

# Design-Based and Model-Assisted Analysis of Complex Sample Surveys with R: an Introduction to the ReGenesees Package. - DRAFT -

Diego Zardetto

Istat - Italian National Institute of Statistics

---

## Abstract

This vignette will provide an overview of the ReGenesees project and a concise description of the main features of the ReGenesees package.

*Keywords:* estimation, calibration, sampling variance, R.

---

## 1. Introduction

### 1.1. R at Istat

Over the last five years, the R system for statistical computing and graphics ([R Development Core Team 2010](#)) has been steadily gaining ground at the Italian National Institute of Statistics (Istat), among both the communities of statistical researchers and software engineers. A major boost in that direction followed a 2005 directive issued by the Italian National Centre for Information Technologies in Public Administration (CNIPA), which pushed Istat to start surveying possible Open Source software alternatives, in order to soften its dependence on proprietary technologies. In this respect the SAS system turned out to be the biggest concern, as it had been used, since the early 80s, not only as a tool for statistical data analysis, but largely as an instrument to develop applications for the different phases of the data production process in most surveys. Such a strong dependence of Istat on SAS was recognized as a risk, as the potential unavailability of the proprietary system at a certain time would have prevented the production processes of official statistics from being carried out. These thoughts, as well as the perspective of a significant cost reduction, were the initial driving factors for studying and experimenting new software solutions based on R. Among the first mature results of this activity, we can list packages **EVER** ([Zardetto 2010](#)) and **StatMatch** ([D’Orazio 2009](#)), both available on CRAN, as well as systems **RELAIS** ([Cibella, Fortini, Scannapieco, Tosco, Tuoto, and Valentino 2010](#)) and **MAUSS-R** ([Buglielli and Pagliuca 2010](#)) both using R as a mathematical engine (although at a different extent) under a Java graphical user interface. Moreover, many ad-hoc procedures, formerly developed in SAS, have been successfully migrated toward R.

The **ReGenesees** package (the acronym stands for “R Evolved Generalized Software for Esti-

mates and Errors in Surveys”), to which this article is devoted, is just the latest outcome of this rich stream of work, and we believe it is not going to be the last one.

## 1.2. Project Background

At present, the **ReGenesees** package is the fundamental building block of a larger, still in-process software project, named – once more – **ReGenesees**<sup>1</sup>. The ultimate aim of the project is to release an R-based, full-fledged software system for design-based and model-assisted analysis of complex sample surveys. A clear-cut two-layer architecture has been designed for the **ReGenesees** system: the application layer will be embedded into the **ReGenesees** package, whereas a second R package, named **ReGenesees.GUI**, is planned to implement the presentation layer of the system (namely a Tcl/Tk GUI). Of course, while the **ReGenesees.GUI** package will require the **ReGenesees** package, it will be possible to use the latter also without the GUI on its top. This means that the statistical functions of the system will always be accessible to users interacting with R through the traditional command-line interface. On the contrary, less experienced R users will take advantage from the friendly “mouse-click” graphical interface.

Till the advent of the new R-oriented production stream we mentioned in Section 1.1, the standard phase of calibration, estimation and assessment of sampling errors in sample surveys was covered at Istat by a SAS application named GENESEES (VV. AA. 2005). The name of the **ReGenesees** project has been chosen precisely to emphasize Istat’ seamless offer of software tools dedicated to that phase, while highlighting at the same time its evolution and enhancement through R. It is worth stressing, anyway, that the **ReGenesees** system is *not* a migration toward R of the former SAS application, but rather the fruit of a new and completely independent software project.

At the very beginning of the **ReGenesees** project, we decided to scan the rich offer of R add-on packages (about 1.500 at that time), in order to verify whether any of them was able to satisfy, at least partially, the typical needs of Istat’ sample surveys. The underlying aim was, of course, code reuse. The beautiful **survey** package written by Thomas Lumley (Lumley 2010, 2004) immediately emerged as the best candidate, and we deeply studied and analyzed its functions. By using data from the Labour Force Survey (LFS) as test bed for calibration and variance estimation, we soon realized (see Scannapieco, Zardetto, and Barcaroli 2007) that **survey** could not be adopted at Istat “as it was”. Indeed, e.g. every attempt of invoking its `calibrate()` function on LFS data invariably led to a memory allocation failure, whatever testing environment (i.e. hardware and operating system configuration) we set up. The point was that, despite being anything but naive, **survey** code was not optimized for processing such huge amount of data<sup>2</sup>.

In a first stage, we tried to overcome **survey** limitations by *locally* modifying and extending its critical functions. For a while we obtained encouraging results, also fruitfully cooperating with the author of the package, as documented in Barcaroli, Scannapieco, Vaccari, and Zardetto (2007). Anyway, notwithstanding the valuable efficiency gain achieved till then, it became

---

<sup>1</sup>Different fonts are being intentionally used in order to typographically distinguish the **ReGenesees** *package* from the **ReGenesees** *system*, to which the former belongs.

<sup>2</sup>Just to get an impression: LFS survey data involve, for each quarterly round, about 200.000 sampling units and about 300 variables; moreover, during the calibration phase, about 200 auxiliary variables inside 21 regional domains are processed, yielding about 4.000 constraint equations to be satisfied in order to match the corresponding known population totals.

clear quite soon that code optimization could not be the solution we were looking for. Indeed, as will be clarified by the discussion we are going to sketch in Appendix ??, enabling **survey** to successfully process Istat data would have required to re-think *globally* the package design, that is its internal structure at a deeper level. Since such a radical remodeling of the **survey** package turned out to fall definitely outside the scope of the author, we decided to start developing a new R package by ourselves. The **ReGenesees** package is the final result of this effort. Moreover, it has to be stressed that, besides the fundamental strong point of being able to successfully handle calibration, estimation and sampling errors assessment for all Istat' large-scale surveys, the **ReGenesees** package also provides a lot of advanced and useful new features that were not covered by **survey**.

## References

- Barcaroli G, Scannapieco M, Vaccari C, Zardetto D (2007). "Migrating a Critical Application toward Open Source: the Istat Experience." In *Proceedings of the 1st International Conference on Methodologies, Technologies and Tools Enabling e Government (MeTTeG 07)*.
- Buglielli MT, Pagliuca D (2010). "MAUSS-R: Multivariate Allocation of Units in Sampling Surveys." Version 0.9.2, URL <http://www.osor.eu/projects/mauss-r>.
- Cibella N, Fortini M, Scannapieco M, Tosco L, Tuoto T, Valentino L (2010). "RELAIS: REcord Linkage At IStat." Version 2.1, URL <http://www.osor.eu/projects/relais>.
- D'Orazio M (2009). "StatMatch: Statistical Matching." R package version 0.8, URL <http://cran.at.r-project.org/web/packages/StatMatch/index.html>.
- Lumley T (2004). "Analysis of Complex Survey Samples." *Journal of Statistical Software*, **9**(1), 1–19. R package version 2.2.
- Lumley T (2010). "survey: analysis of complex survey samples." R package version 3.22, URL <http://cran.at.r-project.org/web/packages/survey/index.html>.
- R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.
- Scannapieco M, Zardetto D, Barcaroli G (2007). "La Calibrazione dei Dati con R: una Sperimentazione sull'Indagine Forze di Lavoro ed un Confronto con GENESEES/SAS." *Collana Contributi Istat*, **4**. In italian.
- VV AA (2005). "GENESEES." Version 3.0, URL [http://www.istat.it/strumenti/metodi/software/produzione\\_stime/genesees/](http://www.istat.it/strumenti/metodi/software/produzione_stime/genesees/).
- Zardetto D (2010). "EVER: Estimation of Variance by Efficient Replication." R package version 1.1.1, URL <http://cran.at.r-project.org/web/packages/EVER/index.html>.

**Affiliation:**

Diego Zardetto

Methodology and Software Division

Istat - Italian National Institute of Statistics

Via Cesare Balbo 16

00184 Rome, Italy

E-mail: [Zardetto@istat.it](mailto:Zardetto@istat.it)