Programmazione efficiente delle chiamate in indagini CATI sulle famiglie italiane

Efficient Call-Scheduling in Italian Household CATI Surveys

Luigi Fabbris, Maria Cristiana Martini[1]
Dipartimento di Scienze statistiche, Università degli Studi di Padova

Abstract: Before, while and after taking a census, sample surveys are needed. These investigations are often conducted by telephone. The methods for efficient call scheduling put forward in this paper are based on a stepwise information setup. This setup includes either the general information on aggregate population, or social and demographic characteristics about sample cases, or the history of telephone calls. Probabilities of contact for each household, which are the basic information for the development of optimal dialling protocols in telephone surveys, are estimated with reference to Italian household surveys.

Parole chiave: Call-scheduling; Household surveys; Check surveys; Post enumeration surveys; CATI

1. Programmazione efficiente delle chiamate telefoniche

I censimenti della popolazione sono generalmente affiancati da una serie di indagini campionarie di preparazione, di arricchimento dei dati rilevati, di controllo della qualità dei risultati. Se i risultati di tali indagini devono essere resi disponibili in tempi stretti, la rilevazione per via telefonica costituisce la prima opzione tra gli approcci di rilevazione.
Nelle rilevazioni telefoniche, si deve spesso ripetere più volte la chiamata delle unità oggetto di indagine per ottenere il contatto. Se, prima di iniziare le chiamate, si conoscesse la probabilità di trovare all'altro capo del filo l'insieme di persone cercate, si potrebbero pianificare le chiamate in modo da massimizzare il numero di contatti utili, o, equivalentemente, da minimizzare le chiamate a vuoto.
L'introduzione di sistemi CATI - *Computer Assisted Telephone Interviewing* - consente la pianificazione efficiente delle chiamate definendo in tempo reale l'ordine delle chiamate sulla base di algoritmi programmati. Gli algoritmi si differenziano per il crescendo di informazioni richieste sul campione oggetto d'indagine.
Per rilevare dati sulle famiglie, l'approccio basilare si riferisce alla situazione in cui non è disponibile alcuna informazione sulle unità da rilevare, ma si dispone del solo dato aggregato delle famiglie che appartengono all'area in esame (par.2). La probabilità di chiamare una famiglia campionaria durante il processo di rilevazione dei dati si determina allora in funzione della frequenza di risposta constatata nel passato presso le famiglie della stessa area, o presso famiglie di aree diverse ma con profili analoghi a quelle su cui si indaga.

[1] Il lavoro è stato svolto unitariamente dai due autori. I par. 1 e 5 della nota sono stati redatti da L. Fabbris e i par. 2, 3 e 4 da M.C. Martini.

Se, in aggiunta, sono note la composizione e certe caratteristiche dei membri delle famiglie campionarie, è possibile prevedere i ritmi di vita nella famiglie e, quindi, il flusso delle presenze nel corso dei giorni (par. 3). In particolare, dopo la rilevazione dei dati censuari risulta disponibile una gran mole di informazioni sulle famiglie censite, e ciò consente di pianificare nel modo più efficiente i contatti durante la realizzazione di indagini di controllo della qualità della rilevazione censuaria.
Un terzo gradino informativo è rappresentato dalla storia delle chiamate, ossia dalle informazioni cumulate durante i tentativi di contatto andati a vuoto, per orario e giorno in cui sono avvenuti i tentativi (par. 4). Per esempio, se sono andati a vuoto già due tentativi durante l'orario di lavoro dei giorni feriali, conviene tentare di sera o di sabato.
In questa nota si applica un modello di stima della presenza in casa delle famiglie (intesa come presenza in casa di uno o più membri di almeno 14 anni) sulla base di:

a) caratteristiche socio-demografiche delle famiglie e dei singoli componenti
b) esiti delle chiamate effettuate fino a quel momento.

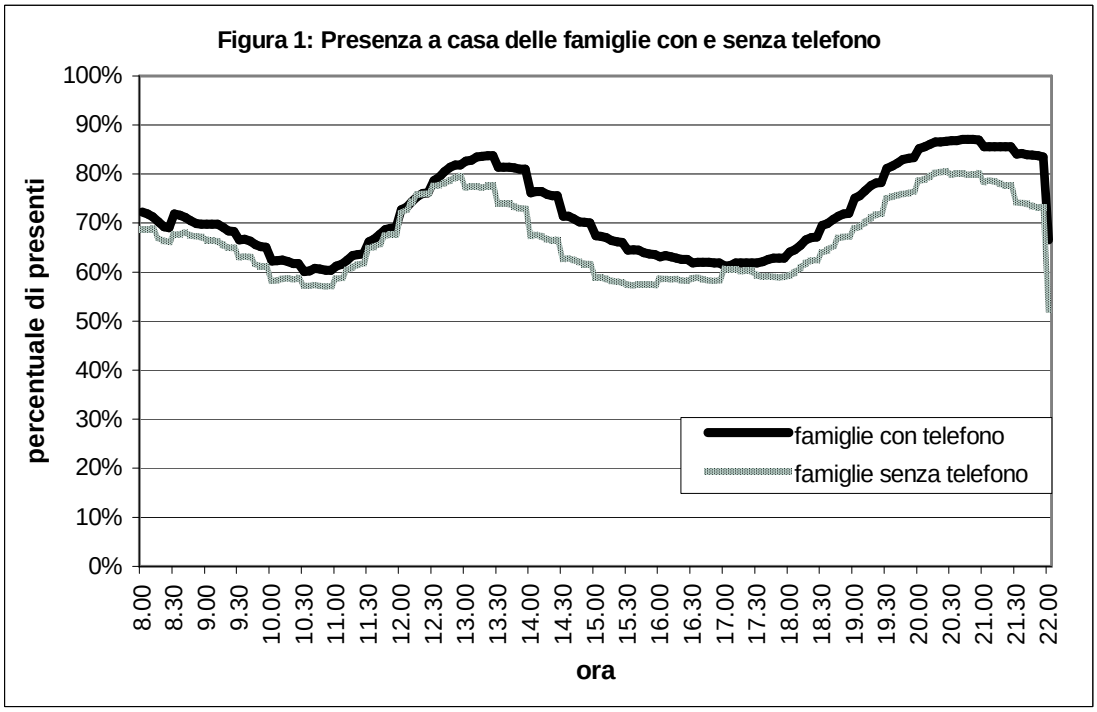
Le stime delle probabilità di presenza a casa danno luogo ad un ordinamento delle priorità di chiamata: in linea di massima, si darà la precedenza in ogni fascia oraria alle unità cui sono attribuite maggiori probabilità di trovare qualcuno a casa.

2. Quadro generale delle distribuzioni di presenza in casa

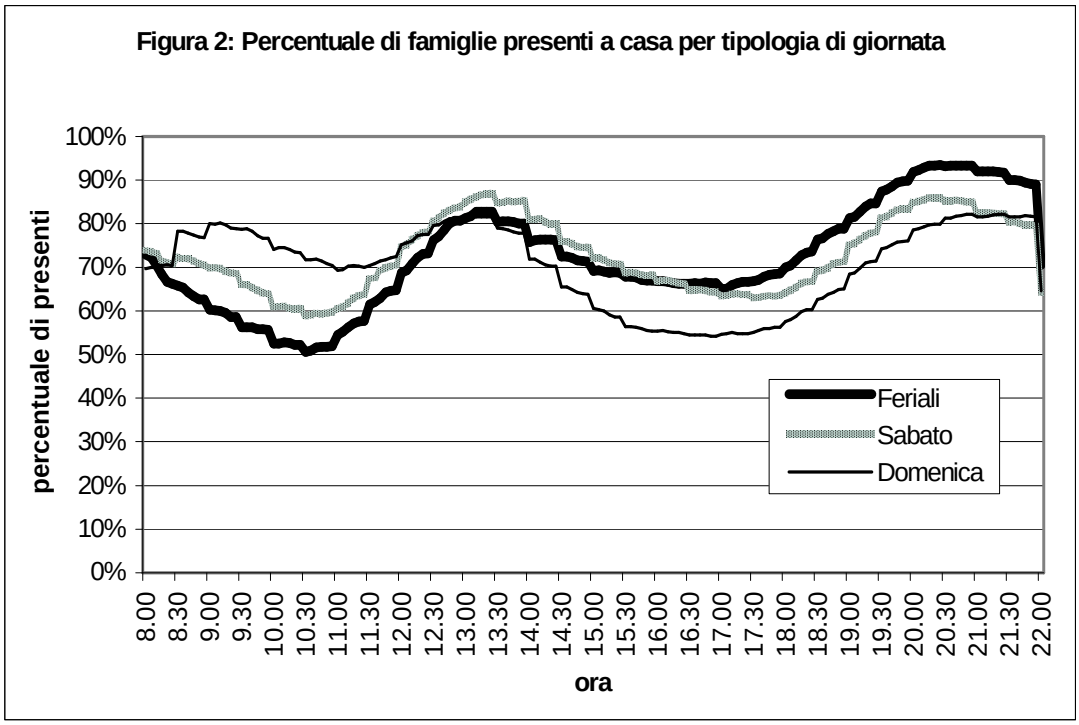
I dati che si analizzano in questo paragrafo e nel successivo sono tratti dai diari sull'uso del tempo allegati al II e III ciclo dell'indagine multiscopo sulle famiglie (ISTAT, 1988). Il campione comprende 13730 famiglie per un totale di 39286 persone, ed è rappresentativo della popolazione nazionale con l'esclusione dei bambini sotto i 3 anni di età e di coloro che vivono in convivenze. L'indagine sull'uso del tempo è stata effettuata tramite diari giornalieri autocompilati che consentono di ricostruire tutte le attività svolte durante la giornata dalle unità oggetto di indagine.
Si confrontano innanzitutto le dinamiche di presenza in casa delle famiglie provviste di telefono con quelle delle famiglie senza telefono[2], quindi, per le famiglie con telefono, si confrontano le distribuzioni di presenza per tipologia di giornata[3] e per grande ripartizione geografica[4]. Si osserva che:

a) *le famiglie in possesso di telefono risultano costantemente più presenti a casa delle famiglie che non possiedono l'apparecchio telefonico nell'abitazione* (Figura 1). Ciò lascia intendere la possibilità che alcune caratteristiche delle persone che trascorrono molto tempo fuori casa siano analoghe a quelle delle famiglie senza telefono. In questo caso, i rischi di distorsione connessi al possesso del telefono andrebbero a sommarsi a quelli legati alla difficoltà di contatto con il rispondente.

[2] Le famiglie senza telefono nel campione sono 2248, mentre 11383 famiglie ne sono provviste; le analisi successive sono svolte su quest'ultimo gruppo.
[3] Il campione è stato ripartito in 3 gruppi: 4828 famiglie hanno compilato il diario per un giorno feriale, 4483 per un sabato e 4419 per una domenica.
[4] Le ripartizioni considerate sono Nord, Centro e Sud (isole comprese); i dati relativi alla ripartizione geografica riguardano soltanto le famiglie intervistate nel II ciclo, ovvero 7056 famiglie.

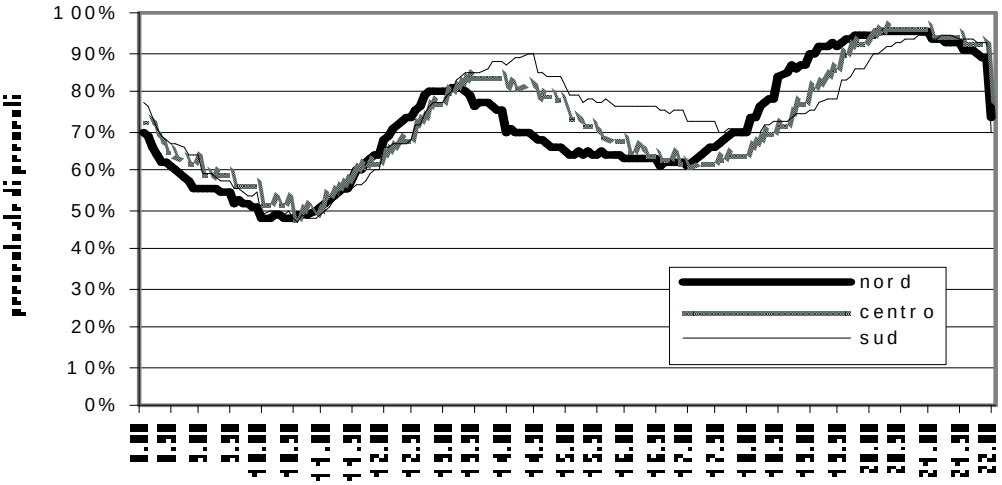


b) *le distribuzioni orarie di presenza risultano diverse per giorno della settimana: feriale, sabato e domenica* (Figura 2). Gli orari ottimali per ottenere un contatto telefonico risultano le fasce orarie del tardo pomeriggio, dell'ora di cena e della prima serata, soprattutto per i giorni feriali, mentre i peggiori risultati si ottengono la mattina dei giorni feriali e la domenica pomeriggio.



c) *vi è uno slittamento in avanti degli orari dei pasti passando da nord a sud* e, di conseguenza, uno slittamento in avanti dei picchi relativi alla presenza a casa in tali fasce (nella Figura 3 si riportano le distribuzioni inerenti ai giorni feriali).

Figura 3: Presenza a casa delle famiglie nei giorni feriali per ripartizione geografica



3. Presenza in casa secondo le caratteristiche della famiglia

Per l'analisi basata sulle caratteristiche delle famiglie si sono scelte variabili demografiche facilmente reperibili nelle indagini di tipo ufficiale: il numero di componenti della famiglia, l'età e il sesso dei componenti.
Si sono accorpate le ore in quattro fasce orarie: una fascia mattutina, una relativa all'ora di pranzo, una pomeridiana e una serale. La presenza nella fascia oraria è stata definita come presenza in almeno metà delle ore comprese nella fascia.
Poiché la variabile dipendente, vale a dire la presenza a casa della famiglia, è una variabile dicotomica, il modello utilizzato è un modello appartenente alla classe dei modelli lineari generalizzati, il modello di regressione logistica:

$$\text{logit}\left[E(\mathbf{Y} \mid \mathbf{X})\right] = \log \frac{E(\mathbf{Y} \mid \mathbf{X})}{1 - E(\mathbf{Y} \mid \mathbf{X})} = \mathbf{X}'\beta$$

dove **Y** è la variabile dipendente dicotomica osservata su n soggetti, **X** è la matrice ($n \times p$) delle variabili esplicative e β è un vettore di parametri.

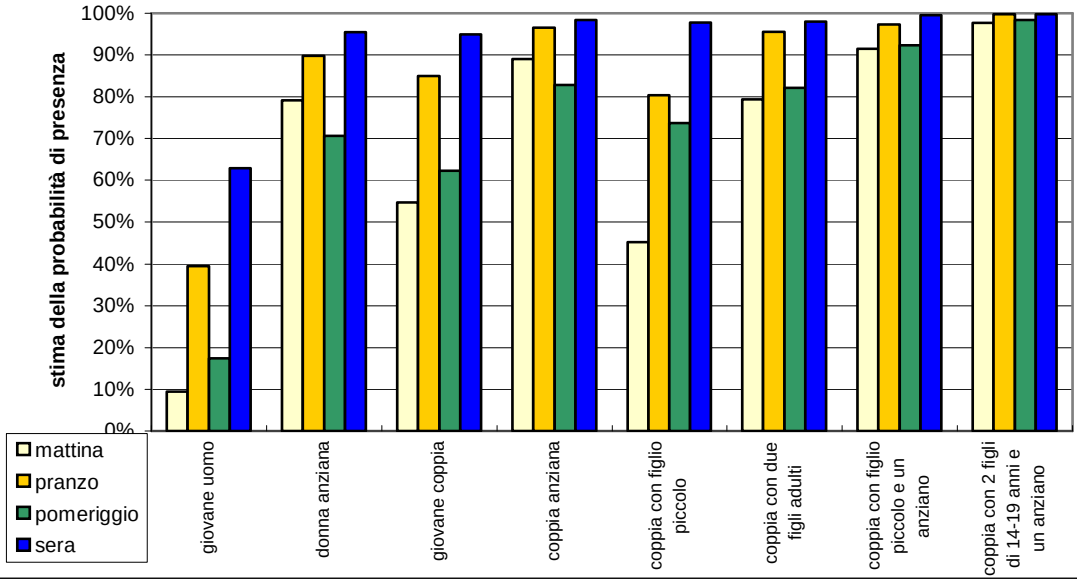
Tabella 1: *Stima dei parametri e relativi standard error (fra parentesi) di modelli logistici, per fascia oraria e giorno della settimana.*

	Giorni feriali				Sabato				Domenica			
Variabili	mattina	pranzo	pomeriggio	sera	mattina	pranzo	pomeriggio	sera	mattina	pranzo	pomeriggio	sera
Intercetta	-2,961 (0,334)	-1,387 (0,317)	-2,133 (0,293)	-0,663 (0,417)	-2,696 (0,308)	-1,763 (0,379)	-2,527 (0,289)	-1,520 (0,339)	-2,153 (0,285)	-0,585 (0,266)	-1,803 (0,266)	-0,710 (0,291)
Famiglia che risiede al Sud o nelle Isole	_____	_____	0,305 (0,135)	-0,540 (0,243)	0,337 (0,158)	0,484 (0,219)	0,242 (0,141)	_____	_____	_____	0,559 (0,150)	_____
Famiglia che risiede al Nord	_____	_____	_____	_____	0,480 (0,147)	_____	_____	_____	_____	-0,124 (0,141)	0,199 (0,132)	_____
Presenza di bambini sotto i 14 anni	-1,138 (0,131)	-1,169 (0,224)	-0,189 (0,157)	_____	-0,519 (0,145)	_____	-0,453 (0,166)	-0,885 (0,225)	-0,482 (0,183)	-1,030 (0,197)	-0,377 (0,133)	-0,621 (0,192)
Presenza di ragazzi in età 14-19	_____	0,700 (0,281)	1,158 (0,218)	0,943 (0,400)	_____	1,461 (0,423)	0,470 (0,203)	0,814 (0,303)	0,551 (0,223)	0,787 (0,251)	0,541 (0,151)	0,557 (0,230)
Presenza di giovani in età 20-29	_____	_____	_____	_____	0,385 (0,131)	0,553 (0,268)	_____	_____	_____	_____	_____	_____
Presenza di persone in età 30-64	_____	_____	_____	0,741 (0,304)	_____	1,106 (0,294)	_____	0,723 (0,211)	0,748 (0,194)	_____	_____	0,753 (0,210)
Presenza di anziani di almeno 65 anni	1,824 (1,184)	1,452 (0,273)	0,916 (0,167)	1,146 (0,364)	1,367 (0,183)	1,894 (0,347)	1,007 (0,178)	1,856 (0,305)	1,308 (0,230)	0,520 (0,201)	0,796 (0,149)	1,892 (0,300)
Presenza in famiglia di femmine adulte	2,071 (0,294)	1,672 (0,274)	1,808 (0,254)	2,459 (0,303)	1,907 (0,238)	1,384 (0,275)	1,907 (0,237)	1,517 (0,255)	1,643 (0,234)	1,115 (0,234)	1,127 (0,234)	0,971 (0,245)
Presenza in famiglia di maschi adulti	0,336 (0,164)	0,632 (0,221)	0,415 (0,171)	1,143 (0,278)	0,549 (0,175)	0,588 (0,247)	0,425 (0,181)	0,434 (0,232)	_____	_____	_____	_____
Numero di componenti della famiglia	0,484 (0,060)	0,584 (0,110)	0,351 (0,075)	_____	0,336 (0,066)	0,312 (0,101)	0,537 (0,084)	0,467 (0,110)	0,509 (0,085)	0,554 (0,091)	0,348 (0,060)	0,337 (0,089)

Le variabili esplicative introdotte nei modelli sono variabili dicotomiche 0-1, con l'eccezione del numero di componenti della famiglia, che è stata trattata come variabile quantitativa intera.
I modelli logistici sono stati ottenuti con criterio di selezione *stepwise* dei predittori (Hosmer e Lemeshow, 1989)[5]. Il metodo di stima utilizzato in questo contesto, e più in generale nei modelli lineari generalizzati, è quello della massima verosimiglianza, che stima i parametri β in modo tale da massimizzare la probabilità dell'insieme di dati osservato.

L'analisi delle stime dei parametri e dei relativi odds-ratio nei modelli relativi a ciascuna ora (Martini, 1998) ha permesso di individuare le variabili che più spiegano il fenomeno nelle diverse fasce orarie, consentendo di ignorare le variabili che non entrano in nessuno dei modelli orari compresi nella fascia, quelle che entrano con segno diverso in differenti modelli compresi nella stessa fascia, o che mostrano discontinuità nelle indicazioni sulla presenza in casa.
Nella Tabella 1 si riportano, per fascia oraria e giorno della settimana, le variabili selezionate e le stime dei parametri per i modelli ottenuti.
Inoltre, si sono stimate le probabilità di presenza per alcune categorie di famiglia (Figura 4).

Figura 4: Stime della probabilità di presenza a casa per alcuni tipi di famiglia



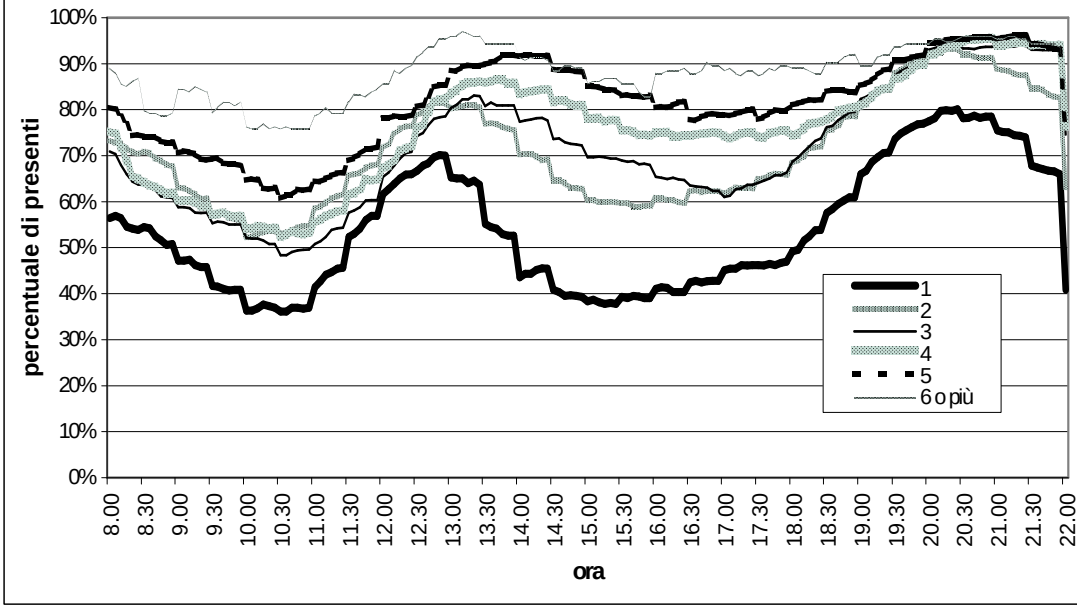
Si osservano forti differenze fra categorie di famiglie:

a) le persone che vivono sole, soprattutto se non sono anziane, risultano molto più difficili da raggiungere rispetto a famiglie numerose (la Figura 5 riporta le distribuzioni di presenza per numero di componenti relative ai giorni feriali);

[5] Il pacchetto statistico utilizzato per le analisi è SAS 6.12.

b) la presenza all'interno del nucleo familiare di donne e anziani garantisce, in ogni momento della giornata, probabilità di trovare a casa almeno un membro della famiglia decisamente maggiori rispetto a famiglie prive di tali figure;
c) la presenza in famiglia di adolescenti aumenta la probabilità di contatto in tutte le fasce orarie tranne che la mattina dei giorni feriali e del sabato, mentre la presenza di componenti nelle fasce d'età centrali si rivela importante soltanto di sera e nei giorni non lavorativi;
d) tutte le categorie di famiglia presentano picchi relativi all'ora di pranzo e all'ora di cena, ma tali picchi risultano particolarmente marcati per certe categorie, mentre le famiglie numerose, in modo particolare se vi sono anziani all'interno del nucleo, non mostrano differenze fra le stime delle probabilità di presenza in fasce orarie diverse.

Figura 5: Presenza a casa delle famiglie nei giorni feriali per numero di componenti



Forzando nei modelli di regressione il livello d'istruzione (descritto dalla variabile categoriale "Massimo titolo di studio raggiunto in famiglia"[6]) e la condizione lavorativa dei componenti della famiglia (descritta da 5 variabili dicotomiche indicatrici della presenza in famiglia di lavoratori autonomi, lavoratori dipendenti, studenti, pensionati e casalinghe) si nota che, mentre il grado di istruzione non covaria con la probabilità di presenza a casa delle famiglie (Martini, 1998), la condizione lavorativa contribuisce a spiegare la variabilità nei modelli relativi ai giorni feriali e alla fascia oraria mattutina del sabato(Tabella 2).
I modelli che comprendono le variabili relative alla condizione lavorativa dei membri della famiglia sono confrontati con i modelli che escludono tali variabili per mezzo della

[6] I titoli di studio rilevati dall'Istat sono stati raggruppati in 3 classi: "Nessun titolo", "Licenza elementare o media" e "Diploma superiore o titolo univerisitario".

percentuale di concordanze tra valori stimati e valori osservati della variabile dipendente "Presenza a casa di almeno un componente della famiglia".

Tabella 2: *Confronto fra le percentuali di concordanze fra osservazioni e previsioni nei modelli costruiti con sole variabili demografiche e nei modelli con variabili sociali.*

Giorno	**Fascia**	**Percentuale di concordanze fra valori osservati e previsti**	
		Solo variabili demografiche	*Anche condizione lavorativa*
Feriale	Mattina	73,3	86,9
	Pranzo	76,7	85,3
	Pomeriggio	72,5	79,1
	Sera	74,9	82,1
Sabato	Mattina	71,2	78,9
	Pranzo	77,7	78,9
	Pomeriggio	72,8	75,2
	Sera	75,8	77,3
Domenica	Mattina	74,1	76,4
	Pranzo	71,9	73,2
	Pomeriggio	67,4	68,9
	Sera	71,3	71,6

4. Presenza in casa secondo la storia delle chiamate

Dopo la prima chiamata telefonica è possibile trarre informazioni dagli esiti dei tentativi di contatto effettuati, al fine di congetturare su ritmi e abitudini delle famiglie, e quindi sulle loro dinamiche di presenza a casa.

I dati considerati (che fanno riferimento alle registrazioni delle chiamate relative alle reinterviste telefoniche di controllo per l'indagine sui consumi delle famiglie del 1996) contengono informazioni su ora, data e giorno di ciascun tentativo di contatto. Si tratta di 5375 tentativi di chiamata effettuati su 3096 famiglie-campione; escludendo le chiamate per cui risulta omesso l'esito e le chiamate iniziali, per le quali non è possibile fare riferimento a una "storia delle chiamate", l'analisi è svolta su 2277 chiamate successive alla prima.

Dato che le interviste sono state svolte soltanto nei giorni feriali e in orari compresi fra le 17.00 e le 19.30 (orari che non presentano differenze apprezzabili dal punto di vista degli esiti dei tentativi), non è possibile trarre informazioni sugli orari e i giorni ottimali, ma soltanto sulle probabilità di contatto conseguenti a storie delle chiamate diverse.

Il modello logistico ottenuto è il seguente[7]:

$$\text{logit } Y = -0{,}3631 - 0{,}3411\, X_1 + 0{,}2866\, X_2 + 1{,}7079\, X_3 - 1{,}0826\, X_2X_3$$

dove:

Y =	percentuale di contatti	X_3 =	esito "occupato" nel tentativo precedente
X_1 =	numero di tentativi precedenti		
X_2 =	cambiamento di fascia rispetto al tentativo precedente	X_2X_3 =	interazione fra X_2 e X_3

[7] Il pacchetto statistico utilizzato in questo caso è SPSS 7.0.

L'analisi degli esiti dei tentativi di contatto evidenzia che:

a) la maggiore probabilità di successo si registra a seguito di un tentativo con esito "occupato", mentre le probabilità di contatto sono inferiori se l'esito è una segreteria telefonica, un fax o un segnale di "libero"
b) le probabilità di ottenere un contatto diminuiscono via via che aumenta il numero di tentativi compiuti
c) dopo un tentativo di chiamata fallito, conviene richiamare in una fascia oraria diversa, a meno che l'esito non sia un segnale di "occupato", che sta, invece, ad indicare la convenienza a riprovare nella stessa fascia oraria.

5. Considerazioni prospettiche

Le strategie di ottimizzazione considerate in questa nota consistono nello stimare, in tempo reale e per ciascuna unità, la probabilità di contatto e, quindi, l'ordine di priorità delle chiamate.

Tuttavia, quando le stime delle probabilità di contatto sono basate su caratteristiche socio-demografiche, vi è la possibilità di escludere sistematicamente le unità che stanno frequentemente fuori casa dalle prime posizioni degli ordini di priorità e, qualora il periodo di rilevazione termini prima di aver completato le interviste, di escludere intere categorie di famiglie dalla rilevazione, con il rischio di distorsione.

Per ovviare a questo problema, è possibile:

a) dedicare alcune fasce orarie o alcuni giorni della settimana al tentativo di contattare le sole unità problematiche;
b) basare le priorità sulla differenza fra la probabilità di contatto nella fascia oraria corrente e quella media, al fine di privilegiare le unità per cui l'attuale fascia oraria comporta il maggior guadagno in termini di probabilità di esito favorevole;
c) combinare le variabili socio-demografiche alla storia delle chiamate nella stima delle probabilità di contatto, aggiornando così le stime delle probabilità di contatto dopo ogni nuovo tentativo.

Le strategie di pianificazione proposte sono in corso di sperimentazione …….

Ringraziamenti

Gli autori ringraziano l'ISTAT nella persona della dott.ssa Giuliana Coccia per aver gentilmente messo a disposizione i dati rilevati nelle reinterviste di controllo relative all'indagine sui consumi delle famiglie del 1996.

Riferimenti bibliografici

Forsman G., Japec L., Lundquist P., Wretman J. (1996) When to Call People in Telephone Surveys, 7th International Workshop on Household Survey Nonresponse, Rome, October 2-4, 1996

Groves R.M., Biemer P.P., Lyberg L.E., Massey J.T., Nicholls II W.L., Waksberg J. (1988) *Telephone Survey Methodology*, Wiley, New York

Hosmer D.W., Lemeshow S. (1989) *Applied Logistic Regression*, Wiley, New York

ISTAT (1988), *Indagine multiscopo sulle famiglie – Istruzioni per l'esecuzione dell'indagine sull'uso del tempo*

Martini M.C. (1998) Modelli statistici per la programmazione ottima delle chiamate in indagini telefoniche assistite da computer, Tesi di laurea in Scienze Statistiche e Demografiche, Università degli Studi di Padova

Stokes S.L., Greensberg B.S. (1990) A Priority System to Improve Callback Success in Telephone Surveys, *Proceedings of the Survey Research Methods Section*, American Statistical Association, 742-747