ISTITUTO NAZIONALE DI STATISTICA

IMPIANTO STATISTICO-METODOLOGICO DELL'INDAGINE DI SIEROPREVALENZA SUL *SARS-COV-2*

Istat

SOMMARIO

EXECUTIVE SUMMARY[1]

Overview e obiettivi

- In considerazione della necessità di disporre con urgenza di studi epidemiologici e statistiche affidabili e complete sullo stato immunitario della popolazione, indispensabili per garantire la protezione dall'emergenza sanitaria inerente l'infezione da virus *SARS-CoV-2*, il Ministero della salute, su indicazione e con l'approvazione della proposta metodologica da parte del Comitato Tecnico Scientifico, ha promosso un'indagine di sieroprevalenza della popolazione, e ne ha condiviso la titolarità con Istat, nell'ambito delle rispettive competenze sanitarie e statistiche.
- Obiettivo principale dello studio è valutare la risposta anticorpale raggiunta a qualche mese dall'inizio conclamato della pandemia nei confronti di *SARS-CoV-2* e le differenze per fascia d'età, sesso, regione di appartenenza, attività economica, e altri fattori di rischio, testando un campione rappresentativo della popolazione per la presenza di anticorpi specifici anti-*SARS-CoV-2* nel siero e determinare la frazione di infezioni asintomatiche o subcliniche. Lo studio ha quindi mirato a: (*i*) valutare il tasso di sieroprevalenza per *SARS-CoV-2* nella popolazione; (*ii*) valutare lo sviluppo della risposta anticorpale a seguito della prima ondata di pandemia e il periodo successivo; (*iii*) disporre di una banca biologica di popolazione per ulteriori valutazioni.

Rilevazione

La realizzazione della rilevazione ha visto tre successive fasi. Nella prima fase si è proceduto alla verifica telefonica della disponibilità delle unità campionate all'effettuazione delle analisi sierologiche, alla somministrazione per via telefonica di un breve questionario e alla definizione di un appuntamento presso un centro prelievo. Nella seconda fase, le unità campionate sono state sottoposte a un prelievo ematico finalizzato alla ricerca di anticorpi specifici anti-*SARS-CoV-2*. La terza fase ha riguardato la refertazione dell'esame, la trasmissione del relativo esito e la consegna dei campioni raccolti alla banca biologica dell'Istituto Nazionale Malattie Infettive «L.Spallanzani».

Errori non campionari

Sia in fase di progettazione che di realizzazione della rilevazione sono state adottate tutte le misure necessarie a contenere gli errori non campionari e a monitorare e gestire le criticità emerse durante il lavoro sul campo.

1 Si ringraziano la Prof.ssa Monica Pratesi dell'Università di Pisa e la Prof.ssa Maria Giovanna Ranalli dell'Università di Perugia per i preziosi suggerimenti e i commenti forniti.

- **Copertura.** La prima fase della rilevazione ha previsto il contatto delle unità campionate per via telefonica. Per limitare le difficoltà provenienti dalla mancanza in Italia di un registro unico dei recapiti telefonici (fissi o mobili) della popolazione, si è prevista per norma l'obbligatorietà per le principali compagnie telefoniche di rilasciare i recapiti dei loro utenti facenti parte del campione. Dopo un processo che si è protratto fino alle fasi finali della rilevazione e che ha coinvolto anche le Regioni e le Province autonome, è stato possibile acquisire i recapiti telefonici del 95,9% del campione estratto.
- **Questionario.** L'intervista è stata realizzata in modalità *Computer assisted*. I quesiti condivisi con il Comitato Tecnico Scientifico consentono di delineare la condizione socioeconomica dell'intervistato, la sintomatologia presente e di rilevare alcuni tra i principali fattori di rischio. In particolare, per la condizione lavorativa sono state rilevate le seguenti informazioni:
  - Svolgimento di ore di lavoro retribuito o presso l'azienda di un familiare nella settimana precedente l'intervista;
  - Condizionale occupazionale (in caso di non svolgimento di ore di lavoro nella settimana precedente);
  - Motivo della non effettuazione di ore di lavoro nella settimana precedente (per gli occupati);
  - Posizione nella professione (per gli occupati).

  Sugli aspetti sanitari sono state rilevate le seguenti informazioni:
- diagnosi di *SARS-CoV-2* e relativa data;
- contatti con persone infette, inclusi il momento di esposizione al contatto (ultimi 14 giorni o prima) e la relazione con la persona infetta (familiare convivente, non convivente, collega, paziente, etc.).
- presenza dei principali sintomi associati a patologie da *SARS-CoV-2* e il periodo della loro manifestazione (ultimi 14 giorni o prima);
- presenza di patologie croniche.

  Sono stati indagati inoltre alcuni comportamenti ritenuti dal CTS rilevanti non solo per il diverso rischio di esposizione al contagio (viaggi all'estero a partire dal mese di febbraio 2020 e paese visitato), ma anche per la necessità di indagare meglio i nessi tra esposizione al rischio, risposta anticorpale e determinati stili di vita (fumo, obesità) o terapie farmacologiche (vaccino antinfluenzale, farmaci antitumorali. È stato possibile identificare attraverso i quesiti sui sintomi la quota di popolazione asintomatica risultata pari al 31,3% delle persone che hanno sviluppato anticorpi.
- **Formazione.** La formazione degli intervistatori è stata curata secondo un approccio centralizzato da personale della Croce Rossa Italiana (CRI) col supporto dell'Istat. Gli operatori formati sono stati circa 1500, di questi oltre 1100 afferenti alla CRI. La partecipazione alle sessioni di training è avvenuta da remoto, attraverso l'accesso a una piattaforma predisposta appositamente da CRI.
- **Monitoraggio.** Per la prima e la seconda fase della rilevazione è stato progettato un articolato sistema di monitoraggio che ha consentito di seguirne l'andamento in tempo reale, ravvisando e intervenendo tempestivamente per risolvere le criticità riscontrate nel *field*. Ogni giorno sono stati elaborati e resi accessibili gli indicatori aggiornati al giorno precedente, consentendo di monitorare, al massimo livello di dettaglio, tutti i possibili esiti del contatto telefonico, le specifiche ragioni della mancata partecipazione alla rilevazione, il mancato rispetto degli appuntamenti fissati per il prelievo, etc. Sono stati calcolati e quotidianamente aggiornati i tassi di completezza, di caduta, di rifiuto, di irreperibilità, di inattività, di pigrizia e di appuntamenti/contatti telefonici andati a buon fine. Tutti i tassi sono stati declinati rispetto a variabili di tipo territoriale e

alle caratteristiche socio-anagrafiche delle unità campionarie (sesso ed età), al fine di gestire eventuali aree di sovraccarico della rete e eventuali differenze di "produttività" attraverso opportuni interventi correttivi (integrazione del team di operatori, ritorni formativi mirati, etc.).

- **Criticità.** Il senso di incertezza e le preoccupazioni hanno indotto molti cittadini, soprattutto nei segmenti più fragili (bambini e anziani), a non collaborare alla rilevazione. Altro elemento di criticità è stato il non pieno allineamento dei centri prelievi con le esigenze provenienti dal *field*. L'elenco iniziale è stato modificato in corso d'opera proprio per meglio rispondere alle esigenze organizzative. Ove possibile, anche con l'aiuto delle informazioni fornite dall'Istat (in termini di distanze delle unità campionate dai centri prelievo), i punti prelievo sono stati collocati in posti raggiungibili dalla gran parte delle unità. Ma non sempre ciò è stato possibile, dunque la distanza dai centri prelievo o la difficoltà di raggiungimento degli stessi è stata talvolta motivo di rinuncia alla partecipazione.

Disegno di campionamento

- **Popolazione di interesse.** Lo studio si rivolge a due differenti popolazioni di interesse – intese come l'insieme delle unità statistiche sulle quali si intende investigare – costituite rispettivamente dagli individui residenti in Italia, ivi compresi i membri permanenti delle convivenze, e i residenti in Italia occupati.
- **Parametri di interesse.** I principali parametri di interesse riguardano la frequenza assoluta e relativa degli individui in funzione del loro stato epidemiologico con riferimento a differenti sotto-popolazioni appartenenti a una specifica partizione del territorio o a specifiche sottoclassi dell'intera popolazione individuate da precise caratteristiche strutturali individuate da variabili demo-sociali (domini di studio).
- **Domini di studio.** I domini territoriali primari sono l'intero territorio nazionale; le macro aree di contagio così definite: (1) "Zona rossa", che comprende le regioni Piemonte, Lombardia, Veneto, Emilia Romagna e Marche; (2) "Resto del nord più centro", che include le regioni Valle d'Aosta, province autonome di Trento e Bolzano, Friuli, Liguria, Toscana, Umbria e Lazio; (3) "Meridione", che coincide con l'usuale ripartizione Sud e Isole; le Regioni Geografiche a cui si aggiungono le Province autonome di Bolzano e Trento. I domini territoriali secondari sono, invece, costituiti dalle Provincie, dai Sistemi Locali del Lavoro (SLL) e dalle Aziende Sanitarie Locali (ASL). I domini strutturali primari, definiti all'interno di ciascun dominio territoriale primario sono costituiti dal sesso e dalle seguenti classi di età (0-17; 18-34; 35 -49; 50-59; 60-69; 70 e più). Inoltre, per quanto riguarda la popolazione degli individui residenti occupati, oltre ai domini strutturali appena elencati, costituisce un dominio di studio anche l'Attività Economica (o ATECO) raggruppata in 4 classi: "Occupati sospesi" (D.P.C.M. 22/03/2020), "Occupati non sospesi Altro", "Occupati non sospesi della Pubblica Amministrazione e Istruzione", "Occupati non sospesi della Sanità". Per questa popolazione i domini strutturali secondari sono rappresentati dall'incrocio delle classi di età per le classi ATECO.
- **Dimensione del campione.** Tenendo conto della natura sia delle variabili *target* sia degli obiettivi di indagine, è stata condivisa la scelta finale adottata di non ricorrere alle sostituzioni al fine di un forte contenimento della distorsione connessa alle mancate risposte totali. Il campione obiettivo (150.000 individui) è stato sovra-dimensionato di circa il 30% (195.000 individui per il campione selezionato).
- **Piano di campionamento.** Il disegno di campionamento è a due stadi di selezione con stratificazione sia delle Unità di Primo Stadio (UPS) sia delle Unità di Secondo Stadio (USS). Le UPS sono i comuni

stratificati all'interno di ciascuna provincia in base alla loro dimensione demografica mentre le USS sono gli individui stratificati sulla base di 6 classi di età (0-17; 18-34; 35-49; 50-59; 60-69 70-+), sesso e 4 macro-aggregazioni dell'attività economica (non occupati, occupati non sospesi del comparto PA e istruzione, occupati non sospesi del comparto sanità, occupati non sospesi di altri comparti, occupati sospesi). Lo studio del campione è stato effettuato sulla base della stima delle prevalenze a livello provinciale fornite dall'ISS ad aprile 2020, con scelte conservative (intese come ipotesi di prevalenze più basse).

- **Allocazione.** L'allocazione del campione di individui e di comuni tra le varie regioni è stata determinata adottando la metodologia di allocazione ottima multivariata e multidominio per disegni stratificati a due stadi.
- **Caratteristiche del campione estratto.** Lo schema a due stadi ha prodotto una buona dispersione spaziale del campione estratto sul territorio, in virtù del fatto che è stato selezionato un numero rilevante di comuni (circa 2.000). È stato inoltre assicurato che tutte le Aziende Sanitarie siano rappresentate nel campione selezionato, e che quasi tutti i 610 Sistemi Locali del Lavoro siano inclusi nel campione stesso (ad eccezione di 82). I risultati dell'allocazione e successiva selezione hanno mostrato, in sintesi, una buona rappresentazione dei territori sub-regionali italiani, in rapporto alle prevalenze stimate e agli errori pianificati, e una soddisfacente copertura del campione a livello comunale, di Aziende Sanitarie e anche di SLL.

Procedura di stima

La procedura di riporto all'universo ha previsto due fasi successive: nella prima, il peso diretto di ciascuna unità rispondente, è stato modificato attraverso un processo di correzione per mancata risposta totale, per rappresentare anche le unità non rispondenti; nella seconda, si è definito il sistema dei pesi finali attraverso un processo di calibrazione dei pesi ottenuti alla prima fase.

- **Mancata risposta.** La percentuale dei rispondenti all'indagine a livello nazionale è pari a circa il 38% del campione iniziale. Inoltre, non tutti i rispondenti all'indagine si sono sottoposti al test sierologico, variabile *target* dello studio epidemiologico, ma soltanto circa il 34% dell'intero campione. Complessivamente, circa 66 mila individui hanno risposto all'indagine e si sono anche sottoposti e hanno avuto un esito al test sierologico.
- **Modello per la non risposta.** La probabilità di risposta è stata stimata attraverso un modello logistico con i seguenti predittori:
  - 21 regioni geografiche (Bolzano e Trento sono state trattate distintamente);
  - 6 tipologie comunali (città metropolitana; corona dell'area metropolitana; minore di 2000 abitanti; tra 2000 e 10000 abitanti; tra 10000 e 50000 abitanti; oltre 50000 abitanti);
  - sesso;
  - sei classi d'età (0-17; 18-34; 35-49; 50-59; 60-69; 70+);
  - quattro classificazioni dello stato ATECO (occupati sospesi, occupati non sospesi PA + Istruzione, occupati non sospesi sanità, altri occupati non sospesi, non occupati);
  - 8 modalità del titolo di studio (Analfabeti, Alfabeti privi di titolo di studi, Licenza di scuola elementare, Licenza di scuola media inferiore, Diploma di scuola secondaria superiore, Laurea o Diploma accademico di I livello, Laurea magistrale/specialistica o Diploma accademico II livello, Dottorato di ricerca)

- tasso di positività comunali, stimato sulla base dei contagi cumulati dall'inizio della pandemia a maggio (previsioni fornite dall'Istituto Superiore di Sanità);
- differenza percentuale dei tassi di mortalità comunali rispetto allo stesso periodo dell'anno precedente;
- numero di tentativi di contatto;
- panel (1 unità appartenente al panel, 0 altrimenti).

• **Fattori di aggiustamento per mancata risposta.** Le probabilità individuali predette, tramite il modello descritto, sono state utilizzate per la costruzione dei fattori di aggiustamento della mancata risposta totale indirettamente: le probabilità individuali predette sono state utilizzate per la costruzione di strati o celle di aggiustamento (response propensity stratification). La costruzione delle celle è stata effettuata con la tecnica dei quintili delle probabilità predette e il fattore correttivo della mancata risposta totale è stato calcolato in ogni cella come inverso del tasso di risposta.

• **Calibrazione.** Nella calibrazione si è tenuto conto dei seguenti totali di popolazione ricavati dal registro base degli individui (RBI):

- distribuzione regionale della popolazione per sesso e classe d'età (0-17; 18-34; 35-49; 50 59; 60-69; 70+);
- distribuzione regionale della popolazione per sesso e stato della ATECO (occupati sospesi; occupati non sospesi, altro; occupati non sospesi, PA + Istruzione; occupati non sospesi, sanità; non occupati);
- distribuzione provinciale della popolazione;
- distribuzione regionale della popolazione per cittadinanza;
- distribuzione della popolazione per ripartizione (Nord-Ovest, Nord-Est, Centro e Mezzogiorno), classe d'età (0-17; 18-34; 35-49; 50 59; 60-69; 70+) e titolo di studio (4 livelli).

Il numero complessivo di totali (vincoli) considerati è 11.100. La funzione di distanza utilizzata è la funzione logaritmica troncata con estremi fissati a 0.74 e 6.50.

• **Errori campionari.** Si fornisce una presentazione sintetica degli errori relativi basata sul metodo dei modelli regressivi basata sulla determinazione di una funzione matematica che mette in relazione ciascuna stima con il proprio errore di campionamento.

PARTE I

LA METODOLOGIA DELL'INDAGINE DI SIEROPREVALENZA SUL *SARS-COV-2*: ASPETTI PROGETTUALI E ORGANIZZATIVI[2]

2 Maria Clelia Romano, Linda Laura Sabbadini.

1. Introduzione

In considerazione della necessità di disporre con urgenza di studi epidemiologici e statistiche affidabili e complete sullo stato immunitario della popolazione, indispensabili per garantire la protezione dall'emergenza sanitaria inerente l'infezione da virus *SARS-CoV-2*, il Ministero della salute, su indicazione e con l'approvazione della proposta metodologica da parte del Comitato Tecnico Scientifico[3], ha promosso un'indagine di sieroprevalenza della popolazione, e ne ha condiviso la titolarità con Istat, nell'ambito delle rispettive competenze sanitarie e statistiche. La realizzazione dello studio è stata prevista dal decreto legge 10 maggio 2020, n. 30, ulteriormente specificata nel protocollo metodologico dal protocollo approvato dal Comitato Tecnico Scientifico[4].

Obiettivo principale dello studio è valutare la risposta anticorpale raggiunta a qualche mese dall'inizio conclamato della pandemia nei confronti di *SARS-CoV-2* e le differenze tra le diverse fasce d'età, sesso, regione di appartenenza, attività economica, e altri fattori di rischio, testando un campione rappresentativo della popolazione per la presenza di anticorpi specifici anti-*SARS-CoV-2* nel siero e determinare la frazione di infezioni asintomatiche o subcliniche.

Lo studio mira quindi a:

- valutare il tasso di sieroprevalenza per *SARS-CoV-2* nella popolazione;
- valutare lo sviluppo della risposta anticorpale a seguito della prima ondata di pandemia e il periodo successivo;
- disporre di una banca biologica di popolazione per ulteriori valutazioni.

2. Il disegno dell'indagine

In ottemperanza a quanto previsto dalla norma, il piano di campionamento (vedi parte II) è stato disegnato in modo da permettere la stima *cross-section* della sieroprevalenza, ritorni di indagine su particolari *target* di interesse, da definire in un secondo momento, e ritorni longitudinali su un sotto-campione rappresentativo dell'intera popolazione[5].

L'impianto metodologico ha previsto la realizzazione della rilevazione in tre successive fasi[6]. La prima fase comprende: a) la verifica telefonica della disponibilità delle unità campionate all'effettuazione delle analisi sierologiche, b) la somministrazione sempre attraverso il telefono di un breve questionario e c) la definizione di un appuntamento presso un centro prelievo. Nella successiva seconda fase, le unità campionate sono state sottoposte a un prelievo ematico finalizzato alla ricerca di anticorpi specifici anti-*SARS-CoV-2*. La terza fase riguarda la refertazione dell'esame, la trasmissione del relativo esito e la consegna dei campioni raccolti alla banca biologica dell'Istituto Nazionale Malattie Infettive «L.Spallanzani».

3 Secondo quanto previsto dall'articolo 2, comma 1, dell'ordinanza del Capo del Dipartimento della protezione civile n. 630 del 3 febbraio 2020.

4 Di cui all'articolo 2 dell'ordinanza del Capo del Dipartimento della protezione civile 3 febbraio 2020, n. 630.

5 Nell'ambito del campione generale è stato inoltre individuato un sotto-campione (definito anticipatorio) di circa ventimila unità, con le medesime caratteristiche del campione totale e che avrebbe dovuto essere utilizzato sia per un'anticipazione del rilascio delle stime sia per ritorni successivi di studio. Nei fatti, le difficoltà organizzative iniziali hanno fatto sì che questo obiettivo non fosse perseguibile: sebbene si fosse data priorità di lavorazione alle unità campionarie incluse in tale sotto-campione si è ritenuto infatti di non rilasciare le stime calcolate solo su tale sotto-insieme.

6 Per ulteriori dettagli si rimanda al Protocollo metodologico per un'indagine di sieroprevalenza sul *SARS-CoV-2* condotta dal Ministero della salute e dall'Istat. https://www.istat.it/it/files//2020/05/Protocollo-approvato.pdf.

Dal punto di vista organizzativo, il disegno di indagine ha previsto, in linea con quanto indicato dalla norma, il coinvolgimento di una molteplicità di figure, a supporto dei titolari della rilevazione (Istat e Ministero della Salute), nella realizzazione delle varie attività delle tre fasi. L'organizzazione del lavoro sul campo è stata curata dalla Croce Rossa Italiana e dalle Regioni e Province Autonome. In particolare la Croce Rossa ha svolto un ruolo cruciale gestendo le interviste telefoniche della fase 1, collaborando con le Regioni per lo svolgimento della fase 2 dei prelievi e gestendo il numero verde per tutta la durata della rilevazione. In particolare, nella fase 1 la CRI ha operato sia attraverso un centro di coordinamento nazionale con funzioni anche di numero verde, sia attraverso i Comitati regionali per la gestione delle interviste sul territorio. Nella fase 2 sono intervenuti i Comitati territoriali per l'effettuazione dei prelievi (anche domiciliari) e di tutte le attività previste a essi connessi, compreso il trasporto (fase 3) dei campioni ematici. Gli operatori impegnati nella rilevazione sono stati in prevalenza volontari CRI, supportati nelle settimane conclusive della rilevazione da operatori specializzati afferenti ai principali Contact Center privati nazionali. Nella fase 2 ovviamente hanno svolto un ruolo cruciale i centri prelievo (afferenti alle Regioni o alla CRI) e, nella fase 3, i laboratori di analisi.

Anche i medici di medicina generale e ai pediatri di libera scelta hanno giocato un ruolo molto importante nel corso della rilevazione. Dalle Regioni hanno ricevuto comunicazione dei nominativi dei soggetti campionati, in modo da poter informare preventivamente e sensibilizzare i propri assistiti alla collaborazione che sarebbe stata loro richiesta. Nel corso della rilevazione hanno così potuto rispondere alle esigenze di rassicurazione da parte dei loro assistiti più titubanti a prendere parte all'indagine, fornendo il loro importante contributo al risultato raggiunto.

Le Regioni e le Province autonome hanno svolto un ruolo chiave nelle varie fasi della rilevazione: segnaliamo in particolare l'individuazione dei Centri prelievo dove potersi recare nella fase 2, la sensibilizzazione attraverso i medici e i pediatri delle unità campione, il recupero di ulteriori recapiti telefonici, il monitoraggio della rilevazione sul territorio di loro competenza.

Tutte le attività relative al lavoro sul campo e alla trasmissione dei dati tra i vari soggetti della rete, sono state gestite attraverso una piattaforma informatica implementata *ad hoc* e in tempi strettissimi dal Ministero della Salute. Attraverso tale piattaforma sono state svolte le attività relative alla gestione degli esiti dei contatti telefonici, all'acquisizione dei dati tramite questionario elettronico, alla condivisione degli indicatori per il monitoraggio del *fieldwork*, allo scambio dei dati tra i vari soggetti coinvolti, etc. La necessità di progettarla e realizzarla in tempi serrati e a ridosso dell'avvio della rilevazione ha imposto una selezione delle funzionalità la cui implementazione fosse prioritaria e possibile nei tempi dati. Tuttavia anche a rilevazione avviata è continuato lo sviluppo evolutivo della piattaforma, in modo da mettere a disposizione ulteriori funzioni, utili a supportare il lavoro degli operatori sul campo (ricerca nominativi, gestione dei contatti, gestione degli esiti, modalità di accesso al questionario, etc.). Per l'accesso alla piattaforma sono state previste utenze nominative e profilate per le diverse funzioni che gli utenti coinvolti nell'esecuzione dell'indagine (call center, unità di prelievo, laboratori, coordinatori regionali e nazionali, Istat, Ministero della salute) erano chiamati a svolgere.

La macchina organizzativa della rilevazione è stata dunque di eccezionale complessità, come eccezionalmente tempestive sono state tutte le attività progettate e implementate per portare a termine una rilevazione di tale importanza per il Paese. Ciononostante grazie a una stretta collaborazione tra tutti i soggetti coinvolti all'interno e all'esterno dell'Istituto è stato possibile disegnare e avviare l'indagine in tempi veramente stretti.

Proprio l'esigenza di condurre la rilevazione in condizioni emergenziali, ha portato a considerare la tecnica telefonica come la più idonea a contattare le unità campionate per raggiungere l'obiettivo, nei tempi dati. Essa, tuttavia, come noto, presenta dei limiti, per la mancanza in Italia di un registro unico dei recapiti telefonici (fissi o mobili) della popolazione. Per tale ragione, già in fase progettuale si è ovviato a questa difficoltà prevedendo per norma l'obbligatorietà per le compagnie telefoniche di rilasciare i recapiti dei loro utenti facenti parte del campione. Di fatto, una volta estratto l'elenco delle unità campione, la prima attività propedeutica al lavoro sul campo (realizzata in pochissimi giorni, a stretto ridosso dell'avvio del lavoro sul campo) ha riguardato proprio la ricerca dei recapiti telefonici, che è stata effettuata sia recuperando i numeri presenti negli archivi già disponibili in Istat (prevalentemente di telefonia fissa), sia attraverso i principali fornitori di telefonia (Tim, Vodafone, Windtre, Iliad, Fastweb) che, in base ad accordi stipulati con il Ministero della Salute, hanno fornito i recapiti di telefonia mobile dei loro utenti rientranti nel campione.

Al termine di questo processo, è stato possibile assegnare un recapito telefonico all'87,6% del campione, mentre una parte non irrilevante del campione (circa il 12,4%) restava ancora senza recapito telefonico e pertanto non contattabile. Di conseguenza si è avviata una intensa collaborazione con le Regioni e le Province autonome chiamate a integrare, anche a rilevazione avviata, le informazioni acquisite nelle modalità precedentemente descritte, attraverso gli archivi a loro disposizione. Al termine di questo intenso lavoro, che si è protratto fino alle fasi finali della rilevazione, per cercare di contenere al massimo la non raggiungibilità delle unità selezionate, solo il 4,1% del campione è rimasto senza recapito e dunque non è stato di fatto contattabile. Si è registrata tuttavia una marcata eterogeneità territoriale: si va dalla copertura totale del Veneto alla mancanza di recapiti per il 13,2% delle unità campionarie in Campania. Dalla tavola seguente risulta evidente l'apporto della rete territoriale attivata per il recupero dei recapiti mancanti e il suo contributo in ciascuna regione per la riduzione delle unità non contattabili.

Tavola 2.1 - Unità campionarie senza recapito telefonico a inizio e fine rilevazione (valori percentuali)

	Inizio rilevazione	Fine rilevazione
PIEMONTE	7,4	5,0
VALLE D'AOSTA	0,5	0,4
LOMBARDIA	15,6	3,9
PROV. AUTON. BOLZANO	4,3	4,2
PROV. AUTON. TRENTO	22,9	8,2
VENETO	9,6	0,0
FRIULI VENEZIA GIULIA	6,0	4,1
LIGURIA	3,4	2,2
EMILIA ROMAGNA	23,7	2,6
TOSCANA	22,7	1,4
UMBRIA	1,4	1,1
MARCHE	1,5	1,3
LAZIO	6,9	6,3
ABRUZZO	4,0	3,8
MOLISE	5,8	5,6
CAMPANIA	22,4	13,2
PUGLIA	0,0	0,0
BASILICATA	8,4	4,5
CALABRIA	16,6	4,6
SICILIA	21,5	7,6
SARDEGNA	23,0	2,8
ITALIA	12,4	4,1

Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

3. Strategie per la prevenzione degli errori non campionari

Nella realizzazione di una rilevazione statistica particolare attenzione va dedicata alla definizione di strategie e criteri atti a prevenire gli errori non campionari, ben sapendo che essi si possono generare in ciascuna delle fasi del processo (dalla progettazione, alla diffusione) e inficiando fortemente la qualità dei dati prodotti. Il lavoro sul campo è un momento estremamente complesso, di conseguenza, molte delle attività dedicate alla prevenzione degli errori non campionari sono concentrate in questa fase. L'acquisizione dei dati costituisce, senza dubbio, uno dei momenti più delicati: gli eventuali errori commessi in questa fase, infatti, difficilmente possono essere sanati senza generare distorsioni nei risultati finali. Ciò è ancora più vero in processi di elevata complessità come l'"Indagine di sieroprevalenza sul *SARS-CoV-2*", il cui impianto è stato progettato nei minimi dettagli, proprio al fine di ridurre gli errori e garantire l'acquisizione di dati di qualità, in linea con gli standard della statistica ufficiale.

L'azione di tutti i soggetti chiamati a collaborare per la realizzazione della rilevazione è stata finalizzata a massimizzare la partecipazione delle unità campionate, in un contesto emergenziale che ha complicato il lavoro di tutti soprattutto per le perplessità dei cittadini chiamati a collaborare e per i tempi serratissimi in cui tutte le operazioni dovevano essere svolte.

La complessità dell'impianto di indagine, la numerosità degli attori coinvolti e la costante attenzione a contenere il rischio di errore non campionario in ciascuna delle fasi del processo di acquisizione dei dati ha indotto ad adottare una serie di misure che hanno consentito di garantire la qualità dei dati raccolti e che hanno riguardato in particolare la formazione accurata della rete di rilevazione, la progettazione di un questionario sintetico e chiaro, la supervisione costante dell'andamento della rilevazione. A seguire sono riportate le azioni messe in atto in ciascuno di questi ambiti per garantire un elevato livello qualitativo delle stime prodotte e contenere al massimo gli errori non campionari. Non va trascurata però l'importante funzione svolta a tal fine anche dalla campagna di comunicazione integrata e congiunta tra Ministero della Salute e l'Istat sia a livello nazionale, sia a livello territoriale. La campagna diffusa sui canali tradizionali e sui new media per raggiungere il maggiore numero di persone possibile, oltre a illustrare gli obiettivi dell'indagine, mirava a sottolineare l'importanza della partecipazione del singolo, per la propria utilità personale e l'utilità della collettività e a dare massima diffusione dei numeri verdi attivati a supporto dei cittadini.

3.1 Il disegno del questionario

Nelle indagini statistiche il questionario è tra le principali cause di errori non campionari[7], per tale ragione la sua progettazione è stata realizzata con molta cura e attenzione, sia per quanto riguarda il wording che la sequenza dei quesiti. L'intervista è stata realizzata in modalità *Computer assisted*, si è pertanto basata su un questionario elettronico che guida automaticamente l'intervistatore nel condurre l'intervista e consente di prevedere controlli già in fase di acquisizione, riducendo il rischio di errore. Ma molto si è fatto anche in fase di formazione, addestrando gli intervistatori a prestare la massima attenzione al contenimento delle mancate risposte parziali, della non corretta interpretazione dei quesiti, così come degli errori di digitazione. Importante per la verifica del corretto funzionamento del questionario elettronico è stata la fase di test funzionali effettuati a stretto ridosso dell'avvio della rilevazione, nel corso della quale è stato possibile correggere tempestivamente errori di implementazione.

7 Sul tema si veda Biemer *et alii*, 1991.

La definizione dei contenuti informativi ha visto un'intensa collaborazione tra l'Istat con il Comitato tecnico scientifico per la definizione della struttura complessiva del questionario, la formulazione dei quesiti e delle relative modalità di risposta, nell'intento di rendere più scorrevole possibile la sua somministrazione da parte degli operatori. La progettazione del questionario è stata guidata dall'intento di contenere il *burden* sul rispondente, aspetto indispensabile soprattutto in una rilevazione come quella in oggetto, in cui l'intervista rappresenta solo il primo momento del coinvolgimento richiesto, per di più meno oneroso, rispetto al successivo prelievo ematico della fase 2. Di conseguenza è stata effettuata una selezione ragionata di tutti i contenuti informativi che pure sarebbe stato utile rilevare, proprio al fine di non abusare della disponibilità delle unità campionate in un'indagine di tale complessità.

La parte anagrafica che consente di accertare e confermare l'identità della persona contattata è costituita da dati precaricati in base alle informazioni presenti nei registri dell'Istituto (vedi par. 4, parte II). Particolare importanza riveste la verifica dell'indirizzo di domicilio da cui sarebbe conseguito l'indirizzamento ai centri di prelievo. In caso di non corrispondenza tra luogo di residenza e luogo di domicilio, o anche di presenza in comuni/regioni diversi da quello di residenza per tutto il periodo di rilevazione, è stato acquisito l'indirizzo in cui l'unità campionata era dimorante, in modo da dare ai comitati regionali CRI la possibilità di individuare i centri prelievi più prossimi.

Ciò ha ovviamente richiesto una forte collaborazione tra tutti i Comitati Regionali per gestire i flussi in entrata e in uscita dal proprio ambito di competenza e di assegnazione (anche con l'ausilio del Contact Center centrale) a centri prelievi operanti in regioni diverse da quelle di residenza. La tavola 3.1 riporta la distribuzione a livello regionale di queste casistiche, che sebbene riguardino complessivamente un numero esiguo di casi (2059, pari all'1,1% del campione) hanno rappresentato un'ulteriore criticità da gestire nelle regioni in cui il numero di prelievi da effettuare è cresciuto rispetto a quanto inizialmente pianificato (per es. Lazio ed Emilia Romagna).

Particolare attenzione ai quesiti dedicati a verificare eleggibilità e identità delle unità contattate è stata necessaria nel caso in cui si trattasse di minori. In questo caso, infatti, a prescindere da chi rispondesse al recapito telefonico è stato necessario parlare con un genitore o con chi ne faceva le veci, per la corretta gestione sia della fase 1 che della fase 2 della rilevazione.

Verificata l'identità del rispondente sono stati somministrati i quesiti condivisi con il CTS e che consentono di delineare la condizione socioeconomica dell'intervistato, la sintomatologia presente e di rilevare alcuni tra i principali fattori di rischio, i cui nessi con la patologia *SARS-CoV-2* sono stati messi in evidenza da diversi studi. In particolare, i primi quattro quesiti mirano a individuare la condizione lavorativa delle unità campionarie con quesiti analoghi a quelli utilizzati nelle indagini campionarie sulle famiglie, ma semplificati per contenere, come si è detto, lunghezza del questionario e *burden* sul rispondente. Sono state pertanto rilevate le seguenti informazioni:

- Svolgimento di ore di lavoro retribuito o presso l'azienda di un familiare nella settimana precedente l'intervista;
- Condizionale occupazionale (in caso di non svolgimento di ore di lavoro nella settimana precedente);
- Motivo della non effettuazione di ore di lavoro nella settimana precedente (per gli occupati);
- Posizione nella professione (per gli occupati).

Tavola 3.1 - Individui campione domiciliati in regione diversa da quella di residenza per regione di residenza e regione di domicilio (valori assoluti e percentuali)

	Residenti domiciliati in altre regioni	Domiciliati ma residenti in altre regioni	Differenza rispetto a prelievi pianificati	
	v.a	v.a	v.a	%
PIEMONTE	50	170	120	1,2
VALLE D'AOSTA	77	12	-65	-1,4
LOMBARDIA	310	344	34	0,1
PROV. AUTON. BOLZANO	34	73	39	0,9
PROV. AUTON. TRENTO	31	11	-20	-0,5
VENETO	89	30	-59	-0,4
FRIULI VENEZIA GIULIA	53	110	57	0,7
LIGURIA	55	38	-17	-0,2
EMILIA ROMAGNA	90	240	150	1,2
TOSCANA	63	158	95	1,2
UMBRIA	79	26	-53	-1,0
MARCHE	73	51	-22	-0,3
LAZIO	93	306	213	1,9
ABRUZZO	72	69	-3	-0,1
MOLISE	148	5	-143	-2,7
CAMPANIA	93	140	47	0,4
PUGLIA	106	113	7	0,1
BASILICATA	170	14	-156	-2,1
CALABRIA	152	58	-94	-1,2
SICILIA	123	61	-62	-0,6
SARDEGNA	98	30	-68	-0,9

Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

Sugli aspetti squisitamente sanitari sono stati rilevate le seguenti informazioni:

- diagnosi di *SARS-CoV-2* e relativa data;
- contatti con persone infette, inclusi il momento di esposizione al contatto (ultimi 14 giorni o prima) e la relazione con la persona infetta (familiare convivente, non convivente, collega, paziente, etc.).
- presenza dei principali sintomi associati a patologie da *SARS-CoV-2* e il periodo della loro manifestazione (ultimi 14 giorni o prima) (Figura 3.1);
- presenza di patologie croniche (Figura 3.2).

Sono stati indagati inoltre alcuni comportamenti ritenuti dal CTS rilevanti non solo per il diverso rischio di esposizione al contagio (viaggi all'estero a partire dal mese di febbraio 2020 e paese visitato), ma anche per la necessità di indagare meglio i nessi tra esposizione al rischio, risposta anticorpale e determinati stili di vita (fumo, obesità) o terapie farmacologiche (vaccino antinfluenzale, farmaci antitumorali), su cui a livello internazionale non sono stati prodotti risultati dirimenti.

Figura 3.1 - A partire dal 1 febbraio 2020, ha avuto uno dei seguenti sintomi? Quando?

	SI	NO	Negli ultimi 14 giorni	Prima degli ultimi 14 giorni
a. Dolori ossei/muscolari	1□	2□	1□	2□
b. Senso di stanchezza	1□	2□	1□	2□
c. Mal di testa	1□	2□	1□	2□
d. Congiuntivite	1□	2□	1□	2□
e. Diarrea	1□	2□	1□	2□
f. Difficoltà a respirare	1□	2□	1□	2□
g. Dolori Addominali	1□	2□	1□	2□
h. Perdita/alterazione del gusto	1□	2□	1□	2□
i. Perdita/alterazione dell'olfatto	1□	2□	1□	2□
l. Mal di gola	1□	2□	1□	2□
m. Febbre	1□	2□	1□	2□
n. Tosse	1□	2□	1□	2□
o. Sindrome di tipo influenzale	1□	2□	1□	2□
p. Nausea/Vomito	1□	2□	1□	2□
q. Confusione mentale	1□	2□	1□	2□

Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

Figura 3.2 - Presenta qualcuna delle seguenti malattie?

	Sì	No	Non sa
a. Diabete	1□	2□	3□
b. Malattie cardiovascolari	1□	2□	3□
c. Deficit immunitari	1□	2□	3□
d. Malattie respiratorie croniche	1□	2□	3□
e. Rinite allergica	1□	2□	3□
f. Malattie renali	1□	2□	3□
g. Malattie autoimmuni Lupus artrite reumatoide malattie croniche intestinali, sclerosi a placche	1□	2□	3□
h. Ipertensione	1□	2□	3□
i. Malattie del sangue	1□	2□	3□
l. Malattie neurologiche (Parkinson, Alzheimer)	1□	2□	3□
m. Tumori	1□	2□	3□

Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

La formulazione di alcuni quesiti è stata anche funzionale all'indirizzamento di un determinato *target* di rispondenti al prelievo a domicilio. Infatti, al verificarsi di alcune specifiche condizioni (presenza di determinati sintomi, diagnosi di *SARS-CoV-2* oppure contatti con persona affetta da *SARS-CoV-2* negli ultimi 14 giorni) era necessario, come convenuto con in CTS, predisporre il prelievo a domicilio, caldamente suggerito anche per i grandi anziani e i portatori di handicap.

L'analisi dei risultati conferma l'efficacia dei quesiti nella misurazione di variabili chiave per la lettura della sieroprevalenza. A titolo di esempio, basti ricordare che è stato possibile identificare attraverso i quesiti sui sintomi la quota di popolazione asintomatica risultata pari al 31,3% delle persone che hanno sviluppato anticorpi. Un dato che consentito di sottolineare l'importanza dell'identificazione immediata delle persone affette dall'infezione, nonché di tutti gli individui con cui, a loro volta, sono entrate in contatto. Anche i quesiti sulla presenza di malattie croniche hanno fornito dati di elevata qualità come, confermato dalla convergenza delle stime con quelle relative

ad altre indagini condotte sullo stato di salute della popolazione. Merita di essere citato anche il dato relativo alla presenza di una quota di rispondenti che pur non essendo risultati positivi al test, hanno dichiarato di avere avuto una diagnosi *COVID*. Si tratta di circa 600 rispondenti risultati negativi al test sierologico e che potrebbero rappresentare uno dei sottoinsiemi di popolazione da approfondire, secondo quanto previsto in fase progettuale, in modo da raccogliere informazioni più puntuali sulla dinamica longitudinale della risposta anticorpale.

3.2. La formazione della rete di rilevazione

La qualità di una rilevazione diretta dipende, in larga misura, dalla qualità della rete di rilevazione e dalle abilità degli attori che la costituiscono. Per tale ragione le indagini che richiedono l'esecuzione di interviste presuppongono, in fase di disegno, accurate procedure di reclutamento, selezione e formazione di tutto il personale coinvolto[8].

Indipendentemente dal fatto che debbano condurre interviste faccia a faccia o telefoniche, gli intervistatori svolgono un ruolo chiave nel motivare le unità di rilevazione a partecipare alla rilevazione e nel raccogliere dati affidabili nel corso dell'intervista. Somministrare un'intervista richiede infatti una serie di competenze, senza le quali la qualità dei dati può essere molto compromessa. L'intervistatore deve contattare l'unità di rilevazione, verificarne l'eleggibilità, spiegare lo scopo dello studio, motivarla a partecipare, porre le domande nel modo richiesto, farla sentire a proprio agio e registrare accuratamente le risposte e qualsiasi altra informazione richiesta[9]. In sintesi, il comportamento dell'intervistatore può avere un effetto diretto sulla mancata risposta e sull'errore di misura e, quindi, contribuire in modo significativo all'errore totale[10].

Esiste dunque un forte legame tra le competenze, la formazione degli intervistatori e la qualità dei dati raccolti[11]. Gli intervistatori non qualificati producono dati di qualità inferiore, con maggiori tassi di mancata risposta totale, maggiore distorsione da desiderabilità sociale e maggiore difficoltà nell'ottenere la cooperazione degli intervistati. La formazione è dunque una fase essenziale per fornire ai diversi attori gli indirizzi tecnici e comportamentali necessari a garantire loro autonomia e preparazione sufficienti per poter espletare le attività correttamente[12]. Per questo necessita di grande attenzione, tempo e capacità specifiche anche in campo didattico.

Proprio per la crucialità del ruolo degli intervistatori nel lavoro sul campo e di una loro adeguata formazione, nella progettazione della rilevazione sulla sieroprevalenza è stata prestata un'attenzione particolare alla definizione dei processi formativi, necessari a garantire l'acquisizione, da parte di tutti gli attori coinvolti, delle competenze necessarie a operare nel rispetto di elevati standard qualitativi. Attraverso la strategia formativa adottata, si è cercato, da un lato, di far assimilare i contenuti dell'indagine e sviluppare sia le capacità

8 Cfr Stiegler & Biedinger, 2016; Istat, 1989; Istat, 2001; Istat, 2020.

9 Il compito che svolgono è dunque estremamente delicato, anche perché possono influenzare le risposte attraverso le loro caratteristiche personali e i loro comportamenti, determinando quello che viene denominato "effetto intervistatore". Esso può manifestarsi in vari modi: per esempio, in alcuni casi le persone intervistate dallo stesso intervistatore tendono ad avere risposte più simili di quanto ci si aspetterebbe (Kreuter, 2008). Le stesse caratteristiche osservabili dell'intervistatore, come l'età o il genere, nonché il comportamento verbale o non verbale potrebbero influenzare il processo di risposta. Infine, anche errori sistematici nella somministrazione dell'intervista (ad es. la lettura errata delle domande) e differenze nella capacità di conquistare la cooperazione degli intervistati possono provocare un effetto intervistatore (West, Kreuter & Jaenichen, 2013; West & Olson, 2010).

10 Si vedano tra gli altri: Blom & Korbmacher, 2011; Durrant & D'Arrigo, 2014; Groves, 2005.

11 Sul tema si vedano Mohorko & Hlebec, 2015; Olson & Peytchev, 2007.

12 La rete non è costituita solo da intervistatori ma da molteplici figure impegnate nelle varie attività: gestione del numero verde, supervisione e monitoraggio degli intervistatori, coordinamento regionale e nazionale, etc. Ognuna di queste figure necessita di formazione adeguata alle attività da svolgere.

di ottenere la collaborazione delle unità campionate, per limitare il numero dei rifiuti, sia le capacità tecniche e metodologiche nella somministrazione dei quesiti (per lo più di natura sensibile). Dall'altro, gli interventi formativi hanno mirato a sviluppare sia sensibilità e capacità per interagire al meglio con gli intervistati e per gestire eventuali criticità, sia la consapevolezza della crucialità del proprio ruolo e l'importanza di svolgerlo con la massima professionalità e attenzione.

La formazione della rete di operatori della CRI e, nelle settimane finali della rilevazione, di personale specializzato di Contact Center privati è stata dunque impostata in modo da dare ai discenti tutte le informazioni per operare e gestire al meglio i contatti con le unità campione. E' stata strutturata in sessioni di formazione plenaria, di carattere prevalentemente teorico e briefing tecnici. Le prime, destinate a gruppi molto ampi di operatori, sono state finalizzate a dare informazioni di carattere generale sull'importanza della rilevazione, gli obiettivi conoscitivi, i contenuti del questionario, il corretto modo di approcciare l'intervista e interagire con i cittadini al fine di acquisire dati di elevata qualità. Particolare attenzione è stata, inoltre, data agli aspetti normativi sulla privacy, sul trattamento dei dati e sull'obbligo di risposta, elementi fondamentali, oltre che per informare l'intervistato, per mettere in condizione l'operatore di gestire tutte le domande con massima professionalità e competenza. Durante le sessioni plenarie sono state impartite indicazioni sulle modalità comportamentali da tenere durante l'intervista e con l'intervistato, così come le strategie da attivare per motivare le unità più reticenti, o per gestire l'interazione con specifici gruppi di popolazione (anziani, persone in condizioni di salute critiche, etc.). Si tratta di una fase della formazione particolarmente delicata poiché, oltre alla trasmissione di nozioni tecniche, si cerca di insegnare a gestire l'impatto psicologico negativo che i rifiuti, insieme alle tante difficoltà che possono presentarsi durante l'interazione, possono indurre negli intervistatori, soprattutto lavorando sotto pressione e con tempistiche stringenti.

Nei briefing tecnici (organizzati per gruppi più piccoli di discenti) si è passati dai contenuti teorici alla pratica, illustrando modalità di accesso alla piattaforma di acquisizione e le principali funzionalità a supporto del lavoro sul campo. Gli operatori hanno avuto modo di esercitarsi mediante interviste simulate guidate e quindi di familiarizzare con gli aspetti, tecnici e non, di gestione dell'intervista, ma anche di migliorare la propria capacità persuasiva e l'abilità nello stabilire un rapporto di fiducia con l'intervistato.

La formazione è stata curata secondo un approccio centralizzato da personale della CRI specializzato in formazione e esperti di comunicazione, col supporto dell'Istat che ha curato la presentazione dei contenuti del questionario e ha illustrato alcune specifiche strategie da adottare al fine di contenere la mancata risposta totale e parziale e favorire l'acquisizione di dati di elevata qualità. La collaborazione tra formatori diversi e la sinergia che ne è scaturita ha consentito di affrontare in modo condiviso, armonizzato sul territorio e tempestivo qualunque problematica evidenziata dagli intervistatori nel corso degli incontri[13].

È stato predisposto e fornito a tutti i partecipanti del materiale integrativo di approfondimento (slide, Faq, manuali operativi) e tutta la documentazione utile a gestire al meglio il contatto con le unità campionate e rispondere ai loro dubbi (Informativa, riferimenti normativi, etc.). Gli operatori formati sono stati circa 1.500, di questi oltre 1.100 afferenti alla CRI. La partecipazione alle sessioni di training è avvenuta (non poteva essere altrimenti dato il periodo) da remoto, attraverso l'accesso a una piattaforma predisposta appositamente da CRI. Ovviamente questo ha comportato talvolta dei rallentamenti, ma fortunatamente non sono emersi problemi tecnici di particolare rilievo e anche la partecipazione attiva, attraverso chat e interventi diretti, è stata molto soddisfacente.

13 Un approccio siffatto è particolarmente importante per evitare disomogeneità e disallineamenti a livello territoriale. A tal proposito si veda Istat (2020).

Aldilà di un primo momento di arruolamento della rete, organizzato come sinteticamente descritto, di fatto il processo di formazione è stato continuo e particolarmente oneroso e ha richiesto un investimento maggiore di quanto inizialmente preventivato. Ciò è accaduto innanzitutto perché è stato evidente dai primi giorni di rilevazione che il numero inizialmente stimato di operatori necessari non sarebbe stato sufficiente a portare a termine la rilevazione nei tempi previsti: è stato pertanto necessario integrare il bacino iniziale con ulteriori operatori, subentranti in aggiunta o in sostituzione di quanti per varie ragioni abbandonavano la rilevazione (non dobbiamo dimenticare che si trattava di volontari, spesso in congedo dal lavoro per effettuare le interviste). Fino ad arrivare alle ultime settimane, in cui sono stati reclutati operatori specializzati in attività di Contact Center operanti per varie società private, chiamati nella fase conclusiva della rilevazione, allo scopo di massimizzarne i risultati. Tutto ciò per dire che tutto il periodo di rilevazione ha richiesto un forte investimento di risorse per svolgere le attività di training degli operatori man mano subentranti.

Ma la formazione è stata continua anche perché non si è limitata al training iniziale. Come sempre accade in indagini complesse come questa, il lavoro sul campo presenta una serie di casistiche talvolta non previste che richiedono una soluzione rapida e condivisa su tutto il territorio. Di conseguenza le riunioni organizzate giornalmente con i *Focal Point* regionali della CRI (coordinatori delle reti regionali) sono state l'occasione per ritornare su molti dei temi affrontati in fase di formazione. In queste occasioni è stato possibile chiarire in particolare come gestire alcune delle criticità emerse fin dai primi giorni della rilevazione e come esitare alcune specifiche tipologie di contatto, la cui gestione non era risultata chiara a molti degli operatori formati.

Nonostante sia stato molto impegnativo, il contatto con la rete operante sul territorio ha consentito, sebbene mediata dai coordinatori territoriali, di acquisire informazioni sull'andamento della rilevazione, che non sempre è possibile desumere dagli indicatori di monitoraggio, e contribuire in maniera fattiva, suggerendo in tempo reale, strategie e azioni da adottare per risolvere le criticità emerse.

3.3. Il monitoraggio del lavoro sul campo

Nella fase del lavoro sul campo è determinante progettare dei criteri e degli strumenti di monitoraggio delle interviste che tengano conto della metodologia di indagine, del tipo di rete di rilevazione e del relativo assetto organizzativo, oltre che delle tempistiche e della popolazione di riferimento. I sistemi di monitoraggio giornaliero permettono di intervenire tempestivamente in itinere durante il lavoro sul campo per migliorare gli standard produttivi e gestire eventuali imprevisti. La progettazione del sistema degli indicatori di monitoraggio è, pertanto, determinante per rendere efficiente il controllo dell'attività degli intervistatori. Poter disporre quotidianamente di un set di informazioni utili a sorvegliare l'intero processo di produzione dei dati è fondamentale durante tutto il corso della rilevazione, per testare la comprensione da parte dei rilevatori delle istruzioni impartite durante la formazione, per individuare eventuali distorsioni nei dati raccolti e più in generale per valutare l'andamento qualitativo dell'indagine. Tale attività di monitoraggio viene effettuata attraverso lo studio accurato di un insieme di report organizzati e strutturati di indicatori statistici che consentono di supervisionare e valutare la qualità del lavoro sul campo[14].

14 La conduzione di innumerevoli ed eterogenee indagini statistiche ha consentito valutare e apprezzare l'utilità degli strumenti di controllo descritti. Il loro impiego, di volta in volta, consente di prevenire problemi nell'esecuzione delle interviste, verificare l'esattezza dei tempi e dei ritmi di esecuzione dell'indagine, appurare l'idoneità e il rispetto delle regole di gestione dei contatti, di monitorare la produttività della rete.

Per la fase 1 e la fase 2 della rilevazione è stato progettato un articolato sistema di monitoraggio che ha consentito di seguirne l'andamento in tempo reale, ravvisando e intervenendo tempestivamente per risolvere le criticità riscontrate nel *field*. Tramite la medesima piattaforma dedicata all'acquisizione dei dati e sviluppata dal Ministero della Salute, ogni giorno sono stati elaborati e resi accessibili gli indicatori aggiornati al giorno precedente, consentendo di monitorare, al massimo livello di dettaglio, tutti i possibili esiti del contatto telefonico, le specifiche ragioni della mancata partecipazione alla rilevazione, il mancato rispetto degli appuntamenti fissati per il prelievo, etc.

A seguire sono riportati alcuni dei principali aspetti che il monitoraggio implementato ha consentito di tenere quotidianamente sotto controllo.

Innanzitutto sono state progettate e predisposte tavole di sintesi in grado di fornire, attraverso pochi indicatori, una visione complessiva dell'andamento della rilevazione nel tempo, con riferimento alle sue diverse fasi. La Tavola 3.2 prodotta con valori assoluti e percentuali non solo a livello nazionale ma anche di singola regione, fornisce una visione immediata della partecipazione delle unità di rilevazione (esiti positivi e negativi), della sua evoluzione nel corso delle settimane e dell'ammontare complessivo del lavoro svolto o che resta ancora da svolgere. La sua lettura è stata utile per tutta la durata della rilevazione, consentendo di mettere subito a fuoco eventuali cali di produttività e di interpretarli alla luce della numerosità degli operatori attivi sul territorio. È infatti attraverso anche la lettura di questi dati che si è evidenziato, dopo il primo mese, il calo della produttività giornaliera collegato al fenomeno degli abbandoni da parte di alcuni operatori e la necessità di intervenire integrando la rete con ulteriore personale. La declinazione a livello territoriale di questa tavola sintetica ha consentito di individuare le aree in cui il fenomeno appena descritto, a titolo esemplificativo, si presentava con maggiore evidenza e di intervenire prioritariamente sulle situazioni locali più critiche.

Tavola 3.2 - Sintesi giornaliera al 6 giugno 2020 (valori assoluti)

DATA	Fase 1 - Call Center						Fase 2 - Prelievo					Fase 3 - Referto
	APPUNTAMENTI PER PRELIEVO FISSATI /INDIVIDUI DISPONIBILI AL PRELIEVO	INDIVIDUI CADUTI (ESITO DEFINITIVO NEGATIVO)	INDIVIDUI CON TENTATIVI DI CONTATTO SENZA ESITO DEFINITIVO	DI CUI CON APPUNTAMENTO TELEFONICO	INDIVIDUI SENZA ALCUN TENTATIVO DI CONTATTO	OPERATORI ATTIVI (con almeno un tentativo di contatto)	PRELIEVO EFFETTUATO	PRELIEVO RINVIATO (CONCORDATA NUOVA DATA O LUOGO)	PRELIEVO NON EFFETTUATO (SENZA NUOVA DATA)	N. UNITÀ DI PRELIEVO ATTIVE (CON ALMENO UN PRELIEVO EFFETTUATO)	N. UNITÀ DI PRELIEVO SENZA NESSUN PRELIEVO	REFERTI VALIDATI E TRASMESSI
Totale	30032	16239	48919	5910	80501	644	13468	314	9784	717	731	502
05-06-2020	2726	1593	8635	1128	80501	465	2038	66	2043	243	252	204
04-06-2020	2830	1852	8703	1150	0	455	1942	76	1999	241	268	236
03-06-2020	2718	1844	8698	1286	0	468	1962	97	1955	231	286	21
02-06-2020	2526	1302	8450	1185	0	383	1088	40	402	88	125	13
01-06-2020	3034	1418	8629	1228	0	464	1537	112	1503	169	236	0
31-05-2020	1984	960	5908	853	0	302	306	20	457	36	122	28
30-05-2020	2467	1204	7179	971	0	413	1358	65	947	137	186	0
29-05-2020	2979	1384	6756	933	0	437	1429	86	1483	191	253	0
28-05-2020	2920	1384	6513	997	0	428	1126	100	1306	172	236	0
27-05-2020	3097	1563	5337	910	0	431	504	59	982	90	231	0
26-05-2020	2746	1885	5809	891	0	439	190	25	360	44	100	0
25-05-2020	1789	1282	4244	746	0	431	0	1	7	0	7	0

Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

La tavola 3.3 è invece un esempio di tavola analitica che ha consentito di dettagliare, al massimo livello, l'esito del contatto telefonico nella fase 1, individuando tutte le possibili casistiche riferite sia agli esiti definitivi (intervista effettuata, intervista non effettuata), sia agli esiti provvisori, seguiti dunque da ulteriori tentativi di contatto.

L'analisi degli esiti provvisori ha consentito di ravvisare alcune criticità legate per esempio agli orari di effettuazione del contatto telefonico e alla necessità di estenderlo alle ore serali per raggiungere fasce di po-

polazione altrimenti difficilmente contattabili. Così come il trend in crescita degli appuntamenti telefonici, fissati su richiesta delle unità campionate per indisponibilità al momento del primo contatto, ha evidenziato, in alcune fasi della rilevazione, la necessità di impegnare gli operatori prioritariamente nella gestione di queste casistiche che, anche per difficoltà di utilizzo della piattaforma rischiavano di non essere adeguatamente trattate.

Analogamente, per monitorare gli esiti definitivi negativi, sono stati rilevati e analizzati anche i motivi che hanno determinato la mancata risposta (non reperibilità, rifiuto, interruzione, etc.), la tipologia di popolazione che ne è risultata coinvolta (in base alle caratteristiche socio-demografiche o a variabili di tipo territoriale) e gli aspetti organizzativi della rilevazione (giorni e orari di rilevazione, gestione degli appuntamenti, etc.). Soltanto in questo modo, è stato possibile identificare le cause delle mancate risposte e adottare, ove possibile, le misure necessarie a limitarle.

Un'altra informazione importante desumibile da questa tavola riguarda la problematica dei numeri errati o inesistenti che, aggiungendosi ai nominativi privi di recapiti telefonici, sono andati a incrementare la quota di unità di rilevazione non contattabili. A fine rilevazione il numero di unità cadute per recapito di telefono errato o inesistente supera le 10 mila unità (pari al 5,6% del campione complessivo). Questa casistica che genera un esito definitivo è stata distinta, come evidente dalla tavola, dalle situazioni in cui alla stessa unità campionaria corrispondono più recapiti di cui uno solo è errato o da quelle in cui attraverso vari canali si riesce a recuperare il recapito telefonico corretto. In entrambi questi casi, risulta chiaro il carico di lavoro aggiuntivo che la rete ha dovuto fare per contattare le unità di rilevazione utilizzando più recapiti o acquisendone di nuovi, prima di arrivare a somministrare l'intervista e fissare un appuntamento per il prelievo.

Data la delicatezza dell'indagine, sono stati monitorati con particolare attenzione i rifiuti a partecipare (Figura 3.3), le cui ragioni sono particolarmente importanti per indirizzare gli operatori all'utilizzo delle strategie più idonee per contrastarli. La reportistica prodotta giornalmente include anche una tavola di approfondimento di questo aspetto che si è rivelata particolarmente utile per individuare le preoccupazioni più diffuse tra i cittadini e gestirle nei modi più appropriati. In particolare, fin dai primi giorni della rilevazione è emerso che la reazione negativa dei cittadini derivava più che da un rifiuto assoluto alla partecipazione, dall'incertezza sull'opportunità di prendervi parte, dalla non evidente rilevanza della loro collaborazione e dalla preoccupazione espressa soprattutto in presenza di minori e anziani.

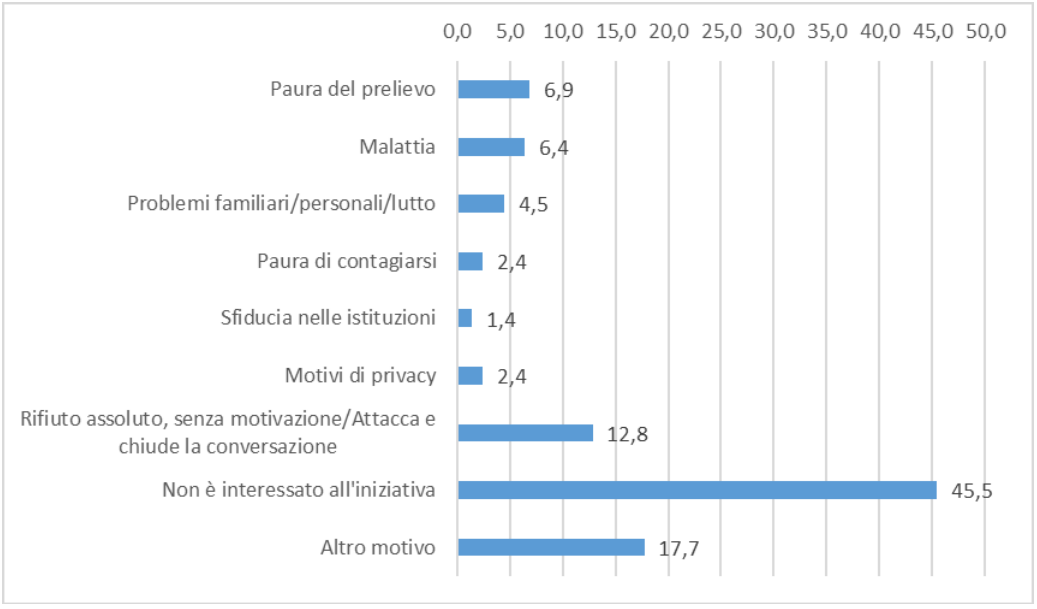
Tavola 3.3 - Tavola analitica sullo stato degli individui al 6 giugno 2020 - FASE 1 (valori assoluti e percentuali)

ESITI	v.a	Per 100 individui teorici	Per 100 individui toccati
1) TOTALE APPUNTAMENTI PER PRELIEVO FISSATI / INDIVIDUI DISPONIBILI AL PRELIEVO)	30032	17,1	31,6
1.1 RISPOSTE DIRETTE da parte di individuo campione	24452	13,9	25,7
1.2 RISPOSTE DA PARTE DI UN FAMILIARE di individuo campione	5580	3,2	5,9
2) TOTALE CADUTE	16239	9,2	17,1
2.1) TOTALE CADUTE PER IRREPERIBILITA	2789	1,6	2,9
A) Numero di telefono inesistente (UNICO RECAPITO)	2042	1,2	2,2
B) Numero errato (corrispondente a persona diversa, sim intestata ad altri, individuo trasferito,individuo sconosciuto, etc.) e NON fornisce recapito (UNICO RECAPITO)	747	0,4	0,8
2.2) TOTALE CADUTE PER ERRORE DI LISTA	455	0,3	0,5
C) Individuo deceduto	101	0,1	0,1
D) Individuo trasferito ESTERO	354	0,2	0,4
2.3) TOTALE RIFIUTI (compresa interruzione definitiva)	12995	7,4	13,7
E) Malato grave	594	0,3	0,6
F) Altri rifiuti	12401	7,1	13,0
3) TOTALE SOSPESE	48919	27,8	51,4
3.1) TOTALE individui SOSPESI RILEVAZIONE IN CORSO	48919	27,8	51,4
G) Libero, cellulare non raggiungibile	27456	15,6	28,8
H) Occupato	1196	0,7	1,3
I) Fax/segreteria	9845	5,6	10,3
L) Il rispondente rifiuta di passare il selezionato e chiude	834	0,5	0,9
M) appuntamento telefonico	5910	3,4	6,2
M1. CHIEDE DI ESSERE RICHIAMATO	4164	2,4	4,4
M2. Al momento assente/non puo rispondere	1746	1,0	1,8
N) acquisito nuovo recapito	439	0,3	0,5
N1. Trasferito altrove in Italia e fornisce recapito telefonico	189	0,1	0,2
N2. Numero errato (corrispondente a persona diversa, sim intestata ad altri, individuo trasferito in Italia, etc.) e fornisce recapito telefonico	250	0,1	0,3
X. Numero inesistente oppure Numero errato (corrispondente a persona diversa, sim intestata ad altri, individuo trasferito, individuo sconosciuto, etc.) e NON fornisce recapito MA SONO PRESENTI ALTRI RECAPITI	1533	0,9	1,6
O) Intervista completata, appuntamento da fissare in regione diversa da quella di residenza	1706	1,0	1,8
3.2) TOTALE individui CADUTE PER FINE PERIODO DI RILEVAZIONE	0	0,0	0,0
P) Fine periodo a disposizione per la rilevazione: iniziati tentativi di contatto	0	0,0	0,0
4) TOTALE individui TOCCATI (1+2+3)	95190	54,2	100,0
Q) Nominativo nuovo, nessun tentativo ancora effettuato	80501	45,8	84,6

Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

Attraverso la lettura di questi dati è stata dunque evidente la necessità di potenziare la campagna informativa sulla rilevazione, in modo da raggiungere quante più persone possibile, e la necessità di gestire alcune situazioni rimandando al confronto con i medici e i pediatri di base, in modo che i più incerti potessero sentirsi rassicurati e rivedere la propria posizione accettando di collaborare, come avvenuto in più casi.

Figura 3.3 - Individui caduti per rifiuto per motivo del rifiuto al 6 giugno 2020 - FASE 1



Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

Rientrano tra gli indicatori calcolati giornalmente anche alcuni dei principali tassi tradizionalmente utilizzati per monitorare non solo la propensione delle unità campionarie a collaborare, ma anche le performance della rete di rilevazione. A tal fine sono stati calcolati e quotidianamente aggiornati i tassi di completezza, di caduta, di rifiuto, di irreperibilità, di inattività, di pigrizia e di appuntamenti/contatti telefonici andati a buon fine[15]. Tutti i tassi sono stati declinati rispetto a variabili di tipo territoriale e alle caratteristiche socio-anagrafiche delle unità campionarie (sesso ed età). In particolare il dettaglio territoriale fornisce indicazioni utili a gestire eventuali aree di sovraccarico della rete e eventuali differenze di "produttività" attraverso opportuni interventi correttivi (integrazione del team di operatori, ritorni formativi mirati, etc.).

Gli indicatori pensati per monitorare la fase 2 hanno consentito di supervisionare i comportamenti delle unità di rilevazione dopo aver rilasciato l'intervista e fissato l'appuntamento per il prelievo (Tavola 3.4). In particolare gli appuntamenti rispettati, modificati e non rispettati senza alcuna comunicazione. Dal monitoraggio quotidiano è emerso che in quest'ultima casistica ricadevano di fatto due diverse circostanze: una era quella prevista di Mancata Risposta Totale (MRT) alla Fase 2 che ha consentito di individuare le unità di rilevazione per le quali essendosi interrotto il processo di acquisizione dei dati sono disponibili solo le informazioni presenti nel questionario. L'altra circostanza riguarda i prelievi effettuati ma non caricati a sistema (per ragioni tecniche o di altra natura). Ovviamente una volta appurata questa commistione sono stati adottati gli opportuni interventi per risolvere i problemi di caricamento da un lato e provare, dall'altro, a ricontattare le unità che non si erano presentate all'appuntamento per il prelievo. Per monitorare questa fase così delicata, gli indicatori sono stati declinati a livello di singolo centro prelievo per cogliere tempestivamente i comportamenti dei cittadini e della rete al massimo livello di dettaglio territoriale e poter implementare le strategie previste (di ricontatto e sollecito).

15 Si vedano in proposito Istat, 2006; Istat 2007a; Istat 2007b.

Tavola 3.4 - Esito dei prelievi per regione al 6 giugno 2020 - FASE 2 (valori assoluti e percentuali)

REGIONE	Prelievi effettuati			Prelievi rinviati (concordata nuova data o luogo)		Prelievi non effettuati		Totale prelievi attesi
	v.a [a]	v.a. campione anticipatorio	per 100 individui con almeno un appuntamento nella regione	v.a	per 100 individui con almeno un appuntamento nella regione	v.a	per 100 individui con almeno un appuntamento nella regione	v.a
PIEMONTE	868	242	63,0	9	0,7	506	36,7	1378
VALLE D'AOSTA	786	309	86,0	4	0,4	126	13,8	914
LOMBARDIA	306	123	29,8	83	8,1	640	62,3	1028
PROV. AUTON. BOLZANO	0	0	0,0	10	7,7	120	92,3	130
PROV. AUTON. TRENTO	750	274	88,4	9	1,1	89	10,5	848
VENETO	0	0	0,0	1	25,0	3	75,0	4
FRIULI VENEZIA GIULIA	1032	183	57,8	9	0,5	755	42,3	1787
LIGURIA	1397	224	75,1	8	0,4	467	25,1	1860
EMILIA ROMAGNA	0	0	0,0	10	2,8	346	97,2	356
TOSCANA	212	70	26,4	5	0,6	589	73,4	803
UMBRIA	26	26	2,8	2	0,2	889	97,2	915
MARCHE	1887	379	87,4	13	0,6	302	14,0	2159
LAZIO	346	113	69,1	19	3,8	143	28,5	501
ABRUZZO	993	265	68,7	14	1,0	441	30,5	1445
MOLISE	295	170	73,9	13	3,3	93	23,3	399
CAMPANIA	1080	127	49,8	12	0,6	1090	50,2	2170
PUGLIA	1297	191	71,2	21	1,2	518	28,4	1822
BASILICATA	1147	144	59,9	5	0,3	834	43,5	1916
CALABRIA	494	118	68,7	13	1,8	217	30,2	719
SICILIA	363	46	41,0	35	4,0	489	55,3	885
SARDEGNA	202	31	14,7	19	1,4	1150	83,9	1371
ITALIA	13481	3035	57,6	314	1,3	9807	41,9	23410

Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

Alcuni indicatori sono stati prodotti anche a livello di singolo operatore in modo da avere il massimo dettaglio informativo sulle performance dell'intera rete di rilevazione. Per tutta la durata della rilevazione, l'attività dei rilevatori è stata costantemente tenuta sotto controllo attraverso l'analisi dei report giornalieri che hanno permesso di identificare situazioni problematiche e di cogliere tempestivamente "eventi sentinella", sintomi di comportamenti anomali, consentendo di riorientare con immediatezza la situazione difficilmente sanabile solo con interventi a posteriori, sicuramente più complessi e meno efficaci, in fasi successive all'acquisizione dei dati.

È stata monitorata anche la diversa produttività in base alla fascia oraria dei contatti telefonici. Questo tipo di informazione è molto importante ai fini di una ottimale organizzazione della rete e della turnazione dei vari operatori che, oltre a garantire la copertura delle ore di contattabilità definite in fase progettuale, deve predisporre il *field* in maniera da massimizzarne la produttività negli orari con maggiore probabilità di successo nel contatto con le unità di rilevazione.

Il sistema di monitoraggio è stato pensato anche per monitorare due importanti aspetti di processo. Il primo riguarda l'ammontare delle unità da raggiungere che, nell'indagine in questione, non corrisponde esattamente al campione estratto. In questa rilevazione, infatti, come si è anticipato, le unità di rilevazione sono diventate contattabili, attraverso la tecnica adottata (telefonica) anche in corso d'opera, man mano che venivano recuperati i recapiti telefonici attraverso tutti i canali attivati (vedi par. 2). Di conseguenza è stato importante

monitorare l'evolversi di questo processo, così come il mutare quotidiano del numero di record lavorabili e l'impatto sui carichi di lavoro della rete.

Analogamente è stato importante monitorare con appositi indicatori i flussi di prelievi da effettuare in regione diversa da quella di residenza, determinando un carico di lavoro non previsto sui centri prelievo della regione di domicilio e potendo determinare criticità anche sul piano organizzativo (vedi par.3.1).

Quelli appena descritti sono solo alcuni dei principali indicatori utilizzati per monitorare in corso d'opera l'indagine sulla sieroprevalenza, attività che ha accompagnato incessantemente la rilevazione fornendo ai titolari dell'indagine e ai supervisori tutte le informazioni utili a ravvisare le criticità e fornire indirizzi utili alla loro tempestiva soluzione.

Insieme al sistema di indicatori di qualità sono state monitorate regolarmente, durante tutta la fase di rilevazione, anche le distribuzioni di frequenza (assolute e percentuali) delle variabili presenti nel questionario La possibilità di disporre tempestivamente dei valori registrati nei vari item di risposta presenti per ogni singolo quesito ha consentito di svolgere un'accurata attività di controllo, soprattutto sulla consistenza reale dei valori assunti nel campione dalle variabili osservate. Questo tipo di controlli permette di evidenziare in corso d'opera sistematicità o distorsioni nelle modalità di risposta, dovute ad esempio a una non corretta formulazione del quesito oppure a un'errata attribuzione della risposta da parte dell'intervistatore, e di intervenire in corso d'opera per eliminare la causa dell'errore. Le distribuzioni di frequenza possono, inoltre, considerarsi o rivelarsi anche degli ottimi strumenti di supporto per la segnalazione di errati funzionamenti del sistema di acquisizione, soprattutto per ciò che riguarda il controllo dei vincoli e delle coerenze interne al questionario elettronico, qualora siano sfuggiti in fase di progettazione e test del questionario.

In estrema sintesi, sia in fase di progettazione che di realizzazione della rilevazione sono state adottate tutte le misure necessarie a contenere gli errori non campionari e a monitorare e gestire le criticità emerse durante il lavoro sul campo. Tra queste va ricordata, in primis, il senso di incertezza e le preoccupazioni che hanno indotto molti cittadini, soprattutto nei segmenti più fragili (bambini e anziani), a non collaborare alla rilevazione. All'opposto si sono registrati casi di persone che si sono fatte avanti, chiamando il numero verde o recandosi presso i Centri di prelievo, per prendere parte alla rilevazione, cosa che ovviamente non è stata possibile, non rientrando nell'insieme dei soggetti facenti parte del campione.

La rete di rilevazione, costituita dai volontari della CRI, ha dato il massimo, confermandosi molto preparata nella gestione dei rapporti con i cittadini in condizioni di emergenza, ma, come comprensibile, era meno avvezza nella gestione di un'indagine statistica, in cui il tempo a disposizione per imparare tutte le strategie comportamentali da adottare e gli strumenti da utilizzare è stato veramente poco.

Altro elemento di criticità è stato il non pieno allineamento dei centri prelievi con le esigenze provenienti dal *field*. L'elenco iniziale è stato modificato in corso d'opera proprio per meglio rispondere alle esigenze organizzative. Ove possibile, anche con l'aiuto delle informazioni fornite dall'Istat (in termini di distanze delle unità campionate dai centri prelievo[16]), i punti prelievo sono stati collocati in posti raggiungibili dalla gran parte delle unità. Ma non sempre ciò è stato possibile, dunque la distanza dai centri prelievo o la difficoltà di raggiungimento degli stessi è stata talvolta motivo di rinuncia alla partecipazione.

16 La metodologia adottata O/D matrix consente di ottenere stime teoriche, calcolate in assenza di traffico alla velocità massima consentita. Essa permette di sapere per ogni percorso (residenza dell'individuo - centro prelievo) distanza e tempo in minuti, espresso in termini di percorrenza stradale.

PARTE II

PROGETTAZIONE DELL'IMPIANTO STATISTICO-METODOLOGICO[17]

17 Michele D'Alò, Claudia De Vitiis, Stefano Falorsi, Andrea Fasulo, Danila Filipponi, Alessio Guandalini, Francesca Inglese, Roberta Radini.

4. Introduzione

I par. 2 richiama gli obiettivi dell'indagine secondo un approccio statistico, funzionale alla definizione della strategia campionaria. I paragrafi successivi, 3 e 4, costituiscono una premessa alla nota metodologica vera e propria in cui si descrivono i criteri sottostanti alle principali scelte metodologiche effettuate e si trattano le modalità di costruzione del frame di campionamento a partire dal Sistema Integrato dei Registri. Il paragrafo 5, conclude il lavoro, riprendendo quanto già trattato secondo lo schema più standard delle note metodologiche Istat, che documentano le rilevazioni dell'Istituto. Questo paragrafo contiene, quindi, anche tutte le tabelle con le descrizioni quantitative del disegno campionario adottato e anche, i grafici sulla distribuzione territoriale del campione selezionato secondo diverse partizioni del territorio.

5. Parametri *target* dello studio epidemiologico

Lo studio si rivolge a due differenti *popolazioni di interesse* – intese come l'insieme delle unità statistiche sulle quali si intende investigare – costituite rispettivamente dagli individui:

- *residenti in Italia*, esclusi i membri permanenti delle convivenze;
- *residenti in Italia occupati*.

L'indagine è di tipo trasversale, in quanto si osservano le caratteristiche della popolazione di interesse con riferimento a un istante di tempo – in questo caso si tratta di un intervallo circoscritto di poche settimane precedenti la data di inizio della rilevazione (25 maggio 2020)[18].

I principali parametri di interesse dell'indagine riguardano la frequenza assoluta e relativa degli individui in funzione del loro stato epidemiologico con riferimento a differenti sotto-popolazioni appartenenti a una specifica partizione del territorio o a specifiche sottoclassi dell'intera popolazione individuate da precise caratteristiche strutturali individuate da variabili demo-sociali. In ogni caso tali caratteristiche individuano quelli che nel seguito saranno denominati *domini di studio*. Per quanto riguarda i domini si ritiene utile una suddivisione in domini *primari* e *secondari*. Per i primi, si vogliono garantire prefissati livelli di precisione delle stime, ritenuti accettabili, compatibilmente con i vincoli di budget. Per secondi, invece, che non influiscono direttamente sulla definizione delle numerosità e, quindi sul budget, si vuole verificare ex-post alla fase di selezione, che il campione effettivamente estratto abbia una copertura tale da garantire, almeno per parte di essi, livelli di precisione accettabili. A tale proposito si precisa che per le indagini su larga scala, errori di campionamento relativi percentuali o Coefficienti di Variazione percentuali (CV%), sotto la soglia del 15-10% sono considerati come livelli di *medio-alta* attendibilità, mentre CV% intorno alla soglia del 33% vengono considerati come livelli di attendibilità *bassa* ma ancora accettabile. Le stime i cui corrispondenti CV% sono superiori alla soglia del 33% sono classificate come aventi livelli di attendibilità *non accettabile*. Per poter tenere sotto controllo gli errori campionari attesi delle stime riferite ai domini primari, le diverse classificazioni che identificano i suddetti domini, sono state considerate nel processo di stratificazione delle unità campionarie ai vari stadi di campionamento. Ciò ha consentito di definire l'allocazione del campione teorico, in modo tale che fossero rispettati i vincoli sugli errori attesi per i diversi domini primari. A tale proposito si rileva, che nella maggior-parte delle indagini campionarie dell'Istat sulla popolazione, i domini di tipo primario sono essenzialmente costituiti da variabili di tipo geografico, preva-

18 Il periodo di rilevazione sul campo è andato dal 25 maggio 2020 al 15 luglio 2020.

lentemente le regioni geografiche insieme alle loro aggregazioni gerarchiche, quali le ripartizioni geografiche e l'intero territorio nazionale. Per tali indagini si considerano, invece, i domini di stima di tipo strutturale – legate alle caratteristiche delle unità della popolazione di interesse – quali, ad esempio, la classe di età, il sesso, il titolo di studio che vengono, tuttavia, considerate come domini di tipo secondario che non vengono introdotti nel processo di stratificazione ma semmai in quello di post-stratificazione per la costruzione delle stime. Nel caso in esame, invece, a differenza delle situazioni più ricorrenti per le indagini campionarie dell'Istat sono stati considerati tra i domini primari anche alcune caratteristiche degli individui, quali il sesso, l'età e la caratteristica di risultare occupato sui registri dell'Istat. Tale scelta è stata dettata dal fatto che questi domini erano ritenuti strategici per l'indagine. Si è ritenuto, pertanto, necessario introdurli tra i domini pianificati al costo di una maggiore complessità del disegno di campionamento a due stadi, caratterizzato sia dalla stratificazione delle unità primarie (i comuni) che dalla stratificazione di quelle secondarie (gli individui).

Ciò premesso, per l'indagine in oggetto, i *domini territoriali primari* sono i seguenti:

- *Intero territorio nazionale*;
- *Macro aree* di contagio[19] così definite: (1) "Zona rossa", che comprende le regioni Piemonte, Lombardia, Veneto, Emilia Romagna e Marche; (2) "Resto del nord più centro", che include le regioni Valle d'Aosta, province autonome di Trento e Bolzano, Friuli, Liguria, Toscana, Umbria e Lazio; (3) "Meridione", che coincide con l'usuale ripartizione Sud e Isole;
- *Regioni Geografiche* a cui si aggiungono le Province autonome di Bolzano e Trento.

I *domini territoriali secondari* sono, invece, costituiti dalle Provincie, dai Sistemi Locali del Lavoro (SLL) e dalle Aziende Sanitarie Locali[20] (ASL).

I *domini strutturali primari*, definiti all'interno di ciascun *dominio territoriale primario* sono costituiti dalle seguenti variabili:

- classi di età (0-17; 18-34; 35 -49; 50-59; 60-69; 70 e più);
- sesso.

Inoltre, per quanto riguarda la popolazione degli *individui residenti occupati*, oltre ai domini strutturali appena elencati, costituisce un dominio di studio anche:

- l'Attività Economica (o ATECO) raggruppata in 4 classi: "Occupati sospesi" (D.P.C.M. 22/03/2020), "Occupati non sospesi Altro", "Occupati non sospesi della Pubblica Amministrazione ed Istruzione", "Occupati non sospesi della Sanità".

Per questa popolazione i *domini strutturali secondari* sono rappresentati dall'incrocio delle classi di età per le classi ATECO.

L'indagine trasversale, compatibilmente con i vincoli di bilancio, ha l'obiettivo di produrre stime dei parametri di interesse, al livello di:

- Regione (totale degli individui della regione), con livelli di attendibilità medio-alta, mantenendo possibilmente un certo livello di attendibilità accettabile anche per i domini strutturali contenuti in ciascuna regione;
- Macro-area di contagio (totale degli individui dell'insieme delle regioni appartenenti all'area), con livelli di

19 Per dare un'idea della prevalenza differenziale tra le regioni del numero di individui "infetti" si ritiene utile fare riferimento ai dati della Protezione civile, aggiornati al 7 Aprile 2020, che sono disponibili al link: https://github.com/pcm-dpc/COVID-19. Si ritiene utile chiarire, che la cosiddetta "Zona rossa", definita in base ai dati in parola era rilevante ad aprile 2020 quando è stato definito il disegno dell'indagine. Tuttavia, con l'evolversi del fenomeno nel tempo la "Zona rossa" ha assunto via via differenti connotazioni del tutto diverse da quella iniziale di aprile 2020.

20 Per il Comune di Milano, su specifica richiesta, sono anche stati considerati i Distretti Sanitari (DS).

attendibilità alta, garantendo, anche, livelli di attendibilità medio-alta per i domini strutturali contenuti in ciascuna Macro-area;

- Nazionale e per ciascun dominio strutturale, con livelli di attendibilità molto alti.

Oltre alla raccolta dei campioni di sangue su cui effettuare il test sierologico prescelto dal CTS, l'indagine ha l'obiettivo di rilevare, mediante un breve questionario un insieme minimo di informazioni ritenute essenziali per arricchire l'informazione epidemiologica di base – vale a dire per definire il "quadro immunitario" e "la diffusione del virus" citati nella norma (cfr. par. 1.1) – con informazioni ausiliarie sulle caratteristiche della popolazione d'interesse. Si osserva che la variabile esito positivo o negativo del test in base al quale viene definita la variabile dicotomica oggetto di indagine individuo *che ha sviluppato anticorpi* si/no deve essere incrociata con ciascuna delle modalità delle variabili che definiscono i domini di stima. In tal senso possiamo dire che l'indagine è *multi-obiettivo*[21] in quanto il principale parametro di interesse viene calcolato con riferimento a differenti sotto-popolazioni di interesse.

6. Impianto generale dell'indagine

L'impianto metodologico dell'indagine è stato discusso sia con gli esperti dell'*Advisory Commitee*[22], istituito appositamente dal Presidente dell'Istat per supportare la progettazione dello studio di sieroprevalenza sul *SARS-COV-2*, che con quelli dell'*Advisory Commitee on Statistical Methods*[23].

Un importante aspetto discusso è stato quello relativo all'utilizzo delle sostituzioni per gestire le Mancate Risposte Totali (MRT). In particolare, tenendo conto della natura sia delle variabili *target* sia degli obiettivi di indagine, è stata condivisa la scelta finale adottata di non ricorrere alle sostituzioni poiché la rilevazione è stata progettata per garantire un alto profilo di qualità alle stime prodotte sia in termini di CV attesi, che di forte contenimento della distorsione connessa alle MRT. Per cercare di garantire comunque il rispetto delle numerosità campionarie teoriche prefissate in fase di disegno il campione di individui pianificato, detto nel seguito campione *obiettivo*, è stato sovra-dimensionato di circa il 30%; quest'ultimo sarà richiamato nel seguito come campione *selezionato*. Si è passati, quindi, da una numerosità iniziale di circa 150.000 del campione *teorico* a una finale intorno ai 195.000 individui per il campione *selezionato*. In assenza di informazioni a *priori* sul tasso di mancata intervista – dovuto sia alle MRT che ai mancati contatti connessi agli errori nel frame di campionamento o all'aggancio dei numeri telefonici da parte dei provider di telefonia mobile e fissa – per un'indagine di questo tipo si è ritenuto ragionevole ipotizzare un tasso di caduta maggiore o uguale del 30%.

21 Il concetto di indagine *multi-obiettivo* può essere ulteriormente esteso sfruttando le informazioni raccolte con il questionario. Si potrebbe valutare, infatti, anche, la possibilità di produrre stime riferite a specifiche sotto-popolazioni, oltre a quelle pianificate, individuate dalle differenti modalità delle variabili contenute nel questionario generando nuove variabili incrocio. Ad esempio, la variabile individuo *ritirato dal lavoro che ha sviluppato anticorpi* si/no. Ciò rappresenterebbe, quindi, una moltiplicazione dei parametri obiettivo per tanti quanti sono gli incroci di interesse che potrebbero essere considerati nel piano di tabulazione. Ovviamente tale possibilità deve essere valutata sulla base degli esiti dell'indagine tenendo conto degli errori campionari delle stime.

22 L'*Advisory Commitee* è stato istituito dal Presidente dell'Istat il 7 Aprile 2020 per supportare l'Istituto nella progettazione dell'indagine di sieroprevalenza in esame e comprende i seguenti esperti esterni: F. Bartolucci, G. Costa, P.D. Falorsi, M. Pratesi, M. G. Ranalli, E. Stanghellini con la partecipazione stabile di Linda Laura Sabbadini, Orietta Luzi e Stefano Falorsi.

23 L'*Advisory Commitee on Statistical Methods* è un organismo collegiale istituito dall'Istat per supportare, migliorare e validare l'introduzione delle innovazioni metodologiche nei processi dell'Istituto. E' composto da esperti metodologi di fama internazionale. La riunione del 13 maggio 2020 è stata dedicata al disegno di campionamento dell'indagine in questione: https://intranet.istat.it/News/Pagine/Comitato-consultivo-per-le-metodologie-statistiche--lindagine-sierologica-focus-dellappuntamento-di-mercoled%C3%AC-13-maggio-.aspx

6.1 Principali scelte metodologiche adottate

Questo paragrafo affronta il problema della formazione del campione. A tale scopo, occorrerebbe focalizzare l'attenzione su molteplici aspetti, interconnessi tra loro, tuttavia ci si limita a dare alcuni cenni su quelli più significativi riguardanti i criteri adottati per:

(1) l'allocazione del campione;
(2) la definizione dello schema generale di campionamento;
(3) la stratificazione delle unità campionarie.

Ulteriori dettagli sui suddetti aspetti saranno forniti nel par. 4.

Per quanto riguarda il punto (1) un elemento di complessità per la definizione degli obiettivi è legato all'esigenza di produrre stime di parametri riferiti a un numero elevato di domini di studio, sia territoriali che strutturali. In tal caso, la ricerca di soluzioni ottime per ciascun dominio può contrastare con l'obiettivo di individuare una soluzione ottima generale. Ciò significa che, al fine di determinare la numerosità campionaria minima atta a consentire il calcolo di stime con predeterminati livelli di precisione, è necessario adottare una soluzione metodologica che tenga conto contemporaneamente di una molteplicità di obiettivi e di vincoli. La soluzione a cui si perviene in tal caso è ottimale in senso globale ma fornisce soluzioni sub ottimali per la pluralità dei domini pianificati, condizionati al mantenimento della efficienza globale e complessiva. Tale metodologia generalizza il metodo proposto da Bethel (1989) per la determinazione della dimensione *ottimale* in un'ottica multivariata, riferita al caso di un disegno a uno stadio stratificato e di un solo dominio territoriale di studio.

Per l'allocazione del campione progettato di 150.000 individui – che come detto sopra corrisponde a un campione selezionato di 195.000 individui – a livello regionale e sub-regionale, sono stati utilizzati i dati ufficiali sulla prevalenza del numero di individui *infetti* a livello di ciascuna regione forniti dall'ISS e relativi al 7 Aprile 2020. A partire da queste informazioni l'Istituto ha proposto tre scenari alternativi. Il primo scenario, caratterizzato dalle prevalenze più basse, inflaziona di cinque volte il dato ufficiale degli *infetti*, il secondo e terzo scenario inflazionano i medesimi dati di 10 e 20 volte.

Per lo studio dell'allocazione campionaria sono stati utilizzati i dati del primo scenario. Questa è stata ritenuta, infatti, la scelta più prudente al fine di riuscire a ottenere stime di buona qualità. Poiché gli scenari si basavano su ipotesi di diffusione della malattia fortemente variabili tra le regioni, essendo il campo di variazione compreso tra 0.13% per il sud e 2.92 % per il nord, al fine di evitare che tutto il campione fosse assorbito dalle regioni con prevalenze basse si è deciso di formare tre classi di prevalenze con un campo di variazione più contenuto compreso tra l'1% e il 2%.

A tale riguardo, un aspetto critico del disegno concerne la definizione di un'indagine volta a osservare un fenomeno mai misurato in precedenza per il quale erano disponibili solo informazioni incomplete e relative a uno specifico sottoinsieme dell'intera popolazione, individuato dal numero ufficiale di individui *infetti* rilevati mediante il "tampone". Il sottoinsieme della popolazione relativo agli individui infetti "asintomatici" e paucisintomatici" a cui non è stato effettuato l'esame del tampone è, infatti, privo di informazioni a priori quantitativamente rilevanti. Di conseguenza appare chiaro che la sola informazione effettivamente disponibile ha fornito una conoscenza sommaria della vera distribuzione sul territorio nazionale del numero di *infetti*, che può essere distorta sia nei livelli che nella variabilità, per effetto delle differenti politiche di effettuazione dei tamponi da parte delle diverse Regioni. Si tratta, inoltre, come è evidente di un fenomeno fortemente concentrato in alcuni ambiti territoriali con una forte dinamica di evoluzione nel tempo e nello spazio. Per questi motivi, le informazioni *ex-ante* sono state

opportunamente trattate al fine di evitare che le numerosità campionarie allocate nei differenti *domini primari* fossero fortemente "disproporzionate" rispetto alla popolazione residente di ciascuno di essi.

Inoltre, per allocare il campione nelle Regioni si è tenuto conto non solo del criterio di "rappresentatività" del campione basato sull'andamento previsto delle prevalenze sul territorio ma anche sul criterio di "proporzionalità", in base al quale la dimensione del campione estratto per ciascuna regione dipende anche dalla popolazione residente in ciascuna Regione. Ciò significa che la dimensione di ciascun campione regionale dipende anche dal peso relativo della Regione in termini di popolazione.

Una volta definito il numero di individui campione da estrarre per ciascuna Regione, l'allocazione nei diversi domini sub-regionali, territoriali e strutturali, è stata effettuata seguendo un criterio di sola "proporzionalità". Ciò è giustificato dalla carenza di informazioni di buona qualità sulla distribuzione delle prevalenze a livello sub-regionale.

Per quanto riguarda il secondo punto dell'elenco, è stato preso in considerazione un disegno campionario di tipo complesso adottato comunemente per le più importanti indagini Istat sulle famiglie. Tale disegno prevede l'estrazione di un campione a uno stadio, in cui le Unità Finali di campionamento (UF) sono costituite dagli individui; per i comuni "grandi", ossia quelli che superano prefissate soglie di popolazione residente. Per i restanti comuni si utilizza, invece, un disegno a due stadi, in cui le Unità Primarie di campionamento (UP) sono i comuni e le Unità Secondarie (US) coincidono con le UF. Lo schema misto, consente di mediare in modo opportuno le proprietà positive e negative dei disegni a uno stadio e due stadi. Come è noto, infatti, ciascuno dei due schemi separatamente considerati, comporta vantaggi e svantaggi sia in termini di aspetti operativi e di costo, che di efficienza attesa delle stime. A parità di unità finali di campionamento, è evidente come uno schema a uno stadio sia preferibile in quanto in grado di garantire una migliore distribuzione spaziale del campione sul territorio, coinvolgendo un numero relativamente più alto di comuni, producendo, per questa via, una maggiore efficienza delle stime prodotte. Dall'altro, lo schema a due stadi consente di gestire meglio la raccolta delle informazioni sul campo tenendo sotto controllo il numero di comuni campione e, per questa via, ridurre i costi di rilevazione[24]. Comunque, la strategia di campionamento ha dato luogo a una buona dispersione spaziale del campione estratto sul territorio.

Più in dettaglio, il piano di campionamento prevede l'applicazione del disegno complesso appena descritto all'interno di ciascuna delle 107 Province in modo tale da garantire, anche per questo importante dominio secondario, una copertura campionaria completa. In particolare, nell'ambito di ciascuna Provincia, i comuni la cui dimensione demografica è maggiore o uguale a una prefissata soglia di popolazione residente sono selezionati nel campione con certezza andando a costituire degli strati a se stanti. Tali comuni sono detti Auto Rappresentativi (AR) e l'insieme che li comprende è detto dominio AR. I restanti comuni, detti comuni Non Auto Rappresentativi (NAR), sono ordinati in base a una graduatoria decrescente in termini di popolazione residente e suddivisi in strati approssimativamente di uguale ampiezza demografica. L'insieme che comprende tutti i comuni di questo tipo è detto dominio NAR. Il disegno, inoltre, prevede:

24 Vale la pena di ricordare, a tale proposito, che per ciascuna variabile di interesse, la maggiore variabilità campionaria dovuta all'adozione del disegno a due stadi cresce all'aumentare del grado di omogeneità interna della variabile osservata sulle UP, misurato statisticamente dal coefficiente di correlazione intra-cluster, e diminuisce al crescere del numero di UP selezionate nel campione. Quindi, quando si progettano indagini basate su disegni a due o più stadi è molto importante avere informazioni ex-ante, il più possibile accurate, sulla variabilità e sul grado di omogeneità interna alle UP, per le principali variabili oggetto di indagine. Ciò consente di valutare l'inflazione della varianza campionaria attesa in relazione a diverse scelte del disegno.

- l'auto ponderazione[25] del campione di individui nell'ambito di ciascuna provincia;
- la selezione, senza re-immissione e probabilità proporzionali all'ampiezza demografica, di un comune campione in ogni strato NAR;
- l'assegnazione di un numero minimo (pari a 50), di individui da intervistare in ciascun comune campione.

Per quanto riguarda il punto (3), relativo ai criteri adottati per la stratificazione dei comuni, si rileva che l'identificazione di un gruppo di comuni di tipo AR, riduce l'esigenza di studiare stratificazioni complesse per i comuni di tipo NAR, in quanto l'inclusione nel campione dei comuni di maggiori dimensioni demografiche riduce l'impatto della stratificazione dei comuni NAR sull'efficienza delle stime riferite all'intero dominio di studio. Tale ragione permette di utilizzare una o pochissime variabili di stratificazione cercando una soluzione di buon senso. Il criterio di formazione degli strati di comuni NAR, rientra nella logica della determinazione ottimale dei confini suggerita da Mahalanobis (1952) e da Hansen *et al.* (1953) che permette di delimitare gli strati in modo che abbiamo una dimensione pressoché costante.

La stratificazione degli individui, all'interno dei comuni, è stata effettuata sulla base dei domini strutturali individuati dall'appartenenza di ciascun individuo a specifiche categorie demo-sociali. Un criterio semplice, sarebbe stato quello di considerare in ciascun comune un'ulteriore livello di stratificazione *strutturale*. Tuttavia, questa soluzione avrebbe comportato, un numero molto elevato (oltre 100.000) di strati molto piccoli, determinando una considerevole inefficienza nelle stime prodotte e, comunque, un innalzamento eccessivo della numerosità campionaria all'interno di ciascuno dei 100.000 strati. Per risolvere questo problema, è stata applicata una tecnica di bilanciamento del campione, nota come metodo dei Quadrati Latini, descritta in Cochran (1977). Nello specifico, si considerano due allocazioni marginali all'interno di ciascuna regione. La prima relativa all'allocazione degli individui campione negli strati di comuni di ciascuna provincia. La seconda relativa agli strati formati dal concatenamento delle variabili che definiscono i domini strutturali.

7. La costruzione della lista di campionamento

La costruzione della lista di campionamento è stata possibile grazie alla disponibilità in Istituto del *Sistema Integrato dei Registri* (SIR). Il SIR, realizzato attraverso l'integrazione di dati derivati dalle fonti amministrative, dalle rilevazioni statistiche e dalle nuove fonti di dati, ha l'obiettivo di produrre statistiche ufficiali in modo unitario e coerente nei diversi domini statistici attraverso l'integrazione concettuale e statistica delle unità che lo compongono.

Nella costruzione della lista di campionamento sono stati utilizzati i seguenti registri del SIR:

- *Registro Statistico di Base degli individui, delle famiglie e delle convivenze* (RBI) che contiene informazioni anagrafiche di base (età, sesso, luogo di nascita, cittadinanza) e il livello di istruzione per tutti gli individui della popolazione italiana riferibile a un certo anno. Inoltre contiene le informazioni relative al luogo di residenza al primo gennaio dell'anno di riferimento del registro.
- *Registro Statistico di base delle unità economiche* (ASIA) che contiene variabili strutturali, quali attività economica, forma giuridica e addetti, per tutte le unità economiche attive sul territorio italiano nell'anno di riferimento del registro.

25 Tutti gli individui di una data provincia hanno la medesima probabilità di essere inclusi nel campione.

- *Registro Statistico tematico sul lavoro* (RTL) che contiene le posizioni lavorative attive nell'anno di riferimento del registro con le relative caratteristiche. Le posizioni lavorative definiscono il legame tra due tipologie di unità statistiche, il datore di lavoro e il lavoratore, legando le caratteristiche del lavoratore, quelle del datore di lavoro e quelle della relazione lavorativa che ne scaturisce.
- *Registro base dei Luoghi* (RBL) che contiene le informazioni relative alle diverse tipologie di unità territoriali: (i) amministrative, come comune, province, regioni, sezioni di censimento, (ii) funzionali, come sistemi locali del lavoro e ASL, (iii) geo-referenziali come coordinate di punti di interesse, di confini di aree, (iv) tabellari come informazione di dettaglio sullo stradario.

I quattro registri considerati sono integrabili attraverso codici pseudo-anonimi che identificano le unità statistiche di base e che consentono di collegare le informazioni tra registri e quindi domini differenti.

L'insieme delle unità statistiche sulle quali l'indagine di sieroprevalenza sul *SARS-COV-2* intende investigare sono gli individui (i) residenti in Italia e (ii) residenti in Italia occupati. Date le popolazioni di interesse dell'indagine, i Registri di riferimento per l'individuazione delle unità statistiche sono RBI e RTL, rispettivamente per l'individuazione degli individui residenti e gli occupati.

Il *Registro Statistico di Base degli individui* è una lista di campionamento consolidata in Istituto da più di tre anni ed è utilizzata anche per la selezione del campione del Censimento Permanente della Popolazione. Il registro è alimentato da tutte le fonti amministrative acquisite dall'Istituto che hanno come unità di riferimento l'individuo e a partire da questa integrazione, utilizzando metodologie *ad hoc*, vengono individuate popolazioni di interesse statistico. La popolazione degli *individui residenti* nei vari comuni italiani è definita sulla base della presenza dell'individuo nelle Liste Anagrafiche Comunali (LAC) e l'Anagrafe Nazionale delle Persone Residenti (ANPR) e il riferimento temporale dell'informazione è il primo gennaio di ogni anno. Oltre all'informazione sulla residenza dalle fonti anagrafiche vengono acquisite le informazioni relative alla composizione familiare oppure l'appartenenza a convivenze, alle relazioni tra gli individui della famiglia con l'intestatario del foglio di famiglia e il loro stato civile. Per ragioni legate ai tempi di acquisizione dell'informazione e di elaborazione statistica la popolazione residente e le relative informazioni anagrafiche sono disponibili con un ritardo di circa 5 mesi rispetto alla data di riferimento dell'informazione.

L'altro registro coinvolto per l'individuazione delle popolazioni *target* dell'indagine è il *Registro Statistico Tematico del Lavoro*. Il registro è pensato come un sistema informativo a supporto dei processi statistici relativi alla domanda e all'offerta di lavoro, retribuzioni, costo del lavoro e ore lavorate. Il registro del lavoro associa le informazioni relative ai due lati del mercato del lavoro: le unità economiche (datore di lavoro), che utilizzano le risorse umane di cui hanno bisogno, e gli individui (lavoratori) disposti a offrire i propri servizi lavorativi.

Gli individui e le unità economiche sono associati mediante relazioni aventi per oggetto un'attività lavorativa, indicate come posizioni lavorative. Le posizioni lavorative sono caratterizzate oltre che dall'unità economica e l'individuo, anche dalla forma di lavoro e dalla relativa data di attivazione e quindi e possibile individuare l'esistenza o meno di una posizione lavorativa per qualsiasi intervallo temporale di interesse. È importante sottolineare che uno stesso individuo, in un determinato periodo di riferimento, può avere più posizioni lavorative associate a diverse unità economiche. Il *Registro Statistico Tematico del Lavoro* viene aggiornato annualmente, sulla base di fonti previdenziali e fiscali. Anche in questo caso, per ragioni legate ai tempi di acquisizione dell'informazione e di elaborazione statistica è disponibile con un ritardo di circa 15 mesi rispetto alla data di riferimento dell'informazione.

Partendo Registro Statistico Tematico del Lavoro è possibile derivare l'universo degli individui occupati come l'insieme degli individui che presentano almeno una posizione lavorativa nell'intervallo di tempo preso in considerazione. Le caratteristiche di lavoro sono derivate a partire dalle posizioni lavorative. Nel caso in cui allo stesso individuo siano associate più posizioni lavorative, la posizione lavorativa prevalente è individuata sulla base delle definizioni ILO ed è quella alla quale si dedica il maggior numero di ore di lavoro o, a parità di ore, quella dalla quale deriva un reddito più elevato.

L'universo di riferimento per la lista di campionamento è stato ricavato a partire dagli individui presenti in RBI e residenti in Italia al 1° gennaio 2019. Inoltre, al fine di ottenere una lista più aggiornata e quindi di ridurre i mancati contatti in fase di rilevazione dell'indagine sono state acquisite e integrate ulteriori fonti quali:

- le fonti anagrafiche riferite al gennaio 2020, anche se non ancora in forma completa e hanno consentito di aggiornare i cambi di residenza in comuni diversi da quelli selezionati dal campione;
- la fonte di anagrafe tributaria delle persone fisiche riferita all'anno 2019 per effettuare l'esclusione delle persone decedute durante il 2019;
- una fornitura *ad hoc* della fonte di anagrafe tributaria delle persone fisiche riferita ai primi tre mesi del 2020, anche questa fornitura è stata utilizzata per identificare le persone decedute.

L'universo degli individui residenti in Italia occupati è stato individuato integrando la lista dei residenti con gli individui occupati derivati da RTL con riferimento all'ultima settimana di dicembre 2018, che rappresenta l'informazione più aggiornata relativa all'occupazione.

L'integrazione degli individui della lista con il *Registro Statistico dei Luoghi* ha consentito di definire i domini territoriali primari quali regioni e *Macro aree* di contagio e secondari quali le Provincie, i Sistemi Locali del Lavoro (SLL), Aziende Sanitarie Locali (ASL) e per il Comune di Milano Distretti Sanitari (DS). L'integrazione con il Registro dei luoghi ha anche permesso di collegare l'indirizzo di residenza anagrafica, che in RBI è espresso con un codice, a un indirizzo normalizzato fino al civico, dal momento che RSBL contiene tutto lo stradario nazionale.

Per quanto riguarda i domini strutturali le classi di età e il sesso sono presenti in RBI, mentre i raggruppamenti di Attività Economica nelle 4 classi: "Occupati sospesi", "Occupati non sospesi Altro", "Occupati non sospesi della Pubblica Amministrazione ed Istruzione", "Occupati non sospesi della Sanità", sono stati ricostruiti integrando attraverso il codice dell'unità economica relativa alla posizione prevalente gli occupati con il *Registro Base delle Unità Economiche*, permettendo di integrare l'informazione relativa all'attività economica principale del datore di lavoro, riferita all'anno 2018.

8. Il disegno di campionamento

8.1 Lo schema di campionamento

Il disegno di campionamento adottato per l'indagine di sieroprevalenza è a due stadi di selezione con stratificazione sia delle Unità di Primo Stadio (UPS) sia delle Unità di Secondo Stadio (USS). Le UP sono i comuni stratificati all'interno di ciascuna provincia in base alla loro dimensione demografica mentre le US sono gli individui stratificati sulla base di 6 classi di età (0-17; 18-34; 35-49; 50-59; 60-69 70-+), sesso e 4 macro-aggregazioni dell'attività economica (non occupati, occupati non sospesi del comparto PA e istruzione,

occupati non sospesi del comparto sanità, occupati non sospesi di altri comparti, occupati sospesi). La numerosità campionaria delle UP è pari a circa 2000 comuni, quasi il 25% dei comuni italiani, mentre la numerosità campionaria delle US è pari a circa 150,000 individui distribuiti su tutto il territorio nazionale.

Per ridurre al massimo le distorsioni e garantire una numerosità adeguata, la numerosità campionaria è stata ampliata del 30% per un totale di circa 195.000 unità.

Le scelte di base del disegno di campionamento possono essere così riassunte:

1. Domini di stima territoriale costituiti dalle Regioni e Province autonome di Bolzano e Trento.
2. Domini strutturali, all'interno di ciascuna Regione geografica, costituiti da 6 classi di età, 0-17; 18-34; 35 -49; 50-59; 60-69; 70 e più; sesso e maggiori e minori di 50 anni e da 4 macro-classi di ATECO (sotto-popolazione degli occupati);
3. Lo studio del campione è stato effettuato sulla base della stima delle prevalenze a livello provinciale fornite dall'ISS ad aprile 2020, con scelte conservative (intese come ipotesi di prevalenze più basse).
4. L'esigenza di gestire al meglio la raccolta delle informazioni sul campo e di tenere sotto controllo i costi di rilevazione, ha fatto prediligere un disegno di campionamento con schemi di selezione a due stadi stratificati; tale schema ha prodotto una buona dispersione spaziale del campione estratto sul territorio, in virtù del fatto che è stato selezionato un numero rilevante di comuni (circa 2.000).
5. E' stato inoltre assicurato che tutte le Aziende Sanitarie (USL) siano rappresentate nel campione selezionato, e che quasi tutti i 610 Sistemi Locali del Lavoro (SLL) siano inclusi nel campione stesso (ad eccezione di 82 SLL); i risultati dell'allocazione e successiva selezione hanno mostrato, in sintesi, una buona rappresentazione dei territori sub-regionali italiani, in rapporto alle prevalenze stimate e agli errori pianificati, e una soddisfacente copertura del campione a livello comunale, di Aziende Sanitarie e anche di SLL.

La metodologia di allocazione adottata per il campione, di tipo multivariato e multi-dominio per disegni stratificati a due stadi, è implementata in un pacchetto R.

Tutto il processo di disegno e selezione dei campioni è basato sull'uso integrato di importanti e consolidate fonti di informazioni esistenti in Istat, sia sulla popolazione italiana e sulle relative caratteristiche socio demografiche, sia sul loro stato occupazionale e attività professionale, sia sulla loro collocazione sul territorio italiano.

8.2 Definizione della dimensione campionaria e sua allocazione nei domini pianificati

L'allocazione del campione di individui e di comuni tra le varie regioni è stata determinata adottando la metodologia di allocazione ottima multivariata e multidominio per disegni stratificati a due stadi implementata nel pacchetto R R2BEAT (Falorsi *et al.*, 2020).

Il pacchetto consente, a partire da alcune informazioni (in questo caso la prevalenza stimata dall'ISS), di determinare la numerosità campionaria necessaria a livello di strato per soddisfare dei vincoli di precisione prefissati.

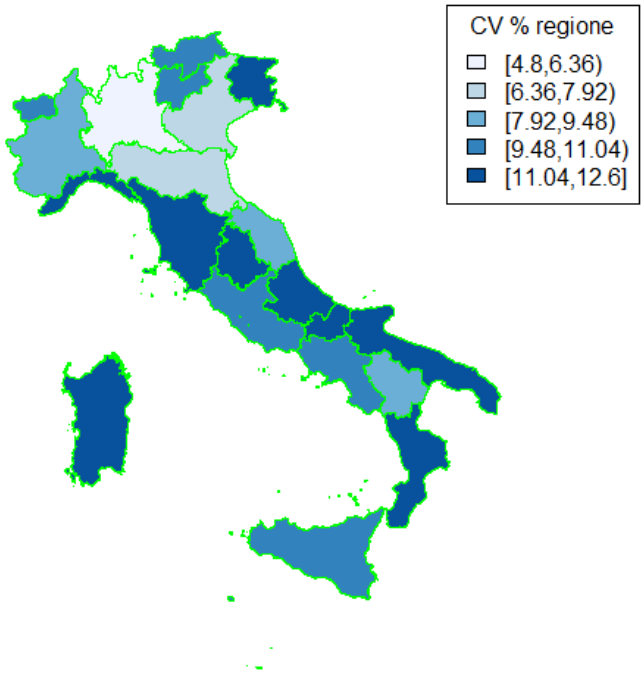
In genere l'allocazione ottima, a parità di precisione, attribuisce maggior campione agli strati con maggior numero di unità nella popolazione, con maggiore variabilità e livelli di stima più bassi. Nella Tavola 8.1 si riportano i livelli di errore prefissati con riferimento ai diversi domini pianificati. Nella Figura 8.1 si riporta la mappa con i CV% regionali attesi.

Tavola 8.1 - Livelli di errore prefissati a livello regionale

Dominio di stima		Prevalenze considerate	Errori fissati (CV%)
	Descrizione		
1	Zona rossa	2.0	12.1
2	Resto del Nord + Centro	1.5	11.1
3	Sud e Isole	1.0	10.1
4	Basilicata	1.0	15.0
5	Italia	1.0	2.3

Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

Figura 8.1 - Errori relativi attesi nelle regioni



Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

L'utilizzo *tout-court* delle prevalenze fornite dall'ISS avrebbe, a questo passo, l'effetto di sovradimensionare il campione nelle regioni con livelli di prevalenza più bassi, a discapito di quelli con prevalenze più alte. Per combattere questo effetto, o meglio, per ottenere in parte l'effetto opposto in modo da riuscire a fornire stime con dettaglio più fine in regioni maggiormente colpite dal *COVID-19*, si è deciso di definire livelli di precisione diversificati a seconda delle macro-aree. Inoltre, per preservare il più possibile la proporzionalità del campione è stata data maggiore importanza al vincolo di precisione relativo alle stime nazionali. *Infine, per ottenere la soluzione definitiva è stato necessario aggiungere un extra-dominio per trattare separatamente la Basilicata e contenere la numerosità campionaria regionale*. Nella Tavola 8.2, viene confrontata a livello regionale, l'allocazione ottima adottata con quella proporzionale.

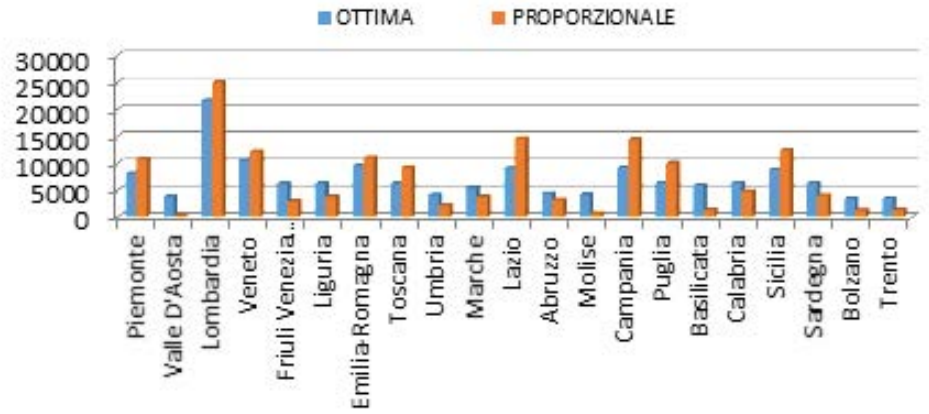
Tavola 8.2 - Distribuzione regionale del campione ottimale e proporzionale

Regione	Popolazione	Allocazione	
		Ottima	Proporzionale
Piemonte	4359336	8108	10793
Valle D'Aosta	126098	3795	312
Lombardia	10087648	21635	25035
Bolzano	530496	3505	1319
Trento	541262	3470	1344
Veneto	4913951	10581	12193
Friuli Venezia Giulia	1215537	6234	3010
Liguria	1550941	6238	3832
Emilia Romagna	4463320	9626	11058
Toscana	3732511	6257	9243
Umbria	883824	4161	2189
Marche	1526444	5405	3781
Lazio	5885023	9042	14607
Abruzzo	1312974	4401	3254
Molise	305741	4160	757
Campania	5815546	9223	14444
Puglia	4031023	6383	10005
Basilicata	562381	5873	1393
Calabria	1944003	6330	4823
Sicilia	5003819	8855	12412
Sardegna	1641298	6337	4072

Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

Come già detto e come si può vedere nella Tavola 8.2 e dalla Figura 8.2, l'allocazione regionale ottenuta è molto vicina all'allocazione proporzionale ma garantisce un minimo di 3.500 unità campionarie anche nelle regioni più piccole (Valle d'Aosta, Bolzano e Trento).

Figura 8.2 - Distribuzione regionale del campione ottimale e proporzionale



Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

8.3 Stratificazione e selezione delle unità campionarie di primo stadio

Nell'indagine in esame, i comuni sono stratificati (all'interno di ciascuna provincia) in base alla loro dimensione demografica e nel rispetto delle seguenti condizioni: (*i*) auto-ponderazione del campione a livello regionale; (*ii*) selezione di un comune campione nell'ambito di ciascuno strato definito sui comuni dell'insieme NAR; (*iii*) scelta di un numero minimo di individui da intervistare in ciascun comune campione (50 individui); (*iv*) formazione di strati aventi ampiezza approssimativamente costante in termini di popolazione residente. Il procedimento di stratificazione, attuato all'interno delle province di ciascuna regione geografica *r* (*r* =1,…,*R*), si articola nei seguenti passi:

1. determinazione di una soglia regionale, ${}_r\lambda$, di popolazione per la definizione dei comuni AR, mediante la relazione:

$$ {}_r\lambda = \frac{{}_r\bar{p}}{{}_rf} $$

in cui, si è indicato con: ${}_r\bar{p}$, il numero minimo di individui da intervistare per comune; ${}_rf$ il tasso di campionamento regionale in termini d i individui campione;

2. ordinamento dei comuni della provincia in ordine decrescente secondo la loro dimensione demografica in termini di popolazione residente;
3. effettuata la stratificazione, i comuni AR sono inclusi con certezza nel campione; per quanto riguarda, invece, i comuni NAR, nell'ambito di ogni strato è stato estratto un comune campione con probabilità proporzionale alla dimensione demografica, mediante la procedura di selezione sistematica proposta da Madow (Cochran, 1977).

Tavola 8.3 - Distribuzione regionale dei comuni e degli individui nell'universo e nel campione

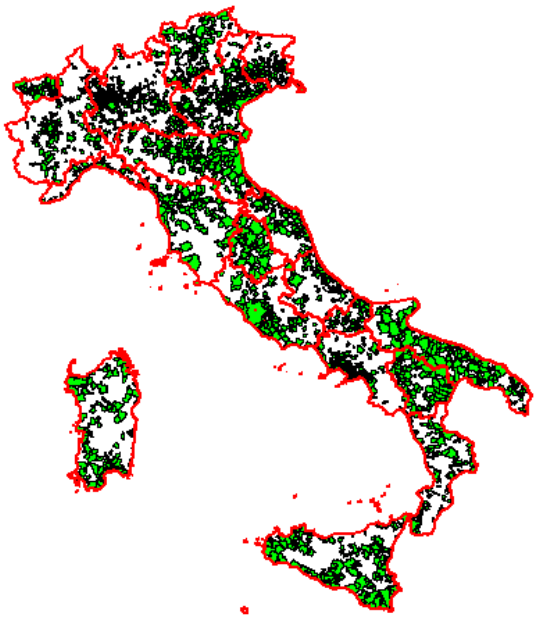
Regione	Comuni		Individui campione
	AR	NAR	
Piemonte	37	80	8108
Valle D'Aosta	29	13	3795
Lombardia	102	224	21635
Bolzano	20	28	3505
Trento	16	32	3470
Veneto	58	114	10581
Friuli Venezia Giulia	42	40	6234
Liguria	29	31	6238
Emilia Romagna	51	70	9626
Toscana	37	51	6257
Umbria	23	16	4161
Marche	36	39	5405
Lazio	35	44	9042
Abruzzo	26	36	4401
Molise	23	28	4160
Campania	63	70	9223
Puglia	44	52	6383
Basilicata	46	25	5873
Calabria	28	68	6330
Sicilia	59	61	8855
Sardegna	35	54	6337
Totale	839	1176	149619

Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

La selezione del campione di comuni è stata ottenuta attraverso il pacchetto R FS4[26].

La Figura 8.3 rappresenta la mappa dei comuni estratti all'interno delle singole regioni.

Figura 8.3 - Distribuzione spaziale dei comuni campionati



Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

8.4 Stratificazione e selezione delle unità campionarie di secondo stadio

Una volta individuata l'allocazione regionale ed estratti i comuni campione è stato utilizzato il metodo dei Quadrati Latini descritto in Cochran (1977) per garantire il rispetto delle due stratificazioni sub-regionali: stratificazione degli individui per regione, classe d'età, categoria ATECO e sesso e dei comuni per provincia e dimensione demografica.

Date le due stratificazioni è stata definito il numero di individui da rilevare di ogni strato (classe d'età, categoria ATECO e sesso) di ciascun comune campione in modo da garantire la coerenza con l'allocazione a livello provinciale.

Le unità campionarie di secondo stadio sono, poi, state estratte casualmente all'interno di ciascuno strato dei comuni estratti.

8.5 Calcolo e presentazione sintetica degli errori attesi

Il campione così definito consente di ottenere delle stime delle prevalenze con livelli di attendibilità medio-alta (CV% al di sotto del 10-15%) in tutte le regioni e per i diversi livelli di prevalenza ipotizzati nella Tavola 8.4. Questo si verifica anche nel caso in cui le prevalenze stimate dovessero essere intorno all'1%, ovvero la situazione con il più alto margine di errore.

26 First stage stratification and selection in sampling, disponibile nella repository Istat <https://www.istat.it/en/methods-and-tools/methods-and-it-tools/design/design-tools/fs4> (ultimo accesso 22 maggio 2020).

Tavola 8.4 - Coefficiente di variazione percentuali (CV%) attesi per diversi livelli di prevalenze per regione

Regione	Popolazione	Campione	Prevalenza						
			0.5%	1%	2%	3%	5%	7%	10%
Piemonte	4359336	8108	15.67	11.05	7.77	6.31	4.84	4.05	3.33
Valle D'Aosta	126098	3795	22.90	16.15	11.36	9.23	7.08	5.92	4.87
Lombardia	10087648	21635	9.59	6.76	4.76	3.87	2.96	2.48	2.04
Bolzano	530496	3505	23.83	16.81	11.82	9.60	7.36	6.16	5.07
Trento	541262	3470	23.95	16.89	11.88	9.65	7.40	6.19	5.09
Veneto	4913951	10581	13.71	9.67	6.81	5.53	4.24	3.54	2.92
Friuli Venezia Giulia	1215537	6234	17.87	12.60	8.87	7.20	5.52	4.62	3.80
Liguria	1550941	6238	17.86	12.60	8.86	7.20	5.52	4.61	3.80
Emilia Romagna	4463320	9626	14.38	10.14	7.13	5.80	4.44	3.72	3.06
Toscana	3732511	6257	17.83	12.58	8.85	7.19	5.51	4.61	3.79
Umbria	883824	4161	21.87	15.42	10.85	8.82	6.76	5.65	4.65
Marche	1526444	5405	19.19	13.53	9.52	7.73	5.93	4.96	4.08
Lazio	5885023	9042	14.84	10.46	7.36	5.98	4.58	3.83	3.15
Abruzzo	1312974	4401	21.26	15.00	10.55	8.57	6.57	5.49	4.52
Molise	305741	4160	21.87	15.43	10.85	8.82	6.76	5.65	4.65
Campania	5815546	9223	14.69	10.36	7.29	5.92	4.54	3.80	3.12
Puglia	4031023	6383	17.66	12.45	8.76	7.12	5.46	4.56	3.75
Basilicata	562381	5873	18.41	12.98	9.13	7.42	5.69	4.76	3.91
Calabria	1944003	6330	17.73	12.51	8.80	7.15	5.48	4.58	3.77
Sicilia	5003819	8855	14.99	10.57	7.44	6.04	4.63	3.87	3.19
Sardegna	1641298	6337	17.72	12.50	8.79	7.14	5.48	4.58	3.77

Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

9. Analisi della copertura del campione estratto nei domini territoriali non-pianificati

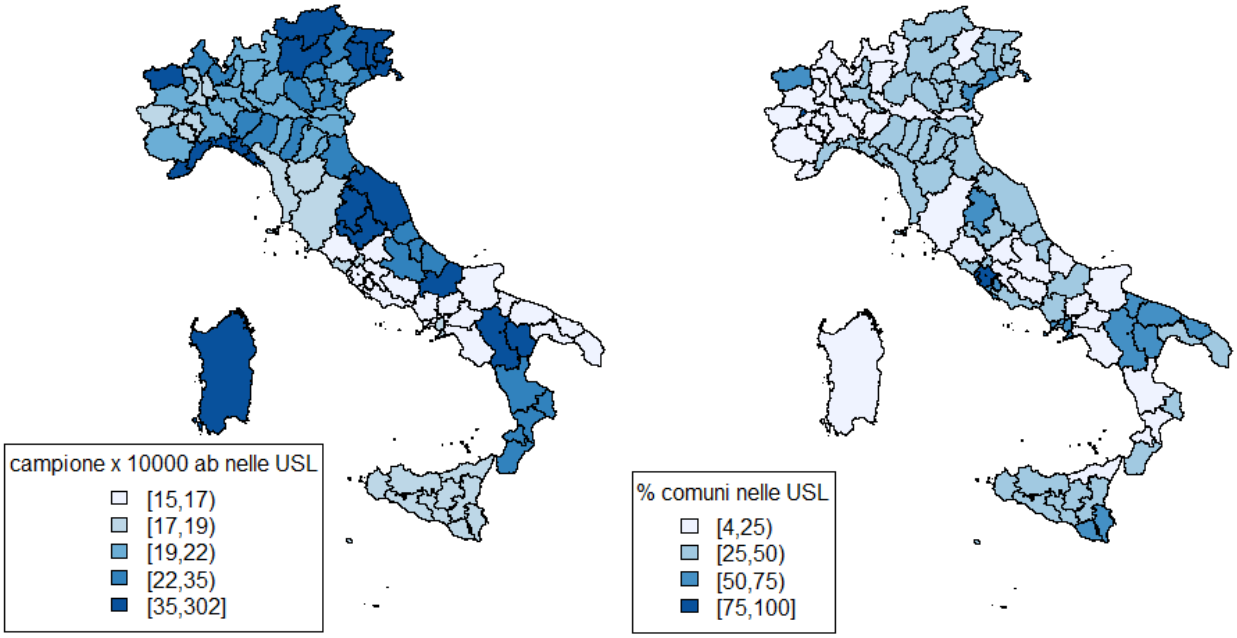
Il campione è stato estratto cercando di ottimizzare la copertura territoriale anche per domini non pianificati. Tali domini di tipo non ammnistrativo sono stati ritenuti rilevanti per lo studio del fenomeno oggetto di rilevazione per la loro natura. In particolare, sono state considerate le Aziende Sanitarie Locali (ASL) in quanto deputate all'erogazione di servizi sanitari e, per tale ragione, è stata considerata la partizione del territorio più idonea all'analisi dell'evoluzione dell'epidemia oggetto di analisi.

Come è possibile notare dalla figura 9.1, tutte le ASL del territorio sono state selezione nel campione, anche quelle del comune di Roma che nel suo territorio ha più di una ASL e che, pertanto ha una copertura del campione in termini di comuni pari al 100%. Nelle restanti ASL si registra una buona copertura dei comuni che le compongono. In termini di individui campionati, le ASL del Nord hanno avuto una copertura maggiore rispetto a quelle del centro sud.

L'idea alla base è stata quella di poter eventualmente applicare metodi di stima più complessi (modelli di stima per piccole aree) per la sieroprevalenza a livello di ASL. A tele proposito occorre ricordare che l'andamento della rilevazione con i tassi di risposta realizzati se da un lato avrebbe richiesto l'applicazione di metodi di stima per piccole aree, per effetto della riduzione delle numerosità campionarie, dall'altro ha reso maggiormente problematica la loro applicazione in quanto le stime basate sui soli dati campionari erano potenzialmen-

te affette da *selection bias*, aspetto che complica notevolmente l'applicazione di questo tipo di metodologie. Ciò ha reso più cauti circa l'effettiva applicabilità di tali metodi in fase di stima.

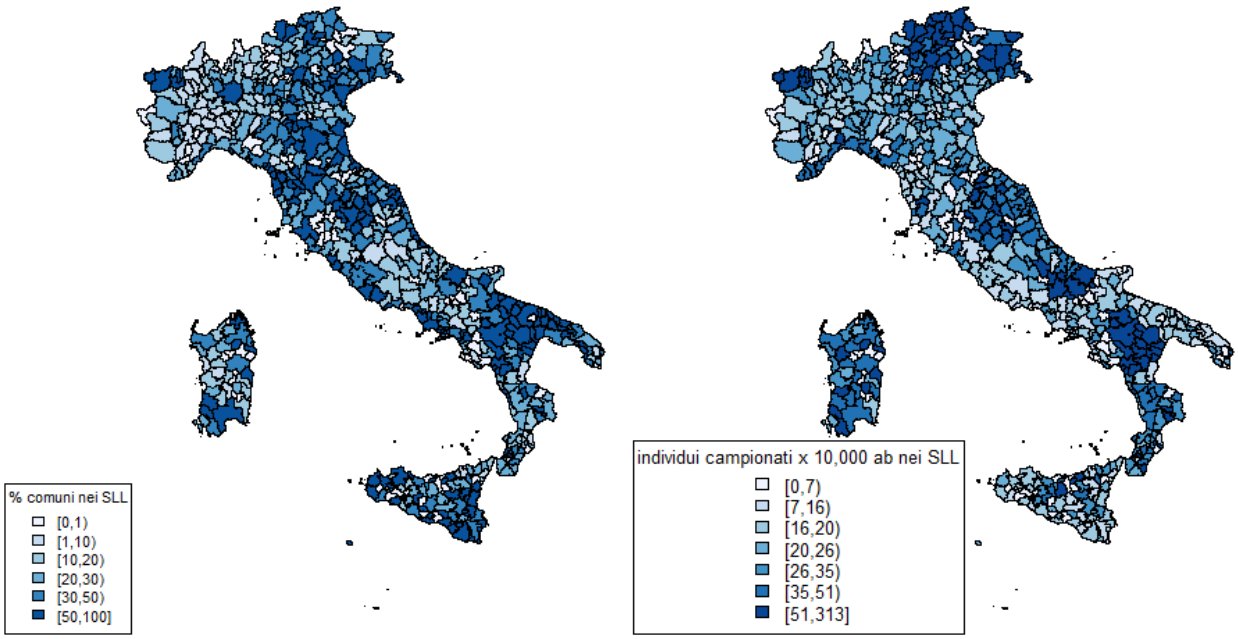
Figura 9.1 - Distribuzione campionaria degli individui e dei comuni campione per ASL



Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

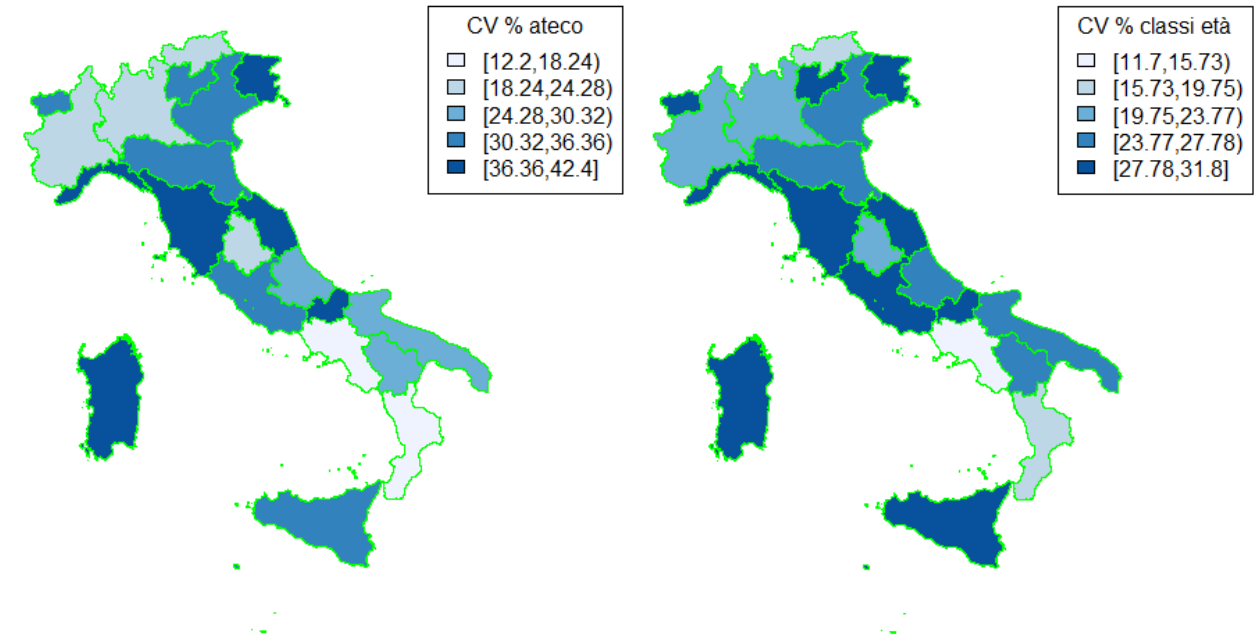
Sempre con l'idea di poter calcolare stime per piccole area senza avere un impatto sulla dimensione campionaria, nella progettazione del campione dell'indagine è stato deciso di fornire una copertura campionaria soddisfacente anche a livello di Sistema Locale del Lavoro (SLL). I SLL sono delle unità territoriali di natura non amministrativa i cui confini sono definiti utilizzando i flussi degli spostamenti giornalieri casa/lavoro tra comuni. Data la loro natura funzionale si è ipotizzato che fossero le aree territoriali particolarmente adatte a verificare come gli spostamenti degli individui tra unità territoriali potesse essere alla base dell'evoluzione dell'epidemia sul territorio. Infatti, i SLL sono quelle aree dove gli individui esercitano la maggior parte delle relazioni sociali ed economiche. Come è possibile notare dalla figura 9.2 la maggior parte dei SLL locali hanno una copertura campionaria, sia in termini di individui, sia in termini di comuni campionati al loro interno. Tale distribuzione del campione ben si presta alla possibilità di stime per piccole aree anche con modelli che prevedano la correlazione spaziale tra le aeree.

Figura 9.2 - Distribuzione campionaria degli individui e dei comuni campione per SLL



Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

Figura 9.3 - Distribuzione regionale degli errori relativi mediani attesi per classi ATECO e classi di età



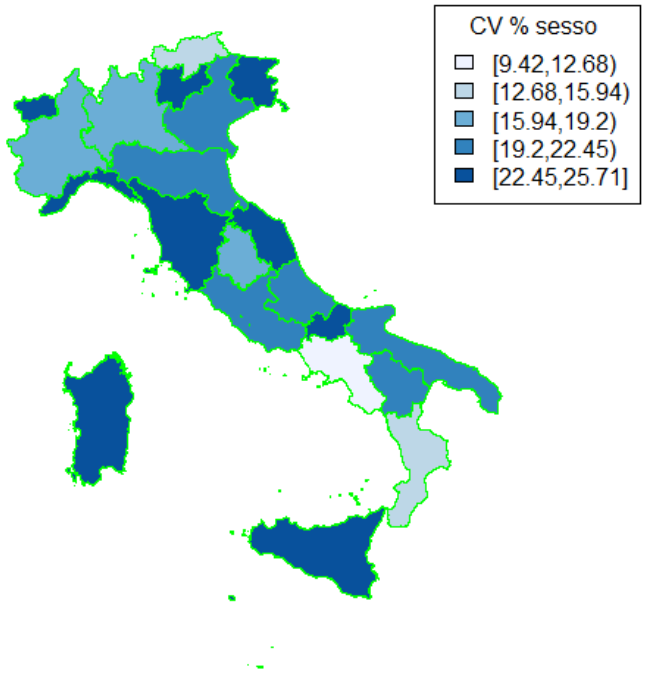
Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

Nella figura 9.3, si riporta la distribuzione degli errori relativi mediani attesi per classi di ATECO e per classi di età all'interno delle singole regioni. Come si può vedere dalle mappe, gli errori relativi mediani attesi sono per alcune regioni troppo elevati. Anche in questo caso è possibile applicare metodi di stima per piccole aree al fine di aumentare l'efficienza delle stime per i domini dati rispettivamente dall'incrocio delle regioni con le

classi ATECO e con le principali classi d'età. L'obiettivo è quello analizzare da un punto di vista epidemiologico la trasmissione della malattia per ciascuna regione all'interno di specifiche categorie di attività e per specifiche classi di età.

Nella figura 9.4 sono riportati i valori mediani attesi percentuali dell'errore relativo per i domini non pianificati dati dall'incrocio delle regioni con le due modalità del sesso. A differenza dei valori dei CV attesi riportate nelle mappe della figura 9.3, in questo caso anche le stime dirette potrebbero fornire per circa la metà delle regioni stime sufficientemente affidabili.

Figura 9.4 - Distribuzione regionale degli errori relativi mediani attesi per sesso



Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

PARTE III

METODOLOGIE DI ELABORAZIONE DATI E CALCOLO DELLE STIME DELL'INDAGINE DI SIEROPREVALENZA SUL *SARS-COV-2*[27]

27 Michele D'Alò, Claudia De Vitiis, Stefano Falorsi, Andrea Fasulo, Danila Filipponi, Alessio Guandalini, Francesca Inglese, Roberta Radini, Enrico Orsini.

10. Introduzione

In questo parte si descrivono le diverse fasi del processo produttivo dedicate al riporto dei dati campionari all'universo e alla loro diffusione. Com'è noto, infatti, il processo standard di elaborazione dei dati di tutte le indagini campionarie su larga scala prevede una serie di *step* successivi a partire dal *dataset* dei dati definitivi, che sono già stati sottoposti a una fase di controllo e correzione. Nel *dataset* dei dati definitivi sono riportate, oltre alle informazioni da registro provenienti dal SIR, utilizzate per progettare l'indagine, anche tutte le informazioni raccolte con l'indagine sulle unità rispondenti. Inoltre, per tutte le unità non osservate appartenenti al campione teorico, il *dataset* contiene le informazioni di dettaglio sul mancato contatto o la mancata risposta. A ciascuna unità del campione teorico è, inoltre, assegnato il corrispondente *peso diretto* legato alla sua probabilità di inclusione nel campione dipendente dal disegno campionario adottato.

Per l'indagine in oggetto, la procedura di riporto all'universo ha previsto due fasi successive. Nella prima fase, il peso diretto di ciascuna unità rispondente, è stato modificato attraverso un processo di correzione per mancata risposta totale, per rappresentare anche le unità non rispondenti. A seguire si è definito il sistema dei pesi finali attraverso un processo di *calibrazione* dei pesi già corretti per mancata risposta totale in base a una o più variabili ausiliare, legate a quelle di interesse, i cui totali di popolazione sono noti da fonti esterne all'indagine (Registri o fonti amministrative). La calibrazione se da un lato porta a stime generalmente più *efficienti*, in termini di variabilità campionaria rispetto a quelle non calibrate – nella misura in cui le variabili ausiliarie sono legate a quelle di interesse – dall'altro garantisce l'uguaglianza tra un insieme di distribuzioni di popolazione note e le corrispondenti stime campionarie finali. Il processo usualmente applicato per la maggior parte delle indagini campionarie dell'Istat prevede un unico passo di correzione dei pesi diretti mediante calibrazione finalizzato a correggere per mancata risposta totale, a tenere sotto controllo una serie di distribuzioni di popolazione note e a migliorare l'efficienza delle stime. Per l'indagine in oggetto, tuttavia, data l'entità del fenomeno della mancata risposta totale e data la delicatezza e la rilevanza delle stime uscenti dall'indagine per la collettività nazionale si è deciso di curare al meglio il processo di riporto dei dati. Per questo motivo sono stati messi in campo due processi separati. Il primo relativo alla correzione per mancata risposta totale, in base alle informazioni disponibili sul campione teorico; il secondo di calibrazione. Infine a seguire sono state calcolate le stime di prevalenza previste dal piano di tabulazione e i modelli per la presentazione sintetica degli errori campionari. In base a detti modelli, a partire dal valore della stima, l'utente è in grado di calcolare in autonomia gli errori campionari e i relativi intervalli di confidenza anche per altre stime di prevalenza non previste inizialmente dal piano di tabulazione.

11. Analisi e trattamento della mancata risposta totale

Nonostante le strategie messe in campo per la riduzione dell'errore non campionario, l'indagine di sieroprevalenza sul *SARS-COV-2* è stata caratterizzata da un tasso di mancata risposta totale alquanto elevato. La percentuale dei rispondenti all'indagine a livello nazionale è pari a circa il 38% del campione iniziale. Inoltre, non tutti i rispondenti all'indagine si sono sottoposti al test sierologico, variabile *target* dello studio epidemiologico, ma soltanto circa il 34% dell'intero campione. Tale risultato è stato determinato dalla combinazione di più cause: il rifiuto di collaborare all'indagine espresso dagli individui contattati e il rifiuto di una parte degli individui intervistati a effettuare il test sierologico.

Come è noto, la mancata osservazione del fenomeno indagato per una parte delle unità campionarie selezionate comporta una riduzione dell'accuratezza complessiva delle stime finali, determinata sia dall'aumento della varianza di campionamento sia dall'introduzione di effetti distorsivi. Quest'ultimi sono tanto più gravi quanto più i rispondenti differiscono sistematicamente dai non rispondenti, rispetto alle caratteristiche di interesse.

Al fine di analizzare le caratteristiche dei rispondenti, sono stati calcolati i tassi di risposta declinati per le variabili a disposizione. Si sottolinea che sono stati considerati come rispondenti tutti i rispondenti all'indagine che si sono anche sottoposti e che hanno avuto un esito al test sierologico pari a circa 66 mila individui.

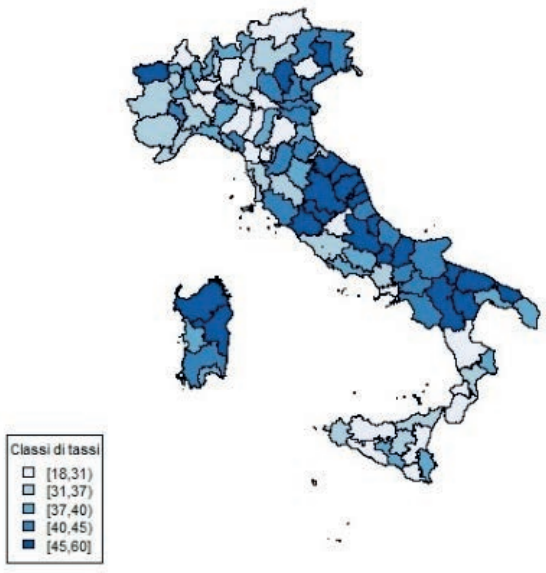
Per l'acquisizione delle variabili ausiliarie per l'analisi della non risposta si è fatto riferimento al *Sistema Integrato dei Registri* (SIR) disponibile in Istituto. In particolare sono stati utilizzati i seguenti registri del SIR:

- *Registro Statistico di Base degli individui, delle famiglie e delle convivenze* (RBI) per l'acquisizione di informazioni anagrafiche di base (età, sesso, luogo di nascita, cittadinanza), il livello di istruzione e luogo di residenza.
- *Registro Statistico tematico sul lavoro* (RTL) per l'acquisizione delle posizioni lavorative attive nell'anno di riferimento del registro con le relative caratteristiche.
- *Registro base dei Luoghi* (RBL) per l'acquisizione delle informazioni relative alle diverse tipologie di unità amministrative territoriali, come comune, provincia e regione.

Ulteriori informazioni disponibili a livello comunale sono state acquisite sui flussi di nati-mortalità dall'*Anagrafe tributaria* e sul numero di casi accertati per infezione da *SARS-COV-2* dall'inizio della pandemia a maggio per comune forniti dall'Istituto Superiore di Sanità.

La figura 1 riporta i tassi di risposta all'indagine per provincia di residenza dell'individuo. È possibile notare una forte variabilità territoriale dei tassi. In particolare si evidenziano bassi tassi di risposta nelle province del nord, nelle quali si è verificato un alto valore di casi di positività cumulati nei mesi da marzo a maggio, e nelle province della Calabria e Sicilia. Più alti tassi di risposta, sebbene sempre inferiori al 50% a esclusione delle Marche e della Basilicata, si osservano invece nelle altre regioni del centro-sud.

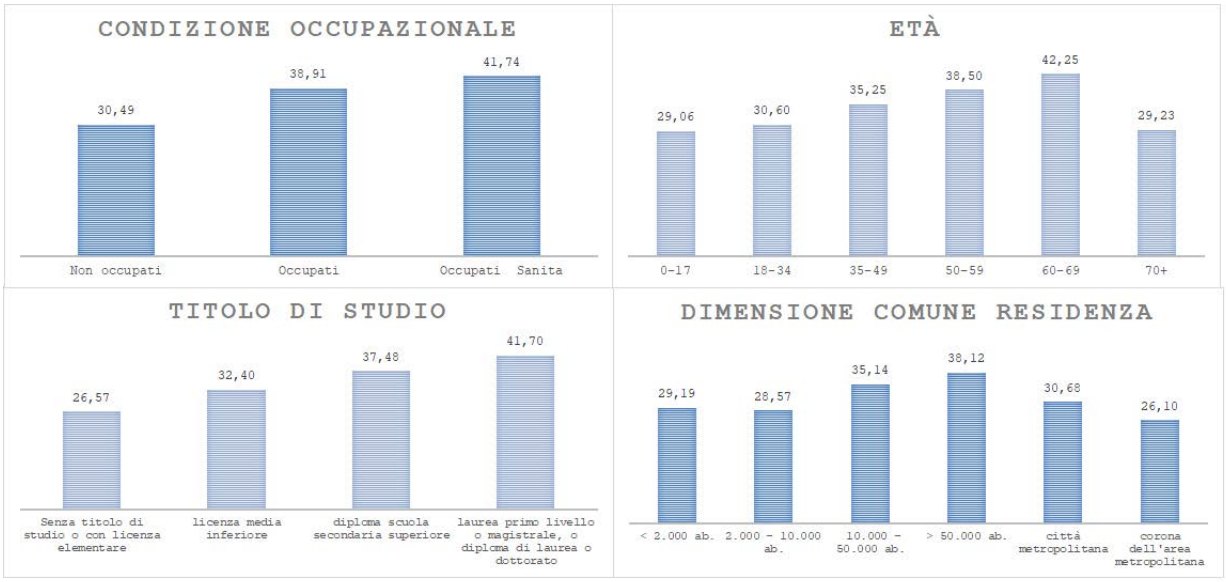
Figura 11.1 - Tassi di risposta indagine sierologica per provincia



Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

La figura 11.2 riporta i tassi di risposta per classe di età, condizione occupazionale, titolo di studio e tipo di comune di residenza. Come spesso accade nelle indagini sociali, la propensione alla risposta è minore per gli individui con un basso titolo di studio, per i non occupati e nelle aree metropolitane. Nello specifico dell'indagine invece, è interessante notare che il tasso di risposta è più alto negli occupati della sanità rispetto al resto degli occupati, spiegabile da una maggiore sensibilità all'argomento da parte degli operatori sanitari, e più basso per gli individui sotto i 17 anni e maggiori di 70 anni, probabilmente spiegabile da una reticenza a effettuare test sierologici nei bambini e negli anziani, e un basso tasso di risposta nei piccoli comuni, determinato dalla maggiore difficoltà nel raggiungere i centri di prelievo.

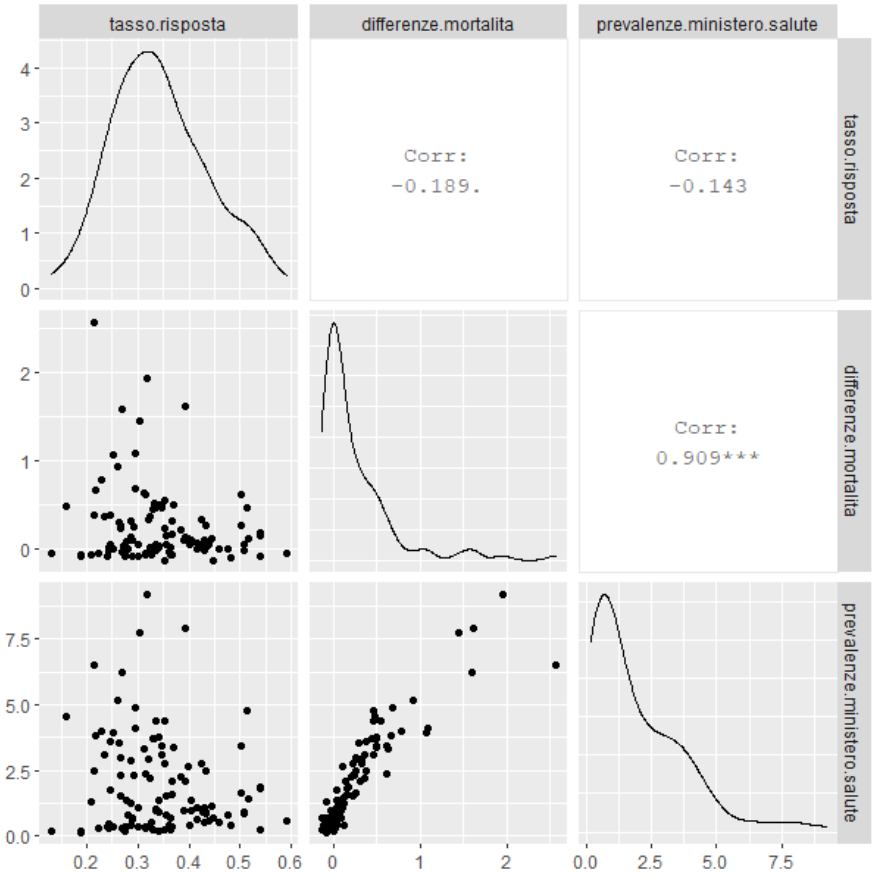
Figura 11.2 - Tassi di risposta per condizione occupazionale, titoli di studio, età e tipologia di comune di residenza



Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

Infine, la figura 11.3 mostra la relazione tra i tassi di risposta provinciali, il numero di casi accertati per infezione da *SARS-COV-2* dall'inizio della pandemia a maggio forniti dall'Istituto Superiore di Sanità per provincia e le variazioni percentuali dei decessi provinciali tra il 2019 e il 2020 registrati all'anagrafe tributaria nei mesi di marzo e maggio.

Figura 11.3 - Scatterplot che mostra le relazioni tra i tassi di risposta, le infezioni da *SARS-COV-2* del ministero della salute e le variazioni percentuali della mortalità 2019-2020



Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

Il grafico mostra un'assenza di relazione tra i tassi di risposta provinciale e le variabili associate ai livelli di sieroprevalenza quali le prevalenze provinciali misurate dal Ministero della Salute e i livelli di mortalità. Inoltre evidenzia, come da aspettarsi, una forte relazione positiva tra le prevalenze e la mortalità a livello provinciale.

Per eliminare gli effetti distorsivi, o almeno attenuarli, la mancata risposta totale osservata per l'indagine di sieroprevalenza è stata trattata in fase di stima con una metodologia che consiste nell'aggiustamento dei pesi campionari iniziali associati alle unità rispondenti in modo da tener conto delle unità non rispondenti. Per lo sviluppo di detta metodologia sono state utilizzate tecniche di riponderazione basate sul "*response propensity method*" (Rosenbaum e Rubin, 1984; Bethlehem *et al.*, 2011). Per l'applicazione del metodo è necessario disporre di informazione ausiliaria nota sia sui rispondenti sia sui non rispondenti all'indagine e stimare – tramite un buon modello di risposta – le probabilità di risposta individuali che sono funzionali alla costruzione di fattori correttivi della mancata risposta totale.

La procedura che ha condotto all'aggiustamento dei pesi campionari per la mancata risposta totale dell'indagine è stata sviluppata in più passi:

- Acquisizione di variabili disponibili per la definizione del modello di risposta.
- Studio dei modelli di risposta e scelta del *working-model*.
- Stima delle probabilità individuali di risposta per la costruzione di fattori correttivi.

11.1 Studio dei modelli di risposta e scelta del *working-model*

Le variabili ausiliarie acquisite a livello individuale tramite i registri disponibili opportunamente riclassificate e quelle acquisite a livello aggregato per comune sono state utilizzate come covariate nello studio dei modelli per la stima delle probabilità individuali di risposta.

La relazione tra la variabile dipendente risposta (R_i =1 per le unità rispondenti e R_i =0 per le unità non rispondenti) e le p variabili (ausiliarie) indipendenti (X_{i1}, …, X_{ij}, …, X_{ip}) è stata studiata tramite regressioni di tipo logistico.

I modelli studiati sono stati valutati sulla base di alcuni principali indicatori e test di significatività:

- Percentuale di Concordanti – per valutare la capacità di stimare la probabilità del verificarsi della risposta.
- Indicatore *Akaike Information Criterion* (AIC) – per valutare la bontà di adattamento ai dati.
- Test di significatività congiunta dei coefficienti (*Likelihood ratio test/score test/Wald test*) – per valutare la capacità esplicativa dei modelli.
- Test di significatività dei singoli coefficienti (*Wald Chi-Square test*) – per valutare la significatività dei singoli coefficienti, ovvero la rilevanza dei corrispondenti regressori nella spiegazione della variabile dipendente.

Sulla base dei suddetti criteri è stato scelto il *working-model* per la stima delle probabilità individuali da utilizzare nella costruzione dei fattori correttivi della mancata risposta totale.

I predittori della probabilità di risposta utilizzati nel modello sono costituiti da:

- 21 regioni geografiche (Bolzano e Trento sono state trattate distintamente);
- 6 tipologie comunali (città metropolitana; corona dell'area metropolitana; minore di 2000 abitanti; tra 2000 e 10000 abitanti; tra 10000 e 50000 abitanti; oltre 50000 abitanti);
- sesso;
- sei classi d'età (0-17; 18-34; 35-49; 50-59; 60-69; 70+);
- quattro classificazioni dello stato ATECO (occupati sospesi, occupati non sospesi PA + Istruzione, occupati non sospesi sanità, altri occupati non sospesi, non occupati);
- 8 modalità del titolo di studio (Analfabeti, Alfabeti privi di titolo di studio, Licenza di scuola elementare, Licenza di scuola media inferiore, Diploma di scuola secondaria superiore, Laurea o Diploma accademico di I livello, Laurea magistrale/specialistica o Diploma accademico II livello, Dottorato di ricerca);
- tasso di positività comunali, stimato sulla base dei contagi cumulati dall'inizio della pandemia a maggio (previsioni fornite dall'Istituto Superiore di Sanità);
- differenza percentuale dei tassi di mortalità comunali rispetto allo stesso periodo dell'anno precedente;
- numero di tentativi di contatto.

La percentuale di concordanti del *working-model* è buona ed è pari al 71,2 per cento, mentre la bontà di adattamento del modello ai dati misurata tramite l'AIC è pari a 223684,29 (il valore più basso rispetto ad altri modelli studiati). I parametri stimati dal modello logistico sono riportati nella tavola 11.1.

Tavola 11.1 - Stima dei parametri e test di significatività dei singoli coefficienti del modello (livello di significatività 0,05)

Parametro		DF	Stima	Errore standard	Chi-quadrato di Wald	Pr > Chi-quadrato
Intercept		1	-0,0266	0,0178	2,2224	0,1360
COD_REG	**1**	1	-0,1900	0,0238	63,8832	<.0001
COD_REG	**2**	1	0,5933	0,0313	360,2814	<.0001
COD_REG	**3**	1	-0,4241	0,0171	617,6200	<.0001
COD_REG	**5**	1	0,2896	0,0190	232,6833	<.0001
COD_REG	**6**	1	0,00316	0,0236	0,0180	0,8933
COD_REG	**7**	1	0,1490	0,0254	34,5061	<.0001
COD_REG	**8**	1	-0,1577	0,0212	55,1055	<.0001
COD_REG	**9**	1	-0,1281	0,0249	26,4376	<.0001
COD_REG	**10**	1	0,5541	0,0289	368,7284	<.0001
COD_REG	**11**	1	0,9530	0,0262	1321,3799	<.0001
COD_REG	**12**	1	-0,2054	0,0222	85,3483	<.0001
COD_REG	**13**	1	0,1883	0,0277	46,1639	<.0001
COD_REG	**14**	1	0,3416	0,0295	134,1401	<.0001
COD_REG	**15**	1	-0,2934	0,0217	182,2538	<.0001
COD_REG	**16**	1	0,6011	0,0253	565,6970	<.0001
COD_REG	**17**	1	0,2946	0,0245	145,0906	<.0001
COD_REG	**18**	1	-0,7006	0,0264	701,9042	<.0001
COD_REG	**19**	1	-0,5864	0,0239	602,6719	<.0001
COD_REG	**20**	1	0,0235	0,0244	0,9231	0,3367
COD_REG	**41**	1	-0,9650	0,0361	714,5947	<.0001
dom6	**1**	1	-0,2825	0,0142	394,1448	<.0001
dom6	**2**	1	-0,0602	0,0148	16,4478	<.0001
dom6	**3**	1	0,1131	0,0176	41,5388	<.0001
dom6	**4**	1	0,0768	0,0102	56,1730	<.0001
dom6	**5**	1	0,0871	0,0103	71,9781	<.0001
NUM_TENTATIVI_CONTAT		1	-0,1804	0,00189	9108,3343	<.0001
STATO_ATECO	**-2**	1	-0,2895	0,0112	670,0118	<.0001
STATO_ATECO	**0**	1	-0,0394	0,0138	8,1456	0,0043
STATO_ATECO	**1**	1	0,0607	0,0124	24,1936	<.0001
STATO_ATECO	**2**	1	0,2583	0,0184	196,7700	<.0001
C_ETA	**1**	1	0,5094	0,0187	744,1972	<.0001
C_ETA	**2**	1	-0,3690	0,0126	864,2164	<.0001
C_ETA	**3**	1	-0,1814	0,0120	230,5118	<.0001
C_ETA	**4**	1	-0,0149	0,0125	1,4242	0,2327
C_ETA	**5**	1	0,2124	0,0127	279,8934	<.0001
SESSO	**F**	1	0,0393	0,00524	56,4232	<.0001
PANEL		1	0,4644	0,0170	748,0251	<.0001
Prevalenze		1	-0,5899	0,3296	3,2023	0,0735
diff_perc		1	-0,0109	0,00532	4,2059	0,0403
COD_TITOLO_STUDIO_S8	**0**	1	-0,7201	0,0289	621,1614	<.0001
COD_TITOLO_STUDIO_S8	**1**	1	-0,9696	0,0822	139,1698	<.0001
COD_TITOLO_STUDIO_S8	**2**	1	-0,4033	0,0301	180,0681	<.0001
COD_TITOLO_STUDIO_S8	**3**	1	-0,1633	0,0197	68,5564	<.0001
COD_TITOLO_STUDIO_S8	**4**	1	0,1102	0,0170	42,0051	<.0001
COD_TITOLO_STUDIO_S8	**5**	1	0,4073	0,0170	571,1895	<.0001
COD_TITOLO_STUDIO_S8	**6**	1	0,6215	0,0281	490,0300	<.0001
COD_TITOLO_STUDIO_S8	**7**	1	0,5331	0,0207	665,0067	<.0001

Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

11.2 Stima delle probabilità individuali di risposta per la costruzione di fattori correttivi

Le probabilità individuali predette, tramite il modello descritto, sono state utilizzate per la costruzione dei fattori di aggiustamento della mancata risposta totale sia direttamente sia indirettamente: nel primo caso i fattori correttivi sono stati calcolati come inverso delle probabilità individuali predette (*response propensity weighting*); nel secondo caso le probabilità individuali predette sono state utilizzate per la costruzione di strati o celle di aggiustamento (*response propensity stratification*). La costruzione delle celle è stata effettuata con la tecnica dei quantili delle probabilità predette e il fattore correttivo della mancata risposta totale è stato calcolato in ogni cella come inverso del tasso di risposta.

Al fine di tenere sotto controllo il rischio di aumentare eccessivamente la variabilità dei pesi campionari, con conseguenze negative sull'efficienza delle stime, sono stati utilizzati criteri diversi per la costruzione delle celle di aggiustamento. In particolare, oltre al fattore correttivo ottenuto tramite il metodo *response propensity weighting*, sono stati considerati i decili, i quintili e i quartili delle distribuzioni delle probabilità individuali predette dal *working-model* definiti sia sull'intero campione teorico sia sui singoli campioni regionali.

Per valutare differenti soluzioni è stata utilizzata la statistica di Kish (1992), che è una misura dell'impatto della maggiore variabilità dei pesi campionari corretti per mancata risposta sulla varianza delle stime. L'analisi comparativa, oltre a permettere la scelta del fattore correttivo con il minor impatto sulla variabilità dei pesi campionari, conferma la robustezza del metodo utilizzato per la correzione della mancata risposta totale dell'indagine.

Il peso campionario aggiustato per mancata risposta totale con un valore più contenuto della statistica di Kish è stato quello ottenuto a partire da classi di aggiustamento determinate sulla base dei quintili delle distruzioni regionali delle probabilità individuali di risposta predette dal *working-model*.

Di seguito si riportano i risultati di alcune statistiche, e i valori dell', calcolate sui pesi campionari da disegno e sui pesi campionari corretti secondo due differenti tecniche di costruzione delle celle di ponderazione (Tavola 11.2).

Tavola 11.2 - Statistiche calcolate sui pesi campionari e sui pesi campionari corretti per mancata risposta totale

	Media	Mediana	Minimo	Massimo	$1 + CV^2$
Pesi campionari	288.43	231.30	0.94	5629.78	1.781168
***Tecnica*:**					
Decili della distribuzione del campione nazionale	778.39	545.62	1.31	16934.94	2.112526
Quintili delle distribuzioni regionali del campione	769.29	543.87	1.39	14533.33	2.082994

Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

12. Definizione dei pesi finali mediante calibrazione

Le stime prodotte per l'indagine sono principalmente stime di frequenze assolute o di frequenze relative riferite per diversi domini (nazionale, regionale, provinciale, area *COVID*, stato ATECO, sesso e classi d'età e alcuni incroci di questi).

Definendo la variabile *Y* come una variabile dicotomica che, dunque, sulla generica unità *k* ($k = 1, \dots, N$) assume valore 1 se l'unità possiede la caratteristica *Y* e 0 altrimenti, la frequenza assoluta può essere scritta come il totale della variabile *Y* nel domino U_d:

$$t_{Y_d} = \sum_{k \in U_d} y_k.$$

La frequenza relativa, dunque, può essere vista come la media del carattere *Y* nel dominio U_d: ed è ottenuta dividendo t_{Y_d} per la numerosità della popolazione *U*:

$$\mu_{Y_d} = \frac{\sum_{k \in U_d} y_k}{N_d} = \frac{t_{Y_d}}{N_d}.$$

Le stime di queste quantità sono ricavate attraverso lo stimatore calibrato (cfr. Deville, Särndal, 1992; Särndal, 2007; Tillé, 2019) che costituisce il principale metodo di stima correntemente utilizzato nella maggior parte delle indagini Istat.

Lo stimatore calibrato del totale è definito come:

$$\hat{t}_{Y_{CAL}} = \sum_{k \in R} y_k \, w_k$$

dove i pesi finali sono determinati attraverso la risoluzione di un problema di minimo vincolato così definito:

$$\begin{cases} min\left\{\sum_{k \in \mathrm{R}} dist(d_k, w_k)\right\} \\ \sum_{k \in \mathrm{R}} \boldsymbol{x_k} \, w_k = \boldsymbol{t_X} \end{cases}$$

in cui d_k è il peso da disegno relativo all'unità *k*-esima che deriva dall'inverso della probabilità di inclusione dell'unità nel campione e dalla procedura di correzione per mancata risposta; $\boldsymbol{t_X}$ è il vettore dei delle variabili di calibrazione e $\boldsymbol{x_k}$ è il vettore delle variabili ausiliarie osservate sulla *k*-esima unità dei rispondenti.

I pesi w_k così ottenuti garantiscono la coerenza con i totali noti delle variabili ausiliarie considerate e, rispetto a una opportuna funzione di distanza prescelta, sono il più vicino possibile ai pesi da disegno. In pratica, il peso indica il numero di unità della popolazione rappresentata dalla generica unità campionaria *k*. Per esempio, se a un'unità campionaria viene attribuito un peso pari a 30, ciò indica che questa unità rappresenta se stessa e altre 29 unità della popolazione che non sono state incluse nel campione.

La procedura di calibrazione è stata svolta con il pacchetto ReGenesees (Zardetto, 2015) implementato in ambiente R. Nella calibrazione si è tenuto conto dei seguenti totali di popolazione ricavati dal registro base degli individui (RBI):

(i) distribuzione regionale della popolazione per sesso e classe d'età (0-17; 18-34; 35-49; 50 59; 60-69; 70+);
(ii) distribuzione regionale della popolazione per sesso e stato della ATECO (occupati sospesi; occupati non sospesi, altro; occupati non sospesi, PA + Istruzione; occupati non sospesi, sanità; non occupati);
(iii) distribuzione provinciale della popolazione;
(iv) distribuzione regionale della popolazione per cittadinanza;
(v) distribuzione della popolazione per ripartizione (Nord-Ovest, Nord-Est, Centro e Mezzogiorno), classe d'età (0-17; 18-34; 35-49; 50 59; 60-69; 70+) e titolo di studio (4 livelli).

Il numero complessivo di totali (vincoli) considerati è 11100. La funzione di distanza utilizzata è la funzione logaritmica troncata con estremi fissati a 0.74 e 6.50.

13. Calcolo e presentazione sintetica della variabilità campionaria delle stime

13.1 Calcolo degli indicatori assoluti e relativi di variabilità campionaria

Al fine di valutare l'accuratezza delle stime prodotte dall'indagine è necessario tenere conto dell'errore campionario che deriva dall'aver osservato la variabile di interesse solo su una parte (campione) della popolazione. Tale errore può essere espresso in termini di errore assoluto (standard error)

$$\hat{\sigma}(\hat{t}_{Y_d}) = \sqrt{\widehat{Var}(\hat{t}_{Y_d})}$$

o di errore relativo (cioè l'errore assoluto diviso per la stima, che prende il nome di coefficiente di variazione, CV)

$$\hat{\varepsilon}(\hat{t}_{Y_d}) = \frac{\sqrt{\widehat{Var}(\hat{t}_{Y_d})}}{\hat{t}_{Y_d}}$$

che spesso viene riportato in valore percentuale (CV%).

Gli errori campionari delle espressioni (1) e (2), consentono di valutare il grado di precisione delle stime; inoltre, l'errore assoluto permette di costruire l'intervallo di confidenza di livello $1-\alpha$, che, quindi, con probabilità $1-\alpha$ contiene il parametro d'interesse. Con riferimento alla generica stima tale l'intervallo di confidenza di livello $1-\alpha$ è:

$$\mathrm{IC}_{1-\alpha} = \left[\hat{t}_{Y_d} - k\,\hat{\sigma}(\hat{t}_{Y_d});\ \hat{t}_{Y_d} - k\,\hat{\sigma}(\hat{t}_{Y_d})\right],$$

ove *k*, nel caso di intervalli di confidenza al 95%, è 1.96 ovvero, pari al valore del ($1-\alpha$ /2) percentile della normale standard.

Di seguito è riportato la tavola 13.1 che fornisce l'errore relativo associato a determinati valori della stima puntuale nei vari domini di studio.

Tavola 13.1 - Modelli sintetici degli errori

DOMINIO					10000	20000	50000	70000	100000	200000	500000	1000000	2000000
	ITALIA	11,99332	-1,36748	0,929	74,021	46,082	24,629	24,629	15,333	9,545	5,102	3,176	2,407
						REGIONI							
1	Piemonte	11,23392	-1,40470	0,899	42,659	26,217	13,775	13,775	8,466	5,203	2,734	1,680	1,264
2	Valle d'Aosta	6,56678	-1,43177	0,886	3,651	2,223	1,154	1,154	0,702	0,428	0,222	0,135	0,101
3	Lombardia	11,07151	-1,37186	0,915	45,755	28,441	15,170	15,170	9,430	5,862	3,127	1,943	1,472
5	Veneto	10,92177	-1,40657	0,893	36,182	22,222	11,666	11,666	7,165	4,401	2,310	1,419	1,067
6	Friuli-Venezia Giulia	9,23261	-1,39454	0,867	16,435	10,136	5,350	5,350	3,300	2,035	1,074	0,663	0,499
7	Liguria	9,37704	-1,36231	0,905	20,492	12,780	6,847	6,847	4,270	2,663	1,427	0,890	0,675
8	Emilia-Romagna	10,98014	-1,38352	0,904	41,426	25,647	13,607	13,607	8,424	5,215	2,767	1,713	1,294
9	Toscana	10,70237	-1,35980	0,886	40,214	25,102	13,463	13,463	8,404	5,246	2,814	1,756	1,333
10	Umbria	8,07687	-1,30061	0,914	14,212	9,055	4,990	4,990	3,179	2,026	1,116	0,711	0,546
11	Marche	9,53435	-1,40351	0,876	18,338	11,274	5,927	5,927	3,644	2,240	1,178	0,724	0,545
12	Lazio	11,08717	-1,37309	0,907	45,852	28,490	15,187	15,187	9,437	5,863	3,126	1,942	1,470
13	Abruzzo	8,86452	-1,34149	0,833	17,455	10,965	5,931	5,931	3,725	2,340	1,266	0,795	0,606
14	Molise	6,26845	-1,20906	0,917	8,771	5,769	3,315	3,315	2,180	1,434	0,824	0,542	0,424
15	Campania	9,62466	-1,22958	0,905	42,737	27,909	15,889	15,889	10,376	6,776	3,857	2,519	1,963
16	Puglia	8,96387	-1,19634	0,920	35,794	23,645	13,668	13,668	9,029	5,964	3,448	2,278	1,787
17	Basilicata	6,72494	-1,24944	0,840	9,150	5,934	3,348	3,348	2,171	1,408	0,794	0,515	0,400
18	Calabria	8,39077	-1,20626	0,838	25,675	16,903	9,726	9,726	6,403	4,215	2,426	1,597	1,250
19	Sicilia	9,70389	-1,23824	0,857	42,726	27,818	15,774	15,774	10,270	6,687	3,792	2,469	1,921
20	Sardegna	7,84526	-1,18745	0,915	21,314	14,123	8,197	8,197	5,432	3,599	2,089	1,384	1,088
41	Bolzano	9,19680	-1,38728	0,850	16,691	10,320	5,466	5,466	3,379	2,090	1,107	0,684	0,516
42	Trento	8,88656	-1,39713	0,887	13,659	8,417	4,438	4,438	2,734	1,685	0,888	0,547	0,412
						AREA *COVID*							
1	Rossa	11,47080	-1,38517	0,918	52,544	32,511	17,236	17,236	10,664	6,599	3,498	2,164	1,635
2	Resto del Nord + Centro	11,51270	-1,39316	0,911	51,716	31,911	16,855	16,855	10,400	6,417	3,390	2,092	1,577
3	Mezzogiorno	9,97216	-1,24815	0,909	46,681	30,288	17,097	17,097	11,093	7,198	4,063	2,636	2,047
						RIPARTIZIONE a 3							
1	Nord	11,71879	-1,39914	0,918	55,772	34,342	18,090	18,090	11,139	6,859	3,613	2,225	1,675
2	Centro	11,18462	-1,37278	0,911	48,211	29,959	15,973	15,973	9,926	6,168	3,288	2,043	1,547
3	Mezzogiorno	9,97216	-1,24815	0,909	46,681	30,288	17,097	17,097	11,093	7,198	4,063	2,636	2,047

Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

13.2 Modello per la presentazione sintetica degli errori

Ad ogni stima $\hat{t}_{Y_d}$ corrisponde un errore campionario relativo $\hat{\varepsilon}(\hat{t}_{Y_d})$; ciò significa che per consentire un uso corretto delle stime sarebbe necessario pubblicare per ogni stima il corrispondente errore di campionamento relativo. Questo, tuttavia, non è possibile sia per limiti di tempo e di costi di elaborazione, sia perché le tavole di pubblicazione risulterebbero appesantite e di non facile consultazione per l'utente finale. Inoltre, non sarebbero comunque disponibili gli errori delle stime non pubblicate, che l'utente può ricavare in modo autonomo.

Per tali motivi si ricorre, in genere, a una presentazione sintetica degli errori relativi basata sul metodo dei modelli regressivi (Wolter, 2007) fondata sulla determinazione di una funzione matematica che mette in relazione ciascuna stima con il proprio errore di campionamento. L'approccio utilizzato per la costruzione di questi modelli è diverso a seconda che si tratti di variabili qualitative o quantitative. Infatti, per quanto riguarda le stime di frequenze assolute (o relative) riferite alle modalità di variabili qualitative, è possibile utilizzare modelli che hanno un fondamento teorico, secondo cui gli errori relativi delle stime di frequenze assolute sono funzione decrescente dei valori delle stime stesse.

Il modello utilizzato per le stime di frequenze assolute, con riferimento al generico dominio d, è del tipo seguente:

$$\log\left(\hat{\varepsilon}^2\left(\hat{t}_{Y_d}\right)\right) = a + b\log\left(\hat{t}_{Y_d}\right)$$

dove i parametri a e b vengono stimati, utilizzando il metodo dei minimi quadrati, su un insieme di stime ottenute dall'indagine (con i rispettivi errori relativi) che coprono approssimativamente l'intervallo di variazione delle stime di frequenze che vengono pubblicate.

Per quanto riguarda la stima della varianza campionaria delle stime di frequenze assolute e relative, al fine di permettere il calcolo degli errori campionari delle stime pubblicate, mediante il metodo sopra descritto, nella tavola 13.1 vengono riportati i valori di a e b e l'indice r^2 che fornisce una misura del grado di rappresentatività degli errori campionari stimati in base al modello.

Inoltre, allo scopo di facilitare il calcolo degli errori campionari, sempre nella tavola 13.1 sono riportati, per i diversi domini territoriali di riferimento delle stime, i valori interpolati degli errori campionari relativi percentuali di alcuni valori tipici assunti dalle stime di frequenze assolute e di totali.

Nella tavola 13.2, infine, sono illustrate le modalità di calcolo per la costruzione dell'intervallo di confidenza al 95% delle stime puntuali riferite al numero di positivi in Italia, Toscana, Area *COVID* Rossa e ripartizione del Nord.

Tavola 13.2 - Calcolo esemplificativo dell'intervallo di confidenza

	NUMERO DI POSITIVI IN ITALIA	NUMERO DI POSITIVI IN TOSCANA
Stima puntuale:	1.482.146	38.041
Errore relativo percentuale (CV%)	2,427	16,213
Errore relativo (CV)	2,427/100 =**0,02427**	16,213/100 =**0,16213**
Stima intervallare:		
Semi ampiezza dell'intervallo	1,96*0,02427*1.482.146 = **70.499**	1.96*0,16213*38.041 =**12.089**
Limite inferiore dell'intervallo di confidenza	1.482.146- 70.499 = **1.411.647**	38.041-12.089 =**25.953**
Limite superiore dell'intervallo di confidenza	1.482.146+ 70.499 **= 1.552.644**	38.041+12.089 =**50.130**

	NUMERO DI POSITIVI IN AREA *COVID* ROSSA	NUMERO DI POSITIVI NEL NORD
Stima puntuale:	1.150.819	1.218.153
Errore relativo percentuale (CV%)	1,963	1,938
Errore relativo (CV)	1,963/100 =**0,01963**	1,938/100 =**0,01938**
Stima intervallare:		
Semi ampiezza dell'intervallo	1,96*0,01963*1.150.819 = **44.296**	1,96*0,01938*1.218.153 = **46.267**
Limite inferiore dell'intervallo di confidenza	1.150.819-44.296 =**1.106.524**	1.218.153-46.267 =**1.171.886**
Limite superiore dell'intervallo di confidenza	1.150.819+44.296 =**1.195.115**	1.218.153+46.267 =**1.264.419**

Fonte: Istat, Rilevazione di sieroprevalenza sul SARS-CoV-2

RIFERIMENTI BIBLIOGRAFICI

Alleva, G., G. Arbia, P.D. Falorsi, V. Nardelli, and A. Zuliani. 2020. "A sample approach to the estimation of the critical parameters of the *SARS-CoV-2* epidemics: an operational design". *Journal of Official Statistics - JOS* (*in corso di pubblicazione*).

Biemer, P.P., R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman (*Eds.*). 2013. *Measurement errors in surveys*. Hoboken, NJ, U.S.: John Wiley & Sons - *Wiley Series in Probability and Statistics*.

Bethel, J. 1989. "Sample allocation in multivariate surveys". *Survey Methodology*, Volume 15, N. 1: 47-57.

Bethlehem, J., F. Cobben, and B. Schouten. 2011. *Handbook of Nonresponse in Household Surveys*. Hoboken, NJ, U.S.: John Wiley & Sons - *Wiley Handbooks in Survey Methodology*.

Blom, A.G., and J.M. Korbmacher. 2011. "Measuring Interviewer Effects in SHARE Germany". *SHARE Working Paper Series*, 03-2011. Munich, Germany: SHARE-ERIC.

Cochran, W.G. 1977. *Sampling techniques, 3rd Edition*. Hoboken, NJ, U.S.: John Wiley & Sons.

Devaud, D., and Y. Tillé. 2019. "Deville and Särndal's calibration: revisiting a 25-years-old successful optimization problem". *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, Volume 28, Issue 4, N. 1: 1033-1065.

Deville, J.-C., and C.-E. Särndal. 1992. "Calibration Estimators in Survey Sampling". *Journal of the American statistical Association*, Volume 87, Issue 418: 376-382.

Durrant, G.B., and J. D'Arrigo. 2014. "Doorstep Interactions and Interviewer Effects on the Process Leading to Cooperation or Refusal". *Sociological Methods & Research*, Volume 43, Issue 3: 490-518.

Grassi, D., e M.C. Romano (*a cura di*). 2020. "L'approccio trasversale alla formazione delle reti di rilevazione". *Letture statistiche - Metodi*. Roma: Istat. https://www.istat.it/it/archivio/243198.

Groves, R.M. 2005. "The Interviewer as a Source of Survey Measurement Error". In Groves, R.M. *Survey Errors and Survey Costs*. Hoboken, NJ, U.S.: John Wiley & Sons - *Wiley Series in Probability and Statistics*.

Hansen, M.H., W.N. Hurwitz, and W.G. Madow. 1953. *Sample survey methods and theory*. Hoboken, NJ, U.S.: John Wiley & Sons.

Istituto Nazionale di Statistica - Istat. 2007*a*. "Come si progetta il monitoraggio del lavoro sul campo di un'indagine sulle famiglie". *Metodi e Norme*, n. 34. Roma: Istat. https://ebiblio.istat.it/digibib/Metodi%20e%20norme/IST0054521Ed2007N34.pdf.

Istituto Nazionale di Statistica - Istat. 2007*b*. "Seminario sulla qualità: l'esperienza dei referenti del sistema informativo SIDI - 2a giornata". *Contributi Istat*, n. 10. Roma: Istat. https://www.istat.it/it/files//2018/07/2007_10.pdf.

Istituto Nazionale di Statistica - Istat. 2006*a*. "Il sistema di indagini sociali multiscopo. Contenuti e metodologia delle indagini". *Metodi e Norme*, N. 31. Roma: Istat. https://www.istat.it/it/files//2014/06/met_norme_06_31_il_sistema_di_indagini_multiscopo.pdf.

Istituto Nazionale di Statistica - Istat. 2006*b*. "La rilevazione sulle forze di lavoro: contenuti, metodologie, organizzazione". *Metodi e Norme*, N. 32. Roma: Istat. https://www.istat.it/it/files/2014/06/met_norme_06_32_rilevazione_forze_lavoro.pdf.

Istituto Nazionale di Statistica - Istat. 2001. "Indagini sociali telefoniche: metodologia ed esperienze della statistica ufficiale". *Metodi e Norme*, N. 10. Roma: Istat.

Istituto Nazionale di Statistica - Istat. 1989. "Manuale di tecniche di indagine. 6 - il sistema di controllo della qualità dei dati". *Note e Relazioni*, N. 1. Roma: Istat.

Kish, L. 1992. "Weighting for Unequal Pi". *Journal of Official Statistics - JOS*, Volume 8, Issue 2: 183-200.

Kreuter, F. 2008. "Interviewer Effects". In Lavrakas, P.J. (*Ed.*). *Encyclopedia of Survey Research Methods*. Thousand Oaks, CA, U.S.: SAGE Publications, Inc.

Mahalanobis, P.C. 1952. "Some aspects of the design of sample surveys". *Sankhyā: The Indian Journal of Statistics*, Volume 12, N. 1/2: 1-7.

Mohorko, A., and V. Hlebec. 2015. "Effect of a first-time interviewer on cognitive interview quality". *Quality and Quantity: International Journal of Methodology*, Volume 49, Issue 5: 1897–1918.

Olson, K., and A. Peytchev. 2007. "Effect of Interviewer Experience on Interview Pace and Interviewer Attitudes". *Public Opinion Quarterly*, Volume 71, Issue 2: 273-286.

Rosenbaum, P.R., and D.B. Rubin. 1984. "Reducing the Bias in Observational Studies Using Subclassification on the Propensity Score". *Journal of the American Statistical Association*, Volume 79, N. 387: 516-524.

Särndal, C.-E. 2007. "The calibration approach in survey theory and practice". *Survey Methodology*, Volume 33, N. 2: 99-119.

Stiegler, A., and N. Biedinger. 2016. "Interviewer Skills and Training (Version 2.0)". *Social Science Open Access Repository - SSOAR*. Mannheim, Germany: GESIS - Leibniz Institute for the Social Sciences.

West, B.T., and K. Olson. 2010. "How Much of Interviewer Variance is Really Nonresponse Error Variance?" *Public Opinion Quarterly*, Volume 74, Issue 5: 1004-1026.

West, B.T., F. Kreuter, and U. Jaenichen. 2013. ""Interviewer" Effects in Face-to-Face Surveys: A Function of Sampling, Measurement Error, or Nonresponse?" *Journal of Official Statistics - JOS*, Volume 29, Issue 2: 277–297.

Wolter, K.M. 2007. *Introduction to Variance Estimation*. Cham, Switzerland: Springer Nature - *Statistics for Social and Behavioral Sciences*.

Zardetto, D. 2015. "ReGenesees: an Advanced R System for Calibration, Estimation and Sampling Error Assessment in Complex Sample Surveys". *Journal of Official Statistics - JOS*, Volume 31, Issue 2: 177-203.