Nota metodologica dell'Indagine Inapp-Plus 2022

1. Il disegno dell'Indagine

L'indagine INAPP PLUS (Participation Labour Unemployment Survey) è una rilevazione nazionale campionaria sull'offerta di lavoro alla sua nona annualità, presente nel Piano Statistico Nazionale dal 2006.

L'obiettivo primario dell'indagine PLUS è quello di fornire stime statisticamente affidabili di fenomeni rari o solo marginalmente esplorati dalle maggiori rilevazioni sul mercato del lavoro italiano (ISTAT, INPS). Infatti, se la indagine sulle Forze di Lavoro (IFL) dell'Istat fornisce regolarmente gli aggregati e gli indicatori ufficiali sul mercato del lavoro (tassi di disoccupazione, occupazione, attività, ecc.), l'indagine PLUS è principalmente rivolta all'approfondimento di specifici aspetti, quali la distribuzione dei contratti (dipendente, autonomo, informale, ecc.), la ricerca di lavoro, la partecipazione lavorativa delle donne, dei giovani e delle persone con più di cinquanta anni, le scelte di pensionamento, istruzione e formazione, le dinamiche inter-generazionali.

L'indagine è stata ideata per analizzare il mercato del lavoro attuale, contraddistinto da una sempre più veloce trasformazione delle modalità di ricerca e svolgimento del lavoro (rimaste pressoché invariate nei decenni passati) in nuove e molteplici forme, in cui i concetti di occupazione e disoccupazione sfuggono oggi sempre più alle categorie tradizionali alle quali eravamo stati abituati. Per affrontare le sfide epistemologhe appena descritte, l'indagine INAPP PLUS si distingue dalle usuali indagini su base familiare per l'assenza di rispondenti proxy e per la capacità di integrare diversi aspetti del mercato del lavoro, spesso analizzati in maniera disgiunta.

Inoltre, al fine di misurare la qualità dell'attuale occupazione con la massima precisione possibile, i dati PLUS sull'occupazione sono ispirati ad un criterio classificatorio lievemente diverso rispetto ai dati RCFL. Infatti, mentre la rilevazione PLUS definisce come occupati e in cerca di lavoro le persone che si *auto definiscono* tali, RCFL segue un percorso che identifica la condizione in base ad alcune informazioni "oggettive" che riguardano: per gli occupati l'aver svolto almeno un'ora di lavoro retribuito nella settimana di riferimento oppure avere un lavoro dal quale si è assenti, con una durata dell'assenza che non supera i tre mesi; per le persone in cerca l'aver compiuto almeno un'azione di ricerca entro le quattro settimane che precedono l'intervista ed essere immediatamente disponibili a lavorare[1]. Da questa dovuta precisazione ne segue come

[1] La definizione di occupato utilizzata da RCFL è cambiata mentre era in corso la rilevazione Plus 2021: la definizione del campione Plus è stata realizzata con la vecchia definizione di occupato mentre per le stime sono stati utilizzati i dati provvisori riferiti alle nuove definizioni. In sintesi, nella nuova definizione, per gli assenti da lavoro la durata dell'assenza (più o meno di 3 mesi) diviene il criterio prevalente per definire la condizione di occupato. Per maggiori dettagli visitare la pagina web: https://www.istat.it/it/archivio/253095

l'impianto ISTAT (EUROSTAT) sembri sottintendere nel proprio meccanismo contatore una certa sovrastima degli occupati e sottostima delle persone in cerca. L'idea generale dell'indagine PLUS di registrare, nel modo più accurato possibile, la condizione *auto percepita* dai soggetti intervistati fa sì che anche la distinzione tra *persone in cerca* ed *inattivi* sia differente da quanto adottato in RCFL.
In particolare:

a) si considerano *persone in cerca*, e quindi attive, alcune tipologie di individui che per l'Istat sono da considerare *inattivi*;
b) non si considerano occupati quei soggetti che svolgono una attività lavorativa che non è, in termini economici e secondo la propria percezione, tale da giustificare la loro inclusione in tale categoria (studenti, pensionati da lavoro e casalinghe – lavoratrici/ori saltuari), considerandoli *occupati non prevalenti*.

In questo modo non vengono considerati, tra gli *occupati,* quelli *con condizione non prevalente* e, tra gli inattivi, i *disoccupati che non rientrano nella classificazione ISTAT*. Queste sotto-popolazioni rappresentano proprio gli ambiti di maggior interesse per lo studio e le politiche per l'attivazione stabile e continuativa. Ovviamente da PLUS è possibile ricostruire gli occupati nelle definizioni ISTAT-EUROSTAT essendo somministrati i quesiti necessari alla loro individuazione. A tale scopo, si è deciso di vincolare i dati PLUS ad alcuni aggregati ufficiali di fonte RCFL, precisando che gli *occupati non prevalenti* e gli inattivi che si dichiarano in cerca sono stati considerati nelle condizioni cui si attribuivano autonomamente.

2. Piano di campionamento e riporto all'universo

La rilevazione Inapp-PLUS 2022 ha raccolto informazioni su circa 46 mila individui intervistati telefonicamente attraverso un sistema CATI ed in assenza di rispondenti proxy. Il questionario, di circa 200 domande complessive, è stato sottoposto ad un campione di persone residenti con età compresa tra 18 e 74 anni.
Per quanto riguarda gli individui 18-74 anni, sono esclusi dal campo di osservazione soltanto gli studenti con età maggiore di 39 anni in modo da poter effettuare analisi sul 99,8% dell'offerta di lavoro nel nostro paese.
Nell'ottica della riduzione del fastidio statistico dell'indagine, il questionario è organizzato in sezioni appositamente studiate per specifici target di popolazione (occupati, disoccupati, giovani, donne, ecc.). Inoltre, a partire dalla sua seconda annualità (2006) una consistente quota di interviste longitudinali (panel), effettuate in anni consecutivi agli stessi individui, è inclusa nel campione (nel 2022 la quota delle interviste panel è pari al 21,5 %)[2].

[2] La presenza di interviste panel contribuisce ad "alleggerire" ulteriormente il questionario, evitando di effettuare domande su fenomeni invariati nel tempo a persone intervistate l'anno precedente

La pianificazione delle interviste da effettuare è stata realizzata sulla base di un campionamento per quote stratificato con definizione di domini di studio parzialmente sovrapposti[3].

Il campione è suddiviso in dieci target fondamentali (o domini) costituiti da:

1. giovani occupati, in età compresa tra 18 e 29 anni;
2. giovani studenti, in età compresa tra 18 e 39 anni;
3. giovani in cerca di occupazione, in età compresa tra 18 e 29 anni;
4. donne attive, in età compresa tra 18 e 39 anni;
5. donne inattive, in età compresa tra 18 e 39 anni;
6. adulti e anziani attivi, in età compresa tra 50 e 74 anni;
7. adulti e anziani pensionati da lavoro, in età compresa tra 50 e 74 anni;
8. in cerca di occupazione 18-74 anni (definizione estesa);
9. occupati 18-74 anni;
10. altri inattivi (non studenti/non pennsionati da lavoro) 18-74 anni.

Allo scopo di poter fornire stime attendibili anche per sottopopolazioni di questi 10 domini (ad esempio, limitatamente a ciascuna delle 20 regioni italiane) si è proceduto alla pianificazione di un campionamento stratificato, dove gli strati - definiti dall'incrocio delle variabili riportate in Figura 1 - costituiscono una partizione del campione e (raggruppamenti di strati) degli stessi domini di studio.

Figura 1 – Variabili di stratificazione campionaria, PLUS 2022

Variabili	Modalità
Regione	Piemonte e Valle d'Aosta, Lombardia, Trentino A.A., Veneto, Friuli V.G., Liguria, Emilia Romagna, Toscana, Umbria, Marche, Lazio, Abruzzo, Molise, Campania, Puglia, Basilicata, Calabria, Sicilia, Sardegna
Tipo di comune	Comune metropolitano, Comune non metropolitano
Sesso	Maschi, Femmine
Età in classi	18-24, 25-29, 30-39, 40-49, 50-64, 65-74
Condizione occupazionale	Occupato, In cerca di occupazione, Studente, Pensionato da lavoro, Altro inattivo (casalinga)

[3] La scelta del campionamento per quote è stata motivata dall'esigenza di ridurre notevolmente la numerosità campionaria necessaria alla produzione di stime statisticamente significative per piccole sottopopolazioni di interesse. Come scelta alternativa si sarebbero potute utilizzare strategie classiche di campionamento a due stadi (es: comuni e famiglie) che però, oltre a richiedere rilevazioni sul campo molto più onerose, avrebbero comportato la rinuncia ad una delle principali caratteristiche dell'indagine PLUS, quella dell'assenza di rispondenti proxy.

Il numero di interviste da realizzare per ciascuno degli strati è stato determinato in modo da fornire stime attendibili per l'intera popolazione di riferimento e per particolari sottoinsiemi d'interesse, attraverso l'implementazione di una procedura di allocazione multi-dominio, basata sulla risoluzione di un problema di minimizzazione vincolata. Più precisamente, sono stati fissati a priori livelli di varianza massima per i domini di interesse elencati sopra e le loro disaggregazioni territoriali per regione e tipo di comune (metropolitano e non).
Formalmente, possiamo rappresentare la numerosità della popolazione in ogni dominio (o sotto dominio) come semplice somma di sottopopolazioni di strato,

$$N_d = \sum_h N_h I_{h,d}$$

dove $I_{h,d}$ è una variabile indicatrice che assume valore 1 (0) se lo strato *h* è (non è) incluso nel dominio *d*. Indicando con p_d la generica stima della proporzione $P_d = N_d / N$ della popolazione nel dominio *d*, la procedura è sintetizzata dal seguente problema di minimo vincolato,

$$\sum_h n_h = \min$$
$$\text{s.t. } V(p_d) \leq V_d^* \quad \forall d$$

dove il limite V_d^* varia opportunamente per diversi sotto domini. La criticità tipica dei campionamenti per quote, rappresentata dalla non conoscenza a priori delle probabilità di inclusione delle unità di rilevazione (ovvero, delle numerosità N_h), è stata superata considerando come popolazione di riferimento quella ottenuta tramite stima dalla rilevazione RCFL dell'Istat.
Relativamente alla rilevazione Plus 2022, vengono riportate in Tabella 1 la numerosità nella popolazione, nel campione teorico, nel campione effettivo, il tasso di sondaggio teorico e il tasso di sondaggio effettivo (per 100.000 residenti), per condizione occupazionale, classe di età e sesso.

Tabella 1 - Plus 2022: numerosità nella popolazione di riferimento, numerosità del campione teorico, dl campione effettivo, tasso di sondaggio teorico e tasso di sondaggio effettivo (per 100.000 residenti), per condizione occupazionale, classe di età e sesso

Condizione occupazionale e classe di età	Maschi					Femmine				
	Numerosità nella popolazione	Campione teorico	Tasso di sondaggio teorico	Campione effettivo	Tasso di sondaggio effettivo	Numerosità nella popolazione	Campione teorico	Tasso di sondaggio teorico	Campione effettivo	Tasso di sondaggio effettivo
Occupato 18-24	629.730	1.226	195	1.228	195	375.080	1.126	300	1.133	302
In cerca (estesa) 18-24	352.924	1.567	444	1.563	443	279.014	1.299	466	1.302	467
Altro inattivo 18-29	219.579	251	114	264	120	412.211	1.147	278	1.144	278
Inattivo studente 18-24	1.004.852	1.295	129	1.451	144	1.158.841	3.401	293	3.379	292
Occupato 25-29	974.915	1.893	194	1.903	195	749.629	1.778	237	1.782	238
In cerca (estesa) 25-29	298.662	1.327	444	1.322	443	267.146	1.241	465	1.251	468
Inattivo studente 25-39	209.951	399	190	403	192	254.461	824	324	818	321
Occupato 30-39	2.719.481	1.129	42	1.147	42	1.979.861	2.695	136	2.706	137
In cerca (estesa) 30-39	412.343	516	125	518	126	479.259	878	183	883	184
Altro inattivo 30-39	186.912	151	81	162	87	810.724	2.141	264	2.132	263
Occupato 40-49	3.624.281	1.128	31	1.155	32	2.747.560	1.130	41	1.177	43
In cerca (estesa) 40-49	396.028	507	128	505	128	519.489	659	127	650	125
Altro inattivo 40-49	264.492	284	107	259	98	1.062.093	866	82	904	85
Occupato 50-74	5.015.843	3.229	64	3.215	64	3.639.351	2.345	64	2.374	65
In cerca (estesa) 50-74	517.701	720	139	697	135	453.727	633	140	622	137
Inattivo ritirato 50-64	791.898	1.175	148	1.147	145	492.162	737	150	736	150
Altro inattivo 50-74	831.095	562	68	520	63	3.968.626	877	22	1.000	25
Inattivo ritirato 65-74	2.616.662	2.342	90	2.307	88	1.923.989	1.716	89	1.816	94

La fase successiva è relativa alla scelta del tipo di stimatore di ponderazione vincolata da adottare ai fini del calcolo del coefficiente di riporto all'universo. In particolare, come per le precedenti rilevazioni, si ricorrerà all'implementazione di metodologie basate sull'utilizzo dello stimatore di regressione generalizzato (GREG estimator). Esso garantisce che le stime delle frequenze assolute delle variabili ausiliarie utilizzate come regressori siano coincidenti con i totali noti osservati nella popolazione ed imposti come vincoli di calibrazione.

Consideriamo una generica variabile di interesse *Y* e definiamo il suo totale sulla popolazione di riferimento *P* come,

$$Y = \sum_{k \in P} y_k$$

dove *k* è una generica unità appartenente a *P*. Seguendo un approccio di tipo totalmente predittivo,[4] lo stimatore $\hat{Y}$ di *Y* è formalmente derivabile attraverso la soma di due semplici componenti:

$$\hat{Y} = Y_S + \tilde{Y}_{\bar{S}} = \sum_{k \in S} y_k + \sum_{k \in \bar{S}} \tilde{y}_k$$

dove Y_S è la parte osservata sul campione e $\tilde{Y}_{\bar{S}}$ è la soma dei valori predetti $\tilde{y}_k$ (con $S \cup \bar{S} = P$). Il passo fondamentale è quello di ipotizzare l'esistenza di una relazione (lineare) tra la generica variabile di studio *Y* e un opportuno insieme di variabili esplicative $\boldsymbol{x}=(x_1,\ldots,x_K)$ tale che,

4 Vedi Dorfman A.H., Royall R.M., Valliant R. (2000).

$$\tilde{y}_k = B' x_k + \varepsilon_k \qquad (1)$$

dove la componente errore ε_k soddisfa le ipotesi standard di omoschedasticità e $\boldsymbol{x}$ è definito come il vettore di *variabili ausiliarie.* Come già anticipato, è stato implementato un approccio basato sullo *stimatore di regressione generalizzato* (GREG), nel caso particolare dei modelli con variabili strumentali.[5] Data la probabilità di inclusione π_k per ogni unità *k* del campione, i pesi base $d_k = 1/\pi_k$, e il conseguente stimatore standard di Horwitz-Thompson $\hat{Y}_{HT} = \sum_s d_k y_k$, è stato possibile definire il seguente stimatore finale;

$$\hat{Y} = \sum_{k \in S} w_k y_k \qquad (2)$$

dove,

$$w_k = d_k \left(1 + \gamma' z_k\right)$$

$$\gamma' = \left(\sum_{k \in U} x_k - \sum_{k \in S} d_k x_k\right)' \left(\sum_{k \in S} d_k z_k x_k^{'}\right)^{-1}$$

$$z_k = x_k$$

L'implementazione di tale procedura consente di:

- garantire la coerenza tra le stime prodotte tramite i dati PLUS e RCFL, sia in termini delle più generali distribuzioni demografiche che riguardo i principali indicatori del mercato del lavoro;
- migliorare l'efficienza (o precisione) delle stime;
- contribuire al controllo della distorsione dei dati campionari dovuti al ben noto fenomeno dell'attrito da selezione (dovuto principalmente alla strategia campionaria per quote, al metodo CATI di effettuazione delle interviste, nonché alla presenza di una consistente quota panel nel campione).

In generale, i risultati di una metodologia di calibrazione sono tanto più robusti quanto più "solida" è la mole di informazione ausiliaria implementata attraverso l'imposizione dei totali noti. Tra i principali prerequisiti desiderabili al fine di ottenere stimatori GREG con varianza piccola, i seguenti quattro aspetti devono essere presi in considerazione nella progettazione PLUS:

- definizione di un affidabile e plausibile modello di regressione di Y (variabili di studio) su X (variabili ausiliarie);
- inclusione di variabili ausiliarie già presenti tra le variabili di stratificazione;
- elevata affidabilità (buona reputazione) della fonte di dati da cui derivare i totali noti da implementare nella regressione.

5 Vedi Deville and Särndal (1992), Särndal e Lundström (2005).

La RCFL dell'ISTAT rappresenta la fonte ufficiale di derivazione di tutti i principali indicatori del mercato del lavoro; le variabili ausiliarie (che generano circa 110 totali noti) prese in considerazione sono: genere, ripartizione geografica, classe di età, livello di istruzione, status professionale e regime orario.

Tabella 2 – Plus 2022: Coefficiente di variazione delle stime finali per domini pianificati di indagine

Dominio pianificato di indagine	Coefficiente di variazione per stime pari a 100.000 individui
giovani occupati, in età compresa tra 18 e 29 anni	10,05
giovani studenti, in età compresa tra 18 e 39 anni	7,22
giovani in cerca di occupazione, in età compresa tra 18 e 29 anni	3,58
donne attive, in età compresa tra 18 e 39 anni	8,48
donne inattive, in età compresa tra 18 e 39 anni	5,81
adulti e anziani attivi, in età compresa tra 50 e 74 anni	15,11
adulti e anziani pensionati da lavoro, in età compresa tra 50 e 74 anni	10,78
in cerca di occupazione 18-74 anni (definizione estesa)	8,1
occupati 18-74 anni	18,46
altri inattivi (non studenti/non pennsionati da lavoro) 18-74 anni	16,35

In Tabella 2 viene riportato il coefficiente di variazione (con livello di significatività del 95%) per ciascun dominio pianificato relativamente a stime pari a 100.000 individui. Se una stima (per esempio il numero di persone occupate con una determinata tipologia contrattuale) sul dominio "occupati 18-74 anni" risulta essere pari a 100mila, allora, in base alla tabella, c'è il 95 per cento di possibilità che il valore vero sia all'interno dell'intervallo 80.000 – 120.000 mila circa.
La strategia campionaria appena descritta massimizza la coerenza tra le stime Plus e quelle RCFL. Il diverso impianto metodologico ed organizzativo alla base delle due indagini può tuttavia causare differenze nella stima di alcuni sub-aggregati all'interno dei domini pianificati. In Tabella 3 sono riportate le distribuzioni degli occupati per settore e per professione stimate da Plus e da RCFL.

Tabella 3 – Confronto tra stime Plus e stime RCFL relativamente alle distribuzioni di frequenza degli occupati per settore economico e professione – Anno 2022

Settore Economico	Plus		RCFL	Professione	Plus		RCFL 22
	v.a.	%	%		v.a.	%	%
Agricoltura, silvicoltura e pesca	**1.367**	**6,0**	**3,8**	**Qualificate e tecniche**	**8.533**	**37,2**	**34,5**
TOTALE INDUSTRIA (b-f)	**4.911**	**21,4**	**26,8**	*dirigenti e imprenditori*	719	3,1	2,7
Industria in senso stretto	3.016	13,2	20,1	*professioni intellettuali*	4.019	17,5	14,7
Costruzioni	1.895	8,3	6,8	*professioni tecniche*	3.795	16,6	17,0
TOTALE SERVIZI (g-u)	**16.638**	**72,6**	**69,4**	**Esecutive**	**3.481**	**15,2**	**12,1**
Commercio	2.471	10,8	13,6	**Qualificate nei servizi**	**3.742**	**16,3**	**18,7**
Trasporto e magazzinaggio	1.376	6,0	5,0	**Operai e artigiani**	**5.229**	**22,8**	**23,0**
Alloggio e di ristorazione	1.750	7,6	6,0	*artigiani, operai specializzati, agricoltori*	4.271	18,6	14,7
Servizi di informazione e comunicazione	674	2,9	2,9	*conduttori di impianti*	958	4,2	8,3
Attività finanziarie e assicurative	1.126	4,9	2,6	**Personale non qualificato**	**1.861**	**8,1**	**10,7**
Servizi alle imprese (l-n)	1.383	6,0	11,4	**Forze armate**	**69**	**0,3**	**1,0**
Servizi generali della PA	1.120	4,9	5,0	**Totale**	**22.916**	**100,0**	**100,0**
Istruzione e sanità (p-q)	2.549	11,1	15,5				
Altri servizi collettivi e personali (r-u)	4.189	18,3	7,4				
Totale (al netto MR)	**22.916**	**100,0**	**100,0**				

Le differenze Plus-Rcfl sono diminuite sensibilmente rispetto a quelle registrate in occasione della rilevazione Plus 2021. Tale miglioramento è dovuto al riavvicinamento, a partire dagli inizi 2022, alle condizioni lavorative pre-pandemiche degli occupati. Il ricorso alle interviste telefoniche su telefono fisso, caratteristica della rilevazione Plus, aveva determinato, durante il periodo di maggiore diffusione dei contagi (2020-2021), una elevata sovrastima delle tipologie di lavoratori che maggiormente hanno fatto ricorso al lavoro da casa, quali i lavoratori nei servizi e quelli occupati nelle professioni esecutive. Le stime Plus 2022 per questi aggregati hanno beneficiato del ritorno alla normalità ma, d'altro canto, è emerso un problema di confrontabilità rispetto alle stime del 2021.

3. Il peso longitudinale

Come già ricordato, l'indagine PLUS è una delle poche fonti informative sul mercato del lavoro italiano in grado di consentire analisi di tipo panel. Per ogni coppia di anni sono stati definiti i pesi di riporto all'universo longitudinale, ottenuti sulla base della stessa metodologia illustrata per il caso cross-section e con l'aggiunta di poche specifiche variazioni.

Il trattamento statistico implementato si basa sullo studio del sottostante modello di mancata risposta panel e di alcuni flussi tra condizioni di interesse primario per l'indagine, quali quello tra disoccupazione e occupazione, da lavoratore "standard" a "non standard", da part-time a full-time, ecc.

Attraverso l'utilizzo di algoritmi di classificazione non parametrica - Classification tree analysis (C&RT) - si individuano delle sottopopolazioni omogenee rispetto ad alcune caratteristiche individuali, rispetto alle quali specificare i totali noti da imporre come vincoli di calibrazione longitudinale.

Il peso panel finale è ottenuto a partire dal peso sezionale corrispondente all'anno di inizio periodo t_0 e applicando un correttore di calibrazione che da una parte tiene conto

degli effetti di attrito, dall'altra sintetizza molta dell'informazione utile ai fini della produzione delle principali stime longitudinali. È interessante notare come la selezione endogena (da modello) delle variabili utili ai fini della spiegazione della mancata risposta panel siano rimaste pressoché invariate nel tempo. Per quanto riguarda, invece, lo studio delle transizioni notevoli si è preferito svincolare gli algoritmi di classificazione dai risultati precedenti, garantendo "omogeneità metodologica" piuttosto che una più stringente stabilità nella definizione delle sottopopolazioni di riferimento.

Bibliografia

Deville J.C., Särndal C.E. (1992), Calibration Estimators in Survey Sampling JASA, 376-382.

Särndal C.E., Lundström, S. 2005. Estimation in Surveys with Nonresponse, Chichester, UK: John Wiley & Sons.

Valliant R., Dorfman A., Royall R.M. (2000), Finite Population Sampling and Inference: A Prediction Approach, New York: John Wiley.

Corsetti G., Mandrone E. (2012), Isfol Plus Survey, in Labour Economics: Plus empirical studies, Temi&Ricerche, ISFOL, 2012.

Corsetti G., Giammatteo M., Martini M. (2010), Monitoring process and non-sampling errors control in PLUS sample survey, Atti della conferenza Q2010 – European Conference on Quality in Official Satistics (Helsinki 4-6 maggio 2010).

Corsetti G., Mandrone E., Spizzichino A. (2015), L'indagine Isfol-Plus, Rivista Italiana di Economia Demografia e Statistica, Volume LXIX, n.1 Gennaio-Marzo 2015.