The Italian National Institute for Statistics (Istat) migration from SAS to R: an Overview

A. Durio and A. Lo Presti

Department of Economics “S. Cognetti de Martiis”, University of Turin, Italy

Third Torino R net meeting

Torino, 19 April 2012

Perché questa presentazione.

A febbraio 2012 Pietro T. ci scrive: “Guardate a pagina 17 dell’allegato bollettino dell’ISTAT, una applicazione che mi pare non banale di R.”

SOFTWARE STATISTICO

IL NUOVO SOFTWARE BASATO SU R PER IL CALCOLO DELLE STIME E DEGLI ERRORI CAMPIONARI

di Diego Zardetto (zardetto@istat.it)

Nel mese di dicembre 2011 l’Istat ha reso disponibile la versione 1.0 del sistema ReGenesees (*R evolved Generalised software for sampling estimates and errors in surveys*). Il software può essere scaricato dal sito web dell’Istat e dal *repository* per il software *open source* della Commissione europea JOINUP.

CORRETTEZZA, SICUREZZA E RIPRODUCIBILITÀ

Le attività di calibrazione dei pesi campionari, il calcolo delle stime di interesse e la valutazione dei relativi errori di campionamento costituiscono uno snodo cruciale del processo di produzione dell’informazione statistica ufficiale.
Si tratta di attività complesse, la cui corretta esecuzione esige una buo-

stime e degli errori campionari. La medesima visione strategica spinge, oggi, gli stessi istituti a rinnovare ed arricchire la loro offerta di sistemi software. Il sistema ReGenesees è il frutto più recente dello sforzo profuso dall’Istat in questa direzione.

REGENESEES IN UN GUSCIO DI NOCE

ReGenesees è un sistema software basato su R per l’analisi *design-based* e *model-assisted* di indagini campionarie complesse. Il nome del sistema intende enfatizzare la continuità dell’offerta di strumenti Istat per la calibrazione ed il calcolo delle stime e degli errori (di qui il richiamo al suo predecessore SAS: Genesees) sottolineando, al contempo, l’evoluzione e l’arricchimento di tale

simbolico il disegno di campionamento ed il modello di calibrazione; è il sistema stesso a trasformare i dati di indagine nelle complesse strutture necessarie a risolvere il problema di calibrazione e a calcolare stime ed errori; oltre a totali, medie e frequenze, ReGenesees consente di calcolare stime ed errori campionari rispetto a quantili, rapporti, e, più in generale, rispetto a qualsiasi stimatore complesso, purché esprimibile come funzione differenziabile di stimatori di Horvitz-Thompson o di Calibrazione. Si noti che tali stimatori complessi possono essere definiti in modo del tutto libero: l’utente deve solo fornire al sistema l’espressione simbolica dello stimatore come funzione matematica. ReGenesees, infatti, è in grado di linearizzare automaticamente gli stimatori complessi, consentendone così la stima della varianza campionaria a costo zero per l’utente.
Entrambe le innovazioni fanno leva su un punto di forza specifico del linguaggio R: la sua capacità di

Agenda

- Come e perché R all'ISTAT

Agenda

- Come e perché R all'ISTAT
- Rassegna dei software sviluppati in ambiente R dall'ISTAT

Agenda

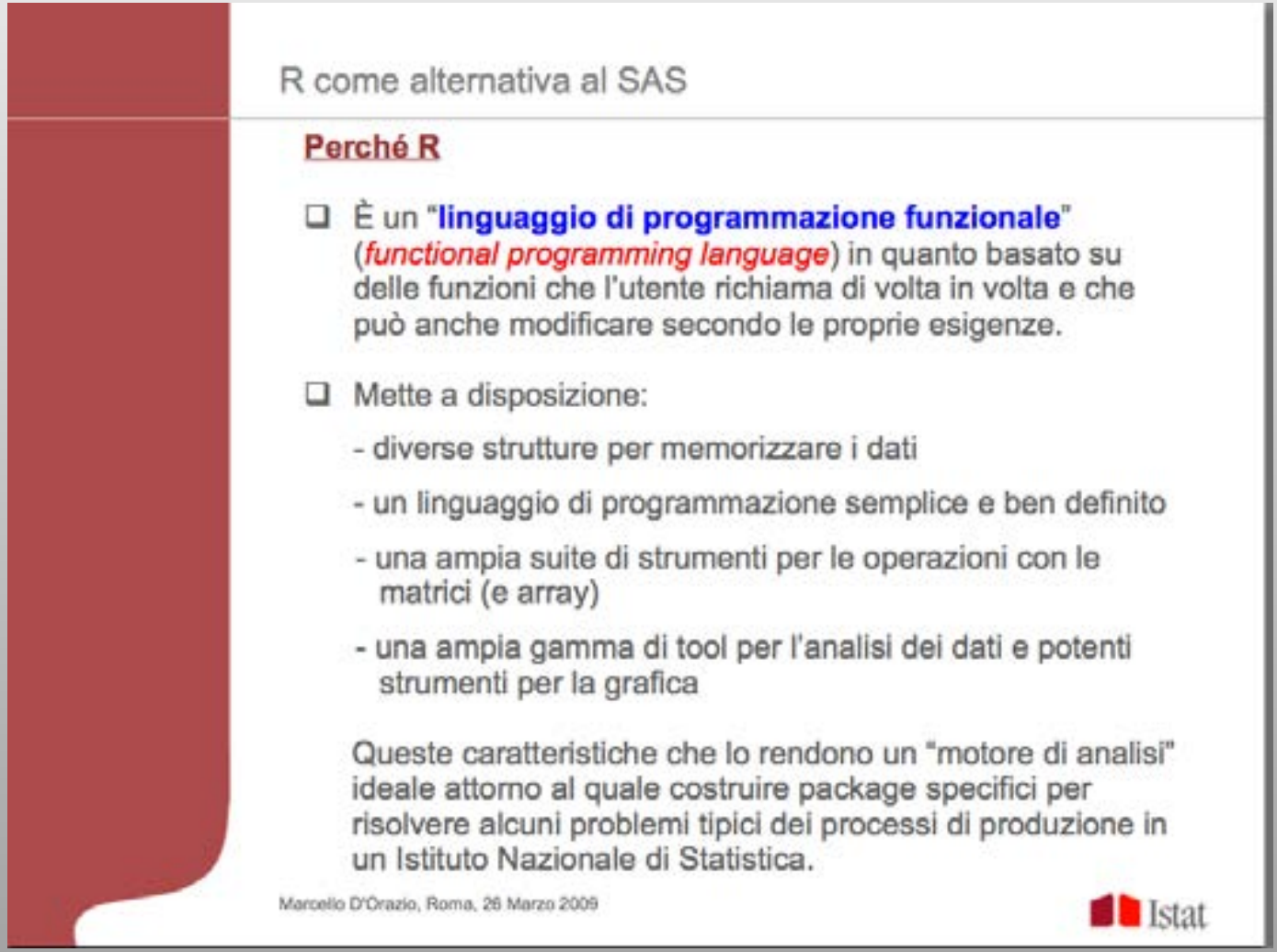
- Come e perché R all'ISTAT
- Rassegna dei software sviluppati in ambiente R dall'ISTAT
- Un esempio di funzionalità dell'ultimo nato: ReGenesees

R all'Istat

- nei primi anni del nuovo millennio il passaggio da SAS ad R ha avuto una forte accelerazione su pressione dell'Authority per l'informatica nella P.A. principalmente con finalità di risparmio.
- già nel 2004, a causa della indisponibilità di procedure SAS adeguate a risolvere specifici problemi, i programmatori dell'istat hanno iniziato a creare packages R.
- al primo seminario sul passaggio da SAS a R la software-house vi ha fatto partecipare il responsabile delle vendite ma, nonostante le offerte di sconto sulle licenze, la spesa di circa $1,000,000$ di euro del 2003 si ridurrà nel 2013 a $400,000$ euro.
- attualmente, a causa delle diverse visioni nel nuovo management dell'istat il processo di passaggio da SAS a R rischia di rallentare anche a fronte della probabile offerta da parte di SAS della vendita dell software che ha sempre dato (dagli anni 70) solo in abbonamento.

Perchè R

Marcello D'Orazio (Ricercatore Istat) nel 2009 illustrava così i motivi della scelta di R quale sofware alternativo a SAS per le esigenze dell'Istituto Nazionale di Statistica



R come alternativa al SAS

Perché R

- È un "**linguaggio di programmazione funzionale**" (*functional programming language*) in quanto basato su delle funzioni che l'utente richiama di volta in volta e che può anche modificare secondo le proprie esigenze.
- Mette a disposizione:
  - diverse strutture per memorizzare i dati
  - un linguaggio di programmazione semplice e ben definito
  - una ampia suite di strumenti per le operazioni con le matrici (e array)
  - una ampia gamma di tool per l'analisi dei dati e potenti strumenti per la grafica

Queste caratteristiche che lo rendono un "motore di analisi" ideale attorno al quale costruire package specifici per risolvere alcuni problemi tipici dei processi di produzione in un Istituto Nazionale di Statistica.

Marcello D'Orazio, Roma, 26 Marzo 2009

Istat

Package utilizzati – (i)

Prima di implementare packages propri, i ricercatori dell'istat hanno individuato una serie di utili strumenti giá disponibili tra le molteplici offerte di R:

Package utilizzati – (ii)

La cooperazione con l'autore di *survey*, per cercare di ottimizzarlo nell'ottica della gestione di grandi moli di dati, ha dato origine al progetto ReGenesees

Formazione del Personale.

FORMAZIONE		
2004	1 corso	*R-base (5gg)*
2006	2 corso	*R-base (5gg)*
2007	2 corso	*R-base (5gg)*
2008	5 corsi	*3 R-base (5gg) + 2 R-programmazione (3gg)*
2009-oggi	2-3 corsi a trimestre	*R-base e/o 2 R-programmazione*

- docenti interni (gruppo informale di esperti)
- ad oggi sono stati formati circa 300 ricercatori
- dal 2008 non viene piú fatta alcuna formazione SAS

Software R sviluppato in Istat

SOFTWARE	FUNZIONI	SITO
MAUSS-R	Piani di campionamento stratificati	JOINUP, ISTAT
SamplingStrata	Ottimizzazione di stratificazioni Selezione del campione	CRAN
RELAIS	Record Linkage	JOINUP, ISTAT
StatMatch	Statistical Matching	CRAN
SeleMix	Editing Selettivo	CRAN, JOINUP
EVER	Stima e valutazione della varianza campionaria	CRAN, ISTAT
ReGenesees	Calibrazione Stima e valutazione della varianza campionaria	ISTAT, JOINUP

- JOINUP: `https://joinup.ec.europa.eu/`
- ISTAT: `http://www.istat.it/it/strumenti/metodi-e-software`
- CRAN: ... chi lo scorda? .... `http://cran.r-project.org/`

MAUSS-R

- Multivariate Allocations of Units in Sampling Surveys: è un sistema sofware per la determinazione dell'allocazione campionaria nel caso multivariato e per più domini di stima per le indagini ad uno stadio di campionamento.
- Implementa un'estensione del metodo di allocazione di Neyman al caso di più variabili e adotta come metodo di risoluzione una generalizzazione della proposta di Bethel (1989) già implementata in MAUSS (sas).
- Per ciascuna delle variabili target è possibile definire la precisione della stima desiderata (in termini di CV) per ogni dominio. Richiede in input prestime di medie e varianza degli strati e restituisce la numerosità campionaria per ciascuno strato, fornisce confronti con altri criteri di allocazione e misura la sensitività.
- è interamente sviluppato in codice R viene rilasciato per ambienti Windows compilato con una interfaccia grafica Java. Una nuova versione è prevista per l'estate.

SamplingStrata

- SamplingStrata: è un pacchetto che consente di determinare la migliore stratificazione per un piano di campionamento, quella che assicura la dimensione minima degli strati garantendo i vincoli di precisione definiti in casi multivariati e multidominio
- L'ottimizzazione è basata sull'uso di un algoritmo genetico e l'allocazione viene fatta secondo l'algorimo di Bethel-Chromy.
- le funzioni del pacchetto consentono di: analizzare i risultati in ogni passo del processo di ottimizzazione, aggiungere strati nel piano di campionamento, selezionare il campione secondo la stratificazione ottima. Nel pacchetto c' è una funzione che consente di creare l'importante data.frame "strata" contenente medie ed errori delle variabil target nei vari strati.
- si tratta di un CRAN package *Author:Giulio Barcaroli* prima versione *29-Apr-2011* ultima *2012-01-24*

- REcord Linkage At IStat: è un sistema sofware con finalità di identificare la stessa entità del mondo reale, che può essere rappresentata da differenti unità in fonti di dati diverse, anche in assenza di un identificativo univoco delle unità.
- per ciascuna delle fasi di Record Linkage (preparazione dei dati, selezione delle variabili di matching, riduzione dello spazio di ricerca delle coppie di abbinamento, ecc...) RELAIS mette a disposizione alcuni tra i metodi più diffusi in letteratura.
- implementato in R interfaccia JAVA e e ha una architettura di database relazionale di ambiente MySQL.

- è un pacchetto che consente l'integrazione statistica di due data-base riferiti alla stessa popolazione che contengono un certo numero di variabili comuni.
- Alcune funzioni consentono l'imputazione dei valori mancanti con metodo del donatore di minima distanza mentre altre forniscono metodi basati sulla calibrazione dei pesi per il matching di data-set di indagini statistiche complesse.
- si tratta di un CRAN package *Author:Marcello D'Orazio* prima versione *10-Oct-2008* ultima *2012-04-04*.

- Selective Editing via Mixture Modelling: è un pacchetto per l'identificazione di outlier e/o errori di editing di dati numerici.
- Il metodo comporta l'assunzione di modelli sia per la distribuzione dei dati non affetti da errore che quella che genera gli errori. Per i dati reali (non affetti da errore) si suppone una Normale (o lognormale) che à la prima componente della mistura, per il sottoinsieme di dati che si ritengono affetti da errori si suppone una distribuzione Normale (la seconda componente della mistura) con media zero e matrice di covarianza proporzionale a quella dei dati reali.
- si tratta di un CRAN package *Author:Ugo Guarnera, M. Teresa Buglielli* prima e ultima versione *2011-09-21*.

- Estimating of Variance by Efficient Replications: è un pacchetto per calcolare stime ed errori standard in indagini campionarie complesse.
- La stima della varianza si basa sul metodo DAGJK (Delete-A-Group Jackknife) che particolarmente utile in caso di indagini complesse e di grande numerosità.
- Il package fornisce la stima degli errori standard e gli intervalli di confidenza per stimatori definiti dall'utente stesso non necessariamente in forma analitica (non espressi come funzioni di valori direttamente osservabili sulle unità statistiche).
- si tratta di un CRAN package *Author:Diego Zardetto* prima *12-Sep-2008* ultima versione *2012-04-10*.

ReGENESEES

- Evolved Generalised Software for Sampling Estimates and Errors in Surveys: è un sistema software per l'analisi design-based e model-assisted di indagini campionarie complesse.
- Non è una semplice migrazione da GENESEES (SAS), risponde ad un radicale cambiamento di logica applicativa che, oltre a consentire un più agevole e sicuro utilizzo del software, garantisce un notevole ampliamento nella scelta di stimatori rispetto ai quali calcolare le stime e gli errori campionari.
- è interamente sviluppato in codice R può essere utilizzato in ambienti Windows e Unix-like. Si articola in due package:
  - ReGenesees $--> $ tutte le funzionalità statistiche (utilizzabile da linea di comando R)
  - ReGenesees.GUI $--> $ interfaccia utente di tipo grafico basata su Tcl/Tk (carica automaticamente ReGenesees all'atto dell'installazione)

ReGENESEES: funzionalità

- Disegni di campionamento complessi
  Disegni a più stadi, stratificati a cluster ; Probabilità di inclusione variabili; Disegni di campionamenti misti
- Calibrazione
  Globale e/o per partizioni; A livello di unità e/o di cluster ;
- Stimatori
  Horvitz-Thompson; Calibration estimator
- Stima della varianza campionaria
  Formulazione multistadio (algoritmo di Bellhouse); Ultimate cluster approximation; GENESEES-like per disegni misti; Linearizzazione di Taylor per stimatori non lineari “smooth”
- Stima ed errori campionari
  Totali; Medie; Distribuzioni di Frequenze; Rapporti fra totali; Quantili
- Stima ed errori per stimatori complessi
  Funzioni differenziabili arbitrarie di stimatori di Horvitz-Thompson; Definibili liberamente dall’utente; Linearizzazione di Taylor automatica
- Stima ed errori campionari per sottopopolazioni

ReGENESEES: un esempio