A Comparative Study of Complex Survey Estimation Software in ONS

Andy FALLOWS
Megan POPE
Jonathan DIGBY-NORTH
Gary BROWN
Daniel LEWIS
Office for National Statistics, UK

ABSTRACT

Many official statistics across the UK Government Statistical Service (GSS) are produced using data collected from sample surveys. These survey data are used to estimate population statistics through weighting and calibration techniques. For surveys with complex or unusual sample designs, the weighting can be fairly complicated. Even in more simple cases, appropriate software is required to implement survey weighting and estimation.

As with other stages of the survey process, it is preferable to use a standard, generic calibration tool wherever possible. Standard tools allow for efficient use of resources and assist with the harmonisation of methods. In the case of calibration, the Office for National Statistics (ONS) has experience of using the Statistics Canada Generalized Estimation System (GES) across a range of business and social surveys.

GES is a SAS-based system and so is only available in conjunction with an appropriate SAS licence. Given recent initiatives and encouragement to investigate open source solutions across government, it is appropriate to determine whether there are any open source calibration tools available that can provide the same service as GES. This study compares the use of GES with the calibration tool 'R evolved Generalized software for sampling estimates and errors in surveys' (ReGenesees) available in R, an open source statistical programming language which is beginning to be used in many statistical offices.

ReGenesees is a free R package which has been developed by the Italian statistics office (Istat) and includes functionality to calibrate survey estimates using similar techniques to GES.

This report describes analysis of the performance of ReGenesees in comparison to GES to calibrate a representative selection of ONS surveys. Section 1.1 provides a brief introduction to the current use of SAS and R in ONS. Section 2 describes GES and ReGenesees in more detail. Sections 3.1 and 3.2 consider methods for analysing and comparing the performance of the two tools using case studies from business and social surveys respectively. The analyses cover a range of issues including use with large datasets and complex calibration problems. Section 3.3 describes more general

comparisons between the uses of each tool. The report finishes with a conclusion and recommendations. Annex A provides a glossary of key terms used in this report.
Key words: *R, GES, ReGenesees, Calibration*

SAS AND R AT ONS

SAS has an established presence in ONS, being designated as the standard software for statistical programming. SAS, and in recent years SAS Enterprise Guide, is widely used within production systems and across the office for analysis and development of methods. ONS has a contract for SAS licences to cover these varied uses.

R has been used since the early 2000s at ONS. When it was first introduced there were only a handful of technical specialists in Methodology using it as a research tool, mainly for spatial analysis. Now in 2015, R is commonly used in Methodology, mainly as a research tool, and its use has spread to other areas of the office.

There is an informal R User Group that provides support within the office and a small R Development Group, consisting of IT specialists, Methodologists and Statisticians from production areas who are exploring the possibilities of using R more widely in production processes. Currently R is directly used in the production process for one statistical output, mortality rates. This uses the MortalitySmooth package for implementing two-dimensional p-splines in the estimation of mortality rates.

1. COMPLEX ESTIMATION SOFTWARE

This section describes both the SAS-based GES – the current tool used for calibration and estimation of survey parameters at ONS – and the R-based ReGenesees. Throughout this work we use GES v4.3 and ReGenesees v1.6.

1.1 GES

GES was developed in the early 1990s by Statistics Canada – the methodological principles that underpin GES are described in Estevao et al. (1995). Briefly, GES is a suite of SAS macros which are used to produce calibration weights, domain estimates and estimates of variance for a variety of user-defined complex estimators produced under a generalised linear regression model.

At ONS, GES is used for a variety of tasks – it can be used to perform complex calibrations using multiple auxiliary variables or to use existing weights to produce domain estimates with accompanying estimates of variability. For example, in the Annual Business Survey (ABS), GES is used for the calculation of domain estimates and associated variances at the national level. In the Business Register and Employment Survey (BRES) however, a modified version of the GES code, known as 'Fast GES'[1], is used to calculate weights that calibrate to employment totals by both

1. Fast GES uses SAS data step and proc summary processing in place of the SAS IML code used to invert matrices of calibration constraints in GES. This solves the problem of large matrices failing to invert due to the high computing resource required to do this in SAS IML.

Government Office Region (GOR) and high level Standard Industrial Classification (SIC), separately, in 30 calibration groups. Fast GES is also used to produce multiple regression coefficients which are used in setting Winsorisation outlier thresholds. The majority of ONS social surveys use GES or Fast GES to calibrate to Age by Sex and Age by Region population totals.

1.2 ReGenesees

The ReGenesees package[1] is for design-based and model-assisted analysis of sample surveys with various complex designs. This comprises one part of the ReGenesees System - the other is the R package ReGenesees.GUI[2], which is the (optional) user-friendly graphical user interface.

There are a number of different complex survey designs which ReGenesees can analyse, including multi-stage, clustered and stratified designs. It can also handle unequally weighted sampling designs with or without replacement, and strata can be either fully enumerated or sampled.

Much of the information below was taken from the ReGenesees User Guide.

1.2.1 Calibration

ReGenesees can produce calibrated weights using any specified model, and can handle multiple auxiliary variables and constraints for the calibration simultaneously. For example, calibration could be both on the totals of two variables A and B within some calibration domain X, and on the total of variable C within another calibration domain Y. The software does this by iteratively solving the calibration problem in each different domain separately. The known population totals of the auxiliary variables for each of the calibration domains can be automatically calculated from a file containing the sampling frame. This means no pre-processing of the data is required. Range constraints can also be imposed on the calibration problem (range restricted calibration). These ensure that the ratios between the final calibrated weights and the initial/direct weights are constrained to be within an interval supplied by the user. Choosing this interval is made easier by a function which can suggest values for which the calibration algorithm is likely to converge.

During calibration, any heteroskedasticity in the model can be accounted for by simply supplying a variable correlated with the variances of the error terms. The result is that the final weights of units with large values (of this variable) will be much closer to their original weights than for units with small values.

Multiple calibrations can be performed consecutively – for example, it is common in social surveys to calibrate first to reduce non-response bias, and then to calibrate again to different population totals to reduce the variance of the estimates. It is also possible to perform multiple calibrations simultaneously – for example, you

1. http://www.istat.it/en/files/2014/03/ReGenesees.pdf

2. http://www.istat.it/en/files/2014/03/ReGeneseesGUI.pdf

may wish to calibrate both on the number of individuals by age and sex and on the number of households within a region.

ReGenesees also allows calibration on incomplete auxiliary information – for example, the population totals could only be known for specific industries within specific categories of a variable. There are also routines for further analysis of the calibration – one example is the calculation of the residuals (with respect to the calibration model) of any variable of interest. After calibration, the resulting weights that would be used in the estimation process can easily be retrieved.

1.2.2 Estimation

ReGenesees can produce estimates, including those for subpopulations, using the basic Horvitz-Thomson or ratio estimators. Estimates of totals, means, frequency distributions, ratios between totals and quantiles and estimates of the multiple regression coefficients of a specified linear regression model can all easily be calculated. Standard errors, coefficients of variation, confidence intervals and the related design effects can also be produced for each of these estimates.

A useful feature of ReGenesees is that estimates, standard errors and confidence intervals (including those for subpopulations) can be calculated for any complex estimator specified by the user – this can be any analytic function of Horvitz-Thompson or calibration estimators of totals or means. This function can also be used to calculate estimates of the population variance from the sample. A further useful feature of ReGenesees is the capability to quickly estimate the covariance and the correlation (including for subpopulations) between any two complex estimators. Once estimates have been calculated, it is also possible to determine the type of estimator (for example total, mean, ratio) that was used in the estimation process solely by supplying the estimates and associated information on the survey design.

1.2.3 Variance Estimation

In ReGenesees there are a number of options to choose from when it comes to variance estimation. For multi-stage sampling designs, the Ultimate Cluster Approximation (Kalton 1979) option can be invoked – this will only take into account the variance arising from the Primary Sampling Units (PSUs) and ignore any information regarding subsequent stages. This is known to underestimate the true variance, but the error becomes negligible when the sampling fractions in the strata are generally small. If this option is not used, then the contribution to the variance from each sampling stage will be accounted for using a recursive algorithm (Bellhouse 1985).

There is also a compromise solution where only the leading contribution to the variance is considered. For fully-enumerated strata, only the contribution from the secondary stages will be retained, whereas for non-fully enumerated strata, only that from the first stage will be retained. There are also a number of ways to deal with variance estimation for non-fully enumerated strata which contain only one observation; these include a strata aggregation technique (combining similar ‘lonely’

strata), an average variance contribution, using the standard deviation from the population means (rather than stratum means), or alternatively they can be ignored completely.

Another feature – currently under development – is the option to fit a suite of Generalised Variance Functions to the calculated statistics (that is, the estimates and standard errors) to determine how the variance varies as a function of the estimates. User-defined variance functions can also be supplied. The best-fitting model can then be chosen and could be used, for example, to summarise standard errors or to enable users to predict approximate standard errors for any given estimate. Diagnostic plots of the model fits can also be easily produced.

2. ANALYSIS

We decided to focus this study on surveys which utilize GES in different ways (calibration weighting/ variance estimation) or demonstrate how GES is applied to different survey designs (single/multi-stage stratified designs; 'cut-off' sampling designs).

The business surveys analysed were the Quarterly Stocks Inquiry (QSI), BRES and ABS. The social surveys were the Labour Force Survey (LFS), the Life Opportunities Survey (LOS) and the International Passenger survey (IPS).

2.1 Business Surveys

2.1.1 Quarterly Stocks Inquiry

QSI collects information from businesses on the total value of stock held at the beginning and end of each quarter. Businesses are selected from the following industries: mining and quarrying, manufacturing, energy, construction, motor trades and retail and wholesale. Businesses in some of these industries are also required to provide a breakdown of the stock value by asset (for example work in progress, goods on hand for sale). There are no direct publications from this survey; however, quarterly and annual changes in inventories are published as part of UK National Accounts. Changes in inventories is used in the compilation of the UK National Accounts expenditure of the Gross Domestic Product.

QSI employs a 'cut-off' sampling technique in which businesses with an employment size below 10 or 20 (depending on the industry) are not sampled. The business population is composed of mainly small businesses and in comparison to larger businesses these often have very little contribution to the overall estimates. By accounting for these small businesses during calibration, the cost of the survey is reduced and estimates are assumed to be representative of the whole population.

For QSI, estimates are produced for each industry by variables reflecting the types of asset that businesses may hold. For example, when businesses within the manufacturing industry are asked to provide an asset breakdown, they are asked to provide 'materials, stores and fuels', 'work in progress' and 'goods on hand for sale'. Estimates are then produced for each of these variables as well as the total for each industry.

GES is not currently used to produce estimates and variances for the QSI – however, for the purposes of this study, QSI data were used to make a straightforward comparison between GES and ReGenesees.

2.1.1.1 Estimates and variances

The majority of business surveys, including QSI, are one stage stratified designs with simple random sampling. A specific SAS macro within GES produces estimates and variance estimates for this type of design. By supplying several different input files to GES, domain estimates and variances can be calculated for one or more variables of interest.

Design weights for each stratum are calculated as the ratio of the number of businesses in the population to the number of businesses in the sample. QSI uses combined ratio estimation, in which calibration weights are calculated within calibration (model) groups rather than within individual strata. To produce these calibration groups, employment size bands are combined at an industry level, with fully enumerated strata kept separate. The following formula is used to calculate calibration weights:

$$g_{ij} = \frac{\sum_{i \in j}^{N} x_{ij}}{\sum_{i \in j}^{n} a_{ihj} x_{ij}}$$

where:

- x is the auxiliary variable
- a is the design weight
- g is the calibration weight
- i denotes each individual business
- h is the stratum
- j denotes the calibration (model) group

To produce the calibration weights for each model group, the auxiliary total must be calculated for the N businesses in the population (including unsampled strata) and the n businesses in the sample. In this particular survey, calibration is very simple as employment is the only auxiliary variable used.

In GES, after calibration weights have been calculated, the following input files have to be supplied to GES:

- sample file containing the raw data including element, stratum and calibration group identifiers
- population counts at stratum level
- design weights and calibration weights and
- the auxiliary variable used for calibration

In addition, two input files are supplied to specify the domains for which estimates and variances are required. Estimates are required by industry, so this variable is specified within the domain input file. The parameter input file allows specification of the estimate required. In this case it is 'total', however GES can also produce other estimates, such as means and ratios. This file also allows specification of the variables that estimates are required for. For QSI, the variables are asset breakdowns by individual industry. GES then has all of the information needed to produce estimates and variance estimates.

In ReGenesees, the first step is to specify the survey design using the *e.svydesign* command. Within this command, the sample dataset is supplied along with the element and stratum identifiers. The design weights are supplied as initial weights as are the population sizes at stratum level. The calibration weights were supplied to GES via an input file; however, in ReGenesees the calibrated weights were calculated using the *e.calibrate* function to ensure that they were consistent with those calculated manually. The *pop.template* routine first specifies the calibration groups and then the *fill.template* function automatically calculates the auxiliary totals from the sampling frame and the template specified. Similarly, these auxiliary totals include small businesses excluded due to the cut-off threshold. By giving these businesses the correct calibration group identifier on the population file, the correct totals are automatically calculated. Calibration groups are formed by combining employment size bands at an industry level, so these small businesses are included within these model groups. Fully enumerated strata are kept separate. This template is then specified in the *e.calibrate* function to calculate the weights. Calibration weights produced in ReGenesees equate to the product of the design and calibration weights supplied to GES.

The information supplied in separate input files to specify domains for estimation and variables in GES can be specified in one function using ReGenesees. The *svystatTM* command calculates the estimates and variance estimates. This command simply requires the calibrated design object output from *e.calibrate*, the type of estimate required (for example total or mean), the estimation domains and the variables that the estimates are required for.

The resulting domain estimates and variances produced using GES and ReGenesees were identical.

2.1.2 Business Register and Employment Survey

BRES has two main purposes. First, it is the official data source for estimates of employment and employees, which are published by industry and detailed geography. Second, the data it collects are used to update business structures and local unit information on the Inter-Departmental Business Register, which is the main sampling frame used for business surveys at ONS.

As briefly mentioned in §2.1, Fast GES – which in this case is BRES-specific – is used to calculate weights which calibrate to both GOR and high level SIC ('section') employment totals separately. In addition, it is used to calculate the coefficients of the multiple regression models used to set the Winsorisation thresholds

which identify outliers for a particular response variable. In §3.1.2.1 and §3.1.2.2, only these two aspects of BRES are investigated. Whilst undertaking this work, the SAS code that prepares the data for input into GES was completely re-written in R for consistency and to ensure that the entire process could be replicated without using SAS in any way.

2.1.2.1 Calibration Weights

When using the BRES-specific version of Fast GES to calculate weights that calibrate to employment totals, a number of pre-processing steps need to be completed.

One of these steps involves the creation of an element level file which contains design weights. These design weights are calculated differently depending on which stratum the business belongs to. For strata which contain large (>250 employment) businesses, the design weight is calculated as the ratio between the population count and the sample count. For the remaining strata, an adjustment is made to account for businesses that have ceased trading since the sample was selected. This adjustment assumes that for each business that has ceased trading, another has begun. This adjustment can be implemented easily in ReGenesees by first calculating these adjusted design weights and then supplying them to the program via the 'weights' argument when specifying the survey design in *e.svydesign*.

Another pre-processing step involves the calculation of the auxiliary (in this case employee) totals that the weights are calibrated to. Separate files need to be created for each of the different types of domain – one file needs to contain employee totals by section and the other by GOR. Each of these files must contain a variable specifying an identification number (as well as the actual section/GOR variable) for each calibration group as well as the auxiliary variable to be used and the calibration total. In addition to this a 'model specification' file must be created, which again defines the auxiliary variables to be used and points to which files contain the calibration totals. The sample file must also contain variables which specify which section and GOR calibration group (in terms of the above identification number) each element belongs to, in addition to the section/GOR variables which are already on these files. The names of these additional variables must also be specified in the 'model specification' file. In the version of the SAS code which is currently used, it takes a large number of SAS data steps and procedures to manipulate the data into the required datasets for input into GES.

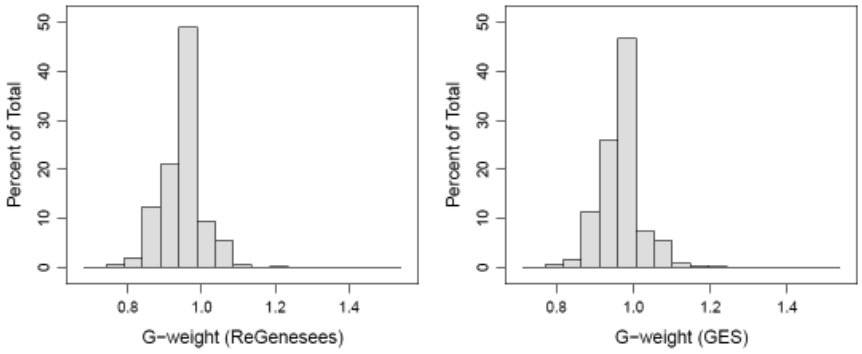
In ReGenesees, the specification of the auxiliary variables and the calculation of the calibration totals is a simple process. There is a routine named *pop.template* which quickly produces an empty template containing every calibration group of interest. This can be created using one short line of code which specifies the calibration model in terms of the two calibration group variables (GOR and section) and the auxiliary variable. The simple inclusion of a '+' symbol between the two calibration variables indicates that the marginal distribution totals should be used rather than the joint distribution totals. This template is then automatically filled with the correct

calibration totals by supplying the sampling frame to the *fill.template* routine using another single, short line of code. If the sampling frame is not available then the template can be filled with the known totals manually.

The next step is to actually perform the calibration and produce the calibrated weights. This is accomplished by the call to GES or by the call to *e.calibrate* in ReGenesees. In SAS, the calibration step is carried out for large and small businesses separately and the results are then combined – this was replicated in ReGenseeses. The calibration procedure was successful – the auxiliary population totals were exactly recovered by using the calculated weights. The distribution of the final calibrated weight divided by the design weight ('g-weight') produced by both the BRES-specific version of Fast GES and ReGenesees were very similar (see Figure 1 and Table 1).

Comparison of the weights produced by GES and ReGenesees

Figure 1



Summary statistics of the weight distributions

Table 1

	GES	ReGenesees
Minimum	0.753	0.723
1st Quartile	0.937	0.916
Median	0.970	0.958
Mean	0.967	0.950
3rd Quartile	0.986	0.979
Maximum	1.498	1.499

2.1.2.2 Multiple Regression Coefficients

The BRES-specific version of Fast GES is also used to estimate the coefficients of a multiple regression model which describes each of the full-time, part-time and total

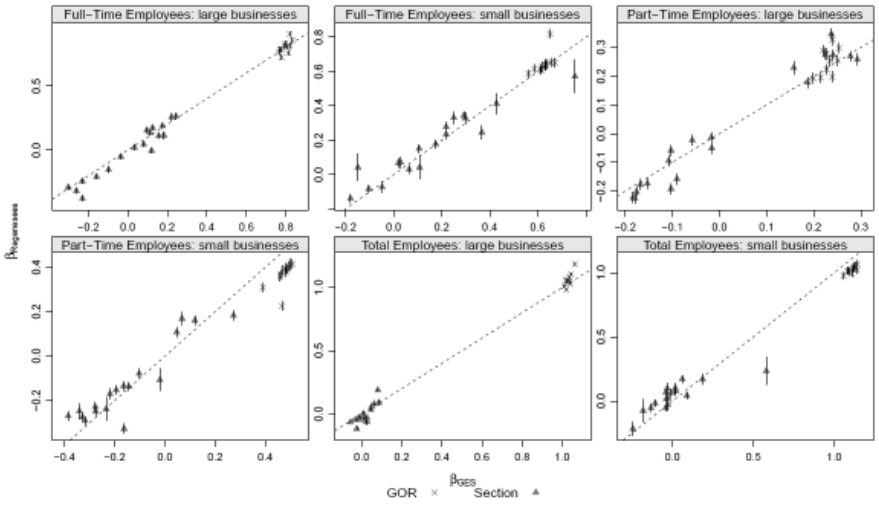
employee variables as a linear function of the number of employees by section and by GOR. The regression coefficients are output from GES and this information is used to set the outlier threshold values in the Winsorisation process.

Regression coefficients of a multiple regression model can be estimated in ReGenesees using the *svystatB* routine. This requires one line of syntax specifying the design object (containing the survey data and sampling design metadata – this is produced via the *e.svydesign* routine) and the regression model whose coefficients are to be estimated. There are various other optional arguments which allow the calculation of the design effect, the standard errors of the estimates of the regression coefficients and the associated confidence intervals (which can be calculated at any user-defined confidence level).

The coefficients (β values) calculated using the BRES-specific Fast GES code and ReGenesees are compared in Figure 2. As mentioned previously, the calibration is performed for large and small businesses separately. An advantage of ReGenesees over GES is that the ReGenesees estimates of the coefficients are accompanied by standard errors (shown in Figure 2) whereas those produced by GES are not. The figure shows that ReGenesees is producing reasonable results when compared to GES, with the regression coefficients for full-time employees being the most consistent.

Comparison of the multiple regression coefficients for full-time, part-time and total employees for large and small businesses separately

Figure 2



2.1.3 Annual Business Survey

The ABS[1] is the largest ONS business survey. It mainly covers the production, construction, services and distribution industries and provides important indicators of economic activity in the UK. These include total turnover, the total value of purchases of goods, materials and services, total employment costs and approximate Gross Value Added at basic prices (aGVA). Much of the ABS information feeds directly or indirectly into the UK National Accounts.

GES is currently used in the ABS system to produce the published national estimates and standard errors for a number of variables down to the four-digit SIC level, as well as at a SIC section by employment size-band level. The specific GES routine used is for a stratified one-stage simple random sampling design without replacement. However, the weights used to calculate estimates are not produced by GES; these are calculated externally and then input into GES to produce the domain estimates and standard errors of interest.

This section therefore concentrates on trying to utilise ReGenesees in a similar way – directly using externally calculated weights to produce the estimates whilst correctly handling the variance estimation by accounting for the survey design. The variables considered in this work are total business turnover and total business purchases. Estimates and standard errors were produced for these variables at a SIC section and division (2-digit SIC) level using the revised 2011 ABS data (results published in June 2014).

2.1.3.1 External Calibration

To produce the external weights, the original sampling strata are first collapsed to form larger strata. Design weights are calculated and adjusted to account for businesses that have ceased trading since the sample was originally taken. Calibration groups are either the strata themselves or are formed by grouping strata within the same industry together across several employment size bands, which can differ across industries. A correction to the calibration weights is also made to account for businesses that have ceased trading. Outliers are identified via both automatic and manual processes and are treated via post-stratification; they are removed from their original design strata and given a weight of one – the weights for their original strata are then recalculated.

In GES, the adjusted design and calibration weights are supplied together with files which identify the calibration groups and the auxiliary variable used in the calibration. The estimates are produced by simply aggregating the weighted responses to the domain level of interest. When calculating variance estimates, GES calculates the residuals in each calibration group separately.

In ReGenesees, to replicate the process the externally calculated final weights (the product of the design and calibration weights) were supplied via the 'weights' argument in the *e.svydesign* routine. Domain estimates and standard errors were then be calculated using the *svystat* routines on the un-calibrated design object produced by

1. http://www.ons.gov.uk/ons/guide-method/method-quality/specific/business-and-energy/annual-business-survey/quality-and-methods/abs-technical-report---june-2014.pdf

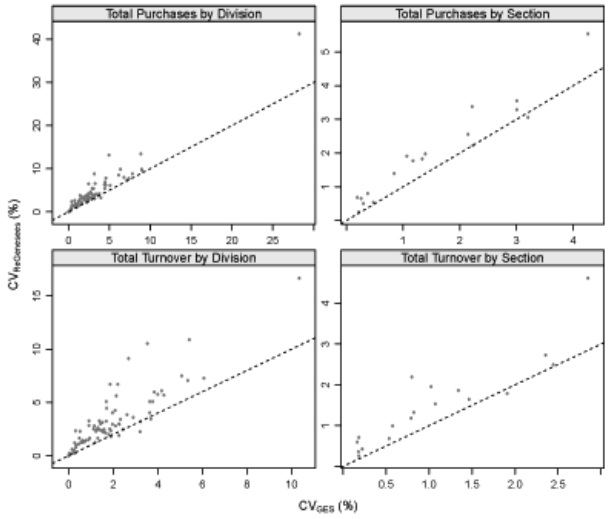
e.svydesign. The resulting estimates were consistent with those produced by GES, as again the weighted responses were simply being aggregated to the domain of interest.

However, because further calibration to the weights was not required, the *e.calibrate* routine was not initially used. This meant ReGenesees was not supplied with the calibration model or the calibration groups used to produce the weights and so during the variance estimation, the residuals were not calculated within each calibration group separately. In general, the resulting variance estimates for a given domain were markedly higher than those produced by GES. This is shown in Figure 3 where a comparison of the coefficients of variation (CVs) for estimates of total purchases and total turnover at the division and section level is made. The dashed line shows equality between the two values.

After further investigation, by performing an *artificial* calibration within ReGenesees the variance estimation could be handled correctly. This step involves first calculating the known auxiliary totals within each calibration group, using the weighted auxiliary variable on the sample file – this can quickly be accomplished through the use of the routine *aux.estimates*. As the weights already calibrate to these totals, *e.calibrate* was then used to perform a calibration leaving the external weights (and therefore domain estimates) unchanged, whilst at the same time providing ReGenesees with the information it required to correctly estimate the variances. After performing this 'artificial calibration' the ReGenesees and GES CVs became consistent, with all CVs being within 0.03 percentage points and the majority being identical.

Comparison between ReGenesees and GES CVs for the estimates of total turnover and total purchases at the SIC division/section level without artificial calibration

Figure 3



2.2 Social Surveys

Social surveys in ONS currently all use some form of GES as part of their processing.

Household surveys are cluster samples from the post code address file (PAF), which is a list of post codes and addresses. Once the addresses are selected all eligible persons within the household are then interviewed (although for some surveys only one person is randomly selected). The weight is constructed by adjusting the design weight for non-response and then calibrating to population totals provided by ONS demography. This weighting is integrated so that all persons within the same household receive the same final weight, meaning that household level estimates can also be formed.

Some surveys have a large number of calibration groups whilst others have groups with a relatively small number of records in them. Bounds are sometimes used to avoid large weights.

2.2.1 Labour Force Survey

The LFS is the largest household survey run by ONS, consisting of approximately 100,000 individuals per quarter (the LFS also forms the majority of the Annual Population Survey and Integrated Household Surveys). It is calibrated to local authority (433 groups), age band by sex (12*2*2 = 44 groups), and age band by sex by region (18*17*2 = 612 groups) to give 1,089 calibration groups in total. Fast GES is currently used to process the survey because of the high amount of memory required to invert the constraints matrix. Further details of the weighting methodology used in the LFS can be found in the LFS User Guide Volume 1[1].

LFS is by far the most resource intensive social survey and was deliberately selected for this analysis to test the memory capabilities of R. The default set up in R caused an error message indicating that the total memory allocation had been reached. Memory.limit was then used to increase the allocation and enabled the successful running of the calibration procedure.

In SAS, there were six seconds of pre-processing of the datasets ready for Fast GES, which took 1 minute and 11 seconds to run. There was then an additional second of running to tidy up the output dataset. To contrast this, the standard use of GES could not process the LFS data and took 18 minutes to determine that it did not have the memory.

In R, there were five seconds of pre-processing (although quite a lot of this had already been done). ReGenesees took 12 minutes to run. It then took 18 seconds to export the output dataset to a CSV file.

2.2.2 Life Opportunities Survey

LOS is a longitudinal survey of impairment, where the same respondents are followed up at each wave. Wave 2 had roughly 26,000 responding individuals (not all Wave 1 respondents were approached for a second interview). The longitudinal

1. http://www.ons.gov.uk/ons/guide-method/method-quality/specific/labour-market/labour-market-statistics/volume-1---2011.pdf

weights were calculated by adjusting the wave 1 weight for loss to follow up and then calibrating back to Wave 1 totals.

Household surveys are normally calibrated to standard age band by sex and region totals, but as part of the development work for Wave 2, a larger number of groups were used in addition to impaired and non-impaired totals to see whether it was possible to improve the match between the Wave 2 and Wave 1 estimates. These additional groups were sex by age band by region (2*5*10 = 100) and impaired/non-impaired (2) to give 102 calibration groups. This survey and example was chosen because some of the calibration groups have only a few records, but have a relatively high total, meaning that the calibration constraints are not easily satisfied. Details of the longitudinal weighting used for the survey can be found online[1].

It was not possible to exactly replicate the GES weighting in ReGenesees because there is no option to limit the absolute value of the calibrated weight (there is however an option to limit the ratio of the pre and post calibrated weights). Without this option several records get some extremely high weights which can result in some estimates getting larger standard errors. However, the final impact on estimates was marginal – a few percentage points at maximum.

To illustrate the above, estimates were formed for a survey variable with three outcomes (A, B and C) using the two sets of weights to demonstrate the impact (Table 2). As can be seen there is a small increase in the CVs. This example is for a variable at the top level – if the estimates are at a domain level then the impact will be greater. Table 3 presents a comparison of run times.

Estimates (%) with different sets of weights

Table 2

	Absolute Limit			Ratio Limit		
	Estimate	**SE**	**CV**	**Estimate**	**SE**	**CV**
A	28.43	0.32	1.13	28.43	0.35	1.23
B	13.09	0.23	1.76	13.05	0.24	1.84
C	58.47	0.36	0.62	58.53	0.40	0.68

2.2.3 International Passenger Survey

The IPS is a survey of passengers at ports across the UK consisting of 62,500 (travel and tourism) interviews per quarter. For airports, timeslots are sampled and for seaports, individual crossings are sampled. There are several initial stages of weighting which account for selection probability, non-response and partial responses. There are then two stages of calibration, the first calibrating to totals – provided by the Civil Aviation Authority (CAA) and the Department for Transport (DfT) – covering the time period where interviewing occurs (known as 'in hours'). The second stage calibrates to all traffic totals – including out of hours and the smaller ports which are not sampled – also provided by CAA and DfT. This survey was selected because of the complexity of having two concurrent calibration stages as well as the fact that there are frequently calibration groups with no sample returns.

1. http://www.ons.gov.uk/ons/guide-method/method-quality/specific/social-and-welfare-methodology/life-opportunities-survey/los-annex-1---weighting---part-2.pdf

Details of the weighting procedures are summarised online[1]. It was relatively straightforward to recreate the results. See Table 3 for a comparison of run times.

2.3 Software Comparison

This section briefly compares GES and ReGenesees in terms of performance (including run times, volume testing and how missing data are handled) and certain aspects of the programming.

2.3.1 Run Times

Table 3 summarises the run times and number of records processed for calibration for each of the social surveys considered in §3.2.

Run times (seconds) and number of records for each survey

Table 3

		R	SAS
LFS	99,793 Records		
	1,089 Constraints		
	Pre-processing	5	6
	Calibration	720	71
	Post-processing*	18	-
	Total	**743**	**77**
LOS	25,780 Records		
	102 Constraints		
	Pre-processing	7	39
	Calibration	4	12
	Post-processing*	36	19
	Total	**47**	**70**
IPS	59,676 Records		
(1)	91 Constraints		
	Pre-processing	4	**21**
	Calibration	5	**58**
	Post-processing	1	**1**
	Total	**10**	**80**
(2)	234 Constraints		
	Pre-processing	76	1
	Calibration	11	69
	Post-processing	1	1
	Total	**88**	**71**

* Includes time taken to write results to CSV file for ReGenesees

From Table 3 it can be seen that when the number of calibration constraints becomes high (greater than around 200), ReGenesees becomes slower than GES/Fast GES, whereas when the number of constraints is low, ReGenesees is the faster option. However, if the comparison is strictly between the standard version of GES and ReGenesees, both pieces of software are extremely similar in their running times.

For ReGenesees, it would seem that the key driver behind the processing time is the number of constraints rather than the number of records. The second stage

1. http://www.ons.gov.uk/ons/guide-method/method-quality/specific/travel-and-transport-methodology/international-passenger-survey/index.html

of calibration for IPS has just over double the number of constraints which has the result of adding 76 seconds to the processing time.

2.3.2 Missing Values

To further test both pieces of software it was decided to try removing all of the records from the dataset belonging to one calibration group, and then separately remove the equivalent total from the population totals to see how the software would react in both situations.

In GES, if there are no records in the survey data matching a total in the population file, then it will simply exclude that total from the population and calibrate to the remainder. If the population is broken down by different variables – for example, age by sex AND region – then this will result in proximal calibration because the sum of the age by sex groups will not equal the region totals. If there are records in the survey data, but no equivalent total in the population file, then the calibration weights will be 1 and so the record will retain its selection weight (GES is assuming that they have been selected with certainty).

In ReGenesees, there is a test to make sure that the survey data and the population totals are consistent (it is suggested that the survey data is used as a template to get the population totals in the correct format), so if there is a population total, but no records in the survey data to match it, then the program will generate an error message stating that the dimensions of the dataframes do not agree. Similarly if there are records in the dataset that do not match a population total then the program will also generate the same error message.

Missing values were introduced to the data in various ways to investigate and compare how ReGenesees and GES would respond. If the variable of interest has missing values then GES automatically ignores them and continues with the processing, generating a warning message informing the user that one or more observation were rejected due to missing values. ReGenesees on the other hand aborts processing and generates an error message. An optional argument can be included which instructs ReGenesees to ignore missing values and only use the non-missing values in the calculations. In this case a warning message will be generated detailing the number of observations which had missing values and which variable they occurred in. If the missing values occur in an auxiliary variable, ReGenesees will fail during the calibration stages (it will not be able to create the population totals template required for the calibration) and generate an error message. These observations will need to be either removed manually or be given a value for ReGenesees to run. GES will also fail if there are missing/zero values for the auxiliary variable (although the BRES-specific 'Fast-GES' will reject them, generate a warning message and continue).

In conclusion, GES will continue to run in the presence of missing values, and the presence of missing values will be unknown to the analysts reviewing results – unless code is written to generate an error – but ReGenesees will generate an error and stop. In terms of missing data then, ReGenesees would be the more prudent choice.

2.3.3 Programming

GES is essentially a suite of SAS macros, each of which accomplishes a specific task (such as weighting or estimation) for a specific survey design. For example, there are different estimation macros for one-stage or multi-stage designs and different macros for different sampling strategies (for example simple random sampling with or without replacement, probability proportional to size). There are also different macros for calculating calibration weights at different levels (for example element or cluster level) for the same survey design.

There is a Graphical User Interface (GUI) for GES which has drop down menus so that the user can select, for example, the survey design, sampling strategy, calibration model and required estimation domains. The GUI will then call the appropriate SAS macros for the survey design and the task at hand. However, ONS still generally create the required input datasets and directly call these macros from within specific SAS programs. Even excluding the macros that have been modified by staff at ONS, there are a multitude of different GES macros, most of which have different inputs (although some of these can be the same). So in addition to ensuring that the correct macro is chosen, there are varying amounts of pre-processing to undertake before the macros can be called.

In contrast, the survey design, sampling strategy, calibration model and the associated constraints are specified using simple formulae in ReGenesees. In fact, the specification of the survey design/calibration problem and the subsequent calibration and estimation stages can all be completed in a few short lines of syntax. There is also the choice of using the ReGenesees.GUI package (the user friendly interface), although a comparison between this and the GES GUI was out of scope of this project.

One issue that was encountered during the course of this work (and which is documented in the ReGenesees user guide) is that of contrast handling. This is the way in which R transforms the symbolic calibration model specified by the user into numeric matrices. By default, R (and also ReGenesees) is set to avoid the complete dummy coding of a factor when it thinks that some of the factor levels would produce redundant columns in the model matrix (i.e. it automatically simplifies the model matrix), reducing the computation time. In some cases, however, this causes unexpected results when using ReGenesees – when the population totals template was produced in §3.1.2.1, some of the calibration groups were missing and the calculated totals were incorrect. This then lead to incorrect calibration weights being produced. This situation can easily be avoided by switching all contrast handling off using a short command. Although this increases the computation time, it guarantees the correct population template and totals will be produced. As it took some time to find this solution, it would seem that perhaps this should be the default setting in ReGenesees, rather than one which needs to be specified by the user.

In conclusion, the amount and complexity of programming required to achieve the same result is higher using SAS and GES when compared to using R and ReGenesees.

CONCLUSION

The aim of this project was to compare two complex estimation software packages – one SAS-based (GES), and the other R-based (ReGenesees). The context is that GES is currently used by ONS to perform a range of calibration and estimation tasks on both business and social surveys. Ultimately, we wanted to determine whether ReGenesees (an open source package) could replace GES, which requires not only a SAS licence but also to be purchased from a vendor (Statistics Canada).

A range of business and social surveys were selected, each of which uses GES for a different task or is unique in some way. ReGenesees was successfully used in place of GES in each situation. Overall, the results from both pieces of software were very similar and each task was easier to implement in ReGenesees. In general, although there are issues related to using ReGenesees (for example, the default way it treats contrast handling), there is more risk attached when using GES, due to the way in which it treats missing values and the amount of pre-processing that is required.

There were, however, some attributes of GES which were not available in ReGenesees, which meant that some results could not be exactly replicated. An example is that in ReGenesees, bounds cannot be set on the absolute value of the calibrated weights, which can result in some extremely large weights and therefore higher standard errors for some estimates (although the impact on the estimates themselves was found to be marginal). There were also numerous attributes of ReGenesees that are not available in GES (see §2.2 for a review of the full functionality of ReGenesees). These include allowing the user to specify complex estimators and to choose between different variance estimation options.

In summary, it seems that it is entirely feasible to use ReGenesees in place of GES at ONS. The issue of whether this would be cost effective, or indeed even viable, when considering the system changes required is something that needs to be thoroughly investigated before any action is taken.

RECOMMENDATIONS

The recommendation of this report is for organisations across the GSS to explore the introduction of ReGenesees as a replacement for GES in a production setting.

These investigations would obviously require compatibility testing, upstream and downstream processing testing, and full end-to-end testing. The specification of this testing should be developed by the IT teams supporting current production systems, but the research team responsible for this project is willing to help and advise wherever needed.

References

1. Bellhouse, D. R., Computing Methods for Variance Estimation in Complex Surveys . Journal of Official Statistics, Vol. 1, No. 3, pp. 323-329, 1985
2. Estevao, V., Hidiroglou, M.A. and Särndal, C.E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. Journal of Official Statistics, 11, pp 181-204.
3. Kalton, G., Ultimate cluster sampling , Journal of the Royal Statistical Society, Series A, 142, pp. 210-222, 1979

Annex A:

Glossary of terms

Auxiliary variable – A variable value that is either known or can be estimated relatively accurately that is then used as a constraint in the estimation of other variables.
Calibration – An estimation procedure which constrains sample-based estimates of auxiliary variables to known totals (or accurate estimates).
Calibration weight – The adjustment made to the initial weight in order to calibrate to known (or accurately estimated) totals – also known as a g-weight.
Coefficient of Variation (CV) – A measure of relative dispersion equal to the standard error of an estimate divided by its arithmetic mean. A CV is a measure used to assess the quality of an estimate.
Confidence interval – An interval within which the true value of a population parameter lies with known (usually high) probability.
Correlation – A measure of association between two variables. It measures how strongly the variables are related or change with each other; this can either be positive or negative.
Covariance – Indicates how two variables are related.
Design weight - The inverse of the probability of selection, also known as an a-weight.
Domain – Any well-defined sub-group of the population.
Generalised Variance Function – A method of variance estimation. This function connects the variance of a survey estimator to the expected estimator.
Heteroskedasticity – Non-uniformity in the variance.
Horvitz-Thomspson – An estimation using only weights which are the inverse of the selection probabilities. This is also known as an expansion estimator.
Marginal distribution – The probability distribution of variables given in a subset. No reference is made to any other variables.
Proximal calibration – When trying to simultaneously calibrate to two conflicting totals, proximal calibration ensures an optimal mid-point is taken.
Residual – The difference between an observed random value and its mean or an estimate of its mean.
Standard deviation – A measure of dispersion of the values of a random variable around their mean. This is the square root of the variance.
Standard error – An indication of the precision of an estimate calculated as the positive square root of the variance of the sampling distribution of a statistic.
Variance – A measure of dispersion of the values of a random variable around their mean. The variance captures the sampling error.
Winsorisation – A method used to detect and treat outliers.