Office for National Statistics

A Comparative Study of Complex Estimation Software

Jonathan Digby-North, ONS

Email: j.digby-north@ons.gsi.gov.uk

Collaborators: Andy Fallows, Megan Pope, Daniel Lewis and Gary Brown

Outline

- Overview of ONS and sampling
- SAS and R at ONS
- GES and ReGenesees
- Analysis/Testing
- Conclusions/Recommendations

Office For National Statistics

National Statistical Institute for UK

UK National Accounts

Decennial UK population census

UK Balance of Payments

Births, marriages and deaths

Price Indices (CPI, RPI)

Labour Market Statistics

Business Activity

Surveys and Sampling

- Use sample surveys to estimate population statistics
  - Timeliness
  - Burden
  - Cost
  - Quality

- Calibration
  - Use known (auxiliary) information about population in estimation
  - Produce weights that recover these population totals

- Complex survey designs – statistical software required for weighting and estimation (GES is standard tool)

SAS and R at ONS

- SAS
  - Standard statistical software at ONS
  - Production systems; Methods development
  - Contract for SAS licences

- R
  - Used since early 2000s
  - Becoming more commonly used in Methodology Dept.
  - R User Group; R Development Group

Generalized Estimation System (GES)

- Suite of SAS-based macros developed by Statistics Canada in 1990s
- Produces calibration weights, domain estimates and standard errors for variety of complex survey designs
- Used in a number of high profile business and social surveys at ONS (e.g. Annual Business Survey, Labour Force Survey)
- Appropriate SAS licence required and software needs to be purchased

→ Explore open source solutions in line with Government IT Strategy and Code of Practice for Official Statistics

CoP and UK Government IT Strategy

principle 7, practice 5: Seek to balance quality (for example, accuracy and timeliness) against costs (including both costs to government and data suppliers), taking into account the expected use of the statistics

"Where appropriate, government will procure open source solutions. When used in conjunction with compulsory open standards, open source presents significant opportunities for the design and delivery of interoperable solutions."

ReGenesees R Package

- R package - ‘R evolved Generalized software for sampling estimates and errors in surveys’
- Calibration tool developed by Diego Zardetto, Italian Statistics Office (Istat)
- Uses similar techniques to GES
- Wide range of functionality

Feasibility study: Could ReGenesees be used in place of GES at ONS?

Analysis

- Compare performance of ReGenesees and GES on a representative selection of ONS business and social surveys:
  - Quarterly Stocks Inquiry
  - Business Register and Employment Survey
  - Annual Business Survey
  - Labour Force Survey
  - Life Opportunities Survey
  - International Passenger Survey
- Assess in terms of viability, accuracy, run times, volume testing, handling of problematic data, ease of programming, functionality...

Analysis – Business Surveys

- Quarterly Stocks Inquiry
  - Cut-off sampling technique
  - Identical domain and variance estimates
- Business Register and Employment Survey
  - Calibrates to region and industry employment totals separately
  - Calculates coefficients of multiple regression models
  - Weights distribution very similar and coefficients successfully calculated
- Annual Business Survey
  - Complex weighting procedure outside GES
  - Artificial calibration enabled correct variance estimation

Analysis – Social Surveys

- Labour Force Survey
  - More than 1,000 calibration groups – resource intensive
  - After increasing memory allocation available to R, calibration was successful
- Life Opportunities Survey
  - Calibration constraints difficult to satisfy
  - ReGenesees unable to replicate GES weighting – cannot set absolute limits on weights distribution
- International Passenger Survey
  - Two consecutive calibration stages
  - Calibration ran successfully

Analysis – Calibration times

Survey	Records	Constraints	Calibration Time (s)	
			ReGenesees	GES
LOS	25,780	102	4	12
IPS	59,676	91 then 234	16	127
LFS	99,793	1,089	720	1080* (71)

*GES took this long to decide it could not process the data due to insufficient memory.

- Calibration faster in ReGenesees than standard GES
- For LFS GES failed after 18mins (modified GES ran in 71s)
- Reading large datasets into R initially took a long time → SQLDF package
- Writing ReGenesees results to CSV files from R took a few minutes (generally longer than the calibration)

Problematic data

	ReGenesees	GES
Missing records	➢ Stops and generates error message.	➢ Excludes total, continues without warning message
Missing calibration total	➢ Stops and generates error message	➢ Assumes a weight of one, continues without warning message
Missing values	➢ Stops and generates error message. Optional argument to override.	➢ Automatically ignores them

→ Less risk with ReGenesees?

Programming

GES

- Suite of many SAS macros
- Estimation or Weighting
  - One stage/Multi-stage
  - Sampling strategy
  - Level of calibration (element/cluster)
- Create many different input files
- Large number of data steps and SAS procedures required

ReGenesees

- Survey design, sampling strategy, calibration model and constraints specified using simple formulae
- One line of syntax to specify model
- Generates groups and totals
- Specification of above plus calibration and estimation stages – few lines of syntax

→ Amount and complexity of programming lower with ReGenesees

Conclusions and Recommendations

- Entirely feasible to replace GES with ReGenesees at ONS - is it viable?
- Cost savings in the long run as SAS requires a licence
- Likely huge costs to make system changes
- Thorough investigations required (to be specified by IT teams):
  - Compatibility testing
  - Upstream/downstream processing testing
  - Full end-to-end testing

Final recommendation: Organisations across the GSS should explore the introduction of ReGenesees as a replacement for GES in a production setting