MSc in Official Statistics
Statistical Computing:
BCS Design

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

British Crime Survey

- Clustered, stratified sample of households
- Single respondent within household, reporting on personal and HH experience
- Complex questionnaire structure
- Various sample boost procedures

- Fieldwork and cleaning done by BMRB

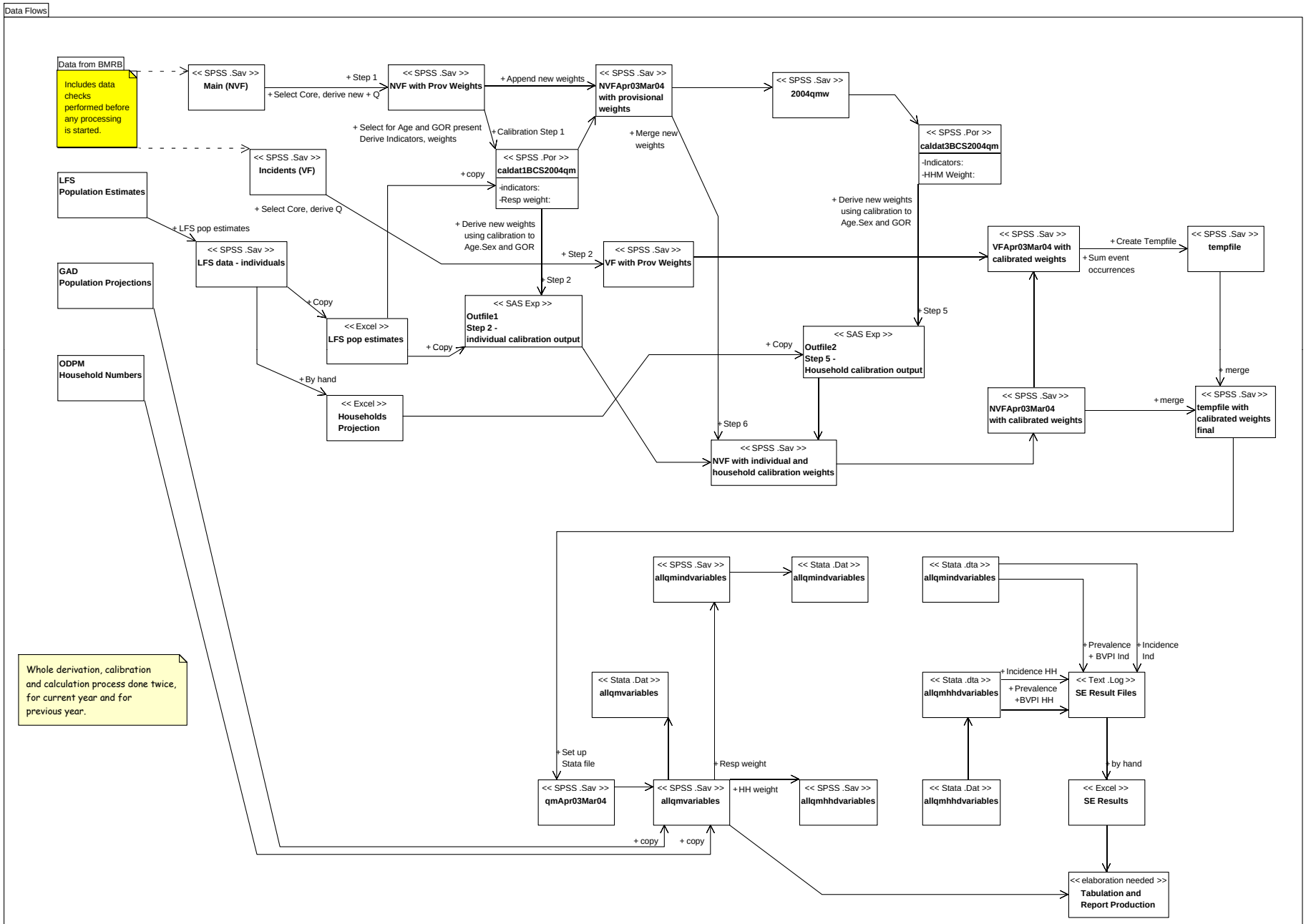
BCS Processing

- Data supplied to BCS group quarterly
  - » 2 SPSS Sav files, containing 12 months data
  - » One file for respondents (NVF), one for each reported incident (VF)
- Population Estimates and Projections obtained from ONS
- Annual publication (long), quarterly update reports (short, unchanging)
  - » Same basic processing, but more analysis for annual report
  - » Always compare current and previous year

Processing Tasks

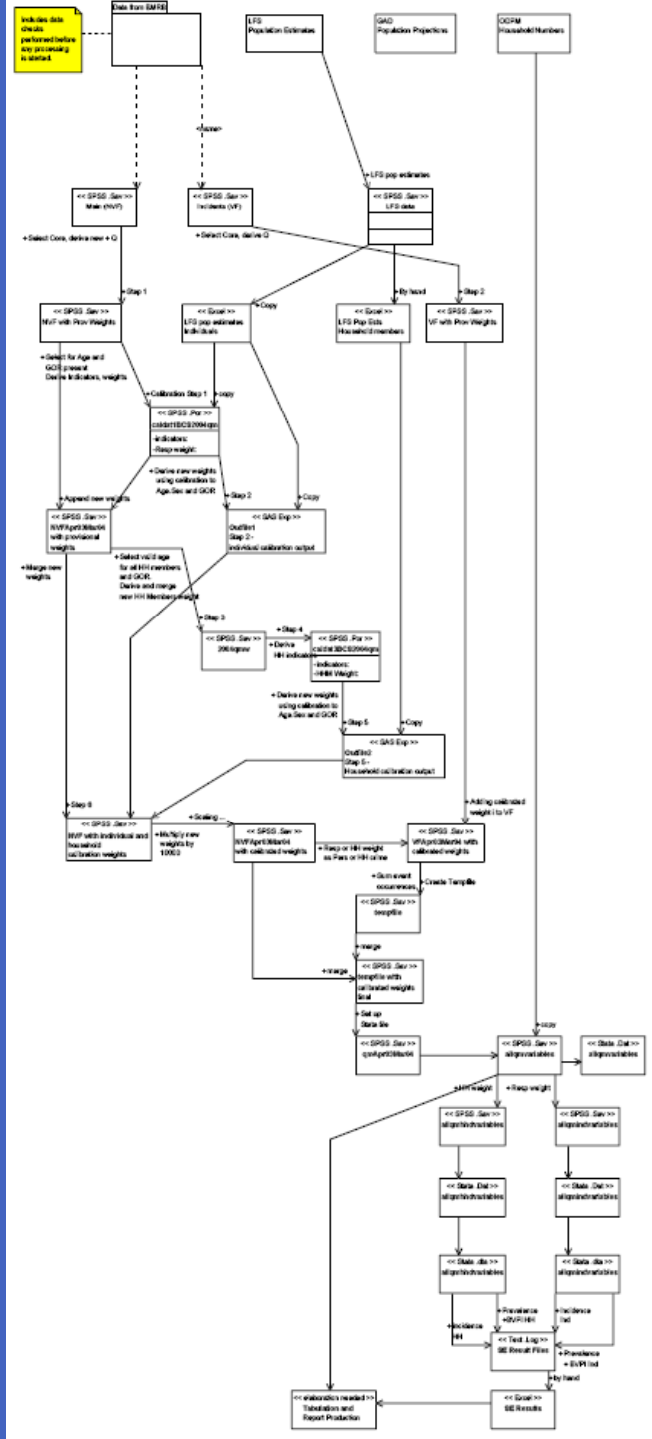
- Most processing done in SPSS
  - » Basic checks on data
    - Overall consistency of distributions with previous
  - » Derive variables used for reporting
  - » Tabulations for reporting
- Calibration weighting done in SAS
  - » Uses Calmar macro (from ONS)
  - » Used to update sampling weights
  - » Using population estimates (age x sex + GOR)
- Sampling error calculations in Stata
  - » Uses rates macro from ONS
- Publication results produced in Excel
- Most staff familiar with SPSS

BCS Data Handling



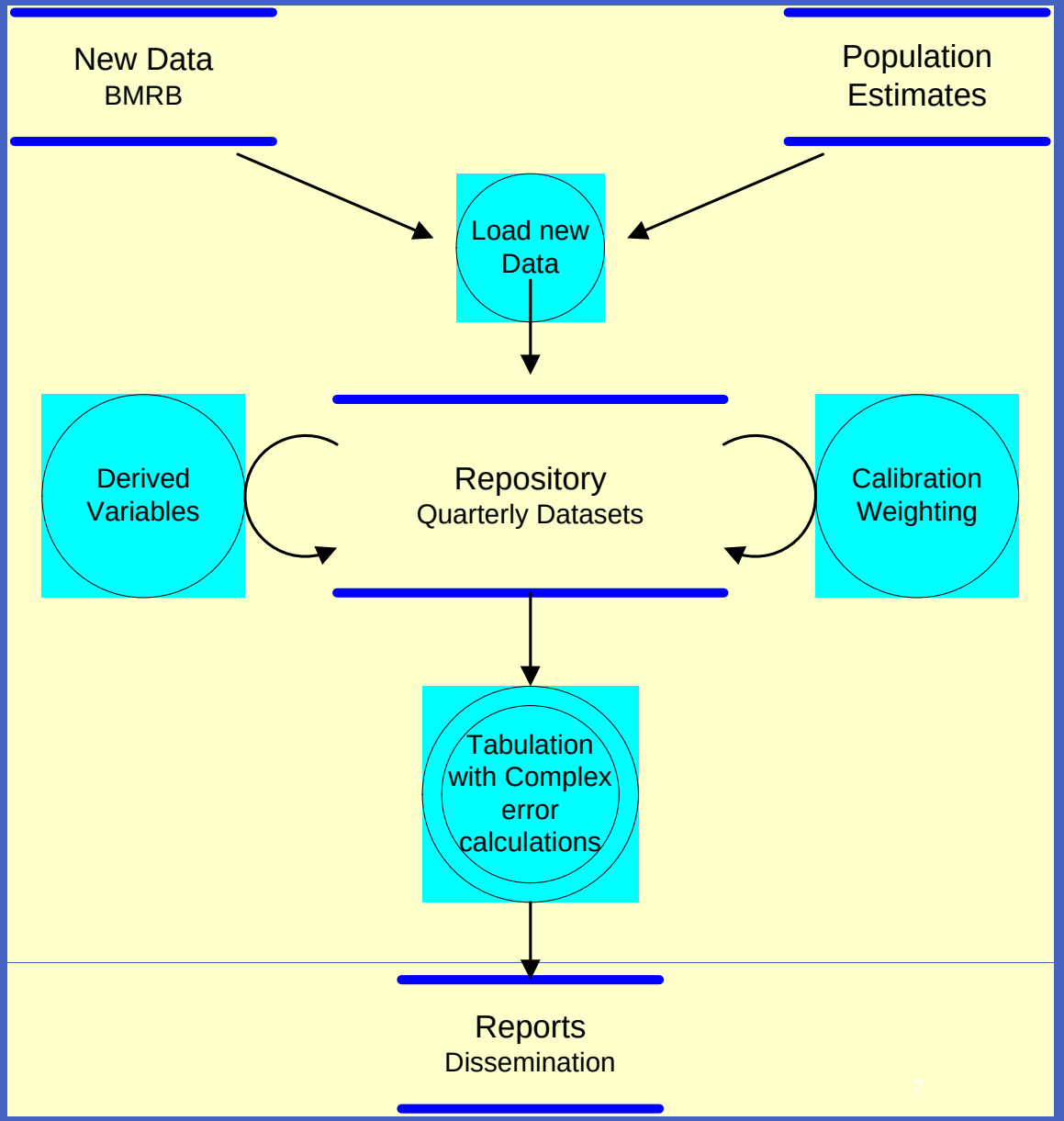
Issues in Processing

- Complex Process
  - » Well documented
  - » Many files
  - » Data Transfers to SAS, Stata
  - » Manual edits to scripts for file names
- Risk of errors
  - » Correct changes in correct places
- Time taken
  - » ~ 2 weeks to complete table production
- Brief
  - » To propose modifications that can be applied to the existing system to improve efficiency and timeliness



BCS Data Flow

- Possible Organisation
  - » Central Repository, containing NVF and VF records for each quarter, plus Pop. Estimates
  - » Derivation and calibration operate on repository, adding new variables to records
  - » Tabulation combines records from selected quarters
  - » Complex SD calculations done in SPSS



Solution Overview

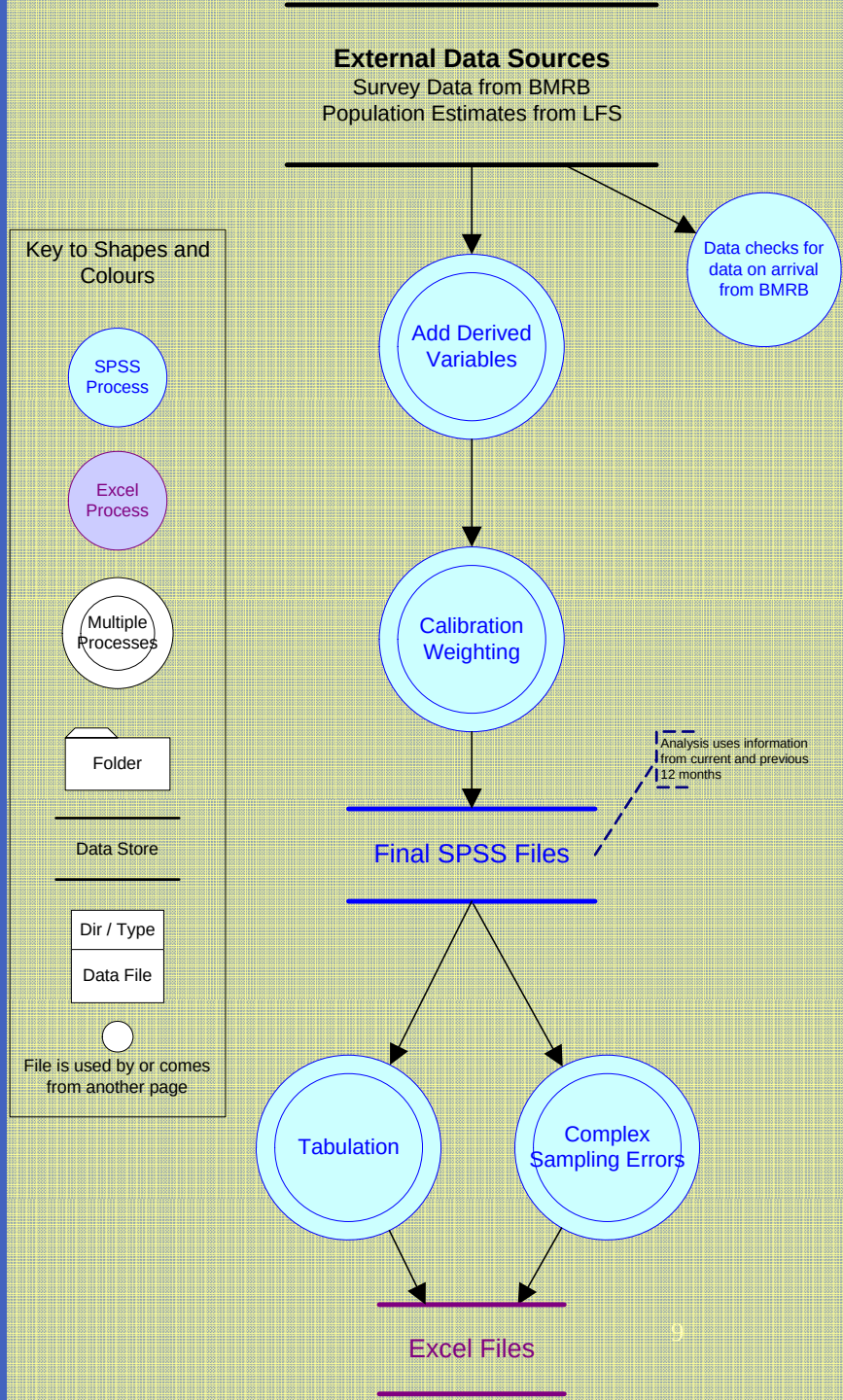
- Simplify directory and file name structure
  - » Easier to understand
- Organise scripts into modules
  - » Can be Included from others
  - » Easier to document and maintain
- Centralise period specification
  - » Use macros to reference
- Focus on SPSS
  - » Need more recent version for Complex Samples
  - » Use g-Calib (Statistics Belgium) for weighting

Process overview

- Derivation
- Calibration
- Tabulation

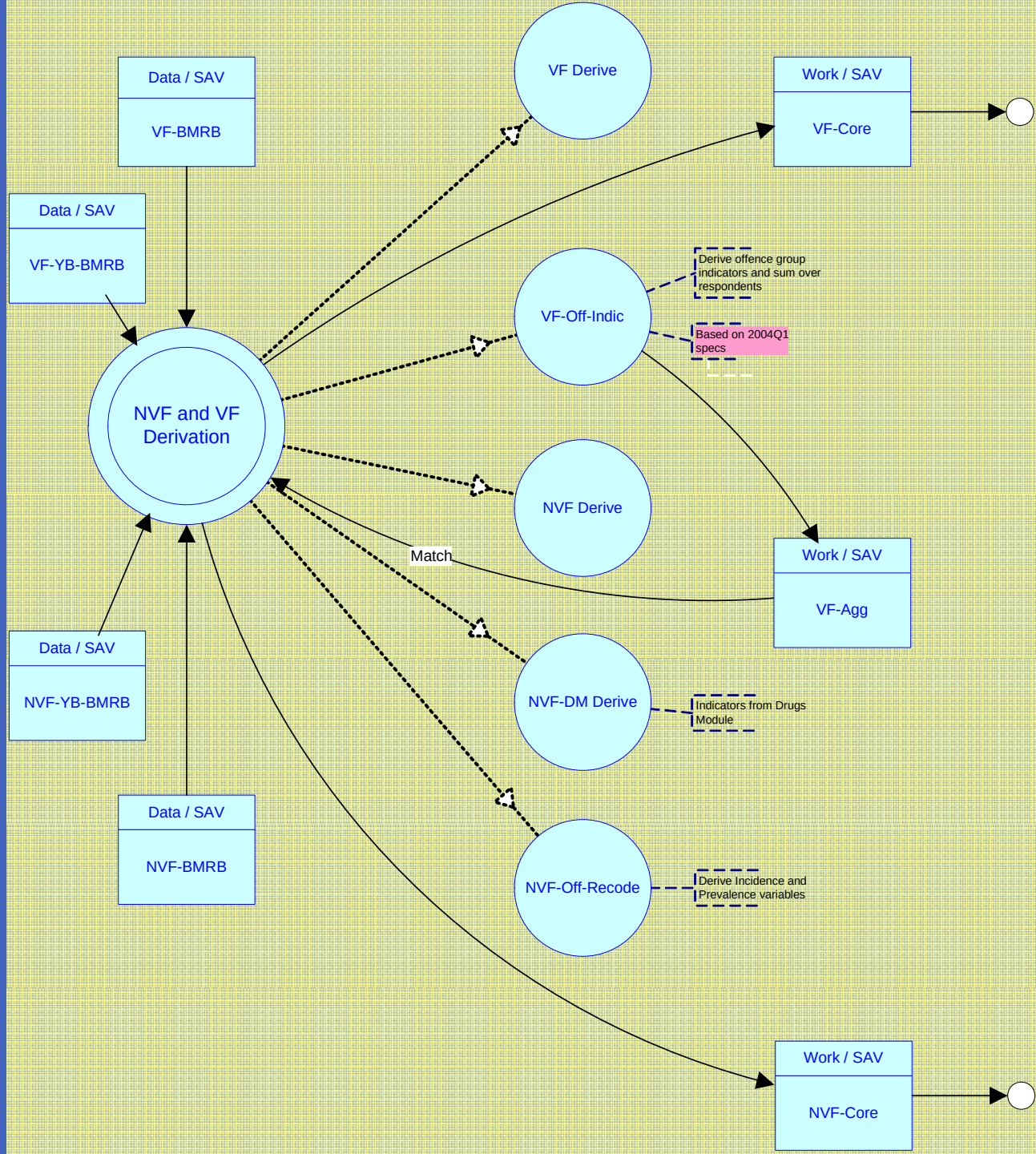
- Most steps involve various sub-processes

- Diagrams use Visio Data-Flow Diagram template

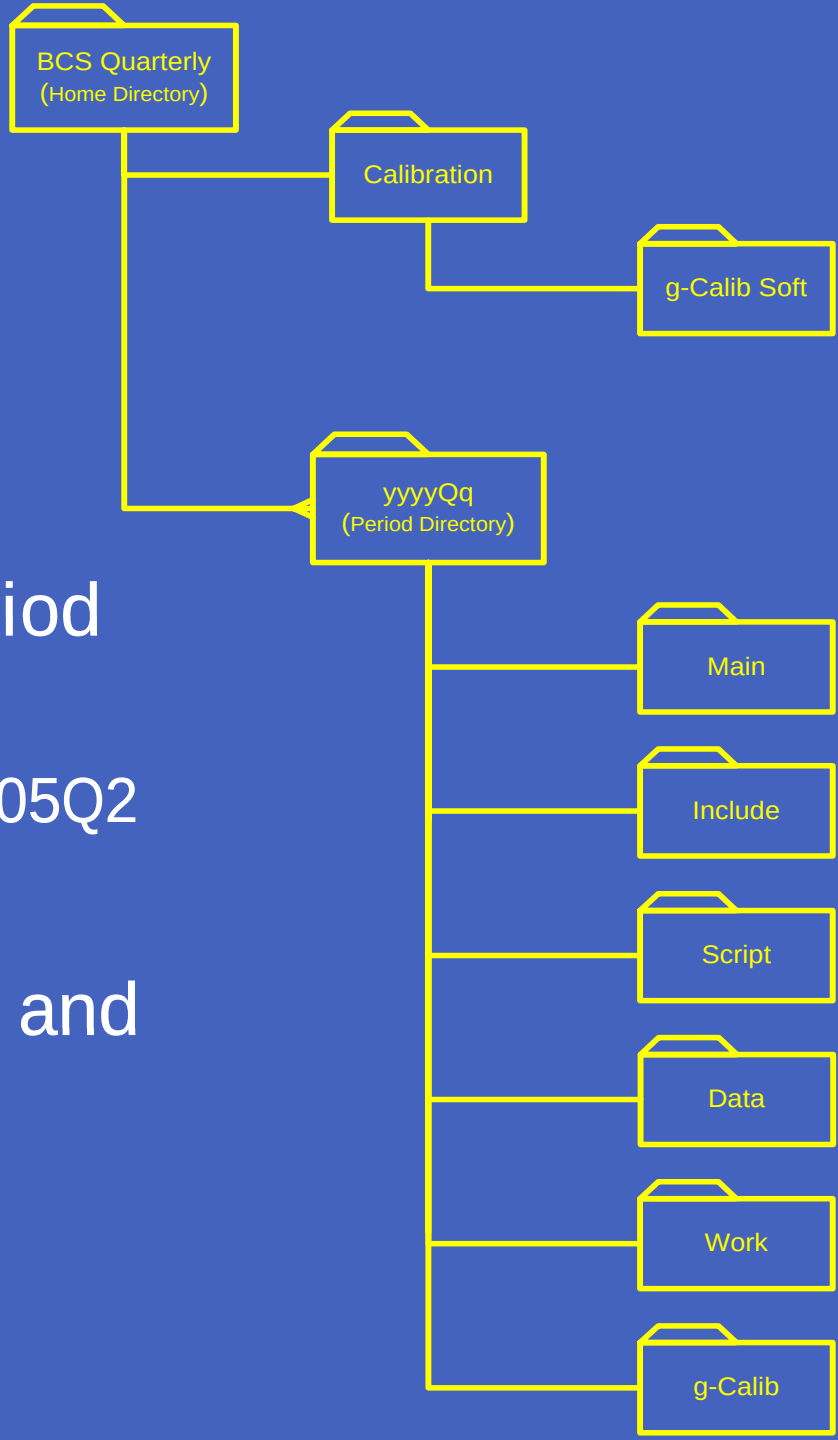


Example: Derivation Steps

- Data files are used
- Modules are included
- Data files are created



Directory and File Structure



- Home directory for each Period
  - » Template called yyyyQq
  - » Copy for each Period – e.g. 2005Q2
- File names do not change
- Separate data from modules and working files

Use of Macros

- All Period directory references are automated
- Set text values in a single place
  - `define mPeriodYear() "2005" !Enddefine.`
  - `define mPeriodQuarter() "2" !Enddefine.`
- Use in other definitions
  - `define mPeriodDir() !Quote(!Concat(!Unquote(!Eval(mPeriodYear)), "Q", !Unquote(!Eval(mPeriodQuarter)))) !Enddefine.`
  - `define mPeriodPath() !Quote(!Concat(!Unquote(!Eval(mBCSHomePath)), "\", !Unquote(!Eval(mPeriodDir)))) !Enddefine.`
- Reference in Scripts
  - `Get File = mPeriodPath + "\Data\NVF-BMRB.sav".`
  - `Execute.`
  - `Select if (SampType = 1).`
  - `Include mPeriodPath + "\include\NVF Derive.sps".`
  - `Execute.`
- Document Files
  - `ADD DOCUMENT`
  - `'Title: ' + mqPeriodTitle + '.'`
  - `'Content: NVF File - after adding derived variables, for ' + mPeriodDir + '.'`
  - `'  Combined Main, Youth and Ethnic samples.'.`

Documentation of processes

- Construct Modules by Function
  - » Easier to maintain
    - Youth and Drugs modules added subsequently
  - » Diagrams useful for overall flows
- Explanation of steps
  - » SPSS Syntax (with comments) acceptable for most basic derivations
  - » Better solution needed for complex transformations and derivations
    - Decision Tables
    - Incidence Tables
  - » Code and Table Generators?

Decision Tables

```
*SOC 100 CODES.
IF any (rsoc1990,100) respsc=1.
IF any
     (rsoc1990,101,102,103,111,113,120,121,122,123,124,125,126,127,130
     ,131,132,139,140,141,142,152,153,154,155) respsc=2.
IF any (rsoc1990,160,169,170,171,173,175,177,179,190,191,199)
     respsc=2.
IF any (rsoc1990,112) respsc=3.1.
IF (any (rsoc1990,110,172,174,176,178)) AND (respseg=1.2 or
     respseg=2.2) respsc=2.
IF (any (rsoc1990,110,172,174,176,178)) AND (respseg=5.2 or
     respseg=1.1 or respseg=2.1 or respseg=12) respsc=3.1.
*SOC 200 CODES.
IF any
     (rsoc1990,200,201,202,209,210,211,212,213,214,215,216,217,218,219
     ,220,221,222,223,224,230,232) respsc=1.
IF any (rsoc1990,240,241,242,250,252,253,260,261,262,290,291,292)
     respsc=1.
IF any (rsoc1990,231,233,234,235,239,251,270,271,293) respsc=2.
*SOC 300 CODES.
IF any (rsoc1990,
     300,301,302,303,304,309,311,312,313,320,330,331,332,340,341,342,3
     43,344,345,346,347,348,349,350) respsc=2.
IF any
     (rsoc1990,360,361,362,363,364,370,371,380,381,382,383,384,385,390
     ,391,392,394,395,396,399) respsc=2.
IF (any(rsoc1990,310,386,387,393)) AND (respseg=1.2 or respseg=2.2)
     respsc=2.
IF (any(rsoc1990,310,386,387,393)) AND (respseg=1.1 or respseg=2.1 or
     respseg=12 or respseg=5.2 or respseg=6) respsc=3.1.
*SOC 400 CODES.
IF any
     (rsoc1990,400,401,410,411,412,420,421,430,440,450,451,452,459,460
     ,461,463,490,491) respsc=3.1.
IF (rsoc1990=441) AND (respseg=2.2 or respseg=8) respsc=3.2.
IF (rsoc1990=441) AND (respseg=1.1 or respseg=2.1 or respseg=12 or
     respseg=10) respsc=4.
IF (rsoc1990=462) AND (respseg=5.2) respsc=3.1.
IF (rsoc1990=462) AND (respseg=6) respsc=4.
……
```

Sequence	RSoc1990	Resp Seg	Resp SC
0	100		1
1	101, 102, 103, 111, 113, 120, 121, 122, 123, 124, 125, 126, 127, 130, 131, 132, 139, 140, 141, 142, 152, 153, 154, 155		2
2	160, 169, 170, 171, 173, 175, 177, 179, 190, 191, 199		2
3	112		3.1
4	110, 172, 174, 176, 178	1.2, 2.2	2
5	110, 172, 174, 176, 178	5.2, 1.1, 2.1, 12	3.1
6	200, 201, 202, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 230, 232		1
7	240, 241, 242, 250, 252, 253, 260, 261, 262, 290, 291, 292		1
8	231, 233, 234, 235, 239, 251, 270, 271, 293		2
9	300, 301, 302, 303, 304, 309, 311, 312, 313, 320, 330, 331, 332, 340, 341, 342, 343, 344, 345, 346, 347, 348, 349, 350		2
10	360, 361, 362, 363, 364, 370, 371, 380, 381, 382, 383, 384, 385, 390, 391, 392, 394, 395, 396, 399		2
11	310, 386, 387, 393	1.2, 2.2	2
12	310, 386, 387, 393	1.1, 2.1, 12, 5.2, 6	3.1
13	400, 401, 410, 411, 412, 420, 421, 430, 440, 450, 451, 452, 459, 460, 461, 463, 490, 491		3.1
14	441	2.2, 8	3.2
15	441	1.1, 2.1, 12, 10	4
16	462	5.2	3.1
17	462	6	4

Incidence Matrix – for Offence Groups

```
if (any(offence, 80, 81, 82, 83, 84, 85, 86)>0)
    p1=number.
if (any(offence, 81, 82)>0) p2=number.
if (any(offence, 80, 83, 84, 85, 86)>0)
    p3=number.
if (any(offence, 51, 52, 53)>0) p4=number.
if (any(offence, 53)>0) p5=number.
if (any(offence, 51, 53)>0) p6=number.
if (any(offence, 51, 52)>0) p7=number.
if (any(offence, 52)>0) p8=number.
if (any(offence, 55)>0) p9=number.
if (any(offence, 61, 63)>0) p10=number.
if (any(offence, 60, 62)>0) p11=number.
if (any(offence, 71, 72)>0) p12=number.
if (any(offence, 60, 61, 62, 63, 71, 72)>0)
    p13=number.
if (any(offence, 60, 61, 62, 63, 71, 72, 81,
    82)>0) p14=number.
if (any(offence, 64)>0) p15=number.
if (any(offence, 50, 55, 56, 57, 58, 65)>0)
    p16=number.
if (any(offence, 51, 52, 53, 60, 61, 62, 63, 64,
    71, 72, 80, 81, 82, 83, 84, 85, 86)>0)
    p17=number.
if (any(offence, 50, 51, 52, 53, 55, 56, 57, 58,
    60, 61, 62, 63, 64, 65, 71, 72, 80, 81, 82,
    83, 84, 85, 86)>0) p18=number.
if (any(offence, 43, 44, 45, 51, 52, 53, 60, 61,
    62, 63, 64, 71, 72)>0) p19=number.
if (any(offence, 31, 34, 35)>0) p20=number.
```

Vari-able	Value	Group																		
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Offence	11																				
	12																				
	13																				
	21																				
	31																				*
	32																				
	33																				
	34																				*
	35																				*
	41																				
	42																				
	43																			*	
	44																			*	
	45																			*	
	50																*		*		
	51				*		*	*										*	*	*	
	52				*			*	*									*	*	*	
	53				*	*	*											*	*	*	
	55									*							*		*		
	56																*		*		
	57																*		*		
	58																*		*		
	60											*		*	*			*	*	*	
	61										*			*	*			*	*	*	
	62											*		*	*			*	*	*	
	63										*			*	*			*	*	*	
	64															*		*	*	*	
	65																*		*		
	67																				
	71												*	*	*			*	*	*	
	72												*	*	*			*	*	*	
	73																				
	80	*		*														*	*		
	81	*	*												*			*	*		
	82	*	*												*			*	*		
	83	*		*														*	*		
	84	*		*														*	*		
	85	*		*														*	*		
	86	*		*														*	*		

Calibration Weighting

- ONS recommendation is to use Calmar macro in SAS
- G-Calib from Statistics Belgium provides equivalent functions
  - » Implemented using Matrix facilities
  - » Controlled by a set of macros
  - » Has front-end to create macro values for exploratory use
- BCS processing is fixed
  - » Standard modules used for macros and data transformation
  - » Identical results for Individuals, small differences for Households

Complex Sampling Errors

- SPSS procedures produce identical estimates of SEs using sample Design Weights
  - » Comparison with Stata
- Calibration Weights combine design weights with data, so are random variables
  - » ONS recommendation is to use special macros in Stata
  - » These use CS regression methods with a linearised ratio estimate of the SEs
- CS estimates for regression available with V13 of SPSS
  - » We worked with Version 12 using CSDescriptives
  - » Use of Calibration weights as though they were Design weights produces slightly larger SEs (i.e. conservative significance levels)

Outcomes

- Performance
  - » New system takes ~1 hour on PC
    - ~ 44500 respondents (4300 variables) and 17500 events (970 variables)
    - Includes SEs and rates but not main tabulations or report production
  - » Easily extended to add Youth and Drug module processing
- Statistical results generally identical or conservative
  - » Some further exploration needed
  - » Specifications for Household Calibration
  - » Further experimentation with CSRegression in SPSS 15 now underway

Summary

- Solution chosen to work well with staff skills
- Needs to be understood and flexible, so that it can be updated
- Importance of good documentation of processes and steps, for updating and because staff change regularly
- Data structure not too complex, so no need for DBMS
  - » But could be used if Repository approach developed