Australian Bureau of Statistics

Automatically protecting user-defined tables and analytic outputs from the Australian Population Census

ESSnet workshop on Statistical Disclosure Control of Census data

Luxembourg April 2012

Melissa Gare

Introduction

- ABS legislative framework
- Overview of disclosure risk associated with release of Census data in tables
- Exploring ABS Census TableBuilder perturbation method and additivity algorithm
- Future research directions – Survey TableBuilder and Analysis Server

Australian Census of Population and Housing

- Conducted every 5 years
- Last held in 2011
- First release of data June 2012

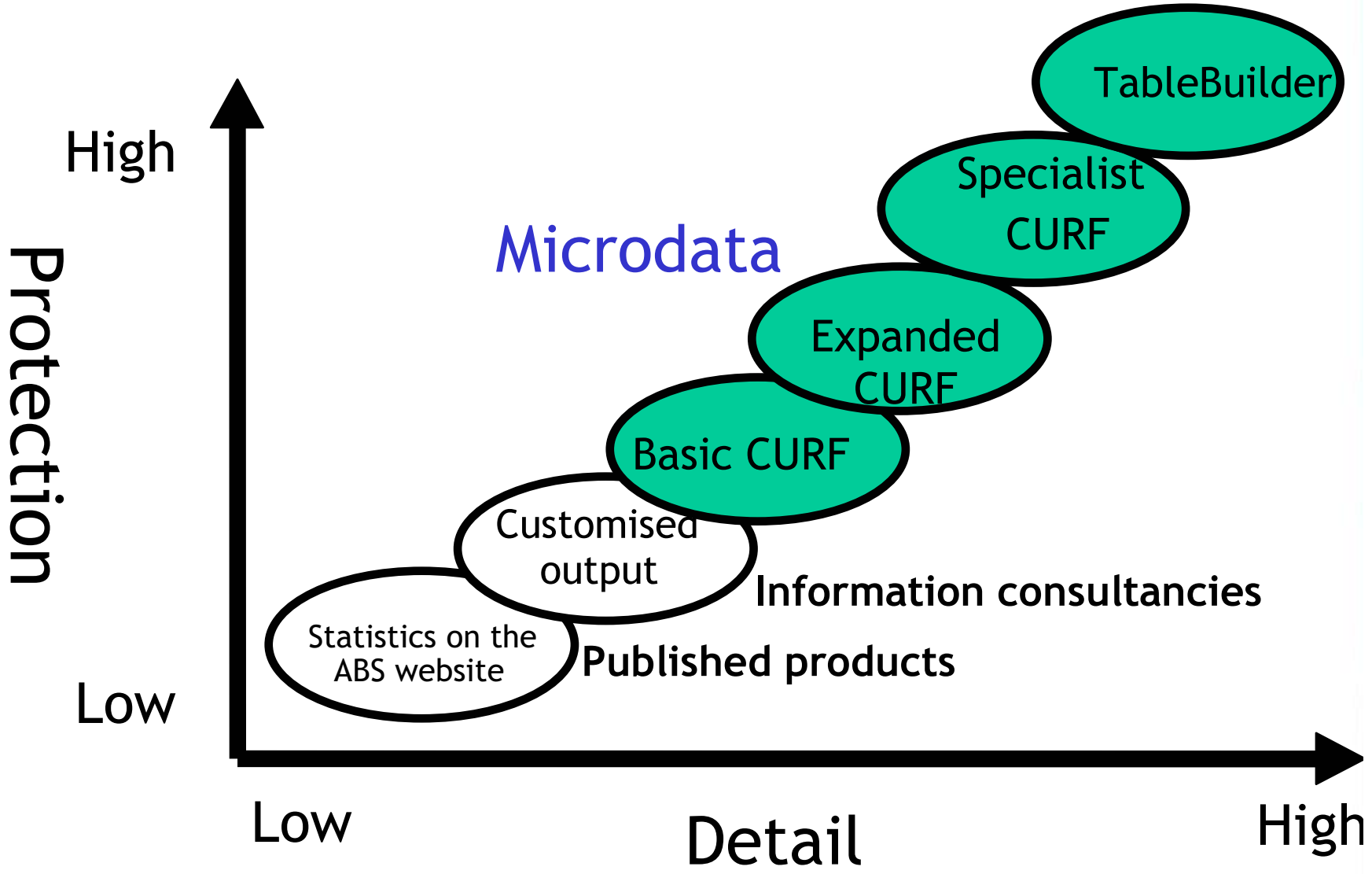
Releasing statistics at the ABS

- The ABS is committed to maximising use of ABS data for informed decision making
- Data at the ABS are collected under the Census and Statistics Act. This legislation requires that any data collected under this authority shall not be released in a manner that is likely to enable the identification of a particular person or organisation
- ABS is supplementing traditional published tables with a web-based application that enables users to create their own queries

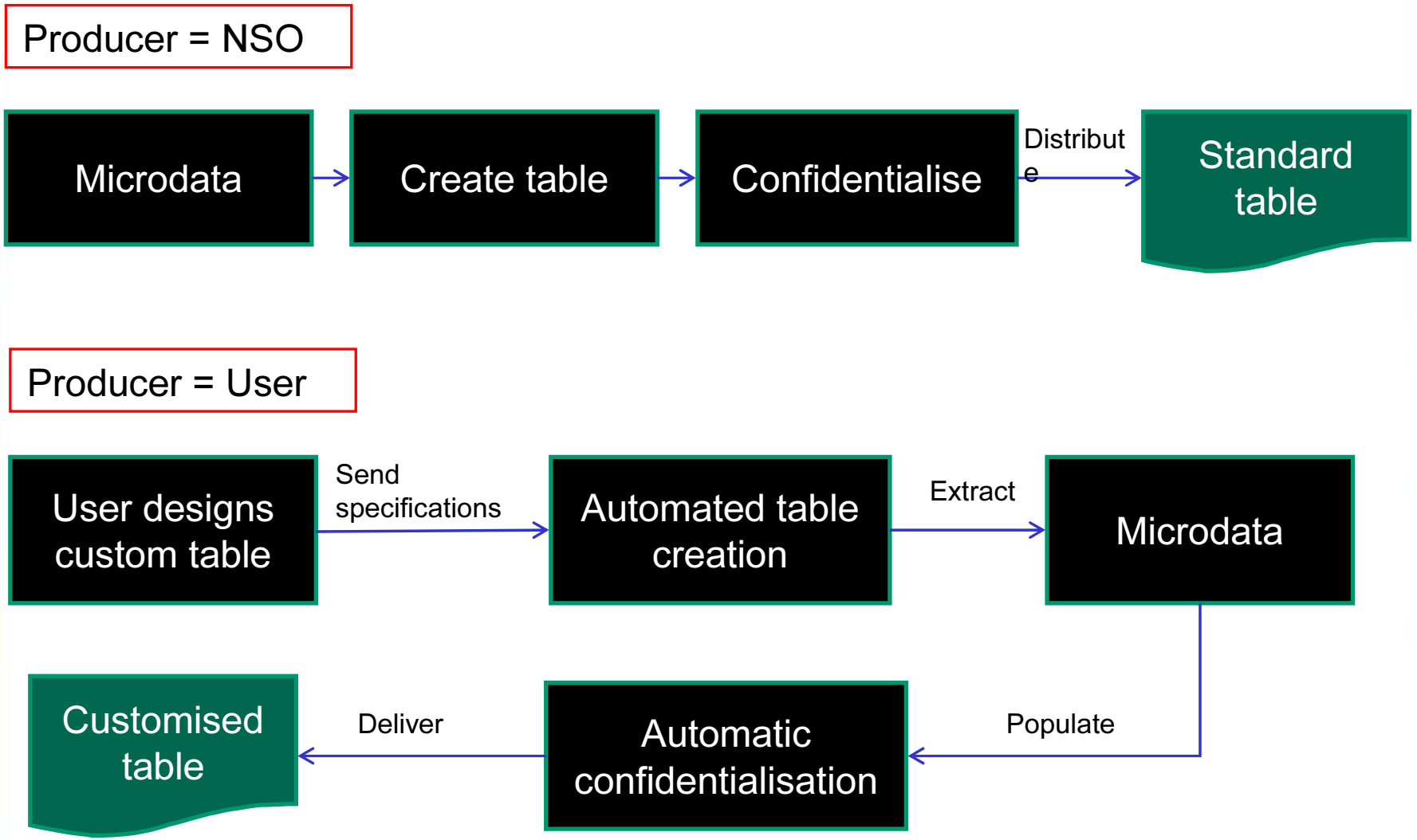
Analysis of Population Census data

- Users' Environment
  - 1% Basic CURF on CD-ROM
- Remote Execution – Remote Access Data Laboratory
  - 5% Expanded CURF. Users can submit programs written in SAS, SPSS and STATA.
- Census TableBuilder
- On-site - ABSDL
  - Access to Expanded CURF
- Special Data Service/Consultancies

Current modes of access to ABS microdata



User customised tables



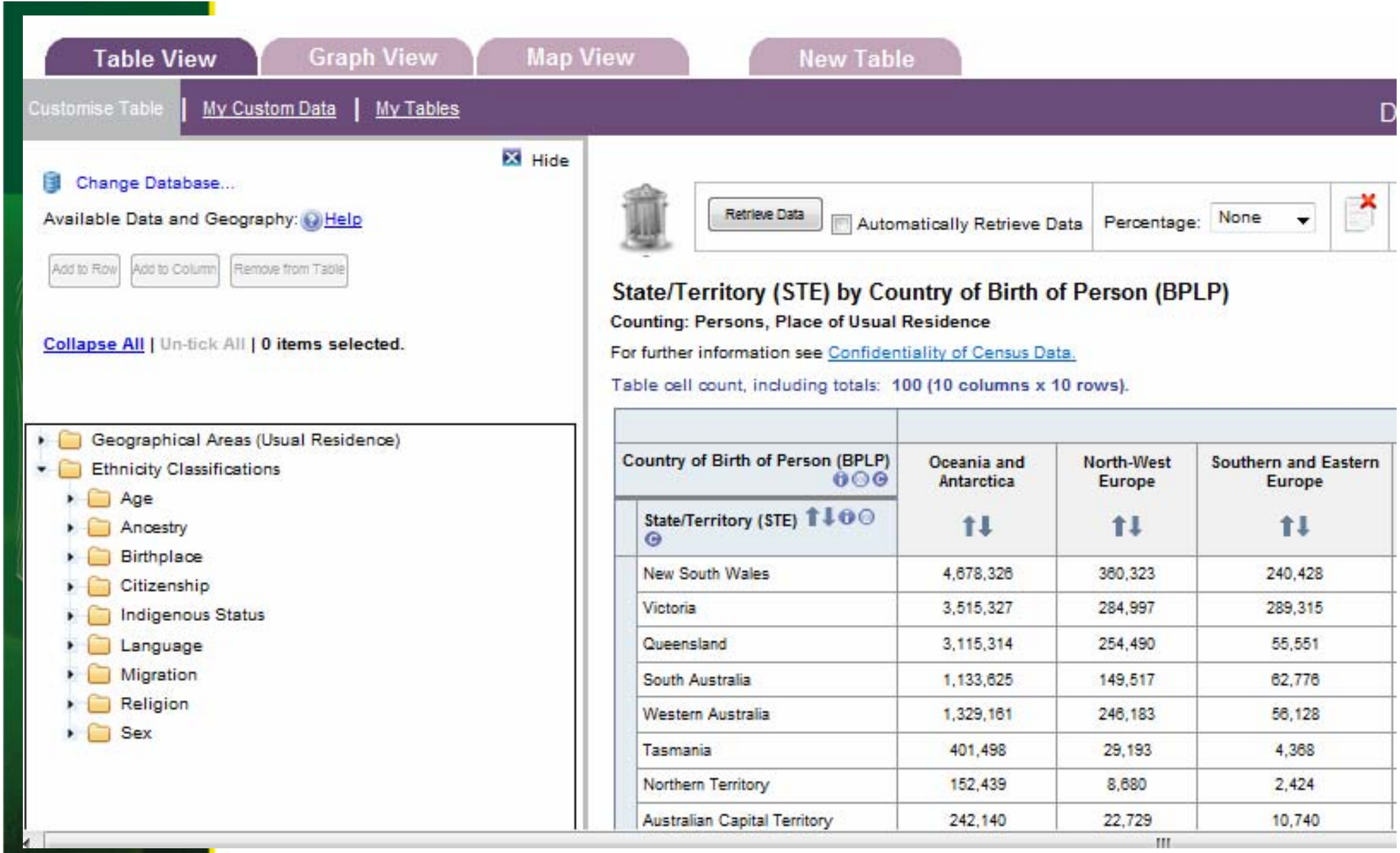


Table View
Graph View
Map View
New Table
Customise Table
My Custom Data
My Tables
Hide
Change Database...
Available Data and Geography: Help
Add to Row
Add to Column
Remove from Table
Collapse All | Un-tick All | 0 items selected.
Geographical Areas (Usual Residence)
Ethnicity Classifications
Age
Ancestry
Birthplace
Citizenship
Indigenous Status
Language
Migration
Religion
Sex
Retrieve Data
Automatically Retrieve Data
Percentage: None
State/Territory (STE) by Country of Birth of Person (BPLP)
Counting: Persons, Place of Usual Residence
For further information see Confidentiality of Census Data.
Table cell count, including totals: 100 (10 columns x 10 rows).
Country of Birth of Person (BPLP)
State/Territory (STE)
Oceania and Antarctica
North-West Europe
Southern and Eastern Europe
New South Wales 4,678,326 360,323 240,428
Victoria 3,515,327 284,997 289,315
Queensland 3,115,314 254,490 55,551
South Australia 1,133,625 149,517 62,776
Western Australia 1,329,161 246,183 56,128
Tasmania 401,498 29,193 4,368
Northern Territory 152,439 8,680 2,424
Australian Capital Territory 242,140 22,729 10,740

Census TableBuilder perturbation method

- Method protects against repeated requests and differencing
- Numeric record key assigned to each unit
- Record keys combined to form cell key
- Cell key and look-up table determine perturbation

Implementing the perturbation method

The following parameters need to be set:

1. the distribution used to generate the record keys;
2. the best way to combine the record keys to create the cell-level keys; and
3. the distribution of the perturbation values held in the look-up table

Census TableBuilder method

- Perturbation look-up table ensures:
  - integer values in output table
  - mean perturbation is zero
  - no negative values or values below a threshold in output table
  - fixed variance
  - fixed maximum for absolute value of perturbation

Distribution of perturbation values

- The ABS distribution satisfies:

a) $E_*\left(e_i^*\right)=0$

b) $Var_*\left(e_i^*\right)=\sigma^2$

c) $\mathrm{Cov}_*\left(e_i^*,e_j^*\right)=0 \quad if \ i\neq j$

d) whenever the same set of records contribute to a cell count, the value for e_i will always be the same

e) e_i^* is an integer

where $Var_*(\ )$ and $E_*(\ )$ are the variance and expectation with respect to the perturbation distribution of e_i^*.

- Entropy (a measure of uncertainty) is maximised subject to variance and bias constraints to create the look-up table.

Australian Bureau of Statistics

Creating look-up tables

- Solve resulting systems of non-linear equations numerically, to calculate a probability distribution
- Populate cells of the look-up table to approximate the distribution

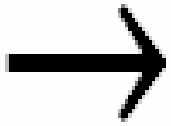
Flexibility of the method

- Marley and Leaver (2011) demonstrate the flexibility of the method through three alternative perturbation distributions:
  - a maximum perturbation of +/- 1 to the unconfidentialised cell counts;
  - a maximum perturbation of +/- 5 to the unconfidentialised cell counts and no confidentialised cell values between 1 and 4 inclusive; and
  - a maximum perturbation of +/- 20, with the constraint that no cell values are unchanged, and that all confidentialised cell values are multiples of 5.

Look-up table 1

- Maximum perturbation of □1

1	26	2
52	3	14
8	0	5

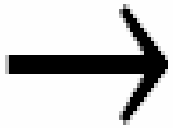


0	26	1
53	4	13
9	0	5

Look-up table 2

- Maximum perturbation of □5
- No perturbed cell values between 1 and 4

1	26	2
52	3	14
8	0	5

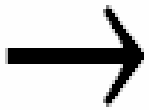


5	23	0
53	0	10
8	0	10

Look-up table 3

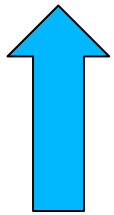
- Maximum perturbation of □20
- All non-zero cells changed
- All perturbed cell values are multiples of 5

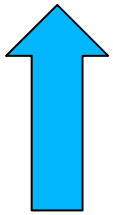
1	26	2
52	3	14
8	0	5

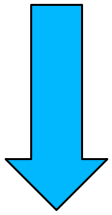


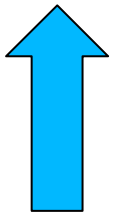
0	45	10
40	10	25
0	0	20

Utility loss versus Disclosure risk

Variance of the perturbation distribution 

 Size of perturbation

Disclosure risk 

 Utility loss increases

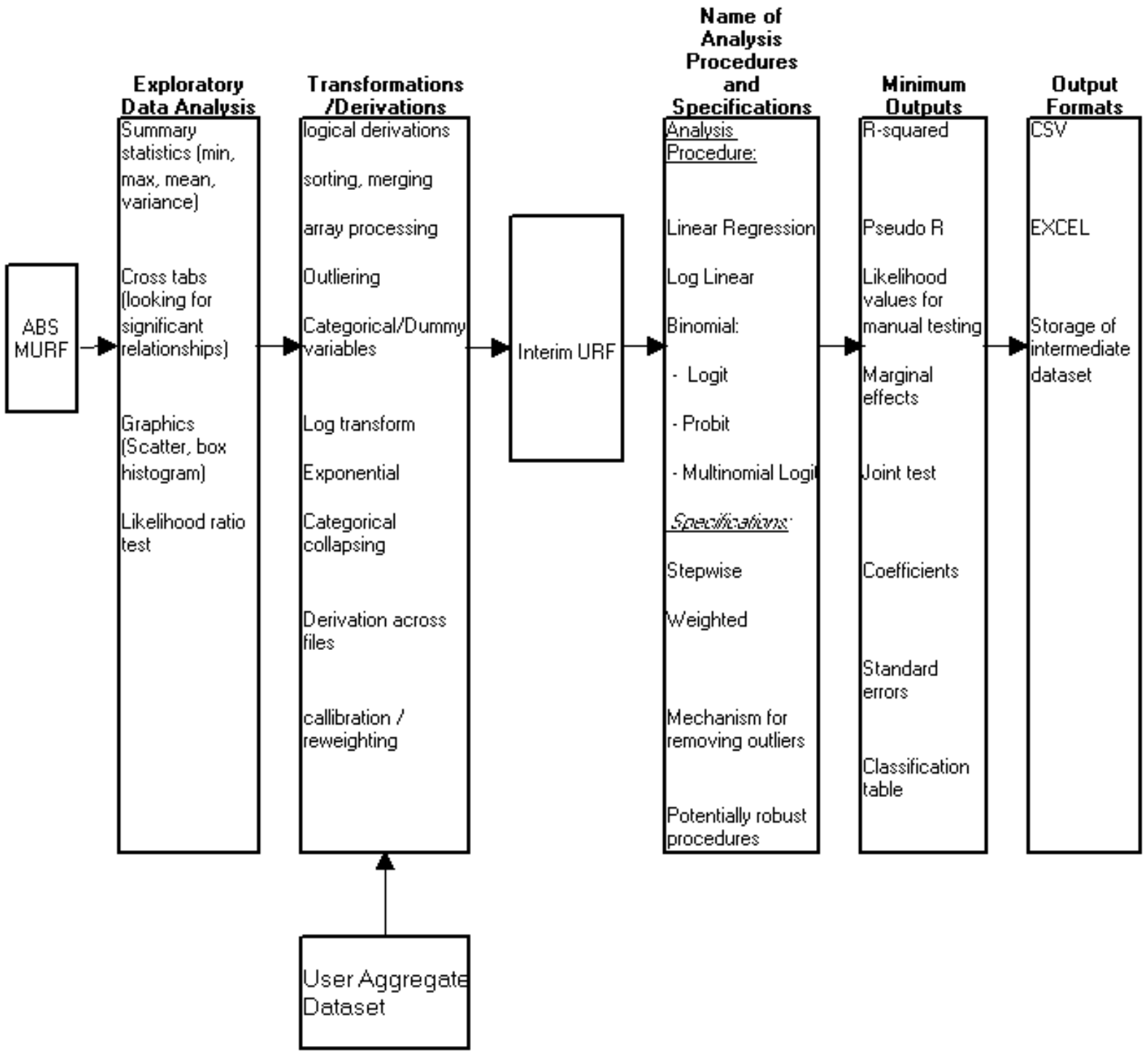
Undesired property

- Following perturbation tables may not be additive
  - Interior and marginal cells are all perturbed independently
- Additivity module
  - Census TableBuilder incorporates a second routine that restores additivity to tables post perturbation
  - Cells appearing in two different tables will be perturbed the same, however may be altered different ways by additivity
  - Uses iterative methods to restore additivity

Future directions

- Extension of perturbation method to weighted counts from surveys
- Extension of perturbation method to continuous outputs such as total, mean, median and quantile bounds
- Creation of an Analysis Service

User Desired Analysis Functionality

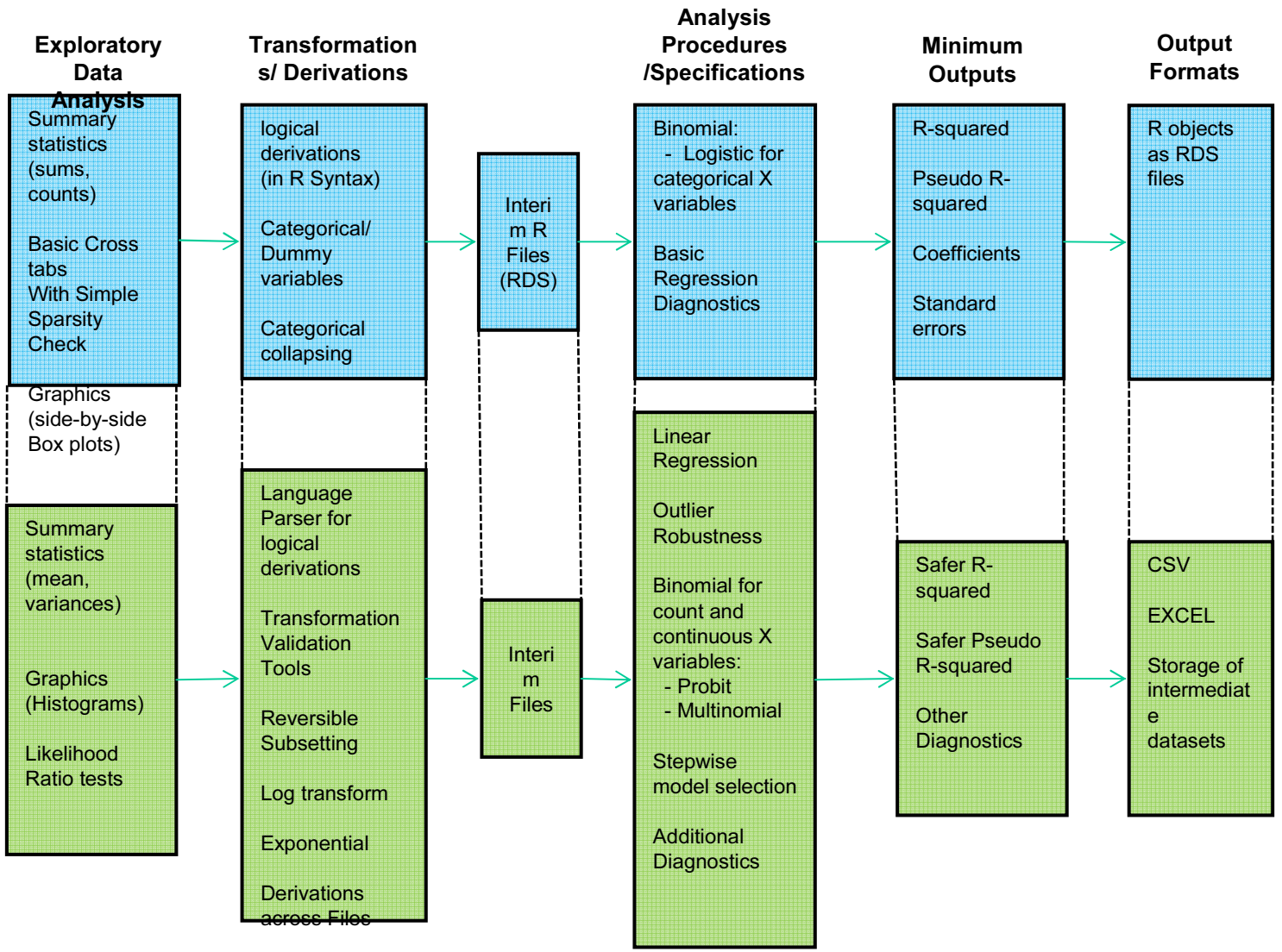


Australian Bureau of Statistics

By June 2011 Demonstrator

By June 2012 Production Version

- Written in R or other software?
- Full User Authentication
- Audit System
- Workflow Control,
- Data Repository Interface
- Metadata Interface
- Survey Table Builder Interface



Disclosure Control for Modelled Outputs

- Based on methodology by Chipperfield et.al.
- Two levels of control are necessary:
  - Parameters and Inferences
  - Diagnostics e.g. residual plots, influence diagnostics
- For parameters:
  - Perturb the score function for estimating the model parameters
  - Calculate the variance post perturbation for inference purpose
- For diagnostics, residual plots:
  - Perturb diagnostics by an amount which would depend on maximum influence a record can have on the diagnostic measure
  - Show box diagrams rather than plots.
- Introduce general restrictions and attack specific restrictions

Australian Bureau of Statistics

Perturbing Score Function

- Instead of solving $\boldsymbol{H(\beta) = 0}$ and releasing $\widehat{\boldsymbol{\beta}}$, the server solves $\boldsymbol{H(\beta) = E^*}$

  and releases the solution $\widehat{\boldsymbol{\beta}}^*$

  - Where $\boldsymbol{E^*}$ are perturbations introduced
  - Perturbations are bound by the maximum influence a record will have on the estimating equation.
- Variance can be calculated:

$$V_{m^*}(\widehat{\boldsymbol{\beta}^*}) = V_{m^*}(\widehat{\boldsymbol{\beta}}) + V_{e^*}(\widehat{\boldsymbol{\beta}^*})$$

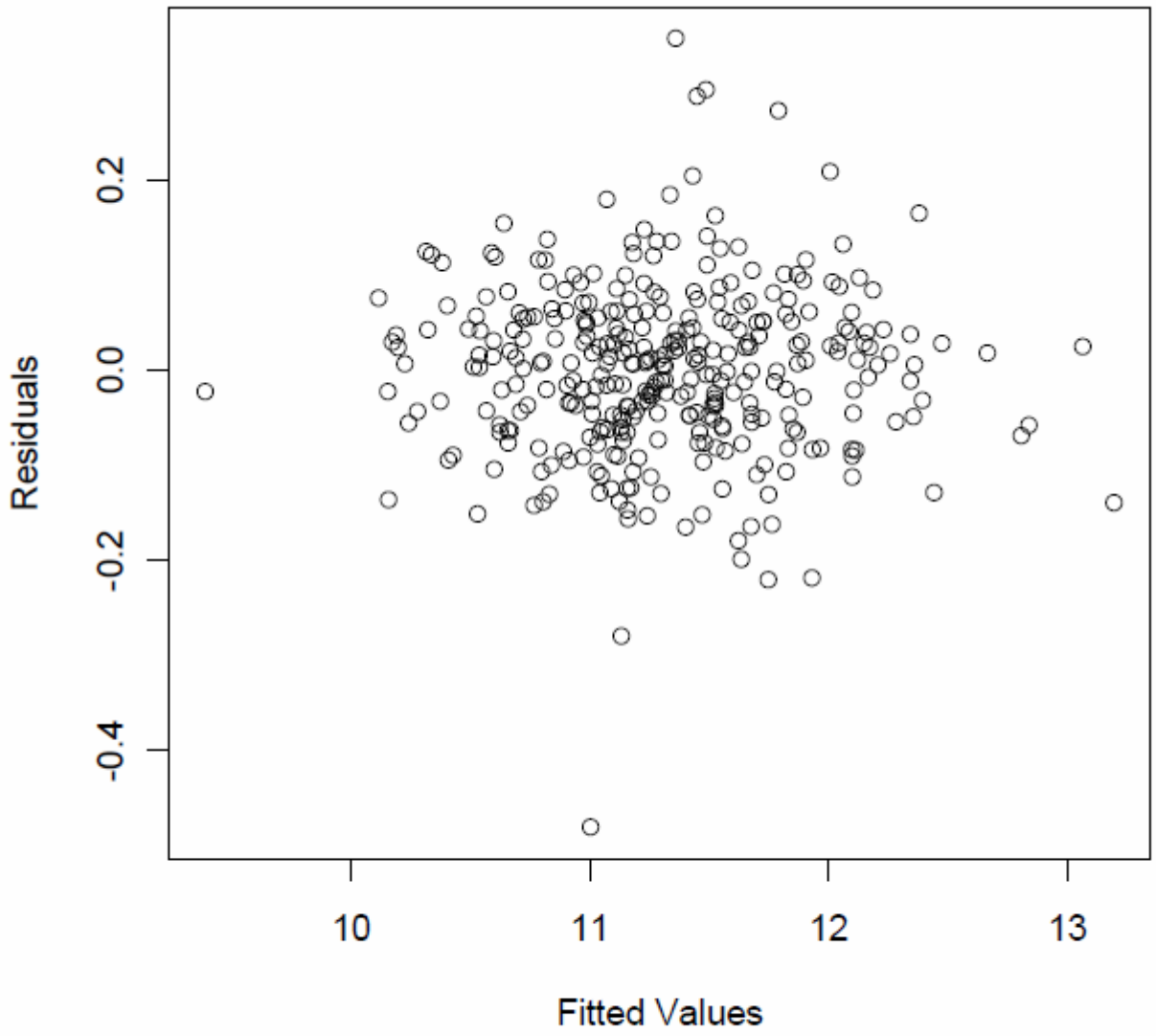


Figure: Standard Residual Plot

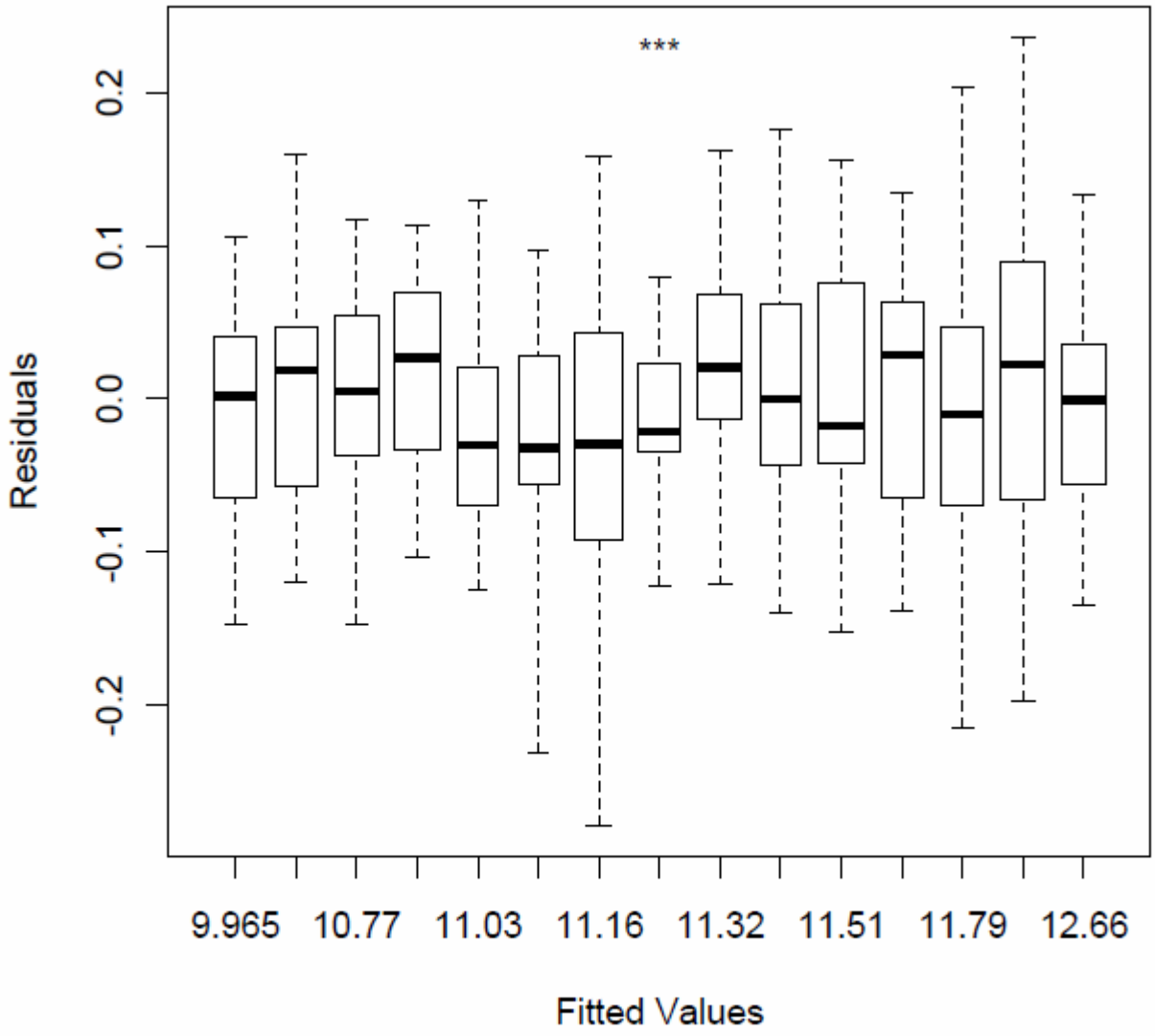


Figure: Confidential Residual Plot

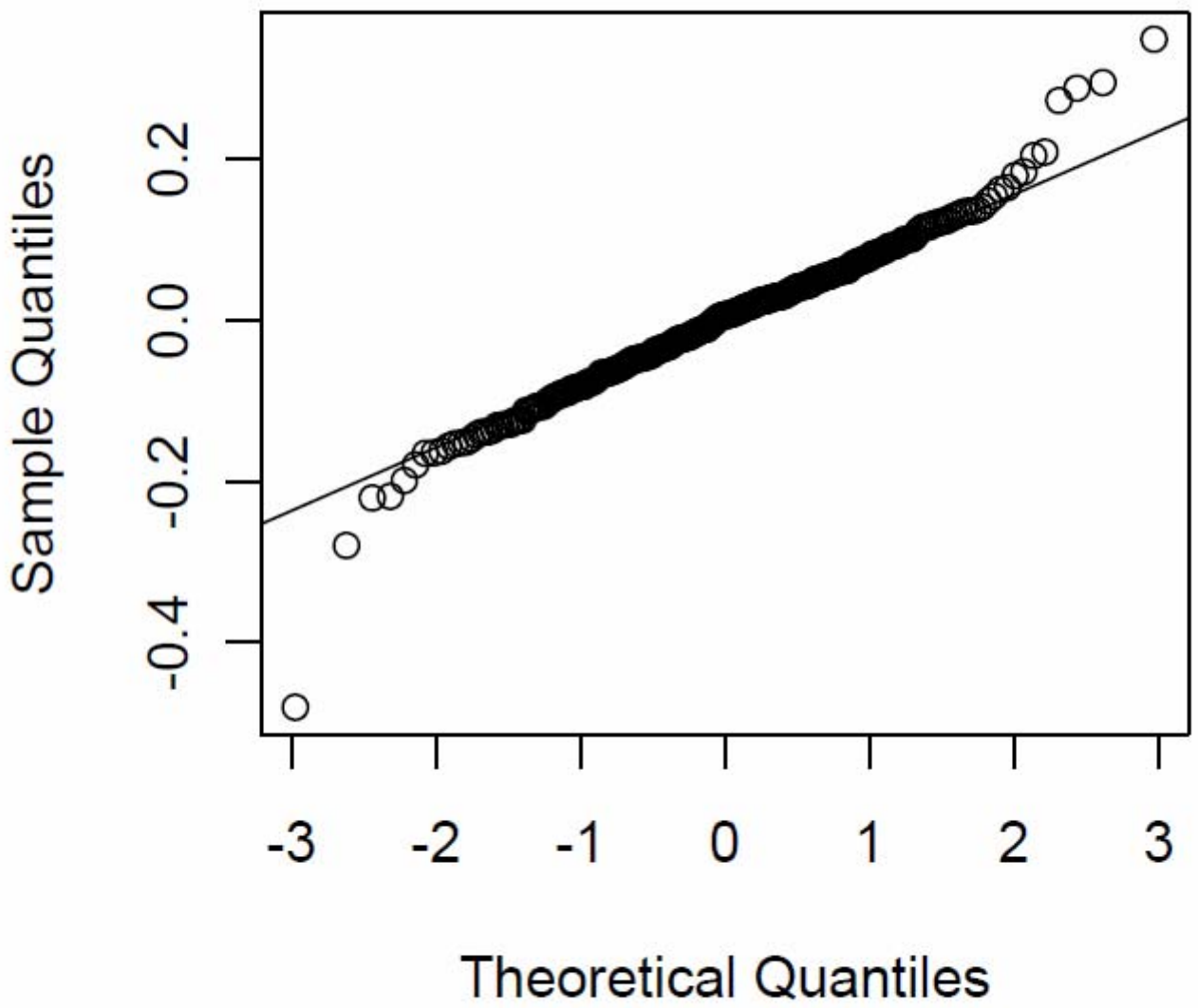


Figure: Standard Q-QPlot

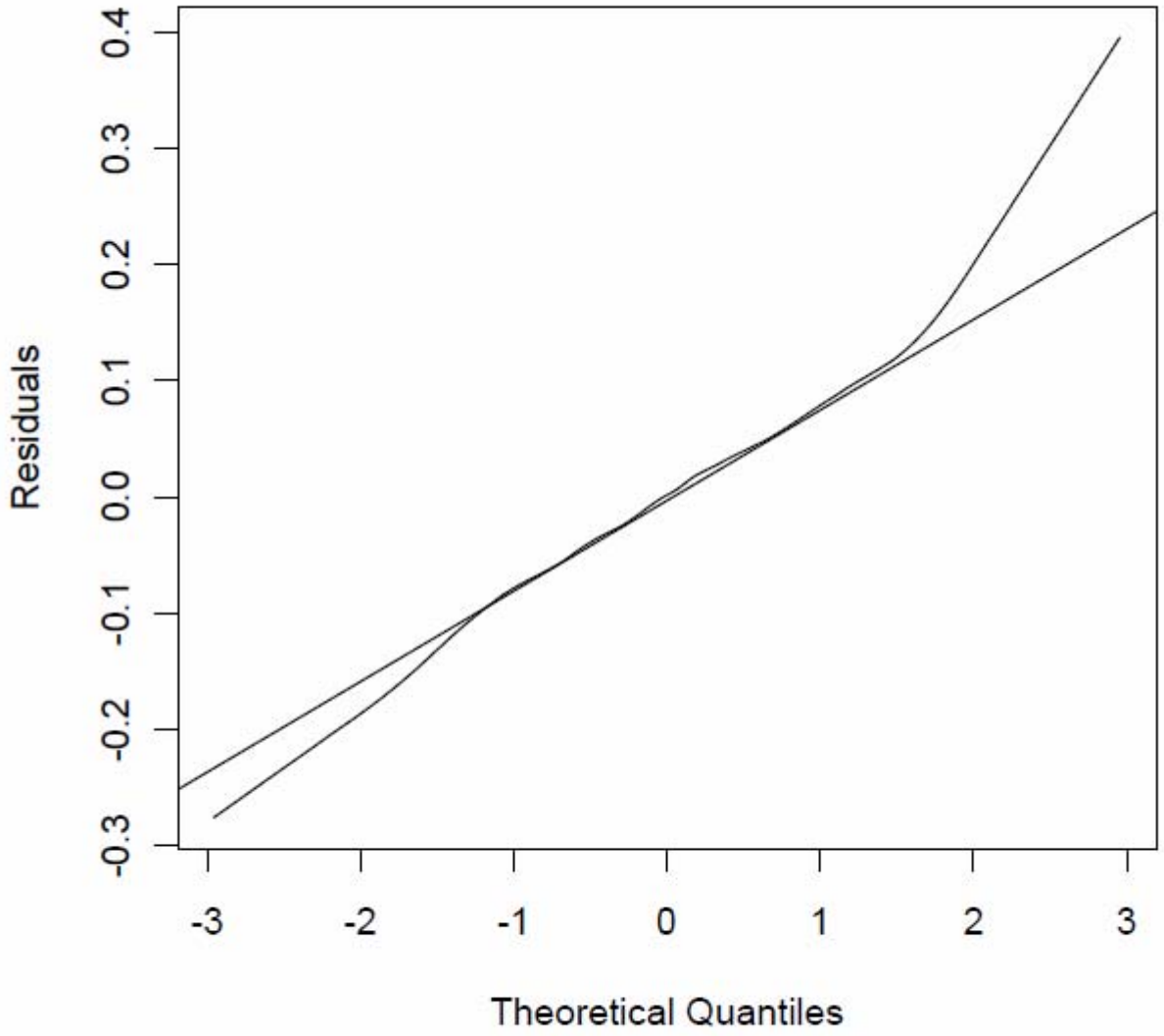


Figure: Confidential Q-Q Plot

Acknowledgements

- ABS Co-authors of the paper and presentation
  - Victoria Leaver
  - Dr James Chipperfield