*Procedure di aggiustamento ICVT*

*3.7.1 Ponderazione*

A questo punto del processo di produzione dei dati si è affrontato il problema delle procedure di aggiustamento dei dati, al fine di bilanciare i possibili effetti distorsivi determinati dai fenomeni di non contatto, di rifiuto e di non risposta parziale, in particolare ai quesiti relativi alle variabili monetarie presenti nel questionario.

Come ricordato nella precedente sez. 3.5 il numero di famiglie effettivamente intervistato è stato sensibilmente inferiore all’obiettivo che ci si era prefissati (2625 famiglie contro le 4000 previste), con andamenti fortemente differenziati per rilevatore e conseguentemente per aree territoriali. In estrema sintesi, possiamo dire che tale risultato è dipeso essenzialmente da due fattori: i) le diverse abilità degli intervistatori ad entrare effettivamente in contatto con le famiglie; ii) la capacità, una volta ottenuto il contatto, di portare a termine positivamente l’intervista. Il primo fattore, misurato dal rapporto tra le famiglie effettivamente contattate (3720) e le famiglie con cui si è tentato di avere un contatto (4799), rappresenta il tasso di contatto ed è risultato pari a 0.76; il tasso di contatto risulta fortemente variabile sia tra i rilevatori, sia tra i quattro supervisori, con rilevanti conseguenze sulla conservazione della capacità autoponderante del campione. Il secondo fattore, denominato tasso di risposta, e calcolato come rapporto tra famiglie effettivamente intervistate (2625) e famiglie contattate, è risultato pari al 70.6 con andamenti difformi tra aree geografiche, rilevatori, e altre variabili relative al contesto abitativo e alle caratteristiche familiari delle famiglie da intervistare.

Come consuetudine nelle indagini campionarie complesse su individui e famiglie, si è pertanto ritenuto necessario ed opportuno procedere alla definizione di un sistema di pesi in grado di attenuare i possibili effetti di distorsione indotti dal processo di selezione che ha presumibilmente caratterizzato le varie fasi della rilevazione.

Prima di entrare nel merito della procedura di ponderazione predisposta all’uopo, si ritiene opportuno sottolineare che alcuni confronti effettuati tra le distribuzioni campionarie e quelle provenienti da fonti esterne (Censimento –dati provvisori-, anagrafi, indagine campionarie Istat e Banca d’Italia) di alcune variabili caratteristiche di individui e famiglie, hanno indicato una più che discreto adattamento del nostro campione (per maggiori dettagli si veda la successiva Tab. 6). La procedura di ponderazione è stata quindi realizzata nei seguenti passi:

* 1. pesi di disegno: i cosiddetti pesi di disegno sono stati definiti come l’inverso del tasso di contatto delle famiglie per strato. Ciò ha permesso da un lato di ripristinare la struttura autoponderante del campione, dall’altro di attenuare i possibili effetti distorsivi indotti dal diverso andamento territoriale dei tassi di contatto e da alcune peculiarità sulle forme di residenza fittizia delle famiglie rilevate in alcune aree della Toscana (ad esempio si sono riscontrate numerose famiglie ufficialmente residenti nei comuni costieri della regione, ma di fatto domiciliate o altrove. Tale insieme di pesi prende il nome di *whstr,* e rappresenta il primo fattore per la definizione dei pesi finali.
	2. pesi di non risposta: si è proceduto quindi alla definizione dei pesi per compensare i fenomeni di non risposta totale all’indagine; ciò è avvenuto attraverso la specificazione di un modello statistico (regessione logistica) che mette in relazione le risposte/non risposte all’indagine con una serie di covariate disponibili sulle famiglie contattate. Dal modello, stimato con una procedura *stepwise*, si evince che la non risposta è stata condizionata dalle seguenti variabili dicotomiche: *i)* il contesto abitativo (l’abitazione di residenza è isolata o situata in un borgo o in un centro abitato senza servizi, [D01X]), *ii)* la tipologia dell’abitazione di residenza (villa, villetta singola bi/trifamiliare, villa a schiera, edificio rurale, casa singola o bifamiliare contigua [D03X]); *iii)* lo stato di conservazione esterno dell’edificio di residenza (buono [D04X1], pessimo [D04X2]); *iv)* l’attività dei supervisori (supervisore 1 [SUPERVX1], supervisore 3 [SUPERVX3]); *v)* la provincia di residenza della famiglia (segnatamente le provincie di Firenze [D48], di Livorno [D49], di Pisa [D50] e di Arezzo [D51]). L’insieme di pesi è definito come l’inverso della probabilità di risposta che è possibile derivare dal modello di regressione logistica stimato (vedi Tab. 7) e prende il nome di *wnr.*
	3. pesi di post-stratificazione: si è proceduto alla post-stratificazione dei dati, sfruttando le informazioni disponibili da fonti esterne, prevalentemente censuarie, su alcune delle caratteristiche rilevate con l’indagine. Questa fase ha riguardato distintamente le famiglie e gli individui. Relativamente alle famiglie il peso è costruito per riportare la distribuzione campionaria delle famiglie per provincia a quella ottenuta con il Censimento della Popolazione (C.P. 2001- dati provvisori). Tale procedimento ha dato luogo all’insieme di pesi *whpv*. Per quanto riguarda gli individui si sono effettuate due distinti passi: la prima post-stratificazione è fatta rispetto alla distribuzione della popolazione toscana per sesso e provincia (C.P.2001), come il solito normalizzato rispetto al totale degli individui del campione, dando luogo ai pesi *whpvs*; la seconda post-stratificazione degli individui è stata svolta per sesso e fascia di età (0-14, 15-29, 30-44, 45-64, >64), questa volta su base regionale: in questo caso i dati di riferimento sono quelli anagrafici al 31/12/2001. Quest’ultimo peso, prende il nome di *whrse*.
	4. pesi campionari finali: dati i risultati dei precedenti passi è stato possibile calcolare i pesi campionari finali per famiglie (*whfam*) e per gli individui (*whind*): il primo è ottenuto come prodotto dei tre pesi familiari (*whstr×wnr×whpv*) e successivamente normalizzato all’effettivo numero di famiglie rispondenti; analogamente, il peso campionario finale per gli individui è calcolato come il prodotto del peso familiare per i pesi di post-stratificazione individuale (*whfam \* whpvs \*whrse*) e successivamente normalizzato per il numero di individui effettivamente rispondenti.
	5. pesi di riporto all’universo: infine, si sono calcolati anche i pesi per il riporto all’universo di famiglie (*whfamun*) ed individui (*whindun*) della Toscana: questi rappresentano i cosiddetti coefficienti di espansione da utilizzare per riportare le numerosità di individui e famiglie del campione alle numerosità dei relativi universi di riferimento.

*Tab. 6: Confronto universo e campione finale, per alcune caratteristiche di individui e famiglie*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variabili  |  | **UNIVERSO**Dati da fonti diverse\* | **CAMPIONE***Non pesate* | **CAMPIONE***pesate[[1]](#footnote-1)* |
| GENERE *\*Censimento 2001, provvisorio* | MF | 48.151.9 | 48.451.6 | **48.1****51.9** |
| ETA’*\*Istat 31/12/00* | 0-1415-2930-4445-64>64 | Tot M F11.6 12.4 10.916.9 17.8 15.922.8 24.0 21.726.7 27.0 26.422.1 18.9 25.1 | Tot M F11.1 11.6 10.616.2 17.3 15.121.8 22.2 21.428.7 29.0 28.622.2 19.9 24.3 | Tot M F**11.6 12.4 10.9****16.9 17.8 15.9****22.8 24.0 21.7****26.7 27.0 26.4****22.1 18.9 25.1** |
| STATO CIVILE*\*Indagine Banca d’Italia, Centro, 2000* | Celibe/nubileConiugatoSep/div.vedovo | M+F34.054.63.18.3 | M+F33.754.82.88.7 | M+F34.853.33.08.9 |
| TITOLO STUDIO*\* F.L. 2001* | Nessuno/lic.elem.Media inf/Qual.pr.DiplomaLaurea e oltre | Tot M F41.5 37.0 45.731.8 35.5 28.420.4 21.2 19.76.3 6.3 6.2 |  Tot M F39.3 34.7 43.728.9 33.3 24.723.9 24.2 23.87.9 7.8 7.9 |  Tot M F39.0 34.2 43.429.2 33.6 25.123.9 24.3 23.67.9 7.9 7.9 |
| CONDIZIONE OCCUPAZIONALE*\*F.L 2001* | OccupatiForza lavoroNon forza lavoro | Tot M F41.3 50.8 32.643.6 52.3 35.456.4 47.7 64.6 |  Tot M F40.4 48.4 32.944.2 51.9 37.155.8 48.1 62.9 |  Tot M F40.4 49.2 32.344.2 52.4 36.655.8 47.6 63.4 |
| NUM. MEDIO COMP. FAMIGLIA*\*CP 2001, provv.**\*BdI, 2000, Centro* |  | 2,502.68 | 2.62 | 2.60 |
| COMPONENTI FAMIGLIA*\*Indagine Banca d’Italia 2000, Centro* | 12345 e oltre | 19.528,923.022.16.5 | 19.530.825.518.55.7 | 19.731.325.118.35.6 |
| N. MEDIO PERCETTORI DI REDDITO PER FAMIGLIA \*BdI, Centro, 2000 |  | 1.89 | 1.79 | 1.77 |

*Tab. 7: Stima del modello di regressione logistica*

|  |  |
| --- | --- |
| Variabile | Stima |
| Intercetta | 0,2463\*\*\* |
| D01X | 0,2247\*\* |
| D03X | 0,3689\*\*\* |
| D04X1 | 0,5528\*\*\* |
| D04X2 | -0,4057\*\* |
| SUPERVX1 | 0,4578\*\*\* |
| SUPERVX3 | 0,8264\*\*\* |
| D48X | -0,3144\*\*\* |
| D49X | -0,3316\*\* |
| D50X | 0,3805\*\*\* |
| D51X | -0,7411\*\*\* |

\*\*\* p-value ≤0.01; \*\* 0.01<p-value ≤0.05; \*0.05<p-value ≤0.10

*3.7.2 Imputazione*

Una volta messe a punto le procedure di aggiustamento dei dati per tener conto dei fenomeni di non risposta totale, è stato affrontato il problema del trattamento delle non risposte parziali, ovvero se e come procedere al completamento di quei record relativi a famiglie e individui rispondenti che presentavano però dati mancanti ad una o più delle variabili rilevate con l’indagine.

Diciamo innanzitutto che la scelta di base è stata quella di intervenire quasi esclusivamente sulle variabili relative alle varie forme di reddito sia familiare che individuale, in quanto ritenute di primaria importanza per lo svolgimento delle previste analisi per lo studio della povertà e delle condizioni di vita. Inoltre, per le altre variabili non reddituali, (i.e. possesso di beni durevoli, condizioni abitative, ecc.) la presenza di mancate risposte è risultata praticamente irrilevante. Nonostante la suddetta autolimitazione sulla scelta delle variabili da imputare, il procedimento di imputazione è stato piuttosto articolato e complesso e sostanzialmente suddivisibile in tre distinti sottoinsiemi di variabili:

1. Imputazione dei redditi da lavoro individuali;
2. Imputazione dei redditi da fabbricati individuali;
3. Imputazione dei redditi presunti (affitto presunto) e della superficie dell’abitazione di residenza;

L’applicazione delle procedure che si andranno nel seguito a descrivere ha comunque permesso di ottenere una matrice dei dati completa, perlomeno per quanto riguarda le tipologie di variabili testè ricordate.

*Tab.8: Distribuzione di frequenza al quesito C3.2 e distribuzione delle non risposte ai quesiti sul reddito*

|  |  |  |  |
| --- | --- | --- | --- |
| Modello di dichiarazione compilato (C.3.2) | Frequenza | di cui non rispondenti ai redditi  | non rispondenti ai redditi (%) |
| Nessuno, solo pensione | 817 | 30 | 3.7 |
| Nessuno, solo stipendio | 264 | 19 | 7.2 |
| CUD 2001 | 370 | 116 | 31.3 |
| 730/2001 | 1653 | 596 | 36.1 |
| UNICO 2001 | 1068 | 569 | 53.3 |
| Nessuna dichiarazione | 174 | 100 | 57.5 |
| *Totale risposte valide* | *4346* | *1430* | *32.9* |
| Missing | 357 | 357 | 100.0 |
| *Totale* | *4703* | *1787* | *38.0* |

1. Imputazione dei redditi da lavoro individuali

Per quanto attiene ai redditi da lavoro si è innanzitutto osservata la distribuzione di frequenza alla domanda C3.2, per noi fondamentale, relativa al modello di dichiarazione fiscale dei redditi compilato, congiuntamente alla successiva compilazione o meno dei quadri richiesti (vedi Tab. 8). Da ciò si evincono due sottoinsiemi di individui da trattare: i) il primo è rappresentato da coloro i quali, pur dichiarando il modello di dichiarazione fiscale compilato, hanno poi rifiutato di fornire i dati reddituali (1430 individui); ii) il secondo (357 individui), formato invece da coloro i quali si sono rifiutati totalmente di rispondere a tutti i quesiti relativi al reddito, compresa la domanda che richiedeva quale modello era stato compilato (*missing* alla domanda C3.2). Si è ritenuto opportuno fare tale distinzione per due ordini di motivi. In primo luogo, l’analisi svolta sui dati di reddito osservati per chi aveva risposto alla C3.2, come si può notare dalla tab.9, evidenziano forti differenziazioni - sia in termini di reddito medio, sia per quanto riguarda la percentuale di mancate risposte - a seconda delle modalità di dichiarazione fiscale seguite. Ciò ci ha fatto propendere, come meglio dettagliato nel seguito, a condurre ulteriori analisi per verificare l’opportunità di operare imputazioni distinte per ognuno dei sei sottogruppi individuati dalle risposte al quesito C3.2. In secondo luogo si deve tener presente che per chi aveva invece rifiutato di rispondere alla domanda C3.2 non si poteva disporre di alcuna informazione su qualsivoglia tipologia di reddito.

Per quanto riguarda il primo gruppo, come precedentemente accennato, prime analisi esplorative effettuate sui dati osservati hanno messo in evidenza che la risposta al quesito C3.2, determina l’individuazione di sottogruppi di percettori piuttosto differenziati sia in termini di reddito medio e della sua variabilità, sia in termini di comportamento di risposta (Tabb. 8 e 9).

*Tab. 9: Reddito individuale netto da lavoro sui dati osservati (annuale in migliaia di lire)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Modello di dichiarazione compilato (C.3.2) | N | Media | Minimo | Massimo | Std | C.V. |
| Nessuno, solo pensione | 787 | 15468 | 1690 | 119603 | 10222 | 0.661 |
| Nessuno, solo stipendio | 245 | 20428 | 4000 | 60000 | 8683 | 0.425 |
| CUD 2001 | 254 | 22893 | 830 | 95986 | 11663 | 0.509 |
| 730/2001 | 1057 | 25000 | 0 | 127239 | 13392 | 0.536 |
| UNICO 2001 | 499 | 29954 | 0 | 451138 | 35143 | 1.173 |
| Nessuna dichiarazione | 74 | 18217 | 700 | 100000 | 17083 | 0.938 |
| Totale | 2916 | 22535 | 0 | 451138 | 18856 | 0.837 |

Pertanto, al fine di verificare la significatività della capacità esplicativa della variabile C3.2 rispetto al reddito, si è proceduto ad un’analisi della varianza, considerando come variabile di classificazione le modalità di risposta al quesito C3.2. Dall’analisi condotta sui dati osservati è emerso come il 10% della variabilità dei dati di reddito risulti significativamente spiegato dalla suddetta variabile. Conseguentemente, si è deciso di procedere all’imputazione dei dati mancanti per ognuno dei sei sottogruppi individuati dalla variabile C3.2, secondo le modalità qui di seguito descritte.

* Per chi aveva dichiarato di avere solo reddito da pensione (C3.2=1), solo stipendio (C3.2=2) oppure nessuna dichiarazione compilata (C3.2=6), si è proceduto ad una semplice imputazione per medie di cella, distintamente per ognuno dei sottogruppi sopra ricordati. Dopo alcune analisi, volte ad individuare quali variabili potessero essere predittive del reddito e nel contempo associate al meccanismo di non risposta si è giunti alla costruzione delle celle di imputazione considerando due variabili dicotomiche: *Sesso* (maschio, femmina) e *Titolo di studio* (basso, alto).
* Per le altre tre modalità di risposta al quesito C3.2 (ovvero C3.2=3=CUD 2001, C3.2=4=730/2001, C3.2=5=UNICO/2001), che presentano percentuali di dati mancanti usualmente più elevate e soprattutto profili reddituali e di variabilità piuttosto differenziati, si è deciso di adottare una strategia di imputazione più complessa. Per ognuno dei tre sottogruppi sopra indicati si è proceduto ad una imputazione per regressione stocastica, sintetizzabile nelle seguenti fasi: i) stima sui dati osservati dei parametri di un modello di regressione lineare (dove la variabile dipendente è rappresentata dal logaritmo del reddito individuale), con variabili esplicative rilevate con l’indagine, ii) modifica dei parametri stimati, mediante l’aggiunta di un opportuna componente casuale; iii) stima provvisoria del logaritmo del reddito da imputare sulla base della regressione con parametri modificati; iv) stima del valore finale da imputare, ottenuta aggiungendo alla stima provvisoria una componente casuale che renda conto della variabilità non spiegata dal modello. Questo procedimento è stato effettuato utilizzando il *software IVE-ware*, che permette anche di imporre dei vincoli sul campo di variazione dei dati imputati. Le variabili esplicative sono diverse nei tre casi così come le capacità esplicative dei singoli modelli (vedi Tab. 10). In particolare, il modello relativo al gruppo CUD presenta un R2adj=0.15, il modello relativo al gruppo 730 un R2adj=0.32, mentre nel caso dell’UNICO un R2adj=0.26.

*Tab.10: Stime dei parametri e livelli di significatività. per l’imputazione del reddito*

 *secondo il modello di dichiarazione*

|  |  |  |  |
| --- | --- | --- | --- |
| Modello di dichiarazione | **CUD** | **730** | **UNICO** |
| Variabile | Stima | Stima | Stima |
| Costante | 15.06891\*\*\* | 15.09531\*\*\*  | 14.48347\*\*\* |
| dsesso (1=maschio) | 0.35609\*\*\* | 0.26180\*\*\*  | 0.29490\*\*\* |
| etax1 (età individuo)  | 0.04108\*\*\* | 0.05662\*\*\*  | 0.06604\*\*\* |
| etax1sq (etax1^2) | -0.00032\*\* | -0.00049\*\*\*  | -0.00053\*\*\* |
| dstu56 (1=diploma univ. e oltre) | 0.72220\*\*\* | 0.65234\*\*\*  | 0.66640\*\*\* |
| dstu23 (1=lic. Media-qual.prof) | 0.41098\*\*\* | 0.27282\*\*\*  | 0.28040\*\*\* |
| dstu4 (1=diploma maturità) | 0.56448\*\*\* | 0.48468\*\*\*  | 0.55356\*\*\* |
| dc3101 (1=dich non congiunta) |  | -0.18906\*\*\*  |  |
| dposdip (1= Dipendente) |  | 0.20760\*\*\*  | 0.21543\*\* |
| Dposind (1=indipendente) |  |  | 0.22164\*\* |
| *Adj R-Sq*  | *0.1525* | *0.3216*  | *0.2567* |

\*\*\* p-value ≤0.01; \*\* 0.01<p-value ≤0.05; \*0.05<p-value ≤0.10

Relativamente al secondo gruppo, ovvero i 357 individui (poco meno dell’8%) che non avevano risposto neppure al quesito C3.2, si è deciso di trattarli ricorrendo alla medesima metodologia di imputazione descritta sopra, ovvero un’imputazione stocastica per modello di regressione.

*Tab.11: Stime dei parametri e livelli di significatività per*

*l’imputazione del reddito dei non rispondenti al quesito C3.2*

|  |  |
| --- | --- |
| Variabile | Stima |
| Costante | 14.72348\*\*\* |
| dsesso (1=maschio) |  0.33579\*\*\* |
| etax1 (età individuo)  | 0.04699\*\*\* |
| etax1sq (etax1^2) | -0.00035\*\*\* |
| dstu56 (1=diploma univ. e oltre) | 0.64877\*\*\* |
| dstu23 (1=lic. Media-qual.prof) | 0.26890\*\*\* |
| dstu4 (1=diploma maturità) | 0.49588\*\*\* |
| Cocc1 (1=occupato) | 0.34397\*\*\* |
| Relpar12 (1=capofamiglia o coniuge) | 0.05544\*\* |
| *Adj R-Sq*  | *0.2577* |

\*\*\* p-value ≤0.01; \*\* 0.01<p-value ≤0.05; \*0.05<p-value ≤0.10

In questo caso, però, il modello di regressione è stato stimato sui dati completati mediante le imputazioni sopra descritte. Come si può vedere dalla Tab. 11 il nuovo modello presenta un valore di adattamento abbastanza soddisfacente (R2adj=0.26).

Completato nel modo sopradescritto il processo di imputazione del reddito individuale, è possibile riassumere nella seguente Tab. 12 l’impatto che tale processo ha avuto sui dati. Come ci si aspettava, le imputazioni per cella (modalità 1, 2, 6) hanno comportato marginali differenze sulle relative medie e variabilità, ad eccezione del gruppo che non aveva presentato alcuna dichiarazione dei redditi, peraltro quantitativamente marginale (il 4,0% circa). Per quanto riguarda le imputazioni per regressione relative alle modalità 3, 4 e 5, si può invece notare come, secondo le aspettative, i dati completati, presentino valori di reddito mediamente superiori a quelli osservati (appare infatti presumibile una maggiore reticenza alla risposta tra i possessori di redditi più elevati). Per quanto riguarda la variabilità dei dati, la situazione si presenta invece di più difficile interpretazione. Per il CUD e il 730 si nota un certo incremento tra dati osservati e dati completati, mentre per quanto riguarda l’UNICO si nota un leggero decremento. In ogni caso si deve sottolineare come la procedura di imputazione seguita non riproduca correttamente al variabilità delle stime, in quanto basata su un'unica realizzazione del processo di imputazione. E’ nostra intenzione, in un prossimo futuro, di affinare tale metodologia, adottando un procedimento di imputazione multipla, che permetta di tener conto più correttamente della variabilità dei dati imputati.

*Tab. 12: Reddito individuale netto da lavoro sui dati osservati e completati*

 *(annuale in migliaia di lire)*

|  |  |  |
| --- | --- | --- |
| Modello fiscale (C.3.2) | Dati osservati | Dati completati |
|  | N | Media | std | c.v. | N | Media | std | c.v. |
| 1=Nessuno, solo pensione | 787 | 15468 | 10222 | 0.661 | 817 | 15526 | 10072 | 0.649 |
| 2=Nessuno, solo stipendio | 245 | 20428 | 8683 | 0.425 | 264 | 20478 | 8389 | 0.409 |
| 3=CUD 2001 | 254 | 22893 | 11663 | 0.509 | 370 | 23024 | 13369 | 0.581 |
| 4=730/2001 | 1057 | 25000 | 13392 | 0.536 | 1654 | 26145 | 14626 | 0.559 |
| 5=UNICO 2001 | 499 | 29954 | 35143 | 1.173 | 1068 | 32800 | 31881 | 0.972 |
| 6=Nessuna dichiarazione | 74 | 18217 | 17083 | 0.938 | 174 | 17720 | 12062 | 0.681 |
| ‘Missing’[[2]](#footnote-2) | 4347 | 24837 | 20277 | 0.816 | 4703 | 24885 | 20249 | 0.814 |

1. *Nella tabella 6, le cifre in neretto indicano le distribuzioni dl campione utilizzate per la post-stratificazione.* [↑](#footnote-ref-1)
2. Per dati osservati si intendono anche quelli completati nelle precedenti fasi di imputazione. [↑](#footnote-ref-2)