Istat
Istituto Nazionale
di Statistica

Metodi e strumenti IT per la produzione statistica

Contenuti dell'area web
Metodi e strumenti IT per la produzione statistica

Per inviare informazioni e/o aggiornamenti
su metodi e strumenti IT scrivere a softgen@istat.it

Ultimo aggiornamento: **Giugno 2017**

INDICE

Metodi e strumenti IT per la produzione statistica

Il Repository dei metodi e strumenti IT per la produzione statistica contiene:

- **metodi statistici**
- **strumenti IT generalizzati**

validati e utilizzati nei processi di produzione delle informazioni statistiche dell'Istat.

Il Repository è organizzato per **fasi del processo produttivo statistico**, in conformità al Generic Statistical Business Process Model (GSBPM) Version 5.0.

FASI DEL PROCESSO

PROGETTAZIONE	METODI	STRUMENTI
RACCOLTA	METODI	STRUMENTI
ELABORAZIONE	METODI	STRUMENTI
ANALISI	METODI	STRUMENTI

Le fasi del processo produttivo rappresentano le "porte" di accesso alle informazioni e ai materiali specifici. Sono prese in considerazione solo le fasi Progettazione, Raccolta, Elaborazione, Analisi ossia le fasi per le quali sono attualmente disponibili metodi e strumenti.

Le pagine delle singole FASI contengono gli elementi informativi più importanti sui sottoprocessi interni alla fase selezionata.

I metodi e gli strumenti di una data fase sono raggruppati per sottoprocessi. I metodi e gli strumenti pertinenti a più di un sottoprocesso sono attribuiti al sottoprocesso prevalente.

Le pagine relative ai METODI contengono elenchi di documenti che descrivono metodi statistici. I documenti sono standard per l'Istat (o per il Sistema statistico europeo), oppure manuali metodologici, pubblicazioni e relazioni redatti da ricercatori Istat (anche in collaborazione con esterni). Per un dato sottoprocesso, i documenti sono visualizzati in ordine cronologico inverso (dal più recente al meno recente) e sono resi disponibili in formato PDF o tramite collegamento a siti esterni.

Le pagine relative agli STRUMENTI contengono elenchi di strumenti IT generalizzati che implementano metodi statistici. Per un dato sottoprocesso, gli strumenti sono elencati in ordine alfabetico. Per ciascuno strumento è fornita una descrizione generale che esplicita le funzionalità implementate e una scheda informativa.

Degli strumenti sviluppati dall'Istat è possibile effettuare il **download**, previa indicazione di un indirizzo e-mail. Per gli strumenti non di proprietà Istat vengono indicate le modalità di reperimento.

Ultimi aggiornamenti

- StatMatch – Pubblicata la versione 1.2.5 del pacchetto R che rende disponibili alcune funzioni per l'integrazione dei dati attraverso lo statistical matching e, come prodotto secondario, la possibilità di imputare i valori mancanti in un data set | **16 giugno 2017**
- ReGenesees – Pubblicata la versione 1.9 del software basato su R per l'analisi design-based e model-assisted di indagini campionarie complesse. La nuova funzione **'trimcal'** consente di effettuare il trimming dei pesi calibrati preservando, al tempo stesso, tutti i vincoli di calibrazione | **8 maggio 2017**
- COMIC – Pubblicata la versione 1.0 del software basato su SAS per la costruzione di indici compositi, attraverso vari metodi di sintesi, e la valutazione della loro robustezza | **15 marzo 2017**

Fase PROGETTAZIONE

Progettazione della lista e del campione

Le attività relative alla progettazione della lista e della metodologia di campionamento fanno riferimento al sottoprocesso 2.4 "*Design frame and sample*" del GSBPM. Più precisamente:

- Costruzione dell'archivio di selezione relativo alla popolazione obiettivo e contenente, per ciascuna unità della popolazione, le informazioni identificative e necessarie per il contatto, eventuali variabili ausiliarie utili per la definizione del disegno (variabili di stratificazione, variabili identificative degli eventuali stadi di selezione);
- Progettazione del disegno di campionamento che, sulla base degli obiettivi conoscitivi specificati nella Fase 1 "Specify Needs" e dei vincoli operativi e di costo, consenta di ottenere stime il più possibile precise.

Progettazione della lista di selezione

Le caratteristiche della lista di campionamento sono rilevanti per la corretta definizione del disegno di campionamento.

È necessario che la lista risponda a criteri di qualità in termini di aggiornamento, copertura e accuratezza delle informazioni in essa riportate. Da un punto di vista strettamente teorico la lista di selezione ideale dovrebbe possedere i seguenti requisiti:

- essere costituita dalle sole unità appartenenti alla popolazione di interesse al momento di riferimento dell'indagine;
- includere ogni unità della popolazione una sola volta;
- contenere dati aggiornati e corretti relativamente alle informazioni identificative (nome e indirizzo) e alle eventuali informazioni descrittive (altri dati strutturali importanti) delle unità.

Le possibili situazioni di allontanamento dalla lista ideale sono:

- *sottocopertura*, nel caso in cui alcuni elementi della popolazione non sono contenuti nella lista e non possono, pertanto, essere inclusi nel campione;
- *sovracopertura*, quando alcuni elementi della lista sono inesistenti e/o non appartengono alla popolazione di interesse;
- *duplicazione di alcune unità*, se alcuni elementi della popolazione sono presenti più volte nella lista;
- *grappoli di unità*, quando alcuni elementi della lista contengono grappoli di elementi della popolazione.

Progettazione e realizzazione del disegno campionario

La progettazione del disegno campionario prevede innanzitutto la seguente attività:

- **definizione dello schema di campionamento** sulla base dei costi legati alla tecnica di rilevazione prescelta e delle informazioni contenute nella lista di selezione

(campionamento a più stadi di selezione, campionamento stratificato). La scelta di un disegno a più stadi di selezione deriva generalmente dalla necessità di concentrare il campione sul territorio al fine di contenere i costi di rilevazione in caso di indagini che prevedono una modalità diretta di somministrazione del questionario (intervista faccia a faccia). La scelta di un disegno campionario stratificato ha come finalità il miglioramento della precisione delle stime. La suddivisione delle unità della popolazione in strati è effettuata sulla base di variabili ausiliarie presenti sulla lista e legate alle variabili oggetto di indagine.

Sulla base dello schema di campionamento adottato, possono essere previste le seguenti fasi:

- **scelta dei criteri di stratificazione** (scelta delle variabili, scelta del numero degli strati, definizione del criterio di formazione degli strati);
- **scelta del metodo probabilistico di selezione delle unità campionarie** (selezione con probabilità uguali, selezione con probabilità variabili). In caso di disegni a due o più stadi la selezione delle unità di primo stadio è effettuata generalmente con probabilità proporzionale ad una misura di ampiezza supposta correlata con le variabili oggetto di indagine.
- **determinazione delle numerosità campionarie** per i diversi stadi di selezione e **allocazione del campione** tra gli strati sulla base dell'errore di campionamento ammesso per le principali stime in relazione ai domini di riferimento e alle sottoclassi di popolazione. Poiché solitamente le indagini sono progettate per la produzione di molteplici stime per diversi domini di interesse, è necessario utilizzare metodologie che affrontino in un'ottica globale il problema della determinazione della dimensione campionaria ottima in presenza di un numero elevato di obiettivi e di vincoli.

Fase PROGETTAZIONE – METODI

- **PROGETTAZIONE DELLA LISTA E DEL CAMPIONE**

Generalized Framework for Defining the Optimal Inclusion Probabilities of One-Stage Sampling Designs for Multivariate and Multi-domain Surveys
2015
Survey Methodology, 41.

MEMOBUST – Handbook on Methodology of Modern Business Statistics:
Statistical Registers and Frames
Sample Selection
2014
MEETS ESSnet MEMOBUST

Riferimenti

Istat. 2008. Strategia di campionamento e precisione delle stime. In "*L'indagine europea sui redditi e le condizioni di vita delle famiglie (Eu-silc)*", Collana Metodi e Norme, n. 37, Istat.

Istat. 2006. Il disegno campionario della nuova indagine e la fase di estrazione. In "*La rilevazione sulle forze di lavoro: contenuti, metodologie, organizzazione*", Collana Metodi e Norme, n. 32, Istat.

Istat. 2006. Il piano di campionamento. In "*Il sistema di indagini sociali multiscopo. Contenuti e metodologia delle indagini*", Collana Metodi e Norme, n. 31, Istat.

Istat. 2006. Strategia di campionamento e livello di precisione delle stime. In "*L'indagine campionaria sulle nascite: obiettivi, metodologia e organizzazione*", Collana Metodi e Norme, n. 28, Istat.

Cicchitelli G., Herzel A., Montanari G.E. 1992. *Il campionamento statistico.* Il Mulino, Bologna.

Särndal C.E., Swensson B., Wretman J. 1992. *Model Assisted Survey Sampling.* Springer-Verlag, New York.

Bethel J. 1989. *Sample Allocation in Multivariate Surveys*. *Survey Methodology* 15:47-57.

Chromy J. 1987. *Design Optimization with Multiple Objectives*. Proceedings of the Survey Research Methods Section*, American Statistical Association, 194-199*.

Cochran W.G. 1977. *Sampling Techniques.* J. Wiley & Sons, New York.

Fase PROGETTAZIONE – STRUMENTI

- **PROGETTAZIONE DELLA LISTA E DEL CAMPIONE**

FS4 (First Stage Stratification and Selection in Sampling). Software generalizzato per la stratificazione e selezione delle unità di primo stadio in disegni di campionamento a due o più stadi, realizzato completamente in linguaggio R e dotato di un'interfaccia utente di tipo grafico (GUI).

MAUSS-R (Multivariate Allocation of Units in Sampling Surveys – versione R con interfaccia Java). Software per la determinazione dell'allocazione campionaria nel caso multivariato e per più domini di stima per le indagini ad uno stadio di campionamento.

Multiway Sample Allocation Pacchetto R che implementa l'allocazione del campione per disegni a uno stadio stratificati a più vie (semplice o con probabilità variabili) e disegni a stratificazione incompleta. L'allocazione permette di controllare gli errori campionari attesi delle stime di molti parametri calcolate su diverse sottopopolazioni di interesse.

SamplingStrata. Pacchetto R per la determinazione della stratificazione ottima in campioni stratificati ad uno stadio.

FS4 (First Stage Stratification and Selection in Sampling)

Descrizione

FS4 è un software generalizzato *open source* per la stratificazione e selezione delle unità di primo stadio in disegni di campionamento a due o più stadi, realizzato completamente in linguaggio R e dotato di un'interfaccia utente di tipo grafico (GUI).

Il software effettua il merge di due dataframe (quello della popolazione di PSU e quello di allocazione) e successivamente, calcola la stratificazione e la selezione di un campione a dimensione fissa di PSU per ciascuno strato realizzata con il metodo Sampford (selezione con probabilità proporzionale alla dimensione e senza reinserimento), implementato nella funzione UPsampford del package R "sampling". In merito alla stratificazione, la funzione realizza, all'interno di ciascun dominio di stima, il calcolo di una soglia dimensionale per una data misura di ampiezza. Le unità di primo stadio (PSU, Primary Sampling Unit) con la misura di ampiezza superiore alla soglia sono identificate come autorappresentative (SR, Self Representative) e ciascuna di esse forma uno strato a sé. Le restanti PSU non autorappresentative (NSR, Non Self Representative) sono ordinate per la misura di ampiezza e suddivise in strati di dimensione approssimativamente costante alla soglia corretta e con PSU aventi dimensioni il più possibile omogenee. FS4 determina, per ogni PSU, la dimensione del campione di unità finali sulla base di dimensioni pianificate in ingresso.

Le caratteristiche principali del package FS4 sono le seguenti:

- è un software flessibile che permette all'utente di:
    - scegliere la misura d'ampiezza per definire la stratificazione e la probabilità di inclusione per il campionamento PPS (Probability Proportional to Size);
    - inserire in input una singola o plurima (variabile tra domini) dimensione pianificata del campione di unità finali da osservare in ogni PSU;
    - lanciare la procedura in due passi separati, consentendo di visionare un primo output e sulla base di questo modificare i parametri di input;
- è un package open source rilasciato con licenza EUPL;
- ha una GUI user-friendly, che ne consente l'utilizzo anche ad utenti non conoscitori del linguaggio R, implementata in Tcl/Tk tramite il package R "tcltk, ma è anche possibile utilizzarlo direttamente dalla linea di comando di R attraverso la funzione Stratsel.

Informazioni

Status:	validato
Autore:	Istat
Licenza:	EUPL-1.1
Codifica GSBPM:	2.4 Design frame and sample 4.1 Create frame and select sample
Linguaggio di programmazione:	R
Versione linguistica della GUI:	EN
Parole chiave:	stratificazione, selezione, campionamento, PSU
Contatto:	nome: Raffaella Cianchetta email: cianchet@istat.it

Reperimento software e documentazione

REQUISITI TECNICI

Il package FS4 richiede l'installazione di R versione 3.0.1 o superiore e dei seguenti R package: tcltk2, svMisc, plyr e sampling.

COPYRIGHT

DISCLAIMER

L'Istat non si assume la responsabilità per risultati derivanti da un uso dello strumento non coerente con le indicazioni metodologiche contenute nella documentazione disponibile.

DOWNLOAD

- Versione 1.0 – Package precompilato: Sistemi Windows
- Versione 1.0 – Sorgenti del package: Sistemi Windows e Unix-like

DOCUMENTAZIONE TECNICA E METODOLOGICA

Reference manual – FS4 v. 1.0

Manuale utente - FS4 v. 1.0

MAUSS-R (Multivariate Allocation of Units in Sampling Surveys)

Descrizione

MAUSS-R è un software generalizzato open source per la determinazione dell'allocazione campionaria, sviluppato utilizzando il linguaggio R e dotato di una interfaccia Java per facilitare le selezioni di dati e parametri di interesse.

La metodologia di riferimento è 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).

L'algoritmo di Bethel generalizzato al caso multivariato permette di calcolare l'allocazione campionaria per disegni campionari stratificati: la dimensione complessiva del campione e l'allocazione nei diversi strati viene determinata basandosi sui vincoli imposti nelle indagini sulla precisione delle diverse stime di interesse.

Il software nella versione corrente riguarda la determinazione dell'allocazione campionaria nel caso multivariato e per più domini di stima per le indagini basate su un unico stadio di campionamento.

L'interfaccia permette:

- la modifica dei valori di default dei parametri di elaborazione (come il minimo numero di unità per strato);
- la gestione del file dei vincoli di precisione e la memorizzazione delle diverse versioni dei vincoli e dei risultati ottenuti;
- l'esecuzione del calcolo dell'allocazione ottimale;
- il confronto con le allocazioni proporzionali ed uguali;
- la visualizzazione dei report relativi alla popolazione, ai risultati dell'elaborazione e al confronto fra i risultati ottenuti al variare dei vincoli

Informazioni

Status:	validato
Autore:	Istat
Licenza:	EUPL-1.0
Codifica GSBPM:	2.4 Design frame and sample
Linguaggio di programmazione:	R, Java
Versione linguistica della GUI:	IT, EN
Parole chiave:	disegno del campione, allocazione ottima, dimensione del campione, Bethel
Contatto:	nome: Maria Teresa Buglielli email: bugliell@istat.it

Reperimento software e documentazione

REQUISITI TECNICI

- Ambiente R versione 2.7.0 o superiore.

- Java 2 Runtime Environment versione 6.0 o superiore (solo per l'interfaccia grafica).

COPYRIGHT

DISCLAIMER

L'Istat non si assume la responsabilità per risultati derivanti da un uso dello strumento non coerente con le indicazioni metodologiche contenute nella documentazione disponibile.

DOWNLOAD

MAUSS-R versione 1.1

DOCUMENTAZIONE TECNICA E METODOLOGICA

Manuale utente – MAUSS-R v. 1.1 – (Inglese)

Manuale utente – MAUSS-R v. 1.1 – (Italiano)

Multiway Sample Allocation

Descrizione

Multiway Sample Allocation è un package R che permette di allocare il campione per un disegno campionario con stratificazione a più vie.

La stratificazione a più vie si ottiene combinando le variabili che definiscono le celle marginali di una tabella di contingenza a più vie. Nell'ambito del campionamento da popolazioni finite le celle marginali sono identificate dalle modalità delle variabili che definiscono le partizioni in domini di interesse della popolazione obiettivo tra loro indipendenti. Ad esempio, la stratificazione ottenuta combinando le modalità di due partizioni in domini di interesse definisce una stratificazione a due vie.

Una partizione è definita indipendente o marginale quando non è possibile ottenere i domini per aggregazione di altri domini definiti in altre partizioni. L'allocazione dei del campione nei domini dipendenti avviene aggregando la numerosità campionaria dei domini marginali.

Il package Multiway Sample Allocation definisce l'allocazione del campione nel rispetto dei vincoli di precisioni delle stime per:

- diversi parametri (totali) di interesse (problema multivariato) a livello di sottopopolazioni o domini di riferimento (problema multidominio);
- disegni di campionamento stratificati a uno stadio semplici o con probabilità di inclusione variabili negli strati a una o più vie, in cui la dimensione campionaria è fissata a livello di strato; disegni a stratificazione incompleta in cui la dimensione campionaria è fissata (approssimativamente a meno di un processo di arrotondamento all'intero superiore/inferiore) a livello di dominio di interesse e non negli strati;
- disegni a probabilità variabili in cui la dimensione campionaria è fissata (approssimativamente) a livello di dominio (l'esecuzione del processo di allocazione soggetta a vincoli computazionali connessi con la dimensione della popolazione e il numero delle stime da considerare).

L'algoritmo che produce l'allocazione (Falorsi e Righi, 2015) è un'estensione di quello di Chromy (1987) e Bethel (1989). Tale algoritmo risolve un problema di ottimizzazione formalizzato secondo una espressione generale della varianza delle stime (Falorsi e Righi, 2015) che dipende dal:

- modello di superpopolazione usato per definire i parametri di input;
- disegno di campionamento implementato.

Il principale output del package è la probabilità di inclusione delle unità della popolazione.

Informazioni

Status:	validato
Autore:	Istat
Licenza:	EUPL-1.1
Codifica GSBPM:	2.4 Design frame and samplee
Linguaggio di programmazione:	R
Parole chiave:	allocazione campionaria, stratificazione a più vie, disegno a stratificazione incompleta
Contatto:	nome: Paolo Righi email: parighi@istat.it

Reperimento software e documentazione

REQUISITI TECNICI

Ambiente R versione 3.1.1 o superiore. Richiede l'installazione del pacchetto MASS di R.

COPYRIGHT

DISCLAIMER

L'Istat non si assume la responsabilità per risultati derivanti da un uso dello strumento non coerente con le indicazioni metodologiche contenute nella documentazione disponibile.

DOWNLOAD

- Versione 1.0 – Package precompilato: Sistemi Windows
- Versione 1.0 – Sorgenti del package: Sistemi Windows e Unix-like

DOCUMENTAZIONE TECNICA E METODOLOGICA

Reference manual - Multiway Sample Allocation v. 1.0

ALTRA DOCUMENTAZIONE

De Vitiis C., P. Righi, M. D. Terribili. 2016. *Optimal sample allocation for the Incomplete Stratified Sampling design*. *Rivista di Statistica Ufficiale*, in corso di stampa.

Falorsi P. D., P. Righi. 2016. A flexible tool for defining optimal sampling designs. *The Survey Statistician*, 73:21-31.

Falorsi P. D., P. Righi. 2015. Generalized Framework for Defining the Optimal Inclusion Probabilities of One-Stage Sampling Designs for Multivariate and Multi-domain Surveys. *Survey Methodology*, 41.

Falorsi P. D., P. Righi. 2008. A Balanced Sampling Approach for Multi-way Stratification Designs for Small Area Estimation. *Survey Methodology*, 34(2):223-234.

SamplingStrata

Descrizione

SamplingStrata è un package R che permette di determinare la stratificazione ottimale di un frame di campionamento, tale cioè da garantire la massima efficienza di un campione (espressa dalla sua dimensione totale e dal costo associato) rispettando i vincoli di precisione posti sulle stime target di una data indagine.

Il package SamplingStrata permette di coprire tutte le fasi dell'attività di disegno e selezione del campione.

Le funzioni offerte dal package possono essere classificate in funzioni di:

1. **preparazione** dell'ambiente di elaborazione
2. **ottimizzazione** della stratificazione e analisi dei risultati
3. **stratificazione del frame** secondo la soluzione ottima trovata
4. **selezione del campione.**

Informazioni

Status:	validato
Autore:	Istat
Licenza:	GPL-2 | GPL-3
Codifica GSBPM:	2.4 Design frame and sample 4.1 Create frame and select sample
Linguaggio di programmazione:	R
Parole chiave:	stratificazione ottima, disegno campionario, allocazione campionaria, algoritmo genetico
Contatto:	nome: Giulio Barcaroli email: barcarol@istat.it

Reperimento software e documentazione

REQUISITI TECNICI

Ambiente R versione 2.15.0 o superiore.

COPYRIGHT

Licenza è distribuito “TAL QUALE”, SENZA GARANZIE O CONDIZIONI DI ALCUN TIPO, esplicite o implicite.

DISCLAIMER

L’Istat non si assume la responsabilità per risultati derivanti da un uso dello strumento non coerente con le indicazioni metodologiche contenute nella documentazione disponibile.

DOWNLOAD

- Versione 1.1 – Package precompilato: Sistemi Windows
- Versione 1.1 – Sorgenti del package: Sistemi Windows e Unix-like

DOCUMENTAZIONE TECNICA E METODOLOGICA

Reference manual – SamplingStrata v. 1.1

Vignettes – SamplingStrata v. 1.1

ALTRA DOCUMENTAZIONE

Barcaroli G. 2014. SamplingStrata: An R Package for the Optimization of Stratified Sampling. *Journal of Statistical Software*, 61(4):1-24.

Ballin M., Barcaroli G. 2013. Joint determination of optimal stratification and sample allocation using genetic algorithm. *Survey Methodology*, 39(2):369-393.

Fase RACCOLTA

Creazione della lista e selezione del campione

Le attività relative alla creazione della lista e alla selezione del campione corrispondono al sottoprocesso 4.1 "*Create frame and select sample*" del GSBPM.

La creazione della lista consiste nella costruzione dell'archivio delle unità appartenenti alla popolazione di interesse. La selezione del campione consiste nella individuazione delle unità campionarie sulla base di uno schema di campionamento predefinito.

Per una data occasione di indagine, la creazione della lista e la selezione del campione sono effettuate in base alle specifiche definite nel sottoprocesso 2.4 "*Design frame and sample*".

Acquisizione dati

La fase di acquisizione dati, intesa nel senso stretto del termine, ossia di raccolta del dato presso le unità rispondenti, è preceduta da un insieme complesso ed articolato di attività necessarie alla definizione di un questionario d'indagine che permetta di raccogliere, per poi di misurare, i diversi aspetti del fenomeno indagato.

Nella schematizzazione del GSBPM, quanto appena detto si traduce in un insieme di sottoprocessi che percorrono il modello dalla fase 1 alla fase 4.

Fase 1 "*Specify needs*". In questa fase dovranno essere realizzati i sottoprocessi dall'1.1 all'1.5 necessari ad individuare gli obiettivi di indagine ed a tradurli in concetti che, da un lato dovranno essere comprensibili e accessibili ai rispondenti, e dall'altro, dovranno essere misurabili e quindi trasformabili in variabili statistiche, che saranno disegnate nella fase 2. In questa prima fase è importante verificare l'esistenza dei dati presso altre fonti (ad esempio dati amministrativi) in modo da ridurre l'insieme delle variabili da rilevare con effetti positivi sia sui costi di rilevazione che sul fastidio statistico.

Fase 2: "*Design*". Al termine della Fase 1, si procederà con la fase di disegno che, per l'acquisizione, riguarda le attività descritte nei sottoprocessi 2.1, 2.2 e 2.3. Le variabili individuate nella Fase 1 permetteranno la progettazione del piano di tabulazione (passaggio utile anche per la fase di disseminazione dei risultati) dal quale saranno desumibili anche le variabili derivate (che non saranno rilevate ossia raccolte in fase di acquisizione) e le classificazioni ufficiali che saranno utilizzate. A questo punto sarà possibile "tradurre" le variabili in domande del questionario, mentre le loro caratteristiche (valori ammissibili) e relazioni tra di esse rappresenteranno, rispettivamente, i controlli di validità ed i flussi di intervista. Il disegno delle variabili dovrebbe essere fatto tenendo in considerazione i metadati esistenti in Istituto al fine di utilizzare definizioni esistenti o di inserirne delle nuove qualora mancanti nel sistema. Questo favorirebbe il processo di standardizzazione, anche in un'ottica internazionale, nonché il riuso di elementi definiti in altri processi di acquisizione.

Il disegno delle variabili dovrà essere fatto in parallelo con il sottoprocesso 2.3 "*Design Collection*", finalizzato ad individuare il/i metodo/i di raccolta dati "più appropriato/i". Questo perché il disegno delle variabili e, quindi, la struttura e la formulazione dei quesiti da inserire nel questionario, è strettamente dipendente dalla tecnica di rilevazione utilizzata. In questo contesto rientra anche la progettazione dei controlli che si vogliono inserire durante la fase di

raccolta, quando la tecnica usata è basata sull'uso del computer (es: CATI, CAPI, CAWI, descritte nella Fase 3). È buona regola, infatti, progettare i controlli tenendo presente l'equilibrio tra qualità del dato e il fastidio statistico del rispondente, che, se elevato, potrebbe influire negativamente sulle mancate risposte totali e/o parziali.

Fase 3 "*Build*": in base alle indicazioni provenienti dalla fase di disegno verranno costruiti gli strumenti per la raccolta dati, sottoprocesso 3.1 "*Build collection instruments*" del GSBPM. La fase di acquisizione avviene attraverso l'utilizzo di uno o più tecniche (strumenti), assistite o non dal computer, con o senza il supporto del rilevatore, nonché attraverso strumenti, quali EDI (*Electronic Data Interchange*), XBRL (*eXtensible Business Reporting Language*), adatti allo scambio di informazioni tra l'ente statistico e le unità di rilevazione (aziende e/o organi istituzionali produttori di dati sotto forma di fonti amministrative o registri). E' in questo sottoprocesso che si effettua anche il test di funzionalità del questionario elettronico. E' raccomandabile stabilire una connessione diretta tra gli strumenti di raccolta ed il sistema di metadati al fine di agevolare la fase di costruzione degli strumenti nonché le fasi successive del processo, come ad esempio la Diffusione.

Tra le tecniche di acquisizione dati, giocano un ruolo sempre più importante quelle assiste dal computer, ossia **CADI** (*Computer Assisted Data Inputing*), **CAPI** (*Computer Assisted Personal Interviewing*), **CATI** (*Computer Assisted Telephone Interviewing*), **CAWI** (*Computer Assisted Web Interviewing*). Come già in parte anticipato, la caratteristica più significativa di queste tecniche consiste nel fatto di permettere l'inserimento già in fase di raccolta dati di tutti quei controlli tipici delle successive fasi di controllo e correzione, inibendo di fatto l'acquisizione del dato errato. Si differenzia la tecnica **CADI** che, usata nell'ambito di rilevazione tramite modelli cartacei, consiste di fatto in un'acquisizione controllata, dove i controlli sono inseriti solo al fine di ridurre quelli di registrazione o anche come supporto alla fase di revisione.

Un'altra peculiarità di queste tecniche, sempre ad eccezione delle **CADI**, è quella di consentire la personalizzazione della formulazione dei quesiti in funzione delle caratteristiche del rispondente (nome, sesso) o di risposte fornite a precedenti quesiti del questionario stesso o di informazioni già disponibili, perché rilevate in precedenti indagini, rendendo così l'intervista più colloquiale e facilitando la disponibilità del rispondente a collaborare. La tipologia e complessità dei controlli inseriti in fase di acquisizione è notevolmente differente tra **CATI** e **CAPI** da un lato e **CAWI** dall'altro:

- per le prime due, la presenza di un rilevatore formato sia sui contenuti di indagine che sulla tecnica di somministrazione del questionario permette sia di inserire una quantità di controlli decisamente più elevata di quanto sia possibile effettuare con il **CAWI** che di fare un uso maggiore di controlli di tipo bloccanti ossia che richiedono la risoluzione delle incompatibilità per poter proseguire con l'intervista;
- nel **CAWI**, invece, dove il questionario è auto-compilato dal rispondente, è necessario non solo contenere il numero di controlli inseriti, ma anche di gestirli con degli avvertimenti che segnalano l'incongruenza senza l'obbligo di sanarla per procedere con l'intervista.

È buona regola in tutti i casi progettare i controlli tenendo presente l'equilibrio tra qualità del dato e fastidio statistico del rispondente, che, se elevato, potrebbe influire negativamente sulle mancate risposte totali e/o parziali.

È bene ricordare di nuovo, che l'adozione di queste tecniche di rilevazione ha già un impatto in fase di progettazione di disegno del questionario di indagine, che dovrà successivamente essere

tradotto in questionario elettronico. Il questionario elettronico deve poi essere approfonditamente testato, non soltanto per verificare le performance dell'applicazione informatica in termini di conformità alle specifiche e dei tempi di risposta, ma anche in termini di gradevolezza e fluidità dell'intervista.

Fase 4 "Collect": Dopo la costruzione e test degli strumenti, può iniziare la fase di raccolta dei dati che, espressa in termini dei sottoprocessi 4.2 "Set-up collection" 4.3 "Run collection" e 4.4 "Finalise collection", include tutte le attività che vanno dalla formazione dei rilevatori fino alla predisposizione di appositi ambienti informatizzati dove saranno caricati i dati raccolti per poi essere sottoposti al successivo passaggio di elaborazione (Fase 5 "Process").

Fase RACCOLTA – METODI

- **CREAZIONE DELLA LISTA E SELEZIONE DEL CAMPIONE**

MEMOBUST – Handbook on Methodology of Modern Business Statistics:
Statistical Registers and Frames
Sample Selection
2014
MEETS ESSnet MEMOBUST

- **ACQUISIZIONE DATI**

La modernizzazione delle indagini via Web sulle imprese - Pratiche Raccomandate per il disegno dei questionari
2015
Gdl Armonizzazione dei questionari di impresa, Istat

MEMOBUST – Handbook on Methodology of Modern Business Statistics:
Questionnaire Design
Data Collection
2014
MEETS ESSnet MEMOBUST

Handbook of Recommended Practices for Questionnaire Development and Testing in the European Statistical System
2006
European project QDet

Riferimenti

Murgia M., A. Nunnari. 2012. Improving the quality of data collection: minimum requirements for generalised software independent from the mode. *Seminar on New Frontiers for Statistical Data Collection*, UNECE, Ginevra, 31 ottobre-2 novembre.

Istat. 2008. *La progettazione e lo sviluppo informatico del sistema Capi sulle forze di lavoro*. Collana Metodi e Norme, n. 36, Istat.

Istat. 2006. *L'indagine campionaria sulle nascite: obiettivi, metodologia e organizzazione*. Collana Metodi e Norme, n. 28, Istat.

Fase RACCOLTA – STRUMENTI

- **ACQUISIZIONE DATI**

Blaise

Sistema per l'acquisizione dei dati assistita da computer.

Blaise

Descrizione

Blaise, sviluppato e commercializzato da Statistics Netherlands, è un sistema per l'acquisizione dati assistita da computer. È composto da una serie di moduli che supportano molte funzioni tra cui:

- la schedulazione delle telefonate per le indagini CATI (*CATI Call Management System)*
- la somministrazione del questionario (*Data Entry*)
- la codifica assistita per le risposte testuali (*Coding*).

La schedulazione delle telefonate è gestita da Blaise attraverso un modulo ad hoc chiamato *CATI Call Management System*, che in termini molto generali, seleziona i numeri (rispondenti) da contattare in funzione di una priorità di chiamata che viene stabilita in base agli esiti del contatto telefonico e del valore dei parametri di schedulazione impostati dal responsabile di indagine. Questi ultimi permettono di stabilire quante volte un numero con un esito provvisorio (libero, occupato, appuntamento) debba essere proposto nell'arco temporale della rilevazione in attesa che si trasformi in un esito definitivo (intervista, rifiuto, interruzione definitiva) o che raggiunga il massimo numero di contatti possibili e venga sostituito con una riserva.

La somministrazione del questionario avviene attraverso il modulo *Data Entry* che permette di gestire tutte le tipologie di domande utilizzate nelle indagini statistiche e di organizzare i quesiti in sezioni. Consente anche di inserire i controlli di flusso, tra domande e tra sezioni, i controlli di consistenza su singole variabili e di coerenza tra due o più variabili. Permette la gestione parametrizzata del testo dell'errore in modo da facilitare il lavoro del rispondente/intervistatore.

Il modulo di codifica assistita basato sui s*ub-strings algorithms***,** consente la navigazione del dizionario informatizzato secondo tre modalità:

- secondo l'albero: il codificatore visualizza in un primo tempo la classificazione a livello dei rami gerarchicamente più elevati (utilizzabile solo nel caso di classificazioni gerarchiche)
- secondo la dizione alfabetica: che può essere realizzata secondo due diverse logiche quella per *trigrammi* (il codificatore digita un testo ed il sistema ricerca ed estrae dal dizionario i testi che hanno in comune con quello digitato uno o più trigrammi), ed una seconda, per *ordine alfabetico,* (il codificatore digita un testo ed il sistema si posiziona nel dizionario nel corrispondente punto dell'ordine alfabetico)
- secondo una procedura mista: il codificatore seleziona un ramo della classificazione e quindi effettua una ricerca testuale nell'ambito di quel ramo

Informazioni

Status:	validato
Autore:	Statistics Netherlands
Codifica GSBPM:	3.1 Build collection instrument 4.2 Set up collection 4.3 Run collection 4.4 Finalise collection 5.2 Classify and code
Parole chiave:	CAPI, CATI, CADI, schedulatore, codifica assistita

Reperimento software e documentazione

Per il reperimento del software e della documentazione tecnica e metodologica è possibile rivolgersi a Statistics Netherlands.

ALTRA DOCUMENTAZIONE

Degortes M., S. Macchia, M. Murgia. 2003. The usage of BLAISE in an integrated application for the Births Sample Survey in CATI mode. In Proceedings of the 8th International BLAISE Users *Conference (IBUC 2003)*, International Blaise Users Group, Copenhagen, 21-23 maggio 2003.

Fase ELABORAZIONE

Integrazione

Record linkage

Il record linkage è un processo importante per l'integrazione di dati provenienti da fonti diverse; esso mira ad identificare i record, riferiti alle medesime unità individuali, collocati nello stesso file (de-deduplicazione) o in file diversi (integrazione di fonti). L'identificazione dell'unità in archivi di diversa natura avviene attraverso chiavi comuni, presenti nei vari file; le chiavi possono essere anche non perfettamente corrispondenti. La complessità del record linkage dipende da molteplici aspetti, principalmente legati all'assenza di identificatori univoci o alla presenza di errori negli identificatori stessi.

Nella statistica ufficiale, l'uso di tecniche di record linkage nei vari processi di produzione è ormai diffuso da diversi anni e molteplici sono i campi di applicazione:

- individuazione dei duplicati in un file di dati individuali;
- studio dell'associazione tra variabili raccolte da fonti differenti;
- identificazione dei casi multipli attribuibili ad un singolo individuo (ad esempio ricoveri, parti, ecc.) in un archivio;
- creazione e aggiornamento di liste per la conduzione di indagini;
- re-identificazione per tutela riservatezza di micro-dati rilasciati per uso pubblico;
- determinazione della numerosità di una popolazione con il metodo cattura-ricattura;
- analisi di dati panel;
- ecc.

Il record linkage è un processo complesso a causa dei numerosi aspetti di natura diversa che lo compongono. Se negli archivi da abbinare sono presenti identificatori univoci allora il problema non ha una grande complessità; in generale però, per analizzare dati privi di identificatori univoci o con identificatori univoci affetti da errore, sono richieste sofisticate procedure statistiche; soluzioni informatiche non banali sono necessarie per gestire e trattare grandi moli di dati, mentre i vincoli legati al tipo di applicazione che si intende effettuare possono comportare la soluzione di complessi problemi di programmazione lineare.

Statistical matching

Lo statistical matching (abbinamento statistico) o data fusion si pone l'obbiettivo di integrare due o più fonti dati relative alla stessa popolazione con l'intento di esplorare le relazioni tra variabili non osservate congiuntamente. Le fonti da integrare osservano unità distinte, come di solito accade quando si vogliono integrare indagini campionarie. La situazione tipica dello statistical matching è quella in cui sono disponibili due fonti dati A e B; in A sono disponibili le variabili X ed Y, mentre in B sono disponibili X e Z; l'obiettivo è quello di studiare la relazione tra Y e Z integrando le fonti dati sulla base delle informazioni in comune X. L'interesse può essere di tipo 'macro' o 'micro'; nel primo caso si vogliono studiare i parametri che sintetizzano la relazione tra Y e Z, per esempio coefficiente di correlazione, coefficiente di regressione, tabella

di contingenza; nel secondo caso invece si vuole ottenere un data set completo (data set sintetico) in cui sono presenti tutte le variabili di interesse, X, Y e Z.

Gli obiettivi del matching possono essere realizzati mediante l'utilizzo di metodi parametrici, non parametrici o misti.

L'approccio parametrico prevede la specificazione di un modello e la stima dei parametri che lo caratterizzano. In assenza di informazioni ausiliarie il modello generalmente assunto si basa sull'assunzione di indipendenza condizionata di Y e Z date le variabili comuni X. Tale assunzione è piuttosto forte e purtroppo nella situazione tipica del matching non può essere verificata tramite un test.

I metodi non parametrici solitamente sono applicati quando si ha un obiettivo micro. I metodi hot-deck (imputazione da donatore) sono fra i metodi non parametrici più utilizzati: si basano sull'imputazione (predizione) della variabile mancante nel data set scelto come ricevente (e.g., il data set A) selezionando dei valori dal data set donatore (B). Operativamente, ad ogni unità del data set A (unità ricevente) viene associata una osservazione in B (unità donatrice) selezionata rispetto al suo grado di similarità calcolato sulla base dei valori della variabile comune X.

In letteratura è stato introdotto anche un approccio misto che prevede un primo passo di imputazione tramite modello parametrico, ed un secondo passo di imputazione non parametrica che fa uso dei valori imputati al primo passo per il calcolo della similarità fra unità riceventi e donatrici.

Vale la pena di osservare che è possibile utilizzare un approccio alternativo basato sulla quantificazione dell'incertezza. Tale approccio non richiede l'introduzione dell'ipotesi di indipendenza condizionata o di informazioni ausiliarie sui parametri non stimabili, i.e., parametri che fanno riferimento alle relazioni fra Y e Z. Lo studio dell'incertezza non conduce però generalmente ad una stima univoca dei parametri quanto piuttosto ad un insieme di stime. L'insieme è composto da tutte le possibili stime dei parametri che fanno riferimento alle variabili Y e Z coerenti con quelle ottenibili dai dati osservati, ovvero quelle che fanno riferimento alle coppie (Y,X) e (Z,X).

L'applicazione del matching a dati provenienti da indagini campionarie complesse pone problemi aggiuntivi. In tali circostanze ai fini dell'inferenza bisogna tener conto del disegno di campionamento prescelto per selezionare il campione nonché di altre metodologie usate per far fronte a problemi di natura non campionaria (copertura e mancate risposte totali).

Codifica delle risposte testuali

La codifica rappresenta una fase del processo di produzione statistica da includere nell'organizzazione dello stesso quando il questionario di rilevazione contiene variabili testuali, ossia domande la cui risposta è un testo libero. Si tratta generalmente di variabili testuali per le quali esiste una classificazione ufficiale (Attività economica, Professione, Titolo di studio, Comune e/o Stato di nascita o residenza) che permette la comparabilità del dato raccolto a livello nazionale e/o internazionale. Codificare vuol dire associare al testo rilevato un codice univoco sulla base dello schema classificatorio di riferimento. Il livello di dettaglio del codice da attribuire al testo dipende dagli obiettivi dell'indagine e/o dal livello di dettaglio richiesto per la fase di Diffusione. La codifica può essere fatta manualmente o attraverso sistemi automatizzati. Nel primo caso avviene al termine della fase di raccolta, mentre nel secondo caso può avvenire anche durante la fase di raccolta dati (anch'essa assistita da computer): si parla, di codifica

assistita se effettuata durante la fase di acquisizione e di codifica automatica se effettuata a posteriori.

In termini di GSBPM, la codifica è un sottoprocesso 5.2 "*Classify and code*" della Fase 5 "*Process*" che include tutte quelle attività cui sottoporre i dati per renderli pronti alla successiva fase di analisi (Fase 6 "*Analyse*"). In realtà, parte delle attività della Fase 5 possono iniziare anche prima che la precedente Fase 4 "*Collect*" sia terminata, proprio come nel caso della codifica assistita. Questo permette di migliorare la tempestività nel rilascio dei dati.

Nella gestione di un'indagine la fase di codifica delle risposte testuali è molto onerosa e se eseguita manualmente è anche poco standardizzabile in quanto il risultato è fortemente influenzato dal codificatore. Infatti, sebbene gli addetti alla codifica siano formati sui principi e sui criteri con cui è costruita ogni classificazione, l'attribuzione di un codice è sempre soggetta al fattore interpretazione, il che può comportare che, a parità di formazione, due codificatori attribuiscano codici diversi allo stesso testo.

L'adozione di software specifici per la codifica comporta vantaggi non soltanto in termini di risparmio di tempi e risorse da dedicare a quest'attività, ma soprattutto garantisce la standardizzazione del processo il che implica un più elevato livello di qualità del processo stesso. La codifica tramite computer può avvenire secondo due modalità:

- **automatica**: il software analizza (in batch) un file contenente l'insieme di risposte testuali raccolte al termine dell'indagine;
- **assistita**: il software costituisce un supporto interattivo per il codificatore/rispondente, facilitando la navigazione nella classificazione di riferimento.

Gli obiettivi propri della codifica automatica e della codifica assistita sono diversi: nel caso della codifica automatica la finalità è di individuare ed estrarre dal dizionario una singola descrizione che realizzi il *match* con quella da codificare; nella codifica assistita può essere opportuno, invece, estrarre dal dizionario un set di descrizioni, anche molto simili tra loro, lasciando poi al codificatore la selezione di quella corretta.

Il punto cardine di qualunque sistema di codifica automatica/assistita è la costruzione della base informativa ovvero del dizionario informatizzato relativo al manuale ufficiale della classificazione di riferimento arricchito, di volta in volta, con i testi rilevati durante le indagini realizzate dall'istituto (e correttamente codificati). Quest'ultimo, però, per essere trattato da un software dovrà essere sottoposto ad una serie di operazioni finalizzate ad includere nei dizionari solo descrizioni che siano sintetiche, analitiche e non ambigue. E' importante sottolineare, inoltre, che anche la ricchezza di testi del dizionario informatizzato impatta direttamente sul tasso di codifica.

I sistemi di codifica si differenziano secondo gli algoritmi di ricerca utilizzati per realizzare il *match* tra le descrizioni-risposta e le descrizioni del dizionario. Tali algoritmi sono riconducibili alle seguenti categorie:

- dictionary algorithms: algoritmi che si avvalgono di parole (o gruppi di parole) particolarmente informative per determinare univocamente l'assegnazione del codice;
- weighting algorithms: ricerca di match esatti o parziali sulla base di funzioni di similarità tra testi dove alle parole è attribuito un peso, empirico o probabilistico, proporzionale al loro grado d'informatività;

- sub-strings algorithms: ricerca di match basati sull'accoppiamento di bigrammi o trigrammi di testo.

Inoltre, nel caso di codifica assistita è possibile navigare nel dizionario secondo tre metodi effettuando:

- la ricerca per ramo: si naviga dentro la struttura gerarchica della classificazione, dal ramo più alto fino a quello più basso (foglia) che rappresenta il codice finale al massimo dettaglio da attribuire al testo da codificare;
- la ricerca alfabetica: si naviga in tutto il dizionario alla ricerca della stringa identica o più simile a quella da codificare;
- la ricerca mista: si naviga per ramo e all'interno del ramo selezionato si procede con la ricerca alfabetica.

La scelta del metodo di navigazione è fortemente influenzata dalla tecnica di acquisizione dati utilizzata, in particolare, se si tratta di una tecnica con o senza intervistatore. In quest'ultimo caso, ad esempio nelle interviste via web, occorre predisporre uno strumento di codifica che sia da un lato facilmente utilizzabile dal rispondente e dall'altro garantisca un'elevata qualità del dato codificato.

Sulla qualità della codifica influisce fortemente il contenuto del dizionario informatizzato nonché la fase di addestramento del software. Sarebbe auspicabile che entrambe, ossia l'aggiornamento del dizionario e delle regole software di *matching*, siano effettuate periodicamente in genere al termine di ogni fase di codifica legata ad una particolare indagine. A tal fine è importante eseguire il controllo sui risultati di un passaggio di codifica automatica/assistita per:

- verificare la qualità dei casi codificati;
- utilizzare i casi di errore di codifica e di fallimento per aggiornare l'applicazione;
- mettere in luce eventuali carenze della Classificazione di riferimento.

Per la valutazione della qualità delle due modalità di codifica, è possibile utilizzare i seguenti indicatori:

Indicatori per la codifica automatica:

- efficacia/tasso di codifica, ovvero la percentuale di testi codificati sul totale di quelli da codificare;
- accuratezza, ovvero la percentuale di codici corretti assegnati sul totale dei testi codificati con l'ausilio del computer;
- efficienza, ovvero il tempo unitario di assegnazione del codice.

Indicatori per la codifica assistita:

- tempo medio per l'attribuzione del codice;
- coerenza tra descrizione testuale rilevata in fase d'intervista e codice attribuito.

Individuazione e trattamento degli errori di misura e delle mancate risposte parziali

Le mancate risposte parziali (MRP) e gli errori di misura sono particolari errori non campionari che vengono individuati e trattati nella fase di controllo e correzione dei dati.

Per errore di misura si intende qui una discrepanza tra valore “vero” e valore “osservato” di una variabile in un'unità, dovuta a qualsiasi difetto del processo di misurazione (rilevazione, codifica, registrazione, ecc.). Mancate risposte ed errori di misura possono compromettere seriamente l'accuratezza delle stime di interesse e dovrebbero essere prevenuti con opportuni accorgimenti nel processo di misurazione. Anche dopo l'adozione di tali accorgimenti è tuttavia inevitabile che una frazione dei dati registrati sia caratterizzata dalla presenza di errori e mancate risposte che richiedono quindi l'utilizzo di appositi metodi di controllo e correzione.

Nell'ambito del GSBPM si distinguono due sottoprocessi: 5.3 “Review and validate” e 5.4 “Edit and impute”, che attengono rispettivamente al processo di verifica della validità dell'informazione a disposizione e al complesso di attività volte alla localizzazione degli errori e alla sostituzione dei valori ritenuti errati con valori plausibili (**imputazione**). Non sempre, nei reali contesti operativi, questi due sottoprocessi sono nettamente distinti.

Individuazione degli errori di misura

I metodi di individuazione degli errori possono essere classificati a secondo delle tipologie di errore per cui sono impiegati. Un primo importante criterio di classificazione degli errori distingue **errori sistematici** ed **errori casuali** (o stocastici o non sistematici). Si dicono **sistematici** quegli errori la cui origine è da attribuirsi a difetti strutturali o organizzativi del processo di produzione dell’informazione statistica, alla struttura del modello, o al sistema di registrazione adottati, e si manifestano nella maggior parte dei casi come deviazioni “in una stessa direzione” dal valore vero di una o più variabili rilevate. Gli errori sistematici vengono generalmente trattati con regole deterministiche basate sulla conoscenza del meccanismo che ha generato l'errore. Tra gli errori sistematici, particolarmente comuni sono gli errori di unità misura per le variabili quantitative.

Si dicono **casuali** quegli errori la cui origine è da attribuirsi a fattori aleatori non direttamente individuabili. A differenza degli errori sistematici, per gli errori casuali l'approccio deterministico è sconsigliabile.

Una importante classe di errori è quella degli errori che si manifestano determinando valori di alcune variabili **fuori dominio**, cioè non appartenenti ad un insieme predefinito di valori ammissibili. Simili a questi errori sono quelli che determinano **incoerenze** nei dati, cioè che possono essere rilevati con l'applicazione di **regole di compatibilità** (**edit**) tra le variabili. Gli errori ritenuti poco influenti che determinano incompatibilità tra gli item osservati sono generalmente "localizzati" usando metodi automatici basati su principi "generali". Un approccio particolarmente diffuso in quest'ambito è basato sul **principio del minimo cambiamento**, secondo il quale per ciascun record con errori, deve essere cambiato il numero minimo di variabili che consenta di rendere il record compatibile rispetto agli edit. Sulla base di questo principio è stata sviluppata la metodologia **Fellegi-Holt** inizialmente limitata alle variabili categoriche e successivamente estesa a quelle numeriche.

Un'altra classe importante di errori include quelli che si manifestano con la presenza di **dati anomali** (**outlier**), cioè unità con caratteristiche significativamente diverse da quelle della maggior parte delle altre unità. Le tecniche di individuazione degli outlier utilizzano di solito, implicitamente o esplicitamente, un modello per i dati "corretti", e cercano di identificare le unità che si discostano dal modello. I metodi di ricerca degli outlier sono anche spesso utilizzate in procedure di **editing selettivo** finalizzate all'identificazione degli **errori influenti**. Il concetto di errore influente, pur essendo collegato a quello di outlier, ne è concettualmente distinto.

Mentre la definizione di outlier è legato esclusivamente al modello assunto (almeno implicitamente) per i dati, quello di errore influente dipende anche strettamente dalla stima di interesse. In particolare valori anomali possono non dipendere da errori influenti, e viceversa, errori influenti possono non determinare valori anomali.

Correzione degli errori e imputazione delle mancate risposte parziali

Qualunque sia la tecnica utilizzata, al termine della fase di individuazione degli errori si pone la necessità di sostituire (imputare) i valori classificati come inaccettabili con valori vicini a quelli veri e di integrare le eventuali informazioni mancanti. L'imputazione rappresenta inoltre la procedura comunemente usata per le mancate risposte parziali. L'uso dell'imputazione è giustificata da una serie di motivi sia operativi sia teorici. In primo luogo, normalmente i dati rilasciati dall'Istat necessitano di essere completi (e coerenti) a livello elementare. Inoltre, l'imputazione consente di applicare all'insieme finale di microdati completi metodi e software standard di analisi statistica.

I metodi di imputazione sono numerosi e implementati in numerosi pacchetti statistici sia di proprietà esclusiva sia liberi. Tra le possibili classificazioni, una particolarmente usata distingue tra **metodi parametrici** che si basano su ipotesi distribuzionali esplicite (ad esempio imputazione per regressione) e **metodi non parametrici** che evitano assunzioni distribuzionali (**hot-deck imputation**, **donatore di minima distanza** ecc.). Inoltre i metodi di imputazione possono essere divisi in metodi deterministici, che a seguito di applicazioni ripetute producono gli stessi risultati, e metodi stocastici, caratterizzati da una certa variabilità degli output.

La scelta del metodo di imputazione dipende dagli obiettivi delle analisi che devono essere effettuate sui dati "completati". Ad esempio, se l'interesse dell'analista è rivolto alla stima di quantità lineari come medie o totali può essere opportuno utilizzare un metodo di imputazione deterministico, mentre se sono di interesse anche altre caratteristiche distribuzionali dei dati (come ad esempio momenti di ordine superiore al primo), un metodo di imputazione stocastico di norma è preferibile.

Calcolo delle stime e degli errori campionari

Le attività che riguardano la produzione delle stime di interesse e la valutazione degli errori campionari fanno riferimento ai sottoprocessi 5.6 "*Calculate weights*" e 5.7 "*Calculate aggregates"* del GSBPM.

Produzione delle stime di interesse

Ogni metodo di stima campionaria è fondato sul principio che il sottoinsieme delle unità della popolazione incluse nel campione deve rappresentare anche il sottoinsieme complementare costituito dalle rimanenti unità della popolazione stessa. Tale principio è generalmente realizzato attribuendo a ciascuna unità inclusa nel campione un peso che può essere visto come il numero di elementi della popolazione rappresentati da tale unità.

Le indagini campionarie condotte dall'Istat sono indagini su larga scala che hanno la finalità di fornire un elevato numero di stime di parametri della popolazione che possono essere di natura differente, quali ad esempio frequenze assolute, totali, proporzioni, medie, ecc.

La stima dei parametri della popolazione può essere effettuata ricorrendo a due diversi approcci di stima:

- **Metodi basati sull'approccio diretto** che usano i valori della variabile di interesse osservati sulle sole unità del campione appartenenti al dominio di interesse. Sono i metodi standard utilizzati dall'Istat e in genere da tutti i più importanti Istituti Nazionali di Statistica per la produzione delle stime delle diverse indagini.
- **Metodi basati sull'approccio indiretto** che utilizzano i valori della variabile di interesse osservati sulle unità del campione appartenenti ad un dominio più ampio contenente il dominio di interesse e/o ad altre occasioni di indagine. Sono utilizzati, usualmente, per problemi di stima particolari, quali ad esempio quelli connessi alla produzione di stime riferite ad aree o domini in cui la dimensione campionaria risulta troppo esigua per la produzione di stime con i metodi diretti.

I metodi diretti

In generale per la stima di un totale si devono eseguire le due seguenti operazioni:

1. determinare il peso da attribuire a ciascuna unità inclusa nel campione;
2. calcolare la stima dei parametri di interesse come somma ponderata dei valori relativi ad una data variabile oggetto di indagine con i pesi determinati al punto 1.

Il peso da attribuire a ciascuna unità è ottenuto in base ad una procedura articolata in più fasi:

1. il *peso iniziale* di ciascuna unità campionaria, definito *peso diretto*, è calcolato in funzione del disegno di campionamento adottato, come reciproco della probabilità di inclusione;
2. il peso iniziale viene corretto in modo da correggere la mancata risposta totale, ottenendo il *peso base*;
3. sono calcolati fattori correttivi del peso base per tenere conto dei vincoli di uguaglianza tra alcuni parametri noti della popolazione e le corrispondenti stime campionarie;
4. il *peso finale* è ottenuto come prodotto tra il peso base e i fattori correttivi.

La classe degli stimatori corrispondente alle operazioni appena descritte è nota come *stimatori di calibrazione* o di *ponderazione vincolata*, in quanto sia la correzione del peso per correggere la mancata risposta totale che la correzione del peso per ottenere la coerenza con parametri della popolazione noti si ottiene risolvendo un problema di minimo vincolato. Più precisamente ciò che si vuole rendere minimo è la distanza tra il peso prima e dopo la fase di calibrazione.

Per quanto riguarda la scelta del metodo di stima il problema principale è quello di individuare uno stimatore che risponda a:

- criteri di efficienza delle stime in termini di bassa varianza campionaria e riduzione della distorsione dovuta alla presenza di *mancate risposte totali e parziali* e di sotto-copertura delle liste di estrazione del campione rispetto alle popolazioni oggetto di indagine;
- criteri di *coerenza* esterna ed interna delle stime. Il problema della coerenza esterna delle stime nasce ogniqualvolta si dispone, da fonti esterne, di totali noti aggiornati sulla popolazione oggetto di indagine. Le stime dei totali prodotte dall'indagine devono in generale coincidere o non discostarsi molto dal valore noto di tali totali. La *coerenza interna* delle stime si ottiene quando tutte le stime (prodotte dall'indagine) di uno stesso

aggregato coincidono tra loro. Questo risultato si può ottenere utilizzando un unico sistema di pesi per il riporto dei dati all'universo.

I metodi di stima basati sulla teoria degli stimatori di ponderazione vincolata soddisfano i suddetti criteri in quanto:

- conducono, generalmente, a stime più efficienti di quelle ottenibili con gli stimatori diretti; l'efficienza è tanto maggiore quanto più alta è la correlazione tra le variabili ausiliarie e le variabili oggetto di indagine;
- sono approssimativamente non distorti rispetto al disegno di campionamento;
- portano a stime dei totali che coincidono con i valori noti di tali totali;
- attenuano l'effetto distorsivo dovuto alla presenza di mancate risposte totali;
- riducono l'effetto distorsivo dovuto alla sotto-copertura della lista da cui è selezionato il campione.

Gli stimatori di ponderazione vincolata sono utilizzati per il calcolo dei coefficienti di riporto all'universo della maggior parte delle indagini campionarie dell'Istat sulla popolazione e sulle imprese.

I metodi indiretti

I metodi di stima indiretta sono utilizzati dall'Istat per dare una risposta concreta alla crescente necessità da parte delle Amministrazioni Locali di ottenere informazioni accurate e riferibili ad aree geografiche, o più in generale a domini, di piccole dimensioni, denominate piccole aree. Le indagini campionarie condotte dall'Istat sono, tuttavia, progettate per fornire informazioni attendibili per i principali aggregati di interesse per domini di stima definiti in fase di progettazione del disegno campionario e non può essere in grado di rispondere in maniera idonea ad obiettivi di stima a livello di dettaglio maggiore.

La soluzione adottata in passato dall'Istat per ottenere stime a livello di dominio non pianificato, è stata quella di aumentare la numerosità delle unità campionarie senza modificare la strategia di campionamento adottata, ossia senza modificare né il disegno di campionamento né lo stimatore utilizzato. Il sovra-campionamento comporta tuttavia sia l'aumento dei costi e degli adempimenti operativi a carico della rete di rilevazione che l'incremento degli errori non campionari dovuto alla difficoltà di tenere sotto controllo indagini basate su campioni troppo ampi. Inoltre, il sovra-campionamento costituisce una soluzione parziale al problema di stima per piccole aree, in quanto non potendo aumentare la dimensione del campione oltre un certo limite rende possibile fornire stime attendibili soltanto per un sottoinsieme delle piccole aree di interesse.

Per tali ragioni l'Istat, fa ricorso a metodi di stima indiretti che si basano:

- sull'utilizzo di informazioni ausiliarie, correlate ai fenomeni oggetto di studio, note a livello delle piccole aree di interesse;
- sull'adozione (implicita o esplicita) di modelli statistici che legano i valori della variabile di interesse a livello di piccola area con i valori della medesima variabile relativi a un'area più grande (macroarea) contenente la piccola area di interesse e/o relativi ad altre occasioni di indagine oltre a quella corrente.

Un problema fondamentale di tali metodi è quello legato al fatto che essi si basano su modelli e pertanto le proprietà dei risultati ottenuti sono legate alla validità del modello ipotizzato. Poiché

una perfetta aderenza del modello alla realtà non è mai verificata tali stimatori sono soggetti a distorsioni non misurabili che introducono forti interrogativi sulla loro utilizzazione nei casi concreti.

Valutazione degli errori campionari

Per la valutazione degli errori campionari delle stime prodotte dalle indagini Istat si fa generalmente ricorso a metodi di calcolo della varianza approssimati. Infatti, per la maggior parte delle procedure di stima impiegate non è disponibile un'espressione analitica dello stimatore della varianza, in quanto:

- le indagini Istat vengono realizzate attraverso disegni di campionamento complessi, in generale basati su più stadi di selezione, sulla stratificazione delle unità e sulla selezione delle stesse con probabilità variabili e senza ripetizione;
- le stime vengono determinate mediante l'utilizzo degli stimatori di ponderazione vincolata i quali sono funzioni non lineari delle informazioni campionarie.

I metodi di stima della varianza campionaria utilizzati generalmente in Istat sono basati sul metodo di linearizzazione di Woodruff (1971) che consente di stimare la varianza campionaria nel caso in cui gli stimatori adottati sono funzioni non lineari dei dati campionari.

Sulla base della suddetta metodologia l'Istat ha sviluppato i software generalizzati GENESEES e ReGenesees, che dispongono di un'interfaccia user friendly e sono correntemente utilizzati per la stima degli errori campionari delle stime prodotte dalle diverse indagini Istat.

Inoltre, mediante tali software, vengono calcolate importanti statistiche che consentono di effettuare un'analisi critica del disegno di campionamento adottato. In particolare è possibile valutare:

- l'efficienza complessiva del disegno di campionamento utilizzato, attraverso il rapporto tra la varianza del campione complesso utilizzato e quella di un ipotetico campione casuale semplice di pari numerosità in termini di unità finali di campionamento;
- l'impatto sull'efficienza delle stime dovuto alla stratificazione delle unità, alla definizione degli stadi di campionamento e alla ponderazione delle unità (effetto della stratificazione, effetto del disegno a più stadi).

È importante fare presente che l'Istat effettua una presentazione sintetica degli errori di campionamento mediante modelli regressivi che mettono in relazione i valori delle stime con i corrispondenti errori campionari. Tali modelli sono utilizzati per corredare le tavole pubblicate con importanti informazioni sintetiche sugli errori campionari.

Fase ELABORAZIONE – METODI

- **INTEGRAZIONE**

MEMOBUST – Handbook on Methodology of Modern Business Statistics
Micro-Fusion
2014
MEETS ESSnet MEMOBUST

State of the art on statistical methodologies for data integration
Methodological developments
2011
ESSnet on Data Integration

Old and new approaches in statistical matching when samples are drawn with complex survey designs
2010
in Proceedings of the SIS Conference, Padua

ISAD Work packages and executive summary
2008
ESSnet Statistical Methodology – Area ISAD (Integration of Survey and Administrative Data)

Metodi statistici per il record linkage
2005
Collana Metodi e Norme, n. 16, Istat

Riferimenti

D'Orazio M., M. Di Zio, M. Scanu. 2006. Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints. *JOS*, 22(1):137-157.

D'Orazio M., M. Di Zio, M. Scanu. 2006. *Statistical Matching: Theory and Practice*. J. Wiley & Sons, Chichester.

- **CODIFICA DELLE RISPOSTE TESTUALI**

Metodi e software per la codifica automatica e assistita dei dati
2007
Collana Tecniche e strumenti, n.4, Istat

Riferimenti

Istat. 2008. *L'ambiente di codifica automatica dell'ATECO 2007. Esperienze effettuate e prospettive*. Collana Metodi e Norme, n. 41, Istat.

Istat. 2005. *Una soluzione per la rilevazione e codifica della Professione nelle indagini CATI*. Collana Contributi Istat, n. 11, Istat.

Macchia S., M. D'Orazio. 2001. A system to monitor the quality of automated coding of textual answers to open questions. *Research in Official Statistics*, 4(2).

- **INDIVIDUAZIONE E TRATTAMENTO DEGLI ERRORI DI MISURA E DELLE MANCATE RISPOSTE PARZIALI**

MEMOBUST – Handbook on Methodology of Modern Business Statistics
Statistical Data Editing
Imputation
2014
MEETS ESSnet MEMOBUST

A Contamination Model for Selective Editing
2013
Journal of Official Statistics. Volume 29, Issue 4, Pages 539-555.

Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys
2007
European project EDIMBUS

Results of the EUREDIT project
Euredit – The Development and Evaluation of New Methods for Editing and Imputation
2003
European project EUREDIT

Riferimenti

De Wall T., J. Pannekoek and S. Sholtus. 2011. *Handbook of Statistical Data Editing and Imputation*. J. Wiley & Sons, Hoboken, N.J.

Di Zio M., U. Guarnera and R. Rocci. 2007. A mixture of mixture models for a classification problem: the unity measure error. *Computational Statistics and Data Analysis*, 51:2573-2585.

Di Zio M., U. Guarnera and O. Luzi. 2005. Editing Systematic Unity Measure Errors Through Mixture Modelling. *Survey Methodology*, 31(1):53-63.

Pannekoek J. and T. De Waal. 2005. Automatic Edit and Imputation for Business Surveys: The Dutch Contribution to the EUREDIT Project. *Journal of Official Statistics*, 21(2):257-286.

Särndal C. E. and S. Lundström. 2005. *Estimation in Surveys with Nonresponse*. J. Wiley & Sons, New York.

Chambers R., A. Hentges and X. Zhao. 2004. Robust automatic methods for outlier and error detection. *Journal of the Royal Statistical Society*, Series A, 167(2):323-339.

Little J. and D. Rubin. 2002. *Statistical Analysis with Missing Data*. J. Wiley & Sons, New York.

Chen J. and J. Shao. 2000. Nearest Neighbor Imputation for Survey Data. *Journal of Official Statistics*, 16(2):113-131.

Schafer J. L. 2000. *Analysis of Incomplete Multivariate Data*. Chapmann and Hall/CRC, New York.

Latouche M. and J.M. Berthelot. 1992. Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys. *Journal of Official Statistics*, 8(3):389-400.

Little R.J.A. 1988. Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3):287-296.

Rubin D. 1987. *Multiple Imputation for Nonresponse in Surveys*. J. Wiley & Sons, New York.

Hidiroglou M. A. and J.M. Berthelot. 1986. Statistical editing and imputation for periodic business surveys. *Survey Methodology*, 12(1):73-83.

Kalton G. and D. Kasprzyk. 1982. Imputing for missing survey responses. In *Proceedings of the section on Survey Research Methods*, American Statistical Association.

Fellegi P. I. and D. Holt. 1976. A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, Applications Section, 71:17-35.

- **CALCOLO DELLE STIME E DEGLI ERRORI CAMPIONARI**

MEMOBUST – Handbook on Methodology of Modern Business Statistics
Weighting and Estimation
2014
MEETS ESSnet MEMOBUST

Results of the SAE project
SAE – Small Area Estimation
2012
ESSnet SAE

Riponderazione
2005
Note metodologiche, Istat

Stime ed Errori
2005
Note metodologiche, Istat

Riferimenti

Istat. 2008. Strategia di campionamento e precisione delle stime. In "*L'indagine europea sui redditi e le condizioni di vita delle famiglie (Eu-Silc)*", Collana Metodi e Norme, n. 37, Istat.

Istat. 2006. La procedura di stima e la valutazione degli errori campionari. In "*Il sistema di indagini sociali multiscopo. Contenuti e metodologia delle indagini*", Collana Metodi e Norme, n. 31, Istat.

Istat. 2006. Strategia di campionamento e livello di precisione delle stime. in "*L'indagine campionaria sulle nascite: obiettivi, metodologia e organizzazione*", Collana Metodi e Norme, n. 28, Istat.

Rao, J. N. K. 2003. *Small Area Estimation*. J. Wiley & Sons, New York.

Cicchitelli G., A. Herzel, G.E. Montanari. 1992. *Il campionamento statistico.* Il Mulino, Bologna.

Deville J.C., C.E. Särndal. 1992. Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87:376-382.

Särndal C.E., B. Swensson, J. Wretman. 1992. *Model Assisted Survey Sampling.* Springer-Verlag, New York.

Cochran W.G. 1977. *Sampling Techniques.* J. Wiley & Sons, New York.

Fase ELABORAZIONE – STRUMENTI

- **INTEGRAZIONE**

RELAIS (REcord Linkage At IStat)

Toolkit per risolvere problemi di record linkage.

StatMatch
Pacchetto R che rende disponibili alcune funzioni per l'integrazione dei dati attraverso lo statistical matching e, come prodotto secondario, la possibilità di imputare i valori mancanti in un data set.

- **CODIFICA DELLE RISPOSTE TESTUALI**

CIRCE (Comprehensive Istat R Coding Environment)

Sistema per la codifica automatica di quesiti rilevati a testo libero.

- **INDIVIDUAZIONE E TRATTAMENTO DEGLI ERRORI DI MISURA E DELLE MANCATE RISPOSTE PARZIALI**

Banff. Software per il controllo e la correzione dei dati (imputazione) per le variabili numeriche e continue.

CANCEIS (CANadian Census Edit and Imputation System). Software per il controllo e la correzione dei dati (imputazione). Basato sulla Nearest-neighbour Imputation Methodology (NIM) è idoneo al trattamento di incoerenze tra i valori di variabili qualitative e quantitative rilevate su unità di rilevazione composte da una o più sotto-unità di livello gerarchico inferiore.

CONCORDJava (CONtrollo e CORrezione dei Dati versione con interfaccia Java). Sistema integrato per il controllo e la correzione dei dati (imputazione). Uno dei moduli (SCIA) implementa la Metodologia Fellegi-Holt per il trattamento di incoerenze tra i valori di variabili qualitative.

SeleMix (Selective editing via Mixture models). Pacchetto R per il trattamento di dati quantitativi, che si pone l'obiettivo di individuare un insieme di unità affette da errori potenzialmente influenti sulle stime di interesse (editing selettivo).

- **CALCOLO DELLE STIME E DEGLI ERRORI CAMPIONARI**

EVER (Estimation of Variance by Efficient Replication)

Package R dedicato al calcolo delle stime e degli errori di campionamento in indagini complesse, mediante metodi di replicazione del campione.

ReGenesees (R evolved Generalised software for sampling estimates and errors in surveys) Sistema software basato su R per l'analisi design-based e model-assisted di indagini campionarie complesse.

RELAIS (REcord Linkage At IStat)

Descrizione

RELAIS è un progetto *open source* avente come obiettivo la definizione di un toolkit per risolvere problemi di record linkage.

Le soluzioni a problemi di record linkage, studiate in letteratura e adottate nella pratica, si rifanno a svariati approcci e metodologie, che coinvolgono soluzioni euristiche, metodi probabilistici, approcci bayesiani, soluzioni basate sulle tecniche di data-mining o machine learning. Tuttavia nessuna delle metodologie o delle tecniche proposte finora per il record linkage ha dimostrato di essere la più efficace o la più efficiente per tutte le diverse applicazioni, anche a causa del fatto che i problemi di record linkage sono fortemente caratterizzati dalla natura dei dati da abbinare e dagli obiettivi dell'abbinamento.

Questa constatazione è alla base della filosofia di RELAIS, che è stato progettato e realizzato con l'intenzione di scomporre l'intero problema di record linkage attraverso l'individuazione delle sue fasi costituenti e di affrontare ciascuna di queste fasi con la tecnica più opportuna, in relazione agli obiettivi dell'applicazione del linkage e alla natura dei dati in esame.

Le principali fasi individuate in un processo di record linkage sono:

- Preparazione dei dati di input (pre-processing);
- Selezione degli attributi identificativi comuni (variabili di matching);
- Scelta della funzione di confronto;
- Riduzione dello spazio di ricerca delle coppie candidate all'abbinamento;
- Scelta del modello di decisione;
- Selezione degli abbinamenti univoci;
- Valutazione dei risultati del record linkage.

Per ciascuna delle fasi individuate sono note e largamente utilizzate tecniche diverse. In funzione della particolare applicazione e dei dati in esame, può essere opportuno iterare e/o omettere alcune fasi, così come preferire in ciascuna fase alcune tecniche rispetto ad altre.

Per ciascuna delle fasi individuate, RELAIS mette a a disposizione alcune tra le tecniche e i metodi più diffusi.

In particolare sono disponibili le seguenti funzionalità:

- Lettura di insiemi di dati da file in formato testuale.
- Metadati per scelta variabili bloccaggio.
- Metadati per scelta variabili di matching.
- Creazione dello spazio di ricerca del processo di linkage come prodotto cartesiano dei record degli insiemi di dati in input.
- Realizzazione del metodo di riduzione dello spazio di ricerca (dato dal prodotto cartesiano dei record di ciascun file coinvolto nel processo di linkage) denominato "blocking" mediante specifica di un'opportuna variabile di bloccaggio.

- Realizzazione del metodo di riduzione dello spazio di ricerca denominato “sorted neighborhood method” mediante specifica di un’opportuna variabile di ordinamento e della dimensione della finestra dei confronti.
- Realizzazione del metodo di riduzione dello spazio di ricerca denominato “nested blocking” mediante la combinazione dei metodi di riduzione “blocking” e “sorted neighborhood”.
- Funzioni di distanza.
- Modello deterministico esatto.
- Modello deterministico con regole e soglie.
- Modello probabilistico di Fellegi e Sunter implementato mediante l’algoritmo EM (Expectation-Maximization), sotto l’ipotesi di indipendenza condizionata delle variabili di matching.
- Modello probabilistico di Fellegi e Sunter, sotto l’ipotesi di indipendenza condizionata delle variabili, con l’acquisizione delle probabilità marginali da file esterno.
- Riduzione da matching N:M a matching 1:1.
- Riduzione 1:1 per modello deterministico.
- Euristiche non globali per riduzione1:1.
- Processamento dei blocchi con one-shot execution.
- Gestione ottimizzata output e residui.
- Gestione dei back-up.
- Gestione ottimizzata per processi di linkage orientati alla deduplicazione.
- Processamento in modalità batch.

Informazioni

Status:	validato
Autore:	Istat
Licenza:	EUPL-1.1
Codifica GSBPM:	5.1 Integrate data
Linguaggio di programmazione:	R, Java
Versione linguistica della GUI:	EN
Parole chiave:	integrazione dati, record linkage probabilistico, comparazione di stringhe, blocking/sorting/indexing, deduplicazione, open source software
Contatto:	nome: Luca Valentino email: luvalent@istat.it

Reperimento software e documentazione

REQUISITI TECNICI

RELAIS 3.x (REcord Linkage At IStat) richiede l'installazione degli ambienti Java, R e di MySQL. Con riferimento all'ambiente Java si richiede l'installazione di Java 2 Runtime Environment 6.0 o superiore; con riferimento all'ambiente R è necessaria l'installazione della versione 2.5.1 o superiore, è infine necessario installare i package R lpSolve (versione 5.5 o superiore) e RODBC; con riferimento all'ambiente MySQL oltre all'installazione di MySQL Server si richiede l'installazione di MySQL ODBC 5.x o superiore.

COPYRIGHT

DISCLAIMER

L'Istat non si assume la responsabilità per risultati derivanti da un uso dello strumento non coerente con le indicazioni metodologiche contenute nella documentazione disponibile.

DOWNLOAD

RELAIS versione 3.0

DOCUMENTAZIONE TECNICA E METODOLOGICA

Manuale utente - RELAIS v. 3.0

ALTRA DOCUMENTAZIONE

Cibella N., G.L. Fernandez, M. Guigò, F. Hernandez, M. Scannapieco, L. Tosco, T. Tuoto. 2009. Sharing Solutions for Record Linkage: the RELAIS Software and the Italian and Spanish Experiences. In A*tti della conferenza NTTS (New Techniques and Technologies for Statistics)*, Eurostat, Brussels, 18-20 febbraio 2009.

Eurostat. 2009. Theory and practice of developing a record linkage software. In "*Insights on Data Integration Methodologies. ESSnet-ISAD workshop, Vienna, 29-30 maggio 2008*". Methodologies and working papers, Eurostat.

Cibella N., M. Fortini, M. Scannapieco, L. Tosco, T. Tuoto. 2007. RELAIS: Don't Get Lost in a Record Linkage Project. In A*tti della Conferenza FCSM 2007,* Federal Committee on Statistical Methodology, Arlington, 5-7 novembre 2007.

Fortini M., P.D. Falorsi, C. Vaccari, N. Cibella, T. Tuoto, M. Scannapieco, L. Tosco. 2006. Towards an Open Source Toolkit for Building Record Linkage Workflows. In A*tti del workshop internazionale IQIS,* Chicago, 30 giugno 2006.

StatMatch

Descrizione

StatMatch è un package aggiuntivo per l'ambiente R che rende disponibili agli utenti R alcune funzioni per l'integrazione dei dati attraverso lo statistical matching e, come prodotto secondario, la possibilità di imputare i valori mancanti in un data set.

Il package contiene sia funzioni che implementano metodi di matching che funzioni di supporto al matching (calcolo delle distanze, ecc.). Ci sono ben tre funzioni dedicate all'applicazione di metodi nonparametrici di matching a livello micro:

- NND.hotdeck: selezione del donatore di distanza minima; implementa numerose funzioni di distanza; permette la definizione di classi di donazione. E' possibile imporre il vincolo di utilizzare il donatore una sola volta (matching constained)
- RNDwNND.hotdeck: selezione casuale del donatore in classi fisse o "mobili". In quest'ultimo caso si può selezionare un donatore a caso tra i k più vicini; scelta a caso di un donatore con quelli a distanza inferiore di una certa soglia, ecc. La selezione del donatore può avvenire con probabilità variabili specificando la variabile contenente i pesi. Sono implementate diverse funzioni di distanza.
- rankNND.hotdeck: selezione del donatore più vicino basandosi sulla distanza calcolata tra i percentili della distribuzione empirica cumulata della variabile continua presente in entrambi i data set. Nel calcolo della distribuzione empirica cumulata è possibile tener conto di un peso diverso da assegnare alle unità. La distribuzione empirica cumulata può essere calcolata in opportune classi di unità.

Queste funzioni possono essere utilizzate per imputare i valori mancanti in un data set attraverso i corrispondenti metodi hotdeck.

Solo la funzione mixed.mtc permette di implementare metodi di matching a livello parametrico macro o misto. La funzione assume che le variabili *X*, *Y* e *Z* si distribuiscano secondo una distribuzione normale multivariata. La stima dei parametri della normale può essere condotta con metodo della massima verosimiglianza o con un metodo basato sulle stime campionarie delle quantità di interesse (medie e varianze).

Due sono le funzioni dedicate alla applicazione di metodi di matching a livello macro in presenza di dati provenienti da indagini campionarie complesse. Tali funzioni si basano su una serie di calibrazioni dei pesi campionari associati alle unità nei data set di origine, secondo le metodologie suggerite da Renssen (1998). Al momento l'applicazione di tali metodi è limitata al caso in cui *Y* e *Z* siano entrambe categoriali e l'obiettivo della stima è la tabella di contingenza *Y* vs. *Z*. In particolare la funzione harmonize.x armonizza la distribuzione marginale/congiunta delle prescelte variabili X in modo che essa sia coerente tra i due data set di origine; successivamente la funzione comb.samples procede a stimare la tabella *Y* vs. *Z* attraverso le metodologie proposte da Renssen sia in assenza di altri fonti dati che in presenza di una ulteriore fonte dati ausiliaria C in cui siano osservate congiuntamente *Y* e *Z* o *X*, *Y* e *Z*.

Infine le funzioni Frechet.bounds.cat e Fbwidths.by.x permettono l'esplorazione dell'incertezza quando tutte le variabili (*X*, *Y* e *Z*) sono categoriali. La prima funzione stima gli intervalli di incertezza per tutte le celle della tabella di contingenza *Y* vs. *Z*. La funzione Fbwidths.by.x è utile quando si è in presenza di numerose variabili comuni X e permette di individuare quali sono le

variabili maggiormente legate alle variabili di interesse *Y* e *Z* e che, in quanto tali, permettono una riduzione dell'ampiezza degli intervalli di incertezza.

Informazioni

Status:	validato
Autore:	Istat
Licenza:	GPL-2 | GPL-3
Codifica GSBPM:	5.1 Integrate data 5.4 Edit and impute
Linguaggio di programmazione:	R
Parole chiave:	statistical matching, data fusion, imputazione hot deck, analisi incertezza
Contatto:	nome: Marcello D'Orazio email: madorazi@istat.it

Reperimento software e documentazione

REQUISITI TECNICI

Il package Statmatch funziona su versioni di R a partire 2.7.0 su qualsiasi sistema operativo (Windows, Mac o Linux) . Richiede che vengano installati e quindi caricati i seguenti package aggiuntivi R: proxy, lpSolve, survey. In alcuni casi, si può rendere necessario disporre degli ulteriori package: optmatch, SDaA, simPopulation, MASS; si fa presente che l'utilizzo del package optmatch è soggetto ad alcune limitazioni.

COPYRIGHT

DISCLAIMER

L'Istat non si assume la responsabilità per risultati derivanti da un uso dello strumento non coerente con le indicazioni metodologiche contenute nella documentazione disponibile.

DOWNLOAD

- Versione 1.2.5 – Package precompilato: Sistemi Windows
- Versione 1.2.5 – Sorgenti del package: Sistemi Windows e Unix-like

DOCUMENTAZIONE TECNICA E METODOLOGICA

Reference manual – StatMatch v. 1.2.5

Vignettes – StatMatch v. 1.2.5

CIRCE (Comprehensive Istat R Coding Environment)

Descrizione

CIRCE, sviluppato in Istat, è un pacchetto software basato su R che ha come scopo l'attribuzione automatica di un codice a partire da un testo. E' un prodotto generalizzato, ossia indipendente dalla classificazione considerata e dalla lingua in cui sono espressi i testi. CIRCE sostituisce Actr v3, adottato in Istat sin dal 1998, ma non più manutenuto dall'Istituto canadese e non più compatibile con le nuove piattaforme software (Windows 7, Windows Server 2008) usate in Istat.

CIRCE ricalca l'algoritmo di *matching* di ACTR v3. Questa scelta è stata dettata dall'esigenza di garantire agli utenti gli stessi livelli di qualità della codifica raggiunti con il precedente sistema ampiamente utilizzato in Istituto.

CIRCE, essendo sviluppato in R, è portabile su diversi ambienti senza necessità di compilazione. Questo ha permesso di realizzare un unico pacchetto di codifica funzionante sia in ambiente Windows che Linux. CIRCE è quindi utilizzabile sia in ambiente pc, attraverso un'interfaccia grafica utente, che in ambiente web, attraverso la "chiamata" ad un web service.

Rientra tra i sistemi basati sui *weighting algorithms*.

Gestisce applicazioni di:

- codifica automatica, ossia codifica di interi file (modalità *batch*);
- codifica interattiva che, con l'ausilio dell'interfaccia grafica, permette di analizzare interattivamente la codifica dei casi singoli;
- codifica web, ossia web service per la codifica di singole stringhe. In quest'ultimo caso è attualmente disponibile un web service dedicato alla codifica dell'Ateco accessibile attraverso la pagina: http://www.istat.it/it/strumenti/definizioni-e-classificazioni/ateco-2007.

A prescindere dal tipo di codifica, il confronto tra la risposta da codificare e le voci contenute nel dizionario informatizzato è preceduto dalla fase di standardizzazione dei testi definita *parsing*. Tale fase è completamente controllata dall'utente che ha il compito di adattarla al particolare contesto applicativo. Lo scopo della fase di *parsing* è quello di rimuovere differenze grammaticali o sintattiche al fine di rendere uguali due descrizioni diverse, ma dallo stesso contenuto semantico. CIRCE mette a disposizione (ad oggi) un set di 14 diverse funzioni di *parsing*, tra le quali: la mappatura dei caratteri, l'eliminazione delle parole o delle stringhe ritenute ininfluenti, l'eliminazione dei prefissi e dei suffissi, il trattamento dei sinonimi.

Successiva alla standardizzazione dei testi è la fase di *matching*. Il testo standardizzato viene confrontato con le descrizioni, anch'esse standardizzate, del dizionario di riferimento. Se il risultato del confronto è un *match* diretto (*direct match*) allora il software assegna un codice univoco. In caso contrario, viene utilizzato un algoritmo basato sui pesi delle parole, per individuare il miglior *match* parziale, fornendo così un *indirect match*.

Essendo un prodotto sviluppato internamente all'Istituto, offre l'opportunità di modifiche e/o aggiunte di nuove funzionalità, sia relative al set di funzioni di *parsing* che all'algoritmo di *matching*.

Informazioni

Status:	validato
Autore:	Istat
Licenza:	EUPL-1.1
Codifica GSBPM:	5.2. Classify and code
Linguaggio di programmazione:	R, VB.NET
Versione linguistica della GUI:	IT
Parole chiave:	codifica automatica, algoritmi codifica pesati
Contatto:	nome: Laura Capparucci email: capparuc@istat.it

Reperimento software e documentazione

REQUISITI TECNICI

- Ambiente R versione 3.1.1 o superiore.

- Windows 7 o superiore.

- Microsoft Framework .net 4 (solo per l'interfaccia utente di tipo grafico).

COPYRIGHT

DISCLAIMER

L'Istat non si assume la responsabilità per risultati derivanti da un uso dello strumento non coerente con le indicazioni metodologiche contenute nella documentazione disponibile.

DOWNLOAD

CIRCE versione 1.0

DOCUMENTAZIONE TECNICA E METODOLOGICA

Manuale utente - CIRCE v. 1.0

ALTRA DOCUMENTAZIONE

Istat. 2007. *Metodi e software per la codifica automatica dei dati*. Collana Tecniche e strumenti, n. 4, Istat.

Istat. 2005. *La codifica delle variabili testuali nel 14° Censimento Generale della Popolazione*. Collana Documenti Istat, n. 1, Istat.

Macchia S., M. D'Orazio. 2001. A system to monitor the quality of automated coding of textual answers to open questions. *Research in Official Statistics*, 4(2).

De Angelis R., S. Macchia, L. Mazza. 2000. Applicazioni sperimentali della codifica automatica: analisi di qualità e confronto con la codifica manuale. *Quaderni di ricerca – Rivista di statistica Ufficiale*, 1.

Banff

Descrizione

Banff è un software generalizzato per il *controllo e la correzione dei dati (imputazione) per le variabili numeriche e continue.*

Banff è stato sviluppato da Statistics Canada in ambiente SAS ed è strutturato secondo la filosofia SAS delle procedure (proc). Per il controllo degli errori nei dati si avvale di regole di consistenza *(edit rules)* che devono essere espresse in forma lineare.

Banff ha una struttura modulare: ogni modulo corrisponde ad una particolare sotto-funzione della struttura generale di un processo di controllo e correzione dati di variabili quantitative:

- definizione dei dati;
- definizione delle regole di consistenza;
- verifica della coerenza delle regole di consistenza;
- localizzazione degli errori;
- identificazione dei valori anomali;
- imputazione.

La localizzazione degli errori è fatta tramite *l'algoritmo di Chernikova* basato sul *paradigma di Fellegi-Holt*, ovvero sul *criterio del minimo cambiamento.*

In generale il paradigma di Fellegi-Holt è ritenuto appropriato per trattare errori di tipo stocastico. Per ogni record che fallisce almeno una regola di consistenza, l'algoritmo identifica il minimo numero di campi da cambiare (imputare) affinchè il record passi tutte le regole.

Per quanto riguarda l'imputazione, Banff implementa diversi metodi:

- Imputazione deduttiva
  Verifica se esiste uno ed un solo valore che, una volta assegnato al campo da imputare, fa si che il record soddisfi tutte le regole di consistenza.
- Donatore di minima distanza
  Viene scelta l'osservazione più vicina all'unità da imputare tra i potenziali donatori, i.e. unità che soddisfano tutte le regole. È importante sottolineare che un potenziale donatore sarà scelto effettivamente come donatore, se il valore imputato farà si che il ricevente passi tutti i vincoli. In altri termini i record imputati tramite il donatore di minima distanza soddisferanno un insieme di regole determinate dall'utilizzatore del software. L'imputazione è congiunta, ovvero una volta che un donatore è stato scelto, tutti i campi da imputare del ricevente saranno riempiti con i valori del donatore stesso.
- Stimatori
  Banff implementa una serie metodi chiamati in senso lato stimatori. Tali metodi vanno dalla sostituzione dei valori mancanti con la media calcolata sui valori osservati, alla predizione dei valori mancanti tramite la regressione.

Il software Banff dà inoltre la possibilità di effettuare altre analisi che possono essere utili per capire e studiare l'impatto del piano di controllo e correzione sui dati (ad esempio: la lista delle regole ridondanti, la frequenza di fallimento degli edit per record ecc.).

Informazioni

Status:	in dismissione
Autore:	Statistics Canada
Codifica GSBPM:	5.3 Review and validate 5.4 Edit and impute
Parole chiave:	editing per variabili numeriche, localizzazione degli errori, principio del minimo cambiamento, donatore di minima distanza

Reperimento software e documentazione

Per il reperimento del software e della documentazione tecnica e metodologica è possibile rivolgersi a Statistics Canada.

Solo per i dipendenti Istat: rivolgersi a Francesco Dell'Orco.

CANCEIS (CANadian Census Edit and Imputation System)

Descrizione

CANCEIS è un software generalizzato per il controllo e la correzione di **variabili qualitative e quantitative**.

È stato sviluppato da Statistics Canada per il controllo e l'imputazione dei dati censuari.

Utilizza la metodologia di imputazione Nearest-neighbour (Nearest-neighbour Imputation Methodology – NIM, precedentemente nota come New Imputation Methodology). L'imputazione dei valori è completamente guidata dai dati disponibili e i valori imputati sono prelevati da un unico record donatore.

Il sistema è stato progettato per individuare donatori per l'intera famiglia (la ricerca è circoscritta a famiglie di uguale dimensione) ed è pertanto idoneo al trattamento di dati con struttura gerarchica. Per ciascuna famiglia errata il sistema prima individua i donatori e poi determina il minimo numero di variabili da imputare, sulla base dei donatori individuati, garantendo azioni di imputazione coerenti con un predefinito insieme di regole di controllo.

Informazioni

Status:	validato
Autore:	Statistics Canada
Codifica GSBPM:	5.3 Review and validate 5.4 Edit and impute
Parole chiave:	editing, imputazione, donatore di minima distanza, variabili qualitative e quantitative

Reperimento software e documentazione

Per il reperimento del software e della documentazione tecnica e metodologica è possibile rivolgersi a Statistics Canada.

ALTRA DOCUMENTAZIONE

Istat. 2007. Indagine sulle Cause di Morte: Nuova procedura automatica per il controllo e la correzione delle variabili demo-sociali. Collana Documenti Istat, n. 6, Istat.

Manzari A., A. Reale. 2001. Towards a new system for edit and imputation of the 2001 Italian Population Census data: A comparison with the Canadian Nearest-neighbour Imputation Methodology. In *ISI World Statistics Congress Proceedings 53rd Session,* International Statistical Institute, Seoul, 2001.

CONCORDJava (CONtrollo e CORrezione dei Dati versione con interfaccia Java)

Descrizione

CONCORDJava è un software *open source* per il controllo e correzione dei dati. L'applicazione integra software precedentemente sviluppati ed utilizzati in Istat: SCIA, RIDA e GRANADA.

L'applicazione, attualmente rilasciata in versione beta per la parte riguardante le correzioni deterministiche, è disponibile per il download nella versione in lingua italiana ed in inglese.

I diversi metodi residenti nel software sono implementati in moduli distinti:

- **SCIA** (Sistema di Controllo e Imputazione Automatici):
  Esegue il controllo e la correzione di variabili qualitative applicando integralmente la metodologia di Fellegi-Holt. Per ciascun record errato il sistema prima individua il minimo numero di variabili da imputare e poi effettua l'imputazione garantendo azioni di imputazione coerenti con un predefinito insieme di regole di controllo.

- **RIDA** (Ricostruzione dell'Informazione con Donazione Automatica):
  Esegue l'imputazione di variabile qualitative e quantitative mediante donatore di minima distanza. Operazioni propedeutiche sono:
  - la classificazione delle unità in esatte ed errate;
  - la loro registrazione in due file distinti;
  - la identificazione dei valori da imputare mediante un predefinito carattere (di errore).

- **GRANADA** (Gestione delle Regole per l'ANAlisi dei DAti):
  Esegue l'imputazione di variabili qualitative e quantitative secondo l'approccio deterministico, ossia mediante l'applicazione di regole del tipo SE [condizione di errore] ALLORA [azione di correzione]. Mediante questo modulo è possibile eseguire anche il solo controllo dei dati (separazione in esatti ed errati) secondo regole di incompatibilità che ammettono operatori logici e aritmetici (e quindi valide per variabili qualitative e quantitative).

Propedeutica ai vari passi è la fase di definizione delle variabili, cioè dei campi del record da sottoporre a controllo, e degli edit o regole di controllo sia formali che sostanziali individuabili a partire dal questionario e dalla conoscenza relativa ai fenomeni indagati.

Informazioni

Status:	validato
Autore:	Istat
Licenza:	EUPL-1.1
Codifica GSBPM:	5.3 Review and validate 5.4 Edit and impute
Linguaggio di programmazione:	Fortran, Java
Versione linguistica della GUI:	EN, IT
Parole chiave:	localizzazione, imputazione, donatore di minima distanza, Fellegi-Holt
Contatto:	nome: Maria Teresa Buglielli email: bugliell@istat.it

Reperimento software e documentazione

REQUISITI TECNICI

dimensione minima hardware
256 Mb memoria RAM
30 Mb su disco C:/

software necessari
Java 2 Runtime Environment 6.0 o superiore

COPYRIGHT

DISCLAIMER

L'Istat non si assume la responsabilità per risultati derivanti da un uso dello strumento non coerente con le indicazioni metodologiche contenute nella documentazione disponibile.

DOWNLOAD

CONCORDJAVA versione 2.2

DOCUMENTAZIONE TECNICA E METODOLOGICA

Manuale utente – CONCORDJava v. 2.2

SeleMix (Selective editing via Mixture models)

Descrizione

SeleMix è un pacchetto R per il trattamento di dati quantitativi, che si pone l'obiettivo di individuare un insieme di unità affette da errori potenzialmente influenti sulle stime di interesse (editing selettivo).

La metodologia sottostante si basa su particolari modelli a classi latenti noti in letteratura come modelli di contaminazione. Specificamente, si assume che i dati "veri" (cioè non affetti da errori), eventualmente in scala logaritmica, siano realizzazioni indipendenti di una distribuzione Gaussiana multivariata, con vettore delle medie che può a sua volta essere espresso come combinazione lineare di un insieme di covariate non contaminate. La natura "intermittente" del meccanismo di errore è catturata da variabili Bernoulliane che hanno il ruolo di indicatori per l'occorrenza di errore su ciascuna unità. Inoltre l'errore è supposto additivo e associato a un vettore Gaussiano a media nulla e matrice di varianza e covarianza proporzionale alla matrice di varianza e covarianza che caratterizza la distribuzione dei dati senza errori. La modellizzazione esplicita della distribuzione dei dati non contaminati e del meccanismo di errore consentono di ricavare la distribuzione dei dati veri condizionatamente ai dati osservati. Sulla base di quest'ultima distribuzione vengono effettuate le previsioni dei valori veri non osservati, e quindi degli errori. Per ciascuna unità, è calcolato un punteggio (*score*) in termini della differenza (eventualmente ponderata col peso campionario) tra valore previsto e valore osservato. Tutte le unità sono quindi ordinate (in modo decrescente) in accordo al proprio punteggio. Supponendo che il parametro di interesse sia una media o un totale di popolazione, la selezione delle osservazioni da sottoporre a revisione interattiva è effettuata considerando la stima dell'errore che rimane nei dati al netto delle unità revisionate. Il numero di unità selezionate secondo tale criterio dipende inoltre da una soglia specificata dall'utente che è legata all'accuratezza della stima che si vuole ottenere.

Sono descritte di seguito le principali funzioni del pacchetto SeleMix:

- **ml.est:** effettua le stime di massima verosimiglianza dei parametri del modello di contaminazione mediante algoritmo ECM e fornisce i valori previsti dei dati "veri" per tutte le unità che sono state usate per la stima. Ritorna anche, per ciascuna unità, le probabilità a posteriori di occorrenza dell'errore e i flag di classificazione *outlier – non outlier* calcolati in base ad una soglia per la probabilità di errore specificata dall'utente. Richiede la specificazione del tipo di modello assunto per i dati veri (normale o lognormale) e alcuni parametri tecnici per l'algoritmo ECM.
- **pred.y**: sulla base di un insieme di parametri del modello di contaminazione, e di un insieme di dati osservati, calcola i valori previsti dei corrispondenti dati veri. Sono ammessi anche valori mancanti per le variabili risposta, ma non per le covariate.
- **sel.edit**: effettua l'Editing Selettivo. Sulla base di un insieme di dati osservati e delle corrispondenti previsioni per i dati veri seleziona le unità da sottoporre a editing interattivo. Richiede in input la soglia di accuratezza desiderata e, se presenti, i pesi campionari associati alle unità. Fornisce il punteggio per ciascuna unità e il rank corrispondente.

Data la possibilità di utilizzare le funzioni del pacchetto anche in presenza di dati incompleti, il software può anche essere usato come strumento di imputazione robusta per dati Gaussiani multivariati.

Informazioni

Status:	validato
Autore:	Istat
Licenza:	EUPL-1.1
Codifica GSBPM:	5.3 Review and validate 5.4 Edit and impute
Linguaggio di programmazione:	R
Parole chiave:	Modelli a classi latenti, editing selettivo, errore influente
Contatto:	nome: Maria Teresa Buglielli email: bugliell@istat.it

Reperimento software e documentazione

COPYRIGHT

DISCLAIMER

L'Istat non si assume la responsabilità per risultati derivanti da un uso dello strumento non coerente con le indicazioni metodologiche contenute nella documentazione disponibile.

DOWNLOAD

- Versione 0.9.1 – Package precompilato: Sistemi Windows
- Versione 0.9.1 – Sorgenti del package: Sistemi Windows e Unix-like

DOCUMENTAZIONE TECNICA E METODOLOGICA

Reference manual – SeleMix v. 0.9.1

Vignettes – SeleMix v. 0.9.1

ALTRA DOCUMENTAZIONE

Barcaroli G., D. Zardetto. 2012. Use of R in Business Surveys at the Italian National Institute of Statistics: Experiences and Perspectives. In *Proceedings of the 4th International Conference of Establishment Surveys (ICES IV)*, American Statistical Association Montréal, 11-14 giugno 2012.

EVER (Estimation of Variance by Efficient Replication)

Descrizione

EVER è un software dedicato al calcolo delle stime e degli errori di campionamento in indagini complesse. Nella versione attuale EVER rende disponibili le seguenti funzionalità principali:

- Replicazione del campione
- Calibrazione dei dati replicati
- Calcolo delle stime, degli errori standard e degli intervalli di confidenza di:
    - Totali
    - Medie
    - Distribuzione di frequenza assoluta e relativa
    - Tabelle di contingenza
    - Rapporti tra totali
    - Quantili
    - Coefficienti di regressione
- Calcolo delle stime, degli errori standard e degli intervalli di confidenza di stimatori definiti dall'utente (arbitrari, anche *privi di una rappresentazione analitica*)
- Stime ed errori in sottopopolazioni.

La tecnica di stima della varianza campionaria implementata nel package EVER si basa sul metodo *DAGJK (Delete-A-Group Jackknife) esteso* proposto da Kott.

Il metodo DAGJK può essere visto come una variante computazionalmente efficiente del tradizionale metodo *jackknife stratificato*. La necessità di costruire una replica dei pesi originali per ogni PSU inclusa nel campione rende, di fatto, irrealistico il ricorso al metodo jackknife tradizionale per indagini "complesse e grandi" (decine di migliaia di PSU in strati numerosi e di dimensione molto variabile). L'utilizzabilità pratica del metodo DAGJK poggia, al contrario, sulla capacità del metodo di costruire – per una vasta gamma di stimatori e di disegni di campionamento – stime degli errori standard (quasi) non distorte anche con un piccolo numero (qualche decina) di repliche.

In aggiunta alla sua peculiare efficienza computazionale, il metodo DAGJK gode dei principali vantaggi comuni ai più diffusi metodi di replicazione del campione.

L'idea base di tutti i metodi di replicazione del campione consiste nello stimare la varianza campionaria di uno stimatore arbitrario mediante una adeguata misura della variabilità delle sue stime su repliche opportunamente costruite di un campione originario. Si tratta, dunque, di metodi intrinsecamente versatili, in grado, cioè, di fornire stime della varianza campionaria senza fare ricorso ad ipotesi restrittive sulla distribuzione dei dati della popolazione e/o sulla forma funzionale degli stimatori. Poiché tutto quello di cui necessitano è (i) la definizione della tecnica di replicazione e (ii) la definizione del metodo di calcolo dello stimatore su un campione, i metodi di replicazione si prestano, fra l'altro, a stimare la varianza di *stimatori privi di una rappresentazione analitica* (non esprimibili, cioè, come funzioni di valori direttamente osservabili sulle unità statistiche).

EVER è concepito per sfruttare appieno la versatilità del metodo di replicazione DAGJK: oltre a coprire gli stimatori di uso più comune nelle indagini campionarie su vasta scala, il package

fornisce, infatti, all'utente uno strumento amichevole per calcolare stime, errori standard ed intervalli di confidenza di stimatori arbitrari, definiti dall'utente medesimo. Questa funzionalità rende il package EVER particolarmente attraente in tutti i casi in cui il metodo di linearizzazione di Taylor per la stima della varianza campionaria sia applicabile solo al prezzo di forti approssimazioni (il problema della stima della povertà relativa è un possibile esempio).

Informazioni

Status:	validato
Autore:	Istat
Licenza:	EUPL-1.1
Codifica GSBPM:	5.6 Calculate weights 5.7 Calculate aggregates
Linguaggio di programmazione:	R
Parole chiave:	calibrazione, calcolo delle stime, replicazione del campione, indagini complesse, stimatori complessi, Delete-A-Group Jackknife
Contatto:	nome: Diego Zardetto email: zardetto@istat.it

Reperimento software e documentazione

REQUISITI TECNICI

Ambiente R versione 2.5.1 o superiore.

COPYRIGHT

DISCLAIMER

L'Istat non si assume la responsabilità per risultati derivanti da un uso dello strumento non coerente con le indicazioni metodologiche contenute nella documentazione disponibile.

DOWNLOAD

- Versione 1.2 – Package precompilato: Sistemi Windows
- Versione 1.2 – Sorgenti del package: Sistemi Windows e Unix-like

DOCUMENTAZIONE TECNICA E METODOLOGICA

Reference manual – EVER v. 1.2

ReGenesees (R evolved Generalised software for sampling estimates and errors in surveys)

Descrizione

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 offerta attraverso R.

Occorre, tuttavia, precisare come ReGenesees non sia il risultato di una semplice migrazione di GENESEES da SAS a R, bensì il frutto di un progetto nuovo e completamente indipendente. ReGenesees risponde, infatti, ad un radicale cambiamento di logica applicativa che, oltre a consentire un più agevole e sicuro utilizzo del software, garantisce un notevole ampliamento della scelta di stimatori rispetto ai quali calcolare le stime e gli errori campionari.

Principali funzionalità statistiche del sistema

- Disegni di campionamento complessi
  - Disegni a più stadi, stratificati, a cluster
  - Probabilità di inclusione variabili, con o senza reintroduzione
  - Disegni di campionamento “misti” (strati AR e NAR)
- Calibrazione
  - Globale e/o per partizioni (per modelli fattorizzabili)
  - A livello di unità e/o di cluster
  - Modelli omo-schedastici e/o etero-schedastici
  - Funzioni distanza: lineare, raking e logit
  - Vincoli di range sui correttori
  - Calibrazione in più passi
  - Trimming coerente dei pesi calibrati
- Stimatori di Base
  - Horvitz-Thompson
  - Calibration Estimators
- Stima della varianza campionaria
  - Formulazione multistadio (algoritmo ricorsivo di Bellhouse)
  - Ultimate-Cluster approximation
  - GENESEES-like per disegni “misti”
  - Linearizzazione di Taylor per stimatori non lineari “smooth”
  - Tecnica di collassamento degli strati per la gestione delle lonely PSU
  - Metodo GVF (Generalized Variance Functions)

- Stime ed errori campionari (errore standard, varianza, coefficiente di variazione, intervallo di confidenza, design effect) per:
    - Totali
    - Medie
    - Distribuzioni di frequenza assoluta o relativa (marginali, condizionate e congiunte)
    - Rapporti fra totali
    - Shares e rapporti fra shares
    - Coefficienti di regressione multipla
    - Quantili (stima della varianza con il metodo di Woodruff)
- Stime ed errori campionari per Stimatori Complessi
    - Funzioni differenziabili arbitrarie di stimatori di Horvitz-Thompson o di Calibrazione
    - Definibili liberamente dall'utente
    - Linearizzazione di Taylor automatica
    - Covarianza e correlazione fra stimatori complessi
- Stime ed errori campionari per sottopopolazioni (domini).

Architettura del sistema

L'architettura del sistema si articola su due package R integrati:

- package **ReGenesees**: implementa lo strato applicativo del sistema, cioè tutte le funzionalità statistiche che il sistema rende disponibili all'utente
- package **ReGenesees.GUI**: implementa lo strato di presentazione del sistema, cioè un'interfaccia utente di tipo grafico basata su Tcl/Tk

Deve essere sottolineato che il package **ReGenesees** può essere utilizzato anche da solo, interagendo con R nel modo tradizionale, cioè da linea di comando. Questa opzione potrebbe rivelarsi necessaria in specifici contesti applicativi (le simulazioni sono un tipico esempio) o apparire comunque preferibile ad utenti esperti del sistema R.

Al contrario, il package **ReGenesees.GUI** *richiede* il package **ReGenesees**, e lo importa automaticamente all'atto del caricamento. La GUI è stata progettata e realizzata con l'intento di rendere quanto più possibile amichevole e semplice l'interazione con il sistema ReGenesees anche ad utenti che non siano esperti di R, né di teoria del campionamento da popolazioni finite.

Informazioni

Status:	validato
Autore:	Istat
Licenza:	EUPL-1.1
Codifica GSBPM:	5.6 Calculate weights 5.7 Calculate aggregates
Linguaggio di programmazione:	R
Versione linguistica della GUI:	EN
Parole chiave:	calibrazione, calcolo delle stime, stima della varianza, indagini complesse, stimatori complessi, linearizzazione automatica
Contatto:	nome: Diego Zardetto email: zardetto@istat.it

Reperimento software e documentazione

REQUISITI TECNICI

Il package ReGenesees richiede l'installazione di R versione 2.14.0 (o superiore).

Il package ReGenesees.GUI richiede l'installazione di R versione 2.14.0 (o superiore) e dei package ReGenesees, tcltk2, RODBC e svMisc.

COPYRIGHT

DISCLAIMER

L'Istat non si assume la responsabilità per risultati derivanti da un uso dello strumento non coerente con le indicazioni metodologiche contenute nella documentazione disponibile.

DOWNLOAD

Package **ReGenesees** (funzionalità statistiche del sistema)

- Versione 1.9 – Package precompilato: Sistemi Windows
- Versione 1.9 – Sorgenti del package: Sistemi Windows e Unix-like

Package **ReGenesees.GUI** (interfaccia grafica del sistema)

- Versione 1.9 – Package precompilato: Sistemi Windows

- Versione 1.9– Sorgenti del package: Sistemi Windows e Unix-like

DOCUMENTAZIONE TECNICA E METODOLOGICA

Reference manual – ReGenesees v. 1.9

Reference manual – ReGeneseesGUI v. 1.9

ALTRA DOCUMENTAZIONE

Fallows A., Pope M., Digby-North J., Brown G., Lewis D. 2015. A Comparative Study of Complex Survey Estimation Software in ONS. Romanian Statistical Review, 3:46-64.

Zardetto D. 2015. ReGenesees: an Advanced R System for Calibration, Estimation and Sampling Error Assessment in Complex Sample Surveys, (extended version). Journal of Official Statistics, 31(2):177-203.

Zardetto D. 2013. ReGenesees: an Advanced R System for Calibration, Estimation and Sampling Errors Assessment in Complex Sample Surveys. In *Proceedings of the 7th International Conferences on New Techniques and Technologies for Statistics (NTTS)*, Eurostat, Brussels, 5-7 marzo 2013.

Barcaroli G., D. Zardetto. 2012. Use of R in Business Surveys at the Italian National Institute of Statistics: Experiences and Perspectives. In *Proceedings of the 4th International Conference of Establishment Surveys (ICES IV)*, American Statistical Association, Montréal, 11-14 giugno 2012.

Fase ANALISI

Preparazione degli output preliminari

In questo sottoprocesso i dati vengono trasformati in prodotti statistici. Questo sottoprocesso comprende la produzione di misure aggiuntive come indici, tendenze o serie destagionalizzate, nonché la registrazione delle caratteristiche di qualità.

A. Indici compositi

Un indice composito (o indice sintetico) è una combinazione matematica (o aggregazione) di un insieme di indicatori elementari (variabili) che rappresentano le diverse componenti di un concetto multidimensionale da misurare (per es., sviluppo, qualità della vita, benessere, ecc.). Un indicatore elementare è un dato 'elaborato' costruito, generalmente, rapportando un dato 'grezzo' ad un altro che ne costituisce una base di riferimento (per es., "reddito pro-capite").

Quindi, gli indici sintetici sono usati per misurare concetti che non possono essere catturati da un unico indicatore.

Generalmente, un indice sintetico dovrebbe essere basato su un quadro teorico che consenta di selezionare, combinare e pesare gli indicatori elementari in modo da riflettere le dimensioni o la struttura del fenomeno che si sta misurando. Tuttavia, la sua costruzione non è semplice e, spesso, richiede una serie di decisioni/scelte (metodologiche o no) da prendere.

La procedura per costruire un indice sintetico prevede i seguenti passi.

1. Definizione del fenomeno da misurare. La definizione del concetto dovrebbe fornire un senso chiaro di ciò che si intende misurare con l'indice sintetico. Essa dovrebbe riferirsi a un quadro teorico, comprendente diversi sottogruppi e indicatori sottostanti. Un aspetto fondamentale riguarda l'identificazione del modello di misurazione, per il quale si distinguono due diversi approcci:
    - modello riflessivo, se gli indicatori sono visti come 'effetto' del fenomeno da misurare, per cui un cambiamento nella variabile latente si riflette in un cambiamento degli indicatori osservati (gli indicatori sono intercambiabili e le correlazioni tra di essi sono spiegate dal modello);
    - modello formativo, se gli indicatori sono visti come 'causa' del fenomeno da misurare, per cui un cambiamento nella variabile latente non implica necessariamente un cambiamento di tutti gli indicatori osservati (gli indicatori non sono intercambiabili e le correlazioni tra di essi non sono spiegate dal modello).
2. Selezione di un gruppo di indicatori elementari. La forza e la debolezza di un indice sintetico riflettono la qualità degli indicatori elementari sottostanti. Gli indicatori dovrebbero essere selezionati in base alla loro rilevanza, validità, tempestività, disponibilità, ecc. La fase di selezione è il risultato di un compromesso tra possibili ridondanze e perdita di informazione. Un approccio statistico alla scelta degli indicatori consiste nel calcolare le correlazioni tra potenziali indicatori e includere quelli meno

correlati tra loro. Tuttavia, il processo di selezione dipende dal modello di misurazione adottato: in un modello riflessivo, tutti gli indicatori devono essere correlati tra loro, mentre in un modello formativo possono essere incorrelati.

3. Normalizzazione degli indicatori elementari. La normalizzazione ha lo scopo di rendere gli indicatori comparabili in quanto essi, spesso, sono espressi in unità di misura diverse e possono avere polarità La 'polarità' (o 'verso') di un indicatore elementare è il segno della relazione tra l'indicatore e il fenomeno da misurare (per es., nella costruzione di un indice sintetico di sviluppo, la "speranza di vita" ha polarità positiva, mentre la "mortalità infantile" ha polarità negativa). Pertanto, è necessario portare gli indicatori a uno stesso standard, invertendo la polarità, laddove necessario, e trasformandoli in numeri puri, adimensionali. Esistono vari metodi di normalizzazione, come la trasformazione in indici relativi (o metodo Min-Max) e la standardizzazione (calcolo dei z-scores).
4. Aggregazione degli indicatori normalizzati. E' la combinazione di tutte le componenti per formare l'indice sintetico (funzione matematica). Tale passo richiede la definizione dell'importanza di ciascun indicatore elementare (sistema di ponderazione) e l'identificazione della tecnica di sintesi (compensativa o non-compensativa). Il sistema più semplice e usato per la definizione del sistema di ponderazione - ma non per questo esente da critiche - consiste nell'assegnare lo stesso peso a tutti gli indicatori. Per quanto riguarda la tecnica di sintesi, si distinguono due approcci:
    - approccio compensativo, se gli indicatori elementari sono considerati sostituibili; gli indicatori elementari sono detti 'sostituibili' se un deficit in un indicatore può essere compensato da un surplus in un altro (per es., un valore basso in "Percentuale di persone che hanno partecipato ad attività spirituali o religiose" può essere compensato da un valore alto in "Percentuale di persone che hanno partecipato a incontri di associazioni ricreative o culturali" e viceversa). In tal caso, si adottano delle funzioni lineari, come la media aritmetica;
    - approccio non compensativo, se gli indicatori elementari sono considerati non-sostituibili; gli indicatori elementari sono detti 'non-sostituibili' se un deficit in un indicatore non può essere compensato da un surplus in un altro (per es., un valore basso in "Letti di ospedale per 1.000 abitanti" non può essere compensato da un valore alto in "Medici per 1.000 abitanti" e viceversa). In tal caso, si adottano delle funzioni non lineari in cui si tiene conto - implicitamente o esplicitamente - dello sbilanciamento tra i diversi valori, in termini di penalizzazione.
5. Validazione dell'indice sintetico. Consiste nel verificare che l'indice sintetico è coerente con il quadro teorico generale. In particolare, occorre valutare la capacità dell'indice di produrre risultati stabili e corretti (Analisi di Influenza e/o Analisi di Robustezza) e la sua capacità discriminante.

B. Destagionalizzazione di serie storiche

La stagionalità, nella dinamica di una serie storica, è quella componente che si ripete ad intervalli regolari ogni anno, con variazioni di intensità più o meno analoga nello stesso periodo (mese, trimestre, etc.) di anni successivi e di intensità diversa nel corso di uno stesso anno. La sua presenza, potendo mascherare altri movimenti di interesse, tipicamente le fluttuazioni cicliche, viene spesso considerata di disturbo nell'analisi della congiuntura economica; essa, ad esempio, rende problematica l'interpretazione delle variazioni osservate su una serie storica tra

due periodi consecutivi dell'anno (cd. variazione congiunturale), essendo queste spesso influenzate in misura prevalente dalle oscillazioni stagionali piuttosto che da movimenti dovuti ad altre cause (come al ciclo economico). Questi ultimi possono essere, invece, correttamente evidenziati calcolando le variazioni congiunturali sui dati destagionalizzati, dai quali, cioè, è stata opportunamente rimossa la componente stagionale.

Tale trasformazione dei dati risulta, quindi, opportuna nell'analisi della congiuntura economica, per poter cogliere in maniera più chiara l'evoluzione di breve termine dei fenomeni considerati. L'impiego di dati in forma destagionalizzata trova, inoltre, ampia applicazione nell'utilizzo congiunto delle statistiche prodotte da diversi Paesi, poiché permette di comparare in maniera più idonea l'evoluzione di diverse serie storiche, ciascuna caratterizzata da uno specifico profilo stagionale.

Un'altra pratica, strettamente connessa alla precedente, è quella di correggere i dati per la cosiddetta componente di calendario, determinata dalla diversa composizione del calendario nei singoli periodi dell'anno, che contribuisce anch'essa ad offuscare il segnale congiunturale di interesse. Il diverso numero di giorni lavorativi o di giorni specifici della settimana in essi contenuti, come anche il modo in cui si collocano, nei periodi messi a confronto, le festività nazionali civili e religiose, fisse e mobili, e gli anni bisestili, possono costituire una fonte di variazione di breve periodo per molte serie storiche. Tali effetti, non necessariamente analoghi tra paesi o settori, inficiano la comparabilità nel tempo dei fenomeni economici e pertanto sono spesso rimossi unitamente alla componente stagionale. Il ricorso a tale trasformazione dei dati consente, in particolare, di cogliere in maniera più adeguata sia le variazioni tendenziali (calcolate rispetto allo stesso periodo dell'anno precedente), sia le variazioni medie annue. In molti casi, accanto ai dati destagionalizzati e corretti, vengono prodotte anche serie storiche al netto dei soli effetti di calendario.

Principali approcci alla destagionalizzazione

Generalmente, l'ipotesi sottostante alla costruzione di una procedura di destagionalizzazione è che ogni serie storica *Yt,* osservata a cadenza infra-annuale (ove $t = 1, 2, \ldots, T$ è un indice temporale), sia esprimibile come una combinazione delle seguenti componenti non osservabili:

1. una componente di trend *Tt*, che rappresenta la tendenza di medio-lungo periodo, talvolta denominata anche ciclo-trend (*CTt*);
2. una componente stagionale *St*, costituita da oscillazioni di periodo annuale;
3. una componente irregolare *It*, dovuta a movimenti erratici, cioè a fluttuazioni di breve periodo non sistematiche e non prevedibili.

Nell'ambito della produzione statistica ufficiale, gli approcci metodologici più diffusi alla destagionalizzazione sono essenzialmente i due, il cui impiego viene anche incoraggiato nelle linee guida europee sulla destagionalizzazione (Eurostat, 2015):

1. Metodi di tipo *Arima model based* (AMB), sviluppati tra gli altri da Burman (1980), Box, Hillmer e Tiao (1978) e Hillmer e Tiao (1982), basati sull'ipotesi che esista un particolare modello statistico parametrico (Arima) in grado di descrivere adeguatamente la struttura probabilistica del processo stocastico generatore della serie storica osservata, essendo

quest'ultima concepita come la parte finita di una particolare realizzazione di un processo stocastico. I filtri lineari utilizzati in questo approccio dipendono, conseguentemente, dalle caratteristiche della serie storica considerata. Questo tipo di approccio metodologico è adottato dalla procedura TRAMO-SEATS (*Time series Regression with Arima noise, Missing observations and Outliers e Signal Extraction in Arima Time Series* – TS), sviluppata da Gómez e Maravall (1996).

2. Metodi filter based (FLB), di tipo non parametrico o semiparametrico, in cui, al contrario, la stima delle componenti avviene senza ipotizzare l'esistenza di un modello statistico rappresentante la serie analizzata ma mediante l'applicazione iterativa di una serie di filtri lineari costituiti da medie mobili centrate di diversa lunghezza. Tali procedure sono dette ad hoc, poiché i filtri adottati derivano da regole meramente empiriche piuttosto che dalla struttura probabilistica del processo stocastico che ha generato la serie. Appartengono a questo gruppo i classici metodi della famiglia X-11 (X11): dai primi X11 e X-11-ARIMA (X-11A), ai più attuali X-12-ARIMA (X-12A) (Findley et al., 1998) e X-13-ARIMA-SEATS (X-13AS) (Findley, 2005), che incorporano al loro interno numerosi miglioramenti rispetto alle precedenti versioni. Tra questi, il ricorso a modelli reg-Arima finalizzato al trattamento preliminare dei dati e a una migliore previsione della serie, che si traduce in un miglioramento dei filtri simmetrici a media mobile impiegati, e cioè, generalmente, in una maggiore stabilità dei fattori stagionali stimati.

In entrambe le metodologie è presente un trattamento preliminare dei dati, in cui avviene la scelta dello schema di scomposizione che lega le diverse componenti della serie storica (additiva, moltiplicativa, log-additiva, ecc.) e sono identificati ed eliminati una serie di effetti, quali i valori anomali (outlier) e quelli legati agli effetti di calendario. È su questa serie corretta preliminarmente che viene condotta la fase successiva che consente di ottenere la serie destagionalizzata (SA). A questa fase segue il reinserimento, nella serie SA, di alcuni elementi identificati nella fase di pretrattamento, attribuiti o al trend (come i cambiamenti di livello) o alla componente irregolare (ad es. gli outlier additivi e i cambiamenti temporanei); vengono invece esclusi dalla serie SA gli effetti di calendario e gli outlier stagionali.

Tutela della riservatezza

La funzione primaria di un sistema statistico pubblico è quella di produrre statistica ufficiale per il proprio paese. Infatti, il Decreto Legislativo 6 settembre 1989, n.322, costitutivo del Sistema statistico nazionale (Sistan), cita: *"L'informazione statistica ufficiale è fornita al Paese e agli organismi internazionali attraverso il Sistema statistico nazionale"* (art.1, comma 2) e ancora *"I dati elaborati nell'ambito delle rilevazioni statistiche comprese nel programma statistico nazionale sono patrimonio della collettività e vengono distribuiti per fini di studio e di ricerca a coloro che li richiedono secondo la disciplina del presente decreto, fermi restando i divieti di cui all'art. 9"* riguardanti il segreto statistico (art. 10 comma 1).

Il Decreto Legislativo n.322/1989, inoltre, afferma che *"i dati raccolti nell'ambito di rilevazioni statistiche comprese nel Programma statistico nazionale non possono essere comunicati o diffusi ad alcun soggetto esterno, pubblico o privato, né ad alcun ufficio della pubblica amministrazione se non in forma aggregata e in modo che non se ne possa trarre alcun riferimento a persone identificabili"*. In ogni caso i dati non possono essere utilizzati al fine di identificare nuovamente gli interessati.

Ulteriori principi, in materia di tutela della riservatezza dei dati, sono stabiliti dal Codice di deontologia e di buona condotta per i trattamenti di dati personali a scopi statistici e di ricerca

scientifica effettuati nell'ambito del Sistema statistico nazionale (Decreto Legislativo 30 giugno 2003, n. 196). In particolare, il Codice definisce il concetto di identificabilità di un'unità statistica, in termini di possibilità, mediante l'uso di mezzi ragionevoli, di stabilire un'associazione significativamente probabile tra la combinazione delle modalità delle variabili relative all'unità statistica e i dati identificativi della medesima. Sono specificati, inoltre, i mezzi ragionevolmente utilizzabili per l'identificazione dell'interessato, quali ad esempio, le risorse economiche, di tempo, la possibilità di incroci con archivi nominativi o altre fonti, ecc.

La traduzione dei concetti enunciati nella legge in regole operative dal punto di vista statistico richiede preliminarmente l'individuazione delle unità statistiche soggette a rischio di identificazione e quindi una precisa definizione di cosa si intenda per violazione della riservatezza. La successiva quantificazione della probabilità di violare la riservatezza consentirà di definire le tecniche più idonee per garantire la protezione del dato.

La definizione di violazione della riservatezza adottata dagli Istituti nazionali di statistica è basata sul concetto di identificabilità di una unità della popolazione osservata (interessato). Indicando col termine *intruso* il soggetto che abbia interesse a violare la riservatezza dei dati rilasciati, si verifica una identificazione quando l'intruso riesca ad associare, con un determinato grado di certezza, le informazioni rilasciate al soggetto interessato. Il rilascio di informazione statistica con dati riservati in nessun caso riguarda i cosiddetti *identificativi diretti* (ovvero le variabili che identificano univocamente l'interessato come ad esempio codice fiscale, nominativo o ragione sociale, indirizzo, ecc.). Il problema si pone per i cosiddetti *identificativi indiretti* (o variabili chiave). Si tratta di quelle variabili che non identificano direttamente l'interessato ma che permettono di circoscriverne la popolazione di appartenenza e che l'intruso utilizzerà per i suoi scopi. Un'identificazione indiretta potrebbe essere determinata, ad esempio, dall'utilizzo combinato di variabili territoriali, attività economica e classe di addetti. Il meccanismo con cui una identificazione può avvenire può essere immediato (ad esempio riconoscimento diretto) o affidato a più o meno complessi algoritmi di abbinamento di informazioni (*record linkage*, *statistical matching* ecc.)

Per limitare il rischio di una identificazione gli Istituti nazionali di statistica possono operare modifiche ai dati (ad esempio ricorrendo a tecniche di perturbazione), oppure agire sugli identificativi indiretti eliminandoli in tutto o in parte, oppure riducendone i dettagli (ad esempio decidendo di non rilasciare il dettaglio *comune* e lasciando al suo posto la variabile *provincia* o *regione*). L'applicazione delle tecniche di protezione, sia per la diffusione di tabelle che per la comunicazione di dati elementari, comporta una riduzione o una modifica del contenuto informativo dei dati rilasciati (perdita di informazione).

A. La violazione della riservatezza nel rilascio di tabelle

Le tabella rappresentano lo strumento maggiormente utilizzato dagli istituti nazionali di statistica per la diffusione dei dati aggregati, ovvero raggruppati in celle definite dagli incroci delle variabili di classificazione. Il concetto di violazione della riservatezza prescinde dal tipo di prodotto utilizzato per la diffusione. Coerentemente con quanto esposto nella sezione precedente, anche nel caso di dati aggregati si verifica una violazione quando è possibile trarre informazioni che consentono l'identificazione dell'individuo. Nella definizione di informazioni "riservate" rientrano anche i dati sensibili e i dati giudiziari (così come sono definiti nel Decreto Legislativo 30 giugno 2003, n. 196, art. 4), mentre non sono considerate riservate le variabili pubbliche (il carattere o la combinazione di caratteri, di tipo qualitativo o quantitativo, oggetto di una rilevazione statistica che faccia riferimento ad informazioni presenti in pubblici registri,

elenchi, atti, documenti o fonti conoscibili da chiunque – definizione contenuta nel Codice di deontologia). Quando si intende rilasciare una tabella una prima valutazione riguarda il contenuto informativo relativo ai dati da pubblicare: se questo non ha carattere riservato non si rende necessario attuare procedure di protezione statistica dei dati, in caso contrario è necessario applicare le regole di tutela della riservatezza. La valutazione del rischio di violazione della riservatezza per dati in tabella avviene per singole celle: quando il valore interno ad una delle celle è riconducibile (con un determinato grado di certezza) al soggetto (ai soggetti) cui il dato stesso si riferisce (cella sensibile), allora la tabella non rispetta le norme sulla tutela della riservatezza.

Il processo volto alla protezione dei dati aggregati prevede diverse fasi. La prima fase definisce l'ambito nel quale si sta lavorando, quali tabelle si intendono trattare e le loro caratteristiche. Quindi si definisce la regola di rischio ovvero il criterio (o i criteri) in base al quale stabilire se una cella è o meno a rischio di violazione della riservatezza. La fase finale riguarda la messa in atto delle procedure per la tutela della riservatezza. Queste dipendono dal tipo di tabelle che si intende rilasciare e da eventuali vincoli di pubblicazione presenti, ma anche dalla tipologia di variabili riservate, nonché dalla complessità sottostante a ogni elaborazione e alla disponibilità di dati.

Anche se alcuni dei principi di seguito descritti, con particolare riferimento alla regola della soglia, sono utilizzabili anche per le tabelle di frequenza, le regole elencate sono riferibili principalmente alle tabelle di intensità. Nel caso di tabelle di frequenza le celle a rischio sono individuate a seguito di una valutazione fatta caso per caso e non facendo ricorso a regole generali come invece avviene per le tabelle di intensità.

Tabelle di Intensità e regole di rischio

Le regole di rischio utilizzate per tabelle di intensità sono quelle basate sulla numerosità della cella (*regola della soglia o frequenza*), e quelle basate su misure di concentrazione (*regola della dominanza e regola del rapporto*). In Istat trova largo impiego la regola della soglia secondo cui una cella è sensibile se il numero di unità in essa contenute è inferiore ad un valore *n* (soglia) fissato a priori. Per poter applicare questa regola a tabelle di intensità è necessario disporre della relativa tabella di frequenza. Dal valore di *n* dipende la protezione che si applica alla tabella: maggiore è il valore soglia maggiore è il livello di protezione applicato. Non esiste un criterio univoco per individuare il valore soglia che dipenderà dallo scenario di intrusione ipotizzato e dai dati trattati. Il valore minimo che può assumere la soglia è pari a tre (come previsto dal Codice deontologico).

La regola della dominanza [(*n,k*)-*dominance*] definisce a rischio una cella se i primi *n* contributori detengono una percentuale del suo valore totale superiore ad una soglia *k*% fissata a priori. Dai due valori di *n* e di *k* dipende il livello di protezione che si vuole applicare alla tabella. Non esistono criteri univoci per fissare i due parametri. In base alle unità statistiche coinvolte e ai livelli di protezione desiderati è possibile definire i parametri individuando una concentrazione massima ammissibile.

La regola del rapporto (*p-rule*) si basa sulla precisione con la quale può essere stimato il valore del primo contributore nell'ipotesi in cui il secondo contributore tenti la violazione. La cella è considerata a rischio se l'errore relativo è inferiore ad una soglia *p* fissata a priori.

Nel caso di tabelle con possibili contributi di segno opposto le regole di rischio basate sulle misure di concentrazione perdono di significato. E' possibile tuttavia la loro applicazione facendo ricorso ai valori assoluti dei contributi.

Operare una violazione della riservatezza in un contesto con possibili contributi negativi risulta molto più complesso. La raccomandazione generale è quella di parametrizzare le funzioni di rischio con valori meno stringenti rispetto al caso di contributi solo positivi.

In caso di tabelle campionarie, ovvero ottenute rilevando dati su un sottoinsieme della popolazione di riferimento, la valutazione del rischio di violazione della riservatezza deve tener conto del piano di campionamento utilizzato. Il valore riportato nelle celle è una stima realizzata estendendo un valore parziale (rilevato nel campione) alla popolazione di riferimento. Le unità rilevate non sono conosciute e anche il vero valore della popolazione non viene rilevato. Per le celle che riportano dati stimati con coefficiente di riporto all'universo maggiore dell'unità il rischio di violazione è contenuto. In questo contesto ipotizzare una violazione della riservatezza appare inverosimile. Tuttavia, specie per tabelle di dati economici, un'attenta valutazione del rischio di violazione della riservatezza si rende necessaria anche nel caso di tabelle campionarie. Infatti, in alcuni casi le unità maggiormente rappresentative (dominanti) vengono incluse nel campione con probabilità certa. Inoltre, nel caso di campioni stratificati, alcune celle sono campionate al 100% e quindi il valore rilevato coincide (a meno di mancate risposte) col valore della popolazione.

Tranne casi particolari in cui il disegno campionario e il numero di unità campionate permettono di ritenere sicura una tabella sotto il profilo della riservatezza, anche alle tabelle campionarie devono essere applicate le regole di riservatezza.

Il criterio utilizzato in Istat considera l'applicazione delle regole di rischio sui valori stimati di cella ottenuti usando i pesi di riporto all'universo. Questo criterio presuppone che le unità campionate siano “simili” a quelle presenti nella popolazione.

Tabelle di frequenza e regola di rischio

Le tabelle di frequenza sono utilizzate soprattutto per rappresentare fenomeni sociali e dati di censimento. Per questa tipologia di tabelle l'unico criterio per stabilire se una cella è o meno a rischio è quello basato sulla numerosità delle celle, non possono infatti essere applicate regole di rischio basate sulle misure di concentrazione. Non esistono regole univoche per stabilire se una tabella di frequenza sia o meno a rischio di violazione della riservatezza. Non sempre infatti una cella con frequenza bassa (esempio pari a 1) indica una cella a rischio, e viceversa non sempre una cella che contiene un elevato numero di unità può essere considerata sicura sotto il profilo della riservatezza statistica.

Come regola generale sono considerate a rischio di violazione della riservatezza le tabelle di frequenza che presentano uno dei casi sotto elencati:

- marginale con meno di tre contributori;
- tutte le unità appartengono ad una unica categoria (*group disclosure*), oppure l'unico contributore di una cella (auto riconoscimento) acquisisce informazioni riservate su tutte le altre unità (concentrate tutte in un'altra cella).

Protezione statistica delle tabelle

Individuate le celle a rischio è necessario modificare la tabella in modo opportuno rendendo *anonime* le informazioni in essa contenute. Le tecniche di protezione dei dati sono molteplici e vanno dall'accorpamento di modalità adiacenti, a metodi basati sulla modifica dei dati originali, all'introduzione di valori mancanti (soppressioni). I metodi utilizzati in Istat sono: la modifica delle modalità delle variabili di classificazione e l'introduzione di valori mancanti.
Un metodo di protezione delle tabelle che non si basa sulla modifica dei valori nelle celle è la definizione di una diversa combinazione delle modalità. Individuata la regola di rischio il metodo consiste nel determinare le modalità in modo tale che la distribuzione del carattere e/o delle unità nelle celle sia tale da non presentare alcuna cella sensibile.

Modificando opportunamente le modalità è possibile, ad esempio, ottenere una tabella che presenti una numerosità minima (ad esempio maggiore o uguale a tre) in ogni cella, oppure una tabella con una predefinita concentrazione massima del carattere in ogni cella.

La modifica delle modalità delle variabili di classificazione è soluzione praticabile solo quando il carattere delle variabili di classificazione è trasferibile, e se le tabelle da rilasciare non devono soddisfare regole rigide dettate da regolamenti che vincolano i dettagli delle variabili di classificazione.
La tecnica relativa all'inserimento di valori mancanti (tecnica di soppressione secondaria) prevede che il valore delle celle a rischio sia soppresso (oscurato). La soppressione operata sulle celle a rischio è anche detta soppressione primaria. Con l'introduzione dei valori mancanti in corrispondenza delle celle *sensibili* non si esaurisce il processo di protezione della tabella. E' necessario prima valutare che le celle soppresse non possano essere calcolate a partire dai dati rilasciati, ad esempio per differenza dai valori marginali. Le soppressioni devono distribuirsi tra le celle della tabella in modo da garantire che la tabella sia protetta adeguatamente secondo i criteri imposti. Quando ciò non si verifica è necessario introdurre ulteriori valori mancanti tra le celle non a rischio: le soppressioni secondarie. In letteratura sono stati proposti diversi algoritmi per la determinazione del tracciato delle soppressioni secondarie. Attualmente in Istat quello maggiormente utilizzato è l'algoritmo HiTas disponibile in alcuni software generalizzati come ad esempio Tau-ARGUS.

Tabelle collegate

Si definiscono collegate tabelle che contengono dati relativi alla stessa variabile risposta e che presentano almeno una medesima variabile classificatrice. Il caso più frequente di tabelle collegate è rappresentato da tabelle con celle in comune, con particolare riferimento ai valori marginali. Il collegamento tra dati statistici può inquadrarsi anche in un contesto più ampio. A volte infatti rilevazioni diverse pubblicano stessi aggregati

L'applicazione delle regole di riservatezza a tabelle collegate implica che informazioni (celle) comuni abbiano assegnato lo stesso status di rilasciabilità.

Per ottimizzare il processo di protezione sarebbe opportuno, laddove possibile, operare contestualmente la protezione di tutte le tabelle collegate.

B. La violazione della riservatezza nel rilascio di dati elementari

I dati elementari possono essere definiti come il prodotto finale di una rilevazione statistica dopo le fasi di progettazione, esecuzione, controllo e correzione. I dati elementari nella fase di diffusione sono un archivio di *record* ciascuno contenente tutte le informazioni validate

(generalmente un sottoinsieme di quelle rilevate) relative a una singola unità statistica. Tali variabili, così come avviene nel caso dei dati aggregati diffusi tramite tabelle, possono essere classificate come variabili chiave in quanto identificativi indiretti, oppure come variabili riservate.
Rispetto al caso di rilascio di tabelle cambiano sostanzialmente sia l'insieme delle variabili chiave che, in generale, saranno più numerose, sia il contenuto di un'eventuale violazione in quanto le variabili riservate nei dati elementari sono presenti tutte insieme. Per contro, il rilascio di microdati riguarda esclusivamente le collezioni campionarie e l'accesso ai file è molto più controllato (per soli motivi di ricerca e dietro la sottoscrizione di un modulo/contratto). Tuttavia, non v'è dubbio che il rilascio di dati elementari è questione più delicata rispetto alla diffusione di tabelle. Per questo sono stati elaborati modelli di misurazione del rischio di identificazione specifici rispetto alle tabelle e spesso basati su modelli probabilistici. I metodi di protezione dei dati elementare sono riconducibili a tre categorie:

- ricodifica di variabili (*global recoding*): consiste nel ridurre il dettaglio di rilascio di alcune variabili (ad esempio l'età in classi quinquennali anziché annuali);
- soppressione di informazioni (*local suppression*): per eliminare caratteristiche che rendono alcuni record più facilmente identificabili;
- perturbazione dei dati pubblicati: con metodi diversi ma con le stesse finalità viste per le tabelle.

Fra le iniziative che riguardano il rilascio "protetto" dei dati elementari vanno annoverati i cosiddetti *Microdata File for Research* (*MFR*), i file ad uso pubblico (mIcro.STAT) ed il Laboratorio per l'Analisi dei Dati ELEmentari (ADELE). I file *MFR* vengono prodotti per rilevazioni statistiche riguardanti sia individui e famiglie sia imprese e sono realizzati specificatamente per esigenze di ricerca scientifica. Il rilascio di tali file è soggetto alla sussistenza di alcuni requisiti relativi sia all'organizzazione di appartenenza sia alle caratteristiche del progetto di ricerca per le cui finalità viene richiesto il file. I file mIcro.STAT sono file ad uso pubblico, ottenuti a partire dai rispettivi *MFR,* opportunamente trattati sotto il profilo della tutela della riservatezza e scaricabili direttamente dal sito Istat.

Il Laboratorio ADELE, attivo a partire dal 1999, è un cosiddetto *Research Data Centre* (RDC) ovvero un luogo "sicuro" cui possono accedere ricercatori e studiosi per effettuare autonomamente le proprie analisi statistiche sui dati elementari prodotti dall'Istituto nazionale di statistica nel rispetto delle norme sulla riservatezza. Principale obiettivo del laboratorio ADELE è offrire a un'utenza esterna "esperta" la possibilità di analizzare dati elementari delle principali indagini dell'Istat, spostando la fase di verifica della tutela della riservatezza sull'output dell'analisi statistica piuttosto che sull'input (come avviene nel caso dei file per la ricerca e per i file ad uso pubblico). La tutela della riservatezza per le elaborazioni effettuate presso il laboratorio ADELE viene garantita sotto diversi aspetti:

- legalmente: l'utente sottoscrive un modulo in cui si impegna al rispetto di norme di comportamento specifiche;
- fisicamente: attraverso il controllo dell'ambiente di lavoro. Il Laboratorio è collocato presso la sede dell'Istat con addetti che attendono al controllo della sala, le operazioni di input e output e l'accesso alla rete esterna sono inibite agli utenti;
- statisticamente: tramite il controllo cui sono sottoposti i risultati dell'analisi dell'utente preventivamente al rilascio.

Fase ANALISI - METODI

- **PREPARAZIONE DEGLI OUTPUT PRELIMINARI**

Costruzione e valutazione di indici compositi

Methods for constructing composite indices: one for all or all for one?
2013
Rivista Italiana di Economia Demografia e Statistica, Volume LXVII n. 2.

Destagionalizzazione di serie storiche

Relazione del team tecnico incaricato della definizione di metodi standard per la destagionalizzazione di serie storiche con metodi implementati in diversi strumenti IT (TS, X12-Arima, X13-Arima-Seats, JDemetra)
2015
Gdl per la definizione di standard per l'Istat, Istat

Destagionalizzazione di serie storiche con metodologia Arima model based (AMB) implementata nel software JDemetra+
2015
Gdl per la definizione di standard per l'Istat, Istat

ESS guidelines on seasonal adjustment (2015 Edition)
2015
Manuals and Guidelines, Eurostat

Riferimenti

Costruzione e valutazione di indici compositi

Massoli P., Mazziotta M., Pareto A., Rinaldelli C. 2014. **Indici compositi per il BES**. *Giornate della Ricerca, Istat*, 10-11 Novembre.

Massoli P., Mazziotta M., Pareto A., Rinaldelli C. 2013. **Metodologie di sintesi sperimentali per i domini del BES**. *XXXIV Conferenza Italiana di Scienze Regionali (AISRE)*, Palermo, 2-3 Settembre.

Mazziotta M., A. Pareto 2011. **Un indice sintetico non compensativo per la misura della dotazione infrastrutturale: un'applicazione in ambito sanitario**. *Rivista di Statistica Ufficiale*, 1:63-79.

Mazziotta C., Mazziotta M., Pareto A., Vidoli F. 2010. **La sintesi di indicatori territoriali di dotazione infrastrutturale: metodi di costruzione e procedure di ponderazione a confronto**. *Rivista di Economia e Statistica del Territorio*, 1:7-33.

OECD. 2008. **Handbook on Constructing Composite Indicators. Methodology and user guide**. OECD Publications.

Aureli Cutillo E. 1996. *Lezioni di statistica sociale. Parte II. Sintesi e graduatorie.* CISU, Roma.

Delvecchio F. 1995. *Scale di misura e indicatori sociali*. Cacucci, Bari.

Silvio-Pomenta J. F. 1973. **Typological study using the Wroclaw Taxonomic Method (A study of regional disparities in Venezuela)**. Social science project on human resources indicators, n. 28, UNESCO.

Harbison F. H., J. Maruhnic, J.R. Resnick. 1970. *Quantitative Analyses of Modernization and Development*. Princeton University Press, New Jersey.

Destagionalizzazione di serie storiche

Eurostat. 2015. ***ESS Guidelines on Seasonal Adjustment (2015 Edition)***. Manuals and Guidelines, Eurostat.

Grudkowska S. 2015. *JDemetra+ Reference Manual, v. 0.1*. Narodowy Bank Polski.

Eurostat. 2014. **ESS Handbook for Quality Reports (2014 Edition)**. Manuals and guidelines, Eurostat.

Findley D. F. 2005. **Some Recent Developments and Directions in Seasonal Adjustment**. *Journal of Official Statistics*, 21(2):343-365.

Giovannini E., D. Piccolo. 2000. *Seasonal adjustment procedures – experiences and perspectives*. Istat.

Findley D. F., B. C. Monsell, W. R. Bell, M. C. Otto, B. C. Chen. 1998. **New Capabilities and Methods of the X-12-ARIMA Seasonal-Adjustment Program**. *Journal of Business & Economic Statistics*, 16(2):127-152.

Gómez V., A. Maravall. 1996. *Programs TRAMO and SEATS, Instruction for User (Beta Version: September 1996)*. Banco de España.

Chen C., L. Liu. 1993. **Joint Estimation of Model Parameters and Outlier Effects in Time Series**. *Journal of the American Statistical Association*, 88(421):284-297.

Hillmer S. C., G. C. Tiao. 1982. **An ARIMA-Model-Based Approach to Seasonal Adjustment**. *Journal of the American Statistical Association* 77(377):63-70.

Burman J. P. 1980. **Seasonal Adjustment by Signal Extraction**. *Journal of the Royal Statistical Society*, 143(3): 321-337.

Box G. E. P., S. C. Hillmer, G. C. Tiao. 1978. **Analysis and Modeling of Seasonal Time Series**. *Seasonal Analysis of Economic Time Series*, 309-334.

Box G. E. P., G. M. Jenkins, G. C. Reinsel. 1970. *Time Series Analysis: Forecasting and Control*. Wiley, Holden Day, San Francisco.

- **TUTELA DELLA RISERVATEZZA**

Disclosure protection of non-nested linked tables in Business Statistics
2009
Supporting paper, Joint UNECE/Eurostat work session on statistical data confidentiality, Bilbao, Spain

Metodologie e tecniche di tutela della riservatezza nel rilascio dell'informazione statistica
2004
Collana Metodi e Norme, n. 20, Istat

Riferimenti

Hundepool A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Sculte Nordholt, K. Spicer, P.P. De Wolf. 2012. *Statistical Disclosure Control*. Wiley & Sons, London.

Fase ANALISI - STRUMENTI

- **PREPARAZIONE DEGLI OUTPUT PRELIMINARI**

COMIC (COMposite Indices Creator)

COMIC è un software per la costruzione di indici compositi, attraverso vari metodi di sintesi, e la valutazione della loro robustezza.

Ranker
Sistema software basato sul linguaggio Visual Basic per l'analisi e la valutazione comparata dei risultati prodotti attraverso diversi metodi di sintesi statistica degli indicatori elementari disponibili in letteratura.

- **TUTELA DELLA RISERVATEZZA**

ARGUS
Software generalizzato per l'applicazione di metodi per la tutela della riservatezza.

COMIC

Descrizione

COMIC è un software per la costruzione di indici compositi, attraverso vari metodi di sintesi, e la valutazione della loro robustezza. COMIC è predisposto per il calcolo degli indici compositi rispetto a più anni di riferimento.

COMIC è un software che consente di:

- acquisire in formato standard (.csv o .xls o .txt o SAS) i valori degli indicatori elementari disponibili rispetto a entità specificate dall'utente;
- standardizzare/normalizzare i valori degli indicatori elementari forniti;
- applicare uno o più metodi di sintesi tra quelli implementati al fine di ottenere l'indice composito rispetto alle entità specificate;
- visualizzare i valori e le graduatorie risultanti dall'applicazione di ogni singolo metodo, in forma sia tabellare sia grafica;
- porre a confronto i valori e le graduatorie ottenute mediante i diversi metodi di aggregazione;
- eseguire l'analisi di influenza e l'analisi di robustezza degli indici compositi ottenuti dal calcolo.

COMIC consente di applicare le seguenti funzioni di aggregazione:

- Media indici 0-1. Media aritmetica degli indicatori elementari trasformati con metodo min-max;
- Media z-scores. Media aritmetica degli indicatori elementari trasformati in scarti standardizzati;
- Indice di Jevons. Media geometrica degli indicatori elementari trasformati in numeri indici;
- Mazziotta-Pareto Index (MPI). Media aritmetica penalizzata degli indicatori elementari trasformati in scarti standardizzati. La penalità è sottratta alla media aritmetica e si basa sulla "variabilità orizzontale" degli indicatori elementari (MPI con penalità negativa);
- Adjusted Mazziotta-Pareto Index (AMPI). Media aritmetica penalizzata degli indicatori elementari trasformati con metodo min-max. La penalità è sottratta alla media aritmetica e si basa sulla "variabilità orizzontale" degli indicatori elementari (AMPI con penalità negativa);
- Indice media geometrica (IMG). Media geometrica degli indicatori elementari trasformati con metodo min-max.

Informazioni

Status:	validato
Autore:	Istat
Licenza:	EUPL-1.1
Codifica GSBPM:	6.1 Prepare draft outputs
Linguaggio di programmazione:	SAS
Versione linguistica della GUI:	IT
Parole chiave:	indici compositi, normalizzazione, aggregazione, graduatorie, influenza, robustezza
Contatto:	nome: Pierpaolo Massoli email: pimassol@istat.it

Reperimento software e documentazione

REQUISITI TECNICI

COMIC richiede SAS per Windows (SAS base)

COPYRIGHT

DISCLAIMER

L'Istat non si assume la responsabilità per risultati derivanti da un uso dello strumento non coerente con le indicazioni metodologiche contenute nella documentazione disponibile.

DOWNLOAD

COMIC versione 1.0

DOCUMENTAZIONE TECNICA E METODOLOGICA

Manuale utente – Comic v. 1.0

ALTRA DOCUMENTAZIONE

Massoli P., Mazziotta M., Pareto A., Rinaldelli C. 2015. COMIC: a tool for composite indices evaluation. Conference Dealing with complexity in society: from plurality of data to synthetic indicators, Padova, 17-18 Settembre

Ranker

Descrizione

Ranker è un sistema software per l'analisi e la valutazione comparata dei risultati prodotti attraverso diversi metodi di sintesi statistica degli indicatori elementari disponibili in letteratura.

Ranker è un prototipo di strumento generalizzato, che consente di:

- acquisire in formato standard (csv o .xls) i valori dei diversi indicatori elementari disponibili per ogni entità, già calcolati e normalizzati;
- effettuare il calcolo, per ogni entità, di uno o più metodi tra quelli implementati;
- visualizzare i valori e le graduatorie risultanti dall'applicazione di ogni singolo metodo, in forma sia tabellare sia grafica;
- porre a confronto le graduatorie mediante i diversi metodi.

Ranker ha implementato otto diversi metodi:

- il metodo Mazziotta-Pareto Index (MPI) nelle due varianti (positivo e negativo);
- il metodo tassonomico di Wroclaw (Wroclaw);
- la media della media dei valori standardizzati (M1Z);
- il metodo delle graduatorie (Grad.RNK);
- il metodo degli indici relativi (IR);
- il metodo della media aritmetica dei numeri indici base media (ANIM);
- il metodo della media geometrica dei numeri indici base media (GNIM);
- il metodo della media quadratica dei numeri indici base media (QNIM).

L'analisi dei dati prevede 5 fasi distinte:

1. L'organizzazione e la lettura dei dati. I dati acquisiti devono essere organizzati in forma tabellare, in una matrice che descrive ciascuna entità territoriale (in riga) in base al valore di tutti gli indicatori selezionati (in colonna). Per ciascun indicatore considerato si provvede, inoltre, a specificare il verso, distinguendo quelli che descrivono un effetto "positivo" rispetto alle dinamiche di sviluppo settoriale e quelli che, al contrario, sono correlati in senso inverso e ai quali corrisponde una graduatoria decrescente delle unità territoriali.
2. La standardizzazione. La standardizzazione è finalizzata a ottenere indicatori depurati dalle specifiche unità di misura, che abbiano eguale ampiezza (per es. tra 0 e 100) o ordine di grandezza (per es. media 0 e scarto 1).
3. L'aggregazione. Una volta caricata la matrice con i dati elementari e dopo aver effettuato la loro standardizzazione, con Ranker, l'applicazione effettua l'elaborazione di tutti i metodi sopra indicati al fine di calcolare i relativi indici sintetici e descrivere ciascuna entità territoriale.
4. La visualizzazione. Attraverso l'applicazione si potrà visualizzare sia i valori ottenuti, sia le graduatorie da essi derivate. Oltre alla descrizione dei valori numerici in forma tabellare, Ranker consente di produrre delle mappe che rappresentano graficamente i risultati ottenuti attraverso ciascun metodo di sintesi degli indicatori elementari.

5. La valutazione dei metodi. Il software consente di valutare comparativamente i risultati prodotti dai diversi metodi: in particolare, l'impatto della scelta del metodo sul risultato finale (la graduatoria).Un primo strumento è dato dalla matrice di cograduazione delle graduatorie ottenute con i diversi metodi. Un secondo strumento utilizzato è la matrice dei diagrammi di dispersione, che può essere prodotta sui valori degli indicatori o, meglio, sulle graduatorie ottenute. Un ultimo strumento è dato dal diagramma di dispersione calcolato su coppie di metodi selezionati, per individuare le entità su cui maggiore è l'impatto della scelta del metodo.

L'Istat rende disponibile sia la versione online del sistema (i.ranker) sia la versione desktop. La versione online implementa i primi cinque metodi sopra descritti e offre la possibilità di visualizzare i valori e la graduatoria risultante dall'applicazione di ogni singolo metodo sia in forma tabellare sia grafica. La versione desktop implementa tutti i metodi sopradescritti ma consente solo la visualizzazione tabellare dei risultati.

Informazioni

Status:	validato
Autore:	Istat
Licenza:	EUPL-1.1
Codifica GSBPM:	6.1 Prepare draft outputs
Linguaggio di programmazione:	Visual Basic
Versione linguistica della GUI:	IT
Parole chiave:	indici sintetici, standardizzazione, aggregazione, graduatorie
Contatto:	nome: Marco Broccoli email: broccoli@istat.it

Reperimento software e documentazione

SISTEMA ONLINE

i.ranker

DOWNLOAD

RANKER versione 1.2

DOCUMENTAZIONE TECNICA E METODOLOGICA

Manuale utente – Ranker v. 1.2

ARGUS

Descrizione

L'Istat ha partecipato al progetto europeo CASC (Computational Aspects of Statistical Confidentiality) il cui obiettivo è stato lo sviluppo del software ARGUS per la tutela della riservatezza nella fase di rilascio dell'informazione statistica.

Il software *ARGUS* si compone di due moduli Mu-ARGUS, per i dati elementari, e Tau-ARGUS, per le tabelle, e contiene molti dei metodi di protezione proposti in letteratura.

Informazioni

Status	validato
Autore	CASC
Codifica GSBPM:	6.4 Apply disclosure control
Parole chiave:	riservatezza, rischio di violazione, valore a rischio, valore soppresso, regola di rischio

Reperimento software e documentazione

Per il reperimento del software e della documentazione tecnica e metodologica è possibile rivolgersi a **CASC**.