L'influenza dei rilevatori sulla qualità dei dati nell'indagine "Servizi resi dalle Pubbliche Amministrazioni e grado di soddisfazione dei cittadini"

Paolo Righi[(*)]

(*) ISTAT – Servizio Studi Metodologici

Sommario

La documentazione ed il miglioramento della qualità dei dati di una indagine statistica rappresenta uno dei principali obiettivi dell'ISTAT. A tale riguardo questo lavoro descrive uno studio degli errori non campionari prodotti nella fase di rilevazione, ponendo particolare attenzione all'attività dell'intervistatore, nell'indagine statistica "Servizi resi dalle Pubbliche Amministrazioni e grado di soddisfazione dei cittadini". Per valutare il livello degli errori non campionari che affliggono l'indagine, sono stati presi in considerazione 15 indicatori misurati su ciascun rilevatore riguardanti l'esito del contatto e cooperazione dell'unità selezionata e la corretta compilazione dei questionari. In una seconda fase, le variabili sono state combinate in un insieme di 6 fattori incorrelati usando la tecnica dell'Analisi in Componenti Principali. Nella terza fase si è verificato il grado di associazione tra i 6 fattori estratti ed alcune caratteristiche dei rilevatori (età, sesso, titolo di studio, occupazione, precedenti esperienze come intervistatore, collocazione geografica e dimensione demografica del comune in cui ha operato).

Abstract

Documenting and improving data quality is becoming one of the main goals of National Statistical Institute. This paper describes a study about non – sampling errors produced in the field activity, with particular attention to the interviewer's activity, in the survey "Italian citizen degree of the satisfaction on the public services". To evaluate the level of non – sampling errors affecting the survey, 15 indicators were chosen for each interviewer regarding both the result of contacting and gaining selected statistic unit's co-operation and the success of respondents in filling in the questionnaires properly. In a second step, the indicators were combined in a set of six factors uncorrelated, by using Principal Component Analysis technique. In the third step it was verified the association between six factors previously identified and some of the interviewer characteristics (age, gender, education, type of employment, previous experience as interviewer, geographical area and number of inhabitants in the municipality where the interviewer operated).

1. Introduzione [(*)]

Durante l'esecuzione di una indagine statistica, tramite questionario, si possono commettere degli *errori* che producono delle differenze tra l'informazione raccolta e il reale valore del fenomeno da osservare, generando degli scostamenti tra le stime dell'indagine ed i parametri da stimare (Hansen et al. 1951).

Gli errori sono generalmente divisi in due categorie: gli *errori campionari* e gli *errori non campionari*. La prima classe comprende errori causati dalla variabilità del fenomeno, dal disegno campionario e dalle procedure di stima, mentre quelli non campionari sono definiti dal complesso dei restanti errori che hanno origine nelle fasi operative dell'indagine.

Esistono varie classificazioni per descrivere più dettagliatamente gli errori di una indagine statistica[1]. Ad esempio Lessler e Kalsbeek (1992), suggeriscono di ordinare l'errore in relazione alla fase d'indagine in cui esso si presenta e, al riguardo, individuano quattro momenti: la costruzione della lista che individua univocamente le unità statistiche della popolazione da campionare (*sampling frame*); la definizione del disegno campionario; l'individuazione delle unità del campione e la sollecitazione a partecipare all'indagine; la raccolta dei dati e la trasposizione su supporto informatico.

Ad ognuna delle quattro fasi citate è associato un tipo di errore e solo nella seconda sono situati gli errori campionari.

Nella prima fase sono individuati gli *errori del tipo di lista* (*frame errors*) anche detti di *errori di copertura* (R. Groves, 1989). Tali errori emergono, ad esempio, quando la lista delle unità statistiche, dalla quale si seleziona il campione, non considera tutto il collettivo di interesse o, viceversa contiene una larga parte di unità non appartenenti al collettivo. Tale evenienza influenza, non solo la correttezza delle stime, ma anche la scelta del disegno campionario da adottare a scapito dell'efficienza delle stime stesse.

[(*)] Desidero ringraziare il dott. Marco Fortini per le osservazioni ed i consigli dati nella stesura del lavoro fermo restando che le opinioni espresse sono di mia responsabilità.

[1] La letteratura ha proposto varie classificazioni per descrivere l'insieme degli errori si ricordano: Zarkovich, 1966; Kish, 1965; Deming, 1960.

Nella seconda fase, come si è detto, si originano gli errori campionari mentre nella terza avvengono gli *errori di non risposta* (Kendall e Buckland, 1960; Kish, 1965; Cochran, 1977, ecc.) che si realizzano quando non si ottengono informazioni riguardo alcune unità del campione selezionato (le unità sono non rispondenti) oppure quando le unità cooperano all'indagine ma non rispondono ad alcune domande (Kalton, 1983; Madow et alt., 1983).

Nella quarta fase avvengono, infine, gli *errori di misura* rappresentati dalla differenza tra il valore della risposta dell'unità e il vero valore che l'unità effettivamente possiede (Hansen et alt., 1951; Sukhatme e Sukhatme, 1970). Quest'ultimo tipo di errore si può presentare sia al momento della raccolta del dato, attraverso l'errata risposta alla domanda del questionario, volontaria o involontaria, sia, ad esempio, al momento della digitazione del questionario per il trasferimento su supporto informatico.

L'errore di rilevazione incide sull'accuratezza dei dati in termini di distorsione e variabilità delle stime (qualità dei dati) e risulta, dunque, rilevante conoscerne le dimensioni e le cause che lo hanno generato. Le misure dirette per valutare l'entità degli errori sono normalmente costose perché occorre ripetere le fasi dell'indagine su un campione dell'indagine principale. In alternativa si può preferire osservare eventi connessi agli aspetti delle qualità dell'informazione statistica prodotta dall'indagine, ad esempio, conteggiando gli errori secondo la loro tipologia su ciascun questionario.

L'obiettivo del presente lavoro è di valutare la qualità dei dati conseguita in una indagine dell'ISTAT in relazione all'operato della rete di rilevazione e più specificatamente a quello degli intervistatori. A questo scopo il lavoro si concentra sugli errori di tipo non campionario ed in particolare sugli errori di non risposta di cui si possono ottenere direttamente delle informazioni esaminando i questionari dell'indagine. Ben più laboriosa sarebbe stata, invece, l'identificazione degli altri tipi di errori come quelli di misura, per i quali sarebbe stata necessaria la ripetizione dell'intervista.

Nel paragrafo successivo sono introdotti con maggior dettaglio gli errori di mancata risposta e approfonditi gli aspetti legati alle loro cause, ponendo l'attenzione a quelle in cui interviene attivamente il rilevatore. Nel paragrafo 3 sono definiti gli indicatori di qualità riferiti a ciascun rilevatore. Ogni indicatore aggrega un serie di informazioni elementari in un'unica misura: ad esempio, dalla variabile indicatrice "mancata intervista" (si - no) riferita al singolo rilevatore per ogni questionario si può definire un indicatore di qualità calcolando il

"tasso di interviste realizzate" dal rilevatore. Nel paragrafo 4 viene descritto il metodo dell'Analisi delle Componenti Principali (ACP). Questa tecnica di analisi dei dati è utilizzata sia per studiare la struttura di correlazione degli indicatori di qualità sia per ottenere un numero ridotto di indicatori sintesi delle variabili di partenza. Nel quinto paragrafo sono esposti i risultati dell'ACP e sono evidenziati i legami tra le caratteristiche dei rilevatori e gli indicatori di sintesi precedentemente identificati. Nel sesto paragrafo sono esposte le conclusioni del lavoro.

2. L'errore di non risposta

L'errore di non risposta o mancata risposta è descritto, in letteratura, secondo numerose e differenti classificazioni[2], ma la maggior parte di esse concordano nel distinguere tale errore secondo due categorie: gli errori di non risposta totale (*unit non response*) generati quando non si effettua l'intervista; gli errori di non risposta parziale (*item non response*), determinati dalla mancata compilazione di alcuni quesiti del questionario ai quali è obbligatorio rispondere.

A questo secondo tipo di errore, sono assimilabili le infrazioni alle norme formali di compilazioni del questionario, in altre parole quelle regole che pongono dei vincoli di compatibilità fra alcune risposte del questionario. Una norma formale di compilazione si presenta attraverso una domanda così detta "filtro", dalla cui risposta dipende la compilazione di uno o più quesiti successivi, legati logicamente ad essa. Si compie, dunque, una infrazione qualora, nonostante la risposta ad una domanda filtro escluda la considerazione di un particolare e successivo quesito, quest'ultimo viene, invece, compilato. Allo stesso tempo l'errore si commette anche non rispondendo ad un quesito legato logicamente ad una domanda filtro sebbene la risposta a quest'ultima ne richieda la compilazione.

Gli errori di non risposta possono avere varie origini. Alcune riguardano gli aspetti relativi alla definizione ed agli strumenti di indagine (il questionario, modalità di contatto dell'intervistato, ecc.), altre dipendono dall'organizzazione centrale dell'indagine, altre ancora

[2] Per avere una bibliografia dettagliata sull'argomento si può vedere il volume di Lessler e Kalsbeek (1992).

dalle organizzazione periferiche, che hanno il compito di istruire e coordinare i rilevatori (ad esempio, i Comuni entrati nell'indagine), nonché dagli stessi rilevatori. Infine, alcune fonti di errore sono dovute dalle caratteristiche e dalla propensione a rispondere dell'intervistato.

Il numero delle mancate risposte totali, sono, ad esempio, influenzate: dalla modalità di contatto dell'intervistato (contatto diretto, telefonico o postale); dalla disponibilità economica e di tempo per condurre l'indagine; dalla precisione dell'indirizzo dell'unità statistica ricavato dalla lista da cui avviene la selezione del campione[3]; dalla esperienza e competenza del rilevatore e del Comune; dalla facilità di raggiungimento della persona da intervistare; dalla presenza e disponibilità a rispondere della persona da intervistare.

La quantità di mancate risposte parziali di una indagine possono, invece, dipendere: dalla modalità di somministrazione delle domande, che può avvenire, ad esempio, tramite il rilevatore oppure attraverso l'autocompilazione del questionario; dalla chiara e corretta definizione delle norme di compilazione; dal vocabolario utilizzato nelle domande; dal numero di domande che richiedono l'uso della memoria; dalla preparazione del rilevatore; dal grado di comprensione della persona intervistata (dovuta all'età, al livello di istruzione, ecc.).

In generale, ogni causa di mancata risposta caratterizza la variabilità del numero di errori o degli indicatori di qualità ad essi associati misurati su ogni intervistatore. Così, ad esempio, gli errori originati dagli organismi periferici presentano una certa variabilità territoriale, mentre gli errori connessi al questionario risultano distribuiti con maggiore uniformità nello spazio.

Appare, pertanto, interessante, per l’obiettivo del lavoro, approfondire l’analisi della variabilità e della correlazione tra gli indicatori di qualità registrati su ogni rilevatore al fine di esprimere delle ipotesi sulle fonti di errore.

3. Definizione di alcuni indicatori di qualità dell'indagine "Servizi resi dalle Pubbliche Amministrazioni e grado di soddisfazione dei cittadini" (anno 1994)

[3] Quando l'intervista non si compie per l'indirizzo sbagliato si presenta un errore di lista.

Il presente lavoro analizza la qualità dell'informazione statistica prodotta dall'indagine dell'ISTAT: "Servizi resi dalle Pubbliche Amministrazioni e grado di soddisfazione dei cittadini" (anno 1994). Il collettivo di riferimento di tale indagine, è rappresentato dalla popolazione italiana maggiorenne e le unità statistiche selezionate provengono dall'insieme dei maggiorenni rispondenti selezionati nel campione dell'indagine statistica Multiscopo sulle famiglie dell'ISTAT (anno 1994). Quest'ultima adotta un disegno campionario stratificato a due stadi nel primo dei quali si seleziona un campione di Comuni e nel secondo un campione di famiglie nei Comuni che rientrano nella rilevazione. L'indagine ha impiegato 509 rilevatori coordinati dai Comuni selezionati ed è stata effettuata mediante questionario che si articola in due sezioni principali: la prima contiene i quesiti per l'intervistato[4], la seconda è, invece, riservata al rilevatore, il quale deve riportare le informazioni relative alle modalità di compilazione del questionario ed eventualmente al motivo che ha indotto l'intervistato a non rispondere. Il compito dei rilevatori nell'indagine è costituito dal contatto con le famiglie, selezione dei rispondenti maggiorenni, consegna e ritiro del questionario lasciando all'intervistato l'onere della compilazione. Il rilevatore è, comunque, tenuto a prestare aiuto nella compilazione qualora fosse richiesto. Poiché l'interesse del lavoro è focalizzato sull'operato del rilevatore i questionari compilati considerati sono quelli in cui è stato necessario l'intervento del rilevatore a supporto della compilazione[5].

Nel dettaglio, il questionario si compone di diversi tipi di quesiti.

Un primo tipo di quesito è corredato da una serie di risposte codificate delle quali ne deve essere scelta una sola qualora la domanda richiede obbligatoriamente una risposta (figura 1). Il secondo tipo di domanda contiene una serie di sottoquesiti, inerenti ad alcuni aspetti del quesito principale (figura 2), ai quali si deve rispondere come nel caso del primo tipo di domande. Il terzo tipo di quesito prevede due opzioni di risposta di cui una deve esprimersi con un valore numerico (figura 3). L'ultimo tipo di quesito è strutturato secondo una raccolta di sottoquesiti uguali a quelli descritti nel terzo caso (figura 4).

Ai fini della definizione degli indicatori di qualità, è utile anche distinguere le domande in obbligatorie o "per tutti" e le domande che possono essere scartate (in funzione delle

[4] In questa sezione il questionario si divide in nove parti inerenti agli uffici anagrafici, uffici delle unità sanitarie locali, uffici postali, altri uffici pubblici, agenzie private, autocertificazioni, medico di famiglia, servizi ospedalieri, notizie sulla compilazione del questionario.

[5] Il numero totale di questi questionari è pari a 6.009.

risposte che sono state date a domande precedenti) per le quali non si deve dare, in ogni caso, una risposta.

Figura 1

"Uffici anagrafici (comunali e circoscrizionali)" (sezione 1).

1.1 Normalmente come viene a conoscenza degli orari e delle procedure necessarie per richiedere certificati, documenti o altri sevizi anagrafici?

Recandosi di persona presso gli uffici....1 □
Informandosi per telefono.....2 □
Attraverso conoscenti già esperti.....….....3 □
In altro modo.....….....4 □

Figura 2

"Uffici anagrafici (comunali e circoscrizionali)" (sezione 1).

1.9 Quali dei seguenti servizi ha richiesto?
(una risposta per riga)

	NO	SI
Rilascio o rinnovo di documenti.....	1 □	2 □
Rilascio di certificati anagrafici.....	3 □	4 □
Autentica di firme o documenti.....	5 □	6 □
Altri servizi.....	7 □	8 □

Nell'insieme dei quesiti, su cui si articola il questionario, si definisce, infine, domanda "filtro", quella che, tra le opzioni possibili di risposta, ne presenta una incompatibile con la formulazione di alcune domande successive le quali, nel caso che questa risposta venga scelta, devono essere scartate (figura 5).

Figura 3

"Uffici anagrafici (comunali e circoscrizionali)" (sezione 1).

1.4 Negli ultimi 12 mesi si è recato almeno una volta presso gli uffici anagrafici per richiedere certificati, documenti o altri servizi per se stesso, per un suo familiare o per altre persone?

NO.....00 □
SI, numero di volte.....N. □ □

Figura 4

Primo quesito della sezione "Altri uffici pubblici " (sezione 4).

4.1 Negli ultimi 12 mesi si è recato almeno una volta presso gli uffici sotto indicati, per richiedere un servizio per se stesso, per un suo familiare o per altre persone?
(una risposta per riga)

	NO	SI, numero di volte
Uffici delle imposte	00 □	N. □ □
Uffici del catasto.....	00 □	N. □ □
Uffici della previdenza sociale.....	00 □	N. □ □
Uffici di collocamento.....	00 □	N. □ □
Uffici della motorizzazione civile.....	00 □	N. □ □
Uffici del pubblico registro automobilistico.....	00 □	N. □ □

Le domande filtro sono generalmente accompagnate da un avviso sulla regola di compilazione del questionario che avverte se è necessario scartare o meno un insieme di domande (figura 5). Tuttavia per alcune domande filtro tale regola è implicita nel testo della domanda e/o nella risposta che viene scelta o, infine, nel testo della domanda che segue il filtro (figura 2 e figura 7).

Effettuate queste distinzioni preliminari sulla natura dei quesiti, l'analisi si è concentrata sulla raccolta di una serie di informazioni relative alla qualità dei dati dell'indagine.

Per quanto riguarda la non risposta totale è necessario specificare la distinzione tra il mancato contatto della famiglia, la quale non ha partecipato all'indagine Multiscopo, e il mancato contatto dell'individuo appartenente ad una famiglia partecipante all'indagine Multiscopo ma non rispondente all'indagine che si esamina nel presente lavoro.

Figura 5

"Uffici anagrafici (comunali e circoscrizionali)" (sezione 1).

1.2 E' a conoscenza dell'orario attualmente praticato dagli uffici anagrafici del comune e della circoscrizione in cui risiede?

NO.....1 □ → passare alla domanda 1.4
SI.....2 □

(Se Si)

1.3 Come trova i giorni e gli orari di apertura degli uffici in relazione alle sue personali esigenze?

Molto comodi.....1 □
Abbastanza comodi.....….....2 □
Poco comodi.....….....3 □
Per niente comodi.....….....4 □

Relativamente alle non risposte del primo tipo si è effettuata l'ulteriore distinzione secondo il motivo della mancata intervista, riassunte in tre modalità:

- la famiglia è assente;
- la famiglia rifiuta di collaborare;
- l'indirizzo della famiglia è errato.

Le mancate interviste per individui di famiglie che hanno accettato di rispondere all'indagine Multiscopo sono state, invece, sintetizzate in due motivazioni:

- l'interessato è assente;

– l'interessato non collabora per altri motivi (tra cui il rifiuto)[6].

In relazione agli errori contenuti nel questionario sono state attivate le norme che regolano la compilazione del questionario della sezione relativa all'intervistato. In particolare, sono state prese in considerazione:

– le mancate risposte ai quesiti per i quali è obbligatorio rispondere (solo nel caso in cui l'intervista è stata effettuata);
– le mancate risposte ai sottoquesiti (solo nel caso delle domande in cui almeno un sottoquesito presenta una risposta, ovvero quando la domanda non è stata scartata);
– le violazioni della regola di compatibilità tra la risposta di una domanda filtro e la compilazione della prima domanda successiva legata al filtro.

Sono state, inoltre, osservate le regole che disciplinano la compilazione delle domande rivolte al rilevatore (figura 6):

– le mancate risposte ai quesiti per i quali è obbligatorio rispondere (solo nel caso in cui l'intervista è stata effettuata almeno ad un componente della famiglia);
– la violazione della regola di compatibilità tra la risposta dell'unica domanda filtro e la domanda successiva inserita in questa sezione del questionario.

Completata la fase di revisione dei questionari, le informazioni ricavate, assunte come variabili indicatrici (ad esempio mancata risposta totale: "si - no"), sono state aggregate al livello del rilevatore (generando, ad esempio, la variabile numero mancate risposte totali, o tasso di mancate risposte totali), ottenendo delle variabili che si definiscono come indicatori di qualità.

Gli indicatori di mancato contatto, calcolati su ogni rilevatore, sono stati:

- x_1: tasso di non risposta delle famiglie per assenza, dato dal rapporto fra il numero delle famiglie non rispondenti perché assente al momento del passaggio del rilevatore sul numero totale di famiglie che il rilevatore deve intervistare;

[6] Si è deciso di accorpare le due modalità di non risposta: "rifiuto" e "altri motivi", vista l'esigua presenza di interviste rifiutate.

- x_2: tasso di non risposta delle famiglie per rifiuto dato, dal rapporto fra il numero delle famiglie che hanno rifiutato di rispondere al rilevatore sul numero totale di famiglie che il rilevatore deve intervistare;
- x_3: tasso di non risposta delle famiglie per non eleggibilità, dato dal rapporto fra il numero delle famiglie non rispondenti perché l'indirizzo è errato sul numero totale di famiglie che il rilevatore deve intervistare;
- x_4: tasso di non risposta degli individui per assenza, dato dal rapporto fra il numero degli individui non rispondenti all'indagine di riferimento (ma componenti di una famiglia rispondente all'indagine Multiscopo) per assenza sul numero totale di individui che il rilevatore deve intervistare;
- x_5: tasso di non risposta degli individui per altri motivi, dato dal rapporto fra il numero degli individui che non hanno risposto all'indagine di riferimento per motivi diversi dall'assenza, sul numero totale di individui che il rilevatore deve intervistare;

Gli indicatori di qualità relativi al questionario si riferiscono in parte alla sezione riservata al rilevatore (decima sezione) e in parte alle sezioni a cura del rispondente.

La parte riservata al rilevatore è costituita da tre domande, di cui la prima e la seconda sono obbligatorie mentre la terza è compilata in dipendenza della precedente risposta (figura 6).

Le variabili calcolate su ciascun rilevatore, sono:

- y_1: numero medio di mancate risposte parziali per questionario sulle domande obbligatorie (primo e secondo quesito) nella parte riservata. L'indicatore è calcolato con il numero totale di mancate risposte parziali per rilevatore sul numero totale di unità statistiche che questo doveva contattare (numero di questionari);
- y_2: tasso di questionari in cui si registra una incompatibilità tra le risposte della seconda e la terza domanda nella parte riservata. L'indicatore è calcolato con il numero di infrazioni della regola formale, che disciplina la compatibilità fra il

secondo e il terzo quesito[7], sul numero totale di unità statistiche che compongono le famiglie contattate dal rilevatore.

Gli indicatori relativi alla parte a cura del rispondente si basa sui questionari compilati in collaborazione del rilevatore[8]. Gli indicatori sono:

- z_1: numero medio di mancate risposte parziali per questionario su domande obbligatorie non composte da subquesiti (quesiti del primo tipo). Sono considerate le domande che non hanno ricevuto risposta nonostante il questionario è stato compilato[9]. Per determinare se il questionario è stato compilato, invece, di considerare la risposta alla seconda domanda della parte riservata al rilevatore (figura 6), si è imposto che almeno una delle domande suindicate abbia avuto risposta[10]. L'indicatore è calcolato con il numero totale di mancate risposte parziali per rilevatore sul numero di questionari compilati;
- z_2: numero medio di mancate risposte parziali per questionario sui subquesiti, evidenziati da un avviso (figura 2) che compongono una domanda in cui il numero dei subquesiti è inferiore o uguale a quattro[11]. Nella costruzione dell'indicatore la mancata risposta parziale sul subquesito si conteggia quando almeno uno degli altri subquesiti è compilato. Se tutti i subquesiti di una domanda non hanno risposta si ipotizza che il questionario non è stato compilato o che la domanda sia stata scartata a causa di una norma formale di compilazione del questionario e, dunque, non rientrano nel conteggio. L'indicatore è calcolato con il numero totale di mancate risposte parziali sui subquesiti (che sono al massimo quattro nella domanda) e il numero totale di questionari compilati;

[7] L'infrazione della regola si presenta quando alla seconda domanda si risponde con una delle prime quattro opzioni proposte e si compila la terza o quando alla seconda domanda si risponde con la quinta opzione e non si compila la terza domanda.

[8] Tale informazione si desume dalla seconda domanda della decima sezione del questionario (figura 6).

[9] Le domande prese in considerazione sono la 1.1, 1.2, 1.4 della prima sezione la 2.1, 2.4, 2.6 della seconda sezione, la 3.1, 3.3 della terza, la 6.1 della sesta sezione, la 7.1, 7.3, 7.4, 7.5, 7.6 della settima sezione, la 8.1 dell'ottava sezione, le domande 9.1 e 9.2 della nona sezione.

[10] Si è seguita questa strada in quanto in alcune occasione la risposta al secondo quesito della parte riservata non corrisponde al reale stato del questionario (ossia il questionario è compilato mentre la risposta alla seconda domanda della parte riservata è la quinta, oppure è una delle prime quattro mentre il questionario non è compilato).

[11] Sono considerate le domande 1.9, 2.11, 8.7.

Figura 6

Domande del questionario della sezione riservata al rilevatore.

1. Il questionario è stato compilato nello stesso giorno in cui è stata effettuata l'intervista?

NO.....1 □
SI.....2 □

2. Il questionario è stato compilato:

Dall'interessato.....1 □
Dall'interessato in collaborazione con il rilevatore.....2 □
Da un familiare.....3 □
Da un familiare in collaborazione con il rilevatore.....4 □
Non è stato compilato.....5 □

3. Se il questionario non è stato compilato indichi per quale motivo:

Assenza dell'interessato.....1 □
Rifiuto.....2 □
Altro motivo.....3 □

- z_3: numero medio di mancate risposte parziali per questionario sui subquesiti che compongono una domanda in cui il numero dei subquesiti è maggiore di quattro[12]. L'indicatore è costruito come il precedente;
- z_4: numero medio di incompatibilità nel questionario fra domande filtro e le successive, quando è presente la nota di avviso relativa alla regola formale di compilazione[13]. Nella costruzione dell'indicatore si conteggiano le infrazioni della regola di compatibilità fra la domanda filtro e la domanda successiva. Se la domanda filtro non è stata compilata non viene considerata. L'indicatore è calcolato con il numero totale di infrazioni sulle coppie di domande (con la domanda filtro compilata) e il numero di questionari compilati;
- z_5: tasso di questionari in cui si registra una incompatibilità tra le risposte delle domande 1.11 e 1.12. Questo indicatore è stato distinto dal precedente perché la domanda 1.11 è

[12] Sono considerate le domande 1.19, 2.19, 3.8, 3.17, 4.1, 5.1, 7.7, 8.4, 8.5, 8.6, 9.3.

[13] Sono considerati i quesiti 1.2, 1.4, 1.17, 2.1, 2.4, 2.6, 2.12, 2.17, 3.1, 3.3, 3.9, 6.1, 6.2, 6.3, 7.1, 7.2, 7.6, 8.1.

priva di una possibile opzione di risposta "non ricordo" per cui l'infrazione della norma di compilazione potrebbe essere stata influenzata da questo errore sul questionario. L'indicatore è calcolato come il precedente;

- z_6: tasso di questionari in cui si registra una incompatibilità tra le risposte del secondo subquesito della domanda 1.9 (figura 2) e la domanda 1.10 (figura 7): le due domande sono disciplinate da una norma formale di compilazione implicita nel testo della domanda successiva al filtro. Se la domanda filtro non è stata compilata non viene registrata alcuna incompatibilità. L'indicatore è calcolato con il numero totale di infrazioni sulla coppia di domande e il numero di questionari in cui è stata compilata la domanda filtro;

Figura 7

Decimo quesito della sezione "Uffici anagrafici (comunali e circoscrizionali)" (sezione 1).
(La domanda segue quella della figura 2)

1.10 Se ha richiesto il rilascio di certificazioni anagrafiche, può indicarne il motivo?

Per rispondere ad una richiesta:
dell'amministrazione comunale .....1 □
di un'altra amministrazione pubblica .....2 □
di un soggetto privato .....….....3 □
Per altri motivo .....….....4 □

- z_7: tasso di questionari in cui si registra una incompatibilità tra le risposte del primo subquesito della domanda 4.1 (figura 4) e il primo subquesito della domanda 4.2 (figura 8). Le due domande sono disciplinate da una norma formale di compilazione di cui non si fa alcun avviso né con una nota né all'interno del testo della domanda successiva al filtro[14]. Se la domanda filtro non è stata compilata non viene registrata alcuna incompatibilità. L'indicatore è calcolato con il numero totale di infrazioni sulla coppia di domande e il numero di questionari in cui è stata compilata la domanda filtro.

[14] Analoghe misure possono calcolarsi per gli altri sei subquesiti, tuttavia l'elevata correlazione tra i sei possibili indicatori (le correlazioni sono mediamente uguali a 0,9) indica che queste variabili forniscono la stessa informazione sulla qualità dell'indagine e, pertanto, cinque di esse sono state escluse dall'analisi.

Infine è stato calcolato un indicatore relativo alla coerenza delle risposte nella sezione riservata al rilevatore e il reale stato del questionario:

* w_1: tasso di questionari in cui la seconda domanda della sezione riservata al rilevatore (figura 6) contraddice il reale stato del questionario. Sono stati conteggiati i questionari che hanno avuto almeno una risposta nelle domande obbligatorie nelle prime nove sezioni e, contemporaneamente, è stata scelta dal rilevatore la quinta opzione alla seconda domanda, o, viceversa, tutte le domande obbligatorie non hanno avuto risposta ed è stata scelta dal rilevatore una delle prime quattro opzioni della suddetta domanda. L'indicatore è calcolato con il numero totale di questionari con incompatibilità sul numero totale di unità statistiche contattate dal rilevatore.

Figura 8

Secondo quesito della sezione "Altri Uffici Pubblici" (sezione 4). (La domanda segue quella della figura 4)

4.2 Ha trovato comodi i giorni e gli orari di apertura degli uffici in cui si è recata?

	Molto	Abba-stanza	Poco	Per niente
Uffici delle imposte.....	1 □	2 □	3 □	4 □
Uffici del catasto.....	5 □	6 □	7 □	8 □
Uffici della previdenza sociale.....1 □	2 □	3 □	4 □	
Uffici di collocamento..	5 □	6 □	7 □	8 □
Uffici della motorizza-zione civile.....	1 □	2 □	3 □	4 □
Uffici de pubblico re-gistro automobilistico.	5 □	6 □	7 □	8 □

Di seguito sono illustrate alcune statistiche descrittive degli indicatori di qualità esaminati.

Tabella 1 - Statistiche descrittive dei 15 indicatori di qualità misurati

Indicatori	Media*	Scarto quadratico medio	Coefficiente di variazione
x_1	4,53	9,56	2,11
x_2	2,58	7,22	2,80
x_3	1,21	4,61	3,81
x_4	0,13	1,01	7,77
x_5	0,14	0,92	6,57
y_1	0,04	0,12	3,00
y_2	0,91	3,05	3,35
z_1	1,26	2,42	1,92
z_2	0,31	0,81	2,61
z_3	0,50	0,89	1,78
z_4	0,14	0,20	1,43
z_5	0,05	0,13	2,60
z_6	6,23	7,83	1,26
z_7	0,84	7,55	8,99
w_1	0,49	2,15	4,39

*La media di una variabile tasso è espressa in termini percentuali

4. L'Analisi delle Componenti Principali

L'Analisi delle Componenti Principali (ACP) è un metodo multivariato che trasforma un set di p variabili statistiche definite a priori, ottenendo un insieme ridotto di k (<p) nuove variabili, dette componenti principali, tra loro incorrelate le quali contengono la maggiore informazione possibile dell'insieme originale di caratteri.

In questo contesto, la misura dell'informazione della variabile corrisponde alla sua variabilità, nel senso che una variabile casuale con una varianza elevata fornisce più informazioni sul fenomeno oggetto di studio rispetto ad una variabile pressoché costante.

Si supponga di aver rilevato su n unità statistiche p variabili quantitative (per l'occasione trasformate in variabili con media nulla) le quali sono ordinate in una tabella con n righe e p colonne:

$$\underline{X} = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \cdots & & \cdots & & \cdots \\ \cdots & & \cdots & & \cdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \cdots & & \cdots & & \cdots \\ \cdots & & \cdots & & \cdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix},$$

in cui la generica riga i-esima indicata con x_i descrive i valori osservati sull'unità statistica i nelle p variabili scarto, mentre sulla generica colonna r-esima x_r sono rappresentate le determinazioni della variabile scarto r nell'insieme delle n unità statistiche.

La definizione della prima componente principale, y_1 , è rappresentata dall'espressione

$$y_1 = a_{11}\, x_1 + a_{12}\, x_2 + \ldots\ldots + a_{1r}\, x_r + \ldots\ldots + a_{1p}\, x_p = \underline{X}\, a_1', \qquad [1]$$

in cui il vettore trasposto $a_1' = (a_{11}, a_{12}, \ldots\ldots, a_{1p})$ dei coefficienti sia scelto in modo tale che la varianza di y_1 (Var (y_1)) sia quella massima.

L'ACP prospetta, dunque, un problema di ottimizzazione in cui la funzione obiettivo, espressa secondo le variabili di partenza, assume la forma

$$\text{Var}\,(y_1) = \sum_r \sum_s a_{1r} a_{1s} s_{rs} = a_1' \underline{\Sigma}\, a_1, \qquad [2]$$

dove il termine

$$s_{rs} = \frac{1}{n} \sum_{i=1}^{n} x_{ir} x_{is}$$

rappresenta la generica covarianza tra le variabili x_r e x_s quando $r \neq s$ e la varianza di x_r quando r=s, mentre

$$\underline{\Sigma} = \underline{X}'\, \underline{D}\, \underline{X} = \begin{bmatrix} s_{11} & s_{12} & \cdots & \cdots & s_{1p} \\ s_{12} & s_{22} & \cdots & \cdots & s_{2p} \\ \cdots & & \cdots & & \\ \cdots & & & \cdots & \\ s_{1p} & s_{2p} & \cdots & \cdots & s_{pp} \end{bmatrix},$$

è la matrice, di ordine p, che contiene le varianze e covarianze del set di variabili e $\underline{D}$ è la matrice diagonale con n righe e n colonne, i cui elementi rappresentano i pesi che vengono attribuiti a ciascuna osservazione (generalmente posti uniformemente pari a 1/n).

Per ottenere il valore massimo della [2] bisogna imporre un vincolo alla configurazione dei pesi, in quanto, in caso contrario, la varianza della combinazione lineare aumenta indefinitamente selezionando coefficienti a_{1r} sempre maggiori.

Il vincolo che si pone sui coefficienti si esprime generalmente tramite il rispetto dell'uguaglianza

$$a_{11}^2 + a_{12}^2 + ... + a_{1p}^2 = a_1' a_1 = 1 \qquad [3]$$

(condizione di normalizzazione).

Il processo di estrazione della prima componente principale basata sulla [2] e [3] conduce alla soluzione del sistema a p equazioni

$$\underline{\Sigma}\, a_1 = \lambda_1 a_1, \qquad [4]$$

definite come equazioni agli autovalori, dove il vettore a_1 viene detto autovettore (o vettore caratteristico) di $\underline{\Sigma}$ e lo scalare λ_1 viene detto autovalore della stessa matrice[15].

Moltiplicando entrambe i membri della [4] per a_1' si ottiene, sotto il vincolo [3], il sistema

$$a_1' \underline{\Sigma}\, a_1 = \lambda_1 = \mathrm{Var}\,(y_1), \qquad [5]$$

il quale evidenzia che i coefficienti della componente principale, a_{1r} , costituiscono gli elementi dell'autovettore associato all'autovalore λ_1 più grande di $\underline{\Sigma}$, il quale, a sua volta, rappresenta, grazie alla [3] la varianza della componente principale.

La seconda componente principale y_2, viene costruita massimizzando la varianza non spiegata dalla prima componente. Al vincolo [3] si aggiunge il vincolo di incorrelazione con la y_1,

[15] Gli autovettori e autovalori di una matrice quadrata $\underline{\Sigma}$ sono rispettivamente tutti e soli i vettori x ($x \neq 0$) e gli scalari λ che soddisfano l'equazione $\underline{\Sigma}x = \lambda x$ detta equazione agli autovalori relativa a $\underline{\Sigma}$. Per le caratteristiche della matrice $\underline{\Sigma}$ (simmetrica e definita positiva) gli autovettori sono ortogonali a due a due mentre gli scalari λ sono reali e non negativi. In particolare il numero degli autovalori strettamente positivi è pari al rango di $\underline{\Sigma}$, ovvero al numero di variabili linearnemente indipendenti, mentre i restanti autovalori sono nulli.

$$a_{11}a_{21} + a_{12}a_{22} + ... + a_{1p}a_{2p} = a_1^{'}a_2 = 0\,, \qquad [6]$$

dove gli a_{2r} rappresentano i coefficienti delle variabili osservate che definiscono la y_2.

Il valore determinato dal procedimento di ottimizzazione individua una varianza pari a λ_2 equivalente al secondo autovalore più grande della matrice $\underline{\Sigma}$, mentre il vettore dei coefficienti corrisponde all'autovettore associato che rispetta il vincolo [3].

Le restanti componenti principali presentano le stesse caratteristiche: la generica r-esima componente principale possiede una varianza λ_r uguale al r-esimo autovalore più grande di $\underline{\Sigma}$ ed è definita con i coefficienti dell'autovettore associato.

I molteplici problemi di massimo vincolato, che si presentano per ogni componente principale, possono essere espressi sinteticamente dall'equazione matriciale

$$\underline{\Sigma}\ \underline{A} = \underline{A}\ \underline{\Lambda} \qquad [7]$$

in cui,

$$\underline{A} = \begin{bmatrix} a_{11} & a_{21} & \ldots & \ldots & a_{p1} \\ a_{12} & a_{22} & \ldots & \ldots & a_{p2} \\ \ldots & & & & \ldots \\ \ldots & & & & \ldots \\ a_{1p} & a_{2p} & & & a_{pp} \end{bmatrix}$$

è la matrice dei p autovettori, mentre

$$\underline{\Lambda} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & 0 & \\ & & \ldots & \\ & 0 & \ldots & \\ & & & \lambda_p \end{bmatrix}$$

è la matrice diagonale dei p autovalori di $\underline{\Sigma}$.

Tuttavia, il numero delle componenti principali non è necessariamente pari al numero delle variabili osservate. Esso, infatti, è determinato dal numero delle variabili osservate incorrelate tra loro. La presenza di una variabile x_r esprimibile come combinazione lineare delle restanti p-1 variabili, determina la presenza di un autovalore nullo indicando che la variabilità dei dati iniziali si può descrivere con p-1 variabili artificiali.

4.1. Scala di misura delle variabili e ACP sulla matrice di correlazione

L'ACP determina diverse soluzioni a seconda che esamini le variabili osservate oppure le loro standardizzazioni, in quanto il processo di massimizzazione della varianza di una combinazione lineare attribuisce in media pesi maggiori (i coefficienti) alle variabili con varianza più elevata e pesi trascurabili per i caratteri con varianza più bassa.

Utilizzare i dati direttamente osservati significa assegnare alle variabili inserite nell'analisi un diverso ordine di importanza. Tale rilevanza espressa dal valore di s_{rr} è tuttavia condizionata soprattutto dalla unità di misura con cui si rileva la variabile x_r e non tanto dal ruolo che essa assume nella spiegazione del fenomeno.

Per eliminare l'influenza della scala di misura, che si presenta ad esempio rilevando il reddito pro-capite (misurato in lire), il numero di disoccupati (misurato in persone) e le esportazioni (misurate in quintali) o qualora si voglia, comunque, assegnare un ugual peso di partenza a tutte le variabili, in quanto non si è in grado di assegnarne a priori una gerarchia, si procede rendendo uniforme le varianza attraverso la standardizzazione delle variabili.

L'espressione

$$y_1 = b_{11}z_1 + b_{12}z_2 + ... + b_{1r}z_r + b_{1p}z_p = \underline{Z}\, b_1 , \qquad [8]$$

dove $z_r = x_r / \sqrt{s_{rr}}$, definisce la prima componente principale sulle variabili con scarto quadratico medio unitario, attraverso un diverso insieme di coefficienti b_{1r} i quali sottostanno alla condizione,

$$b_{11}^2 + b_{12}^2 + ...+ b_{1r}^2 + ...+ b_{1p}^2 = 1 . \qquad [9]$$

La [8] può essere riscritta nella forma

$$y_1 = c_{11}(x_1/s_{11}) + c_{12}(x_2/s_{22}) + ... + c_{1r}(x_r/s_{rr}) + ... + c_{1p}(x_p/s_{pp}) = \underline{X}\,\underline{S^2}\, c_1 , \qquad [10]$$

in cui

$$\underline{S}^2 = \begin{bmatrix} 1/s_{11} & & & \\ & 1/s_{22} & 0 & \\ & & \ldots & \\ & 0 & \ldots & \\ & & & 1/s_{pp} \end{bmatrix},$$

rappresenta la matrice diagonale dei valori inversi delle varianze, mentre il vettore $c_1' = (c_{11}, \ldots . c_{1p})$ è uguale a $\underline{S}^{-1}\, b_1$ dove

$$\underline{S}^{-1} = \begin{bmatrix} \sqrt{s_{11}} & & & \\ & \sqrt{s_{22}} & 0 & \\ & & \ldots & \\ & 0 & \ldots & \\ & & & \sqrt{s_{pp}} \end{bmatrix}.$$

Il vincolo al processo di ottimizzazione della funzione obiettivo è dato dalla [9].

La [10] mette in evidenza che le componenti principali sono combinazioni lineari di variabili ponderate i cui pesi sono dati dalla matrice $\underline{S}^2$. Il metodo rende proponibile un qualsiasi altro sistema di pesi, uno di questi, rappresentato dalla matrice diagonale identità $\underline{I}$, è presente implicitamente nella [1].

Il sistema di equazioni agli autovalori al quale si perviene dalla [8]

$$\underline{R}\,\underline{B} = \underline{B}\,\underline{\Psi} \qquad [11]$$

sostituisce la matrice covarianza $\underline{\Sigma}$ e degli autovalori $\underline{\Lambda}$ con quella delle correlazioni

$$\underline{R} = \begin{bmatrix} 1 & r_{12} & & r_{1p} \\ r_{12} & 1 & & \\ & & \ldots & \\ & & \ldots & \\ r_{1p} & & & 1 \end{bmatrix}$$

e con i corrispettivi autovalori racchiusi nella matrice diagonale $\underline{\Psi}$.

4.2. Proprietà dell'ACP

La soluzione individuata dall'ACP gode di numerose proprietà. Di seguito sono elencate quelle che sono usate direttamente per interpretare i risultati, supponendo che venga impostata la condizione di normalizzazione [3] o [9].

a) Il numero delle componenti principali è uguale al rango della matrice $\underline{X}$. Nel caso in cui n>p il rango di una matrice è dato dalla differenza tra il numero di variabili ed il numero di quelle esprimibile come combinazione lineare delle altre.
b) La somma delle varianze delle componenti principali è pari a quella delle variabili di partenza:

$$\sum_{r=1}^{p} s_{rr} = \sum_{r=1}^{p} \mathrm{Var}(y_r) .$$

c) Si dimostra (si veda ad esempio Jolliffe, 1986) che il coefficiente di correlazione tra una variabile x_r e una componente y_s (chiamato component loading) è data dall'espressione

$$r(x_r, y_s) = \frac{a_{rs} \pm \sqrt{\lambda_s}}{\sqrt{s_{rr}}} .$$

Nel caso di un'ACP su variabili standardizzate il coefficiente diventa

$$r(z_r, y_s) = b_{rs} \pm \sqrt{\psi_s}$$

d) Nell'ACP con variabili standardizzate la varianza della generica componente principale si ottengono come somme dei quadrati dei coefficienti di correlazione

$$\sum_{r=1}^{p} r(z_r, y_s)^2 = \mathrm{Var}(y_s) = \psi_s$$

Il termine $r(z_r, y_s)^2$ rappresenta il contributo assoluto di x_r alla varianza y_s. E' evidente che le variabili con una più alta correlazione (sia positiva che negativa) offrono il contributo maggiore alla variabilità della componente.

e) Le componenti principali essendo combinazioni lineari di variabili centrate hanno media nulla.

f) Nell'ACP con variabili standardizzate la varianza di una variabile si determina come

$$\sum_{s=1}^{p} r(z_r, y_s)^2 = 1$$

Qualora la somma si limita alle prime k componenti principali (k<p) la somma

$$\sum_{s=1}^{k} r(z_r, y_s)^2 \leq 1$$

viene detta varianza della variabile spiegata dalle prime k componenti principali. Di conseguenza la variabilità di una variabile poco correlata con l'intero set di componenti, viene mal rappresentata.

g) Il contributo della s-esima componente principale alla spiegazione della variabilità complessiva è dato dal rapporto

$$\frac{\lambda_s}{\sum_{s=1}^{p} \lambda_s}$$

Il contributo di k componenti principali, ad esempio le prime estratte, alla spiegazione della variabilità complessiva , definito dal rapporto

$$\frac{\sum_{s=1}^{k} \lambda_s}{\sum_{s=1}^{p} \lambda_s}$$

è detto indice di qualità globale.

h) La tecnica garantisce che il sottoinsieme delle prime k (<p) componenti principali estratte contiene il massimo della varianza di $\underline{X}$ che possa essere contenuta da un insieme di k combinazioni lineari e non correlate delle variabili.

4.3. Scelta del numero di componenti principali da analizzare

Il numero delle variabili artificiali che il metodo individua coincide, come più volte accennato, al numero delle variabili iniziali, a meno della presenza di una dipendenza lineare nella matrice $\underline{X}$ che raramente si presenta nei casi concreti.

Tuttavia, l'ottica con cui, generalmente, si utilizza l'ACP è quella di ridurre le dimensioni della matrice dei dati poiché in questo caso si ottiene una reale semplificazione delle informazione che essa contiene.

Si procede, quindi, alla esclusione forzata di alcune componenti con l'obiettivo di preservare il più possibile la varianza delle variabili. La scelta delle componenti principali da scartare ricade, per la proprietà h, sulle ultime variabili artificiali identificate per le quali, come è noto, la varianza che le caratterizza è inferiore a quella delle prime componenti.

L'efficacia della riduzione della dimensione della matrice dei dati, dipende dalle correlazione delle variabili iniziali. Ad esempio nel caso estremo in cui tutte le variabili sono perfettamente correlate una sola componente è sufficiente a spiegare la varianza totale della $\underline{X}$ (proprietà a), all'opposto la perfetta indipendenza delle variabili induce il modello ad identificare le stesse variabili di partenza, senza semplificare la matrice dei dati. Nelle applicazione pratiche la situazione sarà intermedia ai due estremi teorici sopra delineati, restando il fatto che la sostituzione delle prime k componenti principali al posto delle p variabili risulterà tanto più adeguata quanto più ci si avvicinerà al massimo legame lineare tra le variabili.

Per selezionare il numero di componenti principali si possono seguire vari criteri euristici. Tre di questi sono quelli più presenti in letteratura.

Il primo criterio si propone, innanzitutto, di preservare la variabilità dei dati (ad esempio l'80% della varianza totale; Jolliffe, 1986) e seleziona il più piccolo valore k tale che le componenti principali spieghino una varianza ad un livello non inferiore a tale soglia.

Il secondo criterio, si basa sull'osservazione del plot degli autovalori (Cattell, 1966). Questo grafico pone in ascissa il numero d'ordine delle componenti principali, in ordinata gli autovalori associati ed i punti così individuati sono uniti da segmenti. Il numero delle componenti è, dunque, scelto uguale al più piccolo valore k tale che l'andamento della spezzata è fortemente decrescente alla sua sinistra, mentre a destra la pendenza è debole. In questo modo si cerca di includere solo le dimensioni con un contributo esplicato molto più elevato rispetto alle altre variabili artificiali.

L'ultimo criterio, in presenza di variabili standardizzate, elimina la componenti principali con autovalori inferiori all'unità (Kaiser, 1960), in quanto esse contengono meno informazioni di quanto ne presenti una variabile di partenza (la cui varianza è pari all'unità).

Tuttavia è necessario ricordare che per le caratteristiche poco oggettive, tali criteri non devono essere seguiti rigidamente, ma devono essere mediati con altri aspetti dell'analisi, non ultima la natura del campo di studio che si intende esaminare e le risposte che si vogliono ottenere dal modello. Ad esempio se il set del prime k componenti trattenute spiega un basso livello di varianza per una variabile ritenuta fondamentale nell'analisi, si possono inserire quelle componenti che maggiormente contribuiscono ad aumentare tale livello.

4.4. Interpretazione delle soluzioni

Definito il numero delle componenti principali risulta interessante offrire una interpretazione dei fenomeni che queste ultime misurano.

Tale obiettivo si può ottenere considerando i coefficienti di correlazione tra variabili e componenti (proprietà c).

Se la componente principale possiede un elevato coefficiente di correlazione positivo con una variabile il significato della componente sarà simile a quello della variabile in questione. Viceversa se la correlazione è di segno negativo l'interpretazione andrà nel senso

opposto. Nei casi più concreti una componente è spesso correlata ad un gruppo di variabili ed, in questo caso, essa rappresenterà una misura intermedia delle variabili.

L'interpretazione delle componenti principali è agevolata dalla presenza di forti legami lineari (positivi o negativi) tra indicatori e componenti oppure dalla indipendenza degli stessi (correlazione vicino allo zero).

Allo scopo di trovare una soluzione con queste caratteristiche (coefficienti di correlazione tra variabili e componenti elevati o vicini allo zero) è possibile trasformare o "ruotare" la soluzione, determinando un diverso insieme di combinazioni lineari con lo stesso potere esplicativo delle componenti principali.

Tra i vari metodi di trasformazione nell'applicazione del presente lavoro è stato adottato quello che utilizza il criterio *varimax* (Kaiser, 1958). Questo tecnica statistica sceglie fra tutte le possibili soluzioni di una rotazione ortogonale delle componenti principali, quella che ottimizza il criterio varimax, ovvero che renda massima la somma delle varianze dei coefficienti di correlazione al quadrato delle variabili di partenza con la combinazione lineare prescelta, riducendo in questo modo il numero di coefficienti di correlazione intermedi (racchiusi, ad esempio, negli intervalli –0,4;-0,2 e 0,2; 0,4).

5. I risultati

5.1. L'analisi delle componenti principali

L'Analisi delle Componenti Principali (ACP) è stata applicata e con lo scopo esaminare la struttura di correlazione dei 15 indicatori di qualità selezionati e sostituirli con un numero inferiore di misure di sintesi delle informazioni in essi contenute.

La tabella 1 mostra che la prima componente principale estratta presenta una varianza pari a 2,06, raccogliendo il 13,74% della variabilità complessiva dei quindici indicatori di qualità standardizzati; la seconda componente principale con un autovalore pari a 1,77

contribuisce alla spiegazione dell'11,77% della variabilità dei dati. Le restanti componenti principali offrono un contributo esplicativo via via decrescente.

In base a questi valori le prime due componente principale, spiegano il 25,51%, della dispersione totale, aggiungendo la terza la quota sala al 35,40%. La somma delle 15 componente principali riproduce il 100% della varianza totale degli indicatori di qualità.

Data la scarsa capacità da parte delle prime variabili artificiali nel riprodurre la variabilità totale (le prime quattro componenti principali spiegano, infatti, meno del 50% di varianza), emerge che i 15 indicatori di qualità presentano livelli di correlazione, in media, non elevati. Ciò suggerisce che le informazioni che esprimono gli indicatori sono in buona parte indipendenti tra loro.

Comunque, attraverso questo primo risultato si è definito il numero k (<15) di componente principali da mantenere nello studio, in modo tale da sintetizzare le informazioni contenute nei 15 indicatori di qualità, cogliendo le informazioni comuni delle variabili di partenza.

Tabella 1 - Soluzione dell'ACP: varianza spiegata delle componenti principali.

Componenti	Autovalori (varianza)	Varianza spiegata Percentuale	Varianza spiegata Cumulata percentuale
1	2,06	13,74	13,74
2	1,77	11,77	25,51
3	1,48	9,89	35,40
4	1,20	8,00	43,40
5	1,13	7,52	50,92
6	1,10	7,31	58,23
7	0,94	6,30	64,52
8	0,91	6,05	70,57
9	0.86	5,70	76,27
10	0,80	5,35	81,62
11	0,75	5,03	86,65
12	0,69	4,57	91,22
13	0,60	3,99	95,21
14	0,36	2,43	97,64
15	0,35	2,36	100,00

Questa operazione produce sull'analisi un effetto positivo ed uno negativo: da una parte, scartando alcune componenti principali si semplifica l'interpretazione del fenomeno; dall'altra le k componenti principali mantenute nell'analisi non riproducono l'informazione totale delle variabili iniziali e, quindi, il fenomeno descritto dai 15 indicatori non viene esattamente rappresentato dalle nuove misure[16].

Per selezionare il numero delle variabili artificiali sono stati presi in considerazioni i tre criteri illustrati nel paragrafo precedente; alla luce dei valori della tabella 2, si è portati a concludere che: mediante il criterio del livello soglia minimo pari all'80% di varianza spiegata, si devono inserire nella successiva fase di analisi le prime 10 componente principali, per una varianza spiegata dell'81,62%; mediante l'osservazione dello scree-plot (grafico 1) si devono trattenere le prime 4 componenti principali per una varianza spiegata del 43,40%; mediante il criterio di varianza minima della componente pari all'unità, si devono selezionare le prime 6 componenti principali per una varianza spiegata del 58,23%.

Si è, dunque, ritenuto più adatto l'utilizzo del terzo metodo, in quanto sembra essere il migliore compromesso tra sintesi e riproduzione delle informazioni contenute nei 15 indicatori di qualità.

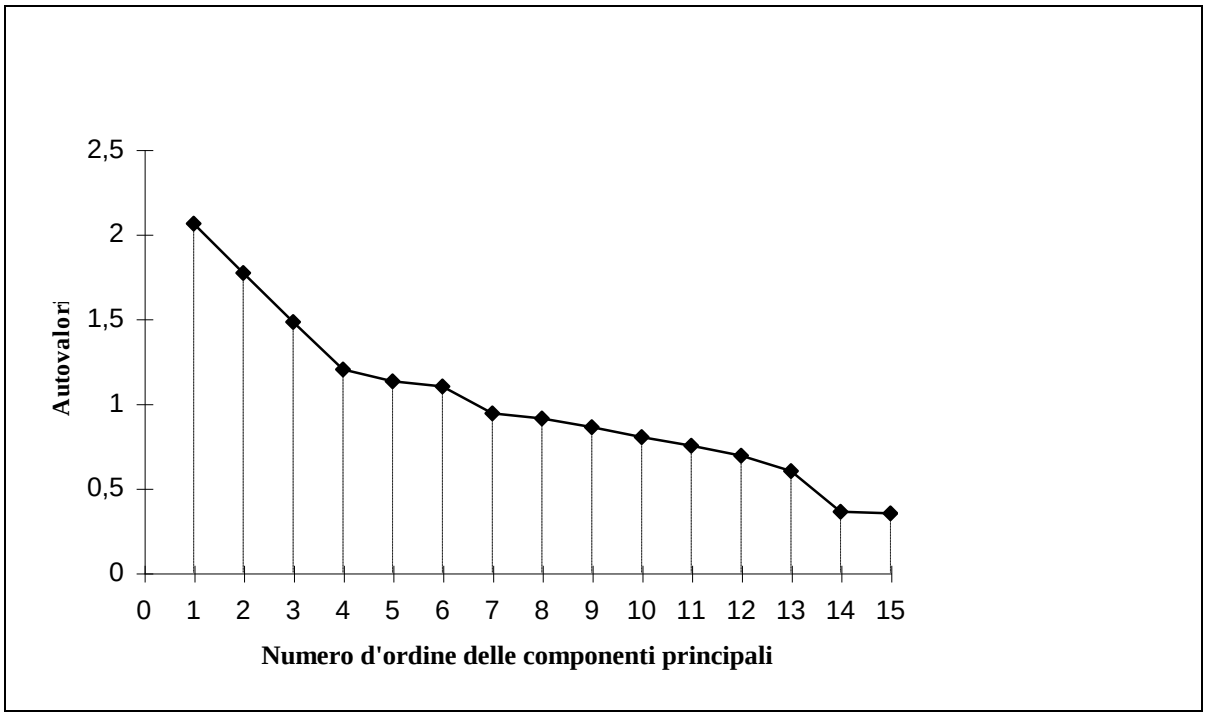
Le componenti principali selezionate essendo costruite dal metodo come combinazioni lineari delle variabili originali, rappresentano a loro volta degli indicatori di qualità dell'indagine. Per ottenere una interpretazione di queste nuove variabili, si è esaminata la struttura di correlazione tra gli indicatori di partenza e le componenti principali (tabella 2).

Da questa analisi si osserva che le prime 6 componenti sono debolmente correlate con le variabili z_6, z_5 e z_4 e la varianza spiegata di ciascuna di esse è inferiore al 50% (rispettivamente pari al 33,6%, 49,6% e 44,9%).

Ciò evidenzia che le nuove variabili colgono una minima parte dell'informazione espressa dai tre indicatori di partenza e pertanto l'interpretazione delle componenti principali è poco legato al loro significato.

[16] A tale proposito è necessario ricordare che, in questo contesto, l'informazione fornita da una variabile sul fenomeno è connessa con la sua variabilità, nel senso che una variabile statistica con elevata variabilità fornisce più informazioni sul fenomeno di quella che manifesta bassa variabilità. Per lo stesso principio, l'informazione contenuta in una matrice di dati è misurabile attraverso la somma delle varianze delle variabili.

Grafico 1 - Soluzione dell'ACP: scree plot



L'interpretazione delle componenti principali, è agevolata dalla presenza di forti legami lineari tra componenti e variabili, sia in senso positivo che negativo, oppure da assenza di correlazione (vicina allo zero). Infatti, ad una componente fortemente correlata con una o più variabili originali è attribuito un significato molto vicino alle variabili stesse, così, ad esempio, la prima variabile artificiale è caratterizzata da una elevata correlazione positiva con le variabili z_3 (0,84)[17], z_2 (0,77), z_4 (0,57) e in misura inferiore con z_5 (0,41) e z_1 (0,37); questi coefficienti suggeriscono che la componente principale rappresenta una misura della qualità dell'indagine nella parte relativa al questionario sottoposta all'intervistato.

Tuttavia, il significato delle altre componenti principali non è altrettanto chiaro. La tabella 2 mostra, ad esempio, che la quinta componente principale presenta coefficienti di correlazione di livello intermedio quali: 0,61 con y_2; 0,46 con y_1; -0,40 con x_2; 0,37 con x_3 attraverso i quali è difficile attribuire un significato alla componente.

A causa di queste complicazioni interpretative, si è deciso di trasformare o "ruotare" ortogonalmente la soluzione, individuando un diverso insieme di combinazioni lineari dei 15

[17] Tra parentesi sono mostrati i coefficienti di correlazione tra la componente principale e la relativa variabile.

indicatori di qualità attraverso il metodo varimax. La tabella 3 illustra i coefficienti di correlazione tra le variabili di partenza e le nuove combinazioni lineari. Queste ultime, non possedendo le proprietà delle componenti principali[18], saranno chiamate, in alcuni casi, fattori.

Tabella 2 - Struttura di correlazione tra gli indicatori e le componenti principali

Indicatori	COMP. 1	COMP. 2	COMP. 3	COMP. 4	COMP. 5	COMP. 6	Varianza spiegata
x_1	-0,206	0,009	0,230	0,628	0,247	-0,128	0,566
x_2	-0,117	-0,236	0,331	0,446	-0,402	-0,221	0,589
x_3	-0,140	-0,264	0,634	-0,148	-0,368	-0,017	0,649
x_4	-0,108	0,647	0,203	0,191	0,110	-0,184	0,554
x_5	0,073	0,641	0,113	-0,157	-0,267	0,200	0,564
y_1	0,176	-0,139	0,175	0,437	-0,217	0,430	0,503
y_2	0,063	-0,024	0,209	0,068	0,612	0,396	0,585
z_1	0,365	0,100	-0,069	0,020	-0,180	0,645	0,596
z_2	0,774	-0,062	0,035	0,239	0,024	-0,087	0,671
z_3	0,835	-0,026	-0,018	0,151	0,048	-0,116	0,737
z_4	0,572	0,120	0,061	-0,056	-0,014	-0,317	0,449
z_5	0,407	0,057	0,354	-0,397	0,035	-0,207	0,496
z_6	-0,055	-0,003	0,341	-0,004	0,460	-0,062	0,336
z_7	0,018	-0,177	0,704	-0,273	0,090	0,179	0,643
w_1	-0,080	0,853	0,205	0,114	-0,085	0,017	0,796

La tecnica di rotazione rispetto alla prima soluzione riduce il numero dei coefficienti di correlazione con valori intermedi[19]. La soluzione proposta in quest'ultima tabella semplifica, pertanto, le interpretazioni dei fattori ai quali sono stati attribuiti i seguenti significati:

a) fattore 1: sembra rappresentare una misura della "non comprensione delle regole di compilazione del questionario che hanno una nota di avviso". Infatti il fattore è fortemente correlato con variabili inerenti alle norme di compilazione esplicitate nei quesiti con avvisi quali z_2 (0,774), z_3 (0,830), z_4 (0,643) e z_5 (0,451), mentre le altre

[18] Le nuove combinazioni lineari nel complesso spiegano lo stesso ammontare di varianza degli indicatori di partenza, tuttavia, la combinazione lineare che dopo la fase di rotazione ha il maggiore livello di varianza non rispetta il criterio di massima varianza possibile spiegata, che si ottiene nella soluzione iniziale dell'ACP.

[19] Effettuando il conteggio dei coefficienti inclusi nell'intervallo (-0,4; -0,2) e (+0,2; +0,4) considerati di livello intermedio, la tabella 2 presenta 23 elementi, mentre nella tabella 3 questi si riducono a 13.

variabili connesse alle norme di compilazione senza un esplicito avviso nei quesiti come z_6 (0,022), y_2 (-0,022) e z_7 (-0,005) sono scarsamente correlate con il fattore.

Tabella 3 - Struttura di correlazione tra gli indicatori e i fattori nella soluzione trasformata dell'ACP con il criterio varimax

Indicatori	FAT. 1	FAT. 2	FAT. 3	FAT. 4	FAT. 5	FAT. 6
x_1	-0,063	0,099	-0,116	0,669	-0,051	0,298
x_2	0,007	-0,059	0,289	0,648	0,042	-0,283
x_3	-0,107	-0,054	0,766	0,175	0,241	-0,126
x_4	0,021	0,681	-0,084	0,182	-0,183	0,130
x_5	0,005	0,653	0,088	-0,257	-0,203	-0,149
y_1	0,057	-0,049	0,085	0,320	0,623	0,014
y_2	-0,022	-0,036	-0,026	-0,047	0,200	0,735
z_1	0,112	0,071	-0,032	-0,257	0,715	0,003
z_2	0,774	-0,082	-0,068	0,097	0,223	0,036
z_3	0,830	-0,073	-0,098	-0,008	0,181	0,027
z_4	0,643	0,096	0,034	-0,075	-0,117	-0,073
z_5	0,451	0,091	0,413	-0,272	-0,184	0,080
z_6	0,022	0,039	0,150	0,091	0,203	0,511
z_7	-0,005	-0,021	0,712	-0,051	0,062	0,358
w_1	-0,030	0,890	-0,040	0,036	0,026	0,010

b) Fattore 2: sembra individuare un indicatore della "indisponibilità del rispondente a partecipare all'indagine", in quanto è fortemente correlato con le variabili w_1 (0,890), x_4 (0,681) e x_5 (0,653), le quali misurano la difficoltà di contattare il rispondente per famiglie che hanno accettato di cooperare nell'indagine mentre è presente una marcata indipendenza lineare con gli indicatori x_1 (0,099) e x_2 (-0,059), di mancata risposta familiare.

c) Fattore 3: può rappresentare una misura degli "errori strutturali dell'indagine" avvenuti sia al livello centrale (questionario) che periferico (errori di lista), poiché esso è fortemente correlato con x_3 (0,766) e z_7 (0,712), variabile che misura un errore nella formulazione delle regole di compilazione del questionario. La correlazione della variabile z_5 (0,413) rafforza tale significato, ricordando, infatti, che questo indicatore individua non solo un aspetto della non comprensione delle norme di

compilazione (fattore 1) ma anche la non corretta formulazione del quesito su cui vengono misurati gli errori[20].

d) Fattore 4: sembra individuare una misura della "difficoltà di contatto della famiglia" visto che si presentano correlazioni elevate con x_1 (0,669) e x_2 (0,648). In base al coefficiente della variabile y_1 (0,320), la quale misura alcuni errori compiuti nella parte del questionario a cura del rilevatore, si può ipotizzare che il mancato contatto sia dovuto, almeno in parte, al rilevatore.

e) Fattore 5: dei sei fattori potrebbe apparire quello di più difficile da interpretazione. Le elevate correlazioni con y_1 (0,623) e z_1 (0,715), rappresentanti alcuni tipi di indicatori di mancate risposte parziali, non sembrano, infatti, conciliarsi con la quasi ortogonalità con altre variabili di mancate risposte parziali, z_2 (0,223) e z_3 (0,181). Tuttavia, queste ultime due variabili misurano le mancate risposte parziali su quesiti che presentano degli avvisi sulle norme di compilazione (fattore 1) e, quindi, possono plausibilmente rappresentare misure della qualità indipendenti dalle prime due (y_1 e z_1).

f) Fattore 6: esso può rappresentare una misura della "non comprensione delle norme di compilazione del questionario implicite nel testo dei quesiti". I coefficienti che suggeriscono tale significato provengono da y_2 (0,735) e z_6 (0,511), variabili che misurano le incompatibilità tra le risposte a particolari domande filtro e le risposte ai quesiti ad essa legate, in cui la norma di compilazione non è esplicitata da un avviso ma è insita nel testo del quesito. La correlazione con z_7 (0,358), conferma questa interpretazione. Questo indicatore, come è stato messo in luce nella descrizione del terzo fattore, rappresenta indirettamente una misura di un palese errore nella costruzione del questionario, ma formalmente rappresenta una misura delle infrazioni di una regola di compatibilità delle risposte in una specifica parte del questionario[21]. Il fattore è pressoché indipendente linearmente con gli indicatori che misurano le

[20] Si ricorda che la variabile rappresenta un tasso di non comprensione della norma di compilazione con nota di avviso su un quesito a risposta chiusa, in cui, però, non è presente una possibile opzione di risposta del tipo "non ricordo".

[21] Tale regola non è evidenziata ne implicitamente nel testo del quesito filtro, ne esplicitamente con un avviso, ma emerge seguendo la sequenza logica tra la risposta alla domanda filtro e le risposte ai quesiti successivi.

incompatibilità di regole di compilazione evidenziate da un avviso e che caratterizzano il primo fattore.

In base a queste interpretazioni si possono trarre alcune prime conclusioni:

1) le misure sugli errori compiuti sul questionario (fattori 1, 5, 6) sono indipendenti dalle mancate interviste (fattori 2 e 4);
2) le misure relative agli errori nel questionario quali mancate risposte parziali (fattore 5), non comprensioni delle regole di compilazione esplicite (fattore 1) o implicite (fattore 3) sono tra loro indipendenti e, quindi il commettere alcuni errori non influenza la presenza di altri tipi errori;
3) il mancato contatto dell'individuo (fattore 2) è indipendente dal mancato contatto della famiglia (fattore 4);
4) gli errori strutturali dell'indagine, causati dagli organi centrali (che contribuiscono alla stesura del questionario) che periferici (che forniscono le liste delle unità selezionate nel campione) sono rappresentabili tramite un unico indicatore di qualità (fattore 3).

5.2. L'analisi della dipendenza semplice

Dopo aver individuato questi sei indicatori di sintesi, nella seconda parte del lavoro si è cercato di evidenziare il legame che essi hanno con le caratteristiche strutturali dei rilevatori, per cercare di far emergere quali ed in che modo gli aspetti che identificano un rilevatore influenzano la qualità dei dati.

Le caratteristiche strutturali sono state suddivise in due classi: quelle intrinseche del rilevatore (sono state considerate età, sesso, grado di istruzione, esperienze in indagini

precedenti dell'ISTAT, condizione occupazionale[22]); quelle ambientali in cui opera il rilevatore (sono state considerate la localizzazione e dimensione demografica del comune[23]).

Si è, dunque, investigato sulla dipendenza tra le variabili strutturali e i sei fattori mediante l'ausilio della statistica-test del χ^2 , tenendo in considerazione che le conclusioni a cui si può giungere attraverso tali analisi devono tenere conto delle possibili interazioni presenti tra le stesse variabili strutturali. Ad esempio, secondo i dati a disposizione emerge che la variabile età è strettamente connessa con la condizione occupazionale del rilevatore, nel senso che all'aumentare dell'età cresce la probabilità che esso sia un dipendente comunale. Pertanto, nell'ipotesi che i valori del primo fattore siano legati all'età del rilevatore, la dipendenza tra le due variabili strutturali può generare indirettamente una dipendenza tra lo stesso fattore e la condizione occupazionale. E' stato, dunque, necessario elaborare analoghe analisi di dipendenza sulle variabili strutturali. Si è, quindi, provato statisticamente che, le variabili età, esperienza e condizione occupazionale, mostrano un grado di dipendenza significativo[24].

Per tutte le altre coppie di variabili strutturali del rilevatore il test indica che le variabili sono statisticamente indipendenti tra loro, ad un livello di significatività del 5%.

Eseguita questa pre-analisi è stata studiata la dipendenza tra variabili strutturali e fattori.

Nella tabella 4 compaiono le probabilità di commettere un errore qualora si consideri provata l'ipotesi di dipendenza tra variabile strutturale e fattore quando, invece, le due

---

[22] L'età si articola nelle classi "18-35 anni" e "36 anni e oltre"; il grado di istruzione si articola nelle modalità "basso" (licenza elementare e media) e "alto" (diploma superiore e laurea); l'esperienza in due modalità, "1 o 2 indagini eseguite" e "3, 4, 5 indagini eseguite"; la condizione occupazionale si suddivide nelle modalità "disoccupato" e "dipendente comunale".

[23] La localizzazione del comune si distingue in Nord, Centro e Sud, la dimensione demografica in comuni con meno di 50.000 abitanti e comuni con più di 50.000 abitanti.

[24]Dal test risulta che l'ipotesi di indipendenza presenta una probabilità di errore di prima specie inferiore al 5%. In particolare, nel caso della coppia delle mutabili età ed esperienza alla modalità "minore ai 36 anni" della variabile età si associa con maggiore frequenza la modalità "una o due indagini" dell'altra variabile (la probabilità dell'errore di prima specie del test sull'indipendenza è pari a 0,1%); nel caso della coppia di mutabili età e stato occupazionale alla modalità "minore ai 36 anni" della variabile età si associa con maggiore frequenza la modalità "disoccupato" dell'altra variabile (la probabilità dell'errore di prima specie del test sull'indipendenza è pari a 1,1%); nel caso della coppia di mutabili esperienza e stato occupazionale alla modalità "una o due indagini effettuate" della variabile esperienza si associa con maggiore frequenza la modalità "dipendente comunale" dell'altra variabile (probabilità dell'errore di prima specie del test sull'indipendenza è pari a 2,1%).

variabili sono indipendenti[25], e sono stati evidenziati quei casi in cui la probabilità di errore è inferiore al 10%.

Dalla tabella si evince che il secondo e il terzo fattore sono indipendenti dalle caratteristiche intrinseche del rilevatore, mentre sembrano dipendere dalla dimensione demografica (per il secondo fattore la probabilità dell'errore di prima specie è α=0,1%; per il terzo fattore è α=0,2%) e dalla localizzazione (per il secondo fattore è α=0,3%; per il terzo fattore è α=0,1%).

Tabella 4 – Probabilità (in percentuale) di commettere un errore nell'ipotizzare la dipendenza tra variabili strutturali e fattori

Variabili	Sesso	Titolo di Studio	Età	Condizione professionale	Esperienza	Dimensione demografica del comune	Localizzazione geografica del comune
Fattore 1	71,7	**8,8**	**2,5**	**0,4**	22,2	70,6	**9,1**
Fattore 2	96,9	67,0	14,9	23,4	56,5	**0,1**	**0,3**
Fattore 3	82,7	12,9	51,8	25,8	20,8	**0,2**	**0,1**
Fattore 4	**4,4**	72,9	12,1	30,9	**0,1**	**0,1**	**0,1**
Fattore 5	59,5	45,4	45,3	82,1	35,9	34,3	12,9
Fattore 6	42,8	11,4	**7,1**	28,7	**0,3**	61,9	**6,2**

Il primo, quarto e sesto fattore presentano valori del χ^2 tali da poter inferire la dipendenza anche con le variabili strutturali intrinseche del rilevatore. In dettaglio, il primo fattore sembra dipendere dal titolo di studio (α=8,8%), dall'età (α=2,5%), dalla condizione occupazionale (α=0,4%) e, in misura più debole dalla localizzazione del comune (α=9,1%). Il quarto fattore appare connesso con la variabile sesso (α=4,4%), esperienza (α=0,1%), dimensione demografica del comune (α=0,1%) e localizzazione (α=0,1%); il sesto fattore sembra dipendere dall'età (α=7,1%), dall'esperienza (α=0,3%) e dalla localizzazione (α=6,2%).

[25] Ciascun fattore è stato suddiviso in classi. Ogni classe contiene circa il 25% delle unità statistiche osservate.

Infine il quinto fattore sembra essere legato ad altre variabili che non sono state inserite nell'analisi anche se è presente una debole dipendenza con la localizzazione del comune (α=12,9%).

Questi risultati aggiungono ulteriori elementi di descrittivi delle 6 combinazioni lineari, in quanto i fattori che apparivano nella fase interpretativa più distanti dalle influenze dirette del rilevatore (il secondo e terzo fattore), si dimostrano indipendenti dalle caratteristiche intrinseche di questi ultimi. Essi indicano, inoltre, in quale direzione hanno agito le peculiarità dei rilevatori sulla qualità dei dati dell'indagine statistica:

a) il primo e il sesto fattore, definiti come misure della qualità della compilazione dei quesiti in relazione alla comprensione delle norme formali presenti nel questionario, sono connessi, soprattutto con il titolo di studio, l'età e la condizione professionale e l'esperienza del rilevatore. Nel dettaglio del primo fattore, la probabilità di avere ottime prestazioni (valori inferiori al primo quartile del fattore) oppure pessime prestazioni (valori superiori al terzo quartile del fattore) è maggiore quando il rilevatore ha un grado elevato di istruzione. Prestazioni di livello intermedio sono, invece, più probabili, con rilevatori in possesso di un titolo di studio basso (grafico 2). La probabilità di presentare valori elevati del fattore (con il valore del fattore superiore al terzo quartile) è superiore per i rilevatori maggiori di 36 anni (grafico 3). La forte dipendenza tra età e condizione occupazionale caratterizza, verosimilmente, la dipendenza tra quest'ultima mutabile e il primo fattore. In questo caso emerge che i rilevatori in condizione di disoccupati presentano probabilità più alte di possedere valori del fattore nelle prime tre classi quartiliche, mentre nell'ultima classe la probabilità prevalente è quella dei dipendenti comunali (grafico 4). Per quanto riguarda il sesto fattore l'osservazione delle probabilità di avere un valore all'interno di una classe quartilica condizionata alle modalità dell'età e dell'esperienza del rilevatore non conduce a conclusioni altrettanto chiare. Considerando le classi estreme del sesto fattore, i rilevatori più anziani (grafico 5) e con un numero di indagini effettuate superiore a tre (grafico 6), sembrano evidenziare prestazioni migliori avendo una probabilità maggiore per la prima classe ed una probabilità

minore per l'ultima classe. Infine, per quanto riguarda la variabile della localizzazione del comune, essa non mostra una chiara tendenza sul sesto fattore.

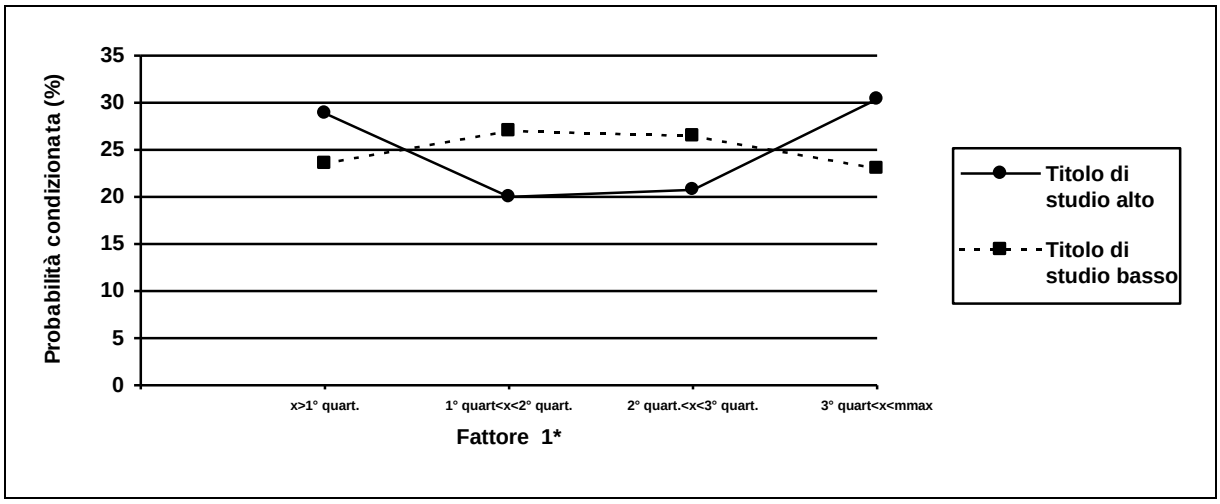
b) Per i fattori relativi alle mancate interviste (secondo e quarto) le variabili ambientali intervengono con maggiore rilevanza. In particolare nel Centro e Sud del Paese sembra più semplice contattare l'individuo di una famiglia (secondo fattore) che ha collaborato nell'indagine Multiscopo. In queste aree la probabilità di trovare rilevatori con valori del secondo fattore inferiore al primo quartile è maggiore che al Nord. Al contrario l'ordine di grandezza delle probabilità delle due aree si inverte qualora si considerino i valori superiori al terzo quartile (grafico 7). Per quanto riguarda il contatto della famiglia (quarto fattore) esso risulta più agevole nel Mezzogiorno che nel Centro e nel Nord. Seguendo l'andamento delle probabilità condizionate, quella di un rilevatore del Sud per un valore del fattore inferiore al primo quartile è superiore rispetto al Nord ed al Centro, ed è minore se, invece, si considerano i casi maggiori del terzo quartile (grafico 8). In entrambe i fattori per i comuni con meno di 50.000 abitanti si hanno maggiori probabilità di entrare in contatto con l'unità selezionata. Ponendo l'attenzione sulle caratteristiche intrinseche del rilevatore si conclude che il mancato contatto di individui appartenenti a famiglie partecipanti all'indagine Multiscopo sia causato, soprattutto, dal rispondente e meno dal rilevatore in quanto il secondo fattore è indipendente da qualsiasi variabile peculiare dell'intervistatore.

   Il quarto fattore appare connesso con il grado di esperienza ed il sesso del rilevatore ed, in particolare, si evidenzia che i rilevatori con minore esperienza offrano maggiori possibilità di contattare la famiglia (grafico 9), mentre dal punto di vista del sesso del rilevatore, la componente femminile mostra maggiori garanzie nel contattare positivamente le famiglie (grafico 10).

c) La dipendenza del terzo fattore esclusivamente con le variabili strutturali di tipo ambientale confermano le considerazioni effettuate in fase di interpretazione del fattore. Infatti, essendo questo una misura degli errori strutturali, è verosimile che le caratteristiche del rilevatore abbiano avuto uno scarso impatto sul fattore. La connessione con le due mutabili ambientali è da attribuire in particolare con gli errori di lista, sui quali possono aver interagito i Comuni, tuttavia, dall'esame delle

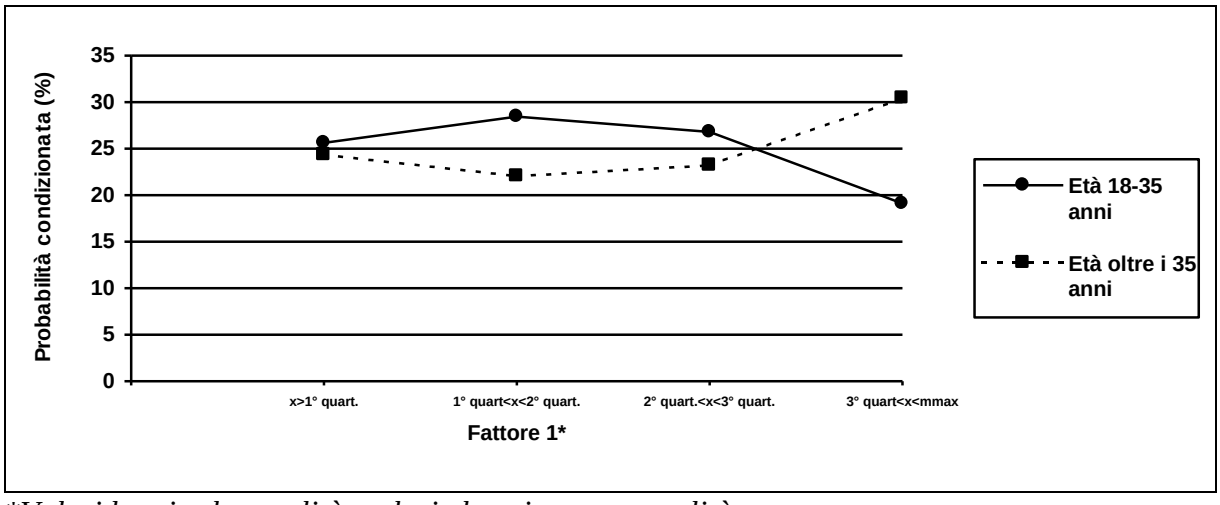
probabilità condizionate non si è in grado di evidenziare particolari associazioni tra le classi del fattore e le modalità delle variabili strutturali.

Grafico 2 – Probabilità del rilevatore di presentare un valore in una determinata classe quartilica del fattore 1 condizionata dal titolo di studio



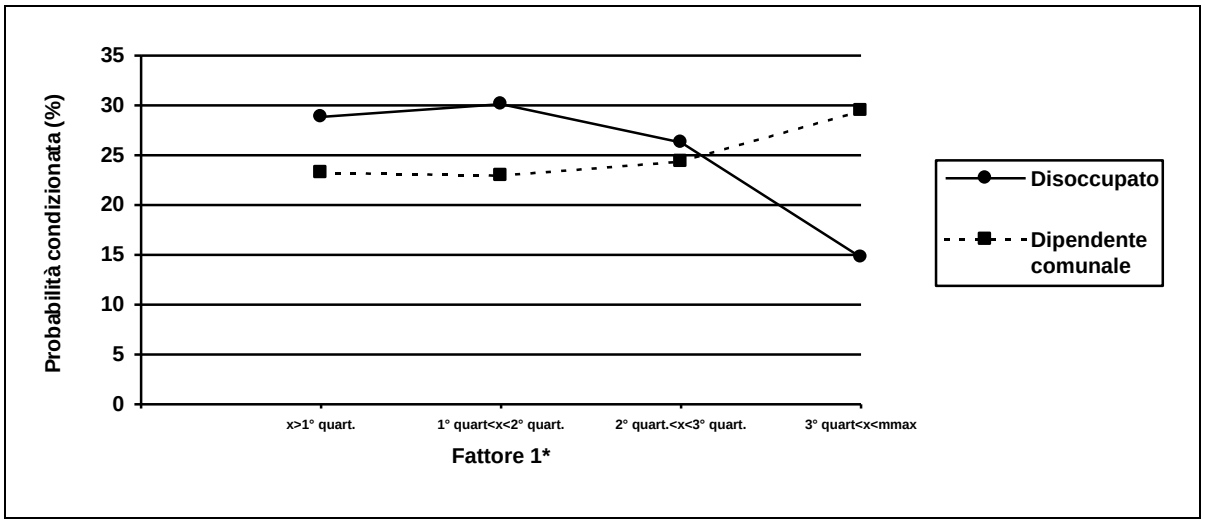
**Valori bassi: alta qualità; valori elevati: scarsa qualità.*

Grafico 3 – Probabilità del rilevatore di presentare un valore in una determinata classe quartilica del fattore 1 condizionata dall'età



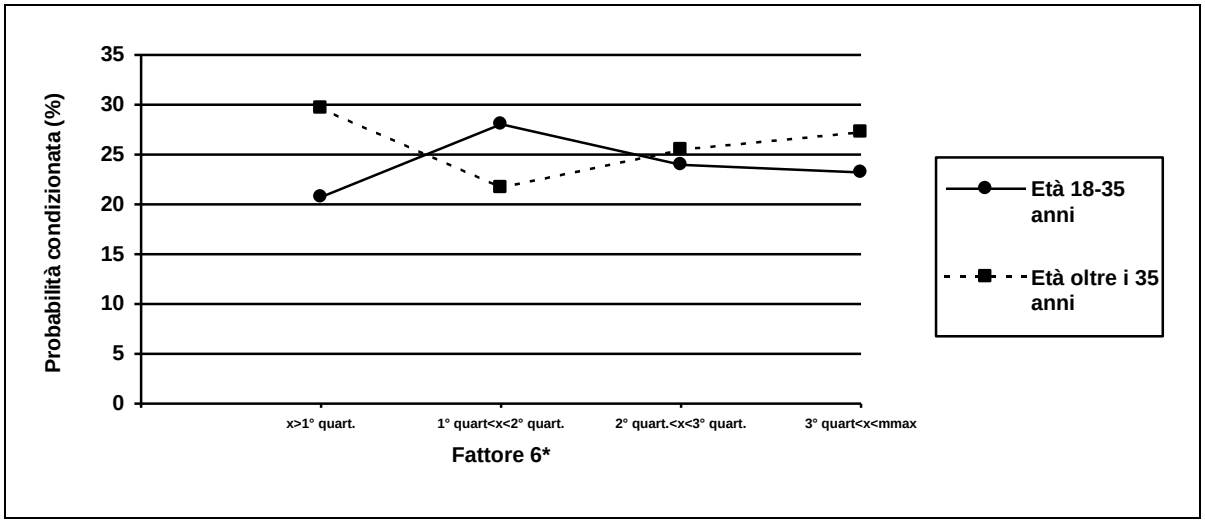
**Valori bassi: alta qualità; valori elevati: scarsa qualità.*

Grafico 4 – Probabilità del rilevatore di presentare un valore in una determinata classe quartilica del fattore 1 condizionata dalla condizione professionale



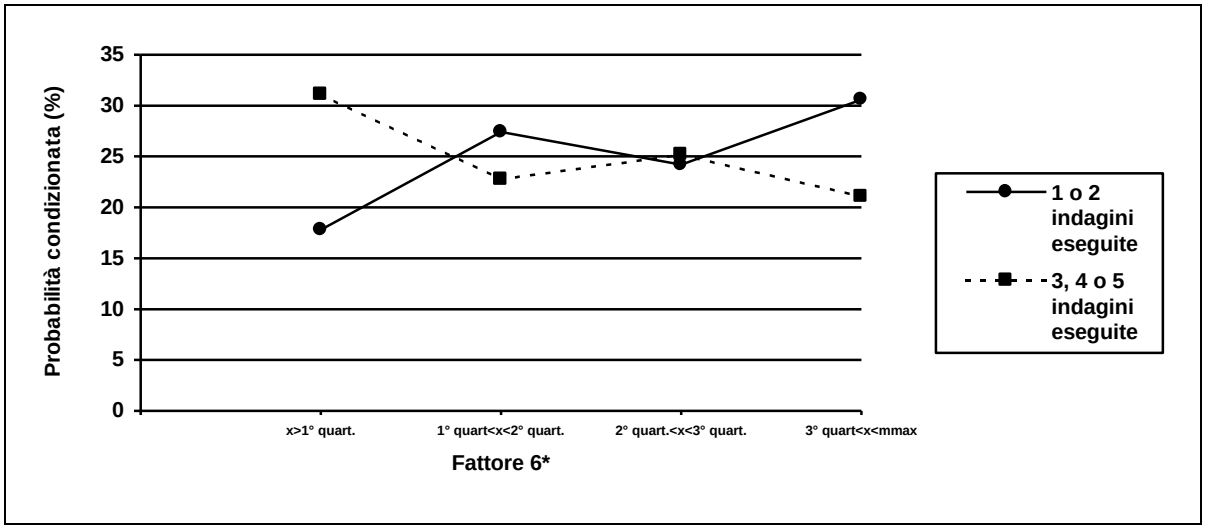
**Valori bassi: alta qualità; valori elevati: scarsa qualità.*

Grafico 5 – Probabilità del rilevatore di presentare un valore in una determinata classe quartilica del fattore 6 condizionata dall'età



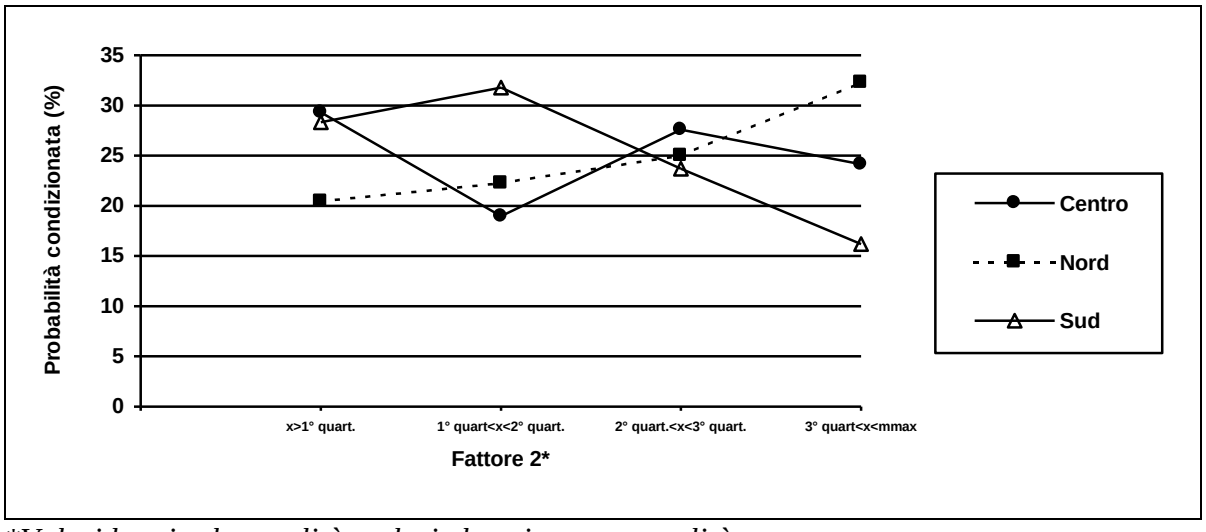
**Valori bassi: alta qualità; valori elevati: scarsa qualità.*

Grafico 6 – Probabilità del rilevatore di presentare un valore in una determinata classe quartilica del fattore 6 condizionata dall'esperienza



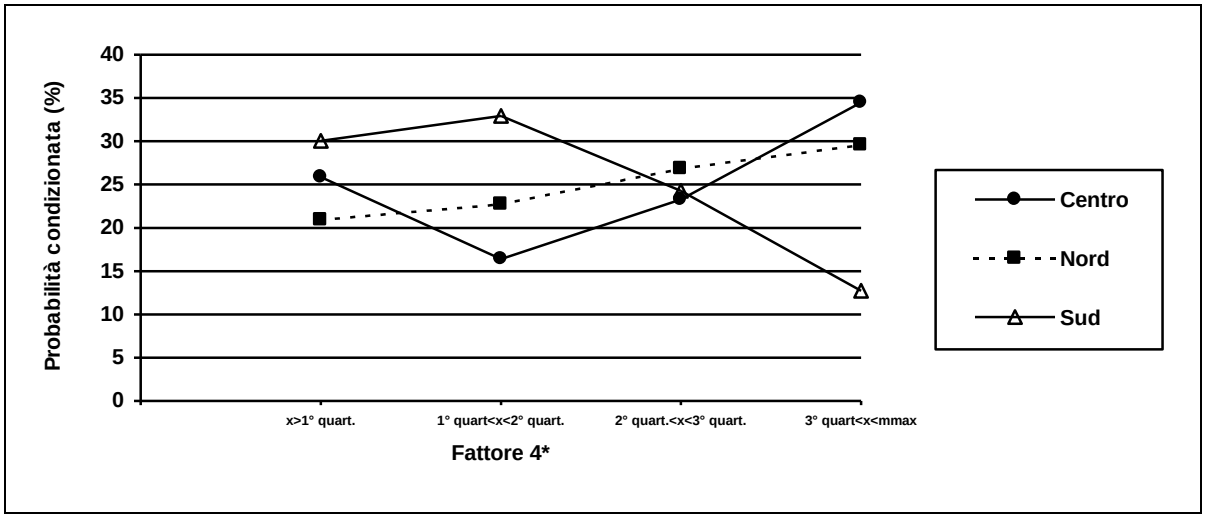
**Valori bassi: alta qualità; valori elevati: scarsa qualità.*

Grafico 7 – Probabilità del rilevatore di presentare un valore in una determinata classe quartilica del fattore 2 condizionata dalla ripartizione in cui si effettua l'intervista



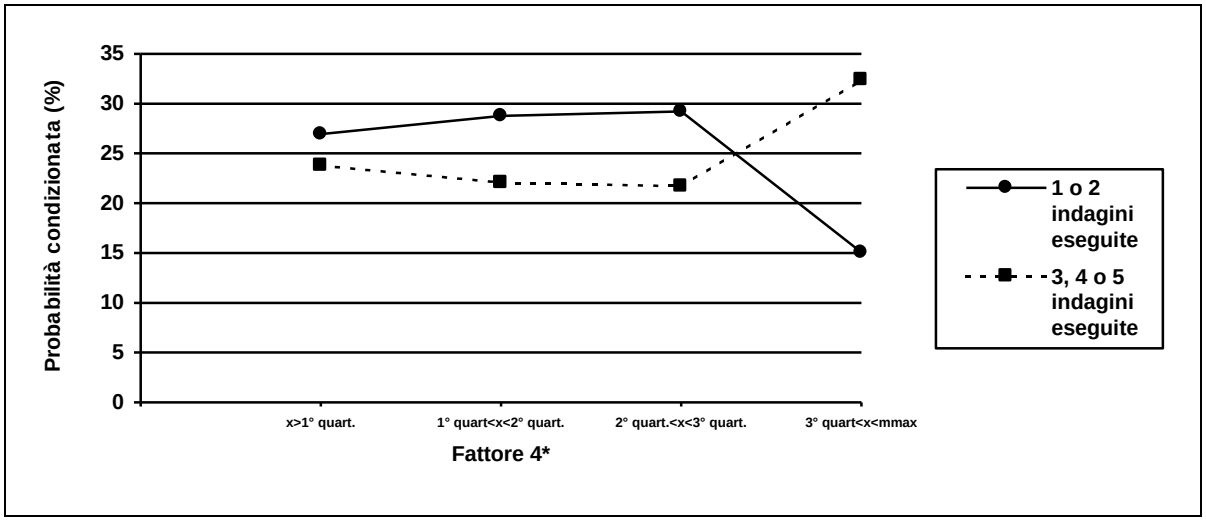
**Valori bassi: alta qualità; valori elevati: scarsa qualità.*

Grafico 8 – Probabilità del rilevatore di presentare un valore in una determinata classe quartilica del fattore 4 condizionata dalla ripartizione in cui si effettua l'intervista



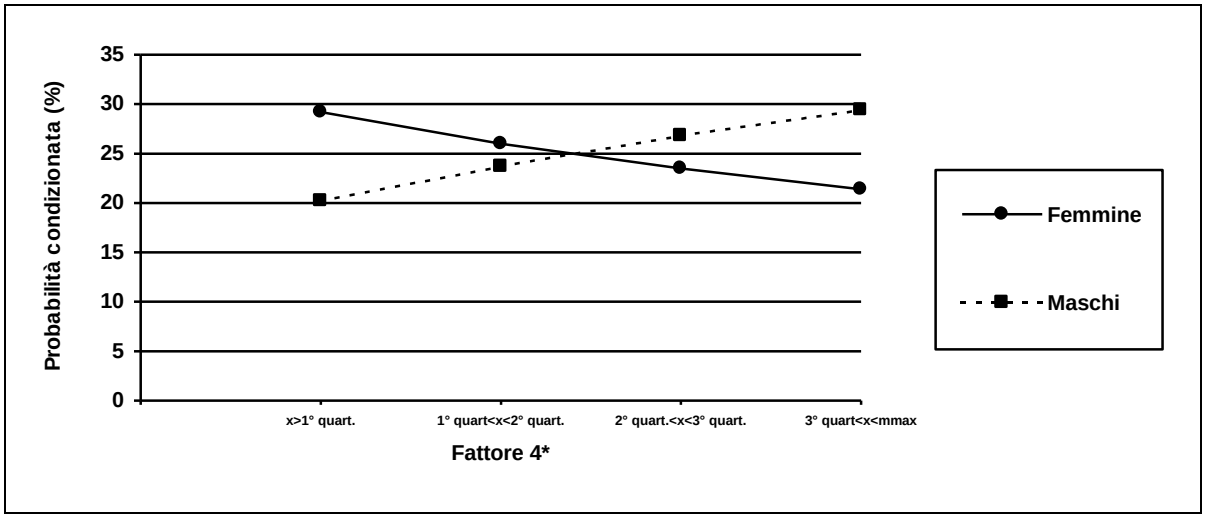
**Valori bassi: alta qualità; valori elevati: scarsa qualità.*

Grafico 9 – Probabilità del rilevatore di presentare un valore in una determinata classe quartilica del fattore 4 condizionata dall'esperienza



**Valori bassi: alta qualità; valori elevati: scarsa qualità.*

Grafico 10 – Probabilità del rilevatore di presentare un valore in una determinata classe quartilica del fattore 4 condizionata al sesso



**Valori bassi: alta qualità; valori elevati: scarsa qualità.*

6. Conclusioni

I risultati della presente analisi consentono di tracciare alcune conclusioni sulla qualità dell'indagine statistica "Servizi resi dalle Pubbliche Amministrazioni e grado di soddisfazione dei cittadini" relative all'operato del rilevatore.

In primo luogo si è potuto constatare che i vari indicatori calcolati, i quali apparentemente sembravano cogliere aspetti molto simili della qualità, si sono dimostrati poco correlati tra loro. Infatti, per spiegare circa il 60% della variabilità delle 15 variabili di partenza sono state necessarie ben 6 componenti principali. Le combinazioni lineari individuate dal metodo ACP (dopo una opportuna fase di trasformazione della soluzione iniziale) hanno mostrato, inoltre, che l'indipendenza non riguarda solo le misure legate alla non risposta totale con quelle relative agli errori compiuti nel questionario, ma caratterizza anche diversi tipi di errori nel questionario.

E’ stato, comunque, possibile definire quali aspetti della non buona qualità dei dati dell'indagine ricadono più propriamente sul rilevatore. Esse si riferiscono: al mancato contatto

della famiglia, alle infrazioni di regole formali di compilazione del questionario esplicitate da una regola di avviso e alle infrazioni alle regole formali di compilazione del questionario implicite nel testo delle domande. E' stato, invece, verificato che altri tipi di errori che inficiano la qualità dell'indagine sono attribuibili ad altri elementi attivi dell'analisi. Il mancato contatto dell'intervistato di una famiglia rispondente all'indagine Multiscopo dipende, ad esempio, dalla persona da intervistare, dalla dimensione demografica e dalla localizzazione geografica comune. Si sono inoltre, palesati errori di tipo strutturali, indipendenti dagli intervistatori, legati alla presenza di incongruità nel questionario ed a errori negli indirizzi delle unità statistiche.

In base ai tre fattori connessi con le caratteristiche del rilevatore, è emerso un profilo dell'intervistatore che ha presentato, in generale, valori degli indicatori di qualità migliori. Questo è di sesso femminile, giovane e disoccupato e con una esperienza rappresentata da una o due indagini compiute.

I rilevatori più esperti (da tre a cinque indagini compiute) presentano, invece, migliori performance nel comprendere delle regole di compilazione non esplicitate con avvisi. Il titolo di studio offre risultati contraddittori: i rilevatori con alto grado di istruzione commettono con maggiori probabilità pochi o molti errori nella compilazione del questionario.

La strategia di analisi adottata per giungere a queste conclusioni si è articolata in più fasi: nella prima sono state costruite delle misure indirette della qualità (indicatori di qualità); nella seconda si è tentato di sintetizzare le informazioni di queste misure mediante l'individuazione di super variabili che hanno colto i principali aspetti della qualità tra loro indipendenti; nella terza fase queste nuove variabili sono state messe in relazione con le caratteristiche del rilevatore.

Tale percorso di ricerca non ha richiamato particolari ipotesi distributive sulle variabili di interesse che, invece, avrebbero richiesto strategie più dirette per raggiungere i medesimi obiettivi. Si può ricordare, ad esempio, il modello statistico dell'analisi fattoriale, che in un'ottica inferenziale si propone di individuare dei fattori latenti (quali possono essere le variabili strutturali del rilevatore) che soggiacciono alla manifestazione delle variabili di partenza (gli indicatori di qualità). Tuttavia, l'ipotesi di normalità sulle variabili analizzate nel presente lavoro, necessaria per applicare il modello, è stata completamente disattesa.

Dall'altra parte l'applicazione di un secondo metodo di analisi dei dati (al pari dell'ACP) come quello delle corrispondenze multiple adatto per valutare contemporaneamente il legame tra indicatori e variabili strutturali del rilevatore, si è dimostrato inefficace a causa del basso grado di coesione degli indicatori di qualità e la conseguente caratterizzazione della soluzione da parte delle sole variabili strutturali.

E' necessario, infine, ricordare che lo scarso livello di associazione degli indicatori è stato certamente influenzato dalla modalità di raccolta dei dati, in quanto nonostante siano stati considerati i questionari compilati con l'aiuto del rilevatore, quest'ultimo non ha necessariamente somministrato il questionario ed, in secondo luogo, a dare le risposte può essere stato una persona diversa dalla persona da intervistare (un familiare).

Questi due elementi hanno certamente attenuato la correlazione tra gli indicatori. Per valutare la bontà della strategia applicata e comprendere con maggiore chiarezza l'influenza del rilevatore sulla qualità dei dati, appare, quindi, necessario analizzare una indagine statistica in cui le domande del questionario sono lette dal rilevatore in modo tale da ridurre il più possibile l'effetto confondente degli errori dovuti al non completo controllo dell'intervista da parte del rilevatore.

Riferimenti bibliografici

CATTELL R.B. (1966), "The Scree Plot Test for the Number of Factors", *Multivariate Behavioral Research*, vol. 1, pp. 245-276.

COCHRAN W.G. (1977), *Sampling Techniques*, 3° ed., Wiley, New York.

DEMING W.E. (1960), *Sampling Design in Business Research*, Wiley, New York.

GROVES R.M. (1989), *Survey Errors and Survey Costs*, Wiley, New York.

HANSEN M.H., HURWITZ W.N., MARKS E.S. e MAULDIN W.P. (1951), "Respons Errors in Surveys", *Journal of the American Statistical Association*, vol. 46, pp. 147-190.

JOLLIFFE I. T. (1986), *Principal Component Analisys*, Verlag - Springer, New York.

ISTAT (1989), *Manuale di Tecniche di Indagine - Il sistema di controllo della qualità dei dati*, Note e Relazioni, vol 6.

KAISER H.F. (1958) “The Varimax Criterion for Analitic Criterion Rotation in Factor Analysis”, *Psychometrika*, vol. 23, pp. 187-200.

KAISER H.F. (1960) “ The Application of Eletronic Computers to Factor Analysis”, *Educational and Psychological Measurement*, vol. 20, pp. 141-151.

KALTON G. (1983), *Compensanting for Missing Survey Data*, Research Report Series: Institute for Social Researche, University of Michigan, Ann Arbor, Michigan.

KENDALL M.G. e BUCKLAND W.R. (1960), *A Dictionary of Statistical Terms*, 2° ed., Oliver & Boyd, London.

KISH L. (1965), *Survey Sampling*, Wiley, New York.

LESSLER J.T. e KALSBEEK W.D. (1992), *Non Sampling Error in Surveys*, Wiley, New York.

MADOW W.G. e OLKIN I. (1983), *Incomplete Data in Sample Surveys*, Academic, New York .

MARDIA K.V., KENT J.T. e BIBBY J.M. (1979), *Multivariate Analisys*, AcademicPress Inc., London.

SUKHATME P.V. e SUKHATME B.V. (1970), *Sampling Theory of Surveys with Applications*, 2° rev. Iowa state Univesity press, Ames, Iowa.

ZARKOVICH S.S. (1966), *Quality of Statistical Data*, Food and Agricoltural Organization of the United Nations, Roma.