Describing Aggregate Data: The Enigma Variations

Presented at IASSIST/IFDO 2001

Wendy L.Thomas

18. May 2001

What are Aggregate Data?

Aggregate data are the result of manipulating microdata by totaling the number of cases meeting specific criteria; by summing microdata variables for specific subpopulations; by listing cases that meet specific criteria.....

Aggregate data are a result set derived through manipulation which have a specific relationship to other result sets derived during the same process.

Why are they so difficult to describe?

—Difficult to provide an abstract definition of what constitutes aggregate data

—Secondary data set —it is the result of manipulating primary (micro) data

—Frequently stored in spreadsheets (grids)

—It is an n-dimensional structure displayed in a 1- or 2-dimensional format

—Discrete cells are used as sources of ‘look-up« information

What more description do they need?

- Logical relationship between cells
- Nature of the relationship
- Additivity
- Location within a physical storage grid
- Description of how they were created
- Directions for deriving from microdata

	Minneapolis	St. Paul	Rochester
Year Built:	xxxx	xxxx	xxxx
Before 1945	xxxx	xxxx	xxxx
1945 to 1964	xxxx	xxxx	xxxx
1965 to 1984	xxxx	xxxx	xxxx
1985 to 1994	xxxx	xxxx	xxxx
1995 to 1999	xxxx	xxxx	xxxx
2000	xxxx	xxxx	xxxx
2001	xxxx	xxxx	xxxx
Number of Units:	xxxx	xxxx	xxxx
1 unit	xxxx	xxxx	xxxx
2 units	xxxx	xxxx	xxxx
3 to 4 units	xxxx	xxxx	xxxx
5 to 14 units	xxxx	xxxx	xxxx
15 or more units	xxxx	xxxx	xxxx

	Minneapolis	St. Paul	Rochester
Year Built:	xxxx	xxxx	xxxx
Before 1945	xxxx	xxxx	xxxx
1945 to 1964	xxxx	xxxx	xxxx
1965 to 1984	xxxx	xxxx	xxxx
1985 to 1994	xxxx	xxxx	xxxx
1995 to 1999	xxxx	xxxx	xxxx
2000	xxxx	xxxx	xxxx
2001	xxxx	xxxx	xxxx
Number of Units:	xxxx	xxxx	xxxx
1 unit	xxxx	xxxx	xxxx
2 units	xxxx	xxxx	xxxx
3 to 4 units	xxxx	xxxx	xxxx
5 to 14 units	xxxx	xxxx	xxxx
15 or more units	xxxx	xxxx	xxxx

AGE by SEX		
	Male	Female
Under 5 years	577	598
5 - 17 years	3673	3899
18 - 64 years	73570	73441
65 years and over	1857	2105

AGE by SEX		
	Male	Female
Under 5 years	577	598
5 - 17 years	3673	3899
18 - 64 years	73570	73441
65 years and over	1857	2105

AGE by SEX		
	Male	Female
Under 5 years	577	598
5 - 17 years	3673	3899
18 - 64 years	73570	73441
65 years and over	1857	2105

AGE by SEX		
	Male	Female
Under 5 years	577	598
5 - 17 years	3673	3899
18 - 64 years	73570	73441
65 years and over	1857	2105

Fishing Vessels				
	Motorized			Sail
	Inboard		Outboard	
	wooden deck	metal deck		
1925	658		56	93
1926	674		54	93
1927	645		62	94
1928	731		78	92
1929	524	226	75	87
1930	542	268	72	86
1931	501	273	76	83
1932	498	278	75	83
1933	443	295	77	79
1934	424	304	74	77
1935	453	354	75	75

Two Approaches

Variable Matrix

- 2 new sections to describe Matrix and dimensions
- Uses var to describe cells and cell coordinates
- Didn t address 2-dimensional storage structures

Terms: matrix, dimensions, coordinates

Cube Description

- New section to describe nested cubes
- Uses var to describe variable (age, race, etc)
- Created new descriptions for 2-dimensional storage structures

Terms: cube, columns, rows, grids

Criteria for an acceptable model

- Describe the logical structure including: full structure, each dimension, each cell, the relationship between all parts, and how they are created
- Describe the physical structure including: multipage grids, irregular grids, how to link them together, and how to access them
- Provide support for the following functions: looking up a specific cell of information, rearranging, collapsing or subsetting a logical structure

Basic Approach

- Separate logical description from physical description
- Incorporate the idea of inheritance
- Look for similarities and describe the dissimilarities
- Focus on using the codebook to describe the logical data set which may have zero or more physical storage structures

Voorburg Compromise

—Logical structure is in section 4.0

—Physical structure is in section 3.0

—Allows for multiple physical instances of a data set OR no physical instance of the data set

—Describes 2-dimensional storage structures

—Describes multidimensional result sets and retains interrelationships within the logical structure

—Retains backward validity with version 1.0

Terms

(everything hinges on our ability to communicate)

- nCube (formerly termed matrix, cube, table)
  - Has 1 to n variables and every cell in the nCube intersects each variable
  - 2 or more nCubes with common variables can be hinged
- Variable (formerly termed variable, matrix dimension, dimension, vector, data item, cell....)
- Grid
  - spreadsheet or any 2-dimensional storage structure
- Basic Layer Sheet
  - Physical layout of data cells within a grid
- Coordinate
  - the logical position of the cell within the nCube

Lessons

- Use language carefully
- Determine what you must be able to do and what you would like to do then build to these criteria
- Build to concepts (example: 1 dimensional and 2 dimensional storage structures)
- Find out how much of what you want to do that you can do with the current model