UNIVERSITÀ DEGLI STUDI DI PADOVA

Facoltà di Scienze Statistiche

Corso di laurea in
Statistica, Popolazione e Società

Tesi di laurea

Il sistema di controllo della qualità nelle indagini campionarie:
un approfondimento sugli errori non campionari

Relatore Prof. Giovanna Boccuzzo

Laureanda Adriana Bortolotti
Matricola N. 543632 - SPT

Anno accademico 2007-08

INDICE

Parte prima

ERRORI NON CAMPIONARI E SISTEMA DI CONTROLLO DELLA QUALITA' NELLE FASI DEL PROCESSO DI INDAGINE

Parte seconda

L'ESPERIENZA DI STAGE PRESSO IL SERVIZIO STATISTICA DELLA PROVINCIA DI TRENTO

Parte prima

ERRORI NON CAMPIONARI E SISTEMA DI CONTROLLO DELLA QUALITA' NELLE FASI DEL PROCESSO DI INDAGINE

1. IL SISTEMA DI CONTROLLO DELLA QUALITA' DELL'INFORMAZIONE STATISTICA

1.1. Il concetto di qualità

E' bene anzitutto definire in modo inequivocabile che cosa s'intende per informazione statistica; essa è il prodotto di un'indagine statistica e può essere scomposta in (Istat, 1989):

a) Microdati, ovvero i singoli record relativi alle unità statistiche;
b) Macrodati, ovvero l'esito di una qualunque operazione effettuata sui microdati;
c) Metadati, ovvero valutazioni circa le diverse operazioni svolte (comprendono le definizioni e classificazioni utilizzate, la stima della precisione dei risultati ecc).

Per poter chiarire il concetto di qualità dell'informazione statistica, è verosimile, nonché semplificante, assimilare quest'ultima ad un qualsiasi altro bene o servizio. Avvalendoci di tale raffronto, la qualità dell'informazione statistica è quindi definibile come il possesso da parte della stessa di quelle proprietà grazie alle quali vengono soddisfatti i bisogni dell'utente. (Istat, 2000). Nel caso specifico si tratta evidentemente di esigenze di natura informativo-conoscitiva, dissimili tra loro in funzione della molteplicità di utenti (committenti, ricercatori, utenti generici...). Va da se che l'identificazione dei possibili fruitori dell'informazione generata con una certa indagine, riveste un'importanza fondamentale ed è requisito primo per la qualità del prodotto finale.
È utile inoltre aver presente che anche l'informazione statistica, come gli altri beni o servizi, è frutto di un processo produttivo (nella fattispecie l'indagine in tutte le sue fasi); ne deriva che si può parlare di qualità sia in relazione al prodotto, sia al processo che l'ha creato, sebbene sia lampante come le due entità siano in stretto collegamento tra loro.

1.2. Le dimensioni della qualità

Alla luce delle considerazioni illustrate, è possibile individuare due dimensioni principali della qualità, le quali sono frazionabili a loro volta:

a) Proprietà complessive dell'indagine
    a) Rilevanza teorica
    b) Rilevanza effettiva
    c) Tempestività
    d) Trasparenza

b) Precisione dei risultati
    e) Precisione campionaria
    f) Precisione non campionaria

E' inoltre opportuno aggiungerne di ulteriori, trasversali alle singole indagini:

c) Altre proprietà
    g) Confrontabilità
    h) Completezza

Così definite (Istat, 1989):

- *Rilevanza teorica*: adeguatezza dell'informazione prodotta in relazione alle esigenze manifestate dall'utenza.
- *Rilevanza effettiva*: grado di utilizzazione concreta dell'informazione prodotta, che dipende dalle scelte in materia di elaborazione e diffusione dei risultati.
- *Tempestività*: lasso di tempo che intercorre tra l'inizio dell'indagine e la disponibilità dei risultati; questa proprietà incide sull'utilità degli stessi, con intensità diversa a seconda dell'uso che ne viene fatto e delle specificità del fenomeno indagato. Tempi e costi di un'indagine si influenzano reciprocamente.
- *Trasparenza*: ha una duplice accezione di accessibilità, cioè di semplicità nel reperimento dell'informazione da parte dell'utente e di chiarezza, ovvero di disponibilità di meta-informazioni, indispensabili per un corretto uso dei risultati.

- *Precisione*: funzione inversa dell'errore, cioè tanto più piccolo è l'errore, tanto più precisi sono i risultati. L'errore è dato dalla differenza tra la stima ed il valore vero (cfr § 1.3); tale differenza può essere dovuta (Istat, 1999):

  i. al fatto che l'indagine è di tipo campionario, ovvero che si dispone di dati relativi ad una sola parte della popolazione di riferimento;
  ii. al fatto che l'indagine reale non corrisponde a quella ideale programmata inizialmente. E' infatti naturale che nelle varie fasi vengano commesse sviste o inesattezze, in modo pressoché proporzionale alla dimensione dell'indagine ed al numero di persone coinvolte. Ne consegue che le rilevazioni censuarie sono immuni da errore campionario, ma sono maggiormente soggette a quello non campionario.

- *Confrontabilità*: possibilità di comparare nel tempo e nello spazio indagini relative ad uno stesso fenomeno; ciò è attuabile adottando quanto più possibile definizioni, classificazioni e standard metodologici comuni.
- *Completezza*: misura di quanto risulta essere esauriente il quadro informativo che si ottiene integrando i risultati delle diverse indagini svolte su uno stesso argomento e di quanto quest'argomento è effettivamente d'interesse per l'utenza, considerate le esigenze conoscitive manifestate dalla stessa.

1.3. L'errore totale

Poniamo la nostra attenzione sulla precisione dei risultati: è stato enunciato che essa è funzione inversa dell'errore totale, il quale esprime la distanza tra il valore osservato nell'indagine ed il valore vero, distanza dovuta ad un complesso di errori di diversa natura. L'idea di "valore vero" è meramente astratta: è immediato comprendere che non si dispone di tale valore, altrimenti l'indagine risulterebbe superflua, dal momento che si propone di stimarlo. Benché possa apparire paradossale, tuttavia, quest'idea ci permette di avere a disposizione un quadro concettuale di riferimento.

Ipotizziamo di effettuare una qualunque operazione di sintesi f (y_1, y_2, …. y_i, …. y_n) sui microdati campionari raccolti in un'indagine (ad esempio una media). La stima risulta essere differente dal valore che avremmo ottenuto utilizzando I) i valori veri

II) di tutte le unità appartenenti all'universo di riferimento.
Formalizzando: f (y_1, y_2, …. y_i, …. y_n) ≠ f (Y_1, Y_2, …. Y_i, …. Y_N).
Tale discrepanza è scomponibile nella somma di più errori: l'errore campionario e l'errore non campionario, i quali a loro volta possono provocare sia distorsioni che errori variabili. È opportuno chiarire tali termini. A tale scopo immaginiamo di ripetere l'indagine più volte sotto le stesse condizioni generali (Istat, 1999):

- *Gli errori variabili* si distribuirebbero casualmente: nelle diverse ripetizioni avremmo cioè valori diversi, ognuno determinazione di una variabile aleatoria di media zero e con una certa varianza.
- *Le distorsioni* rimarrebbero invece costanti e sistematiche: dipendono proprio dalle condizioni generali quindi non cambierebbero nelle diverse ripetizioni e porterebbero o sempre a sottostimare o sempre a sovrastimare il valore vero.

Avvalendoci della doppia classificazione scomponiamo l'errore totale in quattro tipi:

i. Errore variabile campionario: è determinato evidentemente dall'uso della tecnica campionaria ed è esprimibile dalla differenza tra la stima ottenibile con i valori veri delle unità campionarie in una specifica indagine e la media delle stime per tutti i possibili campioni, ossia per tutti i gruppi di identica numerosità ottenibili a partire dalla popolazione di riferimento. Inutile precisare che si tratta di un'operazione puramente virtuale e impossibile da eseguire nel concreto.
   Formalizzando, l'errore variabile campionario è esprimibile come:
   f (Y_1, Y_2, …. Y_i, …. Y_n) - E [f (Y_1, Y_2, …. Y_i, …. Y_n)]
   ponendo y*= f (Y_1, Y_2, …. Y_i, …. Y_n)
   semplifichiamo la scrittura: y* - E [y*]

ii. Distorsione campionaria: è dovuta all'adozione di uno stimatore distorto, cioè la cui media per tutti i possibili campioni non coincide con il valore calcolato sull'intera popolazione. La distorsione è dunque sintetizzabile come:
   E [f (Y_1, Y_2, …. Y_i, …. Y_n)] - f (Y_1, Y_2, …. Y_i, …. Y_N)
   ponendo y*= f (Y_1, Y_2, …. Y_i, …. Y_n) e Y = f (Y_1, Y_2, …. Y_i, …. Y_N)
   riformuliamo in questo modo: E (y*) - Y

iii. Errore variabile non campionario: differenza tra la stima ottenuta con dati relativi ad un campione in una specifica indagine e la media della stima calcolata ipotizzando di intervistare lo stesso campione più volte; è infatti verosimile pensare di non ottenere sempre le stesse identiche risposte (per motivi di svariato tipo). Le variazioni si possono pensare come determinazioni di una variabile aleatoria con media nulla e con una certa varianza.

iv. Distorsione non campionaria: dovuta alla differenza sistematica tra il valore osservato ed il valore vero. Anche immaginando di reintervistare lo stesso campione, tale differenza rimarrebbe costante e con uno specifico segno rispetto al valore vero (o sempre minore o sempre maggiore). Tipico esempio di variabile soggetta a distorsione non campionaria è il reddito, dichiarato sempre inferiore a quello effettivo.

L'errore non campionario è dunque la somma delle due componenti suddette ed è formulabile in termini generali come la differenza tra la stima ottenuta con i dati di un campione ed il valore ottenibile con i valori veri delle medesime unità.
Ovvero: $f(y_1, y_2, \ldots. y_i, \ldots. y_n) - f(Y_1, Y_2, \ldots. Y_i, \ldots. Y_n)$
ponendo $y = f(y_1, y_2, \ldots. y_i, \ldots. y_n)$ e $y^* = f(Y_1, Y_2, \ldots. Y_i, \ldots. Y_n)$
si semplifica così: $y - y^*$
differenza scomponibile in errore variabile e distorsione.
Formalizzando: $y - y^* = [\, y - E(y/c) \,] + [\, E(y/c) - y^* \,]$
dove $E(y/c)$ è la media della stima ottenibile reintervistando lo stesse unità campionarie.

Sintetizzando quanto sopra descritto:

Errore totale [y - Y] = Errore variabile campionario [y* - E (y*)]
+
Distorsione campionaria [E (y*) - Y]
+
Errore variabile non campionario [y - E (y/c)]
+
Distorsione non campionaria [E (y/c) - y*]

1.4. Il sistema di controllo

Il sistema di controllo è costituito da quel complesso di azioni mirate a monitorare e a ridurre l'errore non campionario. Il termine "sistema" non è casuale, ma sta ad indicare un insieme compatto e coerente di controlli. Tali caratteristiche si rivelano indispensabili, dal momento che il sistema si deve applicare ad una realtà, l'indagine per l'appunto, che non è possibile nel concreto scomporre in fasi nitidamente distinte: il sistema di controllo intende quindi affiancare il complesso dell'indagine, supervisionando sia le fasi stesse, che le interrelazioni sussistenti tra esse.
Il fatto che l'errore campionario non sia oggetto di detto monitoraggio può forse apparire strano, eppure, se si riflette sulle strategie utilizzate per il suo controllo, diventa chiara la motivazione dell'esclusione. Basti pensare, ad esempio, alla scelta della numerosità campionaria: nel determinarla, si rende nota la precisione di cui godranno le stime e quindi, specularmente, l'errore campionario da cui saranno affette; anzi, spesso si calcola la numerosità proprio a partire dal livello di fiducia che si desidera ottenere. Un discorso analogo si può fare per quanto riguarda tecniche quali la stratificazione o la stadificazione: si conosce fin dalla loro pianificazione l'impatto che avranno sulla precisione dei risultati. Ulteriori ispezioni durante lo svolgimento dell'indagine risulterebbero dunque assolutamente superflue.
Il sistema di controllo è concepibile come risultato dell'apporto di tre tipi di azioni, differenziate a seconda del momento in cui avvengono e per il fine che si propongono (Istat, 1989).

1.4.1 Le azioni preventive

Il loro intento è di migliorare la progettazione dell'indagine; si collocano dunque in linea di massima temporalmente prima della rilevazione sul campo.
La prevenzione costituisce un obiettivo tanto più rilevante se si considera che la possibilità di modificare le norme in corso d'opera è scarsa ed estremamente costosa in termini organizzativi ed economici. (Istat, 1989)
Fanno parte di questo sottoinsieme tutte quelle accortezze mirate a prevenire l'insorgenza dell'errore o, quantomeno, a diminuirne la probabilità. Si concretizzano nelle azioni più svariate, come ad esempio nell'uso di vocaboli semplici all'interno del

questionario, per evitare dubbi di comprensione dei quesiti oppure in un'attenta selezione dei rilevatori, al fine di ridurre la frazione d'errore a loro imputabile.

1.4.2 Le azioni di monitoraggio e valutazione

Hanno un duplice obiettivo:

a) di monitoraggio delle varie fasi e di costruzione dei relativi indicatori di qualità (es. il tasso di non risposta parziale per le variabili più importanti oppure la durata media dell'intervista)
b) di stima delle componenti dell'errore totale, il cui calcolo, però, implica un ritorno sul campo o comunque operazioni supplementari a quelle ordinarie; in ogni caso sottintende un impegno più oneroso in termini sia economici che organizzativi (es. stima della varianza mediante reintervista delle unità)

Le informazioni che vengono prodotte con queste azioni permettono di quantificare l'errore e forniscono informazioni preziose per perfezionare l'indagine nel caso venisse ripetuta, o per indagini simili. Non è infatti raro che si riesca ad identificare la fonte (o le fonti) che hanno contribuito a determinare l'errore: ad esempio, elaborazioni distinte per rilevatore consentono di rivelare eventuali associazioni tra uno o più rilevatori ed errori riscontrati, come le non risposte.

1.4.3 Le azioni di controllo e correzione

Hanno l'obiettivo di individuare e correggere gli errori e si realizzano contemporaneamente all'indagine; più propriamente potremmo dire che esiste una specifica fase a cui è demandato quasi interamente questo compito: si tratta della fase di revisione.
E' possibile correggere solamente una parte degli errori, ossia quelli derivanti da non risposte, sia totali che parziali, e da incongruenze logiche interne alla variabile o intervariabile. In realtà sarebbe poco appropriato parlare di "correzione", dato che tali azioni non ripristinano il valore esatto, ma si "accontentano" di attribuirne uno il

più possibile simile (cfr § 3.3.2). Ad ogni modo ciò fa sì che l'errore sia minore che se non si fosse operato nessun accorgimento.

2. LA PROGETTAZIONE DELL'INDAGINE

In primo luogo è bene precisare che la progettazione dell'indagine non è compito dei soli statistici, ma deve necessariamente giovarsi dell'apporto di numerose altre professionalità (Istat, 2000). L'indagine non è infatti un processo avulso dalla realtà, al contrario vi è immersa, nel momento in cui la indaga e la esamina. E' indispensabile quindi che nel gruppo di progettazione ad affiancare i tecnico-statistici, vi siano gli esperti del fenomeno oggetto d'indagine; tale gruppo deve ispirarsi irrinunciabilmente al principio dell'interdisciplinarietà, in modo che scelte e soluzioni siano frutto dell'integrazione e della combinazione dei diversi saperi.
Chiarito ciò, si definisce come fase di progettazione la pianificazione del indagine, che nello specifico comprende: la formulazione del modello concettuale, la redazione del questionario e la messa in atto di procedure per il suo controllo, la predisposizione dei piani di lavoro, la scelta del tipo e della tecnica di indagine e dell'eventuale disegno di campionamento. Dal punto di vista della qualità non campionaria, la progettazione costituisce un momento cruciale poiché in essa di fatto si prevengono gli errori che la possono inficiare e viene programmato il loro controllo (Istat, 1989).

2.1. La progettazione concettuale

Questa fase ha come non facile obiettivo quello di costruire un modello concettuale di riferimento, che possiamo a tutti gli effetti considerare le fondamenta dell'indagine, sul quale si compongono tutte le azioni successive. E imprescindibile il rispetto e la conformità a tale modello nel corso di tutte le fasi che susseguono.
A partire da idee approssimative e intenti generici, la progettazione concettuale si propone di esplicitare chiaramente obiettivi, definizioni e classificazioni; si tratta di un'operazione alquanto delicata, data l'inevitabile e profonda connessione con il fenomeno di interesse. Proprio per questo motivo costituisce la fase in cui si commettono la maggior parte degli errori di rilevanza teorica e quindi quella che maggiormente necessita del contributo degli esperti, con il fine di attutire il più possibile il rischio d'errore (Istat, 2000).

Si definiscono:

- *Fenomeno da indagare*: va circoscritto con precisione il dominio di interesse, compatibilmente con le risorse di cui si dispone ed i vincoli a cui si deve sottostare. È importante altresì precisare se si è interessati alla componente statica o dinamica del fenomeno e le ipotesi che si intende sottoporre a verifica.
- *Universo di riferimento*: tutta e sola la popolazione oggetto d'indagine. Va circoscritta univocamente nel tempo e nello spazio. Ad essa sono riferiti i risultati delle analisi.
- *Variabili di studio:* rappresentano lo strumento mediante il quale il fenomeno di interesse può essere investigato. Va prestata attenzione al tipo di variabili che si utilizzano a seconda dell'uso che si prevede di farne (Istat, 2000); ad esempio, se si intende calcolare degli indici statistici che richiedono valori numerici, si deve fare in modo di rilevare le caratteristiche in questione con variabili quantitative.
- *Classificazioni:* a meno che non si decida di optare per risposte aperte, si rende necessario precodificare le modalità che le variabili nominali (sia sconnesse che ordinali) possono assumere, definendo l'insieme delle categorie che ne descrivono i possibili esiti. Naturalmente costringere le risposte all'interno di un elenco predefinito comporta una riduzione d'informazione rispetto alla scelta dei quesiti aperti (tanto più se non si dispongono di informazioni precise sul fenomeno in questione); per contro, questi ultimi sono fonte di complicanze aggiuntive e conseguentemente di errori ulteriori, principalmente per quanto riguarda la loro interpretazione e traduzione in codici elaborabili.

  Nell'articolare le classificazioni occorre ricordare che le modalità devono essere esaustive e mutuamente esclusive, in modo che tutti possano trovare l'esatta opzione che fa al caso loro; inoltre è bene tener presente che modalità numerose e analitiche determinano minor precisione nelle risposte (Istat, 1989).

  Per soddisfare il requisito di confrontabilità tra indagini sarebbe auspicabile rifarsi a classificazioni comunemente utilizzate, ovviamente se ben si adattano agli obiettivi stabiliti; in particolare per le variabili più complesse da definire, esistono delle apposite classificazioni standard cui riferirsi (Ateco, Icd, Isced...)
- *Unità di analisi:* unità a cui si riferiscono le notizie raccolte e, di conseguenza, i risultati prodotti con l'indagine. Possono essere molteplici.
- *Unità di rilevazione:* soggetto a cui viene richiesto di fornire le notizie riferite alla (o alle) unità di analisi. Unità di rilevazione ed unità di analisi possono coincidere (ad

esempio nel censimento il rispondente assume entrambi i ruoli, ma non è l'unica unità di analisi, dal momento che vengono richieste informazioni su altre unità, come la famiglia o le abitazioni).

Nel modello concettuale sono presenti tutte le sopraindicate entità; è chiaro che affinché sia uno strumento prezioso, così come è stato presentato, è fondamentale siano aggiunti anche i legami intercorrenti tra tali entità, in modo da ottenere uno schema logico e coeso, che costituisca di fatto un appoggio imprescindibile durante le fasi dell'indagine.

2.2. Il questionario

Il questionario è il mezzo che permette di rilevare informazioni sulle variabili di interesse presso l'unità di rilevazione. È uno strumento di misura e come tale deve essere standardizzato: a tutti devono essere poste le stesse domande, formulate nello stessa maniera e con le medesime alternative di risposta, così che sia possibile confrontare le informazioni raccolte.
I quesiti per la rilevazione delle variabili di studio costituiscono senza dubbio la parte preponderante del questionario; tuttavia non vanno trascurate altre due componenti importanti (Istat, 1989):

a. I codici identificativi, prerequisito per l'individuazione delle unità e delle loro relazioni (con i famigliari piuttosto che con il rilevatore ovvero il comune di appartenenza).
b. Le variabili di controllo dell'intervista: contenute in una sezione apposita a cura del rilevatore, raccolgono informazioni riguardanti le modalità dell'intervista (dalla durata, alla presenza di estranei, al grado di partecipazione manifestato dal rispondente), utili per derivarne indicatori di qualità.

Dopo di che è conveniente una puntualizzazione: nel linguaggio comune quando si parla di "questionario" generalmente si allude ad un modulo cartaceo da autocompilare; chiariamo che ciò è indiscutibilmente vero, ma non esaurisce il suo significato. Contempliamo invero con tale termine tutti i moduli preposti alla raccolta dei dati, indi compresi lo schema utilizzato nelle interviste telefoniche piuttosto che

quello visibile su pc dai rilevatori che effettuano le interviste faccia a faccia oppure i modelli per la raccolta di dati amministrativi. Detto ciò, nel presente paragrafo ci occuperemo del questionario in questa sua accezione generale, senza scendere in particolarità dovute alla tecnica di intervista, né saranno esaminate tali tecniche.

2.2.1 La redazione del questionario

Un'adeguata formulazione del questionario è cruciale, giacché è con esso che si acquisiscono i dati che saranno utilizzati in tutte le fasi successive, fino al termine dell'indagine. Se il modello concettuale è stato realizzato accuratamente ed è internamente coerente, la strutturazione del questionario ne risulterà notevolmente facilitata e darà luogo ad uno strumento quanto più adeguato agli obiettivi per i quali è stato creato. In particolare il modello si rivela d'aiuto (Istat, 2000):

i. Nella definizione delle sezioni: gli argomenti trattati vanno organizzati per sezioni tematiche, la cui successione deve seguire una certa logica. In questo modo il rispondente si troverà ad effettuare un percorso dotato di senso, riuscendo ad inquadrare all'interno dello stesso i contesti specifici: presumibilmente ciò porterà a meno confusione e quindi a risposte più precise. Tenendo sempre ben presente questa regola generale, è doveroso operare ulteriori accortezze nel caso di quesiti delicati, che vanno sondati preferibilmente alla fine del questionario, per sfruttare la maggior confidenza acquisita, i quesiti che implicano uno sforzo di memoria, che andrebbero collocati a metà circa ed i quesiti che richiedono un'opinione spontanea, che è consigliabile collocare all'inizio, per non influenzare la risposta.

ii. Nella predisposizione delle domande filtro: si dicono domande filtro quei quesiti che danno luogo a percorsi diversificati a seconda della condizione che verificano; servono per porre domande specifiche a differenti sottogruppi (es. su studenti e lavoratori si rilevano caratteristiche distinte) oppure per non sottoporre inutilmente domande quando ciò non ha senso (es. chiedere il numero di figli a coloro che hanno dichiarato di non averne). Le scelte riguardanti la predisposizione dei filtri derivano appunto dal modello concettuale, il quale suggerisce dove vanno inseriti e per quale scopo.

Ci sono altri aspetti a cui è necessario prestare attenzione. Lo scopo è, come per i precedenti, di prevenzione degli errori di rilevanza e di precisione determinabili dal questionario; questi ultimi, però, si potrebbero definire degli "accorgimenti strategici", che poco dipendono dal modello concettuale.

- *Quesiti retrospettivi:*sono costituiti dalle domande che richiedono di ricordare eventi accaduti in un determinato intervallo passato, chiamato periodo di riferimento. Tale intervallo può essere fissato, se delimitato da date, oppure variabile, se dipende dal momento in cui viene chiesto di ricordare l'evento (es. settimana precedente l'intervista). Essendo la memoria soggetta a lacune, è intuitivo comprendere che questi quesiti pongono dei problemi in termini di precisione. L'evento può infatti essere omesso, perché scordato, ed in tal caso si avrà una sottostima del fenomeno in questione oppure essere collocato erroneamente, spostato indietro rispetto al momento in cui è realmente accaduto o in avanti (si parla in tal caso di effetto telescopio). Ciò può comportare una "immeritata" esclusione o inclusione nel periodo di riferimento oppure un'errata collocazione all'interno dello stesso. Quest'ultima svista costituisce fonte d'errore solo se si considerano periodi dettagliati (es. il numero degli eventi in un mese è corretto, ma la distribuzione per settimana è distorta). Esclusioni ed inclusioni scorrette provocano invece delle distorsioni, rispettivamente sottostime e sovrastime.
  È importante cercare di rimediare a tali inconvenienti adoperando delle attenzioni che aiutino la memoria del rispondente, senza influenzarne i ricordi. Anzitutto è opportuno scegliere periodi di riferimento il più possibile vicini all'intervista; se si è interessati a periodi più remoti può essere utile stimolare la memoria fornendo una lista di possibili risposte anziché lasciare aperta la domanda (Es. invece che chiedere che genere di film si è visto nell'ultimo mese, è preferibile fornire una lista dei principali generi cinematografici con le modalità si/no per ognuno di questi).

- *Quesiti proxy:* con tale espressione s'intendono le domande che accettano una risposta anche se fornita da una persona diversa dall'unità designata (nella maggior parte dei casi si tratta di un famigliare). Se il valore rilevato non coincide con quello che avrebbe dato il rispondente "ufficiale", si realizza un errore non campionario, chiamato effetto proxy. Per prevenirlo si può decidere di non accettare risposte se non date di persona, oppure di accettarle limitatamente ad alcuni quesiti, o dopo un

certo numero di ritorni da parte del rilevatore. Ovviamente ciò rischia di incrementare le mancate risposte. È opportuno quindi valutare e soppesare pro e contro, tenendo presente che l'effetto proxy è funzione delle variabili e che quindi ne esistono di poco o per nulla soggette, per le quali cioè è ragionevole supporre che persone vicine al rispondente sappiano fornire informazioni corrette a suo riguardo.

- *Formulazione delle domande:* è opportuno prestare particolare attenzione al vocabolario utilizzato. L'uso di un termine piuttosto che un altro non è una questione trascurabile, tutt'altro, può deviare la risposta in maniera considerevole, come dimostrano numerosi esperimenti condotti da esperti nella comunicazione (Istat, 1989). Come regola generale, vanno usati termini semplici e chiari; la frase va formulata in modo lineare, senza giri di parole e deve contenere tutte le informazioni necessarie per rispondere senza dubbi di comprensione. Per i quesiti delicati può essere utile ricorrere a costruzioni indirette, in modo che il rispondente sia libero di esprimersi senza sentirsi chiamato in causa, ad apposite domande introduttive che evitino di porre "senza preavviso" il quesito problematico e a giustificazioni inglobate nella risposta, tali da permettere una risposta negativa, poiché già legittimata (es. "Le è stato possibile recarsi a votare?")

2.2.2 Il controllo del questionario

Come già evidenziato, un'appropriata formulazione del questionario è decisiva per l'intero processo; pertanto, anche se si sono seguiti tutti i suggerimenti riportati, è opportuno mettere in atto una serie di controlli che testino la validità dello strumento; vanno accertati: la conformità con gli obiettivi conoscitivi dell'indagine, la presenza di tutti i quesiti occorrenti, l'adeguatezza dell'ordinamento e della formulazione dei quesiti, la correttezza delle norme di compilazione e la facilità di gestione per il rilevatore o per il rispondente, qualora il rilevatore non sia previsto (Istat, 2000).

A tale scopo si ricorre a differenti tecniche; nulla vieta di metterle tutte in atto, normalmente, però, viene operata una selezione a partire dall'individuazione degli obiettivi prioritari del controllo, da considerazioni di natura economico-organizzativa, dai tempi a disposizione.

- *Progettazione concettuale*: come ribadito più volte, il questionario discende dal modello concettuale costruito in principio; verificare che ciò si sia effettivamente realizzato costituisce un semplice ma efficace controllo di natura sostanziale.
- *Diagrammazione del questionario*: consiste nella schematizzazione della sequenza dei quesiti mediante un diagramma di flusso; tale rappresentazione permette di esaminare la struttura formale del questionario, con l'obiettivo di appurarne la linearità ed individuare carenze o contraddizioni nelle norme di compilazione.
- *Giudizio degli esperti*: con questo metodo si intende porre a verifica sia gli aspetti formali che di contenuto, rivolgendosi agli esperti del fenomeno per il primo scopo e agli esperti della comunicazione per il secondo. Se ciò è possibile, ci si avvale inoltre di tecniche di laboratorio, che consistono in interviste approfondite realizzate da esperti in grado di cogliere ed interpretare le reazioni e le difficoltà dei rispondenti.
- *Pre-test*: consiste nella somministrazione del questionario ad un campione di ridotta numerosità, di solito ragionato (cioè non estratto casualmente ma costruito ad hoc); l'applicazione di tale tecnica richiede intervistatori esperti ed attenti, in grado di cogliere gli elementi che, esaminati, permetteranno di valutare chiarezza, completezza e semplicità di gestione del questionario. Il pre-test è impiegato di frequente anche per convertire i quesiti aperti in chiusi, soprattutto nel caso in cui si dispongano di scarse informazioni sul fenomeno e perciò sia preferibile sondare il terreno prima di effettuare la precodifica. Limite della tecnica è che individua gli errori, ma non fornisce indicazione per risolverli.
- *Test di alternative*: consiste nel somministrare due o più versioni del questionario, discrepanti per un solo aspetto, ad altrettanti campioni indipendenti, simili per numerosità e caratteristiche rilevanti (nella maggior parte dei casi variabili strutturali, come sesso ed età). Il fattore tenuto sotto controllo può essere l'ordinamento delle domande piuttosto che il livello di dettaglio di una classificazione o altro. Si tratta in pratica di un disegno sperimentale pertanto, sul piano scientifico, è "legittimata" la comparazione dei risultati e diventa così possibile stabilire qual'è l'alternativa migliore. Il test precedente e questo si possono considerare complementari, dal momento che il primo rintraccia gli errori, quest'ultimo li risolve.

2.3. I piani di lavoro

In fase di progettazione vengono predisposti (Istat, 1989):

a) il piano di campionamento
b) il piano di rilevazione
c) il piano di registrazione
d) il piano di revisione
e) il piano di elaborazione
f) il piano di diffusione
g) il piano dei controlli

I piani di lavoro ricalcano non a caso le varie fasi dell'indagine: in essi vengono infatti programmati gli aspetti a tali fasi relativi. Anche la stesura dei piani di lavoro poggia in modo più o meno diretto sul modello concettuale; il caso lampante è probabilmente rappresentato dalla diffusione, la quale non può certo prescindere dagli obiettivi enunciati nel modello, dal momento che ne è specificazione concreta. Ma pure il legame con rilevazione ed elaborazione è di intuitiva comprensione, se si tiene sempre ben presente che il modello fornisce tutte le definizioni basilari, tra le altre quelle di unità di rilevazione e di rilevazione.
Da notare la posizione particolare rivestita dal piano dei controlli: in esso viene infatti organizzato il sistema di controllo dell'informazione statistica, il quale, come illustrato nel capitolo precedente, ha per oggetto sia le fasi che i legami tra di esse. Per questa ragione potremmo definirlo trasversale a tutti gli altri, giacché in esso si organizzano gli strumenti per saggiare la validità di questi e nel contempo anche della coerenza del quadro d'insieme. La coerenza è prerequisito fondamentale, visto che i piani non sono progetti a sé stanti, ma si concatenano e influenzano reciprocamente: il piano di elaborazione è evidentemente vincolato a quello di diffusione, il piano di rilevazione deve servirsi delle informazioni contenute nel piano di campionamento e così via.

2.4. Il disegno d'indagine

Con il termine disegno ci si riferisce alle caratteristiche di un'indagine definite dalla specifica combinazione delle seguenti tre scelte: il tipo di indagine, la strategia di campionamento, la tecnica d'indagine (Istat, 2000).

2.4.1 Il tipo di indagine

Fissati gli obiettivi che si intendono perseguire, occorre stabilire il tipo di indagine più consono a soddisfarli. Una prima sostanziale decisione va presa tra indagine trasversale e indagine longitudinale: la prima mira a produrre statistiche di stato, ovvero riferite ad un dato momento o periodo e quindi si esaurisce in un solo contatto con le unità di interesse; la seconda, invece, essendo orientata alla componente dinamica dei fenomeni, necessita di contatti ripetuti, grazie ai quali raccogliere informazioni su eventuali variazioni ed evoluzioni intervenute. All'interno dei due insiemi la scelta si dettaglia in funzione di altri elementi: le indagini trasversali possono essere occasionali o periodiche, a seconda della frequenza con le quali ci si propone di svolgerle; le longitudinali possono essere semplici o con rotazione del campione, a seconda che, rispettivamente, le unità siano sempre le stesse ad ogni contatto oppure in parte si rinnovino. In questo ultimo caso coesistono la componente statica e quella dinamica ed è quindi possibile compiere entrambe le analisi.

2.4.2 La strategia di campionamento

Prima di tutto va fatta una precisazione, benché di un'ovvietà sorprendente: a rigore si sarebbe dovuto anteporre a questo passaggio la scelta tra indagine totale e non, visto che, palesemente, la strategia di campionamento si mette in atto solo se l'indagine è di tipo campionario. Tuttavia, le difficoltà che comporta il raccogliere informazioni su tutte le unità della popolazione in oggetto (in primis l'insostenibilità economica e il notevole aumento degli errori non campionari) circoscrive di fatto questa modalità ad occasioni di singolare importanza come i censimenti o a casi in cui le informazioni siano già state raccolte per scopi diversi -è il caso delle indagini

amministrative- (Istat, 2000). Non commettiamo dunque una pesante trascuratezza se ci limitiamo alle indagini in cui la strategia di campionamento è presente, poiché rappresentano la stragrande maggioranza dei casi.
È poi opportuno aggiungere un'ulteriore puntualizzazione: il tema del campionamento non sarà trattato in maniera compiuta e diffusa poiché non è argomento di preminente importanza in questa sede; l'attenzione sarà piuttosto rivolta alle ricadute che esso può determinare in termini di errore non campionario.
Detto ciò, si definisce strategia di campionamento quel complesso di operazioni atto a definire un sottoinsieme della popolazione di riferimento, chiamato campione, adatto a rappresentarla (Istat, 1989). Rinviando alla letteratura sull'argomento per quanto attiene alle tecniche impiegate per selezionare le unità costituenti il suddetto campione, indirizziamo piuttosto la nostra attenzione ai supporti fisici sui quali la selezione si basa. Per poter procedere all'estrazione di un sottoinsieme della popolazione, infatti, è indispensabile disporre di un elenco in cui siano registrati tutti i componenti che di essa fanno parte. Questo elenco viene chiamato archivio di base, o lista. Solitamente ci si serve di archivi amministrativi (quali l'anagrafe, il catasto, i registri d'imprese…) o di altra natura, come l'elenco degli abbonati di telefonia fissa; ad ogni modo, nella maggior parte dei casi, esistenti per scopi diversi da quello statistico. Si rende perciò necessario accertare che l'archivio di cui si dispone si presti all'utilizzo come lista di base per il campionamento. A tale scopo devono essere verificate le seguenti condizioni (Istat, 2000):

a) la lista deve contenere per ogni unità le informazioni occorrenti per la localizzazione sul territorio (indirizzo e/o recapito telefono)
b) la lista deve essere completa ed in buono stato di aggiornamento, ovvero deve essere composta da tutte e sole le unità appartenenti alla popolazione di interesse (dette includibili).

Se le caratteristiche della lista non rispondono alle proprietà sopraelencate, si realizzano degli errori non campionari che vanno sotto il nome di errori di copertura, i quali trovano concretizzazione nella successiva fase di rilevazione. È possibile dividerli in due gruppi, a seconda che derivino da (Istat, 2000):

a) sovracopertura della lista, ovvero inclusione di unità non includibili o inclusione ripetuta delle stesse unità. È possibile mettere in atto delle misure per

individuare tali errori durante la rilevazione, inserendo nel questionario appositi quesiti filtro che appurino l'appartenenza alla popolazione o, se il problema è dato dalle duplicazioni, controllando gli elenchi degli intervistati per ciascun rilevatore. Ciò non toglie che i questionari constatati non validi vadano ad incrementare la quota di mancate risposte totali (uno dei principali errori non campionari; verrà illustrato in un paragrafo successivo, dove saranno altresì esaminate le altre fonti da cui può scaturire).

b) sottocopertura della lista, ossia esclusione di unità includibili. Logicamente non è possibile individuare le unità escluse. Questo problema ha risvolti di differente entità a seconda delle caratteristiche che queste possiedono: se non presentano particolari differenze relativamente alle variabili di maggior importanza per l'indagine, allora la somiglianza tra popolazione e lista sopperisce alla mancata coincidenza, garantendo così errori di lieve entità; se, al contrario, l'assenza all'interno della lista non è casuale, gruppi particolari sfuggono alla rilevazione e questa situazione è assimilabile al caso più grave in cui si abbiano mancate riposte totali concentrate su specifiche subpopolazioni. Infine, se le unità sono correttamente incluse, ma in fase di rilevazione risultano irreperibili perché sulla lista non c'è l'indirizzo o questo è erroneo, si determinano delle mancate risposte totali presumibilmente casuali, quindi meno preoccupanti che nella situazione precedente.

Gli errori di lista non hanno ripercussioni sulla sola componente non campionaria: infatti, la riduzione della numerosità del campione dovuta ai questionari non somministrati o non validi impedisce di garantire che l'errore campionario sia quello prestabilito.

È quindi immediato desumere come la strategia di campionamento debba necessariamente delinearsi sulla base delle caratteristiche degli archivi disponibili. Un esempio banale: è noto che il campionamento stratificato sia particolarmente efficiente, cioè permetta, a parità di numerosità campionaria, di ottenere un errore campionario minore; sappiamo che esso prevede una suddivisone del campione in sottogruppi (strati appunto) in funzione di una o più variabili (Istat, 1989). Ora, se la lista utilizzata non contiene per tutte le unità che vi appartengono informazioni su queste variabili, è chiaro che il disegno non è attuabile, quindi la sua efficienza rimane ad un livello meramente potenziale.

In conclusione, è opportuno esaminare la struttura (nonché lo stato di aggiornamento) degli archivi di base disponibili prima di scegliere la tecnica mediante la quale estrarre il campione, perché da ciò può dipendere un aumento dell'errore totale.

2.4.3 La tecnica d'indagine

Con tecnica d'indagine si intende la modalità di contatto e di raccolta delle informazioni presso le unità di rilevazione. Alla base si trova sempre il questionario, solamente è diverso il modo in cui questo viene somministrato: tramite intervista telefonica, a mezzo posta, faccia a faccia, mediante autocompilazione ecc. Non è nostra intenzione dilungarci sulle diverse tecniche, basti aver cognizione del fatto che ognuna di esse ha delle precise caratteristiche, le quali possono conformarsi o meno alla lista in uso, alla strategia di campionamento, come pure ai tempi e ai costi a disposizione: è compito di chi progetta l'indagine scegliere la modalità più opportuna, che limiti quanto più possibile le ricadute in termini di errori non campionari (Istat, 2000).

2.5. Il controllo della fase progettuale: l'indagine pilota

La fase progettuale trova compimento in un ultimo momento di verifica, che permette di saggiare sul campo le soluzioni adottate (Istat, 2000). A tale scopo si ricorre alla cosiddetta "indagine pilota", che potremmo definire come una realizzazione in miniatura dell'indagine, condotta su un campione ridotto, ma in maniera più approfondita. Si colloca solitamente in prossimità dei controlli del questionario (anzi, il pre-test ne fa spesso parte), ma l'obiettivo che persegue è ben più ampio: si propone infatti di ispezionare tutti gli aspetti programmati quindi, oltre appunto al questionario, il grado di efficacia della modalità d'intervista, lo stato effettivo delle liste, i collegamenti tra sede ed organi periferici (rilevatori, comuni ecc), il sistema dei codici identificativi delle unità, l'adeguatezza dei piani di lavoro, eccetera. In sostanza appura l'attuabilità del progetto in una situazione concreta, consentendo l'individuazione di potenziali problemi non accertabili in sede di progettazione astratta (Istat, 1989). Ovvio che in questo caso a seguire la pilota c'è

un secondo passaggio di natura correttiva, così che almeno gli errori rintracciati con essa non siano presenti nell'indagine reale.
Esiste un'altra funzione, benchè secondaria, assolta da questa versione ridotta dell'indagine e cioè quella di stimare la variabilità dei fenomeni di maggior interesse al fine di determinare la numerosità campionaria, qualora non sia possibile far ricorso ad altre fonti.

3. LE FASI OPERATIVE

Predisposta la progettazione dell'indagine e dimostratene la validità mediante l'indagine pilota, si hanno a disposizione tutti gli elementi occorrenti per mettere in pratica ciò che è stato pianificato su carta.
Elenchiamo di seguito le fasi operative attraverso le quali si attua il processo d'indagine. Come già anticipato in diverse occasioni, si tratta di una schematizzazione semplificatrice della realtà: ossia, è vero certamente che è possibile individuare degli insiemi omogenei di operazioni a partire da tutte quelle facenti parte del processo, ma è importante non concepire questa ripartizione come rigida ed indiscutibile poiché, al contrario, ha valenza indicativa ed è suscettibile di modificazioni in funzione delle specificità contestuali. Lo stesso discorso vale per gli errori non campionari che verranno via via affibbiati alle diverse fasi: l'intento è quello di segnalare la fonte principale responsabile di tali errori, che di fatto però si trasmettono di fase in fase o sono imputabili all'interazione tra diverse operazioni e/o persone in esse maggiormente coinvolte.
Detto ciò, distinguiamo (Istat, 1989):

a) la rilevazione
b) la registrazione su supporto informatico
c) la revisione
d) l'elaborazione
e) la validazione
f) la diffusione

3.1. La rilevazione

Nella presente fase si collocano tutte quelle operazioni che vengono svolte dalla rete periferica oppure la riguardano. Sono quindi comprese, oltre alla rilevazione vera e propria (intesa come raccolta dei dati presso le unità), l'estrazione del campione secondo le modalità stabilite nel disegno di campionamento, la pubblicizzazione dell'indagine a livello locale, la selezione e la formazione dei rilevatori, la codifica dei quesiti aperti (Istat, 1989). In ognuna di queste microfasi possono realizzarsi degli

errori non campionari, sia derivanti da carenze o disattenzioni in fase progettuale che propri delle fasi stesse. Ad ogni modo è in questa parte dell'indagine che trovano manifestazione concreta gli errori non campionari più rilevanti, sia per gravità che per numerosità (Istat, 1999):

a) mancate risposte totali, ovvero i questionari non compilati per impossibilità, rifiuto o scarsa motivazione a rispondere da parte delle unità di rilevazione, per irreperibilità delle stesse oppure a causa di errori di lista (di cui si è già ampiamente parlato nel paragrafo dedicato alla strategia di campionamento). Se non sono casuali possono causare distorsioni nei risultati finali.
b) mancate risposte parziali, cioè la non compilazione di una parte dei quesiti contenuti nel questionario a causa di rifiuto, impossibilità o scarsa motivazione a rispondere. Si assimilano ad esse anche i valori fuori campo (ad esempio codice 3 per sesso, che prevede solo i codici 1 e 2) e le incongruenze logiche sia intravariabile (esempio 1800 come data di nascita), che tra variabili (come donna militare di leva). Anche le mancate risposte parziali, se non casuali, possono provocare distorsioni nei risultati, relativamente alle variabili che ne sono interessate. Nel complesso sono ascrivibili in parte al rispondente e alle sue caratteristiche ed in parte all'interazione tra il rispondente, il questionario ed il rilevatore, nonché all'atteggiamento tenuto da quest'ultimo e, di riflesso, alle istruzioni che gli sono state impartite in merito.
c) misurazione non corretta dei fenomeni di interesse, dovuta ad una molteplicità di fattori che è possibile ricondurre essenzialmente al rispondente, al questionario, al rilevatore e all'influenza che egli esercita sul rispondente (Istat, 1989). Citando qualche esempio: disattenzione nella compilazione, lacune nella memoria, mancanza di informazioni, presenza di domande confuse o devianti, presenza di quesiti proxy, condizionamento del rilevatore e/o di terze persone presenti al momento dell'intervista.
d) scorretta codifica delle risposte aperte, ossia erronea interpretazione dell'informazione contenuta esse, che porta a scegliere ed attribuire un codice sbagliato all'interno delle classificazioni utilizzate. Questi errori sono imputabili a coloro che effettuano le suddette operazioni (rilevatori o non). In particolare per quanto riguarda classificazioni complesse ed articolate, come sono tipicamente quelle standard, gli errori sono legati principalmente ai casi meno comuni, a cui è difficile attribuire una collocazione indubbia; a tal proposito è

bene impiegare codificatori esperti, che conoscano molto bene sia il tema di interesse che la struttura della classificazione adottata.

e) Scorretta apposizione o trascrizione dei codici identificativi, che può impossibilitare il riconoscimento delle unità e delle loro relazioni, dei comuni, dei rilevatori e di tutte quelle entità soggette a sistema di identificazione. Questo tipo di errore è dovuto ai rilevatori e/o ai supervisori (spesso i comuni).

In quanto errori non campionari, quelli appena elencati sono oggetto del sistema di controllo della qualità. Vediamo in che modo, prendendo in esame per ora la parte relativa alla prevenzione ed al monitoraggio e rinviando alla fase di revisione per quanto concerne la loro correzione.

3.1.1 La prevenzione degli errori non campionari nella fase di rilevazione

Abbiamo già accennato al fatto che una parte degli errori che si manifestano durante la fase di rilevazione, riflettono in realtà insufficienze di quella progettuale. La prevenzione trova compimento quindi in larga misura durante la programmazione, attraverso i metodi presentati nel capitolo precedente; riepiloghiamoli brevemente: l'attenzione posta nel redigere il questionario, l'applicazione delle tecniche per il suo controllo, la scelta del disegno d'indagine più opportuno, l'effettuazione dell'indagine pilota (Istat, 1989). Oltre a queste, è opportuno effettuare delle altre azioni di prevenzione, anch'esse derivanti da scelte organizzative (contenute nel piano di rilevazione), ma messe in atto durante le fasi operative (Istat, 2000):

- S*elezione e formazione dei rilevatori*, i quali possono determinare, come abbiamo visto, numerosi tipi di errore. È importante che siano informati sulla totalità dell'indagine e non solo in merito alla parte di loro competenza, che colgano la gravità delle conseguenze che gli errori non campionari hanno sui risultati e che siano motivati in questa loro attività. A tale scopo è indispensabile un'attenta selezione e formazione, nonché una costante supervisione, tale da offrire assistenza ed appoggio addizionale rispetto a quanto riportato negli manuali.
  Quest'argomento troverà ampia trattazione, soprattutto in termini di riferimenti pratici, nel corso del capitolo 5.

- *Campagna di sensibilizzazione presso i rispondenti*, che può contribuire a creare un clima favorevole all'indagine, facendo comprendere l'interesse collettivo che riveste, come pure l'importanza di ogni singola collaborazione. L'intento è di ridurre la reticenza e la diffidenza delle persone, in modo da limitare le non risposte e al contempo assicurarsi partecipazioni serie e motivate. Rispondono a questo obiettivo sia la pubblicizzazione più generica, estesa a tutta la popolazione di interesse, sia gli interventi mirati alle sole unità selezionate per il campione, come la lettera di presentazione dell'indagine che preannuncia l'intervista, di solito firmata dal titolare del trattamento dei dati (es. presidente dell'Istat).

3.1.2 Il monitoraggio degli errori non campionari nella fase di rilevazione

Quest'operazione è qui riportata malgrado non avvenga contestualmente alla rilevazione, necessitando di informazioni disponibili in fase revisionale. Ciò nonostante, è indubbia la sua pertinenza tematica. Individuiamo due distinti tipi di azione, richiamando quanto già accennato in precedenza (Istat, 1989):

a) Misura diretta della precisione dei dati mediante la stima dell'errore totale, o meglio della sua componente variabile. Assumendo verosimilmente una semplificazione, cioè che l'errore variabile (approssimato dalla varianza totale) sia scomponibile in campionario (varianza campionaria) e non campionario, quest'ultimo dovuto al rispondente (varianza semplice di risposta) e all'influenza del rilevatore sul rispondente (varianza correlata di risposta), è possibile stimare queste componenti sostanzialmente replicando l'indagine, in modo da poter confrontare le risposte fornite nelle due diverse occasioni. Tale calcolo prevede risorse supplementari rispetto a quelle ordinarie; se queste, come spesso accade, non sono disponibili, il calcolo non viene effettuato, affidando agli indicatori indiretti la misura della precisione.

b) misura indiretta della precisione dei risultati mediante indicatori di qualità. La loro costruzione non comporta operazioni supplementari a quelle usuali dell'indagine e risulta quindi più economica e tempestiva rispetto alle operazioni di cui al punto precedente. Inoltre, visto che gli indicatori sono sempre specifici in relazione ad un

qualche aspetto, collegando questo alla sua fonte (o fonti) d'errore, è possibile la loro individuazione, quindi l'adozione di provvedimenti in merito. In questo senso possiamo intendere questi indicatori anche come "indici di performance" dei diversi oggetti/soggetti coinvolti nella rete di rilevazione (Istat, 1989). Elenchiamo di seguito alcuni indicatori in relazione all'aspetto che trattano:

- *Errori di lista e mancata risposta totale*.

Ponendo: n=numerosità campionaria n_R=questionari compilati
n_{NR}=questionari non compilati r=rifiuti a partecipare
nc=unità non a casa i=unità senza o con indirizzo sbagliato

- Errore di lista = i / n
- Errore complessivo di rilevazione = n_{NR} / n
- Mancata risposta lorda = (r + nc) / n
- Mancata risposta netta = (r + nc) / (n – i)

Se possibile, inoltre, è molto utile effettuare delle analisi sui fattori che influiscono sulla reticenza, esaminando le caratteristiche di chi si rifiuta di partecipare all'indagine; per tale scopo si utilizzano informazioni reperite all'anagrafe e/o i documenti aggiuntivi di rilevazione, compilati dal rilevatore, nei quali vengono rilevate informazioni sui non rispondenti.

- *Mancata risposta parziale*

Ponendo: n_R=risposte compilate nd=risposte non dovute
d=risposte dovute r=rifiuti
fc=valori fuori campo a=valori ammissibili
ac=valori ammissibili compatibili anc=valori ammissibili non compatibili

Suddividendo: dovuti e non a seconda delle regole di compilazione date dai quesiti filtro, ammissibili e non a seconda che stiano nel campo di variazione previsto per quella variabile, compatibili e non in funzione della congruenza con le altre variabili e interna alla variabile stessa (per maggior chiarezza vedi schema 3.3.2).

- Compilazione del quesito = (nd + a) / n_R
- Efficacia dell'intervista = ac / d

- Mancata risposta = (anc + fc + r) / d
- Rifiuto = r / d
- Incompatibilità = i/d

- *Intervista*

Ponendo: n_P=totale rispondenti proxy n_{IR}=totale individui rispondenti
n_R=questionari compilati n=numerosità campionaria
d_i=durata dell'i-esima intervista

- n_P / n_{IR} = percentuale proxy
- $1 - (n_P / n_{IR})$= dimensione campionaria reale
- n_R / n = effettuazione interviste
- ∑i di = durata media intervista

Il calcolo di indicatori di questo tipo separatamente per rilevatore può essere utile per esaminare il loro operato, oppure, distinguendo per comuni piuttosto che per aree socio-demografiche omogenee, si possono ottenere delle prime informazioni circa diverse attitudini alla risposta e/o al grado di collaborazione.

3.2. La registrazione su supporto informatico

La fase di registrazione consiste nella trascrizione delle informazioni presenti nei questionari cartacei su file compatibili con i programmi informatici, in modo che esse possano essere sottoposte a revisione ed elaborazione.
Il tema della registrazione non sarà oggetto di approfondimento in questa sede, fondamentalmente per tre motivi:

a) è minimo rispetto alle altre fonti d'errore
b) sono sempre meno frequenti i casi in cui il questionario sia compilato e registrato manualmente. Grazie all'ampia diffusione della trasposizione dati mediante lettore ottico e le tecniche di rilevazione che prevedono l'ausilio del computer, ovvero CASIC (Computer Assisted Survey Information Collection) come CAPI (Computer Assisted Personal Interviewing) e CATI (Computer Assisted Telephone Interviewing), questa fase, com'è intesa tradizionalmente,

di fatto viene a scomparire. Inoltre, le tecniche citate prevedono l'impiego di programmi predisposti per effettuare dei primi controlli di qualità sui dati, così da ridurre ulteriormente gli errori derivanti da quest'operazione (CGIS, 2000).

c) laddove sia ancora prevista la trascrizione manuale dal cartaceo, questa è spesso appaltata a ditte esterne specializzate, le quali sono normalmente tenute per contratto a rispettare prestabilite soglie d'errore (calcolate in percentuale sui byte digitati), pena il rifiuto dell'intero stock di dati (Istat, 1989); inoltre queste ditte utilizzano software programmati per riconoscere alcuni tipi di errore di immissione ed avvisare in tal caso l'operatore (CGIS, 2000).

Aggiungiamo soltanto che è comunque sempre conveniente effettuare dei controlli mirati per quanto riguarda le variabili strategiche per l'indagine, le variabili strutturali (quelle comunemente usate nelle intestazioni delle tabelle), i codici identificativi.

3.3. La revisione (o editing)

Mediante gli interventi di revisione, le informazioni rilevate e registrate vengono sottoposte a controllo e correzione. In sostanza, in questa fase trova concretizzazione uno degli obiettivi del sistema di controllo, ossia migliorare la qualità dei dati, mediante l'identificazione e la correzione degli errori (cfr § 1.5.3). Come sappiamo, è possibile individuarne solamente una parte, che chiameremo errori individuabili: si tratta delle mancate risposte totali e parziali e delle incoerenze presenti all'interno del sistema di identificazione. Alle non risposte totali solitamente si preferisce porre rimedio con dei pesi correttivi (come sarà spiegato nel capitolo dedicato all'elaborazione), le altre due tipologie d'errore, invece, sono appunto oggetto delle operazioni di editino (Istat, 1989). Sottolineiamo che le mancate riposte parziali sono da intendersi in senso lato, cioè comprensive di modalità fuori campo e incongruenze intra o intervariabile, determinando queste di fatto l'impossibilità di utilizzare il dato.
L'editing può essere effettuato con diversi metodi (Istat, 1999):

a) Manuale: effettuata per intero da esperti del settore che operano direttamente sul questionario

b) Interattivo: i programmi informatici ricercano gli errori e gli esperti trovano la soluzione per correggerli.
c) Automatico: sia la ricerca degli errori che la correzione vengono eseguite tramite procedure informatiche. E' chiaro che questa rappresenta la scelta obbligata in indagini di medio-grandi dimensioni; non dimentichiamo ad ogni caso che è l'esperto ad impartire le norme e selezionare i dati di imput per il software.

Le operazioni revisionali constano di due fasi (Istat, 1989): una prima di natura quantitativa, a cui accenneremo brevemente, la seconda di carattere qualitativo, che tratteremo invece in maniera più approfondita.

3.3.1 La revisione quantitativa

Il controllo quantitativo ha l'obiettivo di verificare che ci sia corrispondenza, nel numero e nelle relazioni, tra unità programmate, unità risultanti dalla rilevazione ed unità presenti nel file registrato.
Per unità, in questo contesto, si intende sia il modello di rilevazione, sia le unità di analisi, sia le istanze che sono coinvolte nell'organizzazione della raccolta e dell'elaborazione dei dati o sono rilevanti per essa. (Istat, 1989).
Rispondono alla definizione e sono dunque da considerarsi tali: lo strato, il comune, l'area interna al comune, il rilevatore, il modello di rilevazione e le unità di analisi.
Ognuna di queste unità è caratterizzata da codici, quindi è identificabile nei tre livelli di controllo sopraelencati; inoltre, il sistema di identificazione è strutturato in modo tale da potervi desumere altresì informazioni circa le relazioni tra queste unità (ad esempio il codice comune che precede il codice questionario o il codice famiglia anteposto a quello individuale). Pertanto il controllo quantitativo si realizza nel concreto come un'insieme di operazioni su detti codici; di analisi e confronto anzitutto ed eventualmente di intervento correttivo e ripristino delle corrispondenze, qualora errori commessi nelle fasi precedenti abbiano determinato incoerenze all'interno del sistema.
Le informazioni necessarie sono contenute nel piano di rilevazione per quanto riguarda numero e relazioni programmate, nei documenti di rilevazione per quanto attiene la situazione a rilevazione effettuata mentre le informazioni relative ai dati

registrati si trovano nel file proveniente dalla fase di registrazione (Istat, 1989). Relativamente, infine, al metodo adottato, benché esso dipenda in larga misura dal tipo di indagine e dalla sua dimensione, normalmente per l'editing quantitativo è di tipo misto, ossia ci si avvale sia degli esperti che dell'apporto dell'informatica.

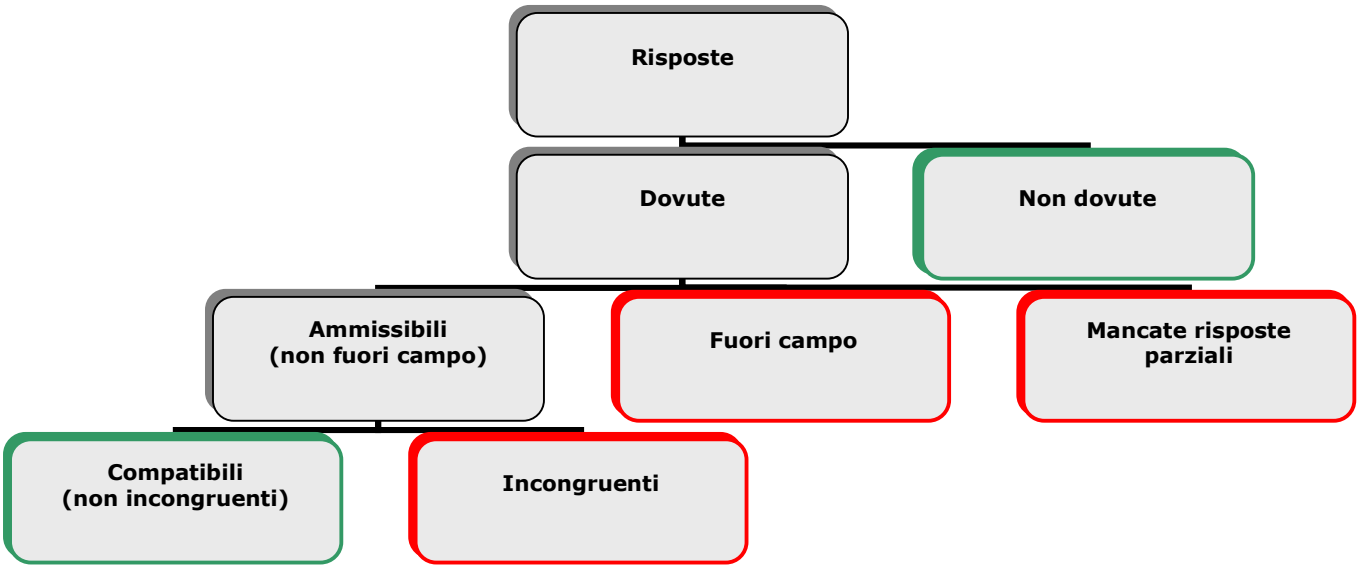
3.3.2 La revisione qualitativa

In questa sede analizzeremo la revisione qualitativa sottintendendola automatica, cioè svolta tramite procedure informatizzate; questo perché tornerà utile ricollegarsi alle seguenti nozioni nella seconda parte dell'elaborato, nel quale sarà presentato un software per il controllo e la correzione dei dati sviluppato e utilizzato dall'Istat (Concord).
Il controllo qualitativo si propone di rintracciare ed eventualmente correggere sia errori relativi a singoli quesiti, che tra quesiti differenti. In particolare (Istat, 1999):

i. Mancate risposte parziali: intese in senso stretto come rifiuti, incapacità o scarsa motivazione a rispondere alle singole domande. Sottolineiamo, nel caso non fosse così scontato, che i filtri escludono una parte dei rispondenti dai quesiti successivi, dando luogo di fatto a delle mancate risposte; esse non sono però da considerarsi errori, poiché rappresentano risposte non dovute, coerentemente con quanto previsto dalle norme di compilazione del questionario.
ii. Valori fuori campo: valori inammissibili in quanto esulano dall'insieme di codici previsti per una data variabile (es. scala di gradimento da 1 a 10, stato civile codificato con 2 codici, uno per libero, l'altro per coniugato, e così via)
iii. Incongruenze logiche all'interno delle singole variabili: valori impossibili o considerati implausibili, cioè esterni ad un ragionevole campo di variazione ammesso per una data variabile (ad esempio un numero di figli spropositato piuttosto che 30 febbraio come data di nascita)
iv. Incongruenze logico-formali tra variabili: abbinamenti impossibili o improbabili tra modalità di due o più variabili relative alla stessa unità (ad esempio un rispondente di otto anni che afferma di avere figli oppure un reddito inverosimile per la professione dichiarata o ancora risposta "No" alla domanda

filtro "E' stato in vacanza nell'ultimo anno?" e a seguire compilazione dei quesiti relativi alle caratteristiche della vacanza ecc..)

Riportiamo di seguito uno schema (tratto da Istat, 1989), che riteniamo essere utile come riferimento circa la scomposizione gerarchica delle risposte ad un qualsiasi quesito del questionario. Le caselle bordate in rosso indicano le risposte non valide, ossia gli errori sopraindicati; quelle bordate in verde, invece, le risposte valide:



Schema 3.3.2: Scomposizione delle risposte al generico quesito i (Istat , 1989)

E' immediato comprendere che, nel caso si verifichino uno o più errori, risulta impossibile, o quantomeno irrealizzabile nel concreto, ripristinare il valore esatto, poiché comporterebbe un ritorno presso le unità non rispondenti o per le quali si sono riscontrate le inammissibilità o le incongruenze. In riferimento all'editing qualitativo, pertanto, è poco opportuno parlare di correzione; nel seguito adotteremmo quindi il termine più appropriato di "imputazione", intendendo con questo che, al posto del valore mancante o non valido, viene forzato un altro valore, che si confida il quanto più possibile simile a quello esatto. Quest'intervento, sebbene rappresenti senza dubbio un'approssimazione, consente tuttavia una riduzione dell'errore rispetto al mantenere i dati "sporchi", così come si presentano una volta rilevati e registrati.

Il primo step consiste nel controllare i dati, ovvero nel ricercare gli errori presenti in essi. A tale scopo è necessario stabilire una serie di condizioni, soddisfatte le quali hanno luogo delle situazioni da considerarsi errate. Tali condizioni vengono esplicitate all'interno di un insieme di regole, che costituiscono il cosiddetto piano di compatibilità (Istat, 1999). Trattasi questa a dire il vero di una denominazione ingannevole, dato che nel piano non vengono descritte le condizioni che realizzano i casi corretti, così come il termine lascia intendere, bensì il contrario; in pratica i casi corretti vengono definiti per negazione, una volta individuati quelli errati.
Detto ciò, è utile suddividere le regole in due grandi insiemi, a seconda della loro natura (Istat, 1999):

a. Regole sostanziali, se riguardano nello specifico il contenuto dei dati, ovvero se concernono il significato che essi hanno su un piano di sostanza logica. Nel concreto queste regole si propongono di mettere in evidenza situazioni impossibili o non plausibili, sia in relazione a singole risposte, segnalando quando queste appaiono per nulla o poco ragionevoli, sia ad accostamenti tra risposte diverse, rintracciando quelli ritenuti inammissibili o inverosimili. In sostanza vanno definiti dei domini plausibili per ogni variabile, sia semplici che condizionati alle modalità assunte da altre variabili (CGIS, 2000); si tratta dunque di una demarcazione in certa misura arbitraria tra plausibile e non, che deriva da conoscenze a priori sulla realtà indagata. Per costruire queste regole in modo più possibile oggettivo, usualmente, ci si rifà ad informazioni sulle distribuzioni, determinando, per ogni modalità, la corrispondente probabilità (condizionata o meno) e quindi escludendo dal dominio quelle modalità che hanno una probabilità talmente bassa da risultare inaccettabile (un esempio di soglia standard è 3σ dalla media).
b. Regole formali, se riguardano la struttura delle risposte, a prescindere dal loro contenuto. In pratica servono per rintracciare la violazione dei filtri, analizzando combinazioni di risposte date da una stessa unità, e i valori fuori campo determinati dal mancato rispetto delle codificazioni, prestabilite per ogni variabile. Queste regole vengono derivate rispettivamente dalle norme di compilazione del questionario, schematizzate come sappiamo mediante diagrammazione, e dal piano di registrazione, nel quali sono esplicitate le decisioni in merito alle codificazioni.

Generalizzando indipendentemente dalla tipologia, una regola si presenta così:
Se [(condizione1),….,(condizione n)] allora [incompatibilità],
dove, al posto dei puntini di sospensione, vanno inseriti gli operatori logici, sia di tipo "or" che "and", a seconda di ciò che la situazione richiede.
L'insieme di tutte le regole costituisce, come detto, il piano di compatibilità. Affinché esso funga da strumento efficace è necessario che risponda a dei requisiti:

a. Completezza, ovvero racchiuda le regole occorrenti per individuare tutti gli errori individuabili. In caso contrario, trascurando una parte di essi, c'è il rischio che la fase di revisione diventi paradossalmente ulteriore fonte di errori.
b. Essenzialità, cioè non contenga regole superflue e ridondanti, non fosse altro per il fatto che queste incidono negativamente sulla velocità di elaborazione del processore. A prescindere da questo, rischiano inoltre di modificare i dati più di quanto non sia strettamente indispensabile.
c. Coerenza, ossia non includa regole tra loro contraddittorie perché ciò in pratica equivarrebbe a non effettuare alcuna revisione o addirittura potrebbe comportare un peggioramento della qualità dei dati (Istat, 2000).

Approntato il piano in modo tale che possieda queste caratteristiche, esso viene applicato ai dati. Mediante quest'operazione, si perviene all'individuazione dei record errati, per i quali cioè si sono verificate le condizioni di errore stabilite nelle regole del piano (Istat, 1999). E' fondamentale comprendere che questo risultato non implica necessariamente che si possano identificare le variabili che saranno da sottoporre ad imputazione: infatti, come già ripetuto in diverse occasioni, ad attivare una regola può essere una sola modalità di una variabile ma anche la combinazione di modalità relative a più variabili. In questo secondo caso non sempre è possibile individuare qual è la variabile errata. Riportando un esempio banale: tra la modalità "Femmina" e la modalità "Militare di leva" sussiste certamente un'incompatibilità, ma non possiamo decidere sulla base di questi elementi quali tra le due è sbagliata e quindi a quale verrà forzato un altro valore. In altri casi, oppure quando la regola è attivata da una modalità singola, può essere invece agevolmente individuata la variabile responsabile dell'errore. (Ad esempio, se una persona intervistata dichiara di non avere figli e a seguire riporta i loro dati anagrafici, è chiaro che ha commesso una distrazione nel rispondere al quesito filtro).

Quanto argomentato costituisce il problema della localizzazione delle variabili da imputare; distinguiamone due casi, a cui corrispondono due diversi modi di procedere (Istat, 2004):

a. Probabilistico: non è possibile dedurre su quale variabile cada la responsabilità dell'errore, per cui la localizzazione è demandata al software. Rinviando al capitolo dedicato a Concord per quanto attiene ai criteri adottati per operare la scelta, basti per ora sapere che viene effettuata in modo da produrre meno modificazioni possibili dell'informazione raccolta
b. Deterministico: la localizzazione non costituisce un problema, perché è immediato comprendere quale sia la variabile sbagliata. In questo caso la regola generalizzata, come sopra definita, continua in questo modo:
   Se [(condizione1),….,(condizione n)] allora [variabile da imputare=x]

A questo punto, compiuta anche la localizzazione (mediante software o meno), si può procedere con l'imputazione delle variabili giudicate non valide. Anche quest'operazione può essere eseguita adottando diversi modi di operare; in particolare va fatta una grande distinzione a seconda che si riesca a capire con certezza quale valore forzare al posto di quello errato oppure no. Nel primo caso si applica il metodo di imputazione deterministica, diversamente uno degli altri metodi che seguono in elenco (Istat, 2004):

a. Deterministico: il valore da sostituire al posto di quello non valido è predeterminato; la generica regola sopra enunciata si completa dunque in questo modo:
   Se [(condizione1),….,(condizione n)] allora [a variabile x imputa valore=y]
   Così strutturata la regola viene chiamata R.I.D., ossia Regola di Imputazione Deterministica (CGIS, 2000). Riprendendo l'esempio di prima, con essa verrebbe forzato il valore "Si" alla variabile che indica se il rispondente ha figli.
b. Da donatore: alle unità affette da errori vengono attribuiti i valori assunti dalle corrispondenti variabili in altre unità, scelte tra quelle con record interamente valido, ossia esente da errori. Viene chiamato ricevente il primo gruppo di unità, donatore il secondo. Esistono diverse applicazioni di questo metodo:
   i. Da donatore cold deck: i record vengono preliminarmente suddivisi in interamente validi e con almeno un errore; per ogni unità con record

non valido viene poi casualmente scelto un record dal gruppo delle unità donatrici, a cui attingere per i valori da imputare (CGIS, 2000).

ii. Da donatore hot deck: i record vengono tenuti tutti insieme e ordinati secondo una variabile significativa, vale a dire che si ipotizza avere una certa influenza su quelle relative al fenomeno oggetto d'indagine (potrebbe essere l'età o il titolo di studio o ancora la ripartizione geografica ecc). Quanto più la variabile è correlata con il fenomeno in questione, tanto più l'ordinamento del file assicura che le unità adiacenti siano simili, cioè che abbiano presumibilmente fornito risposte analoghe (se non altro in misura maggiore di quanto non accada scegliendo a caso tra tutti i possibili donatori). I valori non validi all'interno di un record vengono quindi sostituiti con i corrispondenti dell'unità immediatamente precedente (CGIS, 2000). Due i limiti di quest'approccio: il primo è che si ha a disposizione una sola variabile per ordinare il file, quando è noto che i fenomeni dipendono sempre da una molteplicità di fattori; il secondo è che qualora siano molti i record non validi adiacenti, essi utilizzano tutti quanti i valori dell'ultimo donatore disponibile, creando dei picchi fittizi all'interno delle distribuzioni in corrispondenza di questi valori (Istat, 2004).

iii. Da donatore con distanza minima: la logica alla base di questo metodo è la stessa del precedente, l'unica differenza è che permette di tenere sotto controllo più fattori. Infatti, per ogni unità con record non valido, è possibile scegliere la più simile tra quelle donatrici in riferimento a più variabili, mediante dei calcoli di distanza che esamineremo in dettaglio nel capitolo 6. Lo svantaggio è comune al metodo precedente, ossia uno stesso donatore rischia di essere sfruttato più volte (alcuni programmi permettono di sopperire a questo problema; come vedremo, anche Concord)

c. Mediante regressione: utilizzando i dati dei soli record donatori, si calcolano le funzioni di regressione per tutte quelle variabili che nei record riceventi sono affette da errore. Si cerca cioè di trovare una modello matematico che spieghi ognuna di queste variabili in funzione di altre, dette predittive. Grazie al modello diventa quindi possibile calcolare il valore della variabile che dipende da quelle predittive, solo conoscendo i valori che queste assumono. Quindi, se per i record riceventi sono disponibili queste informazioni (ossia le variabili

predittive non sono affette da alcun errore), applicando la funzione calcolata sui donatori, si ottiene il valore da imputare alla variabile non valida. Il limite è che spesso questo metodo risulta inapplicabile, dato che non sempre le variabili predittive sono disponibili per i record riceventi; inoltre, è necessaria un'attenta analisi del fenomeno per costruire un modello di regressione efficace, che identifichi i fattori che effettivamente hanno maggior influenza sulla variabile.

d. Da valor medio: ogni valore non valido viene sostituito con media (o mediana, o moda, a seconda della natura della variabile), calcolata su tutti i dati validi per quella variabile, oppure in un sottogruppo di questi (scelto logicamente in maniera significativa). È un metodo di semplice applicazione, ma che distorce la distribuzione della variabile, inserendovi picchi artificiali in corrispondenza del valor medio generale o dei valori medi dei vari sottogruppi individuati.

3.4. L'elaborazione

Per elaborazione intendiamo l'insieme di tutte quelle operazioni effettuate sui dati con il fine di acquisire conoscenza in merito al fenomeno oggetto d'indagine (Istat, 2000). Questa fase serve in pratica per rendere realmente utile ciò che è stato rilevato, sfruttando le potenzialità informative insite nei microdati. Parliamo di potenzialità perché è chiaro che risulta arduo, se non impossibile, trarre delle conclusioni sul fenomeno in questione basandosi su un elenco di record individuali, molto spesso di considerevole numerosità. Si rende dunque necessario effettuare delle operazioni che sintetizzino le informazioni contenute in questi record, producendo i cosiddetti macrodati: può trattarsi di semplici statistiche descrittive (come medie, mode, mediane, varianze ecc) oppure di indici più complessi, con i quali sondare le relazioni tra le variabili, o ancora di distribuzioni semplici o congiunte (le quali in pratica mostrano come si ripartiscono le unità tra le varie modalità, o combinazioni di modalità). I macrodati, naturalmente, vanno poi interpretati, desumendovi tutti gli elementi occorrenti per formulare delle considerazioni sulla realtà di interesse; la natura di queste considerazioni dipende dagli obiettivi stabiliti in sede progettuale, quindi di volta in volta esse possono riguardare relazioni tra variabili, confronti spazio-temporali piuttosto che previsioni per il futuro o anche solo semplici descrizioni delle caratteristiche del fenomeno.

Il punto di partenza di questa serie di operazioni è costituito dai microdati, quindi è opportuno in primo luogo valutarne l'adeguatezza. Sappiamo che, grazie agli interventi revisionali, disponiamo di record completi, ossia esenti da mancate risposte parziali, e corretti, rispetto ad incongruenze di qualsivoglia natura. Sappiamo anche, però, che esiste un altro errore non campionario tra quelli individuabili, a cui la fase di revisione non pone rimedio: si tratta delle non risposte totali, ossia dei questionari non compilati integralmente, in corrispondenza dei quali ci sarebbero dovuti essere dei record che invece, per forza di cose, mancano. Quindi, in sostanza, i microdati sono da considerarsi adeguati, ma bisogna ovviare in qualche modo al fatto che sono in numero minore di quanto previsto. Ciò è possibile semplicemente operando un piccolo accorgimento matematico, ossia attribuendo alle unità rispondenti dei coefficienti correttivi, in modo che le loro risposte pesino anche per quelle che avrebbero dato coloro che non hanno partecipato all'indagine. E' indubbiamente più efficace riportare un esempio, seppur da "caso limite": se la dimensione programmata del campione, è di 100 unità, ma la metà di queste rifiuta di compilare il questionario, otteniamo un tasso di non risposta di 50/100=1/2; attribuendo ai rispondenti un coefficiente pari all'inverso del tasso di riposta, quindi uguale a 2, facciamo in modo che ognuno di loro pesi doppio, ossia che 50 persone riescano a rappresentarne 100. Effettuare quest'operazione sottace l'ipotesi che chi non ha compilato i questionari avrebbe dato le stesse risposte di chi invece lo ha fatto. È un'ipotesi forte, che si riesce tuttavia a rendere più verosimile effettuando una preliminare suddivisone del campione in gruppi omogenei (rispetto naturalmente a variabili significative per l'indagine, cioè correlate con il fenomeno trattato) e calcolando i tassi di non risposta, quindi i coefficienti, separatamente per gruppo. Questo fa si che ogni rispondente pesi per i non rispondenti a lui simili, per i quali cioè si può supporre che le risposte sarebbero state analoghe.
Esistono altri coefficienti che vengono applicati alle unità, a cui accenniamo solo brevemente, non perché non siano importanti, tutt'altro, ma per il fatto che dipendono nello specifico da scelte inerenti la strategia di campionamento, che nell'elaborato non sono state trattate debitamente:

i. Coefficienti di riporto all'universo: il campione è scelto in modo da essere rappresentativo della popolazione di interesse, quindi i dati rilevati sulle unità campionarie possono (anzi, devono!) essere riferiti all'intera popolazione. È necessario però ripristinare gli ordini di grandezza, se si vuole che

l'informazione sia "tradotta" in modo corretto. A tale scopo si usano appunto i coefficienti di riporto, i quali fanno sì che il campione pesi anche per le unità della popolazione non estratte. Il calcolo è molto semplice; ipotizzando che su una popolazione di 1000 unità sia selezionato un campione pari ad 1/10, significa che 100 unità ne devono rappresentare 1000, quindi che ogni unità ne deve rappresentare 10: questo è proprio il peso da applicare. Il coefficiente si diversifica poi in funzione della strategia di campionamento scelta.

ii. Coefficienti di correzione: qualora, confrontando le distribuzioni di variabili importanti con quelle fornite da fonti esterne attendibili si dovessero riscontrare delle discrepanze, è conveniente correggere i dati raccolti. Ad esempio, se da dati censuari sappiamo che le persone con titolo di studio alto sono il 30% e nell'indagine risultano invece essere il doppio, ciò provoca distorsioni nei dati, soprattutto se si ha ragione di credere che questa variabile influisca su quelle oggetto d'indagine. Applicando ai rispondenti con titolo di studio alto un coefficiente pari a 30/60=1/2, facciamo in modo che ognuno di loro pesi la metà e, specularmente, applicando ai rimanenti con titolo di studio basso un coefficiente di 70/40=7/4, facciamo si che ognuno di loro pesi più di un'unità (esattamente 3/4 in più). Così facendo sistemiamo la distribuzione, che per un qualche motivo era distorta (forse, in questo caso specifico, a causa della diversa propensione a partecipare all'indagine in funzione del titolo di studio).

iii. Coefficienti previsti dalla modalità di selezione: oltre ai pesi di riporto all'universo, di cui al punto i., la strategia di campionamento comporta a volte di dover calcolare ulteriori coefficienti. Ad esempio, se all'interno dello stratificato si decide di sovrarappresentare un particolare gruppo, bisogna poi sistemare i valori ottenuti in modo che siano rappresentativi della reale quota di quel gruppo nella popolazione. Ovvero, se si estrae tre volte il dovuto, bisognerà poi applicare un coefficiente uguale ad 1/3.

Dopo aver calcolato i coefficienti, si può dunque procedere con le varie operazioni di elaborazione, che portano alla produzione dei macrodati. Questi, spesso sottoforma di tabelle, corredati da opportuni commenti, interpretazioni, considerazioni e grafici, costituiscono il materiale che verrà diffuso.

3.5. La validazione e la diffusione dei risultati

Il materiale che si intende diffondere va prima sottoposto ad un ultimo passaggio di controllo, che serve per validare i risultati, ossia per appurare che essi effettivamente rispondano agli obiettivi preposti. In linea di massima, al di là delle singole indagini, il controllo si rifà in larga misura alle dimensioni della qualità definite nel primo capitolo, verificando che (Istat, 2000):

i. I risultati non contengano delle palesi anomalie in riferimento a confronti con altre indagini simili o con le serie storiche dell'indagine, se ripetuta.
ii. Ci sia coerenza tra le tavole, ossia non si evidenzino contraddizioni tra i medesimi dati presenti in tavole diverse (ad esempio media della popolazione, distribuzioni marginali, dimensione del campione ecc..)
iii. Il materiale effettivamente risponda alle esigenze conoscitive che l'indagine si è proposta di soddisfare e non sia ridondante, ossia appesantito inutilmente; bisogna dunque valutare l'opportunità di aggiungere, rimuovere, sostituire tavole.
iv. Siano acclusi tutti i metadati prodotti nel corso delle varie fasi dell'indagine, in modo che l'utente abbia gli elementi per valutare criticamente la personale utilità dei risultati; sono da intendersi metadati tutti quelli che informano circa l'indagine, quindi i documenti di progettazione, il questionario, la stima diretta della precisione dei risultati (se effettuata), gli indicatori indiretti eccetera…
v. I risultati siano accessibili ai fruitori a cui sono rivolti; ciò si rende possibile grazie a scelte adeguate in merito a modalità di diffusione e pubblicizzazione

Naturalmente sarebbe auspicabile che questo ultimo step rappresentasse solo la riprova della qualità del materiale, giacché gli obiettivi dovrebbero essere già garantiti dall'adeguato svolgimento dell'indagine e delle azioni del sistema di controllo durante il processo.

Parte seconda

L'ESPERIENZA DI STAGE PRESSO IL SERVIZIO STATISTICA DELLA PROVINCIA DI TRENTO

4. L'ENTE OSPITANTE: IL SERVIZIO STATISTICA DELLA PROVINCIA DI TRENTO

Il servizio statistica della Provincia di Trento è stato istituito nel 1981 e provvede all'esercizio delle funzioni provinciali per quanto concerne l'ambito statistico. Adempie agli incarichi di ufficio regionale ISTAT, eseguendo le rilevazioni dell'Istituto che rientrano nelle materie di competenza provinciale, effettua indagini statistiche proprie, con relative documentazioni e pubblicazioni, realizza eventuali lavori statistici commissionati da amministrazioni ed enti pubblici, organizzazioni, associazioni e privati. Predispone altresì il materiale statistico richiesto dagli organi di governo provinciale per i documenti e le relazioni programmatiche e inoltre presiede alle ricerche statistiche curate dagli altri servizi della provincia, nei confronti dei quali ha funzioni di indirizzo e coordinamento (PAT, 2000).

Nel periodo che va da febbraio a maggio 2008 sono stata ospite come stagista del Servizio Statistica della Provincia di Trento; ho avuto modo altresì di svolgere attività di rilevatrice presso il medesimo. Quest'esperienza mi ha dato l'opportunità di saggiare nel concreto degli step di un processo di indagine; in particolare ho potuto approfondire ciò che riguarda le azioni di prevenzione degli errori non campionari nella fase di rilevazione (per i riferimenti teorici vedi §3.1.1) ed ho avuto modo di avvicinarmi ad un software per la revisione qualitativa (per i riferimenti teorici vedi §3.3.2)

5. LE AZIONI DI PREVENZIONE DEGLI ERRORI NON CAMPIONARI NELLA FASE DI RILEVAZIONE

In questo capitolo presenterò le azioni di prevenzione degli errori non campionari che ho potuto sperimentare in prima persona: si tratta di quelle azioni proprie della fase di rilevazione, quindi successive alla redazione e al controllo del questionario ed in generale ad ogni attività che rientra nell'ambito della programmazione dell'indagine.
Saranno dunque oggetto di tale capitolo: la selezione dei rilevatori, ossia come vengono scelti i rilevatori che entrano a far parte dell'archivio della provincia; la giornata di istruzione ai rilevatori, ovvero le nozioni basilari e le regole che vengono impartite per poter svolgere questo lavoro; la formazione specifica, cioè tutte le informazioni necessarie che vengono date in riferimento ad una particolare indagine; l'assistenza, il monitoraggio ed il controllo dei rilevatori ed infine la sensibilizzazione dei rispondenti.

5.1. La selezione dei rilevatori

Un candidato rilevatore deve innanzitutto far pervenire domanda al Servizio Statistica: è sufficiente a tale scopo compilare l'apposito modulo di iscrizione, presente on-line alla pagina del Servizio. Nel modulo vengono richieste, oltre alle generalità anagrafiche, alcune informazioni come la condizione professionale, il titolo di studio, le conoscenze linguistiche ed informatiche, il possesso di patente, eventuali precedenti esperienze come rilevatore. C'è inoltre la possibilità di indicare il comprensorio ed il comune di preferenza lavorativa, nonché la disponibilità nel corso dell'anno.
Le schede dei candidati di norma non vengono sottoposte a pre-selezione; essi vengono contattati, in linea di massima, secondo l'ordine di arrivo della richiesta ed invitati a gruppi presso la sede del Servizio per effettuare delle prove di selezione. Si tratta di due test, sempre più utilizzati per la selezione del personale anche da numerose aziende, soprattutto di medio-grandi dimensioni: uno è di tipo psico-attitudinale, l'altro logico-verbale.
Il primo consta di quasi 200 item, ovvero delle affermazioni con le quali il candidato, pensando al proprio caso, può dichiararsi in accordo oppure in disaccordo; alcune

sono riferite ad un'ipotetica situazione concreta, del tipo "Se la banca mi accreditasse dei soldi per sbaglio, li restituirei", altre sono più generali ed astratte, come ad esempio "Credo che se tutti fossero più sinceri, questo mondo sarebbe migliore", altre ancora indagano sugli atteggiamenti e sulle inclinazioni del candidato, tipo "Mi piace mangiare da solo" o "Non sono solito raccontare fatti personali ad estranei".
Gli item sono accorpabili in gruppi omogenei rispetto al tratto della personalità che indagano, ma nel test sono posizionati in ordine sparso, così che difficilmente si riescono a cogliere le connessioni tra le varie affermazioni mentre lo si compila. In questo modo si ottengono delle risposte spontanee, invece che pilotate ad hoc per preservare la coerenza; ad ogni modo, sono inseriti dei quesiti "trabocchetto" che consentono di smascherare eventuali "forzature strategiche".
Il test compilato, opportunamente analizzato da uno psicologo, permette di delineare una sorta di "profilo psicologico" del candidato, in particolare per quanto riguarda quei tratti di personalità che maggiormente assumono rilevanza in riferimento all'attività di rilevatore (si veda in merito il paragrafo seguente). L'obiettivo è quello di stabilire in che misura la personalità del candidato risponde al profilo richiesto per svolgere questo tipo di lavoro. Questa corrispondenza si verifica per la maggior parte dei candidati, per la ragione che essi tendono ad autoselezionarsi, ovvero, in termini equivalenti, inoltrano domanda per lo più soggetti che si sentono portati (almeno dal punto di vista caratteriale) per questo tipo di attività.
Il secondo test logico-verbale serve invece per accertare che il candidato possegga le abilità minime necessarie per poter fare il rilevatore, cioè sappia destreggiarsi con dimestichezza nel ragionamento verbale e abbia buone capacità logiche, abilità senza le quali diventa alquanto problematico gestire un questionario. Il test intende verificare al contempo che il candidato sappia applicare queste capacità con immediatezza e prontezza; per questo motivo è previsto un vincolo di tempo per la consegna.
Gran parte del test consiste nel completare delle frasi, inserendo una coppia di parole tra quattro disponibili (riportando un esempio banale: "….. sta a cane come miagolare sta a ……." , che richiede l'accoppiata "Abbiaiare-gatto").
Sono poi presenti dei problemi di ragionamento matematico, che consistono principalmente nel completare logicamente delle serie numeriche, come ad esempio: "1, 2, 0, 3, -1, 4, …..".

La correzione del test porta ad attribuire un punteggio ad ogni candidato. Questo, integrato con quanto emerge dal test psico-attitudinale, permette di stilare una sorta di "graduatoria" dei rispondenti secondo il grado di idoneità per questo lavoro.
Infine, ad ogni candidato viene chiesto di presentarsi ai presenti; ciò permette al responsabile della selezione (nonché psicologa) di saggiare delle basilari capacità comunicative e di interazione con degli sconosciuti, come pure di conoscere la motivazione che ha portato a fare richiesta; inoltre, da l'opportunità ai candidati di esprimere delle preferenze o manifestare delle esigenze specifiche, delle più svariate, come ad esempio il privilegiare le telefoniche alle faccia a faccia piuttosto che il negare la disponibilità per le rilevazioni con quesiti sensibili o quant'altro.
I candidati che vengono valutati positivamente nei due test entrano nell'archivio provinciale del Servizio Statistica e sono invitati a partecipare alla giornata di istruzione; in seguito possono dunque essere richiamati per svolgere una o più rilevazioni. La scelta dell'indagine per cui un soggetto viene contattato tiene conto dalle sue caratteristiche (ad esempio il titolo di studio, se può avere una qualche correlazione con il tema dell'indagine), dalle necessità o dai desideri che egli stesso ha espresso e da eventuali annotazioni della psicologa.
I test di selezione illustrati, oltre a poter essere somministrati contemporaneamente a più persone, si basano su punteggi calcolati in modo standard uguale per tutti e sono quindi indubbiamente più imparziali ed oggettivi.

5.2. La giornata di istruzione

I soggetti che hanno superato le prove di selezione sono tenuti a partecipare alla giornata di istruzione presso la sede del Servizio Statistica, affinché acquisiscano le nozioni base e le regole fondamentali per poter intraprendere il lavoro di rilevatore.
Dopo una premessa introduttiva destinata alla presentazione del Servizio, nella quale vengono descritti brevemente i compiti che esso assolve e le rilevazioni che conduce, segue una parte dedicata alle principali normative che regolano l'attività statistica, ossia il decreto 322 del 1989 e la legge 196 sulla privacy del 2003. Naturalmente non vengono sciorinati ad uno ad uno nel dettaglio tutti gli articoli di cui si compongono i suddetti decreti, né si pretende che i neorilevatori li imparino, ma vengono solamente messe in evidenza le parti salienti dei medesimi.

In particolare, viene posta attenzione sui concetti di (PAT, 2000):

- *Segreto statistico*: i dati raccolti vengono usati solo per scopi statistici e non possono essere diffusi se non in forma aggregata, cioè in modo che non sia possibile risalire alla persona fisica o giuridica a cui si riferiscono;
- *Segreto d'ufficio*: tutti addetti agli uffici di statistica non possono comunicare a terzi informazioni di cui sono venuti a conoscenza durante l'esercizio delle loro funzioni;
- *Consenso*: ossia autorizzazione da parte del rispondente al trattamento dei propri dati personali; non è richiesto quando il trattamento è necessario per scopi scientifici e statistici e/o indispensabile per adempiere un obbligo previsto dalla legge;
- *Obbligo di risposta*: obbligo di fornire i dati richiesti (ad eccezione di quelli sensibili) se l'indagine che li rileva è inserita nel Piano Statistico Nazionale ed in una delibera del Presidente del Consiglio oppure nel Piano Statistico Provinciale ed in una delibera del Presidente della Provincia;
- C*asualità dell'estrazione*: i nominativi che entrano a far parte del campione sono scelti in base a metodologie di estrazione assolutamente casuali.

Queste nozioni sono quelle che un rilevatore deve assolutamente conoscere; viene sottolineata più volte l'importanza dell'acquisirle con sicurezza, poiché egli deve essere in grado di spiegarle a sua volta ad eventuali intervistati che chiedano delucidazioni in merito. Non è infatti raro che i soggetti campionati vogliano avere delle informazioni di carattere legislativo, richiedendo in particolare di conoscere i diritti ed i doveri che la normativa vigente prevede: è fondamentale che il rilevatore sappia fronteggiare e risolvere le loro perplessità in modo sicuro e competente.
Chiariti eventuali dubbi dei partecipanti relativamente agli aspetti normativi, la formazione procede con la descrizione del "profilo professionale" ideale di un rilevatore, illustrando le caratteristiche, le attitudini, le capacità che un buon rilevatore dovrebbe possedere, ovvero, riassumendo (PAT, 2000):

- Interesse a relazionarsi con persone estranee e capacità di farlo in maniera disinvolta, creando e mantenendo un clima favorevole, ossia cordiale ma non informale;

- Una buona dose di autocontrollo e pazienza, per gestire i soggetti difficili e le situazioni impreviste e per trattenersi dall'esprimere opinioni personali, che possono influenzare le risposte degli intervistati;
- Buone capacità comunicative per convincere il rispondente a partecipare, per avviare e mantenere una conversazione formale, per motivare il rispondente a portare a termine l'intervista e inoltre padronanza delle informazioni necessarie per rispondere senza alcun dubbio alle domande rivoltegli;
- Buona capacità di coordinamento, concentrazione e memoria, per riuscire a gestire l'intervista con attenzione ed in modo critico (cioè cogliendo le incoerenze e le illogicità delle risposte) e seguire contemporaneamente le istruzioni del questionario.

Queste caratteristiche ed abilità vengono riprese nella parte seguente, nella quale vengono fornite le informazioni su come si deve svolgere un'intervista e le regole di comportamento da seguire. In particolare, viene posto l'accento su (PAT, 2000):

- *La presentazione*: cruciale per minimizzare i rifiuti e per ottenere un maggior grado di collaborazione nel corso dell'intervista. Vengono spiegate le norme base da seguire per un buon approccio: il rilevatore, con il cartellino in vista, deve presentarsi chiaramente, indicando nome e cognome e specificando che rappresenta il Servizio Statistico della Provincia, deve spiegare qual'è il motivo della sua visita e descrivere brevemente scopi e contenuti della rilevazione (o accennare al contenuto della lettera di presentazione se, come spesso accade, questa è prevista per avvisare anticipatamente i soggetti campionati). Inoltre, qualora ne ravveda la necessità, deve rassicurare in merito a riservatezza e casualità dell'estrazione, facendo riferimento alle nozioni sopra menzionate; il consiglio che viene dato è di far leva sull'utilità della collaborazione e mai, se non come "ultima spiaggia", sull'obbligatorietà della risposta (qualora prevista);
- *La neutralità del rilevatore*: il rilevatore non deve influenzare in alcun modo il rispondente; a questo proposito viene raccomandato di leggere le domande e le modalità di risposta in modo neutro, senza particolare inflessioni nel tono di voce; di omettere i pareri personali e di non farli intuire in alcun modo; di controllare il linguaggio non verbale, senza manifestare reazioni di alcun tipo quando viene fornita la risposta. Inoltre, viene fatto divieto di interpretare in modo personale le domande o di fornire chiarimenti con parole proprie, perché ciò può distorcere il

senso originario ed infine viene raccomandato di riportare fedelmente le risposte date ai quesiti aperti, senza riassumere o condensare, per non correre il rischio di alterarne il significato;

- *Il clima formale*: il rilevatore è invitato a tenere un atteggiamento professionale in ogni caso, mantenendo sempre la conversazione su un piano formale. Inoltre, è d'obbligo la cortesia, la pazienza e la calma: non deve mai mostrare fretta, rileggendo le domande se ciò è necessario e mostrandosi sempre disponibile ad aiutare il rispondente se questo è in difficoltà; viene consigliato, però, di non esagerare con la condiscendenza e di non indugiare nel riportare l'attenzione del rispondente sull'intervista, se questo divaga in digressioni di carattere personale o comunque non pertinenti;
- *L'individualità dell'intervista*: è conveniente che durante l'intervista siano presenti solamente il rilevatore ed il rispondente; se non è possibile evitare la presenza di altre persone, queste devono essere caldeggiate a non partecipare e non suggerire, per non influenzare le risposte. Al rilevatore è raccomandato di mettersi sempre di fronte alla persona da intervistare, senza mostrare lo schermo del pc, affinché sia mantenuta sempre viva l'attenzione del rispondente.
- *La discrezione*: il rilevatore deve tutelare sempre la riservatezza dei rispondenti; viene fatto notare che delle azioni banali, spesso commesse soprappensiero, come lasciare il foglio con i nomi delle persone da intervistare sul tavolo piuttosto che mostrare la schermata del pc che riporta gli stessi nominativi, rappresenta a tutti gli effetti una violazione di questa regola ed è quindi necessario prestare attenzione perché ciò non accada.

Tutte le nozioni e le informazioni impartite durante la giornata di formazione sono contenute in un manuale "Intervistare – Come e perché", che viene consegnato ai rilevatori.

5.3. L'indagine "Condizioni di vita delle famiglie trentine"

La mia personale esperienza di rilevatrice è da riferirsi ad un'indagine denominata "Condizioni di vita delle famiglie trentine"; si tratta di un'indagine longitudinale, che ha avuto inizio nel 2005 e viene reiterata da allora con cadenza annuale. Rientra nel Programma Statistico Provinciale e, riconosciutane l'importanza non solo ad un livello

locale, anche nel Programma Statistico Nazionale, il quale contiene tutte le rilevazioni svolte dai soggetti del Sistan considerate di pubblico interesse (Istat, 2008).

5.3.1 Presentazione ed obiettivi

L'obiettivo che la suddetta indagine persegue è di delineare un quadro completo ed approfondito sulle condizioni di vita e sul livello di benessere dei residenti in provincia di Trento. In particolare, l'attenzione è rivolta alle realtà famigliari, esaminate nella loro struttura come pure nella combinazione delle caratteristiche e delle esperienze dei singoli membri che vi appartengono; l'intento, infatti, è quello di investigare, descrivere e comprendere il fenomeno in relazione alle storie di vita famigliari ed individuali. A tal fine, diventa fondamentale anzitutto ricostruire l'intera esistenza degli intervistati, per quanto attiene, naturalmente, ai temi di interesse per l'indagine ed è poi necessario monitorare le modificazioni e le evoluzioni che intervengono nel corso della rilevazione, sia a livello di situazione famigliare, sia a livello di percorsi individuali. Note queste esigenze, è agevole comprendere il motivo di un questionario ricco di quesiti retrospettivi e, per quanto riguarda la tipologia d'indagine, la ragione per cui la scelta sia necessariamente ricaduta su un campione di tipo "panel" (le cui unità vengono contattate per un minimo di quattro anni).
L'indagine è stata promossa dalla presidenza della provincia per supplire alle carenze e alle limitatezze delle basi informative preesistenti sul tema, le quali non riuscivano a soddisfare l'esigenza, sempre più sentita sia in ambito di ricerca che di programmazione politica, di dati maggiormente precisi ed analitici, che permettano non solo di fornire una descrizione del fenomeno, ma anche e soprattutto di seguirne l'evoluzione nel tempo, in relazione a fattori odierni e passati, in modo tale da poter avanzare delle ipotesi su come questi concorrono a determinarlo.
A questo proposito è opportuno puntualizzare che, innegabilmente, alcune indagini Istat raccolgono informazioni su aspetti sociali ed economici di fatto analoghi a quelli previsti da quest'indagine, ma non risulta possibile, utilizzando i dati a cui si perviene con esse, effettuare delle analisi comparabili. Infatti, anche considerando le due rilevazioni più simili a questa (ovvero l'indagine condizioni di vita - eu-silc e l'indagine sui consumi delle famiglie italiane), si evidenzia chiaramente una notevole differenza, ossia che in nessuno dei due questionari compaiono batterie di domande che intendono ricostruire le storie di vita dei rispondenti o, in termini equivalenti, che essi raccolgono informazioni limitatamente a periodi di riferimento coincidenti o prossimi a

quello di rilevazione. Di conseguenza, diventa irrealizzabile effettuare uno studio congiunto dei fattori presenti e passati ed analizzare dunque gli effetti che le vicende famigliari e i "contributi" apportati dalle esperienze dei singoli membri hanno sulle condizioni di vita delle famiglie.
Un ulteriore problema è che l'Istituto, dovendo monitorare l'intero territorio nazionale, si trova costretto ad attribuire alla provincia di Trento un campione di ridotte dimensioni, che per un'analisi a livello locale non garantisce un soddisfacente livello di precisione nelle stime: basti pensare che, a fronte di un campione di circa 600 famiglie per la rilevazione Istat annuale "Indagine sui consumi", "Condizioni di vita delle famiglie trentine" si avvale di un campione di dimensioni circa cinque volte maggiori.

5.3.2 Popolazione di riferimento e disegno di campionamento

La popolazione di riferimento è costituita dalle famiglie di fatto e da tutti gli individui residenti in provincia di Trento appartenenti a quelle stesse famiglie.
Sono quindi escluse le convivenze (cioè le persone coabitanti per motivi particolari, di natura religiosa, militare, detentiva, o assistenziale, come ad esempio gli istituti di pena, le case di riposo ecc…) e si prescinde inoltre dalla famiglia anagrafica (ossia quella risultante dallo stato famiglia in Comune). Non sono poi da considerarsi unità della popolazione di riferimento le persone che coabitano con la famiglia per un motivo di natura sostanzialmente economica, sia esso di lavoro (è il caso delle badanti o dei domestici) o di locazione (come i subaffittuari) e nemmeno coloro che sono presenti in famiglia nel periodo di rilevazione ma appartengono di fatto ad un'altra (ad esempio i bambini in affido temporaneo). Al contrario, vanno rilevati i componenti della famiglia temporaneamente assenti da casa, come gli studenti fuori sede o i ragazzi che prestano servizio militare volontario. Infine, escono dal campione i soggetti o le famiglie che, nel corso dell'indagine, si spostano in un'altra provincia italiana o estera oppure vengono trasferiti in una qualsiasi istituzione, mentre restano nel campione le famiglie che si trasferiscono in un comune interno alla provincia ed i soggetti che escono definitivamente dalla famiglia restando in provincia di Trento; in questo secondo caso si origina quella che viene definita "famiglia split": se essa è formata anche da altre persone, queste diventano nuovi componenti del campione e vanno pertanto intervistati (PAT, 2008).

Il disegno di campionamento è di tipo stratificato, utilizzando come variabile di stratificazione il comune di residenza. Grazie alla cospicua rete di rilevatori disponibile per quest'indagine, si rende possibile coinvolgere ogni comune della provincia, evitando di ricorrere ad campionamento a due stadi, che comporterebbe una ricaduta negativa in termini di "design effect". In ogni strato le unità vengono disposte in ordine alfabetico di via ed estratte in modo sistematico (provocando un'ulteriore implicita stratificazione). Il tasso di campionamento è fisso in ogni strato (pari a 0.015), ossia il numero di unità estratte per comune è proporzionale alla numerosità demografica dello stesso. La numerosità campionaria si attesta a 7730 unità statistiche, per un totale di 2980 famiglie coinvolte.

5.3.3 Modalità d'intervista e questionario

Le unità statistiche della popolazione vengono intervistate dal rilevatore presso la loro abitazione mediante tecnica CAPI (Computer Assisted Personal Interwieving), ossia, in parole povere, viene effettuata un'intervista faccia a faccia con questionario su pc, anziché cartaceo. Il software utilizzato (IdMonitor per questa specifica indagine) gestisce autonomamente i filtri, annullando così il rischio di mancato rispetto delle norme di compilazione ed è inoltre predisposto per compiere un primo controllo di qualità sostanziale dei dati, segnalando al rilevatore eventuali valori inammissibili o sospetti, in riferimento ad impostati intervalli di accettazione.
Il software, inoltre, origina automaticamente il questionario corretto tra quattro possibili in base all'opzione che il rilevatore sceglie, ossia:

a. Questionario per maggiorenni, somministrato a tutti gli individui che al momento dell'intervista hanno superato la maggiore età.
b. Questionario per minorenni, somministrato a tutti gli individui minorenni. Si compone di poche domande, perlopiù di carattere anagrafico; per gli appartenenti alla fascia 14-17 è previsto qualche quesito aggiuntivo nel caso in cui percepiscano un reddito significativo per la famiglia.
c. Questionario proxy: qualora non sia in alcun modo possibile intervistare un componente maggiorenne della famiglia nell'arco dei tre mesi di rilevazione, è acconsentita l'intervista proxy, a patto che ci sia un famigliare in grado di fornire correttamente le informazioni relative all'assente. Il questionario proxy

è una versione estremamente condensata del questionario per maggiorenni standard (consta, in pratica, dei soli quesiti anagrafici e sul reddito).

d. Questionario famigliare: somministrato ad un qualsiasi componente, purché perfettamente informato sulle notizie in esso richieste, ossia relative alla situazione economica generale della famiglia, all'abitazione e alla zona di residenza.

In caso di questionario per maggiorenni sono possibili due percorsi distinti, a seconda che lo stesso sia già stato somministrato negli anni precedenti oppure no: nel primo caso si origina un percorso di aggiornamento, altrimenti uno completo (PAT, 2008).

Il percorso completo riguarda quindi soggetti (ovviamente maggiorenni) che:

i. Non sono mai stati intervistati, in quanto appartenenti ad una famiglia che è entrata a far parte del campione in quello stesso anno
ii. Non sono mai stati intervistati, perché non facevano ancora parte della famiglia negli anni precedenti, ossia sono componenti nuovi (ad es. i genitori anziani che si trasferiscono nella famiglia del figlio oppure il neosposo della figlia che va ad abitare con i suoceri nei primi anni di matrimonio ecc..)
iii. Non sono mai stati intervistati in quanto nuovi componenti di una famiglia split (ad esempio, se la figlia esce dalla famiglia in cui si trovava negli anni precedenti per andare a convivere con il suo ragazzo, quest'ultimo entra a far parte del campione e viene intervistato con percorso completo)
iv. Non sono mai stati intervistati poiché hanno rifiutato di partecipare all'indagine negli anni precedenti
v. Sono già stati intervistati negli anni precedenti, ma con modalità diversa, ossia con questionario per minorenni, se in quell'occasione non avevano ancora compiuto 18 anni, oppure con quello proxy, se erano assenti per l'intero periodo di rilevazione.

Il percorso completo è costituito da numerose domande che ripercorrono a ritroso tutta la vita dell'intervistato, dalla nascita fino al momento dell'intervista.
Il percorso di aggiornamento viene invece effettuato in tutti gli altri casi, cioè per quei soggetti che sono già stati intervistati con il percorso completo in una precedente occasione e quindi hanno già fornito tutte le informazioni di natura

retrospettiva; si tratta dunque semplicemente di registrare eventuali cambiamenti interventi dall'ultima intervista (in realtà una sezione del questionario, quella relativa al reddito, viene somministrata in toto ogni anno, ma ad ogni modo ciò non toglie che l'impegno richiesto all'intervistato, come pure al rilevatore, divenga nettamente più contenuto rispetto al percorso completo).

Esaminiamo infine in dettaglio le sezioni tematiche di cui si compone il questionario per maggiorenni (PAT, 2008):

a. Scheda generale, che raccoglie informazioni anagrafiche. In caso di aggiornamento chiede conferma dei dati invariabili quali luogo e data di nascita ed annota eventuali modifiche in merito a stato civile e cittadinanza.
b. Sezione mobilità geografica, con la quale si intende ricostruire la "storia residenziale" del rispondente, rilevando tutti i differenti comuni in cui ha abitato e in che periodi. In caso di aggiornamento, si registrano solo eventuali cambi di residenza rispetto a quanto dichiarato nell'ultima intervista.
c. Sezione famiglia: parte con dei quesiti che raccolgono notizie della famiglia di origine del rispondente, in riferimento alla sua nascita e a quando aveva 14 anni; successivamente, rileva informazioni su tutti gli episodi significativi (se accaduti) come l'uscita di casa, la formazione di una nuova famiglia, la convivenza o il matrimonio, la separazione, la nascita o l'adozione di figli ecc.. In caso di aggiornamento ci si limita ad annotare eventuali nuovi eventi di questo tipo.
d. Sezione reddito: rileva i redditi percepiti nell'anno precedente a quello dell'intervista, suddividendoli per fonte (da lavoro dipendente, da lavoro autonomo, da pensioni e altro). Questa sezione, dal momento che non indaga a ritroso ma ha come riferimento un anno ben preciso, viene somministrata allo stesso modo nel percorso completo e in quello di aggiornamento.
e. Sezione istruzione e formazione professionale: ripercorre la "carriera formativa" dell'intervistato, ovvero tutti i corsi scolastici frequentati e tutte le esperienze di formazione professionale, nonché tutti i relativi eventuali attestati. Il percorso di aggiornamento rileva solamente se si è verificata qualche novità in campo scolastico/formativo dall'ultima intervista.
f. Sezione lavoro: ricostruisce la "carriera lavorativa" dell'intervistato, intesa come susseguirsi di episodi di lavoro (diversi se cambia il tipo di lavoro, il

datore, il contratto) e di interruzione (ad esempio per disoccupazione, gravidanza, pensione, eccetera) e raccoglie informazioni in merito. In caso di aggiornamento si verifica se è ancora in corso lo stesso episodio dichiarato nell'ultima intervista o se è cambiato e, se si, come.

Riepilogando, il questionario proxy ed il questionario minorenni altro non sono che versioni molto condensate del questionario per maggiorenni, essendo costituito il primo dalla scheda generale e dalla sezione reddito ed il secondo dalla sola scheda generale. Il questionario famigliare, invece, è a sé stante e va somministrato ogni anno sempre uguale.

5.4. La formazione ai rilevatori per l'indagine "Condizioni di vita delle famiglie trentine"

I rilevatori scelti per partecipare all'indagine "Condizioni di vita delle famiglie trentine" sono stati contattati telefonicamente con circa un mese di anticipo; quelli che hanno accettato l'incarico sono stati invitati alle due giornate di formazione. Data la cospicua numerosità (i rilevatori coinvolti erano quasi 50) sono state organizzate tre tornate con altrettanti gruppi; uno di questi riuniva tutte le persone che non avevano mai effettuato l'indagine negli anni precedenti e per le quali, pertanto, la formazione è stata più densa e minuziosa.
Essendo dunque i partecipanti alla loro prima esperienza, è risultata d'obbligo, prima di tutto, la presentazione dell'indagine; in particolare è stato lasciato largo spazio al chiarimento degli obiettivi e a come questi abbiano determinato e giustificato le scelte operate. Particolare attenzione a questo proposito è stato rivolta alla longitudinalità dell'indagine, la quale, come abbiamo visto nel paragrafo precedente, costituisce una prerogativa irrinunciabile per gli scopi prefissati. E' stata evidenziata più volte l'importanza dell'aver colto e compreso bene questo aspetto, dal momento che molte delle domande e dei dubbi degli intervistati vertono proprio intorno ad esso. Inoltre, è stata colta l'occasione per spiegare ai rilevatori il doppio effetto che un'indagine longitudinale sortisce a livello di rispondenti, "preavvertendoli" così di ciò che si sarebbero dovuti aspettare, vale a dire persone con ogni probabilità collaborative, essendosi dimostrate tali negli anni precedenti e di certo meno diffidenti, sapendo già di cosa si tratta, ma anche presumibilmente stanche e

infastidite, facendo parte del campione già da quattro anni. Dunque è stato ribadito, a maggior ragione, quanta rilevanza assume la bravura del rilevatore nello spiegare in modo professionale perché è importante collaborare continuativamente, così da motivare e stimolare i soggetti a partecipare ancora all'indagine.
Lo step successivo è stata l'identificazione della popolazione di riferimento (vedi §5.2.1), la quale ha sollevato delle perplessità in riferimento al concetto di "famiglia di fatto", per la ragione che può divergere da quello di "famiglia anagrafica" e che, a differenza di quest'ultima, non è ufficialmente definita. Per chiarire il concetto sono stati presi in considerazione, ad uno ad uno, tutti gli esempi dubbi ipotizzati dai presenti; è stato consigliato, per risolverli, di tenere sempre ben presente il tema dell'indagine: infatti, se ad interessare sono le condizioni di vita, si può affermare, come regola generale, che un soggetto è incluso nella famiglia se ha influenza sul bilancio famigliare (ad esempio l'universitario fuori sede, anche se non coabitante, è parte della famiglia perché questa lo mantiene). L'identificazione della popolazione di riferimento è quindi proseguita con la specificazione dei concetti di "famiglia trasferita" e "famiglia split" ed infine sono state illustrate le differenze e le caratteristiche delle varie tipologie di questionario e dei due percorsi (completo e di aggiornamento).
Si è poi entrati nel vivo con l'analisi del questionario (quello che racchiude tutti gli altri, cioè per maggiorenni), non tanto in termini di gestione formale, quanto piuttosto sul piano del significato dei quesiti in esso contenuti: sono state quindi lette da cartaceo tutte le domande possibili, senza badare ai salti imposti dai filtri ed è stata spiegata la corretta interpretazione da attribuire ad ognuna. È stato puntualizzato che, qualora fossero possibili altre interpretazioni, il rilevatore deve attenersi esclusivamente a quella fornita, riferendola identica al rispondente, se questo richiede dei chiarimenti in merito (riportando un esempio: in riferimento ai trasferimenti, viene precisato che non sono da intendersi tali gli spostamenti all'interno dello stesso comune oppure tra comuni diversi dello stesso stato estero). Ad ogni modo, viene ricordato che la spiegazione delle domande è riportata sull'apposito manuale, che il rilevatore è tenuto a consultare in caso di dubbio durante le interviste; in esso sono riportate anche alcune informazioni utili che, integrate con i suggerimenti dati durante la formazione, permettono di valutare le risposte fornite dai rispondenti (ad esempio, che una pensione sociale è nell'ordine massimo dei 600-700euro piuttosto che il servizio civile volontario può essere svolto dalle persone in età compresa tra i 18 ed i 28 anni a partire dal 2001, eccetera).

È stato poi raccomandato di attenersi alla formulazione delle domande e di non tralasciare mai parti delle stesse, perché ciò può portare a degli errori non indifferenti. Sono stati evidenziati via via i quesiti che più di altri richiedono attenzione, ad esempio quello che rileva l'importo mensile per compensi extra (come la tredicesima), il quale comporta che il rilevatore divida la cifra comunicata per i mesi dell'anno oppure la domanda sull'eventuale percezione di "altri redditi", la quale esige vengano lette tutte le modalità (assegni famigliari, affitti di terreni..), altrimenti il rispondente è portato istintivamente a rispondere in negativo, e così via.
Continuando, è stato fatto notare che le domande non hanno tutte gli stessi riferimenti temporali e quindi è estremamente importante sottolinearli quando si formulano le domande, affinché l'intervistato abbia chiaro il momento/periodo cui deve fare riferimento per fornire la risposta (PAT, 2008). Le date costituiscono elemento "critico" soprattutto per via dei numerosi quesiti retrospettivi presenti nel questionario: è infatti compito del rilevatore seguire le vicende ricostruite a ritroso dal rispondente e coglierne la coerenza cronologica, accertando ad esempio che, laddove si rilevano le esperienze come insieme di episodi, la fine di un episodio coincida con l'inizio del seguente oppure che la data di conseguimento della licenza elementare sia compatibile con la data di nascita e così via. È stato consigliato di appuntare su un foglio le date più significative comunicate nel corso dell'intervista, così da agevolare la memoria del rispondente e aiutarlo nella ricostruzione della sua storia di vita, evitando al contempo di influenzarne i ricordi (non si fa altro che ribadirgli le risposte fornite ai quesiti precedenti e lo si aiuta nei calcoli). Annotare le date anziché affidarsi alla memoria è tanto più importante quanto più l'intervistato è anziano, diventando sempre più distanti gli eventi che si gli si chiede di ricordare.
Altro elemento a cui è stato sollecitato di prestare molta attenzione è la codifica: sono infatti presenti nel questionario dei quesiti che non accettano una risposta per esteso ed è importante trovare il giusto codice da attribuirvi. A questo scopo il rilevatore è munito di un fascicoletto con i cartellini, suddivisi per argomento e numerati; è stato spiegato come consultarlo e suggerito che può essere d'aiuto, in alcuni casi, esaminarlo insieme all'intervistato. Sempre in merito ai codici, è stata chiarita la differenza tra codice componente, progressivo che serve per identificare un componente all'interno della sua famiglia e relazione di parentela, che invece indica la relazione di ogni componente con la persona di riferimento della famiglia; a riguardo è stato sottolineato che la persona di riferimento è sempre la stessa e non quella che si sta intervistando (ad esempio, se la persona di riferimento è il marito e

ad essere intervistata è la moglie, ipotizzando di dover indicare la madre di quest'ultima, dev'essere ricercato tra i codici quello relativo a "genitore del coniuge di PR"). Infine è stato ribadito che, qualora si trovino nuovi componenti, questi vanno rilevati e ed il loro codice componente è rispettivamente 76-77-78-79.
Terminata l'analisi del questionario e chiariti tutti i dubbi in merito, la formazione è proseguita con una simulazione d'intervista su pc, condotta a turno dai partecipanti e somministrata alle coordinatrici, le quali hanno inventato delle possibili risposte.
La simulazione è servita in primo luogo per familiarizzare con il programma che gestisce l'intervista, scoprendone modalità d'uso, potenzialità e limiti. Dopo aver impartito le istruzioni per avviarlo, è stata illustrata la schermata con i record, facendo notare che per ogni famiglia esiste un record famigliare (caratterizzato dal nome della persona di riferimento con accanto la sigla FAM), tanti record individuali quanti sono i componenti e quattro record vuoti per eventuali nuovi componenti. È stato sottolineato fin dall'inizio che il questionario famigliare va compilato esclusivamente aprendo il record famigliare, regola che il rilevatore deve tenere sempre ben presente, visto che il programma lascia proseguire anche se si agisce diversamente, non essendo predisposto per comunicare questo tipo di errore.
Nel corso dell'intervista sono state via via impartite tutte le regole e suggeriti tutti gli accorgimenti da adottare per gestire un intervista CAPI: è stato fatto notare che le frasi tra parentesi rappresentano dei suggerimenti per il rilevatore o dei filtri per impostare il giusto percorso e pertanto non sono da leggere all'intervistato; in altri casi costituiscono un chiarimento del quesito e quindi vanno specificati solo nel caso se ne presenti la necessità. E' stato mostrato come correggere una risposta data in precedenza e rivelatasi poi sbagliata, evidenziando però che ciò comporta il dover ripercorrere tutto il questionario dal quesito errato in poi; è stato quindi consigliato, per evitare questo inconveniente, di portare sempre avanti l'intervista con una certa sicurezza, chiedendo piuttosto conferma all'intervistato qualora si sospetti errata una risposta da lui fornita. In tal caso viene raccomandato di comportarsi con un certo tatto, incolpando sé stessi con frasi del tipo "Mi scusi, non so se ho capito bene" ed evitando sempre di attribuire lo sbaglio al rispondente.
Nel corso dell'intervista i partecipanti sono stati sollecitati ad inserire volutamente delle risposte errate, in modo da attivare il sistema di avvisatura del programma: è stato mostrato che rilascia un messaggio e non lascia proseguire in caso di risposta impossibile (ad esempio inserendo 13 nel campo del mese) mentre si limita a suggerire una verifica della risposta mediante richiesta di conferma al rispondente,

qualora la situazione gli appaia anomala (ad esempio 100€ come retribuzione mensile). Sono stati inoltre messi in evidenza i casi in cui il programma non dà la possibilità di continuare con l'intervista non a causa di errori veri e propri, bensì solamente per aspetti di natura formale, risolvibili adottando delle piccole accortezze. Ad esempio, nel quesito che rileva la composizione famigliare alla nascita del rispondente, è presente una lista di possibili componenti; se uno o più componenti non erano presenti, in loro corrispondenza va specificato uno zero, altrimenti viene segnalato un errore. Lo stesso dicasi per quei quesiti che necessitano di codice: se viene segnata una risposta per esteso il programma segnala errore e non consente di andare avanti (ad ogni modo, tra parentesi è indicato che bisogna fare riferimento ad un determinato cartellino).

La simulazione d'intervista non è servita solamente per acquisire pratica con il programma, ma ha costituito anche un proseguo della parte teorica precedente: le situazioni inventate ad hoc dalle coordinatrici, infatti, hanno sollevato numerosi altri dubbi, che hanno portato ad ulteriori spiegazioni circa il significato di parecchie domande.

Terminata la simulazione in gruppo, sono stati distribuiti i portatili a tutti i componenti e si è proseguito lavorando a coppie, con simulazioni di tutti i questionari e percorsi possibili. Le coordinatrici hanno monitorato i gruppi e sono intervenute per dubbi e chiarimenti, così che alla fine è stato loro possibile elencare gli errori più frequenti e chiarire quei concetti o quei quesiti che più hanno causato difficoltà.

5.5. L'assistenza, il monitoraggio ed il controllo dei rilevatori

La prevenzione degli errori dovuti ai rilevatori non si è esaurita con le giornate di istruzione e formazione, ma è continuata nel corso del periodo di rilevazione con l'assistenza ed il monitoraggio del loro lavoro e, al termine dello stesso, con il controllo del loro operato.

A scopo di assistenza è stato attivato un numero verde, cui i rilevatori potevano rivolgersi per qualsiasi problema o dubbio. Questi ultimi, per contro, hanno dovuto rendersi rintracciabili via telefono o via mail nei tre mesi di rilevazione per qualunque comunicazione. E' stato loro dovere, inoltre, impegnarsi ad inviare a fine giornata le interviste portate a termine nel corso della stessa, mediante apposito programma. Questo ha permesso in primo luogo di mettere in salvo il materiale raccolto

nell'eventualità che questo andasse perso (per esempio, in caso di guasto al pc) ed inoltre ha consentito il monitoraggio dell'attività degli intervistatori, sia in termini di quantità che di qualità: sono stati sollecitati i rilevatori che in un certo istante avevano svolto meno interviste della media, chiedendo loro il motivo e assicurandosi che riuscissero a consegnare entro il termine previsto e sono stati contattati coloro il cui materiale, ad un primo controllo di qualità, risultava essere affetto da troppi errori.
Tutto ciò ha contribuito a ridurre l'errore non campionario nell'ambito dell'indagine "Condizioni di vita delle famiglie trentine" mentre, differentemente, il controllo a posteriori ha influito sulla qualità delle indagini future: obiettivo del suddetto è stato infatti di valutare la bravura e la professionalità di ogni rilevatore e quindi di disporre degli elementi per soppesare l'opportunità di ricontattarlo o meno in seguito. Il controllo è consistito, oltre che nell'analisi del materiale fornito e nella conseguente costruzione di indicatori circa rifiuti, proxy, incompatibilità o incongruenze, in telefonate a campione presso le famiglie intervistate, per ottenere informazioni sul comportamento tenuto dal rilevatore, chiedendo ad esempio se si è dimostrato professionale, cortese, chiaro, se portava il cartellino, se aveva preventivamente fissato appuntamento telefonicamente e altre informazioni utili per poter giudicare il suo operato.

5.6. La sensibilizzazione dei rispondenti

Nella fase progettuale di "Condizioni di vita delle famiglie trentine" è stato deciso di non pubblicizzare l'indagine presso tutta la popolazione ma di limitarsi ad operare una sensibilizzazione dei soli soggetti estratti per far parte del campione. Ciò si è concretizzato con l'invio di una lettera di preavviso poco prima che iniziasse l'indagine, stampata su carta intestata della Provincia di Trento e firmata dal dirigente del Servizio Statistica.
La lettera presenta l'indagine ed esorta la famiglia alla collaborazione facendo leva sui seguenti aspetti:

- U*tilità delle precedenti collaborazioni*: la famiglia viene ringraziata per la disponibilità dimostrata in occasione delle precedenti interviste, sottolineando che la collaborazione non è stata vana ma è servita per iniziare a costruire la base

informativa per l'analisi del fenomeno, che costituisce appunto l'obiettivo dell'indagine;

- *Tornaconto collettivo*: viene sottolineato che l'indagine viene svolta per soddisfare l'esigenza di decisori pubblici e ricercatori sociali: ciò lascia intendere che essa riveste un interesse pubblico e quindi può avere un tornaconto (seppur indiretto) per tutti i cittadini;
- I*mportanza della longitudinalità*: viene spiegato che è fondamentale seguire l'evoluzione dei fenomeni nel tempo affinché sia possibile spiegarne le cause e quindi che è importante partecipare con continuità;
- *Attenuazione del disturbo*: viene ricordato che l'intervista è più breve rispetto alla prima effettuata, consistendo semplicemente in un aggiornamento e quindi che il disturbo diviene molto più contenuto;
- *Casualità dell'estrazione*: viene ribadito che il nominativo era stato inizialmente estratto dalle anagrafi comunali con procedure del tutto casuali;
- *Segreto statistico*: viene rammentato il significato di segreto statistico, rassicurando il rispondente circa il fatto che i dati da lui forniti non possono essere pubblicati se non in forma aggregata; viene inoltre data informativa rispetto al titolare ed il responsabile del trattamento e precisato che l'indagine fa parte del programma statistico provinciale e del programma statistico nazionale;
- *Obbligo di risposta*: viene chiarito che l'obbligo di riposta è previsto poiché l'indagine è inclusa in un decreto del Presidente della Repubblica e viene fornito il riferimento di tale decreto;
- *Numero verde*: viene messo a disposizione un numero verde a cui il rispondente può rivolgersi in caso di dubbi o delucidazioni in merito all'indagine.

6. LE AZIONI DI CONTROLLO E CORREZIONE DEI DATI IN FASE DI REVISIONE QUALITATIVA: IL SOFTWARE CONCORD

Concord è un programma per il controllo e la correzione dei dati, sviluppato dall'Istat a partire dalla metà degli anni '80; si tratta di un software generalizzato, ovvero applicabile a qualsiasi tipo di indagine (Istat, 1999). Comprende tre diversi moduli: SCIA (Sistema Controllo ed Imputazione Automatici), basato su un approccio di tipo probabilistico, che individua e corregge gli errori, GRANADA (Gestione delle Regole per l'ANAlisi dei Dati), basato su un approccio deterministico, che effettua anch'esso entrambe le operazioni e RIDA (Ricostruzione Informazioni con Donazione Automatica), che corregge soltanto, tramite imputazione da donatore. I moduli sono integrabili tra loro, rendendo possibile l'effettuazione di un percorso che, a partire dagli stessi dati di imput, utilizza tutti e tre gli approcci, sfruttando così i punti di forza di ognuno e trovando il metodo di correzione più adatto ai diversi errori riscontrati. La struttura del percorso è gerarchica, ovvero si parte del modulo probabilistico per passare al deterministico ed infine al donatore. (Istat, 2004)

6.1. Il modello probabilistico - SCIA

Il software SCIA permette di attuare una revisione qualitativa di tipo probabilistico, che si basa sulla teoria sviluppata da due studiosi americani, Fellegi e Holt, nel 1976. SCIA tratta solo variabili qualitative, anche se codificate; per le altre bisogna ricorrere a GRANADA. L'utente deve inserire le variabili che intende correggere, definendone nome, posizione nel file e lunghezza; inoltre va indicato, per ogni variabile, il dominio che essa ammette, così che siano automaticamente determinati i valori fuori campo. Dopodiché vanno esplicitate le regole che costituiscono il piano di compatibilità. Esse vengono chiamate, con la terminologia dei due autori, edit in forma normale e descrivono delle condizioni che, se verificate contemporaneamente (vedi gli and tra parentesi), danno luogo ad una situazione d'errore (Istat, 2004); ogni condizione, a sua volta, può essere definita in termini alternativi (vedi gli or all'interno delle parentesi). Un edit in forma normale è dunque così strutturato (Cd=condizione):

SE [Cd1(Cd_A..OR..Cd_K)…AND…CdY(Cd_A..OR..Cd_K)...AND...CdN(Cd_A..OR..Cd_K)]

ALLORA [incompatibilità].

Naturalmente, questa è una formalizzazione generale che comprende i casi più ampi e complessi; in realtà molte regole si risolvono in costrutti "if-then" più semplici, con una o due condizioni soltanto, ad esempio:

SE [Cd1(maggiorenne=no) AND Cd2(patente=si)] **ALLORA** [incompatibilità]

Le regole definite dall'esperto costituiscono l'insieme iniziale degli edit, il quale viene "aggiustato" dal software attraverso l'eliminazione degli edit contradditori e degli edit ridondanti. Si perviene così al cosiddetto insieme minimale degli edit, che permette l'individuazione dei record errati (per i quali cioè si verificano una o più situazioni d'errore) ma non delle variabili da sottoporre ad imputazione (Istat, 2004). Se cioè, ad esempio, un record attivasse la regola riportata sopra, non sarebbe possibile capire se la variabile da cambiare è "maggiorenne" oppure "patente". A questo scopo è necessario ricavare gli edit impliciti in quelli espliciti, operazione di cui si occupa il software dietro apposito comando, e giungere così all'insieme completo degli edit. La logica è molto semplice; ricorriamo ad un esempio per esporla. Supponiamo che per la variabile maggiorenne siano possibili le modalità si, no; per "condizione professionale" (=condprof) le modalità occupato, in cerca di occupazione, ritirato dal lavoro, studente, altro e per la variabile "Fonte di reddito" (=fontered) le modalità reddito da lavoro, pensione, mantenimento da parte dei familiari, altro. Supponiamo altresì di aver definito le due regole seguenti:

- **SE** [(maggiorenne=no) AND (condprof≠studente)] **ALLORA** [incomp]
- **SE** [(condprof≠occupato) AND (fontered=reddito da lavoro)] **ALLORA** [incomp]

Se le mandassimo in esecuzione, ci verrebbe segnalato questo record:

[maggiorenne=no, condprof=occupato, fontered=reddito da lavoro]

E' importante notare che sarebbe solo la prima regola ad essere attivata, quindi ci limiteremmo a guardare le variabili in essa richiamate, cioè "maggiorenne" e "condizione professionale", e non potremmo capire quali delle due è errata. Se decidessimo di modificare la condizione professionale, verrebbe attivato l'altro edit, perché un non occupato non può avere come fonte di reddito un reddito da lavoro. Modificheremmo allora anche questa seconda variabile, effettuando così un totale di due imputazioni per riportare il record ad una situazione di correttezza.

La situazione cambia se deriviamo la regola implicita nelle due precedenti:

- **SE** [(maggiorenne=no)] AND (fontered=reddito da lavoro) **ALLORA** [incomp]

Lo stesso record di prima attiva ora sia il primo edit che quest'ultimo; diventa dunque chiaro che la variabile da modificare è "maggiorenne", visto che determina due incompatibilità, e il record viene così corretto con una sola modifica.
L'insieme completo degli edit, ossia gli espliciti più gli impliciti, assicura non solo di correggere i record errati, ma di farlo in modo ottimale, ossia senza introdurre ulteriori errori, minimizzando il numero di variabili modificate e mantenendo il più possibile inalterate le distribuzioni originarie delle variabili in questione. La logica alla base di SCIA è che tenta di costruire l'insieme minimale di variabili da modificare prendendo quelle che più probabilmente sono errate; se non ci sono dubbi su quali siano queste variabili, è sbagliato affidare la loro localizzazione al software, che la esegue appunto in questo modo probabilistico, perché ciò potrebbe portare a delle distorsioni. E' invece opportuno in tal caso avvalersi del modulo deterministico, nel quale viene richiesto esplicitamente di chiarire quali variabili sono errate, indi da sottoporre ad imputazione. Inoltre, gli altri errori che è meglio trattare con GRANADA si possono individuare mandando in esecuzione il solo insieme minimale degli edit: vengono infatti prodotte delle statistiche sulle regole attivate che, debitamente analizzate, possono segnalare l'esistenza di eventuali errori sistematici, cioè non stocastici, che è bene sottoporre a passo deterministico. La presenza di errori sistematici si può dedurre dalla ripetuta attivazione di determinate regole, dall'alta frequenza di valori anomali (outlier) e da numerosi errori concentrati su determinate variabili, situazioni che lasciano intuire una motivazione di natura non casuale.
Per tutti gli altri errori si procede invece con l'approccio probabilistico. Dopo aver ricavato l'insieme completo degli edit ed aver localizzato l'insieme minimo delle variabili da modificare per record, viene effettuata l'imputazione di queste variabili. Essa può avvenire per ogni record nei seguenti modi, procedendo per tentativi (Istat, 2004):

i. imputazione congiunta ristretta: considerando le variabili di accoppiamento (ossia non appartenenti all'insieme minimale da modificare), si sceglie tra i donatori quello che in loro corrispondenza assume gli stessi valori assunti dal record errato; se questo donatore esiste, le modalità che esso presenta per le variabili dell'insieme minimale vengono imputate in blocco alle corrispondenti nel record errato.
ii. imputazione congiunta allargata: sempre considerando le variabili di accoppiamento, si sceglie tra i donatori quello che in loro corrispondenza

assume valori appartenenti allo stesso range cui appartengono i corrispondenti valori nel record errato; se questo donatore esiste, le modalità che esso presenta per le variabili dell'insieme minimale vengono imputate in blocco alle corrispondenti nel record errato.

iii. Imputazione sequenziale: se non esistono record in grado di donare in blocco le variabili necessarie, si procede considerando una variabile alla volta, ricercando per ognuna un donatore che abbia un valore ammissibile per il record ricevente.

6.2. Il modello deterministico – GRANADA

GRANADA richiede che sia l'utente a specificare le variabili errate per ogni incompatibilità definita. Pertanto, le regole su cui si basa sono del tipo:
SE [Cd1(Cd_A..OR..Cd_K)…AND…CdY(Cd_A..OR..Cd_K)...AND...CdN(Cd_A..OR..Cd_K)]
ALLORA [variabile errata=x].
In sostanza, la prima parte dell'edit è identica a quella formulata per il modulo probabilistico (tant'è che esiste la possibilità di importare le condizioni di errore da SCIA, senza doverle riscrivere) mentre la seconda parte non si limita come per SCIA a segnalare un'incompatbilità, ma precisa anche quale variabile, tra quelle menzionate nelle condizioni, ne è responsabile.
Basandosi su queste premesse, esistono poi due modi differenti di procedere:

a. Imputazione da valore prefissato: se non ci sono dubbi circa il valore che la variabile errata deve assumere affinché il record sia corretto (ossia se questo valore è unico ed inequivocabile), è possibile effettuare l'imputazione contemporaneamente alla localizzazione, esplicitando il valore puntuale da sostituire a quello sbagliato. In altri termini la regola viene completata in maniera deterministica e si presenta così:
   SE [Cd1(Cd_A..OR..Cd_K)…AND…CdY(Cd_A..OR..Cd_K)...AND...CdN(Cd_A..OR..Cd_K)]
   ALLORA [a variabile errata=x imputa valore=y].
   È bene evitare, o quantomeno limitare, il ricorso a questa prima modalità, in quanto può comportare pesanti distorsioni nelle distribuzioni originarie dei dati.
b. Impostazione dei caratteri d'errore: nel caso in cui, una volta individuata la variabile errata, permangano dei dubbi circa il valore da imputare, ovvero se

esiste più di un valore in grado di riportare il record ad una situazione di correttezza, è opportuno demandare a RIDA la correzione dei dati e circoscrivere l'uso di GRANADA alla sola localizzazione delle variabili errate. Ci si limita in sostanza a segnalare con dei caratteri definiti queste variabili, in modo che sia poi possibile riprenderle in RIDA e sottoporle ad imputazione; per fare ciò basta esplicitare nella seconda parte dell'edit il carattere da attribuire ad ogni variabile errata, ripetendolo in numero pari alla lunghezza della variabile stessa.

In questo secondo caso la regola è dunque del tipo:

SE [Cd1(Cd_A..OR..Cd_K)…AND…CdY(Cd_A..OR..Cd_K)...AND...CdN(Cd_A..OR..Cd_K)]

ALLORA [a variabile errata x imposta caratteri d'errore=$$$$].

6.3. L'imputazione da donatore - RIDA

RIDA non individua gli errori, ma si occupa solamente dell'imputazione: per questo è necessario avvalersi del file di output di GRANADA, nel quale le variabili errate sono segnalate dai caratteri definiti nello step precedente. L'imputazione delle suddette variabili avviene tramite donatore, scegliendo tra le unità non affette da errore quella più simile all'unità ricevente, attraverso il metodo della distanza minima (Istat, 2004).

Si rende opportuno innanzitutto chiarire il concetto di "distanza" (d) tra due unità in riferimento ad una variabile, differenziando in funzione della tipologia della stessa:

a) Variabile qualitativa
   d=0 se le unità presentano la stessa modalità,
   d=1 se presentano modalità diverse
b) Variabile ordinale
   d=0 se le unità presentano la stessa unità,
   d=1 se presentano modalità adiacenti,
   d=2 se tra le due modalità ce n'è una in mezzo e così via fino ad arrivare a d=m-1, (con m=numero di modalità) se le modalità presentate dalle due unità sono agli estremi opposti
   Per imporre d variabile tra 0 e 1, dividiamo per il suo massimo m-1, ossia
   d'= d/(m-1)

c) Variabile quantitativa

   d= |distanza tra le due modalità|

   Per imporre d variabile tra 0 e 1, dividiamo per il suo massimo, cioè la differenza tra il valore più grande riscontrato per quella variabile nel file ed il valore più piccolo.

   d'= |distanza tra le due modalità|/differenza massima

Calcolando la distanza tra due unità in riferimento a più variabili, è possibile pervenire alla distanza totale D che le separa: basta sommare le singole distanze, ponderandole con dei pesi.

$D=\Sigma_i W_i d_i$

con Wì numeri naturali che indicano l'importanza attribuita ad ogni variabile per la quale si è calcolata la distanza singola.

Non si utilizzano ovviamente tutte le variabili disponibili, bensì solo quelle che si reputano significative per l'indagine, tra quelle non affette da errore; tali variabili vengono dette "di matching".

È inoltre possibile stabilire delle variabili "di strato", le quali servono a limitare il calcolo delle distanze ad un sottoinsieme di donatori, ossia quelli che presentano le stesse modalità dell'unità ricevente in corrispondenza di tali variabili (Istat, 2004).

Infine, esistono dei parametri opzionali per "personalizzare" il processo di individuazione del donatore più vicino:

a) Parametro U: fattore di penalizzazione per donatore già utilizzato, il quale fa in modo che, a parità di distanza, si privilegi sempre l'unità che ha "donato" meno volte
b) Parametro R: numero massimo di volte in cui uno stesso donatore può essere utilizzato.
c) Parametro L: massima distanza accettata tra record ricevente e record donatore.
d) Parametro D: minima distanza accettata tra record ricevente e record donatore.

I suddetti parametri, come del resto le variabili di strato, permettono di affinare la ricerca del donatore, ma introducono nel contempo il rischio di non trovarlo.

L’imputazione, se il donatore esiste, avviene in maniera molto semplice: le varabili errate vengono corrette con le modalità che il donatore assume in loro corrispondenza, che vanno in sostanza a sostituire i caratteri d’errore impostati con GRANADA.

7. OSSERVAZIONI

L'esperienza presso il Servizio Statistico della Provincia è stata decisamente costruttiva, prima di tutto perché mi ha dato l'opportunità di vivere in prima persona la realtà di un ufficio statistico, quindi di cogliere nel concreto alcuni aspetti di cui prima ero al corrente ad un livello meramente teorico: mi riferisco alla divisione tra le varie aree di lavoro, l'interconnessione esistente tra queste, la rilevanza assunta da tempi e scadenze, l'impegno in termini di risorse umane, organizzazione e mole di lavoro che sta dietro ad ogni rilevazione, la necessarietà di collaborazioni con enti o soggetti esterni al Servizio, con le inevitabili conseguenze che quest'esigenza implica.
L'esperienza come rilevatrice, soprattutto, si è rivelata molto interessante, poiché mi ha dato modo di valutare il peso assunto da questa figura all'interno del processo di indagine con una cognizione di causa che prima non potevo possedere. Ho avuto modo di riflettere in particolare su quattro aspetti:

i. l'inevitabilità degli errori non campionari imputabili alla fase di rilevazione: nonostante tutti gli sforzi che si possono compiere in termini di prevenzione, è impossibile disporre per la fase di elaborazione di materiale completamente esente da errori derivanti dalla fase di rilevazione, a meno che non si tratti di un'indagine su scala estremamente ridotta. Ovviamente l'annullamento dell'errore non campionario deve pur sempre rappresentare l'obiettivo ideale cui tendere; a mio avviso, ad ogni modo, è sicuramente più corretto in riferimento allo stesso parlare di "minimizzazione".
ii. l'importanza assunta dalla formazione: una formazione pianificata e realizzata in maniera accurata e quanto più possibile esauriente rappresenta senza dubbio il primo passo verso la riduzione dei potenziali errori commettibili dai rilevatori. Nel mio caso specifico, essendo l'indagine sulle famiglie trentine al terzo anno di rilevazione, la formazione è andata certamente migliorando col tempo, aumentando di anno in anno la consapevolezza di quali siano le lacune più sentite, le sviste più commesse, le difficoltà incontrate e quindi dando la possibilità di impreziosire le giornate di formazione con esempi mirati relativi alle precedenti edizioni.
iii. l'importanza dell'assistenza e del monitoraggio: è fondamentale che i rilevatori siano seguiti durante tutto il periodo di rilevazione, dato che è

impossibile prendere in esame nelle giornate di formazione tutte le situazioni ed i casi che si riscontrano poi nel corso dell'attività sul campo, né per la verità prevederli.

iv. l'impagabile beneficio dato dall'ausilio del pc: essendo il questionario dell'indagine sulle famiglie trentine piuttosto lungo, ma soprattutto ricco di quesiti filtro, risulterebbe pressoché ingestibile con tecnica PAPI, o quantomeno tale tecnica comporterebbe un considerevole aumento sia dei tempi, sia degli errori dati dal mancato rispetto delle norme di compilazione. Le tecniche "Computer Asssited" costituiscono dunque sicuramente un prezioso contributo in termini di qualità del materiale raccolto.

BIBLIOGRAFIA

- ISTAT (1989) Manuale di tecniche d'indagine. Volume 6: Il sistema di controllo della qualità dei dati, Note e relazioni, n.1, Roma.
- ISTAT (1989) Manuale di tecniche d'indagine. Volume 4: Tecniche di campionamento - Teoria e pratica, Note e relazioni, n.1, Roma
- PROVINCIA AUTONOMA DI TRENTO (2000), Intervistare – Come e perché, a cura di G. Grandi, Trento
- PROVINCIA AUTONOMA DI TRENTO (2008), Indagine sulle famiglie trentine – Guida per i rilevatori, a cura di G. Grandi, Trento
- ISTAT (1999), Metodi e software per il controllo e la correzione dei dati, a cura di G. Barcaroli, L. D'Aurizio, O. Luzi, A. Manzari, A. Pallara, Documenti Istat, Roma

SITOGRAFIA

- ISTAT (2000), Linee guida metodologiche per rilevazioni statistiche, nozioni metodologiche di base e pratiche consigliate per rilevazioni statistiche dirette o basate su fonti amministrative, di M. Fortini (www.istat.it/strumenti/metodi/lineeguida)
- COMMISSIONE PER LA GARANZIA DELL'INFORMAZIONE STATISTICA (2000), Analisi delle procedure di correzione e imputazione utilizzate dall'Istat nelle principali indagini sulle famiglie – Volume I, Rapporto di ricerca, 00.02, a cura di L. Fabbris, M. E. Graziani, C. Panattoni (http://www.palazzochigi.it/Presidenza/statistica/attivita/rapporti/2000/00.02)
- ISTAT (2004) Concord V. 1.0 - Controllo e correzione dei dati: manuale utente e aspetti metodologici, Tecniche e strumenti, n.1, Roma (www.istat.it/strumenti/metodi/software/individuazione_trattamento_errori/concord)