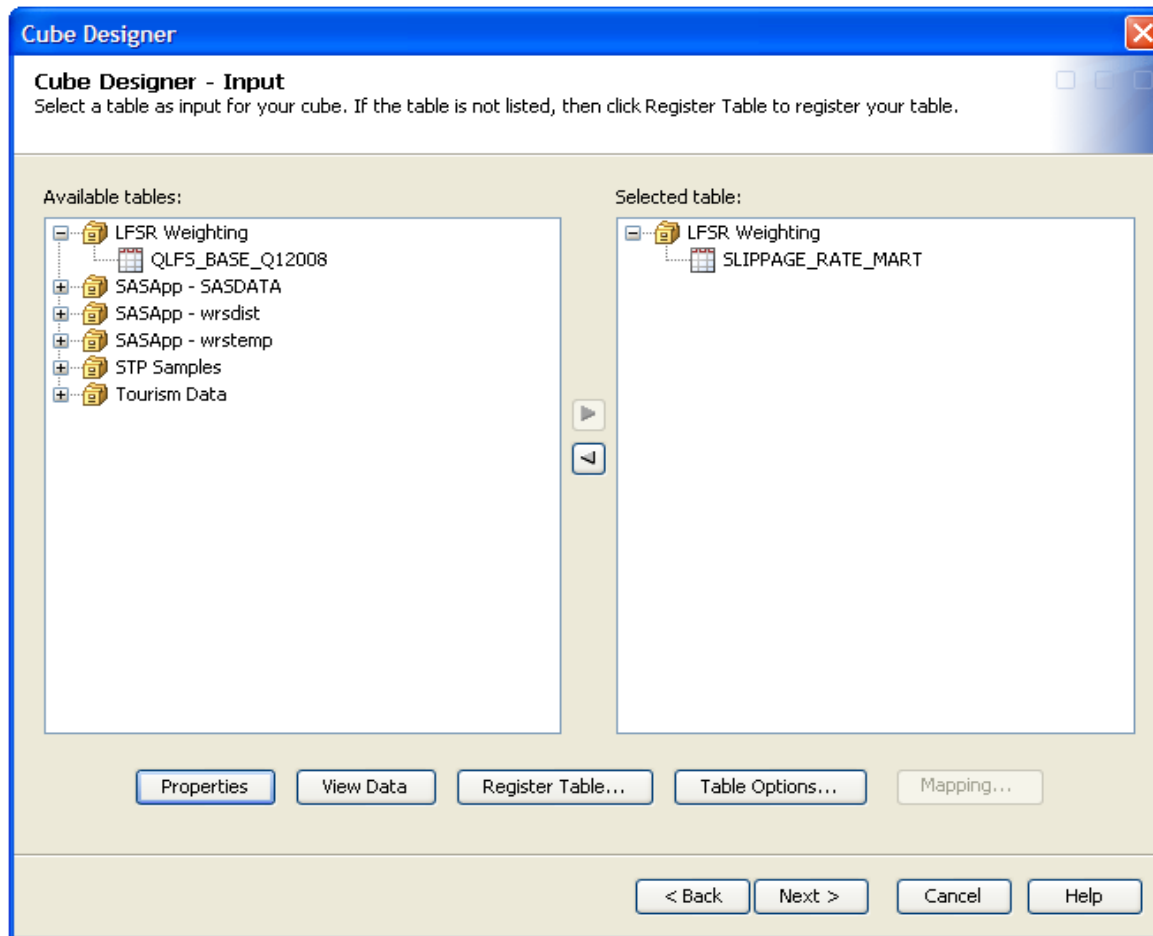
Cube creation: SAS OLAP Cube studio

– Defining dimensions
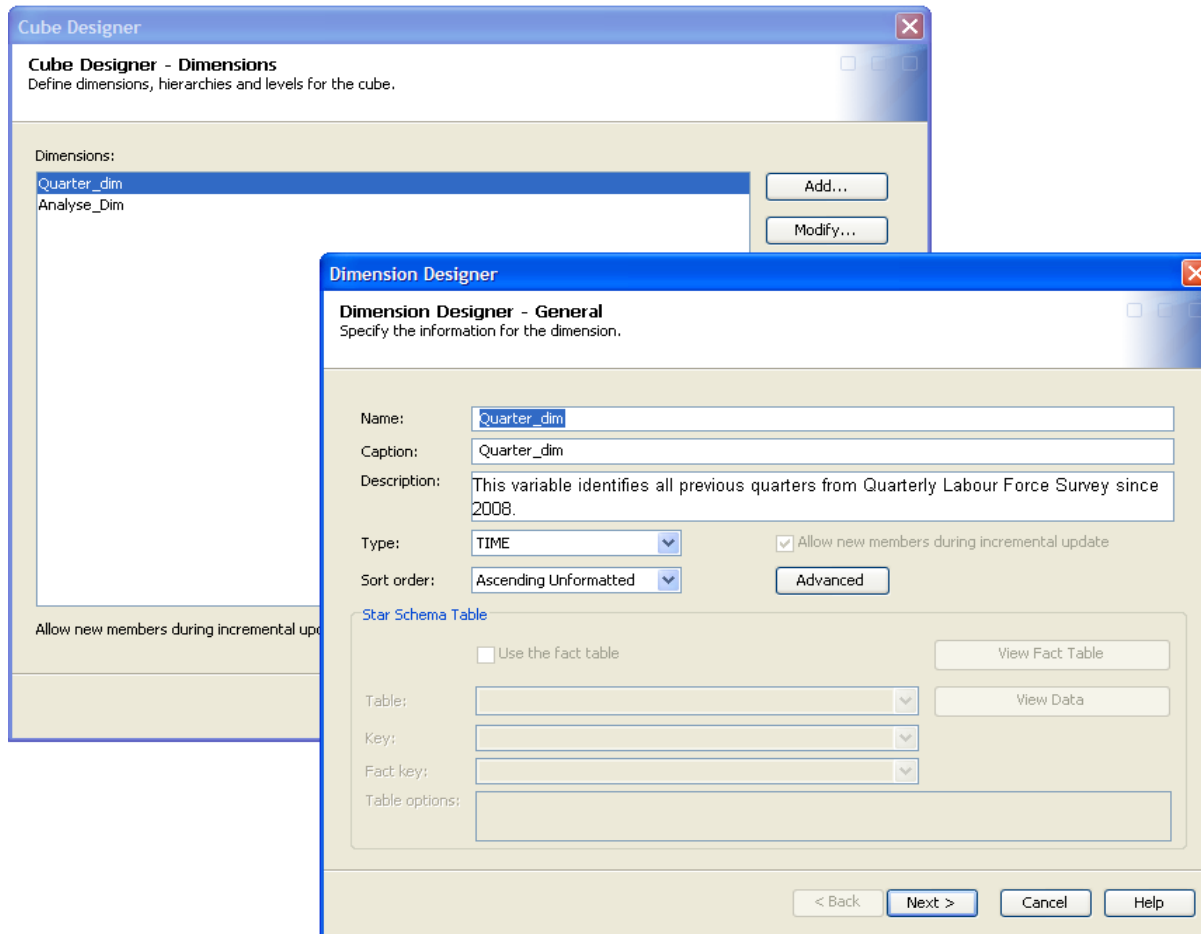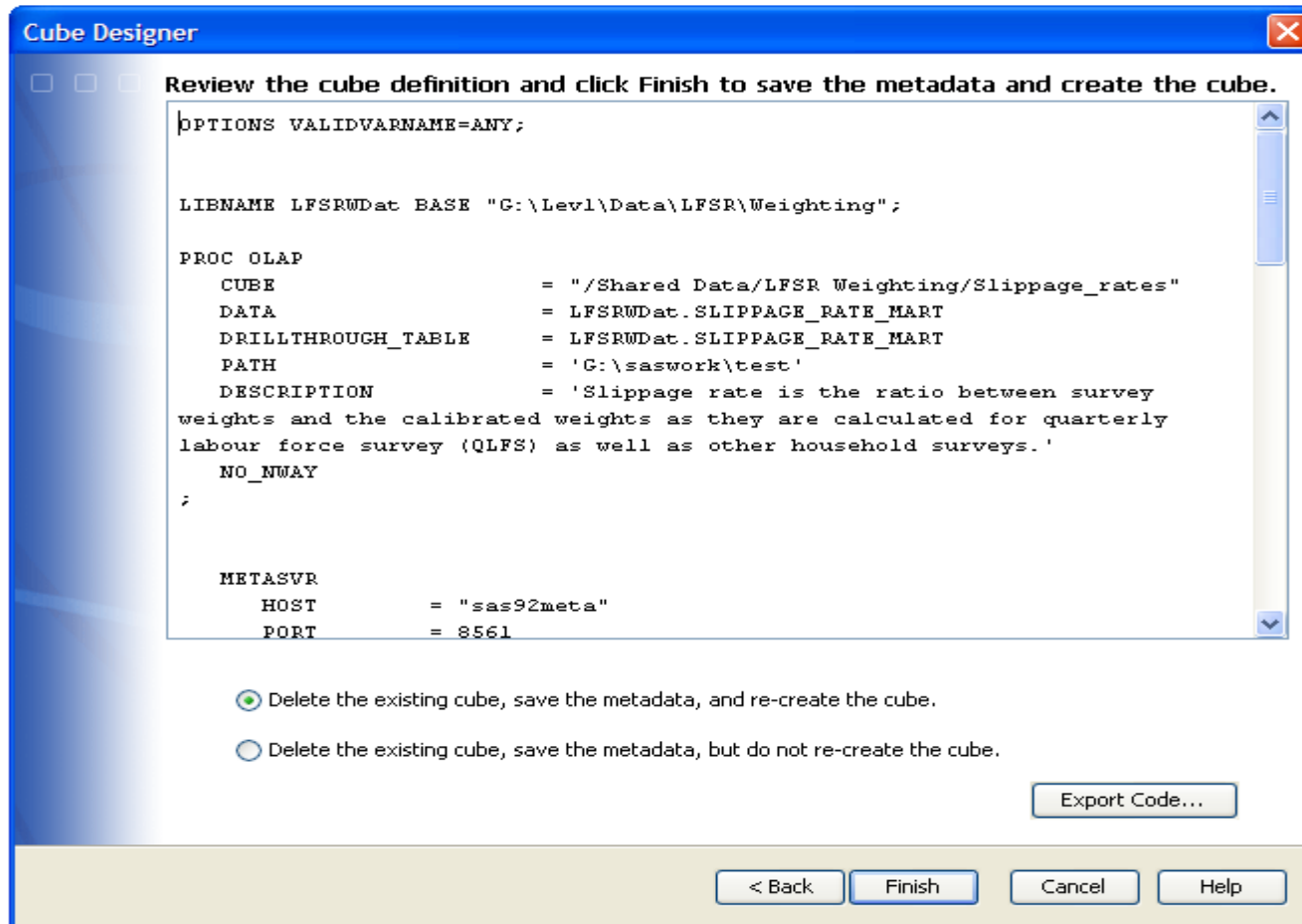
Cube creation: SAS OLAP Cube studio

– Defining measures
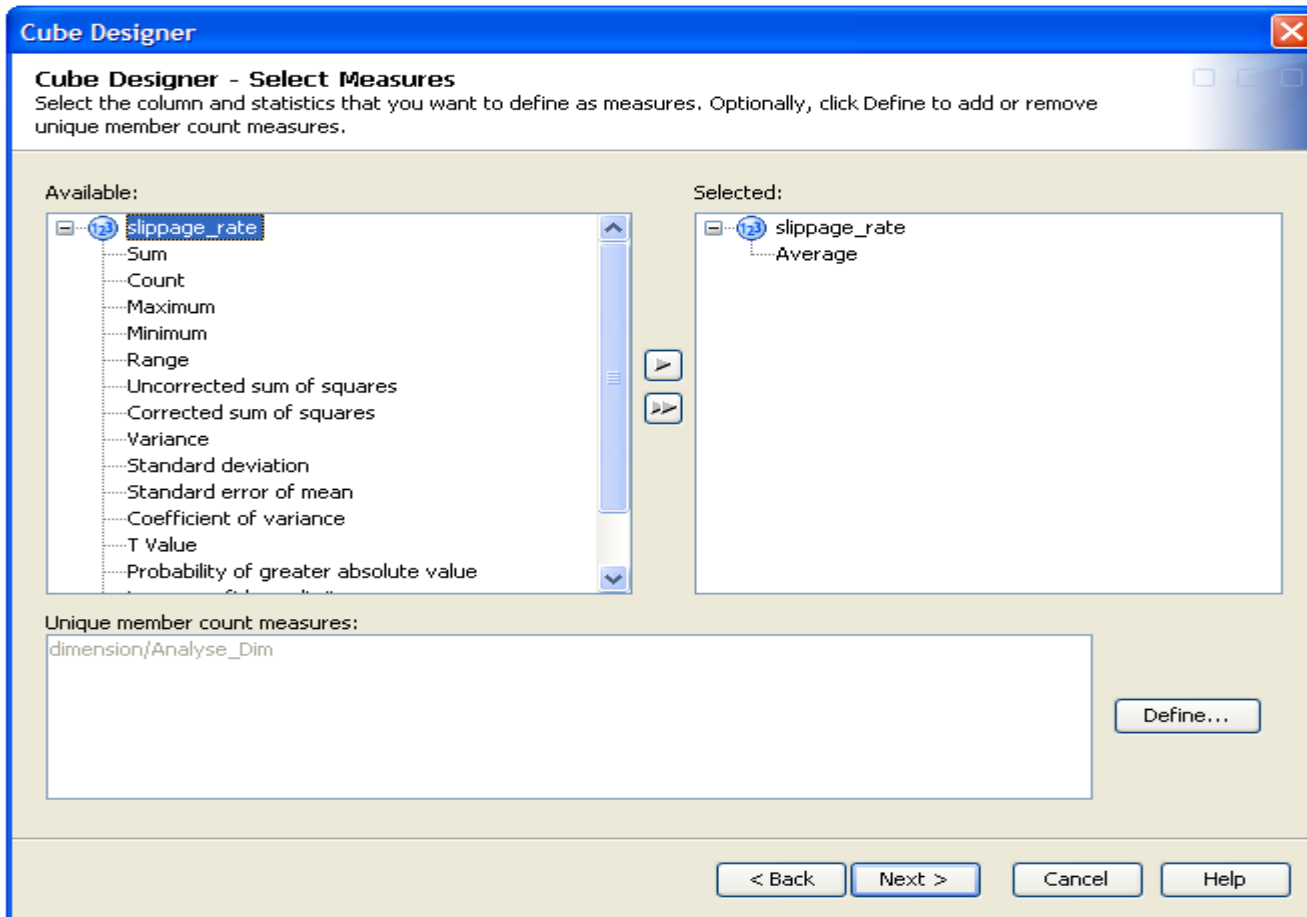
Cube creation: SAS OLAP Cube studio

– Review the cube definition

Thanyani A Maremba

Viewing cubes: SAS Enterprise Guide (Table form)

Editing and Imputation system: Source1 = 02
Weighting System: Source1 = 01

This table accounts for all persons records through the above systems from Q1 2008 to date

Note: Actual figure not displayed due to confidentiality

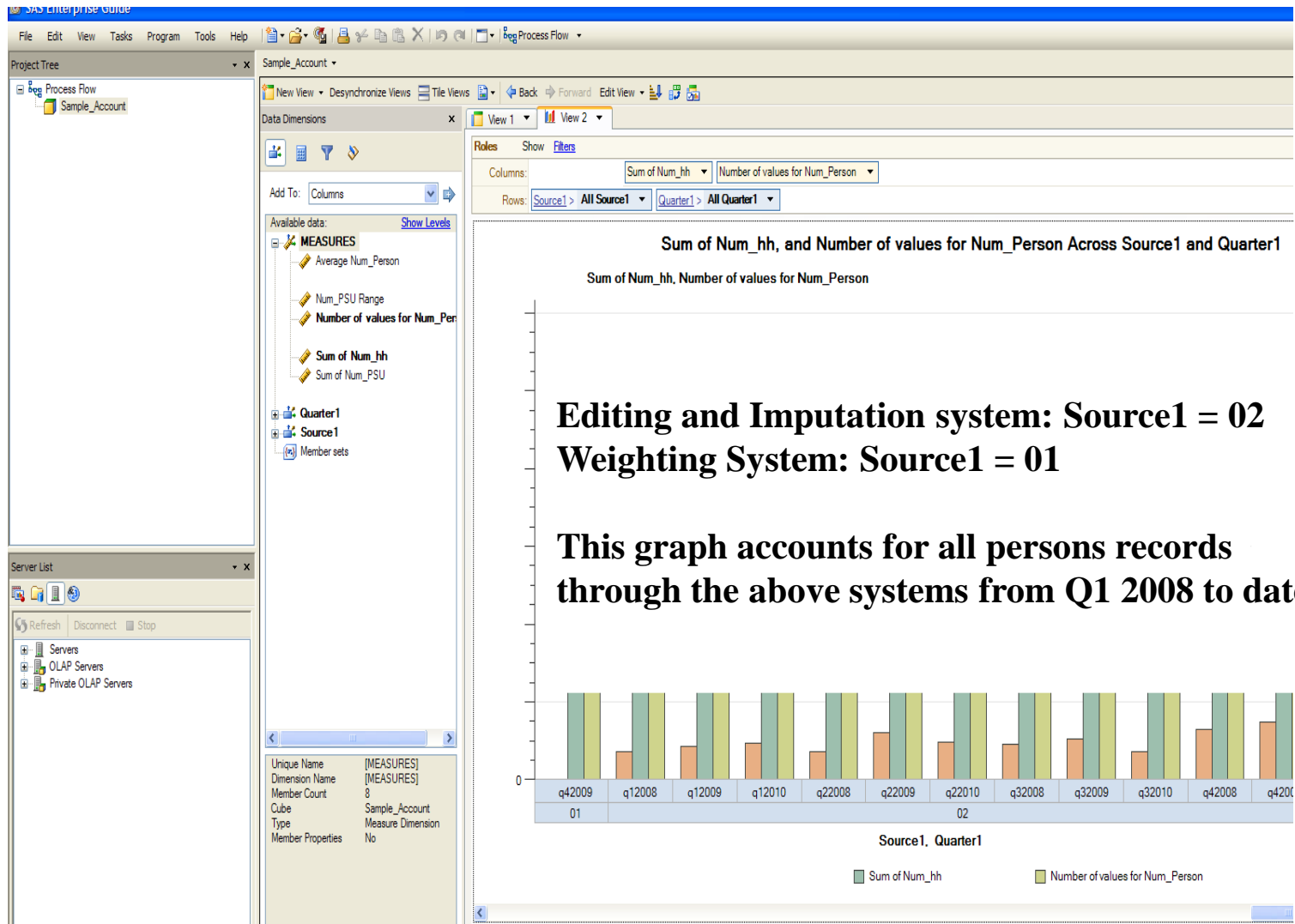## Viewing cubes: SAS Enterprise Guide (Graph form)

Editing and Imputation system: Source1 = 02
Weighting System: Source1 = 01

This graph accounts for all persons records through the above systems from Q1 2008 to date

Viewing cubes: MS Excel (Table and Graph)

- *This trend confirms that people with ages (15 to 34) are hard to count*
- *This is more prevalent with males than females*
- *This is not the findings of this data set only, it is knowledge gained from historical data*

Average_Slippage_rate		
dimension ▼	quarter ▼	Total
G_Female15-34	I_q12010	5.2
	J_q22010	5.1
	K_q32010	5.6
G_Female35-64	I_q12010	3.0
	J_q22010	3.1
	K_q32010	3.8
G_Male15-34	I_q12010	5.4
	J_q22010	5.4
	K_q32010	6.2
G_Male35-64	I_q12010	4.5
	J_q22010	3.7
	K_q32010	4.5

Note: Data for this cube was simulated due to confidentiality

Thanyani A Maremba
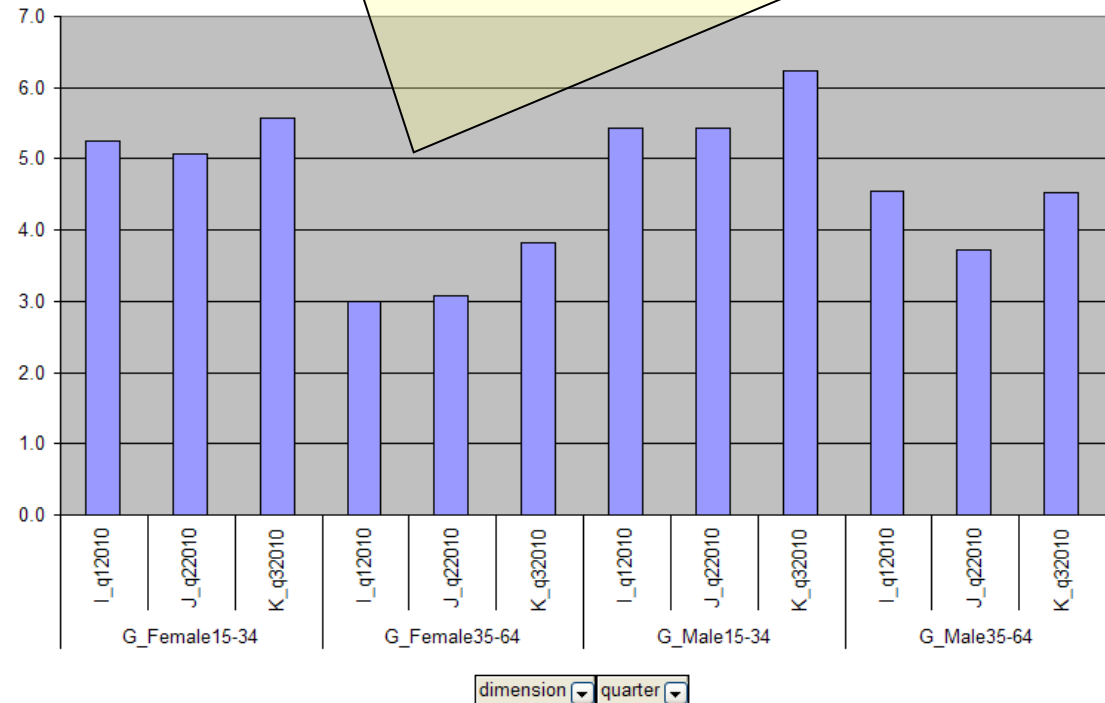
Implementation and results (cont)

Viewing cubes: MS Excel (Table and Graph)

- *Based on the first three quarters of 2008 we observe the similar pattern*
- *There are several queries that can be done on the cube to reveal more patterns in the underlying data set over time*

Average_Slippage_rate		
dimension ▼	quarter ▼	Total
G_Female15-34	A_q12008	3.6
	B_q22008	4.0
	C_q32008	4.3
G_Female35-64	A_q12008	2.1
	B_q22008	2.3
	C_q32008	2.4
G_Male15-34	A_q12008	4.3
	B_q22008	4.7
	C_q32008	4.6
G_Male35-64	A_q12008	2.5
	B_q22008	2.7
	C_q32008	3.0
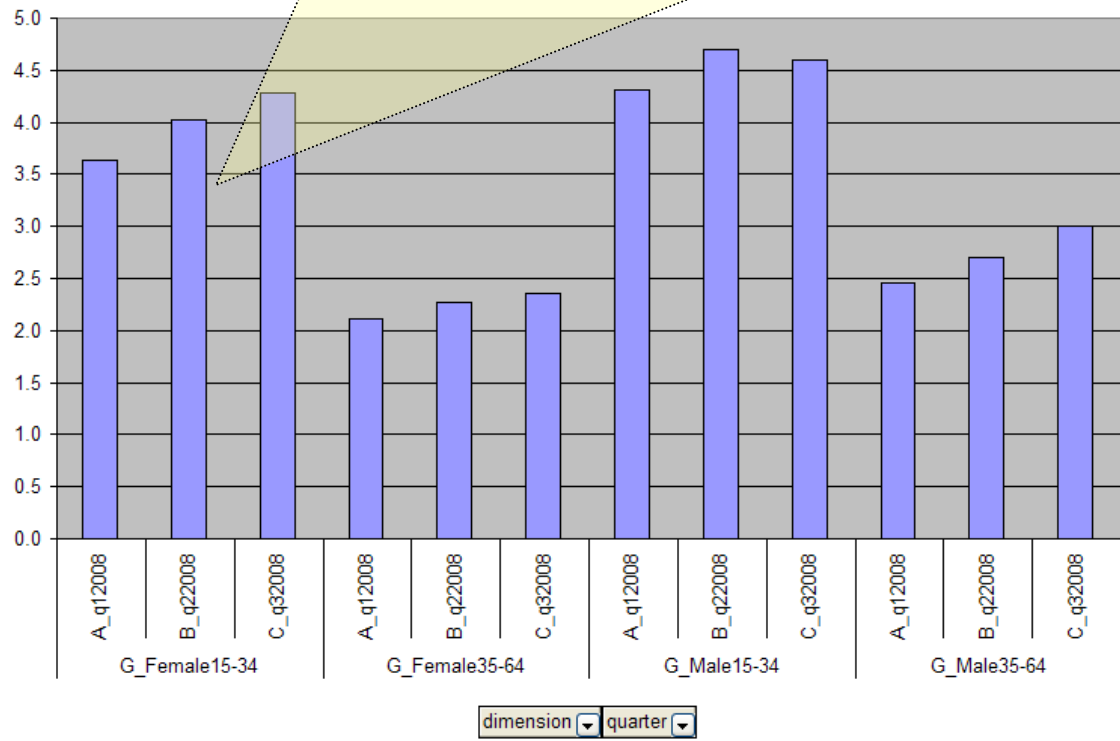
Note: Data for this cube was simulated due to confidentiality

Target audience

- Methodologists and Survey Statisticians
- Team leaders in surveys
- Quality Assurers
- Decision makers in head office and provinces

Thanyani A Maremba

Potential areas of application

- Stats SA has other systems that provide regular reporting facility and are used in survey areas for example, RTMS succeeded by ITS, and other daily web reports from surveys

- There is also electronic products section which deals with final published data and preparing it to be used by external stakeholders

- In light of what we have we look at how OLAP system can be used to analyse quality indicators for other areas in the organisation and these areas are listed below.

  - ❐ Other household surveys
  - ❐ Business surveys
  - ❐ Corporate services
  - ❐ Data processing
  - ❐ Listing
  - ❐ Survey evaluation
  - ❐ Field operations
  - ❐ Census

Thanyani A Maremba

Future improvements

- Interacting with the Geodatabase in reporting geography dimensions.

- Explore other methods of viewing multi-dimensional data.

- To optimally use the technology we have in Stats SA.

- To investigate more about doing analysis and reporting efficiently

Thanyani A Maremba

Conclusions

- We can simply turn statistical data into knowledge by simply organising it in a multi dimensional structure

- A cube can be viewed by multiple users without interfering with the data

- Third party option of using MS Excel to view cubes allows non-SAS users to do independent analysis

- The responsibility of educating users to analyse quality indicators lies in the hand of the methodologists

Thanyani A Maremba

Acknowledgements

- The organisers of the conference

- The methodology team

- ADAPT team

Thanyani A Maremba

Thank you

Thanyani A Maremba