MSc in Official Statistics
Statistical Computing:
The RDBMS approach to Databases

Andrew Westlake

Survey & Statistical Computing

63 Ridge Road, London N8 9NP, UK

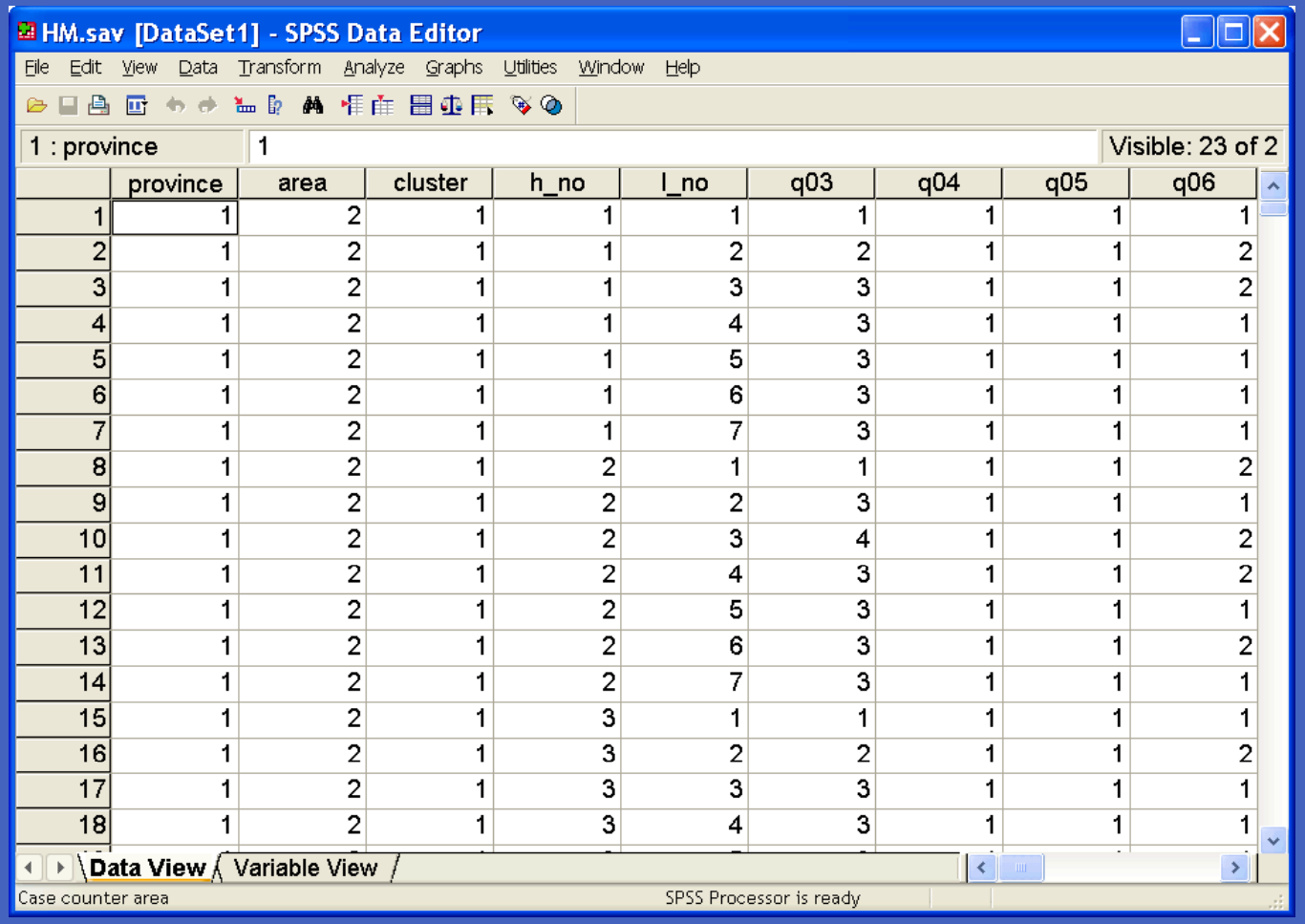
+44 (0) 20 8374 4723

AJW@SaSC.co.uk (E-Mail)

www.SaSC.co.uk

Statistical View of Data

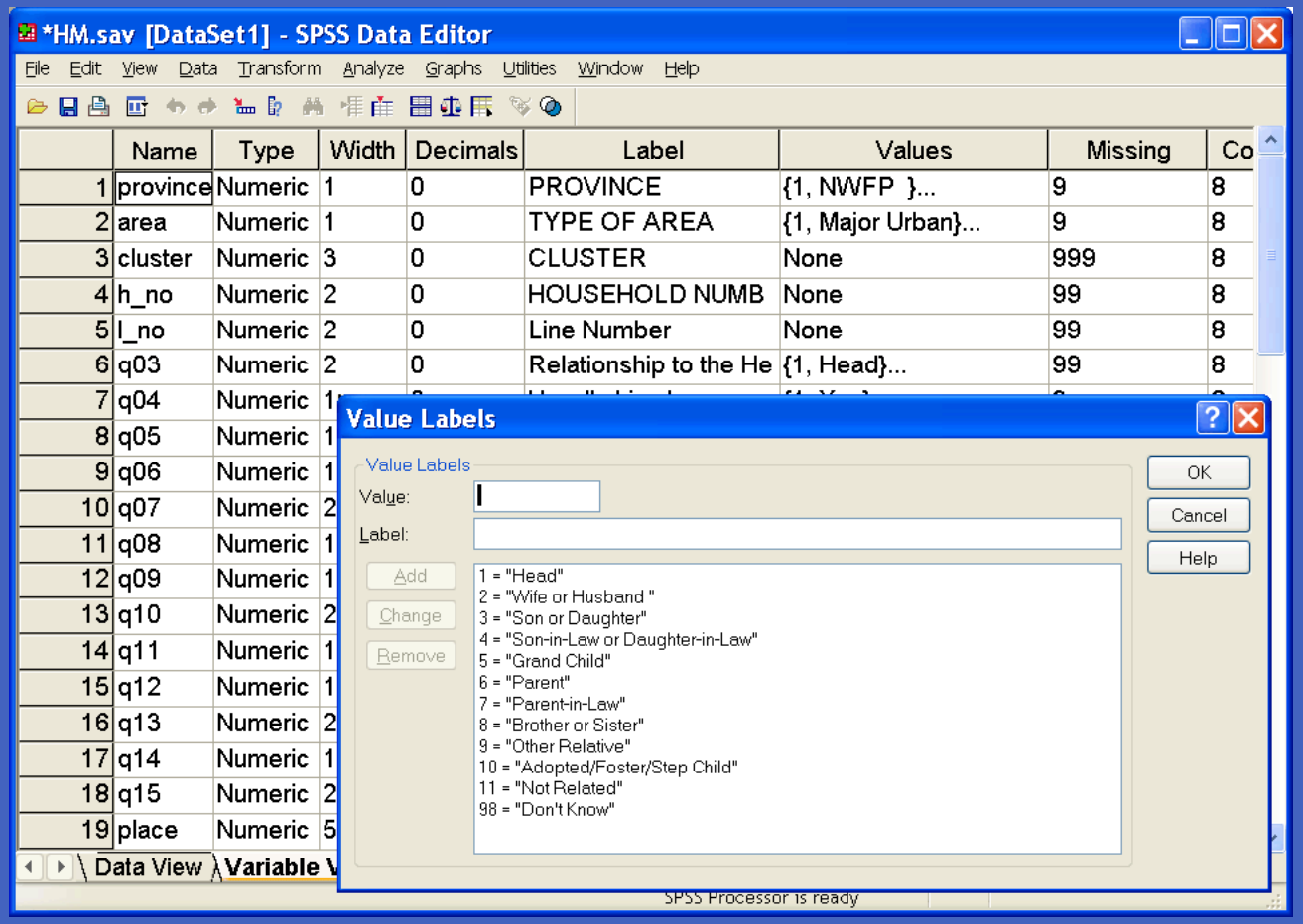
- 'Data matrix', as in SPSS
  - » Columns are Variables
  - » Rows are data values for a respondent (Cases)



	province	area	cluster	h_no	l_no	q03	q04	q05	q06
1	1	2	1	1	1	1	1	1	1
2	1	2	1	1	2	2	1	1	2
3	1	2	1	1	3	3	1	1	2
4	1	2	1	1	4	3	1	1	1
5	1	2	1	1	5	3	1	1	1
6	1	2	1	1	6	3	1	1	1
7	1	2	1	1	7	3	1	1	1
8	1	2	1	2	1	1	1	1	2
9	1	2	1	2	2	3	1	1	1
10	1	2	1	2	3	4	1	1	2
11	1	2	1	2	4	3	1	1	2
12	1	2	1	2	5	3	1	1	1
13	1	2	1	2	6	3	1	1	2
14	1	2	1	2	7	3	1	1	1
15	1	2	1	3	1	1	1	1	1
16	1	2	1	3	2	2	1	1	2
17	1	2	1	3	3	3	1	1	1
18	1	2	1	3	4	3	1	1	1

Statistical Variables

- Data Dictionary for details
  - » Variable and Value labels, etc., in SPSS
  - » Statistical example of Metadata



RDBMS Dataset

- Very similar to Data Matrix
  - » Same structure, different names
  - » Dataset is a Table
  - » Columns are Attributes (or Fields)
  - » Rows are Records

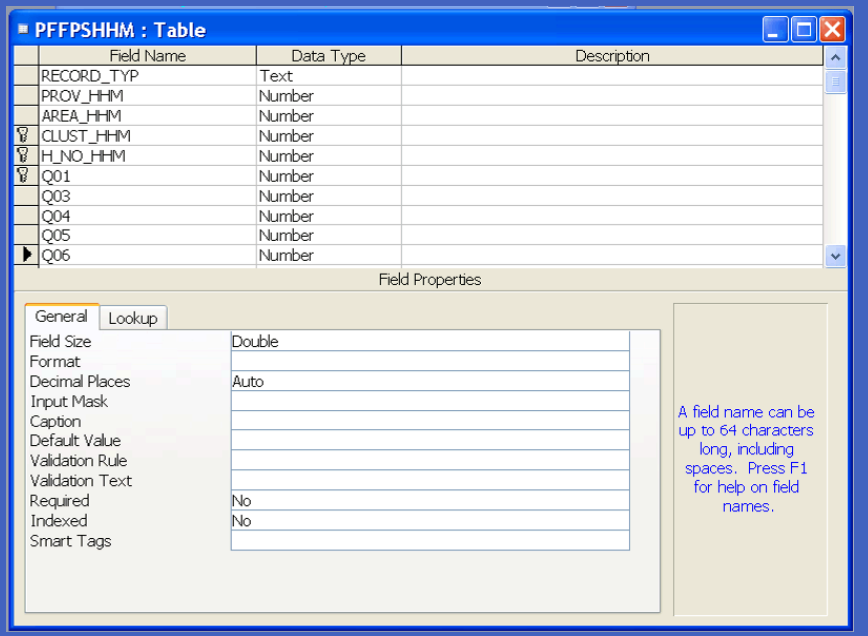
PFFPSHHM : Table

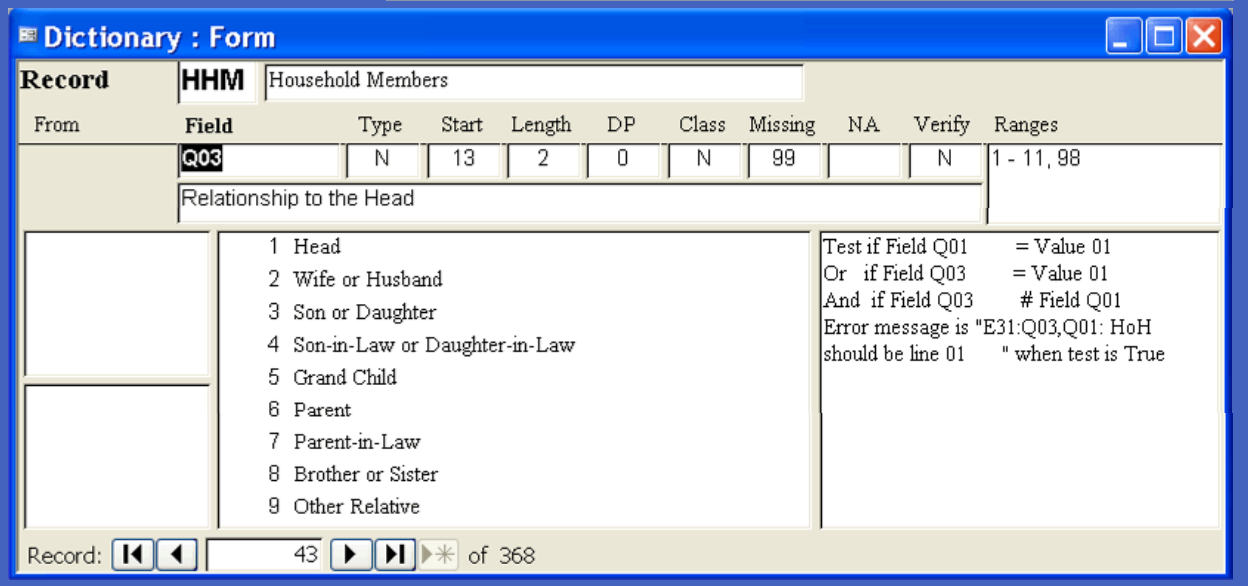
	RECORD	PROV_	AREA_	CLUST	H_NO	Q01	Q03	Q04	Q05	Q06
+	HHM	1	3	3	23	1	1	1	1	1
+	HHM	1	3	3	23	2	2	1	1	2
+	HHM	1	3	3	23	3	3	1	1	2
+	HHM	1	3	5	21	1	1	1	1	1
+	HHM	1	3	5	21	2	2	1	1	2
+	HHM	1	3	5	21	3	3	1	1	1
+	HHM	1	3	5	22	1	1	1	1	1
+	HHM	1	3	5	22	2	2	1	1	2
+	HHM	1	3	5	22	3	3	1	1	1
+	HHM	1	3	5	22	4	3	1	1	1
+	HHM	1	3	5	22	5	3	1	1	1
+	HHM	1	3	5	22	6	3	1	1	1
+	HHM	1	3	5	22	7	3	1	1	2
+	HHM	1	3	5	22	8	3	1	1	2
+	HHM	1	3	5	22	9	3	1	1	2
+	HHM	1	3	5	24	1	1	1	1	1
+	HHM	1	3	5	24	2	2	1	1	2
+	HHM	1	3	5	24	3	3	1	1	1
+	HHM	1	3	5	24	4	3	1	1	2

Record: 1 of 3013

RDBMS Variables

- Variable definition through a table
- Value labels can be supported through forms





Contrasts – Statistical Packages

- Strengths
  - » Statistical Methods
  - » Integration of Statistical meta-data
  - » Missing Value treatment
  - » Familiar terminology
- Weaknesses
  - » Handling multiple datasets, e.g. hierarchies
  - » Data editing – cases and variables
  - » Security and Auditing

Contrasts - RDBMS

- Strengths
  - » Standardised data manipulation and linking
    - Integrated Null concept
  - » Standardised access from other systems, including Statistical Packages
  - » Integrated Programming, Security, Integrity, …
  - » Rich Data Design Methodologies
- Weaknesses
  - » No statistical methods, limited aggregation functionality
  - » No statistical meta-data, missing value
    - Some functionality can be programmed, but not standard
  - » No treatment of macro data
    - Various extensions, but not standard

Roles of RDBMS and Statistical Packages

- Don't choose between, use strengths of both
  - » RDBMS for data storage and management
  - » Statistical package for analysis and presentation
- Rich RDBMS concepts
  - » Useful way to think about all data
  - » Particularly valuable for complex situations
- Example of PFFPS
  - » Pakistan Fertility and Family Planning Survey
  - » Also used in Exercises