Use of R in Business Surveys at the Italian National Institute of Statistics: Experiences and Perspectives

Giulio Barcaroli[1], Diego Zardetto[2]
[1, 2]Istat, Via Cesare Balbo 16, Roma, Italia – email: {[1]barcarol, [2]zardetto}@istat.it

Abstract
Over the last six years, R has been steadily gaining ground in Istat, since a strategic decision to limit dependence on proprietary technologies (like SAS) was taken. A migration activity of our critical IT tools from SAS to R was carried out (we can cite *MAUSS-R* for optimal sample allocation, and *ReGenesees* for the calculation of estimates and sampling errors), and new R packages were developed (e.g. *SeleMix* for selective editing). In particular, *ReGenesees* has been recently experimented on a number of structural business surveys, namely "Information and Communication Technology in Enterprises", "Community Innovation Survey", "Access to Finance" and "Labour Cost". Moreover, in the "Survey on Agricultural Production Prices", various ad-hoc procedures (editing and imputation, calculation of indices), formerly developed in proprietary technologies, have been successfully migrated toward R, in a complex architecture where data reside in an Oracle database and the overall application is web based. A successful test activity has also been carried out for data retrieval and processing from the register of active enterprises (4.5 million units), thus showing that R limitations in handling huge datasets can be overcome.

Key Words: R software, business surveys, selective editing, calibration, estimation, sampling variance

1. Introduction

Since 2006, the R system for statistical computing has been steadily gaining ground at the Italian National Institute of Statistics (Istat), among both the communities of statistical researchers and IT experts. A major boost in that direction was triggered by the strategic decision to start surveying possible open source software alternatives, in order to soften Istat dependence on proprietary technologies (like, e.g., SAS). Such dependence was undesirable not only for related costs, but also for portability and interoperability issues, as both qualities can be guaranteed only by open solutions. Among possible alternatives, quite soon R was identified as the most promising candidate, in particular because of the synergy ensured by the vast community of R developers.

Istat strategy for fostering the usage of R encompassed several different phases. We started with a broad-spectrum study of R, both as a programming language and as a technological environment. Being the outcome of such analysis encouraging, we soon turned to investigate the technical feasibility of a migration towards R of some selected critical Istat applications (e.g. the software system for calibration, estimation and sampling errors assessment). Once such feasibility was established, a rich stream of migration works followed. In the meanwhile, an important activity of training was carried out, so that – up to now – about 300 statistical researchers and IT experts have been given

a basic training on R. This, in turn, stimulated also the birth of several brand new "R-centric" software projects. These projects dealt mainly with the migration of generalised software, formerly developed by using SAS (MAUSS for sampling design and GENESEES for calibration and sampling variance), to new versions making use of R. Also new IT tools have been developed entirely in R, or by making use of R packages (respectively, *SeleMix* and *RELAIS*).

This paper is organised as follows. In paragraph 2 we give a general overview of all the IT tools (packages or generalised software) that, so far, have been developed using R inside Istat, alongside a concise description of methods and applications. In paragraph 3 and 4 we give a more detailed description of two R packages, *SeleMix* and *ReGenesees*, respectively for selective editing and for calibration and estimation, together with related applications to Istat business surveys. In paragraph 5 we report an important experience of use of R as a development technology for the implementation of production procedures in the survey on Agricultural Production Prices: it revealed the possibility to handle data organised in an Oracle database, in a web application. Finally, in paragraph 6 we give hints on future work.

2. R packages and R based tools developed in Istat

A numerous set of R packages and R based tools have been developed in Istat since 2007. Some of them have been placed on the CRAN[1] (in particular, in the Task View "Official Statistics"), while others are available on JOINUP[2], the Open Source Repository of the European Commission, and/or in the Istat official site (ISTAT[3]). They are reported in the next table, together with the indication of the GSBPM[4] sub-process they can be referred to, the main functions they offer, and the websites from where they are downloadable.

Table 1: R packages and R-based tools developed in Istat

Software	GSBPM sub-process	Main Functions	Online Repository
MAUSS-R	2.4 Design Frame and Sample Methodology	Design of Stratified Samples	JOINUP, ISTAT
SamplingStrata	2.4 Design Frame and Sample Methodology 4.1 Select sample	Optimisation of Frame Stratification and Sample Selection	CRAN
RELAIS	5.1 Integrate Data	Record Linkage	JOINUP, ISTAT
StatMatch	5.1 Integrate Data	Statistical Matching	CRAN
SeleMix	5.3 Review, Validate and Edit	Selective Editing	CRAN, JOINUP

[1] http://cran.r-project.org
[2] http://joinup.ec.europa.eu
[3] http://www.istat.it/it/strumenti/metodi-e-software/software
[4] http://www1.unece.org/stat/platform/download/attachments/8683538/GSBPM+Final.pdf?version=1

Software	GSBPM sub-process	Main Functions	Online Repository
EVER	5.6 Calculate Weights 6.2 Validate Outputs	Calibration, Estimation and Sampling Variance Assessment	CRAN, ISTAT
ReGenesees, ReGenesees.GUI	5.6 Calculate Weights 6.2 Validate Outputs	Calibration, Estimation and Sampling Variance Assessment	JOINUP, ISTAT

In the following, we give a concise description of each of them.

Software MAUSS-R
MAUSS-R (Multivariate Allocation of Units in Sampling Surveys – R version) is a tool for defining the sampling design for sample surveys on finite populations, in case of stratified samples with one stage (Buglielli *et al*, 2010[a]), which is the usual design for enterprises surveys. It implements the Bethel algorithm (Bethel, 1989) for multivariate allocation, and extends it to the multidomain case. This means that the user (the statistician planning a survey) can define, for each target estimate, the desired accuracy (expressed as a constraint on its expected coefficient of variation), differentiated by domain of interest. The system offers a solution (i.e. the required sample size together with the allocation of units in the frame strata) that can be acceptable or not, in terms of allowable budget; the user may vary the input parameters; since the system memorises each solution, at the end the user can choose the one that is the best compromise between quality and cost. *MAUSS-R* has been implemented in R (the optimisation algorithm) and Java (the user interface). Thanks to this Java interface, the user does not need to know R language or R GUI to interact with the system. In any case, the Java interface makes use of a package ("mauss"), whose functions can be used independently in R sessions. A new version of this package is being developed in order to cope with the case of two-stage sampling.

Package SamplingStrata
In the field of sampling design (in particular for stratified sampling), this package offers an approach for the determination of the best stratification of a sampling frame, the one that ensures the minimum sample size under the condition to satisfy precision constraints in a multivariate and multidomain case (Barcaroli, 2012). This approach is based on the use of the genetic algorithm: each solution (i.e. a particular partition in strata of the sampling frame) is considered as an individual in a population; the fitness of all individuals is evaluated by calculating (using the Bethel-Chromy algorithm) the sampling size satisfying accuracy constraints on the target estimates. Functions in the package allows to: (a) analyse the obtained results of the optimisation step; (b) assign the new strata labels to the sampling frame; (c) select a sample from the new frame accordingly to the best allocation. There is also a function that allows to build the most important input to the optimisation step, i.e. the ''strata'' dataframe, containing information (means and standard errors) regarding the distributions of the target variables in the different strata, using the sampling frame or using data from previous rounds of the same survey.

Software RELAIS

RELAIS (Record Linkage At Istat) is a toolkit for Record Linkage (Cibella *et al*, 2010). *RELAIS* allows combining techniques for each of the record linkage phases, so that the resulting workflow is actually built on the basis of the requirements of the application at hand. More specifically, the *RELAIS* toolkit is composed by a collection of techniques for each record linkage phase that can be dynamically combined in order to build the best record linkage workflow. *RELAIS* has been implemented in Java and R and has a database architecture (MySQL). Specifically, the estimation phase (EM) for the Fellegi-Sunter probabilistic decision model has been implemented in R as the 1:1 reduction phase that exploits the LP-solve algorithm. The other techniques and GUIs are implemented in Java.

Package StatMatch

StatMatch (D'Orazio 2012) provides some R functions to perform statistical matching, i.e. the integration of two data sources referred to the same target population which share a number of common variables. Some functions can also be used to impute missing values in data sets through hot deck imputation methods. Methods to perform statistical matching when dealing with data from complex sample surveys (via weights calibration) are available too.

Package SeleMix

SeleMix (Selective Editing via Mixture models) is an R package for selective editing (Guarnera, Buglielli, 2011). It includes functions for identification of outliers and influential errors in numerical data. For each unit, it provides also anticipated values (predictions) for both observed and non observed variables. The method is based on explicitly modelling both true (error-free) data and error mechanism. Specifically, true data are supposed to follow normal or log-normal distribution. It is assumed that only a subset of data is affected by error and that the error mechanism is specified through a Gaussian random variable with zero mean vector and covariance matrix proportional to the covariance matrix characterising the true data distribution. A more detailed description of the package, together with its applications, is given in paragraph 3.

Package EVER

EVER (Estimation of Variance by Efficient Replication) is mainly intended for calculating estimates and standard errors in complex surveys (Zardetto, 2012[a]). Variance estimation is based on the extended DAGJK (Delete-A-group Jackknife) technique proposed by Kott. The advantage of the DAGJK method over the traditional jackknife is that, unlike the latter, it remains computationally manageable even when dealing with "complex and big" surveys (tens of thousands of PSUs arranged in a large number of strata with widely varying sizes). In fact, the DAGJK method is known to provide, for a broad range of sampling designs and estimators, (near) unbiased standard error estimates even with a "small" number (e.g. a few tens) of replicate weights. Besides its peculiar computational efficiency, the DAGJK method takes advantage of the strong points it shares with the most common replication methods. As a remarkable example, *EVER* is designed to fully exploit DAGJK's versatility: the package provides the user with a user-friendly tool for calculating estimates, standard errors and confidence intervals for estimators defined by the user themselves (even non-analytic). This functionality makes *EVER* especially appealing whenever variance estimation by Taylor linearisation can be applied only at the price of crude approximations (e.g. poverty estimates).

ReGenesees System

ReGenesees (R Evolved Generalised Software for Sampling Estimates and Errors in Surveys) is a full-fledged R system for design-based and model-assisted analysis of complex sampling surveys (Zardetto, 2012[b]). It handles multistage, stratified, clustered, unequally weighted survey designs. Sampling variance estimation for nonlinear (smooth) estimators is obtained by Taylor-series linearization. Sampling variance estimation for multistage designs can be obtained both under the Ultimate Cluster approximation or by means of an actual multistage computation. Estimates, standard errors, confidence intervals and design effects are provided for: Totals, Means, Absolute and relative Frequency Distributions (marginal or joint), Ratios and Quantiles (variance via the Woodruff method). *ReGenesees* also handles Complex Estimators, i.e. any user-defined estimator that can be expressed as an analytic function of Horvitz-Thompson or Calibration estimators of Totals or Means, by automatically linearising them. All analyses above can be carried out for arbitrary subpopulations. A more detailed description of the system and its applications is given in paragraph 4.

3. SeleMix and its application to Small and Medium Enterprises Survey

SeleMix is an R package that allows to apply a particular method of selective editing, based on *contamination normal models* (Buglielli *et al*, 2010[b]). We set the following conditions:

1. Two sets of variables are observed: q *X*-variables (always correctly measured) and p *Z*-variables (affected by measurement errors), with the special case where no *X*-variables are available;
2. Observed data are characterized by a mixture of distributions, one related to true data, the other related to erroneous data;
3. True (or *non contaminated*) data are represented by a $n \times p$ matrix Z^* of independent realizations of a random p-vector assumed to follow a log-normal distribution whose parameters may depend on the q covariates (*X*-variables). Having set $Y^* = \log(Z^*)$, we can define the regression model:

$$Y^* = XB + U \quad ; \qquad U \sim N(\cdot; 0, \Sigma) \tag{1}$$

4. Erroneous (*contaminated*) data are characterized by an intermittent error mechanism: in other words, we assume the presence of a Bernoulli r.v. I (I=1 if an error occurs, I=0 otherwise) with parameter π being the "a priori" contamination probability. An additional assumption is that errors affect data through an additive mechanism represented by a Gaussian r.v. with zero mean and covariance matrix Σ_ε, proportional to Σ:

$$Y = Y^* + \varepsilon \quad ; \qquad \Sigma_\varepsilon = (\alpha - 1)\Sigma, \ \alpha > 1 \tag{2}$$

So, the error model can be represented by the conditional distribution:

$$f_{Y|Y^*}(y \mid y^*) = (1-\pi)\delta(y - y^*) + \pi N(y; y^*, \Sigma_\varepsilon) \tag{3}$$

where $\delta(t'-t)$ is the delta-function with mass at t.

Under these assumptions, the distribution of observed data can be derived by multiplying the normal density of true data expressed by (1) and the error density expressed by (3), and integrating over Y^*, obtaining:

$$f_{Y|Y^*}(y \mid y^*) = (1-\pi)N(y; B'x, \Sigma) + \pi N(y; B'x, \alpha\Sigma) \tag{4}$$

Parameters of (4) can be estimated by maximizing the likelihood based on n sample units by means of an ECM algorithm (Meng, Rubin, 1993).

It is possible to derive the distribution of true data Y^* conditional on observed data by applying the Bayes formula:

$$f_{Y^*|Y,X}(y^* \mid y,x) = \tau_1(x,y)\delta(y^* - y) + \tau_2(x,y)N(y^*; \tilde{\mu}_{x,y}, \tilde{\Sigma}) \qquad (5)$$

where τ_1 and τ_2 are the posterior probabilities of belonging to true and erroneous data, while:

$$\tilde{\mu}_{x,y} = \frac{y + (\alpha - 1)B'x}{\alpha}; \qquad \tilde{\Sigma} = (1 - \frac{1}{\alpha})\Sigma$$

We can derive the corresponding conditional distribution in the original scale:

$$f_{Z^*|Z}(z^* \mid z) = \tau_1(\ln(z))\delta(z^* - z) + \tau_2(\ln(z))LN(z^*; \tilde{\mu}_{x,\ln(z)}, \tilde{\Sigma}) \qquad (6)$$

that can be estimated by replacing the values of the parameters with the estimates of the $(\mu, \Sigma, \pi, \alpha)$ obtained by the ECM algorithm.

The estimated distribution (6) allows to predict true values z_i for all observations *i=1, ...,n*:

$$\hat{z}_i = E(z_i^* \mid z_i) = \int z_i^* f_{Z^*|Z}(z^* \mid z) dz_i^* \qquad (7)$$

and consequently the expected error:

$$\varepsilon_i = \hat{z}_i - z_i \qquad (8)$$

Let us now suppose that a target estimate is the total T_Z of a variable *Z*, whose estimator is $\hat{T}_Z = \sum_{i \in S} w_i z_i$.

A robust version of this estimator is $\hat{T}_Z^* = \sum_{i \in S} w_i \hat{z}_i$, where the estimated $\hat{z}_i$ have been obtained by using (7). Actually, we are not interested in using this robust estimator, as it is too much dependent on the model assumptions; rather, we use it inside a score function useful to prioritize editing. The score function is the following:

$$SF_i = | r_i | = | \frac{w_i \varepsilon_i}{\hat{T}_Z^*} | \qquad (9)$$

Finally, we introduce the absolute value of the approximated expected residual percentage error in data after removing errors in the units belonging in a subset M of the sample as

$$R_M = | \sum_{i \in M} r_i | \qquad (10)$$

At this point we can define the overall selective editing procedure as the sequence of the following steps:

1. define an accuracy threshold η;
2. order the observations in descending order according to the values of the score function;
3. select the first *k* units for reviewing, so that $\bar{k} = \min\{ k \in (1,...,n) \mid R_{M_j} < \eta, \forall j > k \}$

In order to evaluate this method, a subset of the 2006 *Small and Medium Enterprises* (*SME*) survey data (about 2,650 observations) have been considered. As target *Z* variable, affected by errors, variable *Turnover* has been chosen, while the variables *Labour Cost* and *Number of Employees* have been considered as *X* covariates. Original data *Y** have been contaminated with probability $\pi = 0.05$ by adding a Gaussian disturbance with distribution $N(0, (\alpha - 1)\Sigma)$, where $(\alpha - 1) = 10$ and Σ has been estimated on clean data. The relative error of the total estimate of *Turnover* without editing is about 120%.

Together with the Contaminated Model (CM) method, three other methods have been considered in order to compare results:

1. selective editing based on standard linear regression (LM);
2. selective editing based on robust linear regression (RLM);
3. selective editing based on robust linear regression using clean historical data (HIST).

A Montecarlo simulation has been carried out by replicating the application of the four different methods 1000 times. In each iteration, in units selected for the review the value of *Turnover* has been replaced with the corresponding known true value.
The prefixed threshold was $\eta = 0.02$: this means that at the end of the editing, the expected relative error on target estimate will not exceed 2%. In the following table the results of the application of the four different methods are reported.

Table 2: Results of the simulation

Method	Selected units	Relative bias (%)	Relative RMSE(%)
CM	23.3	1.66	1.90
HIST	20.4	2.46	2.82
RLM	1270.0	-0.09	0.14
LM	1538.6	0.02	0.10

If we analyse these results, we can notice that:

(i) methods based on regression (standard or robust) show a slightly better performance in terms of quality (lower values of relative bias and of relative root mean square error), but at the price of unaffordable cost in terms of reviewed units (in the case of LM 58% of the sample);

(ii) CM and HIST perform almost the same, but the second makes use of historical data in order to determine the cut-off, while the first operates only on current data (with considerable less information); furthermore, CM is fully compliant with the prefixed threshold on the expected relative error (2%), while HIST is not.

We can therefore conclude that in this setting the CM method outperforms with respect to traditional methods based on linear regression, and is comparable to the method that makes use of historical data (which are not required by CM method).

Istat has recently established a working group with the task of assessing the applicability of selective editing methods implemented in R package *SeleMix* under certain specific operational contexts. Among the group's activities, particularly relevant is the analysis of costs and benefits deriving from the use of the tool *SeleMix* in the process of editing data from business surveys. In some experiments, which are based on the availability of both raw and edited data, we consider the latter as "clean" and we plan to compare, in terms of costs of reviewing, editing procedure currently used in production with an alternative procedure that includes the use of *SeleMix*. In practice we try to estimate the savings in resources, currently dedicated to the more expensive activities such as manual review of the questionnaire, follow-up, etc., that the use of the instrument would allow to achieve, without affecting significantly the quality of the estimates. Particular attention is paid to the experiments carried out on incomplete data. In fact these experiments would allow not only to test Selective Editing procedures in the presence of non-response, but also to evaluate the potential of *SeleMix* as a tool for the (robust) imputation of continuous variables. Moreover, also the analysis of the potential of *SeleMix* concerning its capability to exploit the information from external sources is being carried out. This last task is particularly important because it fits into the broader field of research, of great interest to the Institute, which focuses on the use of information from administrative records in the statistical production process.
At present, the group has already carried out some experiments on the 2008 survey on Small and Medium Enterprises (SME 2008) and on the 2008 survey on the use of

Information and Communication Technologies in enterprises (ICT). In both cases we have examined some structural economic variables (Sales, Cost of Labour, Number of Employees, etc.), and experiments were conducted that also use information from administrative sources (Archive of Budgets, Education Sector). External sources have been used as sources of auxiliary information to be included in the model of contamination, and also as elements of comparison to be used for the evaluation of the estimates based on survey data.
Another set of experiments are being conducted on data artificially perturbed in order to evaluate the robustness of the method with respect to departures from the assumptions at the basis of the model that explains the mechanism of error.

Finally, *SeleMix* is going to be submitted to the procedure to become a standard inside Istat. This procedure, recently defined, involves different steps of evaluation and approval so that at the end a method or tool can be considered a mandatory standard. The evaluation is carried out by the "Network of Methodologists", while the approval is delegated to governing bodies of the Institute. Full implementation of the procedure should be completed by the end of 2013.

4. ReGenesees and its usage in Structural Business Statistics

As anticipated in paragraph 2, *ReGenesees* is a full-fledged R system for design-based and model-assisted analysis of complex sampling surveys. In this paragraph, we provide basic background information on the *ReGenesees* project, along with an overview of the main IT features and statistical functionalities of the system. Moreover, we report concisely on Istat Structural Business Statistics surveys that successfully migrated toward the usage of *ReGenesees* so far, highlighting the benefits of such a migration from a user perspective. Lastly, we sketch the encouraging results we obtained by using *ReGenesees* for data retrieval and processing from ASIA (the Istat archive of about 4.5 million Italian active enterprises).

From a historical perspective, the phase of calibration, estimation and assessment of sampling errors has been the first test-bed we adopted to investigate the technical feasibility of a large-scale migration towards R of Istat production processes. At the very beginning, we decided to scan the rich offer of R add-on packages, 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 *survey* package, written by Thomas Lumley (Lumley, 2004), immediately emerged as the best candidate, and we deeply studied and analyzed its functions. Anyway, by using data from the Italian Labour Force Survey (LFS) as test-case, we soon realized that *survey* could not be adopted at Istat "as it was". Indeed, every attempt of exploiting its calibration or variance estimation facilities 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. 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. Anyway, notwithstanding the valuable efficiency gain achieved till then, it became clear quite soon that code optimization could not be the solution we were looking for. Indeed, 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 remodelling 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* system 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* system also provides a lot of advanced and useful new features that were not covered by *survey*.

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* (Zardetto, 2012[b]). A second R package, called *ReGenesees.GUI* (Cianchetta, Zardetto, 2012), implements the presentation layer of the system (namely a Tcl/Tk GUI, see Figure 1 below for sample screenshots). Both packages can be run under Windows as well as under most of the Unix-like operating systems. 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. On the contrary, less experienced R users will take advantage from the user-friendly mouse-click graphical interface.

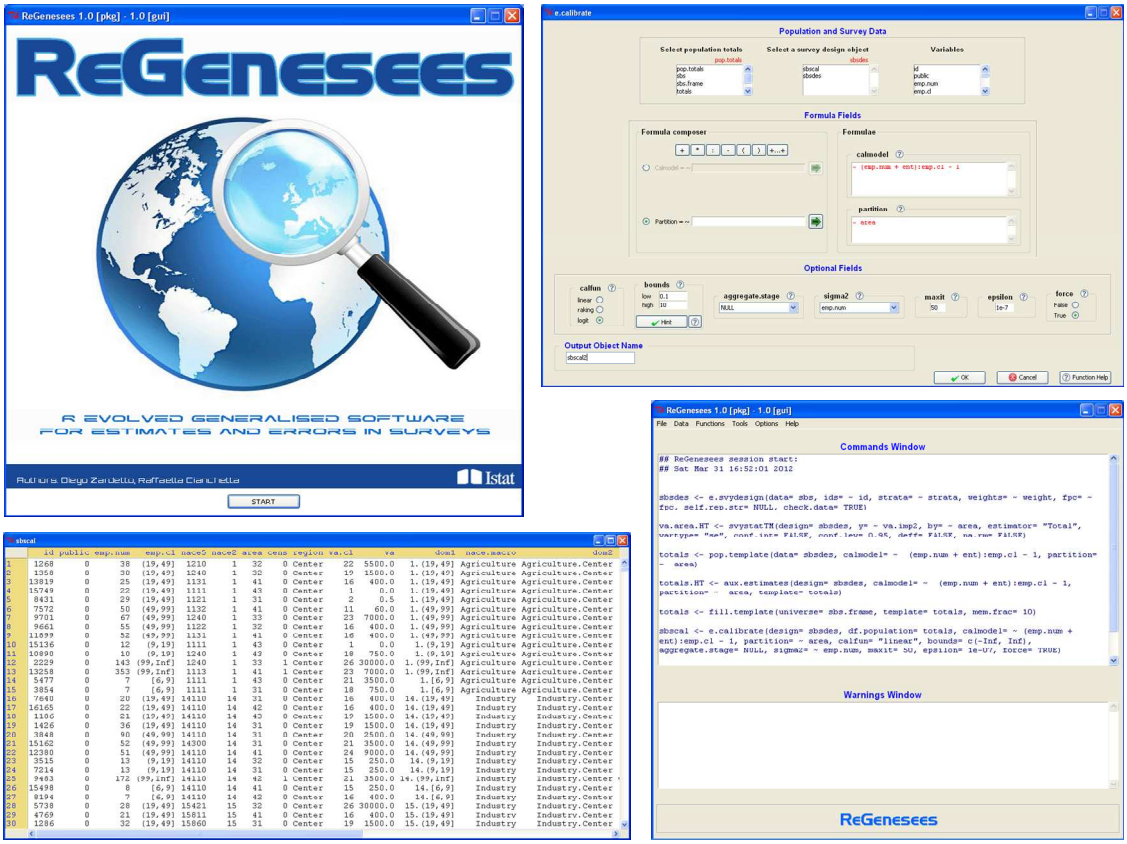

Figure 1: A selection of ReGenesees GUI screenshots

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 in the Box below:

- **Complex Sampling Designs**
  - Multistage, stratified, clustered, sampling designs
  - Unequally weighted sampling, with or without replacement
  - "Mixed" sampling designs (i.e. with both SelfRepresenting and NonSelfRepresenting strata)
- **Calibration**
  - Global and/or partitioned (for factorizable calibration models)
  - Unit-level and/or cluster-level adjustment
  - Homoscedastic and/or Heteroscedastic models
- **Basic Estimators**
  - Horvitz-Thompson
  - Calibration Estimators
- **Sampling Variance Estimation**
  - Multistage formulation (via Bellhouse recursive algorithm)
  - Ultimate-Cluster approximation
  - Taylor-linearization for nonlinear "smooth" estimators
- **Estimates and Sampling Errors (standard erros, variance, coefficient of variation, confidence interval, design effect) for:**
  - Totals
  - Means
  - Absolute and/or relative frequency distributions (marginal and/or joint)
  - Ratios between totals
  - 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
- **Estimates and Sampling Errors for Subpopulations (Domains)**

Box 1: ReGenesees main statistical functions

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 Means, and variance estimation in multistage designs could be tackled only under the Ultimate-Cluster approximation.

The first public release of the *ReGenesees* system for general availability is quite recent, as it dates back to December 2011. Anyway, the software began to spread in Istat since late 2010 during its beta-testing cycle. As a consequence, *ReGenesees* has already been used successfully in production by a small subset of current Istat surveys, including four 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". Basic information on the aforementioned surveys is reported in what follows.

Community Innovation Survey – 2006-2008. The Italian Community Innovation Survey is part of the EU Community Innovation Survey (CIS), carried out on a two-year basis by all European Member States and candidate countries. It covers innovation activities of the Italian enterprises with at least ten employees operating in industry and services, and collects information about new or significantly improved goods or services (product innovations) and new or significantly improved processes, logistics or distribution methods (process innovations). Enterprises with 10-249 employees are sampled, whereas those with at least 250 employees are censused. The sampling design is one-stage stratified simple random sampling, with strata defined by crossing economic activity (NACE), enterprise size and geographical region (NUTS1). The theoretical sample size is about 40,000 units, with a roughly 50% response rate, while the target population size is about 209,000 enterprises. The sample is drawn from the ASIA archive, which is also used to compute population totals to be used as benchmark values for weights calibration.

Labour Cost Survey – 2008. The Labour Cost Survey (LCS) is conducted every four years in all EU Member States on enterprises with at least 10 employees belonging to the sections of NACE Rev. 2 from B to S excluding O. It measures, in both the public and the private sectors, the level and structure of the total expenditure borne for the purpose of employing staff. Labour cost includes compensation of employees (wages and salaries and employers social contributions), vocational training costs, and other expenditures such as recruitment costs and other costs (related to employing labour) paid by the employers. The Italian LCS 2008 results from an integration approach based on Istat statistical survey data, administrative files and tax data. Private enterprises with 10-249 employees are sampled, whereas those with at least 250 employees and those in the public sector are exhaustively observed. The sampling design is one-stage stratified simple random sampling. The sample is drawn by the ASIA archive, from which calibration totals on auxiliary variables are obtained. The planned sample is about 25,000 units (with an approximate response rate of 64%), representing a target population of about 216,000 units.

Access to Finance – 2010. The Eurostat survey on Access to Finance was conducted in 2010 and comprised a sample of about 25,000 enterprises in the 20 participating countries, including Italy. The survey covers small and medium-sized enterprises in terms of employment (with 10 to 249 persons employed). It is a one-off survey, not intended to be repeated and specifically designed to study the reaction of the European enterprises to the financial crises started in 2008. Indeed, information is collected for two significant observation moments: 2007 (considered as a reference point before the crisis) and 2010 (considered a year signalling the end of the financial crisis, at least in some Member States). Information were collected about the perception of finance needs, envisaged finance types and sources, purpose of the finance and potential obstacles to business growth. The Italian survey planned sample was of about 13,000 enterprises, representative of a target population of nearly 107,000. The response rate turned out to be quite low, less than 39% (despite data were collected by online questionnaires).

Information Communication Technology – 2010-2011. The annual ICT survey collects data on the usage of information and communication technology, the internet, e-government, e-business and e-commerce in enterprises. The target population covers all the active enterprises with at least 10 employees. Enterprises with size 10-249 are sampled, whereas those with size 250 or more are all observed. The sampling design is one-stage stratified simple random sampling, with strata defined by crossing economic activity (NACE), enterprise size and geographical region (NUTS1). The theoretical

sample size is about 35,000 units, with a roughly 55% response rate, while the target population amounts to about 212,000 enterprises. The sample is drawn from the ASIA archive, from which calibration totals on auxiliary variables are also obtained.

For all SBS surveys mentioned above, the migration of the standard calibration and estimation procedures from SAS toward *ReGenesees* achieved a significant reduction in both users workload and execution time. In the opinion of survey statisticians involved in the migration, the most relevant advantages have been the following:

(i) *ReGenesees* automates the creation of auxiliary variables on which to calibrate;
(ii) *ReGenesees* assists and drives the users in defining and calculating the corresponding population known totals (thus eliminating the risk of specification errors, especially high in the case of a big number of totals);
(iii) *ReGenesees* prevents the need of developing ad-hoc programs for the estimation of the sampling variance of non-linear estimators (e.g. estimators of ratios between totals).

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, Istat traditional SAS calibration facilities did not gave any support to the users. As a consequence, *ad hoc* SAS scripts for data preparation, transformation and validity check were developed and maintained outside the scope of the calibration system: a time consuming and error prone practice. 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: 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. An even bigger benefit is achieved when computing auxiliary variables totals from external sources, in particular when the sampling frame of the survey is available as a single DB table (this is exactly the case for all Istat SBS surveys whose samples are drawn from ASIA). 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.

Due to well-known R memory limitations and weaknesses in handling huge amounts of data, a lot of effort has been devoted to make as efficient as possible the *ReGenesees* functions performing the above mentioned operations. We were aware that, for sampling frames of several million units and calibration models with several hundred variables, the naive aggregation of the calibration model matrix could be too memory demanding (at least in ordinary PC environments) and determine a memory failure error. Thus we decided to code inside *ReGenesees* the following efficient algorithm: (i) split the sampling frame table in chunks, (ii) generate auxiliary variables based on chunk data, (iii) compute partial sums of auxiliary variables chunk-by-chunk, (iv) update the current auxiliary variables totals by adding progressively such partial sums. Notice that this alternative algorithm is triggered only when it is actually needed. Indeed, *ReGenesees* estimates the memory that would be used to store the full model matrix of the target population and compares it to the maximum memory allocable on the machine. If the latter is not enough, the memory efficient algorithm starts and automatically determines the optimal number of chunks; this happens in such a way that the memory needed to store the partial model matrix of each chunk does not exceed 10% of the maximum

allocable memory. The practical feasibility of this *ReGenesees* function has been experimentally demonstrated in a dedicated test case mimicking the standard calibration process for the Italian Small and Medium Enterprises survey. Specifically, while running *ReGenesees* in an ordinary PC environment, we succeeded in handling about 4,500,000 ASIA records representing the SME target population, and correctly computed the requested calibration known totals.

5. A Case Study: the Agricultural Production Prices information system

In previous paragraphs, a summary description of Istat R-based generalised software tools has been given, along with some illustrative examples of their application to selected Istat business surveys. This paragraph is devoted, instead, to the discussion of an interesting case study in which R has been successfully exploited as a computation engine in a complex and heterogeneous information system.

The survey on Agricultural Production Prices computes two different series of indices: (i) prices of agricultural products sold by farmers; (ii) prices of goods purchased by farmers as means for agricultural production. The purpose of the price indices is to provide information on trends in producer prices of agricultural products and purchase prices of the means of agricultural production. They are intended to permit a comparison of these trends both between the various EU Member States, as well as between different products within a single Member State. In Italy, about 200 (100) distinct products purchased (sold) by farmers are monitored, and the Chamber of Commerce deliver on a monthly basis to Istat more than 3,300 (3,700) price observations for such goods.

Recently, the information system of the Istat survey on Agricultural Production Prices underwent an in depth redesign. Former procedures, mainly developed in SAS but also partly still relying on legacy COBOL routines, were definitely discarded and substituted by a brand new web application (see Figure 2 below). The architecture is a basic 3-tier one: the data storage layer is an Oracle RDBMS, the application layer (deployed in a Windows server OS environment) integrates JSP/Tomcat dynamic Web content technologies with dedicated R modules for statistical computing, and the presentation layer is a Web browser. The application allows remote operators to monitor ongoing procedures, to visualize microdata, statistics and graphics, to edit and impute missing data both interactively and automatically, to calculate micro and aggregated price indices, and to produce quality reports. In particular, R modules, running as batch jobs, are in charge of performing nearest-neighbour missing values imputation, calculating microindices and deriving price indices for disparate product categories and territorial domains. Database connectivity is ensured by using the RODBC package, while transaction integrity for concurrent elaborations is obtained by spawning a distinct R-thread for each operator request.

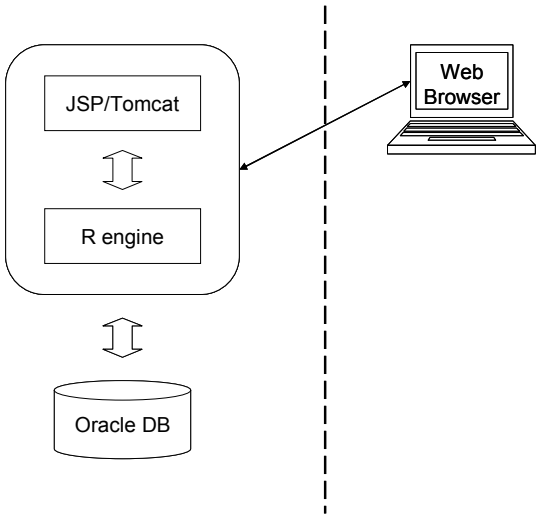

Figure 2: The 3-tier Agricultural Production Prices web application

The lessons learned in developing and maintaining an R-based statistical-core inside the Agricultural Production Prices web application are encouraging. Indeed, on the one hand, R expressive power allows to code even very complex algorithms in extremely compact modules, on the other hand, R portability and native DB connectivity facilities make it possible to deploy such compact modules in (almost) whatever technological environment.

6. Future work and conclusions

A great effort has been dedicated to the introduction of R in Istat as a valid alternative to the proprietary statistical system previously adopted (SAS).
Important results have been obtained: the R culture began to spread inside Istat, and an important internal production of systems and packages developed in R and in other open source technologies allowed to substitute all generalized softwares that were previously implemented in SAS. In economic terms, this made Istat gain a relevant reduction in SAS annual fee (from about € 1,000,000 in 2003 to nearly € 400,000 in 2013, with no reduction in terms of server and client licenses): this would not have been possible without an investment towards alternatives to SAS.
But we cannot hide the fact that, despite the effort in training, R is still far from becoming the standard tool for processing and analyzing data in production sectors. This is due to its well known steep learning curve, and to the fact that SAS has been being used since early 80's, and many people still prefer using it for basic operations on data.
Despite this, we are confident that the use of R will spread more and more. Newcomers from universities are trained to R; the offer of new packages makes R more and more convenient for the variety of solutions it offers; and huge market companies like Google and Oracle invest on R as *the* statistical tool for their applications and environments.
We would like to cite in particular:

1. the Google application interfaces to R, that enable R users to exploit many Google applications;
2. the new Oracle solution for analytics, based on Oracle R Enterprise.

With regard to the first item, Istat is now starting an investigation (in cooperation with University "La Sapienza" of Rome) regarding the possibility to make use of the data on the web, in order to strengthen official statistics. R already offers many packages offering *web scraping* facilities, which will be investigated.
As for Oracle R Enterprise, its very recent release offers a complete integration of R in Oracle databases, thus allowing one solution to the problem of huge data handling which was the well known Achille's heel of R. Istat is going to experiment this solution, and one of the possible areas in which to do it is the one related to Enterprises Frame and Censuses. Apart from Oracle, also the open source Apache Hadoop solution for handling big data will be taken into consideration and experimented.
In any case, it is no more a question of individual choice that a single National Institute can make independently from others. The general trend is towards an integration and sharing of IT tools and software, and fundamental requirements are their portability and interoperability. These requirements exclude proprietary software, and, among open source statistical softwares, R has no competitors.

References

Barcaroli G. (2012). "SamplingStrata: Optimal stratification of sampling frames for multipurpose sampling surveys". R package version 0.9-3, Istat.
http://CRAN.R-project.org/package=SamplingStrata

Bethel J. (1989). "Sample Allocation in Multivariate Survey". Survey Methodology, 15(1): 47-57

Buglielli M.T., De Vitiis C., Barcaroli G. (2010[a]). "MAUSS-R. Multivariate Allocation of Units in Sampling Surveys", Istat.
http://www.istat.it/it/files/2011/02/user_and_methodological_manual.pdf

Buglielli M.T., Di Zio M., Guarnera U., (2010[b]). "Use of Contamination Models for Selective Editing", Q2010, European Conference on Quality in Survey Statistics, 4-6 May 2010, Helsinki.

Cianchetta R., Zardetto D. (2012). "ReGenesees.GUI: a TclTk Interface for the ReGenesees Package". R package version 1.2, Istat.
http://www.istat.it/it/files/2011/12/ReGenesees.GUI_.pdf

Cibella N., Fortini M., Scannapieco M., Tosco L., Tuoto T., Valentino E. (2010). "RELAIS User's Guide", Version 2.2, Istat.
http://www.istat.it/it/files/2011/03/Relais2.2UserGuide.pdf

D'Orazio M. (2012). "StatMatch: Statistical Matching". R package version 1.1.0, Istat.
http://CRAN.R-project.org/package=StatMatch

Guarnera U., Buglielli M.T. (2011). "SeleMix: Selective Editing via Mixture models". R package version 0.8.1, Istat.
http://CRAN.R-project.org/package=SeleMix

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

Meng X.L., Rubin D.B. (1993). "Maximum Likelihood Estimation via the ECM Algorithm: a General Framework". Biometrika, Vol. 80, 267-278.

Zardetto D. (2012[a]) "EVER: Estimation of Variance by Efficient Replication". R package version 1.2, Istat.
http://CRAN.R-project.org/package=EVER

Zardetto D. (2012[b]). "ReGenesees: R Evolved Generalised Software for Sampling Estimates and Errors in Surveys". R package version 1.2, Istat.
http://www.istat.it/it/files/2011/12/ReGenesees.pdf