ReGenesees: an Advanced R System for Calibration, Estimation and Sampling Errors Assessment in Complex Sample Surveys

Diego Zardetto[1]
[1]Istat - Italian National Institute of Statistics, e-mail: zardetto@istat.it

Abstract

ReGenesees is a new system for design-based and model-assisted analysis of complex sample surveys, based on R. As compared to traditional estimation platforms, ReGenesees achieves a dramatic reduction in user workload for both the calibration and the variance estimation tasks. Moreover, it can handle arbitrary Complex Estimators, provided they can be expressed as differentiable functions of Horvitz-Thompson or Calibration estimators of totals. All such innovative features leverage a peculiar strong point of the R programming language, that is its ability to process symbolic information.

Keywords: Complex estimators, symbolic computation, automated linearization.

1. What is ReGenesees

ReGenesees (R Evolved Generalized Software for Sampling Estimates and Errors in Surveys) is a full-fledged R software for design-based and model-assisted analysis of complex sample surveys. This system is the outcome of a long-term research and development project, aimed at defining a new standard for calibration, estimation and sampling errors assessment to be adopted in all Istat large-scale sample surveys.
The first public release of *ReGenesees* for general availability dates back to December 2011. The system is distributed as Open Source Software, under the EUPL license. It can be freely downloaded from Istat website[1], as well as from JOINUP[2] (the collaborative platform for interoperability of the European Commission).
Till the advent of *ReGenesees*, the estimation phase for sample surveys was handled in Istat by a SAS application named GENESEES. The name of the new R-based system has been deliberately chosen to emphasize Istat's seamless offer of software tools dedicated to that phase, while highlighting – at the same time – its *evolution* end *enhancement* through R. It is worth stressing, anyway, that the *ReGenesees* system is not the outcome of a bare migration towards R of its SAS ancestor, but rather the fruit of a new, challenging and completely independent project.

[1] http://www.istat.it/it/strumenti/metodi-e-software/software/regenesees
[2] https://joinup.ec.europa.eu/software/regenesees/description

2. Motivation of the ReGenesees project

The tasks of calibrating survey weights, computing survey estimates, and assessing their precision, constitute a fundamental building block of the production process of official statistics. These are very complex tasks, whose correct execution requires a good knowledge of the underlying statistical theory, full awareness of the adopted sampling plans, and – often – also some insight on the phenomena under investigation. Such skills, which rightfully contribute to define the ideal cultural background of a "good official statistician", cannot now (nor probably will in the future) be entirely superseded by a software, no matter how sophisticated and powerful it may be. However, most NSIs agree in believing that the availability of highly evolved software systems (along with the definition of standard protocols for their optimal usage) is essential to ensure the *accuracy*, the *safety* and the *full reproducibility* of statistical production processes.
In the recent past, this strategic vision drove Istat and many other NSIs to invest in developing in-house softwares dedicated to the estimation phase (think e.g. of GES from Statistics Canada, CLAN from Statistics Sweden, CALMAR and POULPE from INSEE, BASCULA from CBS and GENESEES from Istat). Today, the same strategic vision (even reinforced by the awareness of the ongoing rapid technological change, with its challenges and opportunities) pushes the same NSIs to renew, enrich or even redesign from scratch their software systems. *ReGenesees* is the most recent fruit of the effort made by Istat in this direction.

3. ReGenesees: a paradigm shift

In the design phase of the *ReGenesees* project, it emerged quite soon that it would have been possible to meet the expectations described in Section 2 only trough a radical paradigm shift. As a consequence, *ReGenesees* turns out to be rather different from its SAS ancestor GENESEES (and, by the way, from most of other existing estimation tools) from both the application logic and the user experience standpoints. Indeed, besides allowing to compute estimates and sampling errors for a much wider range of estimators, *ReGenesees* ensures an easier and safer usage and a dramatic reduction in user workload.
In a nutshell (see Sections 3.1 and 3.2 for more on points (1) and (2) below):

(1) User interaction with the new system takes place at a *very high level of abstraction*. *ReGenesees* users, indeed, do not need anymore to preprocess the survey data relying on ad-hoc programs; instead, they only have to feed the software with (i) the data as they are, plus (ii) *symbolic metadata* that describe the adopted sampling design and calibration model[3]. At that point, it is up to the system itself to transform, in an automatic and transparent way, the survey data into the complex data structures required to solve the calibration problem and to compute estimates and errors.

[3] For 'calibration model' we mean the assisting linear regression model underlying a specific calibration problem.

(2) Besides Totals, and Absolute Frequency Distributions (estimators that were already covered by GENESEES), *ReGenesees* allows to compute estimates and sampling errors with respect to Means, Ratios, Multiple Regression Coefficients, Quantiles, and, more generally, with respect to any *Complex Estimator*, provided it can be expressed as a differentiable function of Horvitz-Thompson or Calibration Estimators. It is worth stressing that such Complex Estimators can be defined in a completely free fashion: the user only needs to provide the system with the *symbolic expression* of the estimator as a mathematical function. *ReGenesees*, indeed, is able to automatically linearize such Complex Estimators, so that the estimation of their variance comes at no cost at all to the user.

Existing estimation softwares (Istat traditional SAS system being no exception) generally did not give any support to the users in preparing auxiliary variables and population totals for calibration, nor in deriving the Taylor expansion of non-linear estimators and in computing the corresponding linearized variable for variance estimation purposes. As a consequence, ad-hoc (often very complex) programs for data preparation, transformation and validity check were developed and maintained outside the scope of the estimation system: a time-consuming and error-prone practice. *ReGenesees* frees its users from such needs, with an evident gain in terms of workload reduction, better usability and increased robustness against possible errors.

Interestingly, both the innovative *ReGenesees* features sketched above leverage a peculiar strong point of the R programming language, that is its ability to process *symbolic information*. As a matter of fact, developing the same functionalities in SAS would have been simply impossible: a striking example of what we meant, in Section 2, by 'opportunities of the technological change'.

A technological shift, on the other hand, always involves challenges and some price to pay. The most threatening challenge faced by the *ReGenesees* project has been to demonstrate that an R-based system would have been actually able to manage efficiently the huge amounts of data involved in processing Istat large-scale surveys.

A lot of effort has been invested during the whole development cycle of the new system to meet this challenge. Today, thanks to the empirical evidence and to the reproducible results accumulated during an extensive and thorough testing campaign, we are certain that the challenge has been definitely overcome. Indeed, since its beta release, *ReGenesees* has been successfully tested on both the Labour Force Survey and the Small and Medium Enterprises Survey. When the tasks of calibration and computation of estimates and errors are concerned, these two surveys constitute (each one in its own domain) the most severe test bed available in Istat. Moreover, *ReGenesees* underwent also an independent validation, carried out by colleagues from the UK statistical institute (ONS). A first comparative study, performed on their Life Opportunities Survey, measured *ReGenesees* effectiveness and efficiency using Statistics Canada's GES as a benchmark. The outcome was that *ReGenesees* replicated exactly the results achieved by GES, while ensuring a significant increase in efficiency (in their testing environment, execution times turned out to be halved, on average). This result, in turn, triggered a second ONS initiative: *ReGenesees* was used to calibrate three important surveys[4] for the

[4] Scottish Household Survey, Scottish Health Survey and Scottish Crime and Justice Survey: calibration and estimation for these surveys is currently contracted to three separate external companies.

Scottish Government. Again the results were very satisfactory, and the ONS Methodology Advisory Service suggested *ReGenesees* as a possible "calibration engine" to be adopted in the novel *centralised weighting*[5] framework designed for the Scottish Government.

3.1. Leveraging symbolic information: the Calibration side

Real-world calibration tasks in the field of Official Statistics can simultaneously involve several hundreds of auxiliary variables. Moreover, the construction of such auxiliary variables is in general highly nontrivial, as they need to be carefully derived from the original survey variables according to the (possibly very complex) adopted calibration models. With respect to such operations, traditional calibration facilities (as those listed in Section 2, mostly based on SAS) did not give any practical support to the users, instead devoting dozens of user manual's pages to describe the standard data structures they expected as input. As a consequence, users had to develop customized programs (typically SAS scripts) in order to generate the right input data to feed the calibration system, with 'right' meaning here: (i) appropriate to the survey data and to the calibration task at hand, and (ii) compliant with all the documented rules imposed by the system.
On the contrary, users interact with the *ReGenesees* system at very high level of abstraction, as they only need to specify the calibration model in symbolic way, via R model-formulae[6]: driven by this symbolic information, the system is, indeed, able to transparently generate the right values and formats for the auxiliary variables at the sample level. In addition, *ReGenesees* assists also in defining and calculating the population totals corresponding to the generated auxiliary variables. Indeed, driven again by the calibration model-formula, the system provides the user with a *template* dataset appropriate to store the requested totals. Whenever the actual population totals are available to the user as such, that is in the form of already computed aggregated values (this case typically occurs for household surveys, due to their multistage design), the user has only to fill-in the template. An even bigger benefit is achieved when the sampling frame of the survey is available as a single database table and the actual population totals can be calculated from this source (this is the case of all Istat structural business surveys, whose samples are drawn from ASIA[7]). In such cases, indeed, *ReGenesees* is able to automatically compute the totals of the auxiliary variables from the sampling frame, and to safely arrange and format these values so that they can be directly used for calibration.
The considerations sketched above can be summarized in the following example:

```
sbsdes<-e.svydesign(data=sbs, ids=~id,
        strata=~strata, weights=~weight, fpc=~fpc)
```

1. Persistently bind survey data (sbs) to sampling design metadata;

[5] The Scottish Government Centralised Weighting Project aims at developing *in house* a common framework for calibrating all scottish large-scale population surveys.
[6] These are R objects of class `formula`. Thanks to the flexible syntax of this class and to the powerful semantics of its methods, such objects can be used to specify, in a compact symbolic form, a wide range of statistical models.
[7] The Istat archive of about 4.5 million Italian active enterprises.

```
pop<-pop.template(data=sbsdes,
     calmodel=~((emp.num+ent):(nace2+emp.cl:nace.macro)):region-1)
```

2. Build a population totals template;

```
pop<-fill.template(universe=sbs.frame, template=pop)
```

3. Compute the requested totals from the universe (sbs.frame) and safely fill the template;

```
sbscal<-e.calibrate(design=sbsdes, df.population=pop,
        calfun="linear", bounds=c(0.01,3))
```

4. Perform the calibration task;

The `calmodel` formula in code fragment **2** symbolically specifies a complex calibration problem, with calibration constraints imposed (simultaneously) on the total number of employees (`emp.num`) and enterprises (`ent`) inside domains obtained by (i) crossing `nace2` and `region`, and (ii) crossing `emp.cl`, `nace.macro` and `region`. Given the nature of the `sbs`[8] dataset, such a calibration model translates into 462 different auxiliary variables, whose population totals are transparently computed and arranged in code fragment **3**. Code fragment **4**, first automatically generates the model matrix storing the values of the 462 auxiliary variables for the whole sample, then computes the desired calibrated weights. In conclusion, fragments **1**-**4** show that we passed to *ReGenesees* only the data as they were plus symbolic metadata, without any need of working out the 463 auxiliary variables, nor their population totals.

3.2. Leveraging symbolic information: the Linearization Variance side

The Taylor linearization method is a well-established, approximate tool (Wolter 1985) for estimating the variance of Complex Estimators, namely estimators that can be expressed as non-linear (but smooth) functions of Horvitz-Thompson (HT) estimators of totals:

$$\hat{\theta} = f(\hat{Y}_1,...,\hat{Y}_m) \qquad (1)$$

$$\hat{\theta} \approx \theta + \sum_{j=1}^{m} \left.\frac{\partial f}{\partial \hat{Y}_j}\right|_{\mathbf{Y}} (\hat{Y}_j - Y_j) \doteq \hat{\theta}_{lin} \qquad (2)$$

$$\hat{\theta}_{lin} \approx \sum_{k \in s} d_k \hat{z}_k + const \qquad (3)$$

$$\hat{z}_k = \sum_{j=1}^{m} \left.\frac{\partial f}{\partial \hat{Y}_j}\right|_{\hat{\mathbf{Y}}} y_{jk} \qquad (4)$$

$$\hat{V}(\hat{\theta}) \approx \hat{V}(\sum_{k \in s} d_k \hat{z}_k) \qquad (5)$$

Equation (5) summarizes the "golden rule" of the method. Estimating the variance of the linear approximation of the original Complex Estimator (2) boils down to the much simpler problem of estimating the variance of the HT total of a single artificial variable $\hat{z}$: the so called *linearized variable* (Woodruff 1971) of the Complex Estimator (4).

[8] `sbs` and `sbs.frame` are artificial test datasets provided by the `ReGenesees` package (see Section 5): they represent, respectively, a sample of enterprises and the sampling frame from which the latter has been selected.

The extension to smooth functions of calibration estimators (Särndal 2007) of totals is straightforward (Deville 1999). The "golden rule" (5) still applies, the only relevant change being a different expression for the linearized variable:

$$\hat{\theta} = f(\hat{Y}_1^{CAL},...,\hat{Y}_m^{CAL}) \qquad (6) \qquad\qquad \hat{z}_k = \sum_{j=1}^{m} \frac{\partial f}{\partial \hat{Y}_j^{CAL}}\bigg|_{\hat{\mathbf{Y}}^{CAL}} g_k \hat{e}_{jk} \qquad (7)$$

namely: the value of the original variable y_{jk} has been replaced by the product of the g-weight $g_k = w_k / d_k$ with the estimated *residual* of that variable under the calibration model $\hat{e}_{jk} = y_{jk} - \mathbf{x}_k \cdot \hat{\boldsymbol{\beta}}_j$.

While the mathematical framework outlined by equations (1)-(7) is clear, its software implementation involves some subtle and tricky technical points. For instance, domain estimation of standard errors tends to became cumbersome, especially for complicated functions of calibration estimators. When a *partitioned* calibration is performed (this is by far the computationally most efficient choice for *factorizable* calibration models) the interplay between estimation domains and calibration model reference groups has to be carefully taken into account, as discussed in Estevao, Hidiroglou, and Särndal (1995).

From a software development standpoint, the linearization approach to variance estimation has a fundamental drawback: the Taylor series expansion of a non-linear estimator *does* depend on its functional form *f*. Therefore, being traditional computing environments (e.g. SAS) *unable* to perform *symbolic differentiation*, different programs have to be developed separately for each non-linear statistic *f*. As a direct consequence, traditional systems like those listed in Section 2 suffer two main limits: (i) they support only a quite limited set of non-linear estimators (typically ratios of totals[9]) and (ii) they cannot allow their users to define *their own* Complex Estimators, i.e. statistics which are not built-in. Whenever users of such systems need non-built-in estimators, they have to develop ad-hoc programs to compute the appropriate linearized variables on their own.

ReGenesees overcomes both limits, again leveraging R's ability to process symbolic information. First, we devised a simple syntax[10] for specifying arbitrary Complex Estimators through their functional form, and enabled it by exploiting R methods for manipulating `expression` objects. Then, we used advanced R facilities for calculating symbolic derivatives to develop a sort of "universal" linearization program. Once equipped with it, we were in the position of adding to the system new non-linear estimators almost for free (see Section 4 for a list). Lastly, we engineered our "universal" linearization program, making it friendly and fully visible to the users. The resulting function, named `svystatL`, handles arbitrary user-defined Complex Estimators, as we show below for the geometric mean[11] of `emp.num` on the same data used in Section 3.1:

---

[9] Statistics Sweden's system CLAN is more versatile: it can handle arbitrary rational functions of totals (Andersson and Nordberg 1994). At the other extreme, Istat traditional system GENESEES did not support any non-linear estimator.

[10] The functional form of a complex estimator is specified by an `expression`. Inside it, the estimator of the *total* of a variable is simply represented by the *name* of the variable itself. To represent the estimator of the *mean* of a variable `y`, expression `y/ones` has to be used (`ones` being the convenience name of an artificial variable whose value is `1` for each sampling unit, so that its total estimator estimates the population size).

[11] Recall that the the geometric mean of a non negative variable, say *y*, can be expressed as *exp(mean(log(y)))*.

```
sbsdes<-des.addvars(sbsdes, log.emp.num=log(emp.num))
```

1. Add a new computed variable to the survey data;

```
g<-svystatL(sbsdes, expression(exp(log.emp.num/ones)))
```

2. Estimate the geometric mean and its standard error;

```
print(g)
                          Complex            SE
exp(log.emp.num/ones)     20.51558     0.06077714
```

3. Print on screen the obtained results;

Code fragments **1**-**3** above show that *ReGenesees* was able to estimate the variance of a user-defined Complex Estimator in a completely automated manner, overcoming any need of developing ad-hoc programs.

4. ReGenesees statistical methods in a nutshell

From a statistical point of view, the *ReGenesees* system is very rich and flexible, as it can handle a wide range of sampling designs, calibration models and estimators. A list of the most important methods it implements is reported below:

- **Complex Sampling Designs**
  - Multistage, stratified, clustered, sampling designs
  - Unequally weighted sampling, with or without replacement
  - "Mixed" sampling designs (i.e. with both self-representing and non-self-representing strata)
- **Calibration**
  - Global and partitioned (for factorizable calibration models)
  - Unit-level and cluster-level adjustment
  - Homoscedastic and heteroscedastic models
- **Basic Estimators**
  - Horvitz-Thompson
  - Calibration Estimators
- **Variance Estimation**
  - Multistage formulation (via Bellhouse recursive algorithm)
  - Ultimate Cluster approximation
  - Collapse strata technique for handling lonely PSUs
  - Taylor-linearization for non-linear "smooth" estimators
- **Estimates and Sampling Errors (standard errors, variance, coefficient of variation, confidence interval, design effect) for:**
  - Totals
  - Means
  - Absolute and relative frequency distributions (marginal, conditional and joint)
  - Ratios between totals
  - Multiple regression coefficients
  - Quantiles (variance estimation via the Woodruff method)

- **Estimates and Sampling Errors for Complex Estimators**
  - Handles arbitrary differentiable functions of Horvitz-Thompson or Calibration estimators
  - Complex Estimators can be freely defined by the user
  - Automated Taylor-linearization
  - Design covariance and correlation between Complex Estimators
- **Estimates and Sampling Errors for Subpopulations (Domains)**

It is worth stressing that only a quite limited subset of the statistical methods covered by *ReGenesees* was already available inside its SAS ancestor GENESEES. For instance, the only estimators provided were Totals and Absolute Frequencies, and variance estimation in multistage designs could be tackled only under the Ultimate Cluster approximation.

5. ReGenesees software architecture: a quick overview

The *ReGenesees* system has a clear-cut two-layer architecture. The application layer of the system is embedded into an R package named itself **`ReGenesees`**. A second R package, called **`ReGenesees.GUI`**, implements the presentation layer of the system (namely a Tcl/Tk GUI, see Figure 1 for sample screenshots). Both packages can be run under Windows as well as under Mac and most of the Unix-like operating systems.

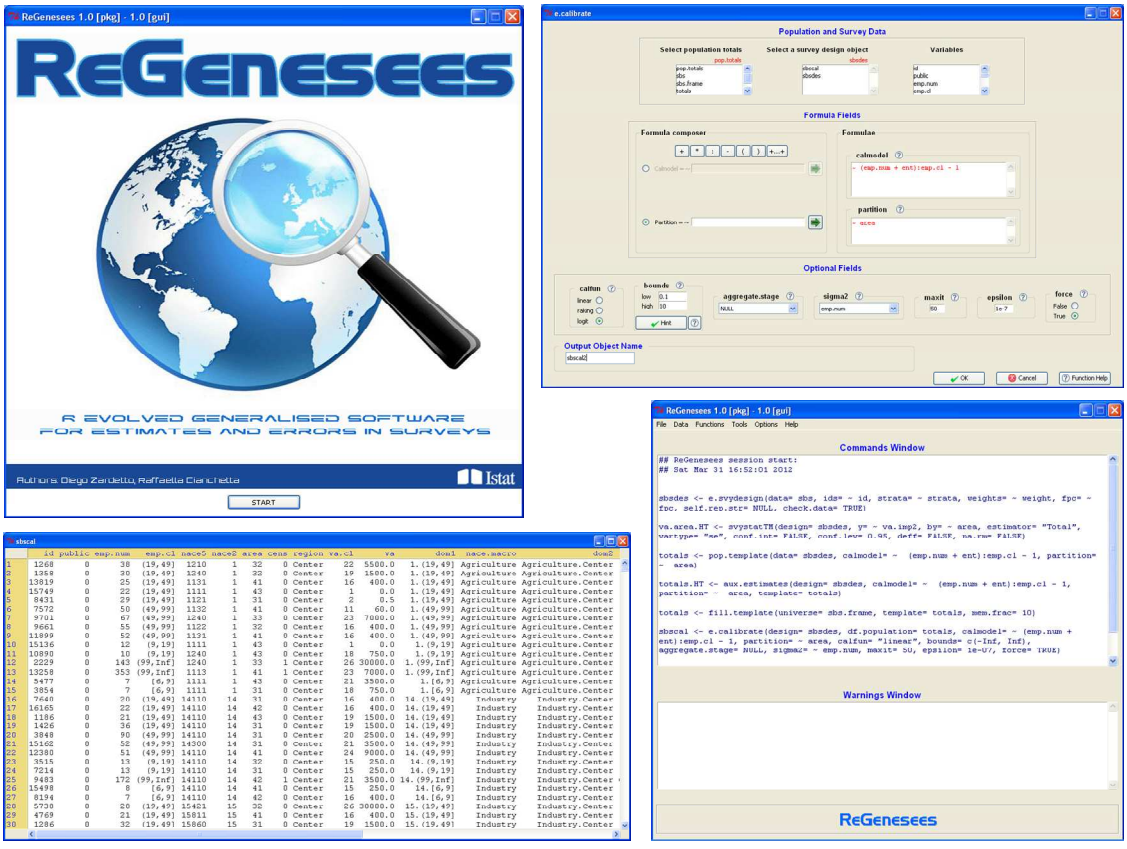

Figure 1: A selection of ReGenesees GUI screenshots

While the `ReGenesees.GUI` package requires the `ReGenesees` package, the latter can be used also without the GUI on its top. This means that the statistical functions of the system will always be accessible by users interacting with R trough the traditional command-line interface (as for code fragments in Section 3.1). On the contrary, less experienced R users will take advantage from the user-friendly mouse-click graphical interface.
It is worth mentioning that, especially in terms of software design principles, the `ReGenesees` package owes a lot to the beautiful, rich and still growing `survey` package written by Thomas Lumley (Lumley 2004). Retrospectively, the original seeds of the *ReGenesees* project can be tracked back to late 2006, when we were trying to optimize `survey` in order to enable its critical functions to successfully process Istat's large-scale surveys. Quite soon, this attempt required us to rethink globally the technical implementation of the package, i.e. its internal structure at a deeper level. Over time, this line of work coagulated into an R package in its own right, with a lot of advanced and useful new features that were not covered by `survey`.

6. Migrating Istat procedures towards ReGenesees

As already said, the first public release of the *ReGenesees* system is quite recent (December 2011). Anyway, the software began to spread in Istat since late 2010, during its beta-testing cycle. Indeed – to date – it has already been used successfully in production by eight Istat large-scale surveys. These include five structural business surveys carried out in compliance with Eurostat regulations: (i) Community Innovation Survey, (ii) Labour Cost Survey, (iii) Access to Finance, (iv) Information and Communication Technology, (v) Continuing Vocational Training Survey, plus three surveys in the social demographic domain: (vi) Time Use, (vii) Employment Perspectives of Young Graduates, (vii) Post Enumeration Survey of the Agricultural Census.
For all the surveys mentioned above, the migration of the standard calibration and estimation procedures from SAS towards *ReGenesees* achieved a significant reduction in both users workload and execution time.

7. Ongoing work and future extensions

Since its beta release, *ReGenesees* has been steadily gaining ground in Istat: to date, as already sketched, it has been successfully integrated in the production workflow of eight large-scale surveys. Moreover, other Istat surveys are migrating towards the new system right now. Internal training courses dedicated to the new R-based system, whose activation is scheduled for this year, will allow an even faster and wider penetration of *ReGenesees* in Istat production processes.
In the meantime, the *ReGenesees* project is in full swing and still growing. *ReGenesees* version 1.3, whose public release took place in mid-November 2012, provides facilities for computing estimates and errors of multiple regression coefficients, as well as for estimating the covariance and correlation of complex estimators (up to previous version 1.2, such methods were indeed kept "hidden" for validation purposes). Furthermore,

facilities implementing the *Generalized Variance Functions* method (Wolter 1985) are currently under development. We are also assessing the feasibility of integrating the EVER package (Zardetto 2012) with *ReGenesees*, thus bringing into the new system the extended DAGJK technique (Kott 2001) for variance estimation: this would be useful whenever the Taylor linearization method can be applied only at the price of crude approximations (e.g. for estimators of at-risk-of-poverty rate and other Laeken indicators). The aforementioned extensions will be included, very likely together with further enrichments, in the next major release of the *ReGenesees* system.

References

Andersson, C., Nordberg, L. (1994), A Method for Variance Estimation of Non-Linear Functions of Totals in Surveys – Theory and Software Implementation. Journal of Official Statistics, 10 (4), 395-405.

Deville, J. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. Survey Methodology, 25 (2), 193-203.

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 (2), 181-204.

Kott, P. S. (2001). The Delete-A-Group Jackknife. Journal of Official Statistics, 17 (4), 521-526.

Lumley, T. (2004). Analysis of Complex Survey Samples. Journal of Statistical Software, 9 (1), 1-19.

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/

Särndal, C.E. (2007). The calibration approach in survey theory and practice. Survey Methodology, 33 (2), 99-119.

Wolter, K.M. (1985). Introduction to variance estimation. New York: Springer.

Woodruff, R.S. (1971). A Simple Method for Approximating the Variance of a Complicated Estimate, Journal of the American Statistical Association, 66 (334), 411-414.

Zardetto, D. (2012). EVER: Estimation of Variance by Efficient Replication. R package version 1.2, Istat, Italy. URL: http://cran.r-project.org/web/packages/EVER/index.html