VAI - Vulnerabilità abitativa e di salute degli Anziani in Italia

WP2 - Indici sintetici VAA/VSA a livello territoriale

D 2.3
Indici sintetici VAA/ VSA a livello macro-territoriale: applicazione tecniche di matching statistico (CEM)

M: Definizione e costruzione indici di VAA e VSA a livello macro-territoriale

A cura di:

Mariateresa Ciommi, Marco Arlotti, Giulia Bettin, Barbara Ermini, Francesca Mariani, Maria Cristina Recchioni, Elena Spina
Dipartimento di Scienze Economiche e Sociali (DiSES), Università Politecnica delle Marche

Luigi Bernardi, Isabella Giorgetti, Matteo Luppi, Gianluca De Angelis

Antonello Alici
Dipartimento di Ingegneria Civile, Edile e Architettura (DICEA), Università Politecnica delle Marche

Emma Espinosa
Dipartimento di Scienze Cliniche e Molecolari (DISCLIMO), Università Politecnica delle Marche

Come citare questo rapporto:
Ciommi, M., M. Arlotti, G. Bettin, B. Ermini, F. Mariani, M. C. Recchioni, E. Spina, L. Bernardi, I. Giorgetti, M. Luppi, G. De Angelis, A. Alici, E. Espinosa (2025), *Indici sintetici VAA/ VSA a livello macro-territoriale: applicazione tecniche di matching statistico (CEM)*, WP2, progetto VAI - Vulnerabilità abitativa e di salute degli Anziani in Italia, Università Politecnica delle Marche.

VAI è una ricerca del progetto "**Age it- Ageing well in an ageing society (AGE-IT)**", codice progetto **PE0000015**, CUP **B83C22004800006**, finanziato nell'ambito del **Piano Nazionale di Ripresa e Resilienza**, **Missione 4** "**Istruzione e Ricerca**" – Componente 2 "dalla Ricerca all'Impresa" – Investimento 1.3, finanziato dall'Unione Europea – NextGenerationEU.

Indice

PARTE 1: LA COSTRUZIONE DEL DATASET INTEGRATO 5

1. **Introduzione** 6
2. **Il matching in assenza di una chiave comune** 7
3. **Il matching statistico** 9
4. **Il Coarsened Exact Matching (CEM)** 11
5. **EHIS + AVQ** 13
6. **EHIS_AVQ+SHARE** 23
7. **EHIS_AVQ_SHARE+EUSILC** 29
8. **EHIS_AVQ_SHARE_EUSILC+SPESE** 41
9. **Commenti conclusivi al matching** 49

PARTE 2: COSTRUZIONE DI VAA E VSA 50

1. **L'indice di Vulnerabilità Abitativa per gli Anziani** 51
2. **L'indice di Vulnerabilità di Salute per gli Anziani** 56

PARTE 3: APPENDICI 63

APPENDICE 1: Il pacchetto MatchIT 64

APPENDICE 2: Lettura dell'output di MatchIT 68

APPENDICE 3: Output del Matching tra EHIS e AVQ 72

APPENDICE 4: Output del Matching tra EHIS_AVQ e SHARE 78

APPENDICE 5: Output del Matching tra EHIS_AVQ_SHARE e EUSILC 86

APPENDICE 6: Output del Matching tra EHIS_AVQ_SHARE_EUSILC e SPESE 90

Introduzione e struttura del documento

Questo documento restituisce il lavoro condotto con riferimento all'attività di ricerca AT 2.3 Definizione e costruzione, a livello geografico aggregato, di due indici sintetici di VAA e VSA, D 2.3, Indici sintetici VAA/ VSA a livello macro territoriale, Milestone M Definizione e costruzione indici di VAA (vulnerabilità abitativa anziani) e VSA (vulnerabilità salute anziani) a livello macro-territoriale.
Più nello specifico, la definizione ed elaborazione degli indici sintetici di VAA e VSA si è basata sua costruzione di un dataset integrato, attraverso l'applicazione di tecniche di matching (CEM).
In questo modo sono stati integrati, in modo originale ed innovativo, i cinque dataset considerati nel progetto (EHIS, AVQ, SHARE, EUSILC, SPESE), in modo da rendere possibile un'analisi empirica, quanto più multidimensionale, delle condizioni di vulnerabilità, sotto un profilo abitativo e di salute, della popolazione anziana
Il periodo di riferimento considerato è il 2019.
Il documento si struttura in 2 Parti principali.
Nella Parte 1, viene ricostruito tutto il percorso di analisi effettuato in termini di costruzione del dataset integrato. Si, considerano, innanzitutto le diverse opzioni metodologiche e il procedimento di analisi, in ultima istanza, adottato.
Viene, inoltrato, presentato il procedimento empirico di aggregazione dei diversi dataset. A questo proposito, il dataset EHIS è stato considerato come dataset di partenza, avendo informazioni particolarmente dettagliate per il dominio salute, nonché informazioni di dettaglio anche per quanto riguarda l'analisi delle condizioni abitative.
L'aggregazione, in base alla selezione di alcune variabili specifiche per ciascun dataset, è stata sviluppata in quattro passaggi: 1) EHIS + AVQ; 2) EHIS_AVQ+SHARE; 3) EHIS_AVQ_SHARE+EUSILC; 4) EHIS_AVQ_SHARE_EUSILC+SPESE.
A partire dalla ricostruzione del dataset integrato, nella Parte 2 del documento vengono presentati gli esiti empirici dell'analisi, in termini di elaborazione di due indici di VAA (vulnerabilità abitativa anziani) e VSA (vulnerabilità salute anziani).
Completano il documento la sezione Appendici, con 5 allegati di dettaglio rispetto al pacchetto di analisi, la lettura dell'output e gli output specifici per ciascuno dei matching eseguito.

PARTE 1: LA COSTRUZIONE DEL DATASET INTEGRATO

1. Introduzione

La fusione di dataset rappresenta una sfida critica in molti ambiti della ricerca scientifica, soprattutto quando si mira a integrare informazioni provenienti da fonti disparate per ottenere una visione più completa e robusta di un fenomeno.

Uno dei problemi principali nel processo risiede nella mancanza di identificatori univoci, le cosiddette "chiavi" (ID) che, quando presenti, permettono di agganciare in maniera univoca i due dataset. Quando appunto questo identificatore univoco è assente, l'allineamento preciso delle osservazioni diventa arduo.

A questo si aggiunge il problema della gestione delle osservazioni mancanti o incoerenti tra i dataset. Questo può portare a una riduzione significativa della dimensione campionaria utile e alla conseguente introduzione di un bias nei risultati.

Occorre, quindi, utilizzare metodologie robuste per garantire l'integrità e l'usabilità del dataset combinato.

Nella prima parte di questo report vedremo una metodologia utile per integrare dataset differenti quando non è presente un ID comune.

2. Il matching in assenza di una chiave comune

Il problema dell'integrazione di dataset in assenza di una **chiave comune esplicita** rappresenta una delle sfide più complesse nella preparazione dei dati scientifici.

Quando non esistono identificatori univoci e condivisi tra i diversi set di dati, le tecniche di *join* tradizionali non sono applicabili, costringendo il ricercatore a esplorare approcci alternativi che richiedono maggiore cautela.

In questi scenari, le metodologie si orientano verso la ricerca di corrispondenze implicite o la fusione basata su presupposti posizionali.

La letteratura propone differenti metodi:

- **Basarsi su Chiavi Sintetiche o Derivate**. In assenza di una chiave preesistente, è talvolta possibile **creare una chiave sintetica** o derivarla da variabili multiple presenti in entrambi i dataset. Ad esempio:
    - **Combinazione di Variabili:** se due dataset contengono entrambi "Nome", "Cognome" e "Data di Nascita" di un individuo, queste tre variabili possono essere combinate per formare una chiave composita (es. "MarioRossi1980-01-15").
    - **Generazione di ID:** Se i dataset si riferiscono allo stesso insieme di entità (es. pazienti, campioni), ma non hanno un ID comune, potrebbe essere necessario generare identificativi unici e poi associare manualmente, o tramite algoritmi di *matching*, le righe corrispondenti.
- **Fusione per Posizione (Bind Orizzontale)**. Questa tecnica è apparentemente semplice da implementare, ma occorre verificare che l'ordine delle righe in entrambi i dataset sia **perfettamente allineato.** In questo modo, **l**e colonne di un dataset vengono semplicemente aggiunte accanto alle colonne dell'altro. Non c'è alcuna verifica basata sui valori dei dati; la corrispondenza è puramente posizionale.
- **Matching Fuzzly (Fuzzy Matching).** Quando le chiavi non sono perfettamente identiche, ma presentano piccole variazioni (errori di battitura, formati leggermente diversi, abbreviazioni), si può ricorrere al *fuzzy matching*. Invece di richiedere una corrispondenza esatta, gli algoritmi di *fuzzy matching* calcolano una "somiglianza" tra le stringhe delle chiavi (es. "Mario Rossi" e "M. Rossi"). Se il livello di somiglianza supera una soglia predefinita, le righe vengono considerate corrispondenti.
- **Tecniche di Integrazione Dati Avanzate.** In contesti di big data o quando la complessità dei dati è elevata, si possono adottare soluzioni più sofisticate, quali:
    - **Record Linkage / Entity Resolution:** Si tratta di processi sistematici per identificare record che si riferiscono alla stessa entità attraverso diversi dataset, anche in assenza di chiavi univoche. Spesso coinvolgono più passaggi, inclusi la normalizzazione dei dati, il blocco (per ridurre il numero di confronti), il *matching* probabilistico e la revisione manuale.

- **Uso di Grafo di Conoscenza (Knowledge Graphs):** Per dati altamente interconnessi, un grafo di conoscenza può aiutare a mappare le relazioni tra entità provenienti da diverse fonti, facilitando l'integrazione semantica.

In questo contesto, verranno utilizzate nella presente analisi le tecniche di Record Linkage avanzate. In particolar modo, ricorreremo a tecniche di Matching.

3. Il matching statistico

Il matching statistico è una metodologia non parametrica ampiamente impiegata nell'inferenza causale, particolarmente in contesti osservazionali dove l'assegnazione randomizzata a gruppi di trattamento e controllo è impraticabile o eticamente non sostenibile.

L'obiettivo primario del matching è replicare, per quanto possibile, le condizioni di un esperimento randomizzato attraverso la selezione di unità di controllo (o comparazione) che siano il più simili possibile alle unità di trattamento rispetto a un insieme di covariate pre-trattamento osservate. Questo processo mira a bilanciare le distribuzioni delle covariate tra i gruppi, riducendo il bias di selezione e consentendo un'inferenza causale più robusta sull'effetto del trattamento.

Il matching crea per ogni unità trattata una controparte non trattata con covariate simili.

Diverse strategie di matching sono state sviluppate per creare gruppi bilanciati:

- **Matching Esatto (Exact Matching):** Consiste nell'abbinare unità di trattamento e controllo che hanno valori identici per tutte le covariate rilevanti. Pur essendo ideale per eliminare il bias, è raramente fattibile in pratica con un numero elevato di covariate o con covariate continue.
- **Matching per Propensity Score (PSM)**: Introdotto da Rosenbaum e Rubin (1983), il PSM è la metodologia più diffusa. Anziché abbinare direttamente sulle covariate multidimensionali, si abbinano le unità basandosi su una singola variabile scalare: il propensity score che rappresenta la probabilità di ricevere il trattamento condizionata sulle covariate osservate. Il PSM riduce un problema di matching multidimensionale a uno unidimensionale. Le tecniche di abbinamento basate sul PS includono:
  - *Nearest Neighbor Matching:* Ogni unità trattata viene abbinata alla/e unità di controllo con il propensity score più vicino.
  - *Caliper Matching*: Simile al nearest neighbor, ma impone un limite (caliper) sulla massima differenza ammissibile nel propensity score per un match.
  - *Kernel Matching*: Utilizza una media ponderata degli outcome delle unità di controllo, dove i pesi dipendono dalla vicinanza del loro propensity score a quello dell'unità trattata.
  - *Stratification Matching*: Divide il campione in strati (quintili, decili) basati sul propensity score, e stima l'effetto del trattamento all'interno di ogni strato, combinando poi i risultati.
- **Coarsened Exact Matching (CEM)**: Questa tecnica, sviluppata da Imai e van Dyk (2005), non si basa sul propensity score, ma opera direttamente sulle covariate. Il CEM discretizza (o "*coarsen*") le covariate continue in categorie significative e poi effettua un matching esatto su queste covariate coarsened. La sua forza risiede nella capacità di preservare il bilanciamento esatto sulle covariate coarsened, garantendo al contempo un bilanciamento approssimativo sulle covariate originali e riducendo la dipendenza da modelli statistici (come la stima del propensity score). Fornisce, inoltre, un'ottima interpretabilità e trasparenza del processo di matching.

- **Optimal Matching**: Mira a minimizzare la distanza totale tra le unità trattate e di controllo, definita su una metrica specifica (es. distanza di Mahalanobis o differenza nei propensity scores). Questo può essere implementato utilizzando algoritmi di ottimizzazione lineare o network flow.
- **Genetic Matching**: Un approccio che utilizza un algoritmo genetico per trovare il set di pesi per ciascuna covariata (o il propensity score) che minimizza il bilanciamento complessivo tra i gruppi di trattamento e controllo, misurato tramite statistiche di bilanciamento (es. differenze standardizzate delle medie).

Tra i vari metodi, ricorreremo alle tecniche di Matching statistico applicando il CEM.

4. Il Coarsened Exact Matching (CEM)

Il Coarsened Exact Matching (CEM) è una metodologia avanzata utilizzata principalmente nel contesto dell'inferenza causale per ridurre lo sbilanciamento tra gruppi (ad esempio, un gruppo di trattamento e un gruppo di controllo) in studi osservazionali.

Il suo obiettivo primario è quello di pre-processare i dati in modo da creare sottoclassi di osservazioni che siano "esattamente abbinate" su versioni "grossolane" delle covariate, facilitando così stime di effetti causali meno dipendenti dal modello.

Il principio sottostante al CEM – quello di trovare corrispondenze esatte su covariate "grossolanizzate" – può essere concettualmente adattato nell'ambito dell'integrazione di dataset, come una strategia per identificare potenziali corrispondenze tra record quando non è disponibile una chiave comune esplicita, agendo come una forma di record linkage basato sulla somiglianza delle caratteristiche.

L'idea è di applicare il CEM per l'Identificazione di Corrispondenze tra Dataset Senza Chiavi Esplicite. In uno scenario in cui due dataset non condividono una chiave comune, ma si sospetta che contengano informazioni sulle stesse entità (es. individui di uno stesso paese, oppure pazienti nel caso di applicazioni in ambito sanitario), il CEM può ispirare un approccio per identificare i "gemelli statistici" tra i due dataset, che potrebbero poi essere considerati per la fusione.

Per poter applicare il CEM in quest'ottica, il primo passo è quello di identificare le variabili (covariate) che sono presenti in entrambi i dataset e che si riferiscono alle stesse caratteristiche dell'entità.

Nel nostro contesto, siamo interessanti a fondere individui, in questo caso le covariate comuni saranno "età", "genere", "regione", "Macroregione", "titolo di studio", "quintile di reddito", "tipo di occupazione".

A questo punto, le variabili continue o con molte categorie vengono "grossolanizzate" in intervalli o categorie più ampie e significative dal punto di vista scientifico.

Questo step viene definito grossolanizzazione (Coarsening) delle Covariate.

Ad esempio, l'età esatta (variabile continua) potrebbe essere trasformata in fasce d'età (es. 65-74 e 75+). La grossolanizzazione rende più probabile trovare corrispondenze esatte. Trovare due individui con esattamente la stessa età (es. 67 anni, 3 mesi, 12 giorni) è difficile; trovarli nella stessa fascia d'età (es. 65-74 anni) è più facile.

Una volta che le covariate comuni sono state grossolanizzate, si procede con un matching esatto. Questo significa che si cercano righe in un dataset che abbiano esattamente gli stessi valori grossolanizzati di una riga nell'altro dataset.

Ogni combinazione unica delle covariate grossolanizzate forma una "cella" o "strato". Solo le righe che appartengono alla stessa cella in entrambi i dataset sono considerate potenziali corrispondenze.

Esempio: Se l'individuo A nel Dataset 1 è [Fascia Età: 65-74, Genere: F, Regione: Abruzzo, Occupazione: Casalinga], si cercherà un individuo B nel Dataset 2 con esattamente [Fascia Età: 65-74, Genere: F, Regione: Abruzzo, Occupazione: Casalinga].

Se una riga in un dataset non trova corrispondenza esatta su tutte le covariate grossolanizzate nell'altro dataset, quella riga non sarà considerata per una fusione basata su CEM.

Un problema comune è che una riga in un dataset potrebbe corrispondere a più righe nell'altro dataset (es. più individui nella stessa "cella" grossolanizzata). In questi casi, è fondamentale applicare criteri aggiuntivi.

Infine, se ci sono troppi pochi match, la grossolanizzazione potrebbe essere stata troppo "fine"; se ci sono troppi match multipli, potrebbe essere stata troppo "grossolana". Il processo può essere iterativo per trovare il giusto livello di grossolanizzazione.

Occorre soffermarci sui limiti di tale approccio: a differenza delle operazioni di join che producono direttamente un dataset unito, l'applicazione del CEM serve a identificare coppie di record corrispondenti. La fusione vera e propria (ad esempio, creando un nuovo dataset con le colonne unite per le coppie identificate) è un passaggio successivo che richiede la gestione delle righe non abbinate e dei potenziali match multipli.

In particolar modo, essendo interessati ad abbinare variabili di tipo binomiale (0= non deprived e 1= deprived), l'imputazione di variabili da un dataset all'altro avviene all'interno di specifici "strati" di dati ottenuti tramite con il (CEM).

L'idea è quella di simulare una variabile nel dataset principale basandosi sulla distribuzione di quella stessa variabile nel dataset secondario, in questo caso sulla proporzione di 1 nel dataset secondario. Ovviamente, se all'interno dello strato la proporzione è simile, questo non è automaticamente verificato a livello globale; pertanto dopo ogni fusione, vengono analizzate e confrontate le distribuzioni nella variabile del dataset secondario e poi in quello principale.

Nei paragrafi seguenti verranno analizzati nel dettaglio i passaggi per la determinazione del Dataset Integrato.

Il dataset di partenza sarà EHIS: tale dataset presente informazioni particolarmente dettagliate per il dominio salute, nonché informazioni di dettaglio anche per quanto riguarda l'analisi delle condizioni abitative. A questo dataset, verranno agganciati, nell'ordine, alcune variabili da AVQ, SHARE, EUSILC e SPESE.

5. EHIS + AVQ

Il dataset di EHIS ha 13722 che corrisponde, considerando i pesi campionari, ad una popolazione di 13576875 individui.
Il dataset AVQ invece ha 11289 unità che corrispondono a 13591336 individui.
Per quanto riguarda le variabili da trasferire da AVQ a EHIS sono state scelte:

- Contesto abitativo (a4_contesto_abitaz)
- Mental Health Index-5 (MHI-5) (s4_mhi5)

Le variabili da importare sono state precedentemente codificate in 0/1 (0= non deprived, 1= deprived)
Inoltre, per ogni dataset, è stato preso in considerazione il vettore del peso campionario.

Sono state utilizzate le seguenti variabili per effettuare il matching:

- Genere (1= Maschio, 0=Femmina),
- Titolo di Studio (0=Nessun titolo, 1=Licenza elementare, attestato di valutazione finale; 2=Licenza media, avviamento professionale; 3=Diploma di maturità, Attestato qualifica professionale triennale, Diploma di 2-3 anni; 4=Diploma di istruzione post-secondaria e terziaria),
- Macroarea (1=Nord Ovest; 2=Nord Est, 3=Centro, 4 =Sud e Isole),
- Regione (01=Piemonte, 02=Valle D'Aosta, 03=Lombardia, 04=Trentino Alto Adige, 05=Veneto, 06=Friuli Venezia Giulia, 07=Liguria, 08=Emilia Romagna, 09=Toscana, 10=Umbria, 11=Marche, 12=Lazio, 13=Abruzzo, 14=Molise, 15=Campania, 16=Puglia, 17=Basilicata, 18=Calabria, 19=Sicilia, 20=Sardegna),
- Condizione Lavorativa (1=Svolge un lavoro o professione; 2=Disoccupato, in cerca di prima occupazione; 3 = Ritirato dal lavoro; 4=Studente, in formazione, stagista o tirocinante non retribuito; 5=Casalingo/a; 6=Permanentemente inabile al lavoro o Altra condizione),
- Età (1=65-74, 2=75 e più).

Dove necessario, le variabili sono state ricodificate in modo da avere le stesse classi.
Infine, i due dataset sono stati combinati, eliminando righe vuote, in AVQ.
Per quanto riguarda la numerosità di EHIS si hanno 13720 che corrispondono a 13575593, mentre su AVQ si hanno 11264 che corrispondono a 13561164 individui.

Prima di procedere con il matching è stata verificata la coerenza delle variabili da usare come link nei due dataset.
Di seguito, per ciascuna variabile, riportiamo le proporzioni in ogni classe

Genere:

```
##       AVQ   EHIS
##   0 0.5554 0.5594
##   1 0.4446 0.4406
```

Titolo di Studio

```
##       AVQ   EHIS
##   0 0.0864 0.0732
##   1 0.4026 0.4023
##   2 0.2510 0.2448
##   3 0.1851 0.1993
##   4 0.0749 0.0804
```

Macroarea

```
##       AVQ   EHIS
##   1 0.2292 0.2378
##   2 0.2098 0.2122
##   3 0.1904 0.1987
##   4 0.3706 0.3513
```

Regione

```
##       AVQ   EHIS
##   01 0.0792 0.0746
##   02 0.0219 0.0227
##   03 0.0821 0.0933
##   04 0.0597 0.0517
##   05 0.0580 0.0584
##   06 0.0401 0.0412
##   07 0.0460 0.0473
##   08 0.0520 0.0609
##   09 0.0635 0.0716
##   10 0.0293 0.0249
##   11 0.0447 0.0405
##   12 0.0529 0.0617
##   13 0.0413 0.0371
##   14 0.0301 0.0231
##   15 0.0653 0.0549
##   16 0.0579 0.0553
##   17 0.0324 0.0299
##   18 0.0423 0.0466
##   19 0.0574 0.0630
##   20 0.0439 0.0412
```

Condizione lavorativa

```
##      AVQ   EHIS
## 1 0.0448 0.0441
## 2 0.0103 0.0032
## 3 0.7049 0.7159
## 5 0.2058 0.1970
## 6 0.0342 0.0398
```

Età

```
##      AVQ   EHIS
## 1 0.4985 0.4915
## 2 0.5015 0.5085
```

Le proporzioni sono abbastanza simili, tali da garantire di poter procedere con l'analisi
Analizzando le classi di ciascuna variabile del matching sono stati creati 7680 combinazioni (la traduzione italiana del termine tecnico è "bidoni") teoriche.
Dopo aver confrontato i risultati di diversi metodi di matching, si è ritenuto come metodo più opportuno quello con il metodo CEM.

Commento ai risultati del matching

L'obiettivo del matching è rendere i gruppi di trattamento e controllo il più simili possibile sulle caratteristiche individuate. L'output completo è riportato in APPENDICE 3.
Per quanto riguarda il bilanciamento delle covariate prima dell'applicazione del matching, quasi tutti i valori della Std. Mean Diff. (Standardized Mean Difference) sono già molto vicini allo zero, il che suggerisce un buon bilanciamento iniziale, o che le differenze tra i gruppi non sono molto ampie su queste covariate prima del matching.
Ad esempio, per Gender0 è 0.0079 e per Gender1 è −0.0079, entrambi molto piccoli. Lo stesso vale per Education, MacroArea, Regione, WorkStatus e Age.
Per quanto riguarda i risultati del bilanciamento dopo che matchit, per tutte le covariate, nelle colonne Std. Mean Diff. (Standardized Mean Difference), eCDF Mean ((Empirical Cumulative Distribution Function Mean), eCDF Max (Empirical Cumulative Distribution Function Max) e Std. Pair Dist., i valori sono esattamente 0.
Il metodo CEM, per sua natura, crea strati (o "celle") in cui tutte le unità (sia trattate che di controllo) hanno valori identici per le covariate utilizzate nel matching. Questo significa che all'interno di ogni strato, il bilanciamento è esatto. Quando i risultati di questi strati vengono aggregati, le differenze standardizzate tra le medie e le differenze nelle distribuzioni cumulative diventano nulle.
Un valore 0 per Std. Mean Diff. indica che le medie di ogni covariata sono perfettamente bilanciate tra il gruppo trattato e il gruppo di controllo nel dataset accoppiato.
Essendo le variabili categoriche sia nel pre-matching che nel post-matching non è applicabile calcolare Var. Ratio, per cui nell'output è visibile un valore "."

Infine, eCDF Mean = 0 e eCDF Max = 0 confermano che le distribuzioni cumulative di ogni covariata sono identiche tra i gruppi accoppiati.

Questo è un indicatore molto forte di un bilanciamento eccellente, anche oltre la semplice uguaglianza delle medie. L'ultima colonna, Std. Pair Dist. = 0 nel caso del CEM, conferma l'assenza di distanza standardizzata tra le coppie o all'interno degli strati.

La sezione "sample size" fornisce informazioni su quante osservazioni sono state utilizzate (o scartate) nel processo di matching.

Il dataset contiene (senza tener conto dei pesi) Control: 11264 e Treated: 13720.

L'ESS (Effective Sample Size) è una stima della dimensione effettiva del campione dopo il matching e la ponderazione, i risultati mostrano i suguenti valori: Control: 9020.39 e Treated: 13433. Osserviamo che un ESS elevato è desiderabile, in quanto indica che si ha ancora un buon potere statistico nel campione bilanciato.

Per il gruppo trattato, l'ESS è molto vicino al numero di osservazioni matched (13433 vs 13433), suggerendo che i pesi non hanno ridotto molto il potere statistico in quel gruppo.

Per il controllo (9020.39 vs 11056), c'è una leggera riduzione, ma è comunque un buon numero.

Per quanto riguarda le osservazioni realmente abbinate, abbiamo Control: 11056 e Treated: 13433. Questo indica il numero di osservazioni (con i loro pesi) che sono state effettivamente utilizzate e bilanciate dal CEM.

Notiamo che un'alta percentuale delle osservazioni originali sono state mantenute nel processo di matching. Per il gruppo di controllo, da 11264 a 11056 sono state matched, mentre per il gruppo trattato, da 13720 a 13433 sono state matched.

Le osservazioni che non hanno trovato una controparte nello strato corrispondente e sono state escluse dal campione bilanciato, sono Control: 208 e Treated: 287.

Infine, osserviamo che nessuna osservazione è stata scartata perché non rientrava in alcuno strato o non aveva covariate valide. Per il CEM, "scartato" significa tipicamente che non esiste uno strato compatibile per un'osservazione (Control: 0 e Treated: 0).

L'analisi congiunta di tutti questi valori suggerisce un ottimo risultato per il bilanciamento ottenuto tramite CEM.

Il matching con metodo CEN, per le 7680 combinazioni teoriche individuati, ne ha selezionate 816 non vuote.

Passo successivo del matching è quello di abbinare, all'interno di ciascun "bidone", le osservazioni delle variabili da esportare da AVQ a EHIS.

Come illustrato nei paragrafi precedenti, poiché le variabili da importare sono di tipo binomiale (0/1), si è deciso di replicare la distribuzione osservata nel "bidone" alle osservazioni dell'altro dataset presenti nello stesso "bidone".

Per valutare l'effettiva bontà del metodo appena illustrato, sono state confrontate, per ciascuna delle variabili importate, le distribuzioni nel dataset integrato con la stessa variabile nel dataset originario.

Di seguito, per il dataset AVQ, riportiamo le differenze tra la variabile originaria e quella nel dataset integrato.

- **Variabile Contesto abitativo (a4_contesto_abitaz)**

```
--- Risultati DatasetIntegrato ---
##   Categoria Frequenza_Ponderata Proporzione Percentuale
##         0          10318930   0.8646547    86.46547
##         1           1615233   0.1353453    13.53453

--- Risultati AVQ ---

##   Categoria Frequenza_Ponderata Proporzione Percentuale
##         0          11605561   0.8538941    85.38941
##         1           1985775   0.1461059    14.61059
```

Questi risultati evidenziano una differenza di 1.08 punti percentuali per quanto riguarda la classe 1

È stato effettuato anche un test del Chi-quadro di Pearson con aggiustamento di Rao & Scott[1] per valutare se esiste o meno una differenza delle proporzioni nei due gruppi.
L'aggiustamento è necessario per tener conto della presenza di pesi campionari. Questo test valuta se la distribuzione delle proporzioni (quanti 0 e quanti 1) della tua variabile binaria è significativamente diversa tra i due dataset, tenendo conto dei pesi.
Formalmente, l'ipotesi nulla è che Non c'è una differenza significativa nelle proporzioni della variabile binaria (v4_a_AVQC) tra il DatasetIntegrato e il dataset AVQ.
In altre parole, l'ipotesi nulla sostiene che la variabile binaria (v4_a_AVQC) e la variabile che indica il dataset di origine (Source) sono indipendenti.
Questo significa che la proporzione di "0" e "1" è la stessa in entrambe le sorgenti di dati (o le eventuali differenze osservate sono dovute solo al caso o alla variazione campionaria).
L'ipotesi alternativa invece, afferma che c'è una differenza significativa nelle proporzioni della variabile binaria (v4_a_AVQC) tra il DatasetIntegrato e il dataset AVQ.
Ovvero, la variabile binaria e la sorgente del dataset non sono indipendenti, il che implica che la distribuzione di 0 e 1 varia tra i due dataset in modo che non può essere attribuito al solo caso.
Se il p-value è inferiore al livello di significatività (es. 0.05), si può concludere che esiste una differenza statisticamente significativa nella distribuzione (proporzioni) della variabile binaria tra i due dataset, tenendo conto dei pesi. Se il p-value è maggiore, non c'è evidenza sufficiente per concludere che le proporzioni siano diverse.
In questo specifico caso, l'output del test del Chi quadro suggerisce che vi sia una differenza tra i gruppi.

[1] Un test analogo è stato effettuato per tutte le variabili nei dataset analizzati. Nei test seguenti non riporteremo, per brevità, la formulazione estesa delle ipotesi nulla e alternativa.

```
## Test Chi-quadro Ponderato (Rao & Scott adjustment):
## Statistica Chi-quadro = 5.56, df = 1
## P-value = 4.46e-02
## Conclusione (alpha = 0.05): Rifiutiamo l'ipotesi nulla.
## Esiste una differenza statisticamente significativa nelle proporzioni tra i due dataset.
```

Sottolineiamo che il valore è abbastanza vicino al valore critico di 0.05 e la differenza nelle percentuali veramente piccola per cui possiamo considerare che nel nuovo dataset, la variabile importata rifletta in maniera corretta la distribuzione nel dataset originale, come evidenziato nella Figura 1.

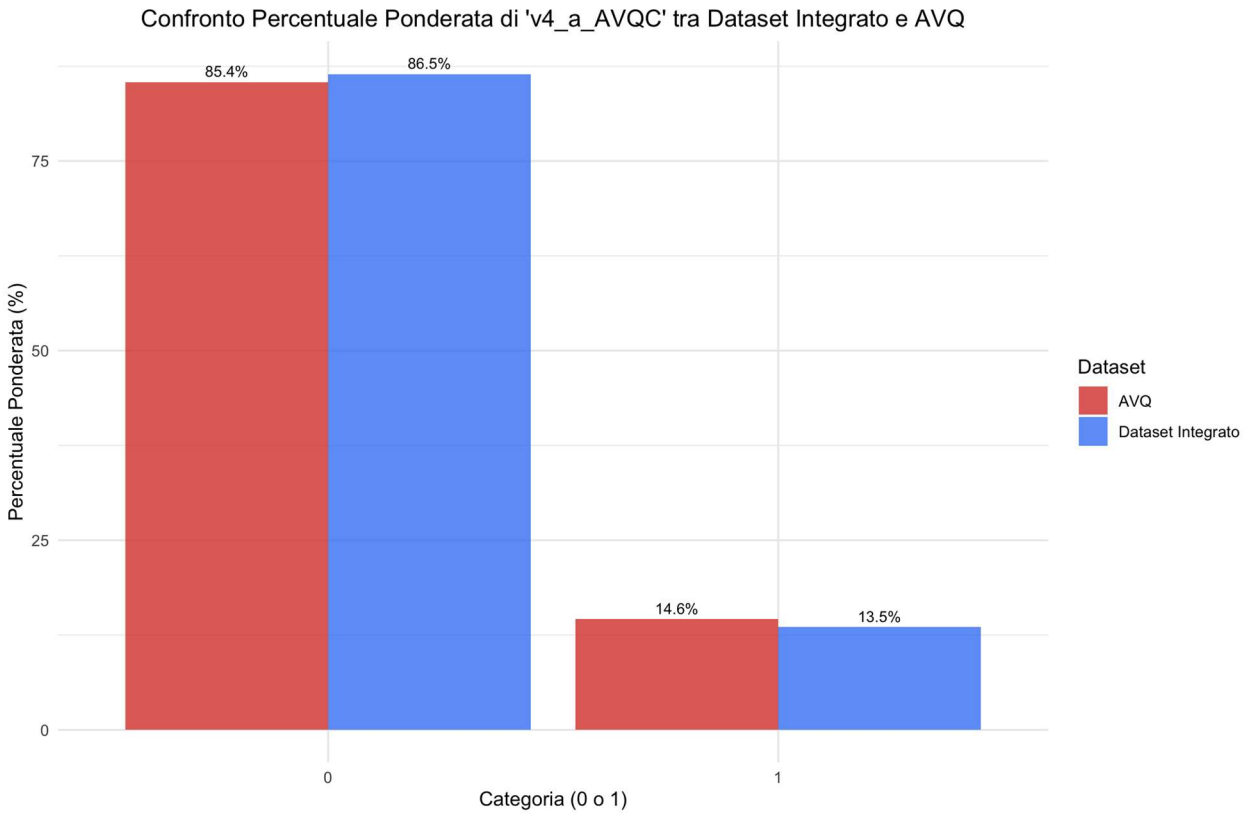

Figura 1: Distribuzione pre e post matching

- **Mental Health Index-5 (MHI-5) (s4_mhi5)**

```
--- Risultati DatasetIntegrato ---
##   Categoria Frequenza_Ponderata Proporzione Percentuale
##         0          8591789   0.7199324    71.99324
##         1          3342373   0.2800676    28.00676

--- Risultati AVQ ---
##   Categoria Frequenza_Ponderata Proporzione Percentuale
##         0          9758006   0.7179578    71.79578
##         1          3833330   0.2820422    28.20422
```

Questi risultati evidenziano una differenza di 0.2 punti percentuali per quanto riguarda la classe 1 Anche, l'output del test del Chi quadro suggerisce che non vi sia una differenza tra i gruppi.

```
## Test Chi-quadro Ponderato (Rao & Scott adjustment):
## Statistica Chi-quadro = 0.11, df = 1
## P-value = 7.69e-01
## Conclusione (alpha = 0.05): Non rifiutiamo l'ipotesi nulla.
## Non esiste una differenza statisticamente significativa nelle proporzioni tra i due dataset.
```

Questo viene anche confermato nella Figura 2

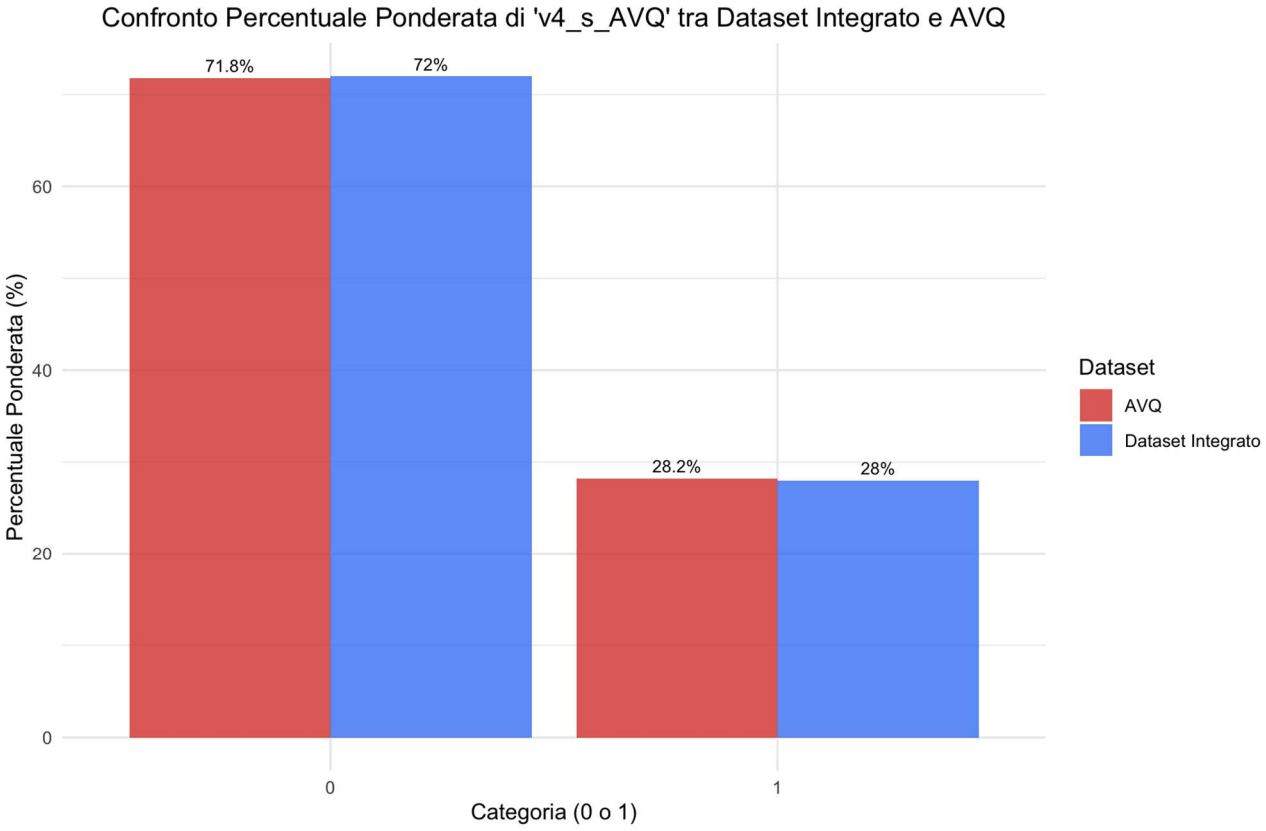

Figura 2: Distribuzione pre e post matching

A seguito del Matching, il dataset di EHIS ha, pertanto, perso 287 osservazioni.

La nuova numerosità del dataset è, dunque, di 13433 che corrisponde ad una numerosità campionaria pari a 13298337.

Di conseguenza, considerando i pesi campionari, si è registrata una perdita percentuale di 2.05. Questo vuol dire che, con il matching, solo il 2% dei dati di EHIS non ha trovato un corrispondente nel dataset di AVQ e, non essendo stati abbinati, sono usciti dal campione.

Questo nuovo dataset integrato EHIS+AVQ rappresenta il punto di partenza per agganciare un nuovo dataset: SHARE.

6. EHIS_AVQ+SHARE

Il dataset EHIS_AVQ ha 13433 righe che corrisponde ad una numerosità campionaria pari a 13298337.
Il dataset SHARE invece conta solo 1554, ma considerando i pesi campionari otteniamo una numerosità confrontabile, ovvero 13126773
Per quanto riguarda le variabili da trasferire da SHARE a EHIS_AVQ sono state scelte:

- social_connectedness (sn_scale_dic):

Le variabili da importare sono state precedentemente codificate in 0/1 (0= non deprived, 1= deprived)
Inoltre, per ogni dataset, è stato preso in considerazione il vettore del peso campionario.
Sono state utilizzate le seguenti variabili per effettuare il matching:

- Genere (1= Maschio, 0=Femmina),
- Titolo di Studio (0=Nessun titolo, 1=Licenza elementare, attestato di valutazione finale; 2=Licenza media, avviamento professionale; 3=Diploma di maturità, Attestato qualifica professionale triennale, Diploma di 2-3 anni; 4=Diploma di istruzione post-secondaria e terziaria),
- Macroarea (1=Nord Ovest; 2=Nord Est, 3=Centro, 4 =Sud e Isole),
- Condizione lavorativa (1=Svolge un lavoro o professione; 2=Disoccupato, in cerca di prima occupazione; 3 = Ritirato dal lavoro; 4=Studente, in formazione, stagista o tirocinante non retribuito; 5=Casalingo/a; 6=Permanentemente inabile al lavoro o Altra condizione),
- Età (1=65-74, 2=75 e più),
- Reddito (Espresso in quintili, 1= I quintile rappresenta i redditi più bassi, … , 5= V quintile, redditi più alti).

Dove necessario, le variabili sono state ricodificate in modo da avere le stesse classi.
Infine, i due dataset sono stati combinati, eliminando righe vuote, in EHIS.
Per quanto riguarda la numerosità di EHIS_AVQ si hanno 134330 che corrispondono a 13298337 mentre su SHARE si hanno 1544 che corrispondono a 13028283 individui

Prima di procedere con il matching è stata verificata la coerenza delle variabili da usare come link nei due dataset.
Di seguito, per ciascuna variabile, riportiamo le proporzioni in ogni classe

Genere

```
##     SHARE   EHIS
##   0 0.5418 0.5603
##   1 0.4582 0.4397
```

Titolo di Studio

```
##   SHARE  EHIS
## 0 0.0611 0.0706
## 1 0.4562 0.4054
## 2 0.2252 0.2455
## 3 0.1660 0.1991
## 4 0.0914 0.0794
```

Macroarea

```
##   SHARE  EHIS
## 1 0.1648 0.2375
## 2 0.1105 0.2119
## 3 0.2431 0.1989
## 4 0.4816 0.3517
```

Condizione lavorativa

```
##   SHARE  EHIS
## 1 0.0284 0.0406
## 2 0.0052 0.0020
## 3 0.7260 0.7300
## 5 0.1883 0.1995
## 6 0.0522 0.0279
```

Età

```
##   SHARE  EHIS
## 1 0.4653 0.4900
## 2 0.5347 0.5100
```

Reddito

```
##   SHARE  EHIS
## 1 0.2516 0.1257
## 2 0.2072 0.2108
## 3 0.2027 0.2167
## 4 0.1956 0.2224
## 5 0.1429 0.2244
```

Le proporzioni sono abbastanza simili, tali da garantire di poter procedere con l'analisi

Analizzando le classi di ciascuna variabile del matching sono stati creati 2000 combinazioni (la traduzione italiana del termine tecnico è "bidoni") teoriche.

Commento ai risultati del matching – round 1

L'output completo è riportato in APPENDICE 4.

Le proporzioni (Means Treated e Means Control) di ciascuna categoria di covariata rispettivamente nel gruppo trattato e nel gruppo di controllo prima del matching sono paragonabili.

Ad esempio, per Gender0 (presumibilmente una categoria di genere, es. "Uomo"), il 56.03% dei trattati sono di questo genere, contro il 54.21% dei controlli.

Per quanto riguarda la Std. Mean Diff. (Standardized Mean Difference): notiamo che, prima del matching, ci sono alcune differenze notevoli. Ad esempio, per MacroArea1 abbiamo 0.1715, per MacroArea2 0.2474, e per MacroArea4 -0.2711.

Ancora più evidente è Income1 con -0.3732, Income5 con 0.1932.

Generalmente, valori vicini allo zero (spesso un valore assoluto inferiore a 0.1 o 0.2, a seconda del campo) indicano un buon bilanciamento. Pertanto, in questo caso abbiamo uno sbilanciamento.

Infine, anche per eCDF Mean (Empirical Cumulative Distribution Function Mean) e eCDF Max (Empirical Cumulative Distribution Function Maximum), metriche che misurano la differenza media e massima tra le funzioni di distribuzione cumulativa empiriche delle covariate tra i gruppi trattato e di controllo, osserviamo valori relativamente alti per le covariate menzionate sopra (es. MacroArea4 a 0.1295, Income1 a 0.1237), confermando il sbilanciamento iniziale.

Per quanto riguarda l'effetto del CEM sul bilanciamento delle covariate dopo il matching, per tutte le covariate categoriche (Gender, Education, MacroArea, Income, WorkStatus, Age), le proporzioni nei gruppi trattato e di controllo sono ora identiche (o quasi, con differenze minime come -0 che indica un valore molto vicino a zero).

Poiché il CEM crea strati in cui le covariate ingrossate sono esattamente le stesse tra i gruppi, la differenza standardizzata delle medie (e di fatto tutte le statistiche di bilanciamento per le covariate categoriale/binarie) diventa zero. Questo indica un bilanciamento perfetto per le covariate specificate, all'interno degli strati definiti dal CEM.

Pertanto, il CEM ha avuto successo nell'ottenere un bilanciamento quasi perfetto per le covariate specificate. Le differenze tra il gruppo trattato e il gruppo di controllo su queste variabili sono state eliminate nel dataset abbinato.

Per quanto riguarda la dimensione dei campioni, si hanno 1544 individui nel gruppo di controllo e 13433 nel gruppo trattato. Per il gruppo di controllo, l'ESS che è una misura della dimensione campionaria effettiva, tenendo conto dei pesi, è 904.55. L'ESS. Per il gruppo trattato, l'ESS è 11039, identico al numero di individui matched, suggerendo che i pesi per il gruppo trattato sono probabilmente uniformi o vicini all'uniformità dopo il matching.

L'output del modello indica che 1502 individui nel gruppo di controllo sono stati abbinati.

Questo significa che, dei 1544 originali, 42 individui di controllo (Unmatched) non hanno trovato una corrispondenza perfetta negli strati creati. 11039 individui nel gruppo trattato sono stati abbinati. Dei 13433 originali, 2394 individui trattati (Unmatched) non hanno trovato una corrispondenza.

Il CEM tende a scartare gli individui che non trovano una corrispondenza esatta negli strati, il che può portare a una perdita di osservazioni, soprattutto se le covariate sono molto diverse tra i gruppi o se le categorie di coarsening sono troppo fini.

Pertanto, il numero di individui che non hanno trovato una corrispondenza negli strati CEM è stato escluso dall'analisi post-matching.

Osserviamo che, in questa fase, 42 controlli e 2394 trattati sono stati esclusi. La perdita di individui trattati è notevole (circa il 17.8%), il che suggerisce che una parte del gruppo trattato non ha caratteristiche comuni con i controlli rimanenti negli strati CEM.

Se da un lato il CEM ha avuto grande successo nel bilanciare le covariate specificate tra i gruppi trattato e di controllo, dall'altro è presente una perdita di osservazioni tale da optare per un raffinamento del matching.

Commento ai risultati del matching – round 2

Come osservato precedentemente, il matching ha determinato 2394 individui di EHIS_AVQ che non hanno trovato un corrispondente nel dataset SHARE. Questo produrrebbe una perdita di individui notevole, circa il 17.8% della popolazione, suggerendo un abbinamento non ottimale.

Visti i risultati del primo matching, è stato necessario, quindi, effettuare un secondo round, prendendo, per quanto riguarda EHIS_AVQ il solo sottogruppo dei non abbinati insieme a tutte le unità di SHARE e utilizzando covariate meno restrittive.

In particolare:

- la variabile Istruzione, da 5 classi è stata trasformata in 3 classi: basso, medio e alto.
- Livello di reddito: dai 5 quintili si è passati a 3 livelli: basso (q1 e q2), medio (q3) e alto (q4 e q5)
- Condizione lavorativo: da 6 classi si è passati a 3: Occupati, Pensionati e Altro

Pertanto, la combinazione delle covariate ha individuato 432 "bidoni" teorici e il processo di matching ne ha individuate 136 non vuote.

Guardando il riepilogo del bilanciamento prima del matching (sezione "Summary of Balance for All Data"), emerge un quadro abbastanza chiaro: i gruppi di trattamento e di controllo erano abbastanza diversi nelle loro caratteristiche iniziali.

Le differenze tra le medie standardizzate erano notevoli per diverse variabili, si veda per esempio la variabile genere o le variabili relative all'area geografica (MacroArea). Anche la fascia di reddito (Income_new) e lo stato lavorativo (WorkStatus_new) presentavano sbilanciamenti significativi, con alcune categorie che mostravano differenze standardizzate ben oltre la soglia accettabile (spesso posta a 0.1 o 0.05).

Dopo il CEM, le misure della differenza tra le funzioni di distribuzione cumulativa empirica (eCDF Mean ed eCDF Max) sono scese a zero, confermando che le distribuzioni delle covariate sono ora perfettamente allineate tra i due gruppi abbinati.

Per quanto concerne la dimensione campionaria, originariamente il dataset presentava 1544 controlli e 2394 trattati (senza considerare i pesi campionari).

Dopo aver eseguito il CEM, si hanno 631 controlli e 1673 trattati dopo il matching. Questa riduzione indica che un numero significativo di osservazioni, in particolare nel gruppo di controllo (circa il 59% dei controlli originali) e una parte notevole nel gruppo trattato (circa il 30% dei trattati originali), non ha trovato una corrispondenza all'interno degli strati definiti dal CEM ed è stato quindi scartato.

Osserviamo inoltre che, sebbene l'output generale di questo secondo step mostri un bilanciamento perfetto (tutti gli Std. Mean Diff. a zero), per alcune categorie delle covariate, come Education_new_Low, MacroArea_2, MacroArea_3, Income_new_High, e Income_new_Low, si osservano ancora piccole differenze standardizzate residue (es. 0.1873 o 0.2002).
Nonostante queste piccole differenze residue, è fondamentale notare che sono comunque enormemente inferiori rispetto agli sbilanciamenti iniziali del dataset originale.
La maggior parte di queste differenze residue è al di sotto o molto vicina alla soglia comunemente accettata di 0.1, indicando comunque un bilanciamento complessivamente molto buono.

Passo successivo del matching è quello di abbinare, all'interno di ciascun "bidone", le osservazioni delle variabili da esportare da SHARE a EHIS_AVQ.
Come illustrato nei paragrafi precedenti, poiché le variabili da importare sono di tipo binomiale (0/1), si è deciso di replicare la distribuzione osservata nel "bidone" alle osservazioni dell'altro dataset presenti nello stesso "bidone".
Questa attribuzione è stata effettuata dopo ciascun step e, alla fine i due dataset sono stati ricombinati per ottenere il file complessivo EHIS_AVQ_SHARE.
Questo dataset consiste in 12712 righe che corrispondono a 12439808 osservazioni. Pertanto, rispetto al dataset iniziale EHIS_AVQ, il dataset integrato ha perso poco più del 6% (6.455909) delle osservazioni iniziali.
Per valutare l'effettiva bontà del metodo appena illustrato, sono state confrontate, per ciascuna delle variabili importate, le distribuzioni nel dataset integrato con la stessa variabile nel dataset originario.
Di seguito, per il dataset SHARE, riportiamo le differenze tra la variabile originaria e quella nel dataset integrato.

- **Variabile social_connectedness (sn_scale_dic):**

```
## --- Risultati Dataset Integrato ---
##   Categoria Frequenza_Ponderata Proporzione Percentuale
##         0          7675518  0.6431552    64.31552
##         1          4258644  0.3568448    35.68448

## --- Risultati AVQ ---
##   Categoria Frequenza_Ponderata Proporzione Percentuale
##         0          8251387  0.6285922    62.85922
##         1          4875387  0.3714078    37.14078
```

Questi risultati evidenziano una differenza di 1.46 punti percentuali per quanto riguarda la classe 1.
È stato effettuato anche un test del Chi-quadro di Pearson come illustrato precedentemente.
In questo specifico caso, l'output del test del Chi quadro suggerisce che non vi sia una differenza tra i gruppi

```
## Test Chi-quadro Ponderato (Rao & Scott adjustment):
## Statistica Chi-quadro = 3.12, df = 1
## P-value = 3.37e-01
## Conclusione (alpha = 0.05): Non rifiutiamo l'ipotesi nulla.
## Non esiste una differenza statisticamente significativa nelle proporzioni tra i due dataset.
```

La figura 3 riporta la distribuzione della variabile importata sia nel dataset originario che in quello integrato.

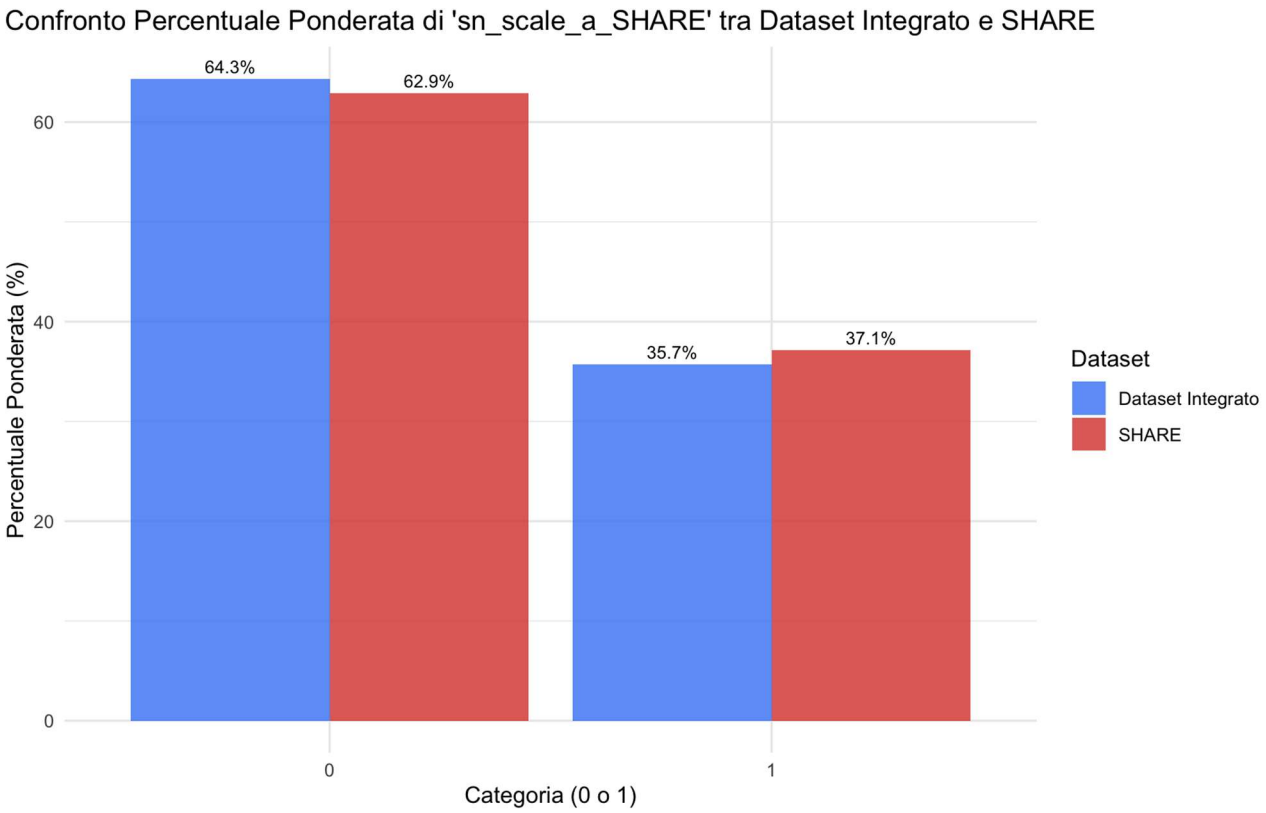

Figura 3: Distribuzione pre e post matching

7. EHIS_AVQ_SHARE+EUSILC

Come indicato sopra, il dataset EHIS_AVQ_SHARE contiene 12712 righe che corrispondono a 12439808 osservazioni.
Il dataset EUSILC, invece, contiene 13585 che corrispondono, applicando i pesi campionari a 14453754 unità.
Per quanto riguarda le variabili da trasferire da EUSILC al dataset integrati sono state scelte:

- Difficoltà a sostenere i costi dell'abitazione (a3_cost_prob_EUSILC),
- Condizione di sotto-occupazione dell'abitazione (a4_under_occ2_EUSILC),
- Rinuncia alle cure mediche specialistiche (s3_no_med_treat2_EUSIL),
- Grado di urbanizzazione (UrbDeg).

Le variabili da importare sono state precedentemente codificate in 0/1 (0= non deprived, 1= deprived). Inoltre, per ogni dataset, è stato preso in considerazione il vettore del peso campionario.

Per effettuare il matching sono state utilizzate le seguenti variabili

- Genere (1= Maschio, 0=Femmina),
- Titolo di Studio (0=Nessun titolo, 1=Licenza elementare, attestato di valutazione finale; 2=Licenza media, avviamento professionale; 3=Diploma di maturità, Attestato qualifica professionale triennale, Diploma di 2-3 anni; 4=Diploma di istruzione post-secondaria e terziaria),
- Macroarea (1=Nord Ovest; 2=Nord Est, 3=Centro, 4 =Sud e Isole),
- Condizione lavorativa (Occupato, Pensionato, Disoccupato e Altro),
- Età (1=65-74, 2=75 e più),
- Reddito (Espresso in quintili, 1= I quintile rappresenta i redditi più bassi, … , 5= V quintile, redditi più alti).

Dove necessario, le variabili sono state ricodificate in modo da avere le stesse classi.
Ne è un esempio la variabile Condizione lavorativa che, dalle 6 categorie iniziali del Dataset Integrato, per essere compatibile con il dataset di EUSILC, è stata ricodificata in 4 classi.
Infine, i due dataset sono stati combinati, eliminando righe vuote, in EUSILC.
Pertanto, il dataset pronto per il matching conta, per quanto riguarda il dataset integrato di 12712 righe che corrispondono a 12439808 individui, mentre in EUSILC si hanno 12712 unità che corrispondono a 14449771 individui.
Prima di procedere con il matching è stata verificata la coerenza delle variabili da usare come link nei due dataset.
Di seguito, per ciascuna variabile, riportiamo le proporzioni in ogni classe

Genere:

```
##    EUSILC   EHIS
## 0 0.5650 0.5553
## 1 0.4350 0.4447
```

Titolo di Studio

```
##    EUSILC   EHIS
## 0 0.0648 0.0730
## 1 0.3612 0.4073
## 2 0.2524 0.2477
## 3 0.2379 0.1993
## 4 0.0837 0.0727
```

Macroarea

```
##    EUSILC   EHIS
## 1 0.2526 0.2259
## 2 0.2297 0.2018
## 3 0.2526 0.2047
## 4 0.2651 0.3676
```

Condizione lavorativa

```
##    EUSILC   EHIS
## 1 0.0485 0.0275
## 2 0.0044 0.0017
## 3 0.7917 0.7601
## 4 0.1554 0.2107
```

Età

```
##   group
##    EUSILC   EHIS
## 1 0.4706 0.4844
## 2 0.5294 0.5156
```

Classe di reddito

```
##    EUSILC   EHIS
## 1 0.2000 0.1295
## 2 0.2000 0.2150
## 3 0.2000 0.2056
## 4 0.2000 0.2243
## 5 0.2000 0.2256
```

Le proporzioni sono abbastanza simili, tali da garantire di poter procedere con l'analisi
Analizzando le classi di ciascuna variabile del matching sono stati creati 1600 "bidoni" teorici.

Commento ai risultati del matching

Prima dell'applicazione del matching, la "Summary of Balance for All Data" ha rivelato differenze significative tra i gruppi, specialmente per quanto riguarda la variabile income (-0.2101) valore che superando le soglie di accettabilità tipiche di 0.1 o 0.25.
Questo potrebbe essere giustificato dal fatto che tale variabile risente di come è stata costruita: nel caso di EUSILC è disponibile la variabile "reddito" e i quintili sono stati costruiti sulla distribuzione ridotta (over 65) e non su tutta la popolazione.
La sezione "Summary of Balance for Matched Data" ha dimostrato l'efficacia del CEM.
Tutte le covariate hanno mostrato Std. Mean Diff. pari a 0 (o quasi zero), indicando un bilanciamento quasi perfetto tra i gruppi trattato e di controllo. Similmente, i valori di eCDF Mean ed eCDF Max erano prossimi a 0, confermando che le distribuzioni delle covariate sono diventate praticamente identiche.

Controllando la dimensione otteniamo i seguenti dati sulle unità utilizzate:

- Originali: 13.579 controlli e 12.712 trattati.
- Appaiate: 12.927 controlli e 12.644 trattati.
- Non appaiate: 652 controlli e 68 trattati.
- Scartate: 0 in entrambi i gruppi.

Complessivamente, 720 unità (652 controlli e 68 trattati) sono state perse nel processo di matching, poiché non hanno trovato una corrispondenza esatta.
Nonostante questa perdita, la maggior parte delle osservazioni è stata utilizzata, mantenendo un'ampia dimensione campionaria effettiva (ESS).
Dopo il matching, le Differenze Medie Standardizzate (Diff.Un) sono risultate molto piccole.
Ad esempio, Income_1 ha mostrato una Diff.Un di -0.0704. Questo valore è ben al di sotto della soglia comune di 0.1, confermano che il CEM ha creato gruppi altamente comparabili.
L'analisi congiunta di tutti questi valori suggerisce, nonostante una moderata perdita di osservazioni non appaiate, un bilanciamento molto elevato, con differenze medie standardizzate trascurabili per tutte le covariate.
Osserviamo che dopo la fase di abbinamento con il CEM, la percentuale di popolazione persa è stata dello 0.63%. Infatti, prima del matching il dataset integrato contava 12439808 individui e dopo il matching, in nuovo dataset ne contiene 12361236.

Il matching con metodo CEM, per le 1600 combinazioni teoriche individuati, ne ha selezionate 579 non vuote.
Passo successivo del matching è quello di abbinare, all'interno di ciascun "bidone", le osservazioni delle variabili da esportare da EUSILC al dataset integrato.

Come illustrato nei paragrafi precedenti, poiché le variabili da importare sono di tipo binomiale (0/1), si è deciso di replicare la distribuzione osservata nel "bidone" alle osservazioni dell'altro dataset presenti nello stesso bidone.

Per valutare l'effettiva bontà del metodo appena illustrato, sono state confrontate, per ciascuna delle variabili importate, le distribuzioni nel dataset integrato con la stessa variabile nel dataset originario.

Di seguito, per il dataset EUSILC, riportiamo le differenze tra la variabile originaria e quella nel dataset integrato.

- **Difficoltà a sostenere i costi dell'abitazione (a3_cost_prob_EUSILC)**

```
--- Risultati DatasetIntegrato ---
##   Categoria Frequenza_Ponderata Proporzione Percentuale
##         0        11313282.7  0.94797461   94.797461
##         1         620879.5  0.05202539    5.202539

--- Risultati EHIS ---"

##   Categoria Frequenza_Ponderata Proporzione Percentuale
##         0         13265126  0.91776337   91.776337
##         1          1188628  0.08223663    8.223663
```

Questi risultati evidenziano una differenza di 3.02 punti percentuali per quanto riguarda la classe 1. È stato effettuato anche un test del Chi-quadro di Pearson con aggiustamento di Rao & Scott come illustrato precedentemente.
In questo specifico caso, l'output del test del Chi quadro suggerisce che vi sia una differenza tra i gruppi.

```
## Test Chi-quadro Ponderato (Rao & Scott adjustment):
## Statistica Chi-quadro = 90.89, df = 1
## P-value = 3.27e-15
## Conclusione (alpha = 0.05): Rifiutiamo l'ipotesi nulla.
## Esiste una differenza statisticamente significativa nelle proporzioni tra i due dataset.
```

Anche se il test rifiuta l'ipotesi di uguaglianza tra i due gruppi, il grafico in Figura 4 mostra che c'è una discreta similitudine tra i due dataset.

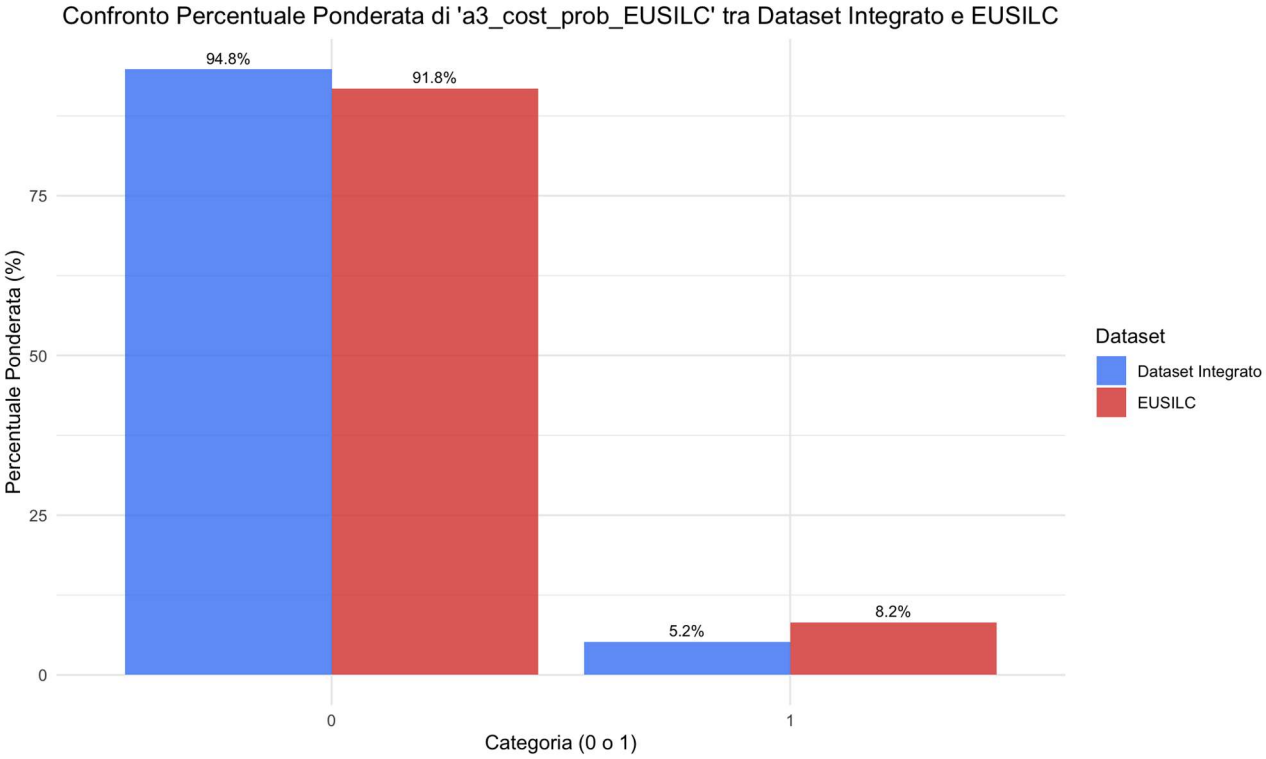

Figura 4: Distribuzione pre e post matching

- **Condizione di sotto-occupazione dell'abitazione (a4_under_occ2_EUSILC)**

```
## --- Risultati Dataset Integrato ---"
##   Categoria Frequenza_Ponderata Proporzione Percentuale
##         0          8926217   0.7479551    74.79551
##         1          3007945   0.2520449    25.20449

## --- Risultati EUSILC ---"
##   Categoria Frequenza_Ponderata Proporzione Percentuale
##         0         10782734   0.7460161    74.60161
##         1          3671020   0.2539839    25.39839
```

Questi risultati evidenziano una differenza di 0.19 punti percentuali per quanto riguarda la classe 1. È stato effettuato anche un test del Chi-quadro di Pearson con aggiustamento di Rao & Scott per valutare se esiste o meno una differenza delle proporzioni nei due gruppi, come illustrato precedentemente.

```
## Test Chi-quadro Ponderato (Rao & Scott adjustment):
## Statistica Chi-quadro = 0.13, df = 1
## P-value = 7.63e-01
## Conclusione (alpha = 0.05): Non rifiutiamo l'ipotesi nulla.
## Non esiste una differenza statisticamente significativa nelle proporzioni tra i due dataset.
```

In questo specifico caso, l'output del test del Chi quadro suggerisce che non vi sia una differenza tra i gruppi. Completiamo le analisi mostrando la distribuzione (Figura 5).

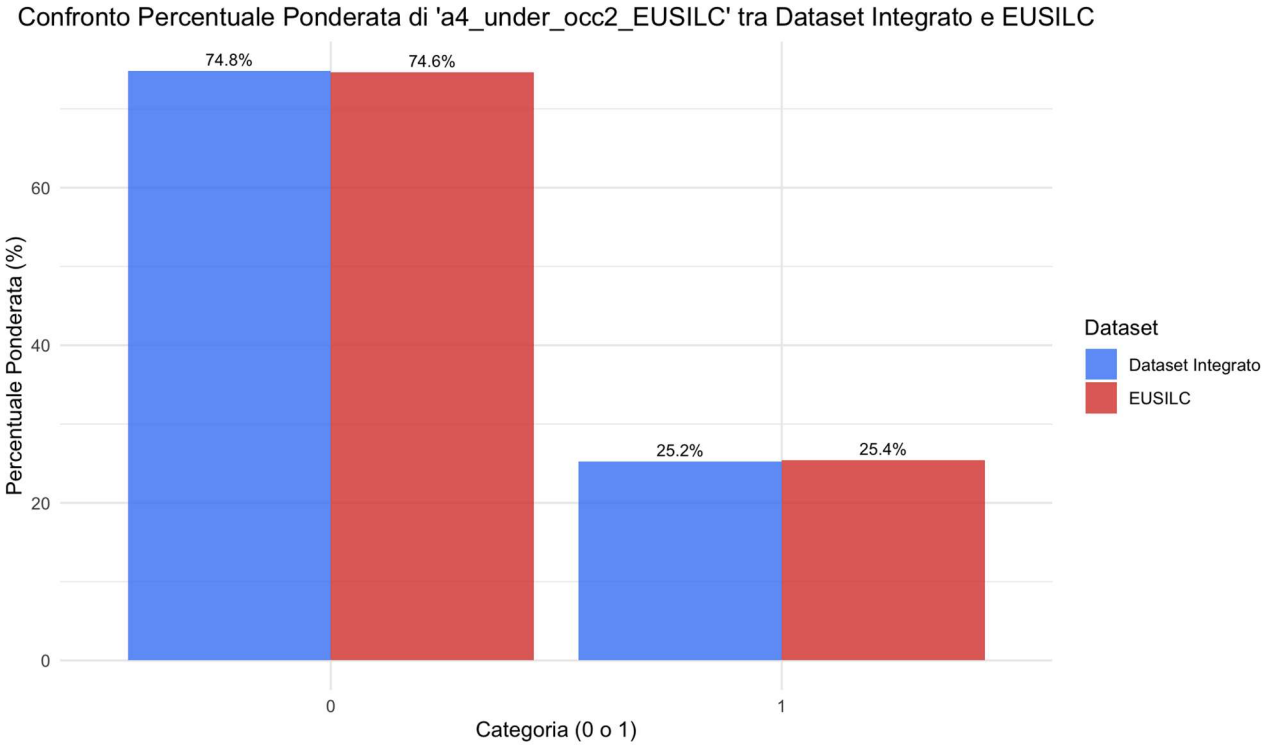

Figura 5: Distribuzione pre e post matching

- **Rinuncia alle cure mediche specialistiche (s3_no_med_treat2_EUSIL)**

```
--- Risultati Dataset Integrato ---
##   Categoria Frequenza_Ponderata Proporzione Percentuale
##         0          11437413  0.95837585   95.837585
##         1            491317  0.04116896    4.116896

## --- Risultati EUSILC ---"
##   Categoria Frequenza_Ponderata Proporzione Percentuale
##         0        13141355.2  0.90920015   90.920015
##         1         658237.4  0.04554093    4.554093
```

Questi risultati evidenziano una differenza di 0.44 punti percentuali per quanto riguarda la classe 1 e 4.92 punti per la classe 0. La discrepanza è data dalla presenza di missing values nel dataset di EUSILC che sono stati poi annullati prima di effettuare il matching.
È stato effettuato anche un test del Chi-quadro di Pearson con aggiustamento di Rao & Scott per valutare se esiste o meno una differenza delle proporzioni nei due gruppi, come illustrato precedentemente.

```
## Test Chi-quadro Ponderato (Rao & Scott adjustment):
## Statistica Chi-quadro = 6.34, df = 1
## P-value = 4.92e-02
## Conclusione (alpha = 0.05): Rifiutiamo l'ipotesi nulla.
## Esiste una differenza statisticamente significativa nelle proporzioni tra i due dataset.
```

Il test suggerisce che esiste una differenza statisticamente significativa tra i due gruppi. La figura 6 completa l'analisi riportando le distribuzioni pre e post matching.

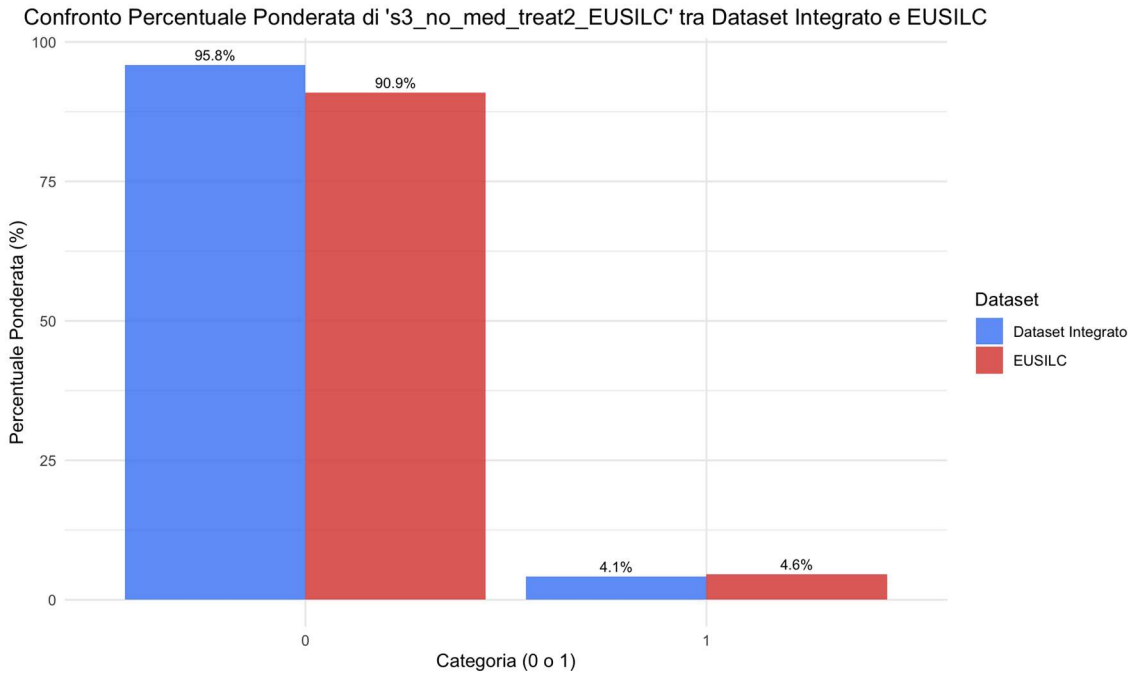

Figura 6: Distribuzione pre e post matching

- **Grado di urbanizzazione (UrbDeg)**

```
## --- Risultati Dataset Integrato ---
##   Categoria Frequenza_Ponderata Proporzione Percentuale
##         0          9776884    0.819235     81.9235
##         1          2157279    0.180765     18.0765
## --- Risultati EUSILC ---
##   Categoria Frequenza_Ponderata Proporzione Percentuale
##         0         11897280   0.8231273    82.31273
##         1          2556474   0.1768727    17.68727
```

Questi risultati evidenziano una differenza di 0.39 punti percentuali sia per quanto riguarda la classe 1 che per la classe 0. La discrepanza è data dalla presenza di missing values nel dataset di EUSILC che sono stati poi annullati prima di effettuare il matching.
È stato effettuato anche un test del Chi-quadro di Pearson con aggiustamento di Rao & Scott per valutare se esiste o meno una differenza delle proporzioni nei due gruppi, come illustrato precedentemente. Il test conferma che non ci sia differenza statisticamente significativa, come evidenziato anche in Figura 7.

```
## Test Chi-quadro Ponderato (Rao & Scott adjustment):
## Statistica Chi-quadro = 0.66, df = 1
## P-value = 4.85e-01
## Conclusione (alpha = 0.05): Non rifiutiamo l'ipotesi nulla.
## Non esiste una differenza statisticamente significativa nelle proporzioni tra i due dataset.
```

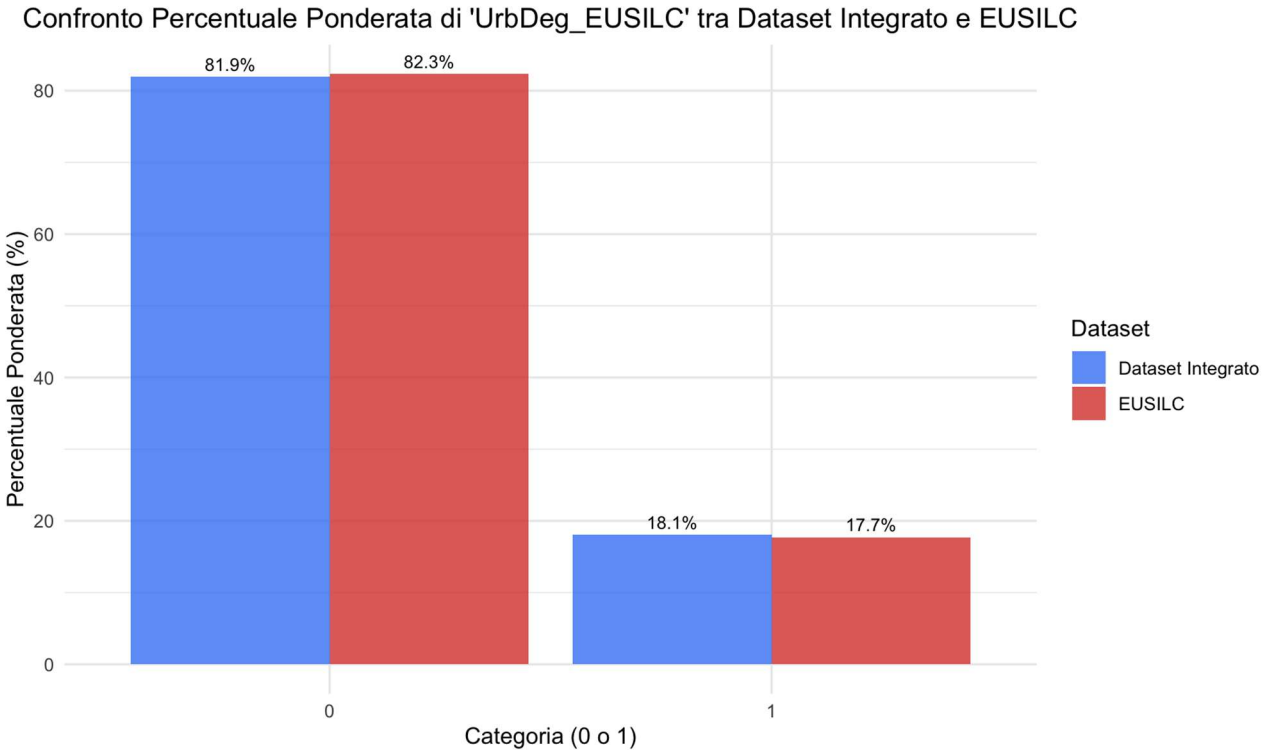

Figura 7: Distribuzione pre e post matching

8. EHIS_AVQ_SHARE_EUSILC+SPESE

Come ultimo dataset, consideriamo il dataset "Spese delle Famiglie" (Acronimo SdF ma che, nel corso del presente report, chiameremo SPESE).
Il dataset contiene 10894 righe che corrispondono a 13686175 unità.
Per quanto riguarda il dataset integrato, dopo aver unito ad EHIS, nell'ordine AVQ, SHARE e EUSILC, si hanno 12644 righe per un totale di 12361236 individui ultra sessantacinquenni.
Per quanto riguarda le variabili da trasferire da SPESE al dataset integrato sono state scelte:

- Assenza di condizionatore (Vabi_cond)
- Povertà Assoluta (povassc)
- Povertà Relativa (poveri)

Le variabili da importare sono stare precedentemente codificate in 0/1 (0= non deprived, 1= deprived). Inoltre, per ogni dataset, è stato preso in considerazione il vettore del peso campionario.
Sono state utilizzate le seguenti variabili per effettuare il matching:

- Genere (1= Maschio, 0=Femmina),
- Titolo di Studio (1=Fino alla licenza elementare, 2= Licenza media, 3= Diploma e oltre.
- Macroarea (1=Nord Ovest; 2=Nord Est, 3=Centro, 4 =Sud e Isole),
- Regione (01=Piemonte, 02=Valle D'Aosta, 03=Lombardia, 04=Trentino Alto Adige, 05=Veneto, 06=Friuli Venezia Giulia, 07=Liguria, 08=Emilia Romagna, 09=Toscana, 10=Umbria, 11=Marche, 12=Lazio, 13=Abruzzo, 14=Molise, 15=Campania, 16=Puglia, 17=Basilicata, 18=Calabria, 19=Sicilia, 20=Sardegna),
- Condizione lavorativa (1= occupato, 2= disoccupato, 3=pensione e 4= altra),
- Età (1=65-74, 2=75 e più),
- Reddito (Espresso in quintili, 1= I quintile rappresenta i redditi più bassi, … , 5= V quintile, redditi più alti).

Dove necessario, le variabili sono state ricodificate in modo da avere le stesse classi.
Per esempio, per quanto riguarda la variabile "Titolo di studio" questo ha richiesto una ricodifica sul dataset Integrato, con una riduzione delle classi. Discorso simile è stato fatto per la variabile "Condizione lavorativa" dove dalle 6 classi del dataset integrato si è passati a 4 classi.
Infine, i due dataset sono stati combinati.
La numerosità non è cambiata perché non erano presenti valori mancanti.
Prima di procedere con il matching è stata verificata la coerenza delle variabili da usare come link nei due dataset.
Di seguito, per ciascuna variabile, riportiamo le proporzioni in ogni classe:

Genere:

```
##    SPESE   EHIS
##  0 0.5509 0.5561
##  1 0.4491 0.4439
```

Titolo di Studio

```
##    SPESE   EHIS
##  1 0.4532 0.4794
##  2 0.2502 0.2481
##  3 0.2966 0.2725
```

Macroarea

```
##    SPESE   EHIS
##  1 0.2411 0.2261
##  2 0.2089 0.2020
##  3 0.2046 0.2053
##  4 0.3453 0.3666
```

Regione

```
##     SPESE   EHIS
##  01 0.0558 0.0742
##  02 0.0254 0.0215
##  03 0.1161 0.0857
##  04 0.0554 0.0493
##  05 0.0621 0.0548
##  06 0.0375 0.0397
##  07 0.0438 0.0447
##  08 0.0540 0.0582
##  09 0.0579 0.0745
##  10 0.0291 0.0259
##  11 0.0330 0.0426
##  12 0.0846 0.0622
##  13 0.0245 0.0394
##  14 0.0325 0.0244
##  15 0.0618 0.0563
##  16 0.0640 0.0577
##  17 0.0322 0.0317
##  18 0.0366 0.0485
##  19 0.0598 0.0660
##  20 0.0340 0.0427
```

Condizione lavorativa

```
##    SPESE   EHIS
##  1 0.0520 0.0263
##  2 0.0016 0.0010
##  3 0.7372 0.7631
##  4 0.2093 0.2096
```

Età

```
##    SPESE   EHIS
##  1 0.5288 0.4827
##  2 0.4712 0.5173
```

Livelli di reddito

```
##    SPESE   EHIS
##  1 0.1622 0.1294
##  2 0.2249 0.2146
##  3 0.2221 0.2063
##  4 0.1883 0.2237
##  5 0.2025 0.2260
```

Le proporzioni sono abbastanza simili, tali da garantire di poter procedere con l'analisi.
Analizzando le classi di ciascuna variabile del matching sono state individuate 19200 classi teoriche ("bidoni") per il matching.
Il matching, anche in questo caso è stato effettuato in R, con la funzione MatchIt metodo CEM.
Di seguito il commento ai risultati.

Commento ai risultati del matching

L'obiettivo del matching è rendere i gruppi di trattamento e controllo il più simili possibile sulle caratteristiche individuate. L'output completo è riportato in APPENDICE 6.
Ricordiamo che il metodo Coarsened Exact Matching (CEM) utilizzato per fare il matching funziona utilizzando le categorie esistenti per le singole variabili e, quindi, eseguendo un matching esatto su queste variabili raggruppate.
Ciò significa che per ogni strato abbinato, le unità trattate e di controllo avranno esattamente gli valori raggruppati per tutte le covariate specificate.
L'output in Appendice 6 mostra sia i risultati prima del matching (Summary of Balance for All Data") che dopo il matching (Summary of Balance for Matched Data).
L'efficacia del metodo è dal confronto tra queste tabelle.
Per quanto riguarda il bilanciamento delle covariate prima del matching ricordiamo che uno Std. Mean Diff. vicino a 0 indica un buon bilanciamento.
Generalmente, valori inferiori a |0.1| o |0.25| sono considerati accettabili, con valori più piccoli che indicano un bilanciamento migliore. Osservando la tabella "All Data", alcune variabili come

Regione03 (-0.1085), Regione12 (-0.0927), Income1 (-0.0978), Income4 (0.0849), WorkStatus21 (-0.1607) e Age1 (-0.0923) mostrano già un certo squilibrio, con WorkStatus21 il più sbilanciato a -0.1607.

Per quanto riguarda la Var. Ratio (Variance Ratio), poiché le variabili usate sono di tipo categorico, osserviamo un ".". Infine, eCDF Mean / eCDF Max (Empirical Cumulative Distribution Function) misurano la massima differenza tra le funzioni di distribuzione cumulativa empiriche di ciascuna covariata nei gruppi di trattamento e controllo. Valori vicini a 0 indicano un buon bilanciamento sull'intera distribuzione della covariata, non solo sulla media.

La “Summary of Balance for Matched Data” riporta l'impatto del CEM.

Osserviamo che per tutte le covariate elencate, le medie/proporzioni sono identiche (ad esempio, Gender0 è 0.5518 in entrambi i gruppi, Education21 è 0.4864 in entrambi i gruppi, ecc.).

Inoltre, per ogni singola covariata, lo Std. Mean Diff. è 0 (o molto vicino a 0, rappresentato come -0).

Questa è la caratteristica distintiva dei metodi di matching esatto come il CEM, che mirano a un bilanciamento perfetto all'interno di ogni strato e eCDF Mean / eCDF Max sono tutti 0.

Questi ultimi valori confermano che le intere distribuzioni delle covariate sono perfettamente bilanciate tra i gruppi di trattamento e controllo dopo il matching.

Concludendo, il CEM ha raggiunto con successo un bilanciamento perfetto su tutte le covariate specificate.

La sezione riguardante la dimensione nel campione riporta prima la numerosità delle unità (senza considerare i pesi) presenti nel dataset: 10894 unità nel gruppo di controllo e 12644 unità nel gruppo di trattamento inizialmente.

Matched (ESS) (Effective Sample Size) rappresenta la dimensione campionaria di un campione non ponderato che avrebbe la stessa potenza statistica del campione abbinato: 6524.89 per il gruppo di controllo; mentre per quello trattato 12092.

Il metodo ha portato a 10191 unità dal gruppo di controllo che sono state abbinate e 12092 unità dal gruppo di trattamento che sono state abbinate. Di conseguenza, le unità che non hanno trovato un corrispettivo (Unmatched) sono 703 unità dal gruppo di controllo e 552 unità dal gruppo di trattamento. Infine, nessuna unità è stata scartata (Discarded).

Notiamo che un numero significativo di unità è stato abbinato: 12092 unità trattate e 10191 unità di controllo.

Alcune unità (703 di controllo, 552 trattate) non sono state abbinate.

L'ESS per il gruppo di controllo è notevolmente inferiore al conteggio grezzo Matched (6524.89 vs 10191).

Questo suggerisce che i pesi (combined$w) hanno avuto un effetto sostanziale: probabilmente ciò indica che le unità di controllo abbinate avevano pesi più variabili rispetto alle unità trattate abbinate, o che alcuni pesi erano piuttosto piccoli per le unità di controllo mantenute.

L'ESS per il gruppo trattato è molto vicino al conteggio abbinato, il che implica che i pesi per le unità trattate erano meno variabili o più vicini a 1.

Infine, la tabella di bilanciamento sui dati abbinati mostra che tutti i valori nella colonna Diff.Un sono estremamente vicini a zero.

Ad esempio: Gender ha un valore di -0.0043, Education2_1 di 0.0248, Income_1 di -0.0374, e Age_2 di 0.0412.

Questi valori indicano che le proporzioni dei gruppi trattato e di controllo sono quasi identiche per tutte le covariate dopo il matching.

L'analisi congiunta di tutti questi valori suggerisce un ottimo risultato per il bilanciamento ottenuto tramite CEM.

Il matching con metodo CEM, per le 19200 combinazioni teoriche individuati, ne ha selezionate 1518 non vuote.

Passo successivo del matching è quello di abbinare, all'interno di ciascun "bidone", le osservazioni delle variabili da esportare da SPESE al dataset integrato.

Pertanto, il dataset integrato completo consta di 12092 righe che corrispondono a 11934162 individui. Questo ultimo step ha determinato una perdita di 3,46% delle unità.

Come illustrato nei paragrafi precedenti, poiché le variabili da importare sono di tipo binomiale (0/1), si è deciso di replicare la distribuzione osservata nel "bidone" alle osservazioni dell'altro dataset presenti nello stesso "bidone".

Per valutare l'effettiva bontà del metodo appena illustrato, sono state confrontate, per ciascuna delle variabili importate, le distribuzioni nel dataset integrato con la stessa variabile nel dataset originario.

Di seguito, per il dataset SPESE, riportiamo le differenze tra la variabile originaria e quella nel dataset integrato.

- **Povertà Assoluta (povassc)**

```
--- Risultati Dataset Integrato ---"
##   Categoria Frequenza_Ponderata Proporzione Percentuale
##         0        11397339.6  0.95501799   95.501799
##         1          536822.6  0.04498201    4.498201

## --- Risultati SPESE ---
##   Categoria Frequenza_Ponderata Proporzione Percentuale
##         0         55275163  0.92368116   92.368116
##         1          4567091  0.07631884    7.631884
```

La variabile povertà assoluta presenta una differenza di 3.13 punti percentuali.
Il tet Chi-quadro di Pearson con aggiustamento di Rao & Scott suggerisce che esite una differenza statisticamente significativa nei due gruppi, anche se tali differenze non sono sostanziali (Figura 8).

```
## Test Chi-quadro Ponderato (Rao & Scott adjustment):
## Statistica Chi-quadro = 113.05, df = 1
## P-value = 9.39e-25
## Conclusione (alpha = 0.05): Rifiutiamo l'ipotesi nulla.
## Esiste una differenza statisticamente significativa nelle proporzioni tra i due dataset.
```

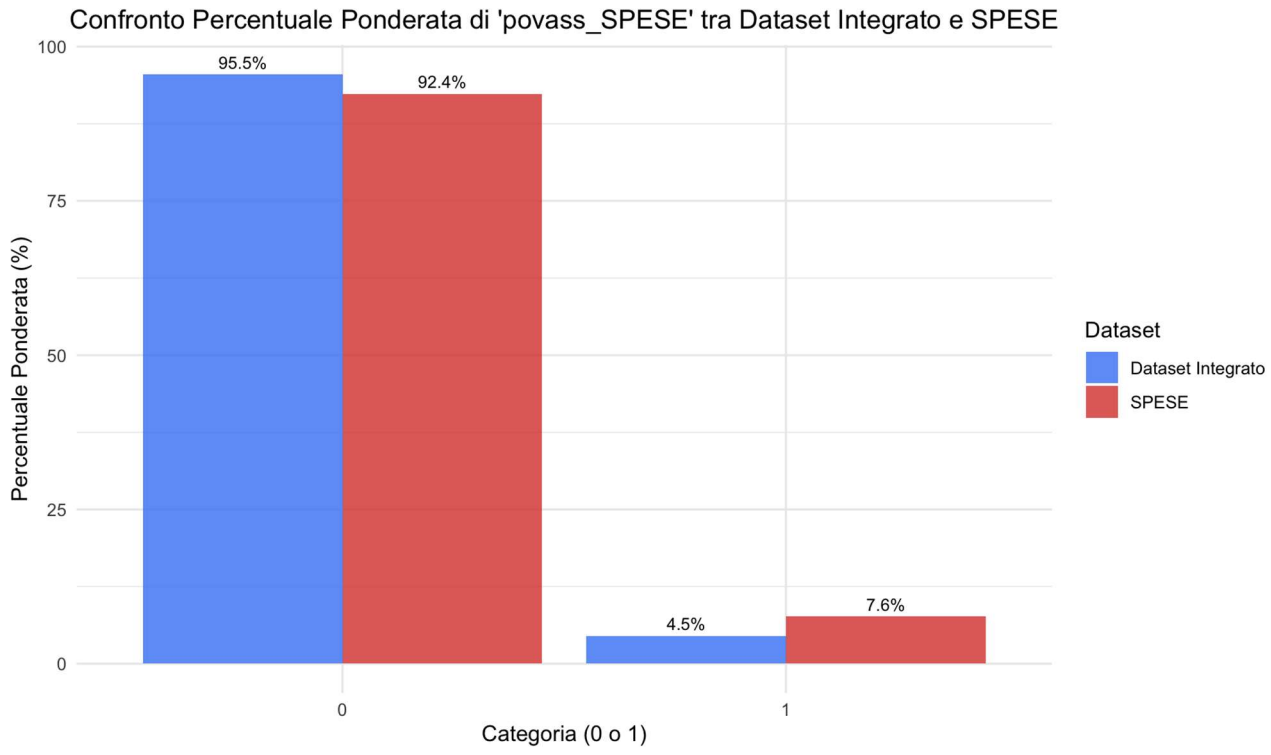

Figura 8: Distribuzione pre e post matching

- **Povertà Relativa (poveri)**

```
--- Risultati DatasetIntegrato ---
##   Categoria Frequenza_Ponderata Proporzione Percentuale
##         0         10889151  0.91243529   91.243529
##         1          1045011  0.08756471    8.756471

--- Risultati SPESE ---
##   Categoria Frequenza_Ponderata Proporzione Percentuale
##         0         51054113   0.8531449    85.31449
##         1          8788141   0.1468551    14.68551
```

Questi risultati evidenziano una differenza di quasi 6 punti percentuali. Le differenze sono riportate anche in Figura 9. L'output del test del Chi quadro di Pearson con aggiustamento di Rao & Scott suggerisce che vi sia una differenza tra i gruppi.

```
## Test Chi-quadro Ponderato (Rao & Scott adjustment):
## Statistica Chi-quadro = 226.09, df = 1
## P-value = 5.32e-48
## Conclusione (alpha = 0.05): Rifiutiamo l'ipotesi nulla.
## Esiste una differenza statisticamente significativa nelle proporzioni tra i due dataset.
```

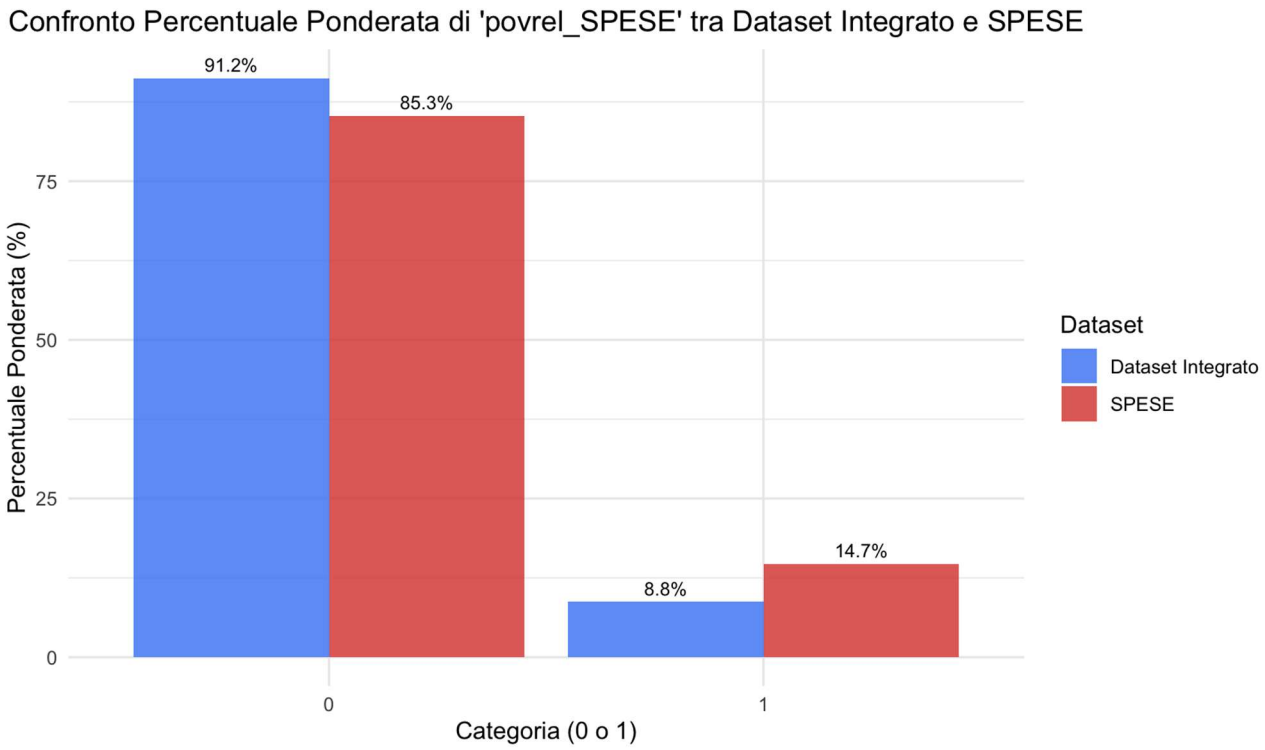

Figura 9: Distribuzione pre e post matching

- **Assenza di condizionatore (Vabi_cond)**

```
--- Risultati DatasetIntegrato ---
##   Categoria Frequenza_Ponderata Proporzione Percentuale
##         0          4907678   0.4112294    41.12294
##         1          7026484   0.5887706    58.87706

--- Risultati SPESE ---
##   Categoria Frequenza_Ponderata Proporzione Percentuale
##         0         27140090   0.4535272    45.35272
##         1         32702164   0.5464728    54.64728
```

Questi risultati evidenziano una differenza di poco più di 4 punti percentuali (Figura 10).
Il test del Chi-quadro di Pearson con aggiustamento di Rao & Scott evidenzia una differenza statisticamente significativa nelle proporzioni tra il dataset integrato e quello delle SPESE

```
## Test Chi-quadro Ponderato (Rao & Scott adjustment):
## Statistica Chi-quadro = 55.05, df = 1
## P-value = 7.45e-13
## Conclusione (alpha = 0.05): Rifiutiamo l'ipotesi nulla.
## Esiste una differenza statisticamente significativa nelle proporzioni tra i due dataset.
```

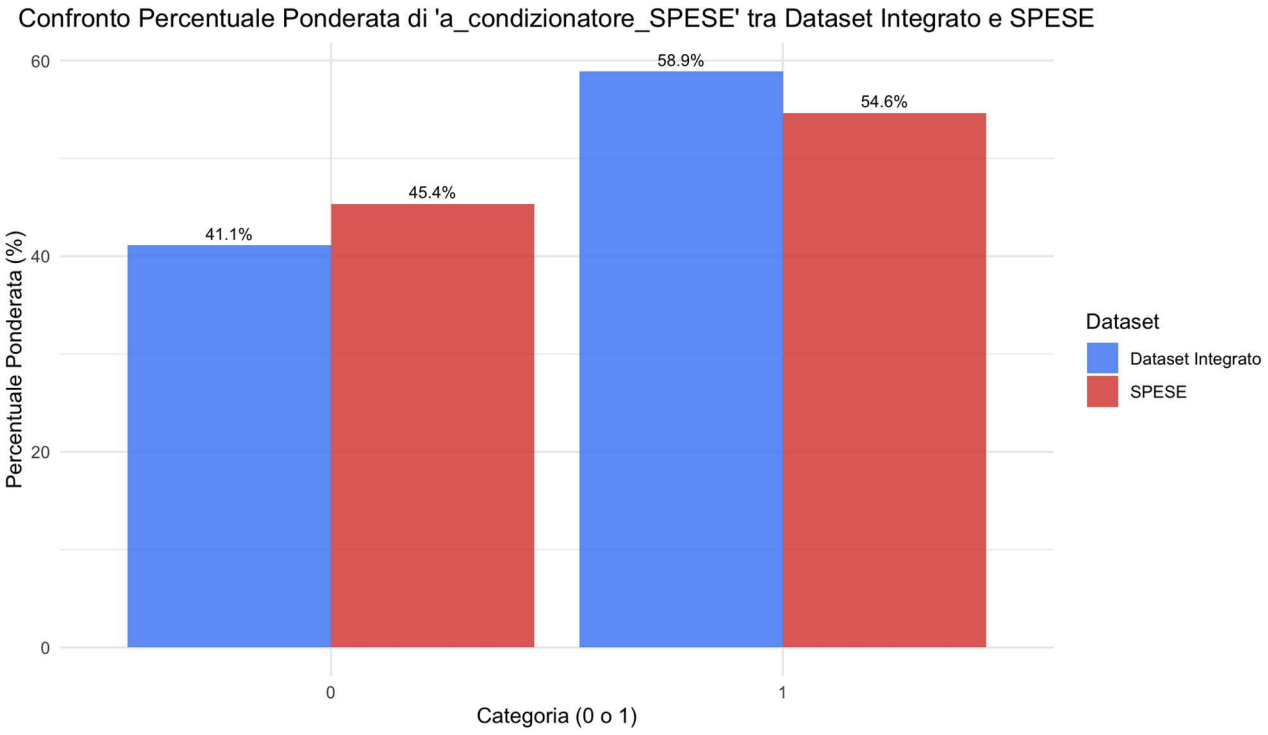

Figura 10: Distribuzione pre e post matching

9. Commenti conclusivi al matching

Nei paragrafi precedenti abbiamo illustrato in dettaglio i passi che hanno portato alla costruzione del dataset integrato.

Dal primo dataset, EHIS, sono state nell'ordine agganciate variabili appartenenti ai seguenti dataset: AVQ, SHARE, EUSILC e SPESE.

Il dataset originario aveva 13722 righe per un totale di 13576875 individui.

L'operazione di link tra i dataset ha comportato una perdita complessiva del 12.09934% delle osservazioni che sono passate a 11934162 (12092 righe).

Questo, sebbene sembrerebbe un valore non trascurabile, è giustificato dal fatto di aver aggiunto diverse variabili, integrando aspetti cruciali per una definizione e misurazione multidimensionale della vulnerabilità nella popolazione anziana, sia abitativa che sociale.

Pertanto, il dataset integrato contiene 11 variabili per la dimensione abitare (VAA) e 14 per la dimensione salute (VSA). Inoltre, sono state individuate 9 variabili per analizzare i possibili sottogruppi nel calcolo dell'indicatore.

Nella sezione successiva i due indici VSA e VAA saranno calcolati per alcune delle variabili utilizzate per creare il dataset integrato.

PARTE 2: COSTRUZIONE DI VAA E VSA

1. L'indice di Vulnerabilità Abitativa per gli Anziani

Procediamo al calcolo dell'indice di Vulnerabilità Abitativa per gli anziani (VAA).
L'analisi è stata condotta sul dataset Integrato ottenuto dalla ripetizione della procedura di matching tra diversi dataset: EHIS, AVQ, SHARE, EUSILC e SPESE come descritto nella sezione precedente che contiene 11 variabili (Tabella 1).

Tabella 1: VAA

DATASET	**Num. Variabili**	**Variabili**
EHIS	6	1. Difficoltà di accesso all'abitazione (A1_diff_accesso_abitaz) 2. Problemi all'abitazione (A2_prob_abitaz) 3. Aiuto vicinato (A3_aiuto_vic) 4. Spese alte (A4_spese_alte) 5. Persona isolata (A5_pers_isolata) 6. Abitazione distante dai familiari (A6_dist_fam)
AVQ	1	7. Contesto abitativo (a4_contesto_abitaz)
SHARE	1	8. Scale of social connectedness (sn_scale)
EUSILC	2	9. Difficoltà a sostenere i costi dell'abitazione (cost_prob) 10. Condizione di sotto-occupazione dell'abitazione (under_occ2)
SPESE	1	11. Assenza di condizionatore (Vabi_cond)

La Tabella 2 riporta il calcolo della vulnerabilità (VAA) accompagnato dai valori di incidenza (H) e intensità (A) per ciascun valore di k (da 1 a 11)[2].
I valori dell'indice vengono riportati a livello nazionale, di macroarea e regionale. Per quanto riguarda il dettaglio per macroarea e regione, i codici identificativi di riferimento sono i seguenti:

- Macroarea (1=Nord Ovest; 2=Nord Est, 3=Centro, 4 =Sud e Isole);
- Regione (01=Piemonte, 02=Valle D'Aosta, 03=Lombardia, 04=Trentino Alto Adige, 05=Veneto, 06=Friuli Venezia Giulia, 07=Liguria, 08=Emilia Romagna, 09=Toscana, 10=Umbria, 11=Marche, 12=Lazio, 13=Abruzzo, 14=Molise, 15=Campania, 16=Puglia, 17=Basilicata, 18=Calabria, 19=Sicilia, 20=Sardegna).

[2] Ciommi, M., F. Mariani, M.C. Recchioni, (2025), *Rapporto di revisione e individuazione tecnica per analisi empirica*, WP2, progetto VAI - Vulnerabilità abitativa e di salute degli Anziani in Italia, Università Politecnica delle Marche.

Tabella 2: VAA per l'Italia

Indicatore	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	k=11
H	0,9639	0,8298	0,5885	0,3315	0,1467	0,0499	0,0113	0,0025	0,0003	0,0000	0,0000
A	0,2758	0,3057	0,3565	0,4214	0,4942	0,5712	0,6590	0,7399	0,8182	0,0000	0,0000
MPI	0,2659	0,2537	0,2098	0,1397	0,0725	0,0285	0,0075	0,0018	0,0003	0,0000	0,0000

Tabella 3: VAA per macroarea

Macroarea	Indicatore	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	k=11
1	H	0,9609	0,8189	0,5645	0,2906	0,1202	0,0338	0,0066	0,0010	0,0000	0,0000	0,0000
	A	0,2646	0,2947	0,3455	0,4142	0,4858	0,5659	0,6506	0,7273			
	MPI	0,2542	0,2413	0,1950	0,1203	0,0584	0,0191	0,0043	0,0007			
2	H	0,9456	0,7824	0,5241	0,2846	0,1142	0,0443	0,0106	0,0035	0,0000	0,0000	0,0000
	A	0,2605	0,2958	0,3520	0,4188	0,5011	0,5744	0,6660	0,7273			
	MPI	0,2463	0,2315	0,1845	0,1192	0,0572	0,0255	0,0071	0,0025			
3	H	0,9715	0,8426	0,6127	0,3437	0,1578	0,0557	0,0098	0,0020	0,0008	0,0000	0,0000
	A	0,2804	0,3094	0,3573	0,4234	0,4939	0,5661	0,6623	0,7654	0,8182		
	MPI	0,2724	0,2607	0,2189	0,1456	0,0779	0,0315	0,0065	0,0015	0,0007		
4	H	0,9725	0,8592	0,6315	0,3861	0,1812	0,0629	0,0167	0,0034	0,0005	0,0000	0,0000
	A	0,2911	0,3175	0,3664	0,4260	0,4965	0,5752	0,6577	0,7410	0,8182		
	MPI	0,2831	0,2728	0,2314	0,1645	0,0899	0,0362	0,0110	0,0025	0,0004		

Tabella 4: VAA per regione

Regione	Indicatore	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	k=11
01	H	0,9719	0,8509	0,6145	0,3423	0,1585	0,0448	0,0082	0,0000	0,0000	0,0000	0,0000
	A	0,2798	0,3067	0,3547	0,4198	0,4850	0,5621	0,6364				
	MPI	0,2719	0,2609	0,2179	0,1437	0,0769	0,0252	0,0052				
02	H	0,9857	0,8837	0,6293	0,3652	0,1436	0,0496	0,0000	0,0000	0,0000	0,0000	0,0000
	A	0,2820	0,3040	0,3534	0,4117	0,4860	0,5455					
	MPI	0,2779	0,2687	0,2224	0,1504	0,0698	0,0271					
03	H	0,9526	0,7971	0,5315	0,2558	0,1026	0,0258	0,0040	0,0000	0,0000	0,0000	0,0000
	A	0,2548	0,2867	0,3391	0,4107	0,4809	0,5595	0,6364				
	MPI	0,2427	0,2285	0,1803	0,1051	0,0493	0,0144	0,0025				
04	H	0,9849	0,8795	0,5985	0,3361	0,1268	0,0384	0,0061	0,0024	0,0000	0,0000	0,0000
	A	0,2744	0,2964	0,3502	0,4106	0,4881	0,5656	0,6719	0,7273			
	MPI	0,2702	0,2607	0,2096	0,1380	0,0619	0,0217	0,0041	0,0017			
05	H	0,9291	0,7525	0,4810	0,2738	0,1124	0,0451	0,0131	0,0062	0,0000	0,0000	0,0000
	A	0,2557	0,2944	0,3579	0,4223	0,5066	0,5844	0,6795	0,7273			
	MPI	0,2376	0,2215	0,1722	0,1156	0,0569	0,0263	0,0089	0,0045			
06	H	0,9499	0,7693	0,5235	0,2806	0,0986	0,0354	0,0100	0,0000	0,0000	0,0000	0,0000
	A	0,2553	0,2939	0,3465	0,4103	0,4963	0,5711	0,6364				
	MPI	0,2425	0,2261	0,1814	0,1151	0,0490	0,0202	0,0063				
07	H	0,9731	0,8416	0,5963	0,3263	0,1070	0,0444	0,0159	0,0093	0,0000	0,0000	0,0000
	A	0,2722	0,3005	0,3494	0,4128	0,5137	0,5969	0,6893	0,7273			
	MPI	0,2649	0,2529	0,2083	0,1347	0,0549	0,0265	0,0109	0,0067			
08	H	0,9531	0,7972	0,5535	0,2862	0,1181	0,0475	0,0093	0,0019	0,0000	0,0000	0,0000
	A	0,2639	0,2977	0,3488	0,4198	0,4997	0,5668	0,6549	0,7273			
	MPI	0,2515	0,2374	0,1930	0,1202	0,0590	0,0269	0,0061	0,0014			
09	H	0,9787	0,8323	0,5981	0,3374	0,1733	0,0659	0,0101	0,0019	0,0000	0,0000	0,0000
	A	0,2784	0,3114	0,3622	0,4313	0,4954	0,5621	0,6535	0,7273			
	MPI	0,2725	0,2592	0,2166	0,1455	0,0858	0,0370	0,0066	0,0014			
10	H	0,9923	0,9127	0,6489	0,3762	0,1383	0,0446	0,0211	0,0061	0,0000	0,0000	0,0000
	A	0,2877	0,3049	0,3549	0,4144	0,5017	0,6009	0,6627	0,7273			
	MPI	0,2855	0,2782	0,2303	0,1559	0,0694	0,0268	0,0140	0,0044			
11	H	0,9741	0,8790	0,6483	0,3530	0,1198	0,0470	0,0074	0,0000	0,0000	0,0000	0,0000
	A	0,2826	0,3034	0,3466	0,4085	0,4958	0,5597	0,6364				
	MPI	0,2753	0,2667	0,2247	0,1442	0,0594	0,0263	0,0047				
12	H	0,9611	0,8276	0,6076	0,3404	0,1601	0,0520	0,0084	0,0019	0,0019	0,0000	0,0000
	A	0,2801	0,3106	0,3572	0,4235	0,4910	0,5667	0,6775	0,8182	0,8182		
	MPI	0,2692	0,2570	0,2170	0,1442	0,0786	0,0295	0,0057	0,0016	0,0016		
13	H	0,9891	0,9391	0,7907	0,5197	0,2487	0,1151	0,0472	0,0132	0,0027	0,0000	0,0000
	A	0,3369	0,3500	0,3816	0,4383	0,5196	0,5952	0,6669	0,7458	0,8182		
	MPI	0,3332	0,3287	0,3017	0,2278	0,1293	0,0685	0,0315	0,0098	0,0022		
14	H	0,9952	0,9246	0,6709	0,3921	0,1405	0,0494	0,0105	0,0000	0,0000	0,0000	0,0000
	A	0,2908	0,3060	0,3530	0,4101	0,4933	0,5647	0,6364				

	MPI	0,2894	0,2830	0,2368	0,1608	0,0693	0,0279	0,0067				
15	**H**	0,9762	0,8327	0,5732	0,3325	0,1478	0,0526	0,0085	0,0016	0,0000	0,0000	0,0000
	A	0,2724	0,3037	0,3588	0,4212	0,4931	0,5629	0,6532	0,7273			
	MPI	0,2659	0,2529	0,2057	0,1400	0,0729	0,0296	0,0056	0,0011			
16	**H**	0,9651	0,8286	0,5739	0,3245	0,1543	0,0440	0,0094	0,0000	0,0000	0,0000	0,0000
	A	0,2731	0,3032	0,3570	0,4218	0,4860	0,5649	0,6364				
	MPI	0,2636	0,2512	0,2049	0,1369	0,0750	0,0248	0,0060				
17	**H**	0,9874	0,9183	0,7173	0,4632	0,2185	0,0971	0,0287	0,0012	0,0000	0,0000	0,0000
	A	0,3160	0,3329	0,3752	0,4315	0,5074	0,5735	0,6403	0,7273			
	MPI	0,3120	0,3057	0,2692	0,1998	0,1109	0,0557	0,0184	0,0009			
18	**H**	0,9871	0,8924	0,6959	0,4286	0,1995	0,0605	0,0140	0,0000	0,0000	0,0000	0,0000
	A	0,3019	0,3243	0,3645	0,4218	0,4885	0,5665	0,6364				
	MPI	0,2980	0,2894	0,2537	0,1808	0,0975	0,0343	0,0089				
19	**H**	0,9678	0,8712	0,6673	0,4359	0,2170	0,0773	0,0228	0,0059	0,0013	0,0000	0,0000
	A	0,3068	0,3308	0,3763	0,4313	0,4995	0,5807	0,6652	0,7480	0,8182		
	MPI	0,2970	0,2882	0,2511	0,1880	0,1084	0,0449	0,0151	0,0044	0,0011		
20	**H**	0,9577	0,8494	0,6143	0,3760	0,1662	0,0509	0,0159	0,0066	0,0000	0,0000	0,0000
	A	0,2883	0,3135	0,3638	0,4216	0,4947	0,5857	0,6743	0,7273			
	MPI	0,2761	0,2662	0,2235	0,1585	0,0822	0,0298	0,0107	0,0048			

2. L'indice di Vulnerabilità di Salute per gli Anziani

Procediamo al calcolo dell'indice di Vulnerabilità di Salute per gli anziani (VSA).
L'analisi è stata condotta sul dataset Integrato ottenuto dalla ripetizione della procedura di matching tra diversi dataset: EHIS, AVQ, SHARE, EUSILC e SPESE come descritto nella sezione precedente che contiene 14 variabili (Tabella 3).

Tabella 5: VSA

DATASET	**Num. Variabili**	**Variabili**
EHIS	11	1. Numero limitazioni nelle attività di cura della persona (ADL) (S1_adl) 2. Numero limitazioni nelle attività quotidiane strumentali di tipo domestico (IADL) (S2_iadl) 3. Gravi difficoltà motorie (S3_diff_motorie) 4. Gravi difficoltà sensoriali (S4_diff_sensoriali) 5. Numero malattie croniche diagnosticate da un medico (S5_croniche) 6. Giudizio sullo stato di salute in generale (S6_perc_salute) 7. Difficoltà a ricordare o a concentrarsi (S7_diff_ricordare) 8. PHQ8 - Patient Health Questionnaire, Depression Scale (S8_phq8) 9. OSS3 - Oslo Social Support Scale (S9_osss3) 10. Persona isolata (S10_pers_isolata) 11. Bisogno di aiuto (S11_bisogno_aiuto)
AVQ	1	12. Mental Health Index-5 (MHI-5) (s4_mhi5)
SHARE	1	13. Scale of social connectedness (sn_scale)
EUSILC	1	14. Rinuncia alle cure mediche specialistiche (no_med_treat2)
SPESE	0	

La Tabella 6 riporta il calcolo della vulnerabilità (VSA) accompagnato dai valori di incidenza (H) e intensità (A) per ciascun valore di k (da 1 a 14)[3].

[3] Ciommi, M., F. Mariani, M.C. Recchioni, (2025), *Rapporto di revisione e individuazione tecnica per analisi empirica*, WP2, progetto VAI - Vulnerabilità abitativa e di salute degli Anziani in Italia, Università Politecnica delle Marche.

I valori dell'indice vengono riportati a livello nazionale, di macroarea e regionale. Per quanto riguarda il dettaglio per macroarea e regione, i codici identificativi di riferimento sono i seguenti:

- Macroarea (1=Nord Ovest; 2=Nord Est, 3=Centro, 4 =Sud e Isole);
- Regione (01=Piemonte, 02=Valle D'Aosta, 03=Lombardia, 04=Trentino Alto Adige, 05=Veneto, 06=Friuli Venezia Giulia, 07=Liguria, 08=Emilia Romagna, 09=Toscana, 10=Umbria, 11=Marche, 12=Lazio, 13=Abruzzo, 14=Molise, 15=Campania, 16=Puglia, 17=Basilicata, 18=Calabria, 19=Sicilia, 20=Sardegna).

Tabella 6: VSA per l'Italia

Indicatore	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	k=11	k=12	k=13	k=14
H	0,8823	0,6549	0,4355	0,3023	0,2180	0,1609	0,1165	0,0784	0,0468	0,0234	0,0083	0,0028	0,0004	0,0000
A	0,2373	0,2948	0,3714	0,4406	0,5004	0,5513	0,5981	0,6459	0,6962	0,7493	0,8131	0,8662	0,9286	0,0000
MPI	0,2093	0,1931	0,1617	0,1332	0,1091	0,0887	0,0697	0,0506	0,0325	0,0175	0,0067	0,0024	0,0003	0,0000

Tabella 7: VSA per macroarea

Macroarea	Indicatore	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	k=11	k=12	k=13	k=14
1	H	0,8656	0,6091	0,3726	0,2525	0,1799	0,1231	0,0829	0,0441	0,0243	0,0090	0,0021	0,0005	0,0000	0,0000
	A	0,2117	0,2708	0,3520	0,4175	0,4707	0,5231	0,5688	0,6295	0,6769	0,7347	0,8022	0,8571		
	MPI	0,1833	0,1649	0,1311	0,1054	0,0847	0,0644	0,0472	0,0277	0,0164	0,0066	0,0017	0,0004		
2	H	0,8615	0,6135	0,3953	0,2514	0,1717	0,1218	0,0876	0,0543	0,0296	0,0121	0,0036	0,0022	0,0019	0,0000
	A	0,2161	0,2746	0,3473	0,4235	0,4874	0,5408	0,5846	0,6364	0,6908	0,7597	0,8659	0,9180	0,9286	
	MPI	0,1862	0,1685	0,1373	0,1065	0,0837	0,0659	0,0512	0,0346	0,0204	0,0092	0,0031	0,0020	0,0017	
3	H	0,8699	0,6479	0,4204	0,2846	0,2018	0,1536	0,1132	0,0776	0,0485	0,0241	0,0100	0,0036	0,0000	0,0000
	A	0,2345	0,2903	0,3701	0,4445	0,5096	0,5574	0,6034	0,6508	0,6984	0,7546	0,8119	0,8571		
	MPI	0,2039	0,1881	0,1556	0,1265	0,1028	0,0856	0,0683	0,0505	0,0339	0,0182	0,0081	0,0031		
4	H	0,9163	0,7216	0,5208	0,3849	0,2873	0,2198	0,1632	0,1212	0,0742	0,0415	0,0150	0,0046	0,0000	0,0000
	A	0,2705	0,3242	0,3942	0,4577	0,5161	0,5649	0,6122	0,6512	0,7016	0,7481	0,8074	0,8571		
	MPI	0,2479	0,2340	0,2053	0,1762	0,1483	0,1242	0,0999	0,0789	0,0521	0,0310	0,0121	0,0039		

Tabella 8: VSA per regione

Regione	Indicatore	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	k=11	k=12	k=13	k=14
01	H	0,8693	0,6328	0,3954	0,2654	0,1955	0,1337	0,0852	0,0430	0,0283	0,0064	0,0039	0,0000	0,0000	0,0000
	A	0,2185	0,2734	0,3518	0,4192	0,4669	0,5177	0,5685	0,6356	0,6690	0,7582	0,7857			
	MPI	0,1899	0,1730	0,1391	0,1113	0,0913	0,0692	0,0484	0,0273	0,0189	0,0049	0,0031			
02	H	0,8848	0,6280	0,3573	0,2155	0,1485	0,1187	0,0811	0,0406	0,0096	0,0065	0,0000	0,0000	0,0000	0,0000
	A	0,2011	0,2541	0,3383	0,4200	0,4805	0,5115	0,5499	0,5996	0,6909	0,7143				
	MPI	0,1779	0,1595	0,1209	0,0905	0,0714	0,0607	0,0446	0,0244	0,0066	0,0046				
03	H	0,8617	0,5904	0,3553	0,2449	0,1753	0,1206	0,0825	0,0417	0,0204	0,0109	0,0015	0,0008	0,0000	0,0000
	A	0,2077	0,2704	0,3547	0,4180	0,4706	0,5220	0,5652	0,6291	0,6893	0,7299	0,8247	0,8571		
	MPI	0,1790	0,1596	0,1260	0,1024	0,0825	0,0630	0,0466	0,0262	0,0140	0,0079	0,0013	0,0007		
04	H	0,8193	0,5531	0,3386	0,1998	0,1454	0,0972	0,0666	0,0424	0,0145	0,0045	0,0013	0,0000	0,0000	0,0000
	A	0,1990	0,2604	0,3349	0,4186	0,4685	0,5236	0,5673	0,6056	0,6715	0,7358	0,7857			
	MPI	0,1631	0,1440	0,1134	0,0837	0,0681	0,0509	0,0378	0,0257	0,0097	0,0033	0,0011			
05	H	0,8719	0,6158	0,3752	0,2181	0,1512	0,1143	0,0712	0,0471	0,0291	0,0116	0,0015	0,0000	0,0000	0,0000
	A	0,2054	0,2611	0,3369	0,4252	0,4870	0,5289	0,5896	0,6354	0,6751	0,7234	0,7857			
	MPI	0,1791	0,1608	0,1264	0,0928	0,0736	0,0605	0,0420	0,0299	0,0196	0,0084	0,0012			
06	H	0,8312	0,5479	0,3295	0,2279	0,1313	0,0804	0,0584	0,0394	0,0166	0,0099	0,0054	0,0027	0,0000	0,0000
	A	0,1960	0,2604	0,3383	0,3936	0,4730	0,5463	0,5907	0,6343	0,7206	0,7727	0,8220	0,8571		
	MPI	0,1629	0,1427	0,1115	0,0897	0,0621	0,0439	0,0345	0,0250	0,0120	0,0077	0,0044	0,0023		
07	H	0,8741	0,6394	0,3999	0,2584	0,1631	0,1070	0,0793	0,0593	0,0343	0,0065	0,0000	0,0000	0,0000	0,0000
	A	0,2142	0,2666	0,3407	0,4100	0,4826	0,5483	0,5903	0,6206	0,6565	0,7143				
	MPI	0,1872	0,1705	0,1363	0,1059	0,0787	0,0587	0,0468	0,0368	0,0225	0,0047				
08	H	0,8687	0,6439	0,4486	0,3046	0,2114	0,1476	0,1182	0,0690	0,0372	0,0149	0,0058	0,0048	0,0048	0,0000
	A	0,2368	0,2945	0,3605	0,4297	0,4932	0,5520	0,5826	0,6415	0,7013	0,7884	0,9037	0,9286	0,9286	
	MPI	0,2057	0,1896	0,1617	0,1309	0,1043	0,0815	0,0689	0,0443	0,0261	0,0118	0,0053	0,0045	0,0045	
09	H	0,8515	0,6188	0,3915	0,2514	0,1897	0,1430	0,1018	0,0678	0,0478	0,0209	0,0069	0,0017	0,0000	0,0000

	A	0,2259	0,2840	0,3659	0,4504	0,5039	0,5519	0,6018	0,6529	0,6870	0,7436	0,8033	0,8571		
	MPI	0,1923	0,1757	0,1432	0,1132	0,0956	0,0789	0,0613	0,0442	0,0328	0,0155	0,0055	0,0015		
10	**H**	0,8823	0,6805	0,4363	0,2931	0,1955	0,1522	0,0959	0,0680	0,0494	0,0282	0,0045	0,0025	0,0000	0,0000
	A	0,2338	0,2820	0,3599	0,4310	0,5035	0,5451	0,6137	0,6604	0,6938	0,7321	0,8258	0,8571		
	MPI	0,2063	0,1919	0,1570	0,1263	0,0985	0,0830	0,0588	0,0449	0,0343	0,0206	0,0037	0,0022		
11	**H**	0,9133	0,7009	0,4587	0,3308	0,2385	0,1778	0,1309	0,0741	0,0349	0,0262	0,0117	0,0030	0,0000	0,0000
	A	0,2425	0,2944	0,3744	0,4362	0,4945	0,5414	0,5818	0,6446	0,7265	0,7543	0,8039	0,8571		
	MPI	0,2215	0,2063	0,1717	0,1443	0,1179	0,0962	0,0761	0,0477	0,0254	0,0198	0,0094	0,0026		
12	**H**	0,8693	0,6494	0,4288	0,2952	0,2016	0,1550	0,1198	0,0879	0,0528	0,0253	0,0127	0,0055	0,0000	0,0000
	A	0,2386	0,2952	0,3735	0,4455	0,5198	0,5686	0,6098	0,6497	0,7018	0,7658	0,8167	0,8571		
	MPI	0,2074	0,1917	0,1602	0,1315	0,1048	0,0881	0,0730	0,0571	0,0370	0,0194	0,0104	0,0047		
13	**H**	0,9276	0,7226	0,5202	0,3528	0,2507	0,1906	0,1486	0,1142	0,0741	0,0433	0,0146	0,0027	0,0000	0,0000
	A	0,2589	0,3121	0,3779	0,4555	0,5247	0,5775	0,6196	0,6557	0,7013	0,7429	0,7991	0,8571		
	MPI	0,2401	0,2255	0,1966	0,1607	0,1316	0,1101	0,0921	0,0749	0,0520	0,0321	0,0116	0,0023		
14	**H**	0,8975	0,7169	0,5234	0,3863	0,2845	0,2147	0,1669	0,1296	0,0771	0,0451	0,0166	0,0047	0,0000	0,0000
	A	0,2756	0,3271	0,3952	0,4594	0,5215	0,5750	0,6169	0,6505	0,7044	0,7481	0,8059	0,8571		
	MPI	0,2474	0,2345	0,2068	0,1775	0,1484	0,1234	0,1030	0,0843	0,0543	0,0337	0,0134	0,0040		
15	**H**	0,9138	0,7057	0,4839	0,3544	0,2682	0,2129	0,1560	0,1221	0,0738	0,0352	0,0135	0,0042	0,0000	0,0000
	A	0,2614	0,3174	0,3974	0,4643	0,5217	0,5644	0,6139	0,6456	0,6940	0,7501	0,8078	0,8571		
	MPI	0,2388	0,2240	0,1923	0,1645	0,1399	0,1202	0,0958	0,0788	0,0512	0,0264	0,0109	0,0036		
16	**H**	0,9227	0,6951	0,4877	0,3561	0,2573	0,1964	0,1605	0,1126	0,0728	0,0373	0,0215	0,0026	0,0000	0,0000
	A	0,2572	0,3180	0,3925	0,4584	0,5247	0,5766	0,6098	0,6565	0,7030	0,7603	0,7942	0,8571		
	MPI	0,2373	0,2210	0,1914	0,1632	0,1350	0,1133	0,0978	0,0739	0,0512	0,0283	0,0171	0,0022		
17	**H**	0,9370	0,7682	0,5969	0,4736	0,3599	0,2606	0,1862	0,1337	0,0851	0,0419	0,0156	0,0082	0,0000	0,0000
	A	0,2948	0,3439	0,4015	0,4503	0,5023	0,5576	0,6092	0,6520	0,6980	0,7548	0,8233	0,8571		
	MPI	0,2762	0,2642	0,2397	0,2133	0,1808	0,1453	0,1134	0,0872	0,0594	0,0316	0,0128	0,0070		
18	**H**	0,9334	0,7369	0,5265	0,3929	0,2947	0,2334	0,1695	0,1229	0,0759	0,0451	0,0154	0,0048	0,0000	0,0000
	A	0,2718	0,3252	0,3980	0,4606	0,5188	0,5613	0,6113	0,6536	0,7043	0,7464	0,8081	0,8571		

	MPI	0,2537	0,2396	0,2096	0,1809	0,1529	0,1310	0,1036	0,0803	0,0535	0,0336	0,0125	0,0041		
19	**H**	0,9124	0,7187	0,5447	0,3998	0,3033	0,2155	0,1531	0,0968	0,0575	0,0374	0,0121	0,0031	0,0000	0,0000
	A	0,2704	0,3241	0,3819	0,4427	0,4926	0,5479	0,5965	0,6527	0,7083	0,7433	0,8037	0,8571		
	MPI	0,2467	0,2329	0,2080	0,1770	0,1494	0,1181	0,0913	0,0632	0,0407	0,0278	0,0097	0,0026		
20	**H**	0,8693	0,6328	0,3954	0,2654	0,1955	0,1337	0,0852	0,0430	0,0283	0,0064	0,0039	0,0000	0,0000	0,0000
	A	0,2185	0,2734	0,3518	0,4192	0,4669	0,5177	0,5685	0,6356	0,6690	0,7582	0,7857			
	MPI	0,1899	0,1730	0,1391	0,1113	0,0913	0,0692	0,0484	0,0273	0,0189	0,0049	0,0031			

PARTE 3: APPENDICI

APPENDICE 1: Il pacchetto MatchIT

Il pacchetto **MatchIt** in R è uno strumento fondamentale per l'inferenza causale in studi osservazionali, in particolare per l'implementazione del **matching**.

L'obiettivo principale del matching è quello di bilanciare le covariate pre-trattamento tra i gruppi trattati e di controllo, in modo da rendere i gruppi comparabili e ridurre il bias da confounding. Questo consente di stimare in modo più affidabile l'effetto causale di un trattamento o di un'esposizione.

Ecco una panoramica delle sue funzionalità chiave e del suo utilizzo:

Perché usare MatchIt?

Negli studi osservazionali, i partecipanti non sono assegnati casualmente al gruppo di trattamento o di controllo. Ciò significa che i gruppi possono differire sistematicamente su diverse caratteristiche (covariate), rendendo difficile attribuire le differenze nell'outcome al solo trattamento.

MatchIt affronta questo problema creando sottoinsiemi di dati "bilanciati", in cui i gruppi di trattamento e di controllo sono il più possibile simili su tutte le covariate rilevanti.

I quattro passaggi fondamentali di un'analisi di matching (secondo MatchIt):

1. **Pianificazione:** Definire l'effetto causale che si vuole stimare (ad esempio, l'effetto medio del trattamento sui trattati - ATT, o l'effetto medio del trattamento sull'intera popolazione - ATE), scegliere le covariate da bilanciare e considerare il tipo di matching appropriato.
2. **Matching:** Implementare il matching vero e proprio utilizzando la funzione matchit().
3. **Valutazione del bilanciamento:** Verificare se il matching ha effettivamente ridotto lo sbilanciamento delle covariate tra i gruppi. MatchIt fornisce strumenti per questa valutazione.
4. **Stima dell'effetto del trattamento:** Analizzare il dataset bilanciato per stimare l'effetto causale e la sua incertezza.

Funzionalità principali del pacchetto MatchIt:

1. Funzione matchit(): Il cuore del pacchetto

Questa è la funzione principale per eseguire il matching. Richiede una formula che specifica la variabile di trattamento e le covariate da bilanciare, e il dataset.

Sintassi di base: m.out <- matchit(treat ~ x1 + x2 + x3, data = mydata, method = "nearest", distance = "glm")

- treat: La variabile binaria che indica lo stato di trattamento (ad esempio, 1 = trattato, 0 = controllo).
- x1 + x2 + x3: Le covariate pre-trattamento che si desidera bilanciare.
- data: Il dataframe contenente le variabili.
- method: Il metodo di matching da utilizzare.
- distance: Il metodo per stimare il punteggio di propensione (se applicabile al metodo di matching scelto).

2. Metodi di Matching supportati:

MatchIt offre una vasta gamma di metodi di matching per soddisfare diverse esigenze:

- **Nearest Neighbor Matching ("nearest"):** Abbina ogni unità trattata alla sua unità di controllo "più vicina" in base a una misura di distanza (spesso il propensity score). Può essere 1:1, K:1, con o senza reimmissione, e con o senza calibro (caliper).
- **Optimal Pair Matching ("optimal"):** Cerca di creare coppie trattato-controllo minimizzando la somma totale delle distanze all'interno delle coppie. Spesso più efficiente del nearest neighbor.
- **Full Matching ("full"):** Crea sottoclassi di unità trattate e di controllo, garantendo che ogni unità trattata e di controllo sia inclusa in una sottoclasse. Questo massimizza l'uso dei dati.
- **Exact Matching ("exact"):** Abbina le unità solo se hanno valori identici su tutte le covariate specificate. Molto forte in termini di bilanciamento, ma può portare a una grande perdita di dati.
- **Genetic Matching ("genetic"):** Utilizza un algoritmo genetico per trovare un set di pesi per le covariate che minimizza lo sbilanciamento complessivo.
- **Propensity Score Subclassification ("subclass"):** Divide i dati in sottoclassi (quintili, ad esempio) in base al propensity score, e poi bilancia le covariate all'interno di ogni sottoclasse.

3. Stima del Punteggio di Propensione (distance):

La maggior parte dei metodi di matching si basa sulla distanza tra le unità, e il punteggio di propensione è una misura di distanza comune. MatchIt consente di stimare il propensity score in vari modi:

- **Regressione logistica ("glm"):** Il metodo più comune, che stima la probabilità di ricevere il trattamento date le covariate.
- **Alberi di regressione ("rpart"):** Utilizza alberi di decisione per stimare i punteggi.
- **Machine learning avanzato ("gbm", "nnet", "bart"):** Offre opzioni per utilizzare metodi più sofisticati come Gradient Boosting Machines, Reti Neurali o Bayesian Additive Regression Trees per stimare i propensity score.
- **Distanza di Mahalanobis ("mahalanobis"):** Non si basa sul propensity score, ma calcola una distanza multivariata direttamente sulle covariate.

4. Valutazione del Bilanciamento:

Dopo aver eseguito il matching, è cruciale valutare quanto bene le covariate siano bilanciate. MatchIt fornisce strumenti per farlo:

- **summary() sull'oggetto matchit:** Restituisce statistiche di bilanciamento (ad esempio, differenze medie standardizzate, ECDF Mean, ECDF Max) sia per i dati originali che per quelli abbinati. Valori più piccoli indicano un migliore bilanciamento.
- **plot() sull'oggetto summary.matchit:** Genera i "Love plots" (grafici delle differenze medie standardizzate), che visualizzano il bilanciamento delle covariate in modo intuitivo. Aiuta a identificare quali covariate potrebbero essere ancora sbilanciate.
- **Integrazione con il pacchetto cobalt:** cobalt è un pacchetto complementare a MatchIt, specificamente progettato per la valutazione e la reportistica del bilanciamento, offrendo grafici e statistiche ancora più dettagliate e personalizzabili.

5. Estrazione dei dati abbinati:

Una volta che il matching è completato e il bilanciamento è soddisfacente, è possibile estrarre il dataset abbinato per l'analisi dell'outcome.

- **match.data() (o match_data()):** Estrae un nuovo dataframe contenente solo le unità abbinate, con l'aggiunta di variabili come i pesi di matching e l'indice delle sottoclassi (se applicabile). Questi pesi sono cruciali per una corretta stima dell'effetto causale.

6. Stima dell'effetto del trattamento:

Dopo aver ottenuto il dataset abbinato, è possibile utilizzare qualsiasi funzione statistica standard in R (ad esempio, lm(), glm()) per stimare l'effetto del trattamento. È importante tenere conto dei pesi di matching e della struttura del matching (ad esempio, le sottoclassi o le coppie) per ottenere stime corrette degli errori standard. Il pacchetto MatchIt fornisce indicazioni su come fare ciò, spesso suggerendo l'uso di modelli di regressione ponderati o robusti.

Esempio di utilizzo in R

```
Carica il pacchetto
library(MatchIt)

# Carica un dataset di esempio (ad esempio, lalonde dal pacchetto MatchIt)
data("lalonde", package = "MatchIt")

# 1. Valutazione iniziale del bilanciamento (prima del matching)
# Non viene eseguito alcun matching, solo la stima del propensity score
```

```
m.out0 <- matchit(treat ~ age + educ + race + married + nodegree + re74 + re75,
                  data = lalonde,
                  method = NULL, # Nessun matching
                  distance = "glm") # Stima del propensity score con regressione logistica
summary(m.out0)
plot(summary(m.out0)) # Love plot

# 2. Esecuzione del matching (es. nearest neighbor 1:1 senza reimmissione)
m.out1 <- matchit(treat ~ age + educ + race + married + nodegree + re74 + re75,
                  data = lalonde,
                  method = "nearest",
                  distance = "glm",
                  replace = FALSE, # Senza reimmissione
                  ratio = 1) # 1:1

# 3. Valutazione del bilanciamento dopo il matching
summary(m.out1)
plot(summary(m.out1)) # Love plot per i dati abbinati

# 4. Estrazione del dataset abbinato
matched_data <- match.data(m.out1)

# 5. Stima dell'effetto del trattamento sul dataset abbinato
# (Esempio con un modello lineare, considerando i pesi)
# Per gli errori standard corretti, spesso si usano metodi robusti o bootstrap
# oppure si considera la struttura delle coppie/sottoclassi.
fit_matched <- lm(re78 ~ treat + age + educ + race + married + nodegree + re74 + re75,
                  data = matched_data,
                  weights = weights)
summary(fit_matched)
```

APPENDICE 2: Lettura dell'output di MatchIT

In generale, la funzione utilizzata è del tipo:

```
matchit_fit_cem_s <- matchit(trattato ~ Gender + Education + MacroArea + Regione + WorkStatus
+ Age,
                             data = combined_small_s,
                             method = "cem",
                             weights = combined_small_s$w)
```

Di seguito analizziamo tutte le componenti:

- **matchit_fit_cem_s**: Questo è il nome dell'oggetto che conterrà i risultati del matching. Sarà un oggetto di classe matchit, ed è utile per estrarre informazioni sul bilanciamento, i pesi e le unità accoppiate.
- **trattato ~ Gender + Education + MacroArea + Regione + WorkStatus + Age**: Questa è la **formula** che definisce il modello di matching.
    - **trattato**: Questa è la tua **variabile di trattamento** (o outcome). matchit la userà per distinguere i soggetti nel gruppo di trattamento da quelli nel gruppo di controllo. Si presume che trattato sia una variabile binaria (es. 0/1, Falso/Vero).
    - **Gender + Education + MacroArea + Regione + WorkStatus + Age**: Queste sono le tue **variabili covariate** (o predittori). Sono le caratteristiche che matchit cercherà di bilanciare tra il gruppo trattato e il gruppo di controllo. Nel contesto del CEM, queste variabili verranno "grossolanamente categorizzate" (coarsened) se continue (come Age) o usate direttamente se già categoriche, e poi il matching avverrà sulle combinazioni esatte di queste categorie.
- **data = combined_small_s**: Questo specifica il **dataframe** da cui matchit deve prendere tutte le variabili (trattamento e covariate).
- **method = "cem"**: Questo è l'argomento chiave che indica l'utilizzo del **Coarsened Exact Matching**. Il CEM differisce da altri metodi (come il Propensity Score Matching) perché non si basa su uno score, ma discretizza le covariate e poi trova corrispondenze esatte all'interno di questi strati discretizzati. Questo garantisce un bilanciamento teoricamente perfetto all'interno di ogni strato per le variabili coarsened.
- **weights = combined_small_s$w**. Questo vettore specifica che nel modello sono presi in considerazione i pesi campionari e verranno incorporati nel processo di bilanciamento, ad esempio, ponderando le osservazioni quando calcola le statistiche di bilanciamento o quando determina i bin per il CEM. L'obiettivo è garantire che il bilanciamento avvenga su una popolazione che rifletta il disegno campionario originale.

Dopo aver eseguito questa riga di codice, l'oggetto matchit_fit_cem_s conterrà:

- **Informazioni sulle unità abbinate**: Quante unità di trattamento e controllo sono state mantenute dopo il CEM.
- **Strati (Subclasses)**: L'assegnazione di ogni osservazione a uno specifico strato CEM, basato sui valori "coarsened" dei covariati.
- **Pesi del matching**: Vettori di pesi che, se usati in una successiva analisi (es. regressione), bilanceranno le covariate nei gruppi di trattamento e controllo, tenendo conto anche dei tuoi weights di input.
- **Informazioni sul bilanciamento**: Utilizzando summary(matchit_fit_cem_s), potrai vedere le metriche di bilanciamento (es. Standardized Mean Difference) per ogni covariata prima e dopo il matching. Ti aspetteresti un bilanciamento molto buono (SMD vicini a zero) dopo il CEM.

Una volta eseguito, l'output sarà suddiviso in diverse sezioni:

1. Sezione "Summary of Balance for All Data":

Questa sezione mostra il **bilanciamento dei covariati *prima* del matching**.

- **Means Treated e Means Control**: Medie di ciascuna covariata per il gruppo trattato e il gruppo di controllo.
- **Std. Mean Diff (Standardized Mean Difference - SMD)**: Indica la differenza media tra i gruppi per ogni covariata, standardizzata. La presenza di valori *alti* (es. > 0.1 o > 0.2), indicando uno squilibrio iniziale.
- **Var. Ratio (Variance Ratio)**: Rapporto tra le varianze delle covariate nei due gruppi.
- **eQQ Mean e eQQ Max**: Misure basate sui quantili empirici, che indicano differenze nella forma delle distribuzioni, non solo nelle medie. Valori elevati indicano differenze significative nelle distribuzioni prima del matching.

2. Sezione "Summary of Balance for Matched Data":

Questa valuta l'efficacia del CEM, mostrando il bilanciamento delle covariate *dopo* il matching.

- **Means Treated e Means Control**: Medie delle covariate nei gruppi trattato e di controllo *dopo il matching*.
- **Std. Mean Diff (Standardized Mean Difference - SMD)**: Per il CEM, ci si aspetta che questi valori siano **estremamente vicini a zero, idealmente inferiori a 0.05 per tutte le covariate.** CEM è un metodo molto stringente, e se il bilanciamento non è quasi perfetto, significa che ci sono ancora differenze significative tra i gruppi.
- **Var. Ratio (Variance Ratio)**: Ci si aspetta che questi valori siano **molto vicini a 1, idealmente tra 0.8 e 1.25.** Se sono distanti da 1, significa che le varianze delle covariate sono ancora diverse tra i gruppi accoppiati.

- **eQQ Mean e eQQ Max**: Questi valori dovrebbero essere **molto piccoli e vicini a zero** dopo il matching, indicando che non solo le medie, ma anche le distribuzioni complete dei covariati sono simili tra i gruppi.

3. Sezione "Sample Sizes":

Questa tabella riassume il numero di unità in ogni gruppo prima e dopo il matching.

- **Original**: Il numero di unità nel gruppo trattato e di controllo nel dataset originale.
- **Matched**: Il numero di unità nel gruppo trattato e di controllo che sono state **mantenute** (cioè, hanno trovato una corrispondenza in uno strato CEM) e che verranno utilizzate nell'analisi successiva.
- **Cosa guardare**:
    - **Perdita di osservazioni**: Quante unità del gruppo trattato e di controllo sono state scartate? Un CEM molto aggressivo (con molti covariati o cutpoint molto fini) può portare a una perdita significativa di osservazioni. Se si perdono troppe unità, la potenza statistica delle analisi post-matching potrebbe essere ridotta e la generalizzabilità dei risultati limitata alla sottopopolazione che è stata effettivamente accoppiata. Non c'è una "percentuale ideale" di perdita, ma è un trade-off tra bilanciamento e generalizzabilità.

4. Sezione "L1 Imbalance":

Questa metrica offre una misura globale dello squilibrio congiunto di tutti i covariati.

- **Before Matching**: Il valore L1 prima del matching.
- **After Matching**: Il valore L1 dopo il matching.
- **Cosa guardare**: Il valore di **L1 dopo il matching dovrebbe essere molto più basso** rispetto a prima, idealmente tendente a zero. CEM è specificamente progettato per minimizzare questa statistica.

Criteri di Bontà riassuntivi per CEM:

1. **SMD Post-Matching molto bassi**: Per tutte le covariate, gli Std. Mean Diff nella sezione "Matched Data" devono essere *vicinissimi allo zero* (es. < 0.05, idealmente < 0.01). Questo è il segnale più forte di un buon bilanciamento.
2. **Var. Ratio vicini a 1**: Sempre nella sezione "Matched Data", i rapporti di varianza devono essere vicini a 1 (es. 0.8 - 1.25).
3. **eQQ e L1 molto bassi Post-Matching**: Confermano il bilanciamento delle distribuzioni.
4. **Numero di Unità Matched sufficiente**: Assicurati di non aver perso un numero eccessivo di osservazioni, compromettendo la potenza statistica o la generalizzabilità.

Se questi punti non sono soddisfatti, potrebbe essere necessario:

- riconsiderare le tue variabili usate per agganciare i dati
- effettuare un nuovo matching sui dati non abbinati ridefinendo nuovi cutpoint

APPENDICE 3: Output del Matching tra EHIS e AVQ

```
matchit_fit_cem <- matchit(trattato ~ Gender + Education + MacroArea + Regione + WorkStatus +
Age,
                           data = combined_small,
                           method = "cem",
                           weights = combined_small$w)

summary(matchit_fit_cem)

##
## Call:
## matchit(formula = trattato ~ Gender + Education + MacroArea +
##     Regione + WorkStatus + Age, data = combined_small_s, method = "cem",
##     weights = combined_small_s$w)
##
## Summary of Balance for All Data:
##            Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## Gender0           0.5595        0.5556          0.0079          .    0.0039
## Gender1           0.4405        0.4444         -0.0079          .    0.0039
## Education0        0.0733        0.0860         -0.0490          .    0.0128
## Education1        0.4023        0.4025         -0.0004          .    0.0002
## Education2        0.2448        0.2511         -0.0147          .    0.0063
## Education3        0.1993        0.1854          0.0348          .    0.0139
## Education4        0.0804        0.0750          0.0198          .    0.0054
## MacroArea1        0.2378        0.2292          0.0202          .    0.0086
## MacroArea2        0.2122        0.2098          0.0058          .    0.0024
## MacroArea3        0.1988        0.1904          0.0209          .    0.0083
## MacroArea4        0.3512        0.3706         -0.0405          .    0.0193
## Regione01         0.0746        0.0792         -0.0176          .    0.0046
## Regione02         0.0227        0.0219          0.0050          .    0.0007
## Regione03         0.0933        0.0821          0.0384          .    0.0112
## Regione04         0.0517        0.0597         -0.0357          .    0.0079
## Regione05         0.0585        0.0580          0.0021          .    0.0005
## Regione06         0.0411        0.0401          0.0049          .    0.0010
## Regione07         0.0473        0.0460          0.0062          .    0.0013
## Regione08         0.0609        0.0520          0.0370          .    0.0088
## Regione09         0.0716        0.0635          0.0317          .    0.0082
## Regione10         0.0249        0.0293         -0.0285          .    0.0044
## Regione11         0.0405        0.0447         -0.0214          .    0.0042
## Regione12         0.0617        0.0529          0.0367          .    0.0088
```

```
## Regione13       0.0371      0.0413      -0.0221       .  0.0042
## Regione14       0.0231      0.0301      -0.0465       .  0.0070
## Regione15       0.0550      0.0653      -0.0452       .  0.0103
## Regione16       0.0553      0.0579      -0.0112       .  0.0026
## Regione17       0.0299      0.0324      -0.0148       .  0.0025
## Regione18       0.0466      0.0423       0.0204       .  0.0043
## Regione19       0.0630      0.0574       0.0234       .  0.0057
## Regione20       0.0412      0.0439      -0.0139       .  0.0028
## WorkStatus1     0.0441      0.0448      -0.0036       .  0.0007
## WorkStatus2     0.0032      0.0101      -0.1223       .  0.0069
## WorkStatus3     0.7159      0.7053       0.0236       .  0.0106
## WorkStatus5     0.1970      0.2056      -0.0216       .  0.0086
## WorkStatus6     0.0398      0.0342       0.0287       .  0.0056
## Age1            0.4915      0.4981      -0.0133       .  0.0067
## Age2            0.5085      0.5019       0.0133       .  0.0067
##           eCDF Max
## Gender0     0.0039
## Gender1     0.0039
## Education0  0.0128
## Education1  0.0002
## Education2  0.0063
## Education3  0.0139
## Education4  0.0054
## MacroArea1  0.0086
## MacroArea2  0.0024
## MacroArea3  0.0083
## MacroArea4  0.0193
## Regione01   0.0046
## Regione02   0.0007
## Regione03   0.0112
## Regione04   0.0079
## Regione05   0.0005
## Regione06   0.0010
## Regione07   0.0013
## Regione08   0.0088
## Regione09   0.0082
## Regione10   0.0044
## Regione11   0.0042
## Regione12   0.0088
## Regione13   0.0042
## Regione14   0.0070
```

```
## Regione15     0.0103
## Regione16     0.0026
## Regione17     0.0025
## Regione18     0.0043
## Regione19     0.0057
## Regione20     0.0028
## WorkStatus1   0.0007
## WorkStatus2   0.0069
## WorkStatus3   0.0106
## WorkStatus5   0.0086
## WorkStatus6   0.0056
## Age1          0.0067
## Age2          0.0067
##
## Summary of Balance for Matched Data:
##           Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## Gender0          0.5603        0.5603               0          .         0
## Gender1          0.4397        0.4397               0          .         0
## Education0       0.0706        0.0706               0          .         0
## Education1       0.4054        0.4054              -0          .         0
## Education2       0.2455        0.2455               0          .         0
## Education3       0.1991        0.1991               0          .         0
## Education4       0.0794        0.0794               0          .         0
## MacroArea1       0.2375        0.2375               0          .         0
## MacroArea2       0.2119        0.2119               0          .         0
## MacroArea3       0.1989        0.1989               0          .         0
## MacroArea4       0.3517        0.3517               0          .         0
## Regione01        0.0752        0.0752              -0          .         0
## Regione02        0.0218        0.0218               0          .         0
## Regione03        0.0936        0.0936               0          .         0
## Regione04        0.0517        0.0517               0          .         0
## Regione05        0.0593        0.0593               0          .         0
## Regione06        0.0403        0.0403               0          .         0
## Regione07        0.0469        0.0469               0          .         0
## Regione08        0.0605        0.0605               0          .         0
## Regione09        0.0720        0.0720              -0          .         0
## Regione10        0.0247        0.0247              -0          .         0
## Regione11        0.0408        0.0408               0          .         0
## Regione12        0.0614        0.0614              -0          .         0
## Regione13        0.0375        0.0375               0          .         0
## Regione14        0.0231        0.0231               0          .         0
```

```
## Regione15       0.0546       0.0546            0       .       0
## Regione16       0.0558       0.0558            0       .       0
## Regione17       0.0301       0.0301           -0       .       0
## Regione18       0.0468       0.0468            0       .       0
## Regione19       0.0631       0.0631            0       .       0
## Regione20       0.0408       0.0408            0       .       0
## WorkStatus1     0.0406       0.0406            0       .       0
## WorkStatus2     0.0020       0.0020            0       .       0
## WorkStatus3     0.7300       0.7300            0       .       0
## WorkStatus5     0.1995       0.1995            0       .       0
## WorkStatus6     0.0279       0.0279           -0       .       0
## Age1            0.4900       0.4900            0       .       0
## Age2            0.5100       0.5100            0       .       0
##           eCDF Max Std. Pair Dist.
## Gender0         0         0
## Gender1         0         0
## Education0      0         0
## Education1      0         0
## Education2      0         0
## Education3      0         0
## Education4      0         0
## MacroArea1      0         0
## MacroArea2      0         0
## MacroArea3      0         0
## MacroArea4      0         0
## Regione01       0         0
## Regione02       0         0
## Regione03       0         0
## Regione04       0         0
## Regione05       0         0
## Regione06       0         0
## Regione07       0         0
## Regione08       0         0
## Regione09       0         0
## Regione10       0         0
## Regione11       0         0
## Regione12       0         0
## Regione13       0         0
## Regione14       0         0
## Regione15       0         0
## Regione16       0         0
```

```
## Regione17        0           0
## Regione18        0           0
## Regione19        0           0
## Regione20        0           0
## WorkStatus1      0           0
## WorkStatus2      0           0
## WorkStatus3      0           0
## WorkStatus5      0           0
## WorkStatus6      0           0
## Age1             0           0
## Age2             0           0
##
## Sample Sizes:
##                Control Treated
## All            11264.    13720
## Matched (ESS)  9020.39   13433
## Matched        11056.    13433
## Unmatched        208.      287
## Discarded          0.        0
```

```
matched_data_cem <- match.data(matchit_fit_cem)
```

```
tabella_bilanciamento_matched_data_cem  <- bal.tab(trattato ~ Gender + Education + MacroArea
+ Regione + WorkStatus + Age,
                                    data = matched_data_cem )
tabella_bilanciamento_matched_data_cem
```

```
## Balance Measures
##               Type Diff.Un
## Gender      Binary -0.0051
## Education_0  Binary -0.0134
## Education_1  Binary -0.0007
## Education_2  Binary -0.0060
## Education_3  Binary  0.0126
## Education_4  Binary  0.0075
## MacroArea_1  Binary  0.0060
## MacroArea_2  Binary  0.0027
## MacroArea_3  Binary  0.0092
## MacroArea_4  Binary -0.0179
## Regione_01   Binary -0.0045
```

```
## Regione_02   Binary -0.0002
## Regione_03   Binary  0.0105
## Regione_04   Binary -0.0078
## Regione_05   Binary  0.0014
## Regione_06   Binary  0.0003
## Regione_07   Binary  0.0002
## Regione_08   Binary  0.0089
## Regione_09   Binary  0.0079
## Regione_10   Binary -0.0040
## Regione_11   Binary -0.0033
## Regione_12   Binary  0.0087
## Regione_13   Binary -0.0032
## Regione_14   Binary -0.0061
## Regione_15   Binary -0.0108
## Regione_16   Binary -0.0026
## Regione_17   Binary -0.0019
## Regione_18   Binary  0.0043
## Regione_19   Binary  0.0053
## Regione_20   Binary -0.0029
## WorkStatus_1 Binary -0.0014
## WorkStatus_2 Binary -0.0022
## WorkStatus_3 Binary  0.0127
## WorkStatus_5 Binary -0.0074
## WorkStatus_6 Binary -0.0018
## Age_2        Binary  0.0055
##
## Sample sizes
##     Control Treated
## All   11056   13433
```

APPENDICE 4: Output del Matching tra EHIS_AVQ e SHARE

ROUND 1

```
matchit_fit_cem <- matchit(trattato ~ Gender + Education + MacroArea + Income + WorkStatus +
Age,
                  data = combined,
                  method = "cem",
                  weights = combined$w)

summary(matchit_fit_cem)
```

```
##
## Call:
## matchit(formula = trattato ~ Gender + Education + MacroArea +
##     Income + WorkStatus + Age, data = combined, method = "cem",
##     weights = combined$w)
##
## Summary of Balance for All Data:
##           Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## Gender0          0.5603        0.5421          0.0366          .    0.0182
## Gender1          0.4397        0.4579         -0.0366          .    0.0182
## Education0       0.0706        0.0589          0.0454          .    0.0116
## Education1       0.4054        0.4573         -0.1056          .    0.0518
## Education2       0.2455        0.2254          0.0468          .    0.0201
## Education3       0.1991        0.1665          0.0817          .    0.0326
## Education4       0.0794        0.0920         -0.0464          .    0.0125
## MacroArea1       0.2375        0.1645          0.1715          .    0.0730
## MacroArea2       0.2119        0.1108          0.2474          .    0.1011
## MacroArea3       0.1989        0.2435         -0.1118          .    0.0446
## MacroArea4       0.3517        0.4812         -0.2711          .    0.1295
## Income1          0.1257        0.2494         -0.3732          .    0.1237
## Income2          0.2108        0.2085          0.0056          .    0.0023
## Income3          0.2167        0.2034          0.0324          .    0.0133
## Income4          0.2224        0.1949          0.0661          .    0.0275
## Income5          0.2244        0.1438          0.1932          .    0.0806
## WorkStatus1      0.0406        0.0285          0.0612          .    0.0121
## WorkStatus2      0.0020        0.0052         -0.0708          .    0.0032
## WorkStatus3      0.7300        0.7254          0.0104          .    0.0046
## WorkStatus5      0.1995        0.1885          0.0276          .    0.0110
## WorkStatus6      0.0279        0.0525         -0.1490          .    0.0245
```

```
## Age1            0.4900       0.4670          0.0460         .        0.0230
## Age2            0.5100       0.5330         -0.0460         .        0.0230
##           eCDF Max
## Gender0     0.0182
## Gender1     0.0182
## Education0  0.0116
## Education1  0.0518
## Education2  0.0201
## Education3  0.0326
## Education4  0.0125
## MacroArea1  0.0730
## MacroArea2  0.1011
## MacroArea3  0.0446
## MacroArea4  0.1295
## Income1     0.1237
## Income2     0.0023
## Income3     0.0133
## Income4     0.0275
## Income5     0.0806
## WorkStatus1 0.0121
## WorkStatus2 0.0032
## WorkStatus3 0.0046
## WorkStatus5 0.0110
## WorkStatus6 0.0245
## Age1        0.0230
## Age2        0.0230
##
## Summary of Balance for Matched Data:
##           Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## Gender0          0.5287        0.5287               0         .          0
## Gender1          0.4713        0.4713               0         .          0
## Education0       0.0492        0.0492               0         .          0
## Education1       0.4326        0.4326               0         .          0
## Education2       0.2556        0.2556               0         .          0
## Education3       0.1924        0.1924               0         .          0
## Education4       0.0702        0.0702               0         .          0
## MacroArea1       0.2206        0.2206              -0         .          0
## MacroArea2       0.1901        0.1901               0         .          0
## MacroArea3       0.2100        0.2100               0         .          0
## MacroArea4       0.3793        0.3793               0         .          0
## Income1          0.1308        0.1308               0         .          0
```

```
## Income2         0.2221       0.2221          0      .      0
## Income3         0.2165       0.2165         -0      .      0
## Income4         0.2262       0.2262          0      .      0
## Income5         0.2044       0.2044          0      .      0
## WorkStatus1     0.0146       0.0146          0      .      0
## WorkStatus2     0.0008       0.0008          0      .      0
## WorkStatus3     0.7945       0.7945          0      .      0
## WorkStatus5     0.1727       0.1727          0      .      0
## WorkStatus6     0.0175       0.0175          0      .      0
## Age1            0.4850       0.4850          0      .      0
## Age2            0.5150       0.5150          0      .      0
##           eCDF Max Std. Pair Dist.
## Gender0          0           0
## Gender1          0           0
## Education0       0           0
## Education1       0           0
## Education2       0           0
## Education3       0           0
## Education4       0           0
## MacroArea1       0           0
## MacroArea2       0           0
## MacroArea3       0           0
## MacroArea4       0           0
## Income1          0           0
## Income2          0           0
## Income3          0           0
## Income4          0           0
## Income5          0           0
## WorkStatus1      0           0
## WorkStatus2      0           0
## WorkStatus3      0           0
## WorkStatus5      0           0
## WorkStatus6      0           0
## Age1             0           0
## Age2             0           0
##
## Sample Sizes:
##             Control Treated
## All          1544.    13433
## Matched (ESS)  904.55   11039
## Matched       1502.    11039
```

```
## Unmatched      42.     2394
## Discarded       0.        0

matched_data_cem <- match.data(matchit_fit_cem)


tabella_bilanciamento_matched_data_cem  <- bal.tab(trattato ~ Gender + Education + MacroArea
+ Income + WorkStatus + Age,
                                    data = matched_data_cem )
tabella_bilanciamento_matched_data_cem

## Balance Measures
##               Type Diff.Un
## Gender      Binary  0.0119
## Education_0  Binary -0.0047
## Education_1  Binary -0.0308
## Education_2  Binary  0.0286
## Education_3  Binary  0.0246
## Education_4  Binary -0.0177
## MacroArea_1  Binary  0.0595
## MacroArea_2  Binary  0.0776
## MacroArea_3  Binary -0.0330
## MacroArea_4  Binary -0.1041
## Income_1     Binary -0.1162
## Income_2     Binary  0.0137
## Income_3     Binary  0.0094
## Income_4     Binary  0.0305
## Income_5     Binary  0.0626
## WorkStatus_1 Binary -0.0081
## WorkStatus_2 Binary -0.0005
## WorkStatus_3 Binary  0.0514
## WorkStatus_5 Binary -0.0191
## WorkStatus_6 Binary -0.0238
## Age_2        Binary -0.0270
##
## Sample sizes
##     Control Treated
## All    1502   11039
```

ROUND 2

```
matchit_fit_cem_step2 <- matchit(trattato ~ Gender + Education_new + MacroArea + Income_ne
w + WorkStatus_new + Age,
                 data = combined_step2_pulito,
                 method = "cem",
                 weights = combined_step2_pulito$w)

summary(matchit_fit_cem_step2)

##
## Call:
## matchit(formula = trattato ~ Gender + Education_new + MacroArea +
##     Income_new + WorkStatus_new + Age, data = combined_step2_pulito,
##     method = "cem", weights = combined_step2_pulito$w)
##
## Summary of Balance for All Data:
##                       Means Treated Means Control Std. Mean Diff. Var. Ratio
## Gender0                      0.7059       0.5421          0.3596          .
## Gender1                      0.2941       0.4579         -0.3596          .
## Education_newHigh            0.1220       0.0920          0.0917          .
## Education_newLow             0.4495       0.5162         -0.1342          .
## Education_newMedium          0.4286       0.3918          0.0742          .
## MacroArea1                   0.3154       0.1645          0.3247          .
## MacroArea2                   0.3120       0.1108          0.4344          .
## MacroArea3                   0.1479       0.2435         -0.2695          .
## MacroArea4                   0.2247       0.4812         -0.6145          .
## Income_newHigh               0.5217       0.3387          0.3663          .
## Income_newLow                0.2607       0.4579         -0.4493          .
## Income_newMedium             0.2176       0.2034          0.0346          .
## WorkStatus_newOther          0.0835       0.0576          0.0936          .
## WorkStatus_newRetired        0.4327       0.7254         -0.5906          .
## WorkStatus_newWorking        0.4837       0.2170          0.5338          .
## Age1                         0.5129       0.4670          0.0920          .
## Age2                         0.4871       0.5330         -0.0920          .
##                       eCDF Mean eCDF Max
## Gender0                  0.1638   0.1638
## Gender1                  0.1638   0.1638
## Education_newHigh        0.0300   0.0300
## Education_newLow         0.0667   0.0667
## Education_newMedium      0.0367   0.0367
```

```
## MacroArea1            0.1509   0.1509
## MacroArea2            0.2013   0.2013
## MacroArea3            0.0957   0.0957
## MacroArea4            0.2565   0.2565
## Income_newHigh        0.1830   0.1830
## Income_newLow         0.1972   0.1972
## Income_newMedium      0.0143   0.0143
## WorkStatus_newOther   0.0259   0.0259
## WorkStatus_newRetired 0.2926   0.2926
## WorkStatus_newWorking 0.2667   0.2667
## Age1                  0.0460   0.0460
## Age2                  0.0460   0.0460
##
## Summary of Balance for Matched Data:
##                       Means Treated Means Control Std. Mean Diff. Var. Ratio
## Gender0                      0.7310       0.7310               0        .
## Gender1                      0.2690       0.2690               0        .
## Education_newHigh            0.0891       0.0891              -0        .
## Education_newLow             0.4704       0.4704              -0        .
## Education_newMedium          0.4405       0.4405               0        .
## MacroArea1                   0.2612       0.2612               0        .
## MacroArea2                   0.2785       0.2785               0        .
## MacroArea3                   0.1698       0.1698               0        .
## MacroArea4                   0.2905       0.2905               0        .
## Income_newHigh               0.5774       0.5774               0        .
## Income_newLow                0.2887       0.2887               0        .
## Income_newMedium             0.1339       0.1339               0        .
## WorkStatus_newOther          0.0735       0.0735               0        .
## WorkStatus_newRetired        0.5338       0.5338               0        .
## WorkStatus_newWorking        0.3927       0.3927               0        .
## Age1                         0.4806       0.4806               0        .
## Age2                         0.5194       0.5194               0        .
##                       eCDF Mean eCDF Max Std. Pair Dist.
## Gender0                       0        0               0
## Gender1                       0        0               0
## Education_newHigh             0        0               0
## Education_newLow              0        0               0
## Education_newMedium           0        0               0
## MacroArea1                    0        0               0
## MacroArea2                    0        0               0
## MacroArea3                    0        0               0
```

```
## MacroArea4                0      0          0
## Income_newHigh            0      0          0
## Income_newLow             0      0          0
## Income_newMedium          0      0          0
## WorkStatus_newOther       0      0          0
## WorkStatus_newRetired     0      0          0
## WorkStatus_newWorking     0      0          0
## Age1                      0      0          0
## Age2                      0      0          0
##
## Sample Sizes:
##              Control Treated
## All          1544.     2394
## Matched (ESS)  130.86    1673
## Matched       631.     1673
## Unmatched     913.      721
## Discarded       0.        0
```

```
tabella_bilanciamento_matched_data_cem_step2  <- bal.tab(trattato ~ Gender + Education_new
+ MacroArea + Income_new + WorkStatus_new + Age,
                                    data = matched_data_cem_step2 )
tabella_bilanciamento_matched_data_cem_step2
```

```
## Balance Measures
##                      Type Diff.Un
## Gender               Binary -0.0448
## Education_new_High   Binary  0.0368
## Education_new_Low    Binary -0.1873
## Education_new_Medium Binary  0.1505
## MacroArea_1          Binary  0.0774
## MacroArea_2          Binary  0.1280
## MacroArea_3          Binary -0.1710
## MacroArea_4          Binary -0.0344
## Income_new_High      Binary  0.2002
## Income_new_Low       Binary -0.1851
## Income_new_Medium    Binary -0.0151
## WorkStatus_new_Other   Binary -0.0358
## WorkStatus_new_Retired Binary -0.0637
## WorkStatus_new_Working Binary  0.0995
## Age_2                Binary -0.0352
```

```
##
## Sample sizes
##     Control Treated
## All     631    1673
```

APPENDICE 5: Output del Matching tra EHIS_AVQ_SHARE e EUSILC

```
matchit_fit_cem <- matchit(trattato ~ Gender + Education + MacroArea + Income + WorkStatus2 +
Age,
                           data = combined,
                           method = "cem",
                           weights = combined$w)

summary(matchit_fit_cem)

##
## Call:
## matchit(formula = trattato ~ Gender + Education + MacroArea +
##     Income + WorkStatus2 + Age, data = combined, method = "cem",
##     weights = combined$w)
##
## Summary of Balance for All Data:
##              Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## Gender0             0.5553        0.5649         -0.0193          .    0.0096
## Gender1             0.4447        0.4351          0.0193          .    0.0096
## Education0          0.0730        0.0648          0.0315          .    0.0082
## Education1          0.4073        0.3611          0.0938          .    0.0461
## Education2          0.2477        0.2524         -0.0108          .    0.0047
## Education3          0.1993        0.2379         -0.0966          .    0.0386
## Education4          0.0727        0.0837         -0.0425          .    0.0110
## MacroArea1          0.2259        0.2526         -0.0638          .    0.0267
## MacroArea2          0.2018        0.2297         -0.0696          .    0.0279
## MacroArea3          0.2047        0.2526         -0.1187          .    0.0479
## MacroArea4          0.3676        0.2651          0.2126          .    0.1025
## Income1             0.1295        0.2000         -0.2101          .    0.0705
## Income2             0.2150        0.1999          0.0366          .    0.0151
## Income3             0.2056        0.2000          0.0139          .    0.0056
## Income4             0.2243        0.2000          0.0582          .    0.0243
## Income5             0.2256        0.2000          0.0612          .    0.0256
## WorkStatus21        0.0275        0.0485         -0.1290          .    0.0211
## WorkStatus22        0.0017        0.0044         -0.0647          .    0.0027
## WorkStatus23        0.7601        0.7917         -0.0738          .    0.0315
## WorkStatus24        0.2107        0.1554          0.1356          .    0.0553
## Age1                0.4844        0.4704          0.0281          .    0.0141
## Age2                0.5156        0.5296         -0.0281          .    0.0141
##              eCDF Max
```

```
## Gender0       0.0096
## Gender1       0.0096
## Education0    0.0082
## Education1    0.0461
## Education2    0.0047
## Education3    0.0386
## Education4    0.0110
## MacroArea1    0.0267
## MacroArea2    0.0279
## MacroArea3    0.0479
## MacroArea4    0.1025
## Income1       0.0705
## Income2       0.0151
## Income3       0.0056
## Income4       0.0243
## Income5       0.0256
## WorkStatus21  0.0211
## WorkStatus22  0.0027
## WorkStatus23  0.0315
## WorkStatus24  0.0553
## Age1          0.0141
## Age2          0.0141
##
## Summary of Balance for Matched Data:
##           Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## Gender0          0.5561       0.5561           0       .      0
## Gender1          0.4439       0.4439          -0       .      0
## Education0       0.0713       0.0713           0       .      0
## Education1       0.4081       0.4081          -0       .      0
## Education2       0.2481       0.2481          -0       .      0
## Education3       0.1997       0.1997           0       .      0
## Education4       0.0728       0.0728           0       .      0
## MacroArea1       0.2261       0.2261          -0       .      0
## MacroArea2       0.2020       0.2020           0       .      0
## MacroArea3       0.2053       0.2053          -0       .      0
## MacroArea4       0.3666       0.3666           0       .      0
## Income1          0.1294       0.1294           0       .      0
## Income2          0.2146       0.2146           0       .      0
## Income3          0.2063       0.2063           0       .      0
## Income4          0.2237       0.2237          -0       .      0
## Income5          0.2260       0.2260          -0       .      0
```

```
## WorkStatus21      0.0263      0.0263          0      .      0
## WorkStatus22      0.0010      0.0010          0      .      0
## WorkStatus23      0.7631      0.7631          0      .      0
## WorkStatus24      0.2096      0.2096          0      .      0
## Age1              0.4827      0.4827          0      .      0
## Age2              0.5173      0.5173          0      .      0
##           eCDF Max Std. Pair Dist.
## Gender0          0          0
## Gender1          0          0
## Education0       0          0
## Education1       0          0
## Education2       0          0
## Education3       0          0
## Education4       0          0
## MacroArea1       0          0
## MacroArea2       0          0
## MacroArea3       0          0
## MacroArea4       0          0
## Income1          0          0
## Income2          0          0
## Income3          0          0
## Income4          0          0
## Income5          0          0
## WorkStatus21     0          0
## WorkStatus22     0          0
## WorkStatus23     0          0
## WorkStatus24     0          0
## Age1             0          0
## Age2             0          0
##
## Sample Sizes:
##             Control Treated
## All          13579.    12712
## Matched (ESS)  9263.64   12644
## Matched      12927.    12644
## Unmatched      652.       68
## Discarded        0.        0

matched_data_cem <- match.data(matchit_fit_cem)

tabella_bilanciamento_matched_data_cem  <- bal.tab(trattato ~ Gender + Education + MacroArea
```

```
+ Income + WorkStatus2 + Age,
                                  data = matched_data_cem )
tabella_bilanciamento_matched_data_cem

## Balance Measures
##               Type Diff.Un
## Gender       Binary  0.0106
## Education_0  Binary  0.0059
## Education_1  Binary  0.0417
## Education_2  Binary -0.0064
## Education_3  Binary -0.0378
## Education_4  Binary -0.0034
## MacroArea_1  Binary -0.0221
## MacroArea_2  Binary -0.0191
## MacroArea_3  Binary -0.0531
## MacroArea_4  Binary  0.0943
## Income_1     Binary -0.0704
## Income_2     Binary  0.0118
## Income_3     Binary  0.0108
## Income_4     Binary  0.0205
## Income_5     Binary  0.0273
## WorkStatus2_1 Binary -0.0062
## WorkStatus2_2 Binary  0.0003
## WorkStatus2_3 Binary -0.0569
## WorkStatus2_4 Binary  0.0628
## Age_2        Binary -0.0252
##
## Sample sizes
##     Control Treated
## All   12927   12644
```

APPENDICE 6: Output del Matching tra EHIS_AVQ_SHARE_EUSILC e SPESE

```
matchit_fit_cem <- matchit(trattato ~ Gender + Education2 + Regione + MacroArea + Income + WorkStatus2 + Age,
                           data = combined,
                           method = "cem",
                           weights = combined$w)

summary(matchit_fit_cem)

##
## Call:
## matchit(formula = trattato ~ Gender + Education2 + Regione +
##     MacroArea + Income + WorkStatus2 + Age, data = combined,
##     method = "cem", weights = combined$w)
##
## Summary of Balance for All Data:
##              Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## Gender0             0.5561        0.5509          0.0103          .    0.0051
## Gender1             0.4439        0.4491         -0.0103          .    0.0051
## Education21         0.4794        0.4532          0.0524          .    0.0262
## Education22         0.2481        0.2502         -0.0049          .    0.0021
## Education23         0.2725        0.2966         -0.0540          .    0.0240
## Regione01           0.0742        0.0558          0.0701          .    0.0184
## Regione02           0.0215        0.0254         -0.0270          .    0.0039
## Regione03           0.0857        0.1161         -0.1085          .    0.0304
## Regione04           0.0493        0.0554         -0.0285          .    0.0062
## Regione05           0.0548        0.0621         -0.0318          .    0.0072
## Regione06           0.0397        0.0375          0.0115          .    0.0023
## Regione07           0.0447        0.0438          0.0044          .    0.0009
## Regione08           0.0582        0.0540          0.0181          .    0.0042
## Regione09           0.0745        0.0579          0.0631          .    0.0166
## Regione10           0.0259        0.0291         -0.0199          .    0.0032
## Regione11           0.0426        0.0330          0.0479          .    0.0097
## Regione12           0.0622        0.0846         -0.0927          .    0.0224
## Regione13           0.0394        0.0245          0.0765          .    0.0149
## Regione14           0.0244        0.0325         -0.0528          .    0.0081
## Regione15           0.0563        0.0618         -0.0237          .    0.0055
## Regione16           0.0577        0.0640         -0.0271          .    0.0063
## Regione17           0.0317        0.0322         -0.0029          .    0.0005
```

```
## Regione18       0.0485     0.0366      0.0552      .  0.0119
## Regione19       0.0660     0.0598      0.0250      .  0.0062
## Regione20       0.0427     0.0340      0.0432      .  0.0087
## MacroArea1      0.2261     0.2411     -0.0359      .  0.0150
## MacroArea2      0.2020     0.2089     -0.0173      .  0.0069
## MacroArea3      0.2053     0.2046      0.0017      .  0.0007
## MacroArea4      0.3666     0.3453      0.0441      .  0.0212
## Income1         0.1294     0.1622     -0.0978      .  0.0328
## Income2         0.2146     0.2249     -0.0250      .  0.0102
## Income3         0.2063     0.2221     -0.0392      .  0.0159
## Income4         0.2237     0.1883      0.0849      .  0.0354
## Income5         0.2260     0.2025      0.0563      .  0.0235
## WorkStatus21    0.0263     0.0520     -0.1607      .  0.0257
## WorkStatus22    0.0010     0.0016     -0.0166      .  0.0005
## WorkStatus23    0.7631     0.7372      0.0610      .  0.0259
## WorkStatus24    0.2096     0.2093      0.0007      .  0.0003
## Age1            0.4827     0.5288     -0.0923      .  0.0461
## Age2            0.5173     0.4712      0.0923      .  0.0461
##           eCDF Max
## Gender0     0.0051
## Gender1     0.0051
## Education21  0.0262
## Education22  0.0021
## Education23  0.0240
## Regione01    0.0184
## Regione02    0.0039
## Regione03    0.0304
## Regione04    0.0062
## Regione05    0.0072
## Regione06    0.0023
## Regione07    0.0009
## Regione08    0.0042
## Regione09    0.0166
## Regione10    0.0032
## Regione11    0.0097
## Regione12    0.0224
## Regione13    0.0149
## Regione14    0.0081
## Regione15    0.0055
## Regione16    0.0063
## Regione17    0.0005
```

```
## Regione18     0.0119
## Regione19     0.0062
## Regione20     0.0087
## MacroArea1    0.0150
## MacroArea2    0.0069
## MacroArea3    0.0007
## MacroArea4    0.0212
## Income1       0.0328
## Income2       0.0102
## Income3       0.0159
## Income4       0.0354
## Income5       0.0235
## WorkStatus21  0.0257
## WorkStatus22  0.0005
## WorkStatus23  0.0259
## WorkStatus24  0.0003
## Age1          0.0461
## Age2          0.0461
##
## Summary of Balance for Matched Data:
##              Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## Gender0             0.5518        0.5518               0          .         0
## Gender1             0.4482        0.4482               0          .         0
## Education21         0.4864        0.4864               0          .         0
## Education22         0.2430        0.2430               0          .         0
## Education23         0.2706        0.2706               0          .         0
## Regione01           0.0771        0.0771               0          .         0
## Regione02           0.0214        0.0214               0          .         0
## Regione03           0.0892        0.0892              -0          .         0
## Regione04           0.0504        0.0504              -0          .         0
## Regione05           0.0562        0.0562               0          .         0
## Regione06           0.0400        0.0400               0          .         0
## Regione07           0.0452        0.0452               0          .         0
## Regione08           0.0592        0.0592               0          .         0
## Regione09           0.0759        0.0759               0          .         0
## Regione10           0.0254        0.0254               0          .         0
## Regione11           0.0402        0.0402               0          .         0
## Regione12           0.0641        0.0641               0          .         0
## Regione13           0.0375        0.0375               0          .         0
## Regione14           0.0224        0.0224               0          .         0
## Regione15           0.0567        0.0567               0          .         0
```

```
## Regione16       0.0576       0.0576            0       .       0
## Regione17       0.0299       0.0299            0       .       0
## Regione18       0.0436       0.0436            0       .       0
## Regione19       0.0655       0.0655            0       .       0
## Regione20       0.0425       0.0425            0       .       0
## MacroArea1      0.2330       0.2330            0       .       0
## MacroArea2      0.2058       0.2058           -0       .       0
## MacroArea3      0.2056       0.2056            0       .       0
## MacroArea4      0.3557       0.3557            0       .       0
## Income1         0.1247       0.1247            0       .       0
## Income2         0.2192       0.2192            0       .       0
## Income3         0.2110       0.2110            0       .       0
## Income4         0.2258       0.2258            0       .       0
## Income5         0.2194       0.2194            0       .       0
## WorkStatus21    0.0185       0.0185            0       .       0
## WorkStatus22    0.0002       0.0002            0       .       0
## WorkStatus23    0.7778       0.7778            0       .       0
## WorkStatus24    0.2034       0.2034            0       .       0
## Age1            0.4797       0.4797            0       .       0
## Age2            0.5203       0.5203           -0       .       0
##           eCDF Max Std. Pair Dist.
## Gender0          0          0
## Gender1          0          0
## Education21      0          0
## Education22      0          0
## Education23      0          0
## Regione01        0          0
## Regione02        0          0
## Regione03        0          0
## Regione04        0          0
## Regione05        0          0
## Regione06        0          0
## Regione07        0          0
## Regione08        0          0
## Regione09        0          0
## Regione10        0          0
## Regione11        0          0
## Regione12        0          0
## Regione13        0          0
## Regione14        0          0
## Regione15        0          0
```

```
## Regione16         0           0
## Regione17         0           0
## Regione18         0           0
## Regione19         0           0
## Regione20         0           0
## MacroArea1        0           0
## MacroArea2        0           0
## MacroArea3        0           0
## MacroArea4        0           0
## Income1           0           0
## Income2           0           0
## Income3           0           0
## Income4           0           0
## Income5           0           0
## WorkStatus21      0           0
## WorkStatus22      0           0
## WorkStatus23      0           0
## WorkStatus24      0           0
## Age1              0           0
## Age2              0           0
##
## Sample Sizes:
##                Control Treated
## All            10894.    12644
## Matched (ESS)  6524.89   12092
## Matched        10191.    12092
## Unmatched      703.      552
## Discarded      0.        0
```

```
matched_data_cem <- match.data(matchit_fit_cem)

tabella_bilanciamento_matched_data_cem  <- bal.tab(trattato ~ Gender + Education2 + MacroArea + Income + Regione + WorkStatus2 + Age,
                                   data = matched_data_cem )
tabella_bilanciamento_matched_data_cem
```

```
## Balance Measures
##                Type Diff.Un
## Gender         Binary -0.0043
## Education2_1   Binary  0.0248
## Education2_2   Binary -0.0019
```

```
## Education2_3  Binary -0.0229
## MacroArea_1   Binary -0.0063
## MacroArea_2   Binary  0.0045
## MacroArea_3   Binary -0.0032
## MacroArea_4   Binary  0.0050
## Income_1      Binary -0.0374
## Income_2      Binary -0.0100
## Income_3      Binary -0.0046
## Income_4      Binary  0.0367
## Income_5      Binary  0.0153
## Regione_01    Binary  0.0223
## Regione_02    Binary -0.0027
## Regione_03    Binary -0.0280
## Regione_04    Binary -0.0032
## Regione_05    Binary -0.0023
## Regione_06    Binary  0.0038
## Regione_07    Binary  0.0022
## Regione_08    Binary  0.0062
## Regione_09    Binary  0.0167
## Regione_10    Binary -0.0025
## Regione_11    Binary  0.0066
## Regione_12    Binary -0.0241
## Regione_13    Binary  0.0121
## Regione_14    Binary -0.0100
## Regione_15    Binary -0.0059
## Regione_16    Binary -0.0079
## Regione_17    Binary -0.0021
## Regione_18    Binary  0.0069
## Regione_19    Binary  0.0032
## Regione_20    Binary  0.0087
## WorkStatus2_1 Binary -0.0099
## WorkStatus2_2 Binary  0.0001
## WorkStatus2_3 Binary  0.0049
## WorkStatus2_4 Binary  0.0049
## Age_2         Binary  0.0412
##
## Sample sizes
##     Control Treated
## All   10191   12092
```