

Roma, 16 Aprile 2025

Povert  multidimensionale. I dati sulla distribuzione congiunta di reddito, consumo e ricchezza delle famiglie

Istat | Aula Magna | via Cesare Balbo, 14 Roma

Il metodo Istat: EU-SILC come archivio di riferimento

Marcello D'Orazio

Istat | Direzione centrale per la metodologia e il disegno dei processi statistici (DCME)

Statistical Matching all'Istat

«**Statistical matching**» (SM) : un ampio insieme di metodi per integrare due fonti dati (tipicamente indagini campionarie rappresentative della stessa popolazione obiettivo) per **studiare la relazione tra variabili non osservate congiuntamente**

Lunga esperienza Istat su SM:

- primi esperimenti avviati nel 2006 (Coli et al, 2006)
- lavori su metodi di SM (D'Orazio, Di Zio e Scanu, 2006a; 2006b; 2017; 2019, 2024)
- applicazioni di matching SILC-HBS (Donatiello et al, 2014; 2016; 2022)
- Sviluppo software: package R **StatMatch** (D'Orazio, 2025)

Statistical Matching: esercizio Istat (1/2)

Obiettivo: studiare la relazione tra **Reddito**, **Spese** e **Ricchezza** (ICW: *Income, Consumption and Wealth*) delle famiglie italiane

Difficoltà/vincoli nell'approccio Istat:

- Le indagini da integrare sono tre (e non due): SILC, HBS e SHIW
- E' necessario creare un dataset «sintetico»: dataset che includa tutte le variabili di interesse (oltre ad altre rilevanti per le analisi finali) → da fornire ad Eurostat
- Il dataset sintetico deve essere creato a partire da SILC (*recipient*) → setting «preferito» da Eurostat

Primo esperimento condotto con dati del 2016 (Donatiello et al, 2025)

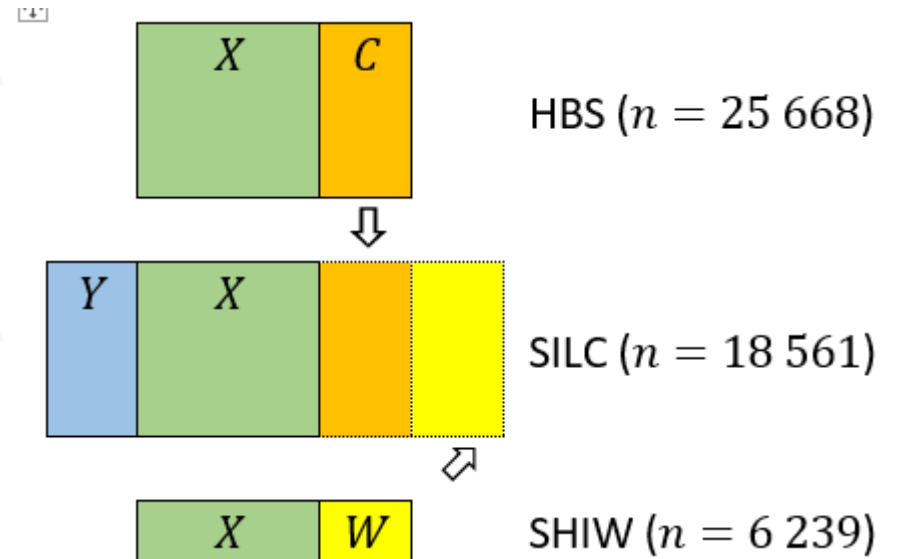
Statistical Matching: esercizio Istat (2/2)

- Statistical Matching a livello Micro

- SILC è il *recipient* in cui imputare:

Fase 1) spese totali (C) da HBS (donor 1)

Fase 2) ricchezza (W) da SHIW (donor 2)



- Metodo di imputazione: si imputa valore osservato su **donatore a distanza minima** (*Nearest Neighbour Donor*)

SM Fase 1: Matching SILC-HBS

Si imputano in SILC le spese totali (C) osservate in HBS. Si utilizza **NND «modificato»**:

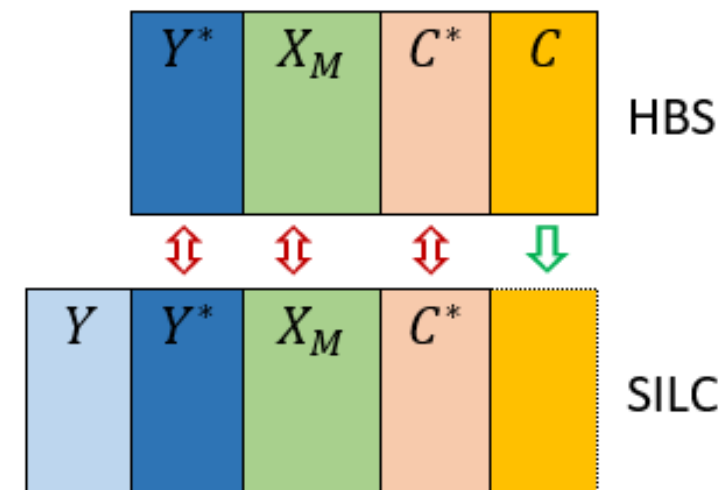
1.1) famiglie divise in sotto-insiemi in base a:

- Ripartizione geografica (5 cat) e Titolo godimento abitazione (2 cat)

1.2) in ogni sotto-insieme si calcola distanza SILC-HBS con **dist. di Gower «robusta»** su:

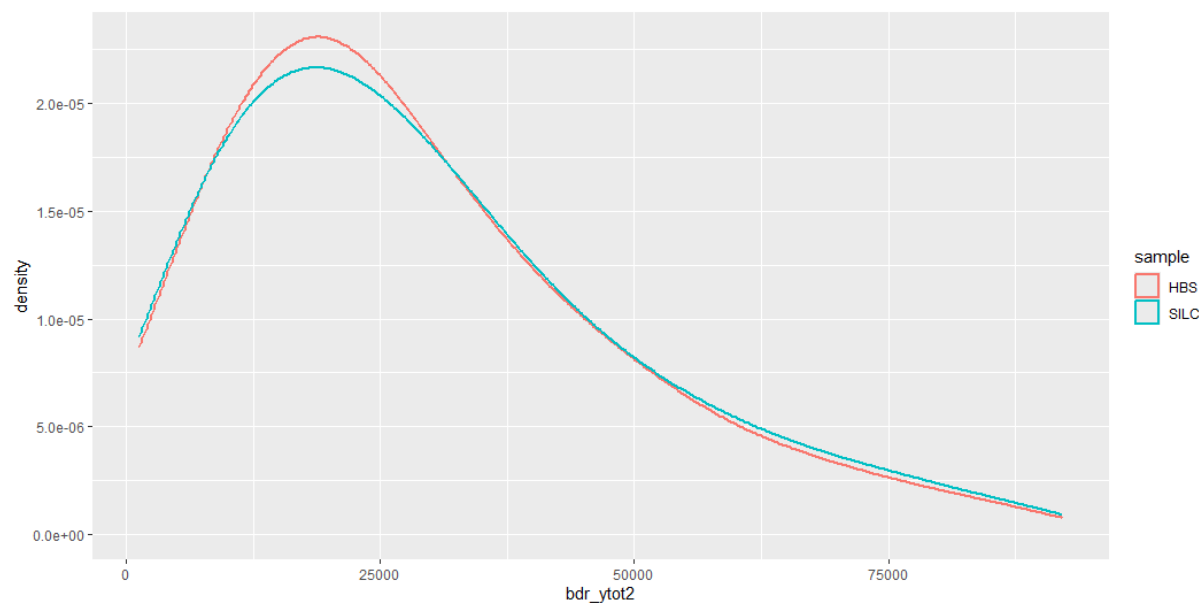
- **Reddito da archivio** (Y^*) (aggiunto con record linkage a entrambi) (pos. su ecdf)
- **«proxy» spese tot** (C^*) (modulo ad hoc in SILC) (pos. su ecdf)
- Metri quadri abitazione
- No. Occupati in famiglia (0-3)
- No. Minorenni in famiglia (0-3)
- No. Anziani in famiglia (0-3)
- Livello istruzione pers. riferimento (4 cat)

1.3) per ogni fam. SILC si individuano le $k=4$ fam HBS più vicine e tra queste si sceglie quella che minimizza la distanza in valore assoluto tra decile reddito e decile di spesa

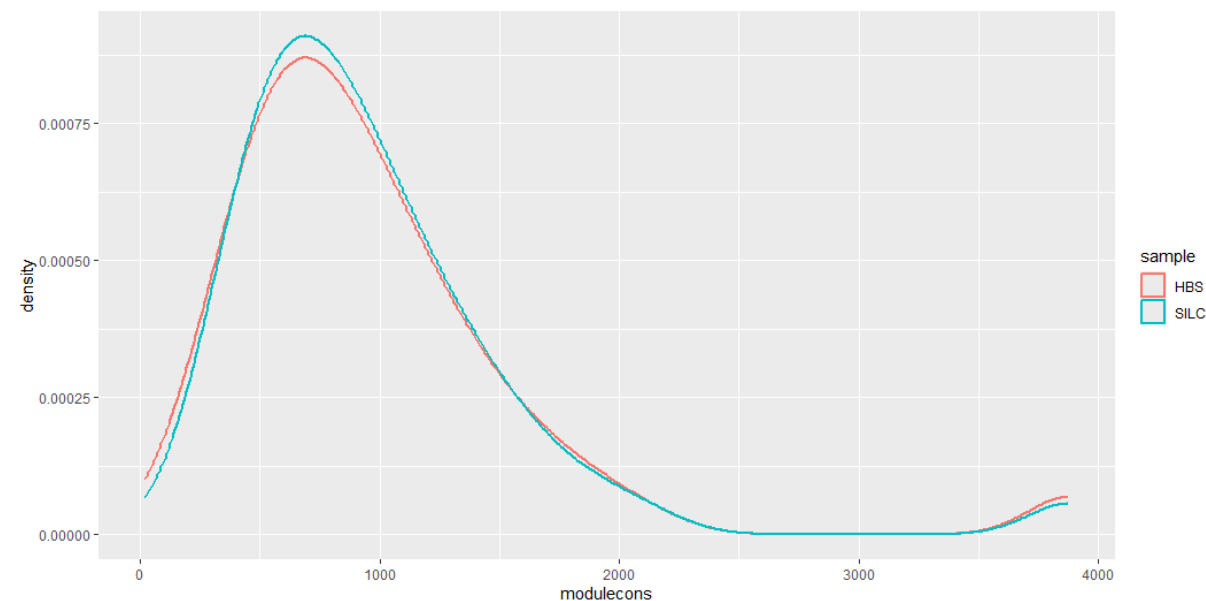


SM Fase 1 - SILC-HBS: variabili chiave

Reddito da archivio (Y^*) (aggiunto con record linkage a entrambi)

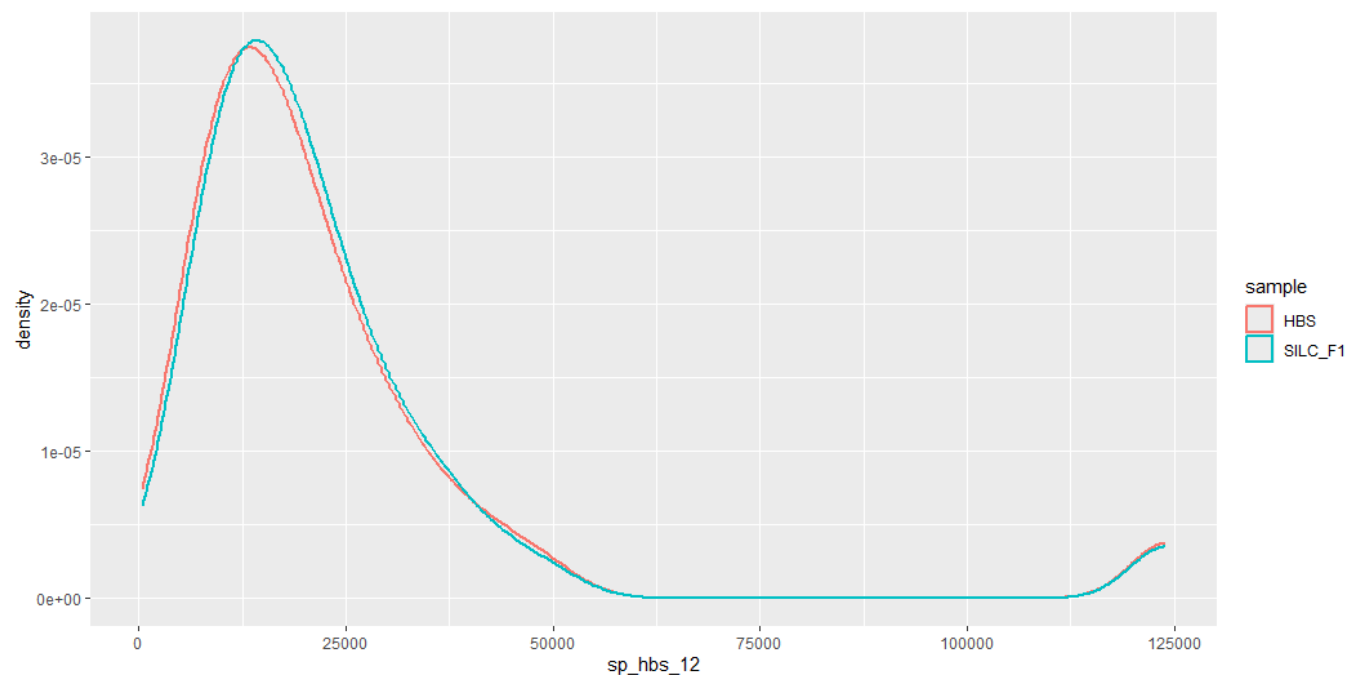


«proxy» spese tot (C^*) (modulo ad hoc in SILC)

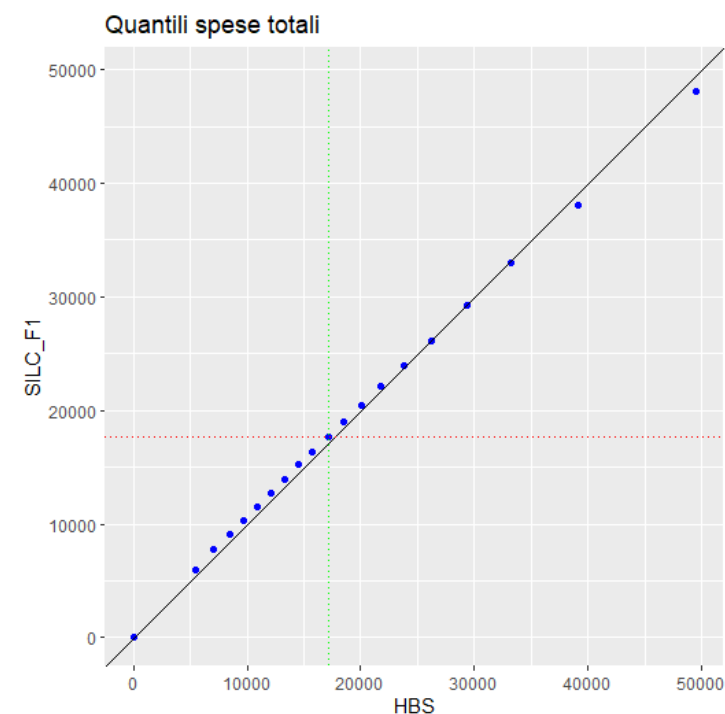


SM Fase 1 - SILC-HBS: distr. Marginale Spese Imputate in SILC

Distribuzione marginale spese imputate in SILC Vs. distr. spese in HBS



Quantili spese imputate in SILC Vs. quantili in HBS



SM Fase 2: matching SILC_F1-SHIW

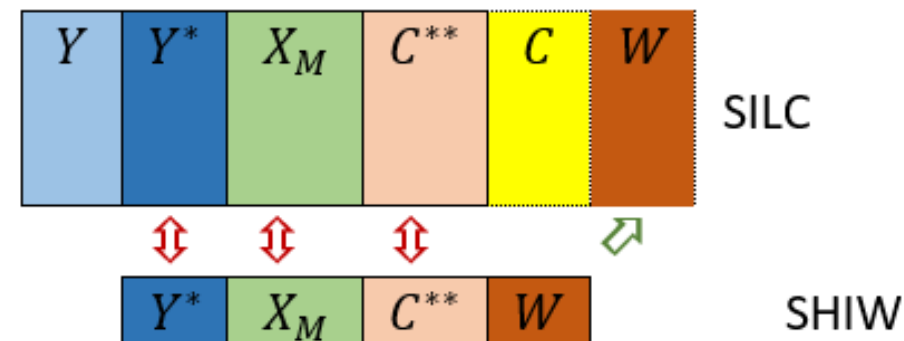
Si imputa in SILC la ricchezza (W) osservata in SHIW. Si utilizza **NND**:

2.1) famiglie divise in sotto-insiemi in base a:

- Ripartizione geograf. (5 cat), Titolo godimento abitazione (2 cat), rischio povertà (0,1)

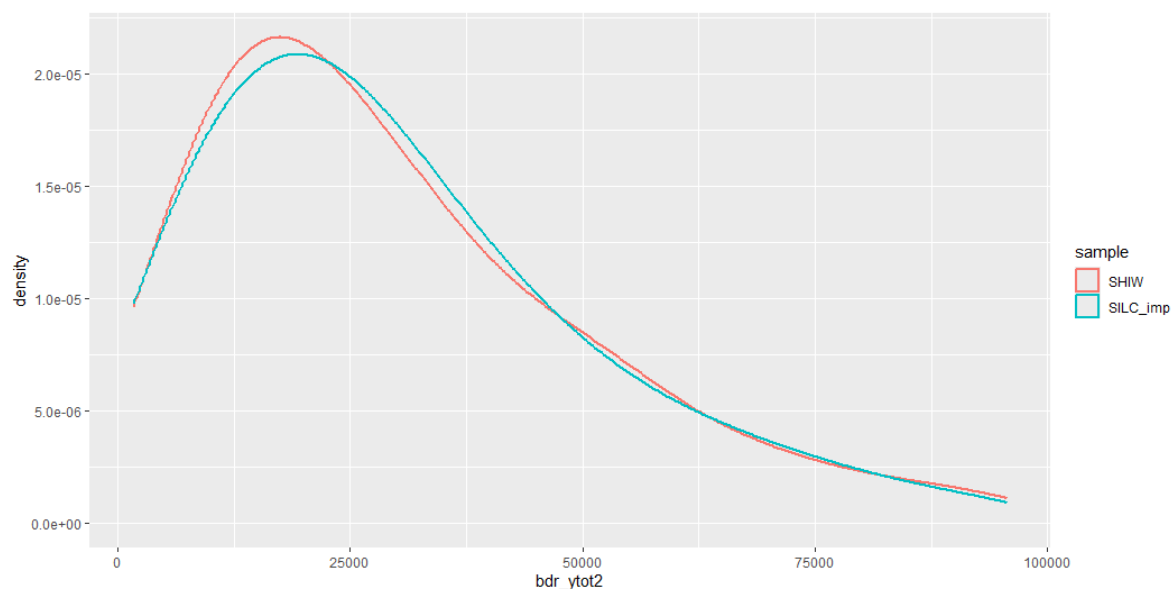
2.2) calcolo distanza tra fam. SILC e fam. SHIW con **dist. di Gower «robusta»** su:

- **Reddito da archivio** (Y^*) (aggiunto con record linkage a entrambi)
- **spese per alimentari e utenze** (C^{**})
- Metri quadri abitazione
- Affitti figurativi

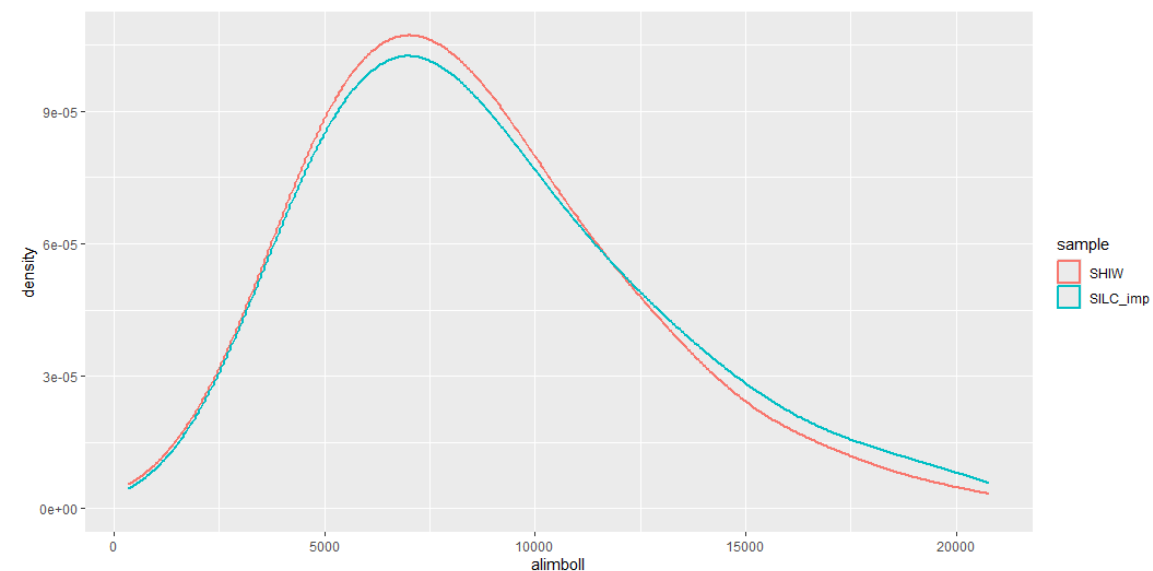


SM Fase 2 – SILC_F1-SHIW: variabili chiave

Reddito da archivio (Y^*) (aggiunto con record linkage a entrambi)

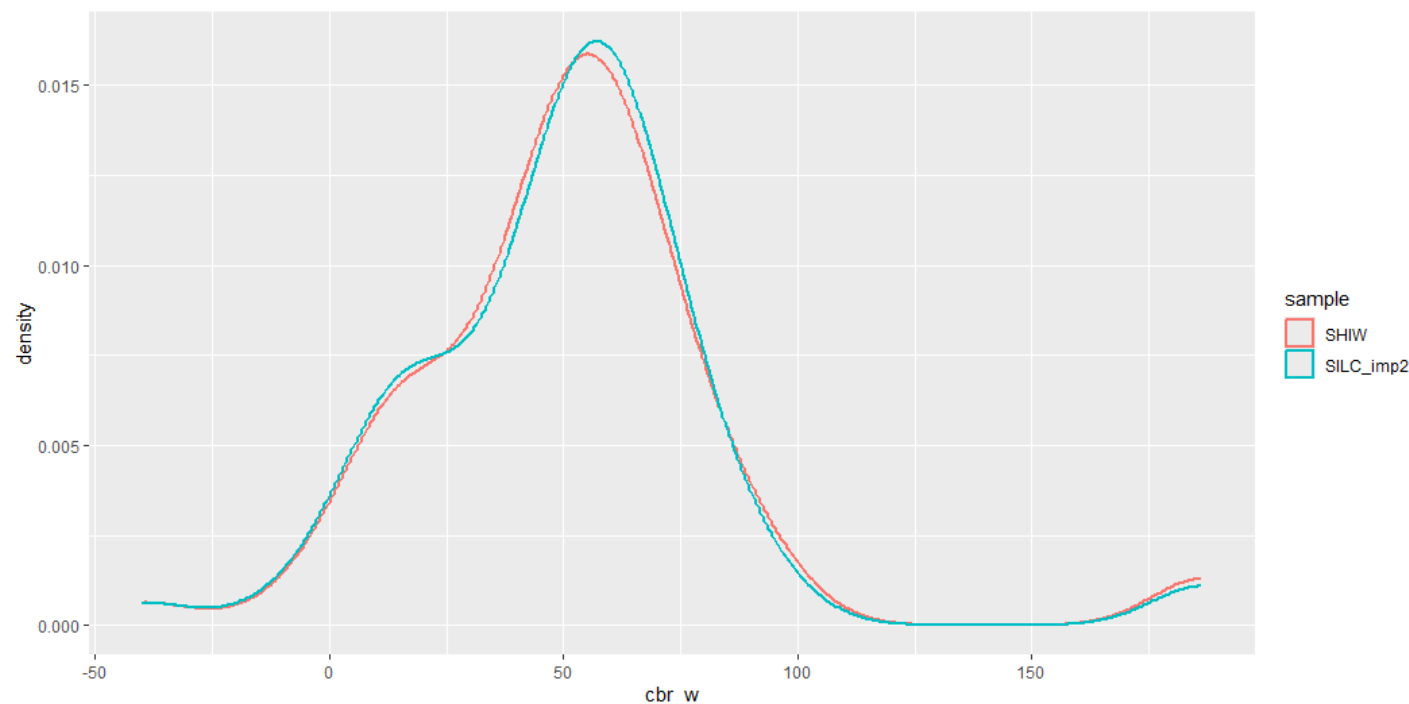


spese alimentari e utenze (C^{**})

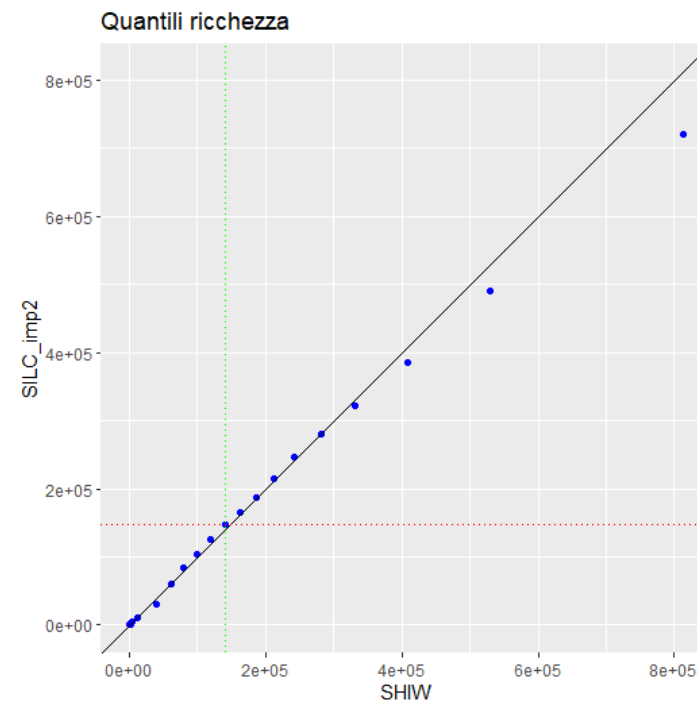


SM Fase 2 – SILC_F1-SHIW: distr. Marg. Ricchezza Imp. in SILC

Distribuzione marginale ricchezza imputata in SILC Vs.
distr ricchezza in SHIW (cubic root)



Quantili ricchezza imputata in SILC Vs.
quantili in SHIW



SM esercizio Istat: Conclusioni

- L'applicazione di SM ai dati delle indagini riferiti al 2020 è stata più problematica di quella con dati 2016 (si veda Donatiello et al, 2025), i dati 2020 scontano diverse problematiche legate al periodo pandemico
- Il setting «suggerito» da Eurostat pone la necessità di due matching distinti, cosa non usuale e più complessa rispetto ad un singolo step di matching (SILC è molto più grande di SHIW)
- Il matching SILC-HBS è più «semplice» poiché già «previsto» in fase di disegno di SILC (raccolta info sui consumi con un modulo ad hoc del questionario SILC)
- Il matching SILC_F1-SHIW sconta qualche difficoltà dovuta all'assenza di una proxy della ricchezza sebbene si disponga di ottimi predittori (legati all'abitazione, ecc.)
- Il dataset finale sintetico (SILC integrato con spese e ricchezza) agevola le analisi complesse ma NON permette di includere nelle analisi le variabili non usate nel processo di matching

grazie

Marcello D'Orazio | marcello.dorazio@istat.it

Statistical Matching: alcuni riferimenti bibliografici

- Coli, A., Tartamella, F., Sacco, G., Faiella, I., D'Orazio, M., Di Zio, M., Scanu, M., Siciliani, I., Colombini, S., Masi, A. (2006) "La costruzione di un archivio di microdati sulle famiglie italiane ottenuto integrando l'indagine ISTAT sui consumi delle famiglie italiane e l'Indagine Banca d'Italia sui bilanci delle famiglie italiane". *ISTAT, Documenti*, N. 12/2006.
- Donatiello G., D'Orazio M., Frattarola D., Rizzi A., Scanu M., Spaziani M. (2014) "Statistical Matching of Income and Consumption Expenditures", *International Journal of Economic Sciences*, Vol. III, No.3/2014
- Donatiello G., D'Orazio M., Frattarola D., Rizzi A., Scanu M., Spaziani M. (2016) "The role of the conditional independence assumption in statistically matching income and consumption", *International Journal of the IAOS*, 32, pp. 667-675
- Donatiello G., D'Orazio M., Frattarola D., Spaziani M. (2022) "The joint distribution of income and consumption in Italy: an in-depth analysis on statistical matching". *Rivista di Statistica Ufficiale - Review of Official Statistics*, N. 3/2022, pp. 77-109
- Donatiello G., D'Orazio M., Neri A., Loschiavo D., Tullio F. (2025) "The relationship between income, consumption and wealth: methods and results of a first experimental integration of multiple households surveys in Italy". *Statistical Journal of the IAOS*, DOI: 10.1177/18747655251315305
- D'Orazio M (2025). *StatMatch: Statistical Matching or Data Fusion*. R package version 1.4.3, <<https://CRAN.R-project.org/package=StatMatch>>
- D'Orazio, M and Di Zio, M and Scanu, M (2006a), "Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints", *Journal of Official Statistics*, 22, pp. 137-157.
- D'Orazio, M, Di Zio, M and Scanu, M (2006b) *Statistical Matching: Theory and Practice*. Wiley, Chichester
- D'Orazio M., Di Zio M., Scanu M. (2017) "The use of uncertainty to choose matching variables in statistical matching". *International Journal of Approximate Reasoning*, 90 (2017), pp. 433-440
- D'Orazio M., Di Zio M., Scanu M. (2019) "Auxiliary variable selection in a statistical matching problem", in Zhang L.C. and Chambers R.L (eds.) *Analysis of Integrated Data*. Chapman and Hall/CRC, pp. 101-120
- D'Orazio M., Di Zio M., Scanu M. (2024) "Ask the Experts: State of play on statistical matching with a focus on auxiliary information, complex survey designs and quality issues". *The Survey Statistician*, 89, pp. 47-58

SM Fase 1 - SILC-HBS: variabili chiave e correlazione con target

Correlazioni su SILC

Spearman

	C*	Y*
Y*	0.4914	
Y=HY020	0.5607	0.8435

Pearson

	log (C*)	log (Y*)
log (Y*)	0.3829	
log (Y)	0.3600	0.5500

Correlazioni su HBS

Spearman

	C*	Y*
Y*	0.4889	
C=spese_HBS	0.8470	0.4802

Pearson

	log (C*)	log (Y*)
log (Y*)	0.3942	
Log (C)	0.6921	0.3523

SM Fase 1 - SILC-HBS: Valutazione Approx. Incertezza

Considerando:

- Trasn. Log per le variabili continue
- Distribuzione congiunta Gaussiana Multivariata

Con i dati a disposizione si stima (approssimativamente) che

$$-0,05 \leq \rho_{\log(Y), \log(C)} \leq 0,74$$

Con valore centrale $\tilde{\rho}_{\log(Y), \log(C)} = 0,35$

Ossia la stima sotto l'ipotesi di indipendenza tra Y e C condizionatamente alle prescelte variabili di matching

E' un intervallo ampio con notevole incertezza motivo per cui è stato modificato il metodo NND con una sorta di k-NN

SM Fase 2 – SILC_F1-SHIW: correlazioni variabili chiave

Correlazioni con Z=W su SHIW

Spearman

	irs	mq	C**	Y*
mq	0.5642			
C**	0.5227	0.4606		
Y*	0.6044	0.4738	0.6409	
W	0.8000	0.6060	0.5512	0.6370

Pearson

	log(irs)	log(mq)	log(C**)	log(Y*)
log(mq)	0.3930			
log(C**)	0.3251	0.4622		
log(Y*)	0.3535	0.4325	0.5867	
log(W)	0.6311	0.4531	0.4448	0.4679