Una strategia di controllo e correzione e di editing selettivo basato sui dati amministrativi.

Seminario: “Innovazioni metodologiche e di processo in una rilevazione multi-source su imprese e istituzioni: la Struttura delle retribuzioni e del costo del lavoro 2012”

Maria Teresa Buglielli, Ugo Guarnera, Marilena A. Ciarallo, Ciro Baldi

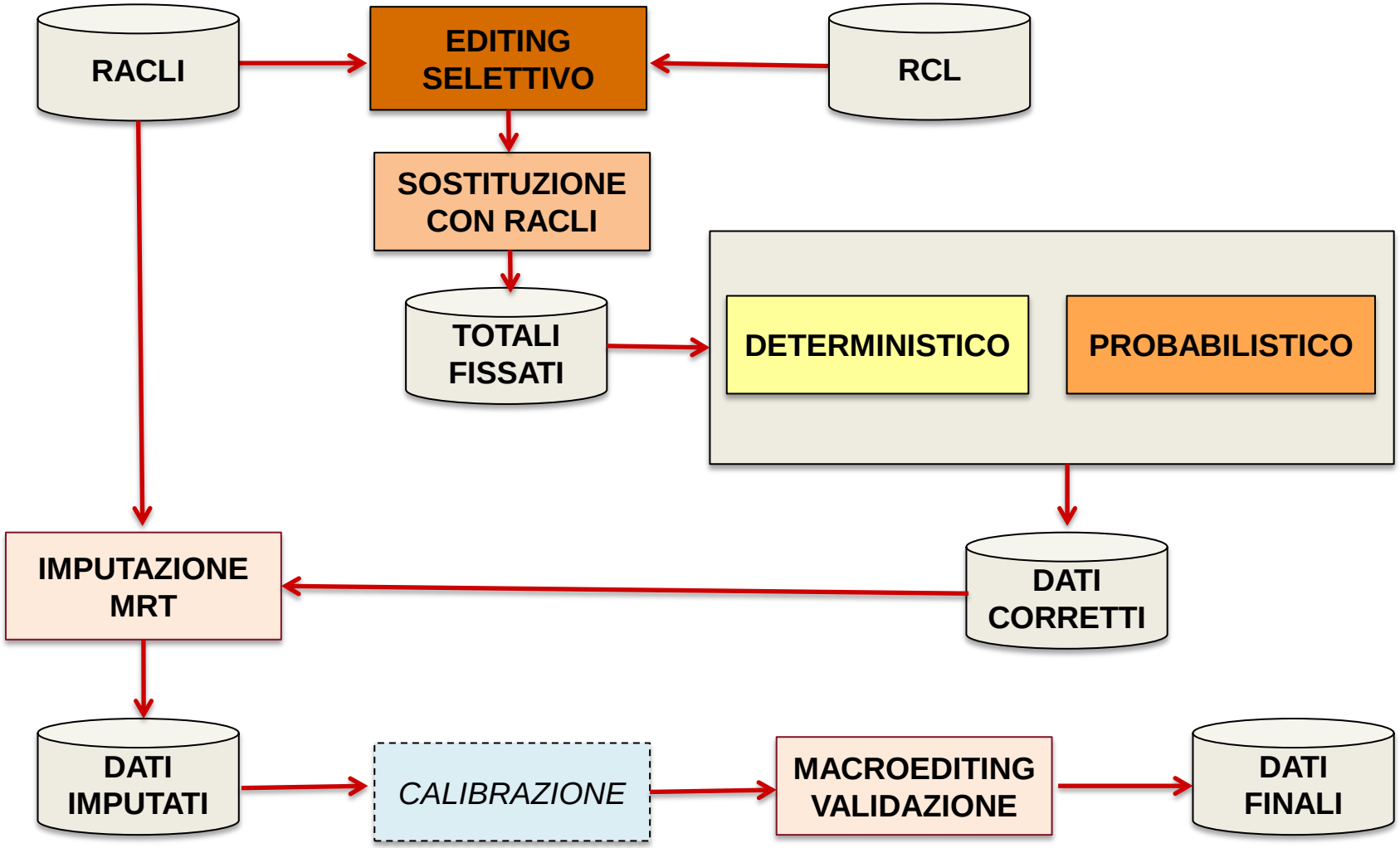
Istat – Aula Magna, 17 febbraio 2015

Istat

Indice

Schema generale del processo di C&C

Schema generale del processo di C&C



Editing selettivo

- Insieme di tecniche per l'identificazione degli errori influenti
- Obiettivo: dividere l'insieme dei dati in due sottoinsiemi
  - un insieme «critico» contenente gli errori potenzialmente influenti (editing interattivo)
  - un insieme contenente record supposti corretti o contenenti errori non influenti (editing automatico)
- Elementi chiave:
  - ordinamento delle unità sulla base di una «funzione punteggio»
  - valore di soglia che determina il numero id unità da revisionare interattivamente

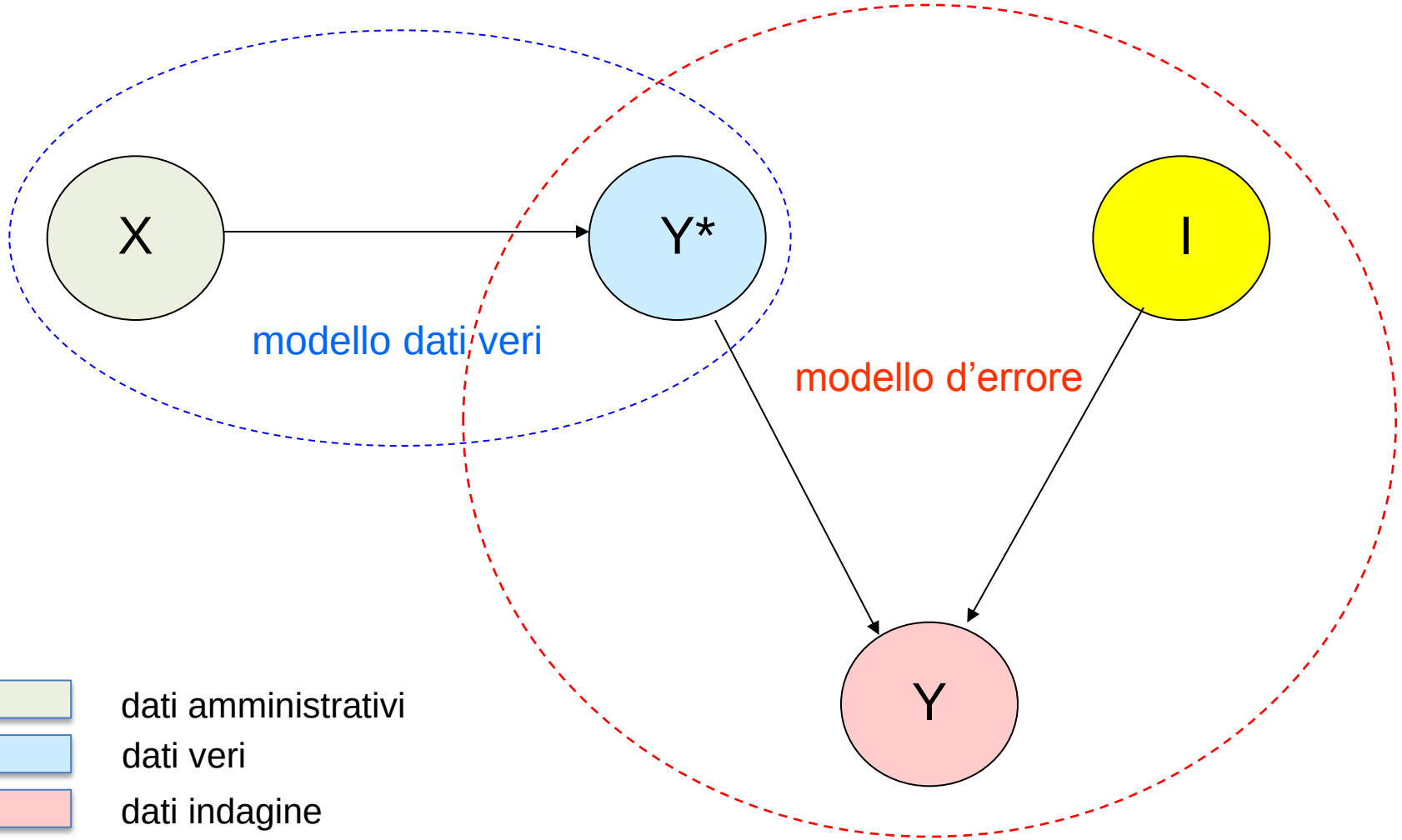
Editing selettivo basato su modello a classi latenti (Selemix)

MSS ha sviluppato un metodo di editing selettivo basato su:

1) modellazione esplicita sia dei dati sia del meccanismo di errore (via modelli mistura)
2) definizione della funzione punteggio in termini della differenza tra valore osservato e valore atteso della distribuzione dei dati veri condizionatamente ai dati osservati

Questo approccio permette di mettere in relazione la soglia di errore per la revisione interattiva all'errore residuo atteso nei dati

Il modello



dati amministrativi
dati veri
dati indagine
indicatore d'errore

Dati fonte RACLI nell'editing selettivo

Variabili X (RACLI)

- numero dipendenti
- retribuzioni totali
- ore retribuite

Variabili Y (RCL)

- ore retribuite
- ore lavorate
- retribuzioni totali
- .....

Gli errori potenzialmente influenti sono stati individuati utilizzando diversi modelli che mettono in relazione le variabili elencate.

Due passi di ES

1. selezione dei record per il ricontatto
2. selezione dei record per la sostituzione dei totali da RACLI

Editing selettivo

Caratteristiche della selezione dei record per ricontatto

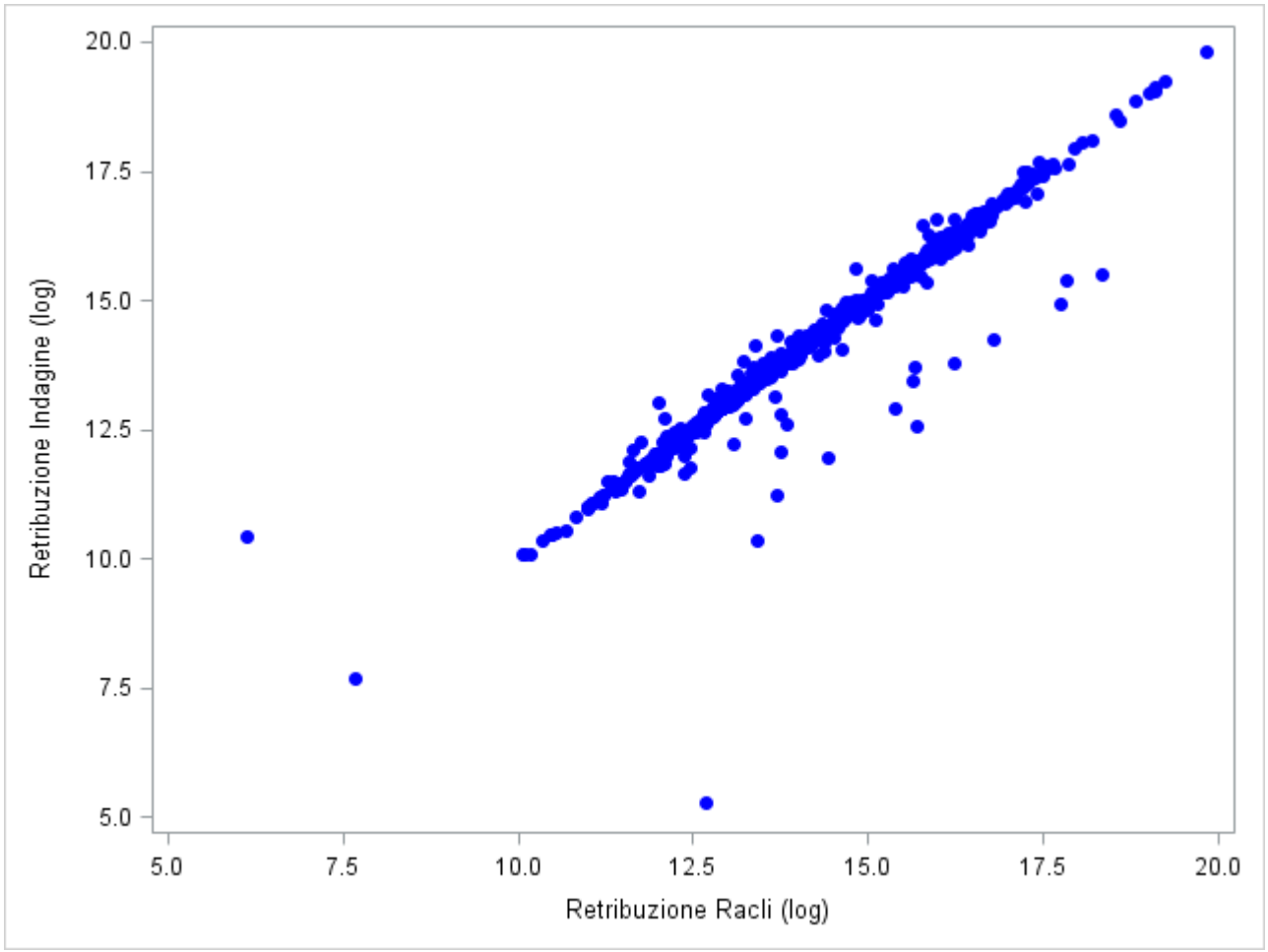
- ✓ in corsa
- ✓ sia a livello di impresa che di unità territoriale
- ✓ diversi modelli per le anomalie di ore retribuite e retribuzioni totali
- ✓ individuazione errori influenti per sezione di ateco
- ✓ retribuzione totale e ore retribuite di fonte amministrativa come covariate

Editing selettivo

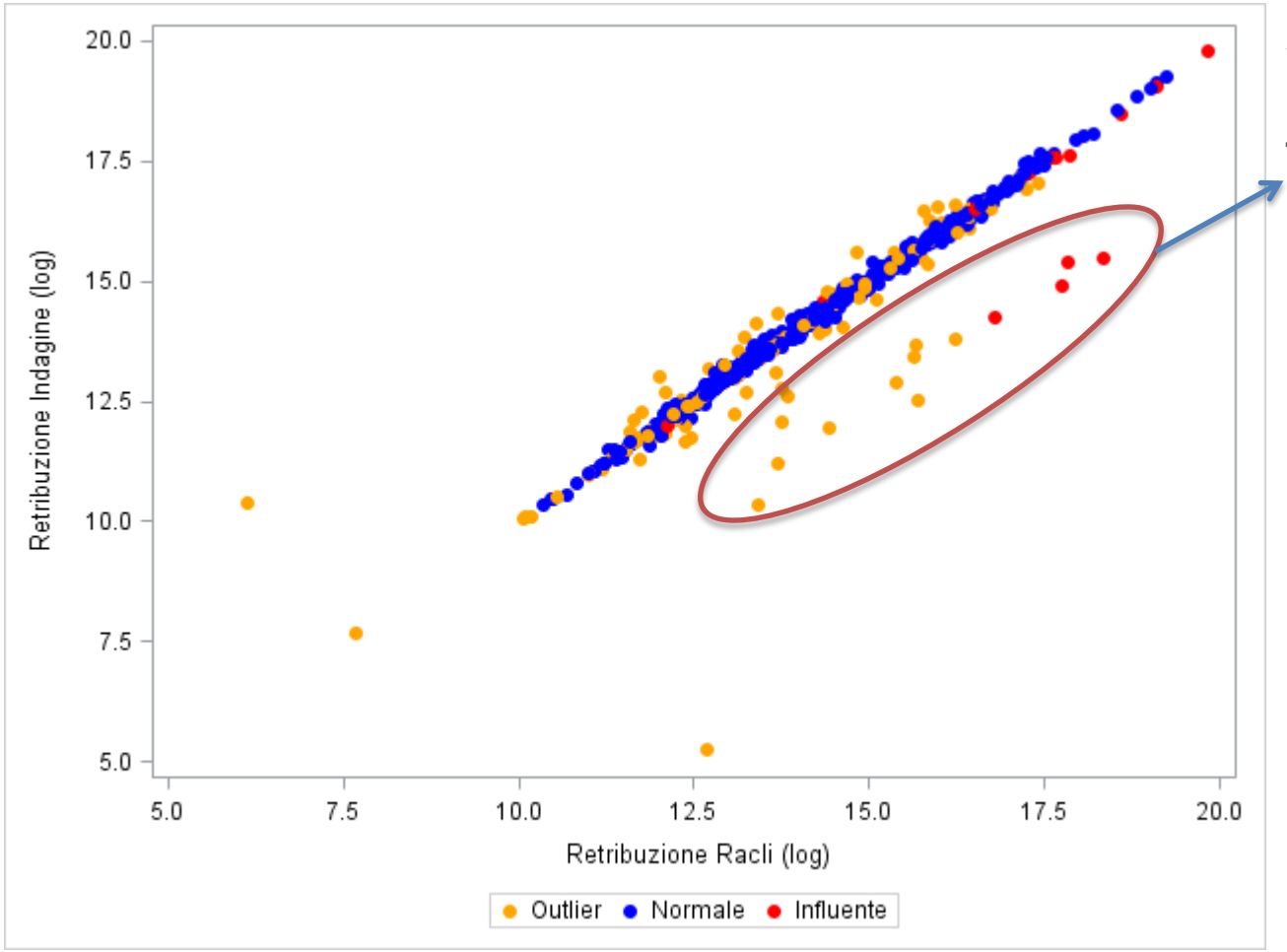
Risultati selezione per ricontatto su 11303 imprese

modello	X	Y	N. outlier	% outlier	N. influenti	% influenti
Struttura retrib.	retrib. RACLI	tredicesima contributi TFR	1513	13,39	132	1,17
Ore	Ore retribuite RACLI	Ore lavorate e retribuite RCL	562	4,97	60	0,53
ALL	dipend. ore retrib, retrib. totali RACLI	dipendenti, ore retrib. retrib. totali RCL	1309	11,58	170	1,50
Totale imprese			2363	20,90	281	2,49

Outlier e influenti per retribuzione - Sezione N



Outlier e influenti per retribuzione - Sezione N



Errore sistematico: sono le imprese interinali che hanno fornito i dati solo per il personale di staff

Sostituzioni con RACLI

IDEA

I dati di fonte amministrativa forniscono i totali di riferimento per dipendenti, retribuzioni, ore retribuite

Individuazione delle imprese con due metodi

- ✓ editing selettivo
  - per raggruppamento di contratti
  - a livello di unità territoriale
  - per imprese non ricontattate o nuove rispondenti
- ✓ confronti deterministici

Sostituzioni con RACLI - confronti deterministici

- Il numero dei dipendenti se:
  - Errore localizzato
  - Si migliora l'accostamento delle variabili procapite con RACLI

- I valori delle unità per cui a livello di impresa:
  - valori non rientrano in range di accostamento
  - imprese non soddisfano i criteri eccezione

- I range di accostamento tengono conto di:
  - Dimensione di impresa
  - Differenti soglie per le ore

- I criteri eccezione tengono conto di:
  - Possibili retribuzioni sopra il massimale imponibile
  - Retribuzioni accostate con i bilanci ma non con Racli (pochissimi casi)
  - Unità non misurate bene da RACLI (es. APSP)

Editing automatico

Correzioni deterministiche

- errori sistematici: si manifestano nei dati in modo consistente in una precisa direzione (retribuzioni per ferie,…)

Correzioni probabilistiche

- causati da fattori accidentali

Correzioni probabilistiche

Metodologia Fellegi-Holt

sviluppata nel 1976 per il trattamento delle variabili qualitative (con edit logici) e poi estesa al trattamento di variabili quantitative (con edit aritmetici: disuguaglianze lineari, rapporti, ..)

Principi:

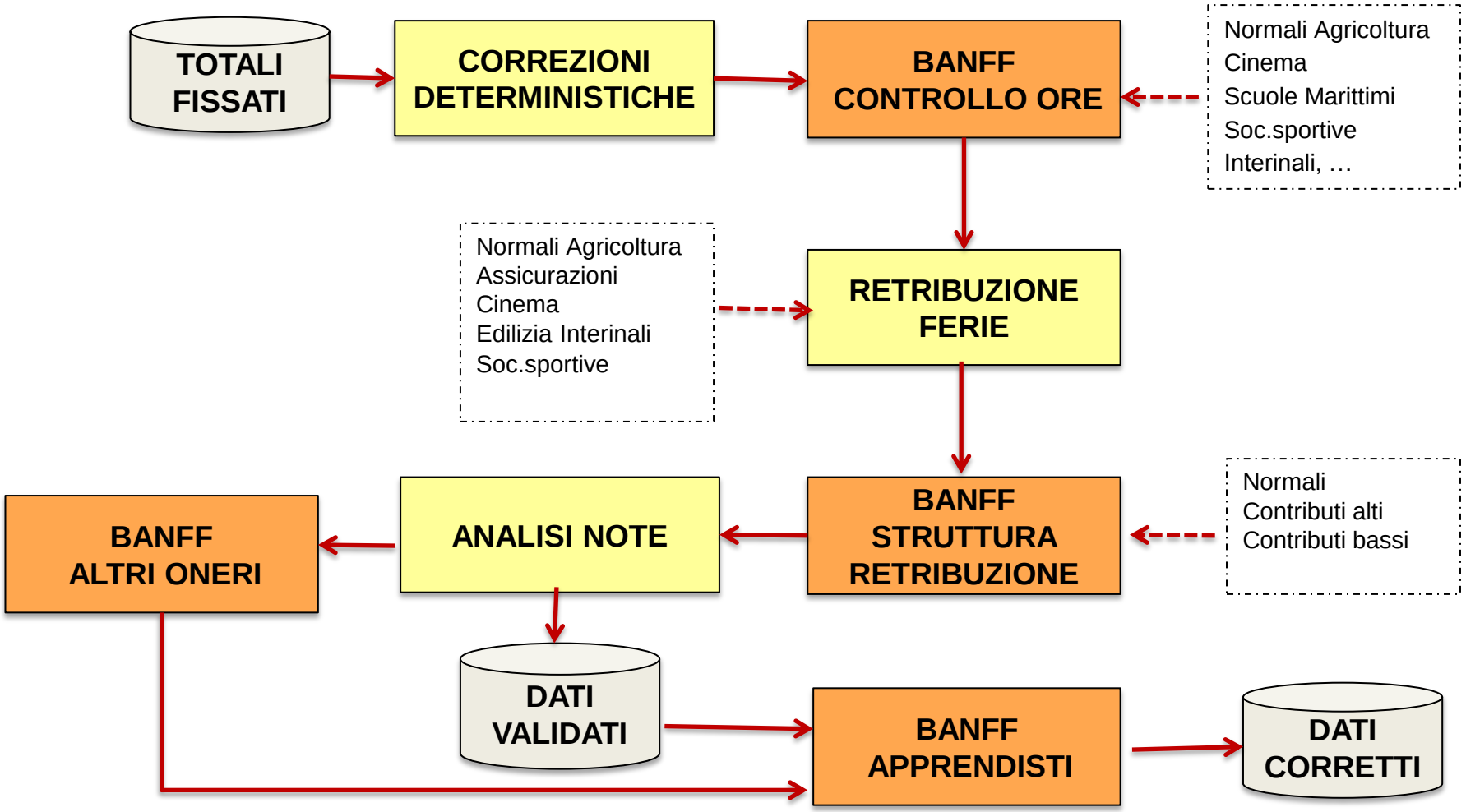
1. Le regole di incompatibilità (edit) devono essere soddisfatti cambiando il minimo numero di campi (principio del **minimo cambiamento**).

2. Le regole di imputazione devono essere derivate automaticamente da quelle di controllo.

3. Le distribuzioni marginali e congiunte dei dati non erronei devono essere preservate il più possibile.

I. P. Fellegi and D. Holt (1976) ”A Sistematic Approach to Automatic Edit and Imputation”, Journal of the American Statistical Association, vol 71, n353, 17-35.

Schema editing automatico



Correzioni probabilistiche

Metodologia Fellegi-Holt per le variabili quantitative

Software disponibili:

SAS

BANFF	insieme di procedure per l'individuazione di outlier e il controllo e la correzione di variabili numeriche continue sulla base di edit lineari

Pacchetti R

editrules	check e localizzazione sulla base degli edit
HotDeckImputation	imputazione col metodo del donatore
VIM	imputazione col metodo del donatore

Alcune correzioni deterministiche

- Riallineamento dati rilevati con ASIA per imprese soggette a trasformazioni giuridiche

- Problemi legati ai dati per ripartizioni
  - alcune correzioni di ripartizione errate
  - imprese che hanno centralizzato i valori delle retribuzioni e costo del lavoro

- Problemi legati a valori non imponibili
  - annullamento di valori non imponibili dove ‘impossibili‘
  - scambio tra imponibile e non imponibile..

- Azzeramento valori negativi indebiti

Correzioni deterministiche: le retribuzioni per ferie

- Per circa il 50% delle imprese valori mancanti o implausibilmente bassi per retribuzioni per ferie, festività e permessi
- Probabilmente a causa delle mancanza dell'informazione dai sistemi informativi
  - come per la nostra busta paga, in cui non c'è conteggio separato delle ferie
- Numero troppo elevato per una correzione con donatore
- **Ricostruzione a partire dalle ore retribuite per ferie, festività e permessi** (novità introdotta nel questionario)

$$retribuzioni\ per\ ferie, festività\ e\ permessi$$

$$= \frac{retribuzioni\ "regolari"}{n.\ ore\ ordinarie} n.ore\ retribuite\ per\ ferie, festività\ e\ permessi$$

Correzioni probabilistiche - Edit per BANFF

- correzioni probabilistiche per sezione di questionario, condizionate alle correzione precedente
- edit diversi per diversi raggruppamenti di contratti e segnali CIG
- definiti in un file excel e gestiti automaticamente dal programma SAS

sezione	tipo	Normali
ore	edit	oreretft <= 2350 * dipft;
ore	edit	oreretft >= 1200 * dipft;
ore	edit	oreretft + oreindmft + orecigft <= 2500 * dipft;
ore	edit	oreretft + oreindmft + orecigft >= 1400 * dipft;

Tabella EXCEL per la gestione degli Edit

Percorso differenziato di C&C in presenza di note

Esempio:' Nel settore agricolo viene applicato uno sgravio del 68% sui contributi a carico ditta per le aziende ubicate nelle aree di cui all'obiettivo 1 della UE'.

Fasi:

- Lettura ed analisi di tutte le note in corrispondenza di valori anomali (ad es. dell'incidenza contributiva)
- Estrazione di parole chiave (es. sgravio, fiscalizzazione, agevolazione, decontribuzione…)
- Procedura che fa uso di espressioni regolari per identificare le unità «da salvare»
- Percorso diversificato nell'ambito delle procedure di C&C

Analisi delle note presenti sul questionario. Esempi

Rapporto tra oneri sociali e retribuzioni molto alto

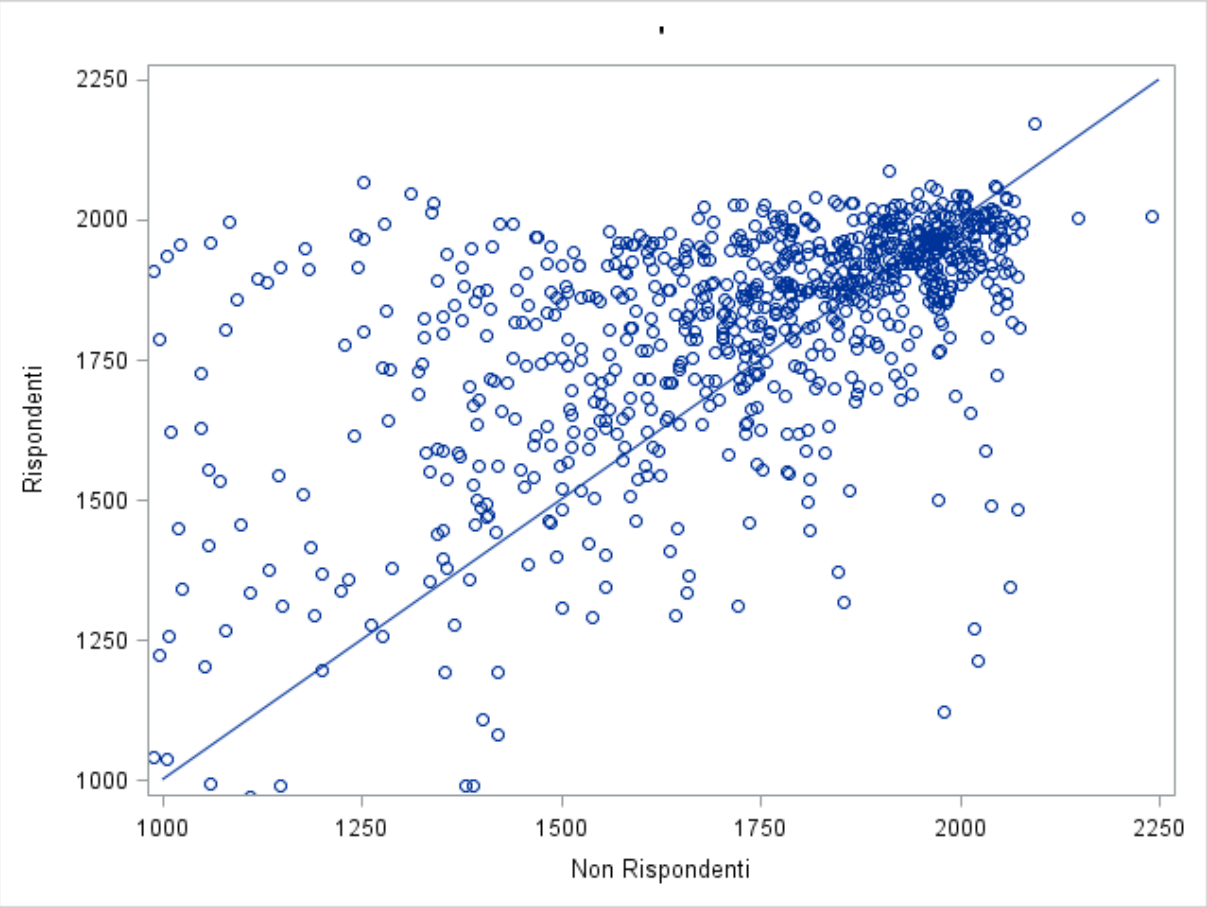
a. Elevata incidenza di personale in CIG (importi per TFR e Contributi sociali versati, e retribuzioni basse)
b. Personale in mobilità che ha usufruito di incentivi all'esodo
c. Imprese che implicano lavori rischiosi e, quindi, soggetti a premi assicurativi elevati contro gli infortuni sul lavoro

Caso di rapporto tra oneri sociali e retribuzioni molto basso

a. Settori che non prevedono il TFR (Attività artistiche, sportive) e/o con retribuzioni molto più elevate rispetto ai limiti fissati per la contribuzione ordinaria.
b. Imprese appartenenti a zone montane o altre aree 'particolari' che beneficiano di agevolazioni contributive per aree svantaggiate.

Analisi delle non risposte

Ore procapite RACLI – Rispondenti vs Non Rispondenti



Imputazione delle MRT

1. Le variabili principali sono state attribuite da RACLI:
   - Dipendenti totali, full-time e part-time, apprendisti
   - Ore retribuite totali, full-time e part-time, apprendisti
   - Retribuzioni imponibili totali e degli apprendisti

2. Le altre variabili sono attribuite con donatore di minima distanza, tramite rapporto con le variabili principali

La distanza è calcolata su: retribuzioni imponibili, dipendenti ed ore retribuite, FT,PT e apprendisti, flag che segnala la situazione di alta CIG

In gruppi formati da imprese con:
- stesso raggruppamento di CCNL
- flag di struttura occupazionale (presenza di tipi di lavoratori - FT,PT ed apprendisti)

Controllo degli errori influenti residui (Macro – Editing)

Macro editing effettuato in maniera Top-down sui seguenti indicatori:

- Ore lavorate per dipendente
- Retribuzioni orarie
- incidenza dei contributi sociali totali rispetto alle retribuzioni;

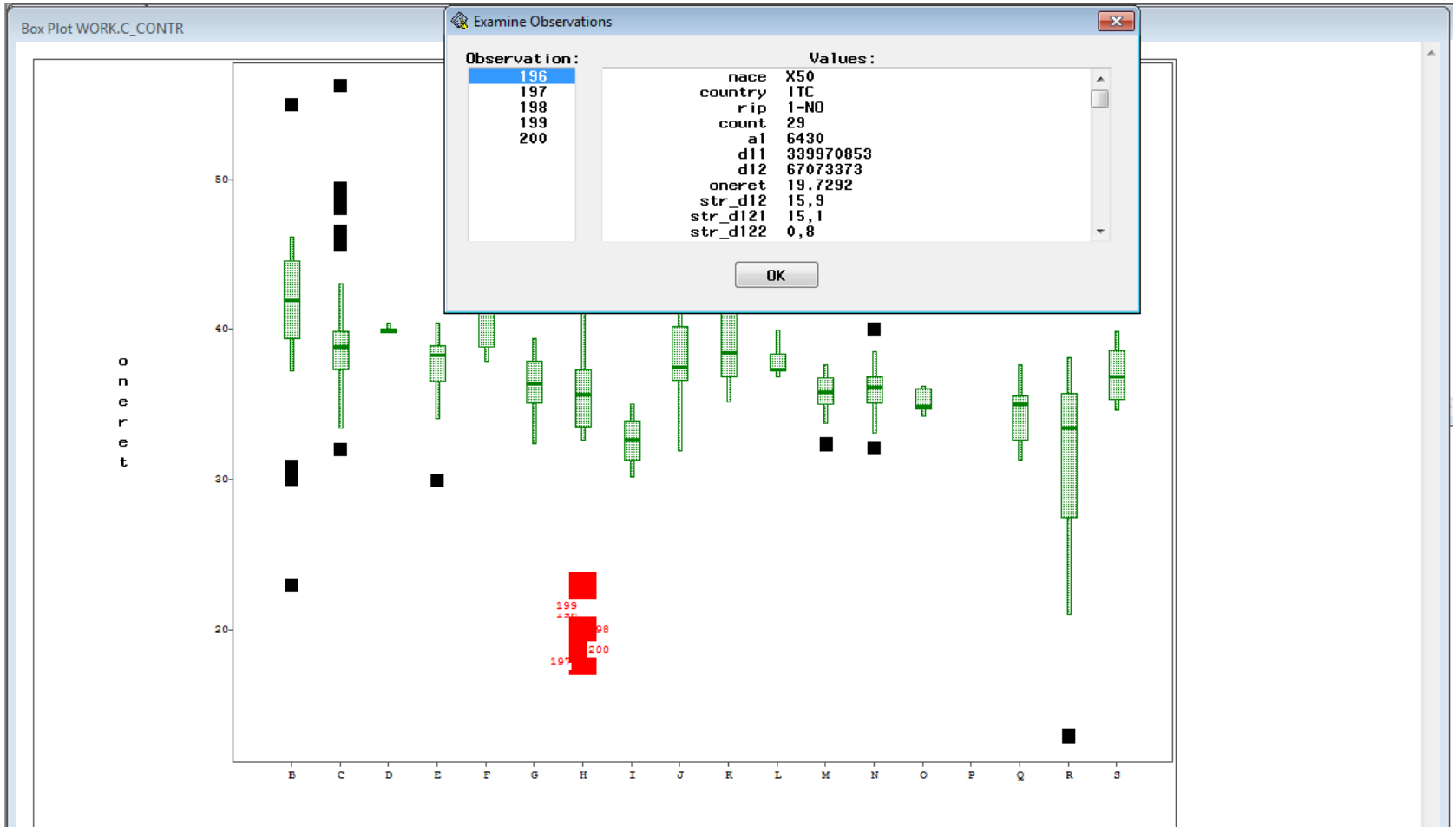
Scopo: individuare unità ancora affette da possibili errori di misura ovvero giustificare le anomalie

Fasi:

- Analisi grafica delle distribuzioni di parametri per celle di stima
  - Divisione ateco *x* ripartizione x classe dimensionale
- Selezione di celle «outlier»
- Analisi delle unità all'interno delle celle, partendo dalle più influenti

Controllo degli errori influenti residui (Macro – editing)

Selezione degli aggregati outliers



Controllo degli errori influenti residui (Macro – editing)

Analisi degli aggregati outliers

- Nell'esempio, Divisione 50 «Trasporto Marittimo e per vie d'acqua».
- Particolare per specificità dovute alla tipologia contrattuale
  - molte società beneficiano di riduzioni contributive.
- Dati ritenuti validi

ALTRI ESEMPI DI SETTORI PARTICOLARI:

- Unità appartenenti al settore «Stampa e Servizi connessi alla stampa» .
  - Il CCNL prevede un contributo obbligatorio aggiuntivo ad un fondo specifico.
- Sezione «R» – gli outliers corrispondono ad alcune società di calcio (alta retribuzione, bassi contributi).

Riassumendo…il valore aggiunto dell'uso di RACLI

- Confronto di variabili compresenti nell'indagine e nel registro ed individuazione valori anomali
  - Ricontatto delle imprese influenti
  - Sostituzione con i valori RACLI per le imprese non influenti

- Imputazione delle MRT

- Gruppi di edits e di imputazione per raggruppamenti di Contratti collettivi nazionali (CCNL), invece che Ateco. In teoria più determinanti per le variabili target

- Informazioni più affidabili sulla CIG per differenziare le procedure di C&C a seconda dell'intensità di CIG
  - Es. Edits sulle ore e sui contributi differenziati per imprese ad alta cig
  - Gruppo di imputazione differenziato

Grazie
per la pazienza!