La codifica delle variabili testuali nel 14° Censimento Generale della Popolazione

Autori

*Francesco Cuccia **, Simone De Angelis **, Antonio Laureti Palma**,*

*Stefania Macchia *, Simona Mastroluca**, Domenico Perrone *,*

2004

* ISTAT – Servizio Metodologie e Tecniche per la Produzione Statistica

** ISTAT – Servizio Censimenti della Popolazione e delle Abitazioni

Sommario

Il documento si propone di descrivere la metodologia innovativa adottata in occasione del 14° Censimento Generale della Popolazione, inerente il trattamento delle variabili rilevate con quesiti a testo libero. Più in dettaglio, si ripercorre l'analisi preliminare, a seguito della quale è stato deciso di adottare la codifica automatica, si documentano le attività finalizzate alla costruzione delle basi informative per ciascuna variabile oggetto di codifica ed i risultati delle sperimentazioni effettuate. Viene quindi delineata la strategia organizzativa implementata per questo censimento e si descrive l'architettura software per la gestione della codifica in Istat. Da ultimo, si riportano i risultati ottenuti sia dal punto di vista quantitativo che qualitativo .

Abstract

The document has the purpose of describing the innovative methodology adopted during the 14° General Population Census, concerning the treatment of free texts variables. More in details, the preliminary analysis which led to the decision to select the automated coding technique is delineated, the activities aimed at implementing the data base for each variable to be coded are documented and the results of a series of previous tests are shown. Then, the organizational strategy set up for this Census and the software architecture for the management of the coding activity made in Istat are described too and, finally, the results are shown, both from a quantitative and from a qualitative point of view.

Indice[1]

[1] Il lavoro è frutto dell'attività di ricerca congiunta degli autori. In ogni caso, ai soli fini dell'attribuzione, i paragrafi 1.1 ed 1.2 sono da attribuirsi a D.Perrone; 1.3, 1.4, 1.5 e 5.4 a S.Macchia; il capitolo 2 e i paragrafi, 5.1 e 5.3 a S.Mastroluca; i paragrafi 3.1 e 3.2 a A.Laureti Palma; i paragrafi 3.3, da 4.2 a 4.5 ed il 5.2 a S.De Angelis; i paragrafi 4.1 e 4.6 a F. Cuccia. Il paragrafo su Conclusioni e prospettive è infine da attribuirsi a S. Macchia e S.Mastroluca.

1 Il software ACTR, gli ambienti di codifica sviluppati ed i test effettuati

1.1 1l sistema ACTR

Il sistema ACTR v3 (Automated Coding by Text Recognition), sviluppato e commercializzato da Statistics Canada, rientra tra i sistemi basati sui cosiddetti "*weighting algorithms"*. E' un software per la codifica automatica di variabili testuali di tipo generalizzato, in altre parole indipendente dalla classificazione considerata e dalla lingua in cui sono espressi i testi. Il sistema gestisce applicazioni di codifica completamente automatiche (modalità *batch*) e, con l'ausilio di un'interfaccia grafica, permette di analizzare interattivamente la codifica di casi singoli. ACTR è stato utilizzato in molteplici applicazioni nell'ambito della statistica ufficiale inclusi i censimenti (Tourigny e Moloney, 1995).

Trattandosi di un sistema generalizzato, la costruzione dell'ambiente informativo (il cosiddetto "d*izionario"*) relativo a ciascuna classificazione da trattare è a carico dell'utente. La base per la costruzione del dizionario informatizzato è il manuale ufficiale della classificazione di riferimento; quest'ultimo, però, per poter essere trattato da un software, deve essere sottoposto ad una serie di operazioni dovute al fatto che le descrizioni testuali delle classificazioni ufficiali sono state pensate per essere utilizzate dai codificatori, che possono fare deduzioni o riferirsi alle loro conoscenze personali. Dovendo invece affidarsi ad un software, è necessario fornire a quest'ultimo tutti gli elementi ai quali farebbe automaticamente riferimento una mente umana. Bisogna quindi effettuare una serie di operazioni finalizzate ad includere nei dizionari processabili esclusivamente descrizioni sintetiche, analitiche e non ambigue. Inoltre l'applicazione di ACTR richiede all'utente un'attività di adattamento alla lingua. Infatti, le regole di *parsing* (con cui si intende il processo di standardizzazione dei testi) e il dizionario della classificazione, che sono parte integrante dell'applicazione di codifica, dipendono dalla grammatica e sintassi specifiche della lingua e dal particolare tipo di risposta da codificare. Questa attività di adattamento, denominata "addestramento" del sistema, consiste nell'individuare gli opportuni correttivi al dizionario o alla strategia di *parsing* per ridurre al minimo i casi di fallimento. La performance del sistema ACTR non può quindi essere valutata "a priori", essa è necessariamente vincolata al particolare contesto applicativo e alla qualità della fase di addestramento.

La logica di base del sistema ACTR è ispirata alla metodologia sviluppata originariamente presso US Census Bureau (Hellerman, 1982) ed utilizza degli algoritmi d'abbinamento di dati testuali messi a punto successivamente da ricercatori di Statistics Canada (Wenzowski, 1988).

Anche in ACTR, come negli altri sistemi di codifica automatica, il confronto tra la risposta da codificare e le voci contenute nel dizionario della classificazione è preceduto dalla fase definita *"parsine"*; tale fase è completamente controllata dall'utente che ha il compito di adattarla al particolare contesto applicativo (lingua, classificazione, tipologia di rispondente). La peculiarità di ACTR rispetto ad altri sistemi è che, mettendo a disposizione fino a 14 diverse funzioni di *parsing*, consente una notevole flessibilità e possibilità di personalizzazione del processo di standardizzazione.
Il dizionario della classificazione, è sottoposto al processo di *parsing* e caricato in un database di sistema, allo scopo di ottimizzare i tempi d'accesso e di ricerca. Gli eventuali duplicati ed i casi d'incongruenza (codici differenti associati alla stessa descrizione testuale) della classificazione sono segnalati ed esclusi dal database. In questa fase vengono anche calcolati i pesi (P) associati alle singole parole contenute nel dizionario, che sono proporzionali all'informatività

della parola stessa, ovvero alla sua capacità di discriminare univocamente un codice:

$$P(W_i) = 1 - \log(Nw_i)/\log(N) \qquad (1)$$

dove Nw_i è il numero di codici contenenti la parola i-esima (W_i) mentre N è il numero totale di codici contenuti nel database. Nella fase di codifica vera e propria la risposta testuale è confrontata con il database alla ricerca di un abbinamento esatto (*direct match*), ovvero del testo che abbia tutte le parole in comune con la risposta, che dà luogo inequivocabilmente all'assegnazione di un codice unico. Se il tentativo fallisce viene ricercato un abbinamento parziale (*indirect match*), in questo caso il software individua, tramite una misura della similarità tra testi (S) di tipo empirico, "il codice" o "i codici" del dizionario con descrizione più simile alla risposta fornita dal rispondente. Tale misura è funzione del numero di parole in comune e del loro grado d'informatività all'interno del dizionario di riferimento:

$$S = 10\,(a + 2b)/3 \qquad (2)$$

$$a = 2\,N_C/(N_R + N_D)$$

$$b = 2\Sigma i\, P(W_i^C)/\,(\Sigma j\, P(W_J^R) + \Sigma j\, P(W_J^D))$$

dove N_C è il numero di parole in comune tra risposta e testo del dizionario, N_R e N_D rappresentano il numero totale di parole contenute rispettivamente nella risposta e nel testo del dizionario, mentre $P(W_i^C)$, $P(W_J^R)$ r $P(W_J^D)$ rappresentano i pesi definiti nella (1) rispettivamente della i-esima parola in comune (W_i^C), della j-esima parola contenuta nella risposta testuale (W_J^R) e della j-esima parola contenuta nel testo del dizionario (W_J^D). La misura di similarità espressa nella (2) assume valori compresi nell'intervallo [0,10] i cui estremi corrispondono ad un abbinamento testuale nullo (S=0) o ad un abbinamento esatto (S=10). Il sottoinsieme dei testi del database con almeno una parola in comune con la risposta vengono ordinati per misura S decrescente (S1>S2>…>Sn). La regione di accettazione per la misura di similarità è data dalle relazioni (3) ed è costruita utilizzando tre parametri soglia: Smin, Smax e ΔS, che rappresentano rispettivamente le soglie minima e massima di accettazione, e la minima distanza richiesta tra testo a punteggio massimo (S1) e successivo (S2). I possibili risultati del sistema ACTR si suddividono quindi in:

$S_1 > Smax$ e $(S_1 - S_2) > \Delta S$ (codice *unico*) (3a)

$S_1 > Smax$ e $(S_1 - S_2) \leq \Delta S$ (codici *multipli*) (3b)

$Smin < S_1 \leq Smax$ (codici *possibili*) (3c)

$S_1 \leq Smin$ (casi *falliti*) (3d)

Se è soddisfatta la condizione (3a) la voce del dizionario a punteggio massimo (S1) è dichiarata "vincente", il codice che le è associato è unico e viene assegnato in modo completamente automatico. I rimanenti casi necessitano, invece, della valutazione da parte di codificatori (o di programmi ausiliari) che selezionino il codice corretto tra quelli proposti dal sistema. I valori dei parametri soglia sono fissati dall'utente in funzione dei suoi obiettivi di qualità: valori alti

elevano l'accuratezza (percentuale di codici unici corretti) dei risultati a scapito dell'efficacia (percentuale di codici unici assegnati); quindi la scelta dei valori ottimali si gioca sul bilanciamento tra questi due aspetti della qualità dei risultati.

Tra le funzionalità più avanzate del sistema ACTR si citano l'uso dei campi "filtro" e dei "contesti multipli" di codifica. I campi filtro possono essere utilizzati per restringere la ricerca del codice all'interno di sottoinsiemi di classi specificate dall'utente. Il sistema ACTR consente inoltre di gestire applicazioni in cui la ricerca del codice è effettuata su più ambienti (costituiti ciascuno da un dizionario e dai relativi file di *parsing*) analizzati in cascata (se la ricerca fallisce nel primo contesto si passa al secondo e così via). L'impiego di "contesti multipli" di codifica è particolarmente utile nella codifica di testi bilingui, ma si estende a tutte le situazioni in cui sia possibile e vantaggioso trattare con più metodologie alternative i casi non risolti dal sistema (ad esempio con dizionari ausiliari).

1.2 La costruzione degli ambienti di codifica

Sono stati predisposti gli ambienti applicativi per cinque variabili:

- Professione
- Attività Economica
- Titolo di studio
- Comune
- Stato Estero/Cittadinanza.

La base per la costruzione degli ambienti, per ciascuna di queste variabili, è, ovviamente, il manuale ufficiale di ciascuna classificazione, opportunamente rielaborato per renderlo gestibile dal software e per avvicinarlo al modo di esprimersi dei rispondenti.

Si è proceduto quindi a:

- Semplificare le descrizioni → spesso descrizioni che riassumono più di un concetto sono associate a singoli codici, mentre tipicamente il rispondente si riferisce a concetti singoli. Per esempio: la nostra classificazione delle professioni assegna un unico codice a "matematici e statistici", mentre il rispondente potrebbe rispondere soltanto "matematico", oppure soltanto "statistico", a seconda della sua specializzazione. In casi come questo, è necessario suddividere la frase in due o più descrizioni ed associarle allo stesso codice.
- Definire i sinonimi → in questo contesto, per sinonimi si intende sia l'equivalenza di parole che esprimono lo stesso concetto (ad es.: le parole "scarpe" e "calzature"), sia termini specifici che la classificazione in oggetto riconduce a un termine generico. Infatti, le classificazioni contengono spesso parole generiche che si riferiscono a categorie, mentre i rispondenti possono utilizzare parole specifiche. Per esempio: al posto di "agricoltore di cereali", il rispondente potrebbe rispondere "agricoltore di grano". E' stato quindi necessario esplicitare la lista dei termini puntuali riconducibili al concetto generico.
- Rielaborare le classi aperte → abitualmente le classificazioni prevedono descrizioni del tipo "Altri...", intendendo "qualcos'altro rispetto ai concetti già specificati". Per esempio, "Altri operai specializzati", laddove, differenti tipi di operai specializzati sono già stati elencati in precedenza. Si è tentato quindi di elencare i concetti ai quali le classi aperte si riferiscono.
- Eliminare le clausole di escluso → chiaramente un sistema di codifica automatica non può ragionare in termini di esclusione, quindi non può capire il significato di "escluso…", "a parte…" ed altre clausole affini che escludono determinate categorie da alcune classi. Sono state quindi eliminate queste clausole dai dizionari ed è stato verificato che i concetti "esclusi" da una classe fossero inclusi in un'altra.

- Integrare i dizionari con materiale di riferimento → i dizionari processabili sono stati ampliati con descrizioni provenienti da note esplicative degli stessi manuali delle classificazioni oppure da altre classificazioni correlate. Per esempio, ogni volta che la classificazione delle Attività economiche cita un elemento riguardante la produzione di una "certa categoria di prodotti" (che riassume una lista implicita) è stata utilizzata la classificazione dei prodotti per elencare esplicitamente la lista di prodotti.

Il passaggio successivo per l'arricchimento dei dizionari è quello che utilizza come fonte le risposte empiriche precodificate fornite nell'ambito di precedenti indagini nelle quali è stato rilevato il fenomeno. Questa attività è stata effettuata per quelle variabili per le quali le risposte testuali erano state registrate e quindi erano direttamente elaborabili. A tale proposito, è stato verificato che il valore aggiunto fornito da questa operazione è stato significativo in termini di successo della codifica, in quanto è questa la fonte che maggiormente tiene conto del modo di esprimersi fornito dai rispondenti. In tale fase, inoltre, sono stati inclusi testi contenenti i più frequenti errori ortografici verificatisi nel corso di precedenti rilevazioni.

Infine è prassi che ulteriori integrazioni vengano effettuate durante l'utilizzo delle stesse applicazioni di codifica, analizzando in corso d'opera i testi ai quali il sistema non ha assegnato un codice per carenze dei dizionari stessi ed utilizzandoli per l'arricchimento degli ambienti applicativi.
Nell'ambito delle attività propedeutiche alla codifica del censimento, tale attività è stata effettuata sia nel corso delle indagini pilota che durante le applicazioni per la codifica delle "convivenze" (cfr. par. 5.3).

Nella seguente tabella sono riportate le dimensioni dei dizionari per la codifica, evidenziando il numero di descrizioni presenti originariamente nei manuali ufficiali delle classificazioni ed il numero di descrizioni conseguente alle diverse fasi di arricchimento dei dizionari stessi.

Tabella 1.1 Dimensione dei dizionari conseguente alle diverse fasi di arricchimento

Classificazione	*Numero di descrizioni del manuale ufficiale della classificazione*	*Numero di descrizioni del dizionario elaborabile*
Professione	6.300	20.213
Attività Economica	1.668	27.306
Titolo di studio	915	2.291
Stato estero/Cittad.	212	4.645
Comune	8.091	62.397

Si entra ora in dettaglio sulle fonti utilizzate per la costruzione dei diversi ambienti applicativi per ciascuna classificazione.

Professione: il manuale utilizzato è quello della classificazione Istat 2001 (CP01). Soltanto la rielaborazione di questa fonte ha portato ad un incremento nel dizionario di circa 1000 descrizioni. Sono inoltre state utilizzate le risposte fornite in alcune indagini, nel corso delle quali il sistema di codifica ha costituito un'innovazione nel processo di gestione dell'indagine, oppure è stato utilizzato sperimentalmente:

- Indagine di qualità del censimento della popolazione del 1991
- Indagine sulla salute (1994)
- Indagine sulle Forze di Lavoro (4 trimestri 1998)
- I e II indagine pilota del censimento della popolazione del 2001.

Attività economica: il manuale utilizzato è quello della classificazione ATECO1991; in questo caso, la rielaborazione del manuale ha portato ad un dizionario di 6273 descrizioni. Questo incremento è dovuto al fatto che i testi originari associati a ciascun codice erano estremamente complessi e facevano spesso riferimento a concetti esplicitati in corrispondenza di altri codici, per cui la rielaborazione ha prodotto un elevato numero di testi brevi e semplici associati a ciascun codice. Inoltre sono state utilizzate le risposte fornite nel corso di una serie di indagini, per le quali, come per la Professione, il sistema di codifica ha costituito un'innovazione nel processo di gestione dell'indagine, oppure è stato utilizzato sperimentalmente:

- Indagine di qualità del censimento della popolazione del 1991
- I e II indagine pilota del censimento della popolazione del 2001
- Censimento intermedio dell'Industria, indagine Short Form
- Censimento intermedio dell'Industria, indagine Long Form.

Titolo di studio: il manuale utilizzato è quello della classificazione Istat 1998/99, coerente con l'ultima versione della International Standard Classification of Education (ISCED 97). Sono inoltre state utilizzate le risposte fornite nelle seguenti indagini, nel corso delle quali il sistema di codifica è stato utilizzato sperimentalmente:

- Indagine di qualità del censimento della popolazione del 1991
- I e II indagine pilota del censimento della popolazione del 2001

Stato estero/cittadinanza: la classificazione utilizzata è quella Istat 2001. Sono inoltre state inserite nel dizionario una serie di stringhe concernenti:

- alcune denominazioni di uso locale (per es. Bosnia i Hercevogina per Bosnia-Erzegovina, Al Magrib per Marocco, Rossijskaja Federacija per Federazione Russa);
- alcune denominazioni non più in uso (per es. ex Birmania per Myanmar, ex Zaire per Repubblica Democratica del Congo, ex Ceylon per Sri Lanka);
- alcune denominazioni alternative a quella ufficiale Istat (per es. Russia Bianca per Bielorussia);
- alcune denominazioni riferite a paesi soppressi in seguito ad unificazione, a cui è stato attribuito il codice dell'attuale paese di appartenenza, per esempio: DDR / RDT / Repubblica Democratica Tedesca con il codice della Germania, Yemen del Nord / Yemen del Sud con il codice dello Yemen, ecc.
- alcune denominazioni riferite a territori, possedimenti o parti degli Stati Esteri;
- il nome della capitale di ciascuno stato e di alcune altre città;
- i codici ISO a una, due o tre lettere;
- gli aggettivi maschile e femminile di cittadinanza.

Infine sono stati predisposti altri quattro dizionari che il sistema consulta 'a cascata' , uno per ciascuna delle seguenti lingue:

- inglese
- francese
- tedesco
- spagnolo.

Comune: la classificazione utilizzata è quella Istat 2001. Nel dizionario i codici associati alla dizione di ciascun comune comprendono sia il codice del comune che della provincia corrispondente. Il numero di descrizioni contenute nell'applicazione è così elevato perché sono state utilizzate una molteplicità di fonti che hanno consentito di disporre di una base informativa che tenga conto non soltanto del modo in cui i rispondenti esprimono il nome del comune di pertinenza, ma anche di uno "storico" di questa variabile. Sono infatti stati predisposti tre dizionari che il sistema

consulta a cascata che contengono rispettivamente:

Dizionario 1:

- attuale denominazione degli 8100 comuni italiani, a volte opportunamente rielaborate (es. per i comuni relativamente ai quali denominazione ufficiale prevede due nomi, per ottimizzare la ricerca, alla dizione ufficiale sono state aggiunte, ovviamente con lo stesso codice, due descrizioni, una per ciascun nome, associate, ovviamente allo stesso codice).
- vecchia denominazione di alcuni comuni ossia di quelli che dal 1991 ad oggi hanno cambiato nome (fermi restando il codice provincia e il codice comune)
- denominazioni dei comuni ricavate dalle risposte fornite nell'ambito di una precedente indagine ISTAT
- denominazioni doppie, una in italiano e una in tedesco, per i comuni in provincia di Bolzano.

Dizionario 2:

contiene comuni non più esistenti in quanto aggregati ad uno o più comuni oppure passati ad uno Stato Estero, comuni che hanno cambiato provincia e rioni. In maggior dettaglio le informazioni presenti sono:

- denominazioni di comuni soppressi ed aggregati ad un altro comune
- denominazioni di comuni soppressi ed aggregati a più comuni
- denominazioni di comuni passati ad un altro Stato (si tratta di comuni ubicati al confine, ceduti generalmente allo stato della ex Jugoslavia);
- comuni che hanno cambiato provincia in seguito all'istituzione di nuove province o al cambiamento di alcune sigle di provincia (per ottimizzare la codifica è stato previsto di inserire anche i comuni con la vecchia sigla della provincia, fermo restando il codice numerico corrispondente alla provincia attuale)
- denominazioni dei rioni, ossia dei quartieri che compongono le grandi città.

Dizionario 3: contiene le denominazioni delle località.

1.3 Test delle applicazioni di codifica e risultati ottenuti in precedenti indagini

Come già accennato, in fase di costruzione degli ambienti di codifica, i dati rilevati nel corso di diverse indagini sono stati utilizzati per testare i risultati ottenibili dalle applicazioni di codifica delle diverse variabili e, nello stesso tempo, per arricchire i dizionari con le descrizioni che, nonostante avessero un contenuto informativo sufficiente per l'attribuzione di un singolo codice, il sistema non era riuscito a codificare a causa, prevalentemente, di carenze dei dizionari stessi.

I risultati dei test sono stati valutati in termini di efficacia, nota in letteratura come *"Recall rate"* (percentuale di testi codificati automaticamente sul totale dei testi da codificare) ed accuratezza, nota in letteratura come "P*recision rate"* (percentuale di testi codificati correttamente sul totale di testi codificati automaticamente).

Tabella 1.2: Numero di descrizioni da codificare per ciascun campione utilizzato per i test degli ambienti di codifica automatica e risultati in termini di recall rate e precision rate

Variabile	***Survey***			
	CP/IQ	*Salute (1994)*	*Forze di lavoro (4 trimestri 1998)*	*Censimento Intermedio dell'industria (Ind. Short form)*
Professione				
Num. di testi	5.869	33.735	356.231	
Recall	72,5	72,3	72,0	
Precison	90,0	97,0	97,3[2]	
Attività Economica				
Num. di Testi	6.288			1.793
Recall	54,5			47,0
Precison	85,0			88,2
Titolo di studio				
Num. di Test	2.169			
Recall	86,6			
Precison	99,7			

La prima esperienza è stata effettuata utilizzando come benchmark file un campione di 9.000 famiglie del Censimento della Popolazione del 1991, successivamente intervistate per l'indagine di qualità sul censimento stesso (*CP/IQ).* Questo campione è stato molto utile per l'aggiornamento delle applicazione di codifica, perché consentiva di individuare immediatamente il codice corretto con cui confrontare il codice attribuito da ACTR, senza dover ricorrere al giudizio del codificatore esperto (ciò avveniva ogni volta che i due codici assegnati alla stessa risposta dall'operatore del censimento e poi da quello dell'indagine di qualità coincidevano). Relativamente invece agli altri campioni utilizzati, tutti i casi di divergenza tra i codici assegnati da ACTR e quelli attribuiti manualmente sono stati sottoposti a codificatori esperti.

I risultati ottenuti sono stati considerati incoraggianti, sia rispetto all'efficacia che all'accuratezza. Per quanto attiene il *recall rate*, infatti:

- i risultati sono coerenti, con quelli ottenuti da altri uffici di statistica (Lyberg and Dean, 1992);
- relativamente alla Professione, i risultati hanno dimostrato che un'applicazione costruita utilizzando in primis un campione molto piccolo ha risposto bene anche quando utilizzata per codificare un campione di dimensioni decisamente maggiori;
- relativamente invece all'Attività Economica, le performance sono state inferiori rispetto alla Professione; ciò è principalmente imputabile alla difficoltà degli individui intervistati nelle indagini sulle famiglie a comprendere il significato della variabile richiesta e alla conseguente fornitura di risposte generiche e/o imprecise. Quindi, poche di queste risposte sono state utilizzate per arricchire il dizionari, il che ha influito anche sui risultati ottenuti successivamente sull'indagine *short form* del censimento intermedio dell'industria, che ha invece costituito un fonte molto ricca per integrare l'applicazione.

Riguardo il *precision rate* si specifica che:

- i risultati sono positivi, visto che, per tutte le variabili, sono superiori a quelli ottenibili con la codifica manuale (De Angelis R., Macchia S. and Mazza L., 2000)

[2] Effettuata a campione (Macchia, D'Orazio, 2000)

- la stima dell'accuratezza della codifica della Professione rilevata dall'indagine sulle *Forze di lavoro* è stata effettuata a campione, non essendo possibile esaminare una così grande mole di dati, adottando un disegno campionario, che ha costituito poi la base della metodologia implementata per i dati del censimento della popolazione (cfr. par. 1.5).

A seguito delle sperimentazioni descritte, la codifica automatica ha iniziato ad entrare come innovazione di processo nell'ambito di alcune indagini.
Ci si riferisce in particolare a:

- Censimento intermedio dell'industria (Indagine Long Form)
- Indagine pilota sulle forze di lavoro.

L'adozione della codifica automatica nel corso del censimento intermedio dell'industria (*Long Form*) è stata senz'altro la prima esperienza di notevole rilievo. Alle imprese veniva richiesto di segnalare il proprio settore di Attività Economica, qualora fosse diverso da quello prestampato sul modello di rilevazione; i testi da sottomettere a codifica automatica sono stati 70.236. L'efficacia è stata in media del 58,8%, variando dal 49,8% ed il 67,7%; questa variazione è derivata dal fatto che l'ambiente è stato arricchito in corso d'opera, utilizzando gli output di ciascun passaggio di codifica. Al termine di questa attività, infatti, il *recall rate* è salito al 70%.
I risultati della codifica automatica ottenuti sui dati della prima indagine pilota sulle Forze di lavoro (all'epoca l'indagine era in fase di ristrutturazione) sono riportati nella tabella sottostante. In particolare, sono stati intervistati due campioni indipendenti di 1.000 famiglie ciascuno, l'uno intervistato in modalità CATI (Computer Assisted Telephone Interviewing) e l'atro CAPI (Computer Assisted Personal Interviewing).

Tabella 1.3: Risultati della codifica automatica sui dati dell'indagine pilota sulle Forze di Lavoro

	Recall (%)	*Precision (%)*
Professione	66,7	99,0
Attività Economica	43,5	85,0

Come si può vedere, i risultati in termini di accuratezza sono coerenti con quelli ottenuti in fase di test; relativamente all'efficacia sono lievemente inferiori. Ciò può essere imputabile alla diversa tecnica di rilevazione adottata (soprattutto nel caso del CATI, la necessità di scrivere velocemente la risposta testuale può aver influito negativamente) e ad una formazione degli intervistatori non specifica sulle classificazioni.
In sintesi, quindi, è stato concluso che, anche se con la codifica automatica, non si riesce ad assegnare un codice a tutti i testi da codificare, la qualità dei risultati ottenibili è sempre superiore a quelli della codifica manuale; a questo si aggiungono gli inopinabili vantaggi in termini di tempo e risorse da dedicare. Ciò ha portato a ritenere fattibile l'adozione di questa tecnica per il Censimento della Popolazione e a pianificare ulteriori test da effettuarsi sui dati delle due indagini pilota.

1.4 I risultati della codifica automatica nelle due indagini pilota del Censimento della Popolazione

Nel corso della prima indagine pilota, svoltasi nell'ottobre 1998, sono state sperimentate due innovazioni tecnologiche: la lettura ottica e la codifica automatica. Per entrambe queste tecnologie è stato previsto in parallelo un processo tradizionale (*data entry* e codifica manuale). I quesiti relativi a Professione ed Attività Economica sono stati strutturati come segue:

- un quesito a risposta chiusa in cui era richiesto all'intervistato di individuare il ramo di pertinenza
- un quesito con risposta a testo libero.

Sono state sottoposte a codifica le variabili: Professione, Attività Economica e Titolo di studio.

Le applicazioni di codifica per le prime due variabili citate sono state impostate in modo da effettuare una prima ricerca nei dizionari condizionata dal quesito sul ramo (codifica con "filtro") e, in caso di impossibilità di individuare un codice, una seconda ricerca su tutto il dizionario.

Si riportano nella seguente tabella i risultati ottenuti in termini di efficacia ed accuratezza, quest'ultima messa a confronto con la codifica manuale.

Tabella 1.4: Risultati della codifica sulla Prima Indagine Pilota del Censimento della Popolazione

	Recall della codifica automatica (%)	*Precision della codifica automatica (%)*	*Precision della codifica manuale (%)*
Titolo di studio	75,7	99,7	73,5
Professione	65,5	98,1	64,5
Attività Economica	51,2	93,7	56,1

Anche a seguito di questo test, i risultati in termini di efficacia hanno confermato quelli ottenuti nel corso delle precedenti applicazioni e il livello di accuratezza della codifica automatica si è dimostrato notevolmente superiore a quello della codifica manuale.

Un aspetto piuttosto problematico, limitatamente a Professione ed Attività Economica, è stato invece quello della coerenza tra codice assegnato e ramo selezionato nel quesito a risposta chiusa. E' stata infatti rilevata incoerenza nel 31,6% dei casi per la Professione e nel 43,9 per l'Attività Economica. Ciò è stato imputato principalmente a due fattori: innanzi tutto alla complessità delle due classificazioni, che rende comunque difficile l'individuazione di un ramo nell'ambito del quale ritrovare la propria Professione e/o Attività Economica; in secondo luogo, la strutturazione che è stata data ai due quesiti pre-codificati che aveva portato, probabilmente, ad un eccessivo livello di dettaglio (35 modalità per la Professione e 30 per l'Attività Economica). Se a tutto ciò si aggiunge la scelta che era stata effettuata di non inserire i due elenchi di modalità di risposta di questi quesiti nel questionario (per non appesantirlo in funzione della lettura ottica), ma di consegnarli come allegato, si spiegano anche gli elevati tassi di non risposta ai due quesiti (36% per Professione e 45,7% per Attività Economica).

Relativamente invece alla lettura ottica, ne è stato misurato l'impatto rispetto alla codifica automatica, sottomettendo ad ACTR anche i dati acquisiti con questa tecnologia. I risultati in termini di *recall* sono stati molto inferiori a quelli ottenuti sui dati acquisiti con il *data entry* tradizionale (tab. 1.4), ma questo è stato imputato principalmente al fatto che l'applicazione messa a punto per la lettura ottica, sperimentata per la prima volta su questo questionario, avrebbe dovuto essere maggiormente personalizzata in modo tale da ridurre il tasso di errore nel riconoscimento dei caratteri alfabetici. E' stato quindi messo a punto un *error profile* (Balestrino R., Reale A., 2000) di cui si è poi tenuto conto nella definizione del progetto censuario.

In sintesi, a seguito di questa pilota l'idea di abbandonare la codifica manuale a favore della codifica automatica è diventata una scelta strategica. Con la seconda indagine pilota, quindi, ci si è posti l'obiettivo di ottimizzare il processo; la prima attività è stata quella di ristrutturare i due quesiti pre-codificati di Professione ed Attività Economica, riducendo il numero di modalità di risposta (10 per la prima e 29 per la seconda), inserendole nel questionario, anziché rimandare ad un allegato e semplificando la terminologia adottata.

E' stato inoltre deciso di utilizzare la codifica automatica per tutte le variabili rilevate a testo libero, per cui si è provveduto ad implementare gli ambienti applicativi di altre due variabili: Comune e Stato Estero/Cittadinanza (cfr. par. 1.2).

La reimpostazione dei due quesiti pre-codificati ha portato ad una riduzione dei tassi di non risposta (scesi al 26,8% per Professione ed a 30,5% per Attività Economica). Relativamente all'incoerenza tra questi quesiti ed i codici assegnati dal sistema alle risposte testuali, è stato quantificato soltanto quello dell'Attività Economica, che si è attestato sul 23,7%. Relativamente alla Professione, non è stato possibile effettuare questa valutazione perché, coerentemente con quanto si sarebbe successivamente attuato per il Censimento, il quesito pre-codificato è stato definito in funzione della classificazione ISCO 88 COM, in modo tale che i dati da diffondere in prima battuta fossero già confrontabili a livello internazionale, mentre non era ancora pronta la nuova classificazione nazionale (CP 2001 - che, al quarto digit, garantisce una completa confrontabilità con la classificazione internazionale), rispetto alla quale allineare l'ambiente di codifica.

In merito ai risultati della codifica automatica (cfr Tab. 1.5), è evidente come la complessità delle classificazioni di riferimento e la variabilità linguistica dei rispondenti influenzino soprattutto l'efficacia; laddove infatti la variabilità linguistica è inferiore e c'è minore margine di opinabilità sull'interpretazione da attribuire alle risposte i *recall rate* sono più elevati (oltre il 94% per il Comune) e i *Precision rate* raggiungono il 100%.

Tabella 1.5: Risultati della codifica sulla Seconda Indagine Pilota del Censimento della Popolazione

	Recall (%)	*Precision (%)*
Titolo di studio	87,0	99,0
Professione	68,8	96,8
Attività Economica	51,9	90,0
Stato estero	83,2	100,0
Comune	94,5	100,0

A seguito di queste sperimentazioni, è stata confermata l'adozione per il Censimento della Popolazione della codifica automatica per tutte le variabili rilevate a testo libero. Le attività sono quindi proseguite nell'ottica di semplificare ulteriormente i quesiti pre-codificati di Professione ed Attività Economica, di aggiornare l'ambiente applicativo della Professione con la nuova classificazione e di arricchire ulteriormente i dizionari con testi che il sistema non è riuscito a codificare nel corso delle due indagini pilota.

1.5 La metodologia per il controllo di qualità della codifica

Il controllo di qualità della codifica è indubbiamente una fase essenziale da mettere a regime nell'ambito delle attività censuarie. E' evidente che, data la mole di dati da trattare, non sarebbe stato possibile sottoporre ai codificatori esperti la totalità dei testi codificati automaticamente, così come è stato fatto nella maggioranza dei test. La qualità doveva quindi essere controllata a campione e, per la definizione del disegno di campionamento, si è tenuto conto di principalmente di due esigenze:

- si sarebbe dovuto garantire l'esattezza della codifica delle risposte più frequenti
- sottomettere più volte ai codificatori esperti testi uguali, forniti da rispondenti diversi, sarebbe stato un dispendio di tempo e di risorse.

E' stato quindi definito un disegno campionario (Macchia S., D'Orazio M. 2002), già sperimentato nel corso del test dell'applicazione di codifica automatica della Professione su quattro trimestri delle Forze di Lavoro, che, piuttosto che tenere conto di variabili territoriali, si basa sulla frequenza con cui i testi "uguali" sono stati rilevati; in pratica si utilizza un campione casuale di testi stratificato in funzione della classe di frequenza degli stessi.

Il primo passaggio è quindi quello identificare i testi "uguali", laddove per "uguali" non si intende "identici", perché ciò significherebbe considerare "diversi" testi che si differenziano tra loro per elementi non significanti, quali articoli, preposizioni o elementi correlati con genere e numero. A tal fine, i testi vengono sottoposti ad un "*parsing grezzo*", che si limita a depurarli dai già citati elementi non significanti e, soltanto a questo punto, vengono conteggiati i testi 'diversi' tra di loro, secondo questa accezione. Vengono quindi suddivisi per classi di frequenza e, nell'ambito di ciascuno strato (che si fa corrispondere con la classe di frequenza), si estrae un campione casuale semplice (senza ripetizione). La numerosità campionaria viene calcolata indipendentemente da strato a strato (Cochran, 1977), in funzione del *precision rate* atteso per ciascuno strato. Quest'ultimo è stato fissato uguale per ciascuno strato, mentre il margine di errore viene impostato in modo da decrescere progressivamente per le classi di frequenza più elevate. Ciò garantisce stime maggiormente accurate per i testi 'diversi' più significativi (che pesano di più in termini di frequenza).

Si riportano, a titolo esemplificativo, i risultati dell'applicazione di questa metodologia ottenuti sui dati della citata indagine sulle Forze di Lavoro: partendo da un campione di 356.207 testi originari, sono stati codificati automaticamente 256.113, che, a seguito del "*parsing grezzo*" sono risultati corrispondere a 19.404 testi "diversi" tra di loro. In questa esperienza è stato deciso di non sottoporre a verifica dei codificatori esperti i testi codificati con punteggio 10, ossia quelli che, a seguito del *parsing*, erano identici ai testi del dizionario elaborabile (cfr. par. 1.1). L'universo di cui misurare l'accuratezza, si riduceva quindi a 13.821 testi.
Sono quindi state definite le classi di frequenza riportate nella seguente tabella e, in funzione del *precision* rate stabilito e dei margini di errore per ciascuno strato, è stata calcolata una numerosità campionaria di 938 testi.

Tabella 1.6: Campione estratto per l'analisi di qualità della codifica dell'indagine sulle Forze di Lavoro 1998

Classi di frequenza	*Numero di testi 'differenti'*	*Precision rate atteso*	*Margine di errore*	*Numerosità campionaria degli strati*
1	10.007	75,0%	±5,0%	148
2	1.756	75,0%	±5,0%	138
3–5	1.187	75,0%	±4,5%	160
6–10	473	75,0%	±3,0%	222
11–50	349	75,0%	±2,5%	221
51–100	33	75,0%	±1,0%	33
101–1.000	16	75,0%	±1,0%	16
Totale	13.821			**938**

L'applicazione di questa metodologia ai dati del censimento della popolazione, come descritto nel cap.5, ha portato alla definizioni di classi di frequenza che tengano conto delle peculiarità di ciascuna variabile; inoltre, al fine di evidenziare eventuali imprecisioni dei dizionari elaborabili, si è deciso di comprendere negli archivi dai quali estrarre ciascun campione anche i testi codificati con punteggio 10.

2 La strategia organizzativa per il trattamento delle variabili testuali

2.1 Le ipotesi di lavoro

Il 14° Censimento Generale della Popolazione e delle Abitazioni che si è svolto il 21 ottobre 2001 è stato caratterizzato da innovazioni di prodotto e di processo. Per la prima volta la maggior parte dei dati rilevati, ovvero quelli relativi alle persone residenti in famiglia (56.594.021), sono stati acquisiti tramite la lettura ottica e non attraverso il tradizionale *data entry* così come la codifica delle stringhe alfabetiche contenute sia nei Fogli di Famiglia che nei Fogli di Convivenza, in passato espletata dagli operatori manuali degli 8.100 Uffici Comunali di Censimento consultando le classificazioni ufficiali afferenti alle singole variabili, è stata effettuata attraverso software di codifica automatica in parte in *outsourcing* e in parte all'interno dell'Istat.

Obiettivo era da un lato sollevare i comuni italiani, già pesantemente gravati dai numerosi compiti previsti dalla complessa macchina censuaria, da un'attività particolarmente onerosa sia in termini di tempo che di risorse umane, dall'altro trovare strade alternative affinché l'Istituto potesse comunque gestire l'ingente mole di informazioni da codificare. In fase di progettazione dell'ultimo Censimento sono state pertanto analizzate le varie opportunità di soluzione del problema, valutando sia l'entità del carico di lavoro sia le performance di software di codifica automatica già sperimentati sui dati di altre indagini effettuate all'interno dell'Istat.

Pertanto, considerato che:

- sommando le variabili alfabetiche previste per i 20 milioni di famiglie e quelle relative alle 500.000 Convivenze stimate, l'analisi dei tempi e delle dotazioni umane e informatiche necessarie aveva portato a risultati tali da escludere a priori l'eventualità di evadere interamente in Istat l'attività di codifica
- le sperimentazioni effettuate per la codifica automatica dei testi attraverso il software canadese ACTR aveva prodotto risultati incoraggianti sia in termini di quantità che di qualità dei codici attribuiti con riferimento a stringhe acquisite tramite *data entry* e non anche attraverso la lettura ottica

si è deciso di procedere appaltando allo stesso consorzio di ditte incaricate dell'acquisizione dati delle persone in famiglia anche l'attività di codifica dei testi relativi alla stessa categoria di soggetti (ad eccezione dei residenti in famiglia sloveni e di quelli rilevati su modelli integrativi[3]) e di codificare altresì nell'ambito della Direzione Centrale Censimento della Popolazione, Territorio e Ambiente dell'Istat tutte le variabili contenute nei modelli di rilevazione delle Convivenze.

In particolare l'attività di codifica curata dall'Istituto su testi registrati sarebbe stata effettuata attraverso il software di codifica automatica ACTR, mentre la Elsag avrebbe potuto procedere utilizzando il/i software ritenuti più opportuni ma comunque basati sui dizionari, uno per ogni variabile, messi a punto e forniti dall'Istat stesso[4].

[3] Si tratta di modelli di rilevazione delle persone in famiglia che i comuni hanno perfezionato in ritardo rispetto ai tempi previsti per la consegna alla Elsag, capogruppo del consorzio di ditte aggiudicatario della gara d'appalto per l'acquisizione dati, e che, come da istruzioni, i comuni stessi hanno inviato direttamente all'Istat per il trattamento delle informazioni.

[4] Cfr par. 1.2

Non solo, nella fase immediatamente precedente la data del Censimento, sono state consegnate alla ditta esterna anche delle linee guida sugli step da seguire nella fase di attribuzione dei codici al fine di garantire, per quanto possibile, un livello di omogeneità accettabile nell'ambito di attività espletate da organi diversi.

In fase di progettazione, con riferimento all'appalto in *outsourcing* per la codifica dei testi, sono stati anche fissati alcuni parametri a cui la Elsag doveva attenersi, relativi sia alla percentuale di codifiche da garantire all'Istituto sia al livello di accuratezza (qualità) dei codici forniti. Di seguito sono riportati tali parametri stabiliti in funzione dei risultati ottenuti all'interno dell'Istituto attraverso l'applicazione di software di codifica automatica su dati di indagini effettuate in precedenza.

Tabella 2.1 - Parametri per l'attività di codifica in outsourcing

Variabili	*Livello minimo di assegnazione del codice*	*Livello minimo di accuratezza*
Comune	95%	99%
Stato estero	90%	98%
Titolo di studio	80%	98%

Fonte: A.Gaucci – CAPITOLATO TECNICO – Allegato allo schema di contratto per la fornitura di servizi relativi all'acquisizione dei dati del Censimento generale della popolazione e delle abitazioni del 2001 mediante tecniche di lettura ottica e di riconoscimento dei caratteri

Come si evince dalla tabella, non sono state stabilite percentuali in relazione alle variabili Professione e Attività Economica comunque rilevate nel fogli di Famiglia. Infatti, data la complessità dell'operazione in termini di attribuzione dei codici alle descrizioni fornite dai cittadini (particolarmente articolate vista la natura delle variabili trattate) e la conseguente entità dei costi da sostenere, si è deciso di non procedere in prima battuta con l'assegnazione dei codici, bensì di prevedere a posteriori la codifica di un campione, stratificato per regione, dei testi acquisiti tramite lettura ottica.

Tale ipotesi di lavoro era peraltro avvalorata dal fatto che, per la prima volta in occasione di un Censimento, nel 2001 nei modelli di rilevazione sono stati inseriti due quesiti pre-codificati, uno relativo ai settori di Attività Economica ed uno inerente l'attività lavorativa svolta, la cui diffusione, a livello universale, avrebbe comunque assicurato la pubblicazione di dati relativi alla Professione e all'Attività Economica, seppur con un dettaglio inferiore rispetto a quello raggiungibile attraverso la codifica delle specifiche fornite per esteso dai rispondenti. Infatti, le modalità del quesito pre-codificato sull'attività lavorativa approssimano il primo digit della classificazione internazionale ISCO 88 COM, mentre quelle relative ai 28 settori di Attività Economica approssimano i primi due digit della classificazione italiana delle Attività Economiche 1991. A tale riguardo, benché in fase di progettazione non fossero stati individuati in via definitiva i responsabili e quindi gli autori dell'attività di codifica del campione appena descritto, era comunque opportuno inserirla tra i compiti da evadere presumibilmente all'interno dell'Istituto, progettando quindi un sistema in grado di supportare anche la realizzazione di questa ulteriore elaborazione.

2.2 La progettazione del sistema

La realizzazione del sistema di codifica dei dati relativi al Censimento della Popolazione 2001 è stata preceduta da un'accurata analisi del carico di lavoro da espletare, delle risorse umane disponibili, delle dotazioni informatiche e, naturalmente, dei tempi previsti per la conclusione dei lavori. Come già sottolineato, compito della Direzione Centrale Censimento della Popolazione, Territorio e Ambiente dell'Istat era curare la codifica delle variabili testuali contenute all'interno dei modelli di rilevazione delle Convivenze ovvero Comune, Stato Estero, Titolo di studio, Professione e Attività e Economica[5] e di un residuo delle stringhe inerenti le stesse variabili[6] relative ai Fogli di famiglia che la ditta esterna non aveva codificato, rispettando comunque i parametri minimi di assegnazione richiesti riportati nella tabella 2.1.

Nella fase di progettazione della procedura informatizzata si è dovuto tener conto in primo luogo del fatto che, benché il software di codifica automatica ACTR garantisse la risoluzione di un elevato numero di casi in tempi brevissimi (ACTR analizza ogni singola stringa in 15 *msec* e, in occasione dei test effettuati prima del Censimento, ha fatto rilevare una performance pari al 77,8%[7] di testi codificati sul totale delle stringhe processate) era comunque indispensabile prevedere un passaggio di codifica manuale ("*on line*") che assicurasse un'attribuzione rapida e con uno standard di qualità elevato dei codici non assegnati in *batch*. Il Censimento della popolazione infatti, come noto, si basa sull' "autocompilazione" dei questionari da parte di tutti i cittadini indipendentemente dall'età, dal grado sociale e dal livello di istruzione e, di conseguenza, spesso le informazioni fornite sono caratterizzate da terminologia variegata, a volte anche bizzarra, e comunque non sempre approssimabile alle dizioni inserite nei dizionari sui cui si basano i software per l'assegnazione dei codici in automatico.

In secondo luogo, data l'ingente mole di dati da codificare e le scarse risorse umane disponibili da impegnare nell'attività di codifica, era indispensabile prevedere la realizzazione di un sistema quanto più "flessibile" che consentisse, tra le altre cose, la lavorazione da parte di più operatori indipendenti l'uno dall'altro di testi afferenti ad una stessa variabile ed allo stesso strato territoriale (Provincia o Comune) senza il rischio di codificare contemporaneamente la medesima stringa e senza la necessità di decidere a priori una suddivisione del lavoro basata o sulla tipologia della variabile o sul territorio di riferimento.

2.3 Le variabili da codificare

Sulla base della strategia organizzativa pianificata, il sistema informatizzato doveva innanzitutto essere perfezionato per la codifica delle variabili testuali contenute nei Fogli di Convivenza acquisite tramite *data entry*, alcune delle quali caratterizzate da peculiarità di cui tener necessariamente conto nella fase di progettazione della procedura. In particolare il Comune e lo Stato Estero erano presenti nei modelli di rilevazione in diverse accezioni ovvero:

- per lo Stato Estero
    - Stato Estero di nascita
    - Stato Estero di cittadinanza
    - Stato Estero di cittadinanza precedente

[5] Nei modelli di rilevazione delle persone residenti in Convivenza (mod.Istat CP.2) non sono stati inseriti i quesiti pre-codificati relativi all'attività lavorativa ed al settore di attività economica ed era quindi necessario procedere con la codifica delle specifiche fornite per esteso dai rispondenti acquisite tramite il tradizionale data entry

[6] ad eccezione di Professione ed Attività Economica

[7] si fa riferimento alla media delle percentuali di codifica ottenute attraverso l'utilizzo di ACTR sui dati della seconda indagine pilota pre-censuaria effettuata a ottobre 2000

- Stato Estero di dimora abituale nel 2000

- per il Comune
  - Comune di nascita
  - Comune di dimora abituale nel 2000.

Per il Titolo di studio, inoltre, si doveva prevedere, sia nella fase *batch* che in quella manuale, l'assegnazione di un codice in funzione delle risposte fornite ai quesiti pre-codificati sul grado di istruzione inseriti nel questionario. Infatti i rispondenti, in occasione dell'ultima rilevazione censuaria, erano tenuti a fornire in primo luogo l'indicazione del titolo di studio più elevato conseguito ("Grado di istruzione") selezionando una delle 16 modalità proposte (ad esempio "Laurea"), la durata del corso di studi (da specificare solo nel caso di diplomi di scuola secondaria superiore conseguiti presso istituti professionali, scuole magistrali o istituti d'arte che prevedono corsi di 2-3 anni o 4-5 anni) e, di seguito, a descrivere, attraverso un quesito a testo libero, la tipologia del titolo stesso (ad esempio "Laurea in matematica"). Si trattava in sostanza da un lato di prevedere un dizionario dei titoli di studio contenente non solo l'insieme di tutte le stringhe processabili ma, accanto a ciascuna di esse, un filtro, ovvero un sorta di etichetta corrispondente alle modalità dei quesiti pre-codificati a cui afferiva quel determinato titolo di studio, dall' altro di istruire il sistema affinché l'assegnazione del codice, sia in automatico che nella fase *computer assisted*, garantisse, se possibile, la coerenza con la risposta fornita alla domanda sul grado di istruzione.

Infatti il sistema di istruzione italiano è caratterizzato dal fatto che omonimi titoli di studio possono essere conseguiti a seguito della frequenza di vari tipi di istituti scolastici o, nell'ambito dei medesimi istituti, a seguito della frequenza di corsi di diversa durata e quindi le modalità selezionate dai rispondenti nell'ambito dei quesiti pre-codificati dovevano costituire necessariamente vincoli determinanti per la corretta attribuzione del codice.

Tabella 2.2 – Esempi di associazione tra titoli di studio e codici contenuti nel dizionario

Modalità selezionata al quesito relativo al "titolo di studio più elevato conseguito"	*Durata del corso di studi*	*Descrizione del titolo di studio*	*Codice*
9 (Diploma di scuola secondaria superiore conseguito presso un Istituto professionale)	1 (2-3 anni)	Segretaria d'Azienda	310406
9 (Diploma di scuola secondaria superiore conseguito presso un Istituto professionale)	2 (4-5 anni)	Segretaria d'Azienda	410405

Lo stesso procedimento doveva essere seguito anche per il Comune in relazione al quale l'assegnazione del codice doveva essere subordinato all'indicazione della provincia fornita dai rispondenti. In Italia, infatti, sono assai frequenti i casi di omonimie dei comuni e, pertanto, la codifica di tale variabile in funzione della provincia specificata era indispensabile per garantire uno standard di qualità elevato del dato censuario.

2.4 Gli obiettivi del sistema

La fase di progettazione del sistema informatizzato per la codifica delle variabili testuali da effettuare *in house* è stata caratterizzata dall'individuazione di tutte le funzionalità che il sistema stesso doveva garantire ovvero:

- primo passaggio di codifica *batch* attraverso ACTR per tutte le variabili; in particolare per il Ttitolo di studio ed il Comune il primo passaggio di codifica in automatico doveva essere subordinato alle risposte fornite dai cittadini ai quesiti pre-codificati strettamente connessi ai testi da codificare (grado di istruzione e durata del corso di studi per la variabile Titolo di studio, provincia per la variabile Comune);
- secondo passaggio di codifica *batch* per le variabili Titolo di studio e Comune senza vincoli per tutte quelle descrizioni a cui nel primo passaggio non è stato attribuito alcun codice;
- interfaccia "*friendly*" per la gestione di tutti quei casi che, non codificati in automatico, avrebbero necessitato dell'intervento dell'operatore manuale;
- gestione centralizzata e aggiornamento dinamico dei dizionari in corso d'opera per garantire *performance* sempre crescenti del *software* di codifica automatica;
- possibilità da parte degli operatori distribuiti su diverse postazioni di poter lavorare contemporaneamente la stessa variabile afferente al medesimo strato territoriale (ad esempio il Comune), di consultare rapidamente gli stessi dizionari utilizzati per la codifica *batch* e di visualizzare alcune delle risposte fornite dai cittadini agli altri quesiti contenuti nei modelli di rilevazione attinenti alla variabile oggetto di codifica *step* indispensabile per risolvere i casi più complessi.

Spesso, infatti, come già sottolineato, trattandosi di un questionario "autocompilato", le descrizioni acquisite, più che le dizioni contenute nelle classificazioni ufficiali, approssimano modi di dire, sono grammaticalmente scorrette o, comunque, generiche al punto tale da compromettere sia la codifica in automatico delle stringhe sia l'intervento degli operatori manuali se messi in condizione solo di consultare le classificazioni ufficiali e i dizionari. Era indispensabile, pertanto, prevedere alcune forme di ausilio per cercare di dirimere nella fase "*on line*" il numero più elevato di casi nel più breve tempo possibile e con uno standard di qualità comunque elevato. E' proprio per questo che la progettazione dell'interfaccia per la codifica del tipo *computer-assisted* ha comportato una serie di analisi e riflessioni particolarmente accurate volte ad identificare le varie forme di assistenza agli operatori più idonee a garantire percentuali di attribuzione dei codici prossime al 100%.

Innanzitutto è stata prevista l'opportunità sopra citata di visualizzare alcune delle indicazioni specificate dai rispondenti in corrispondenza di variabili strettamente connesse a quella oggetto di codifica (ad esempio il Titolo di studio e l' Attività Economica per la codifica della Professione o lo "Stato Estero di nascita" per la codifica dello "Stato Estero di cittadinanza" se diverso da quello italiano) come supporto agli operatori nei casi di testi particolarmente generici o incompleti.

In secondo luogo è stata messa a punto una funzione di ricerca per "*keyword"* attraverso cui l'operatore manuale sarebbe stato in grado di codificare in maniera sequenziale tutte le descrizioni contenenti una determinata parola chiave con un conseguente significativo risparmio in termini di tempo ed un basso rischio di errore in termini di attribuzione di codici diversi a stringhe analoghe.

E’ stata inoltre predisposta una funzione di “prenotazione” delle stringhe particolarmente problematiche attraverso cui l’operatore era in grado di accantonare momentaneamente il testo al fine di poter effettuare ulteriori approfondimenti sull’argomento o contattare codificatori esperti nella materia di interesse, per poi avere la possibilità di selezionare nuovamente il testo ed assegnare il codice ritenuto idoneo al termine delle ricerche e delle consultazioni effettuate.

Alla fase di progettazione è seguita quella di realizzazione del sistema e quindi di test nel corso dei quali è stato possibile apportare alcuni perfezionamenti non preventivati in precedenza. E’ stata, ad esempio, resa disponibile per l’operatore manuale l’opportunità di consultare il dizionario degli Stati Esteri e tutte le variabili di ausilio ad esso correlate anche nell’ambito della codifica del Comune e viceversa. Infatti, analizzando alcuni dei testi rilevati, ci si è resi conto che, così come avvenuto in occasione del Censimento ’91, a volte i rispondenti nello spazio riservato alla specifica dello “Stato Estero di nascita” indicano il “Comune di nascita” e in quello riservato al Comune lo Stato Estero. Si trattava di casi da non poter risolvere automaticamente perché il rischio di commettere errori era troppo elevato, ma che dovevano potersi gestire almeno nella fase *on line*.

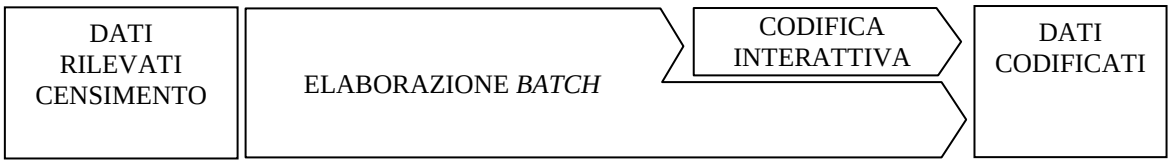
Infine, è stata valutata l’opportunità di implementare in corso d’opera e in maniera dinamica i dizionari con le risposte empiriche più ricorrenti selezionate dagli operatori attraverso la gestione centralizzata dei dizionari stessi. Arricchire i dizionari con il patrimonio costituito dalle empiriche rilevate in occasione di un Censimento significa garantire *performance* sempre più elevate del software di codifica automatica e quindi una sostanziosa riduzione di uomini e mezzi impiegati per l’attività oggetto di studio.

3 Il Sistema per la gestione della codifica realizzato in ISTAT

L'attività di codifica svolta per il censimento generale della Popolazione all'interno dell'Istituto, ha coinvolto diversi operatori per un elevato numero di stringhe da codificare. A supporto di tale attività è stato realizzato un sistema per la gestione della codifica che integra il software ACTR con il sistema informatico del censimento.

Il sistema del censimento può essere visto come un flusso transazionale che produce altri flussi di dati lungo cammini diversi. L'attività di codifica si colloca a valle del sistema di ricezione ed è parallela alle attività di correzione strutturale dei dati dell'indagine. Le stringhe da codificare vengono individuate e distinte a seconda del modello e del progetto di codifica[8] per predisporle alla successiva codifica. L'output del sistema di codifica, invece, determina le informazioni di input per i sotto sistemi di correzione dati, rispettivamente per ogni gruppo di variabili oggetto del processo di correzione e diffusione.

Il software ACTR disponibile per il censimento è di tipo *stand-alone,* mentre l'attività di codifica deve essere parallelizzata su più operatori, causa il grande volume di dati da elaborare. A tale scopo il sistema distribuito è realizzato disaccoppiando le funzioni di elaborazione automatica (fase *batch* di ACTR) da quelle interattive, consentendo quindi, di centralizzare l'attività automatica e distribuire sugli operatori l'attività manuale attraverso una gestione multi-utente della codifica. Questo ha consentito di realizzare un flusso di lavorazione con un alto grado di parallelizzazione ottimizzando i tempi di produzione e con un alto grado di flessibilità.



Schema 1: Flusso di lavorazione del sistema di codifica.

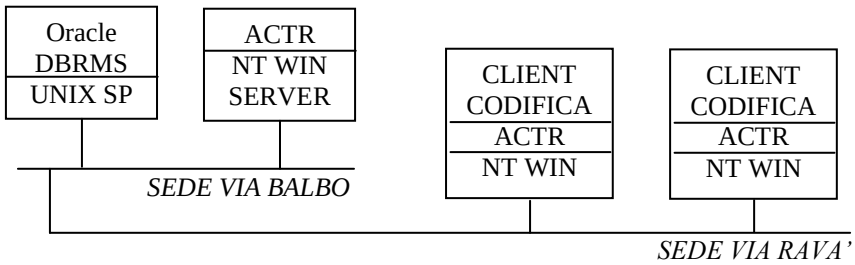
3.1 Architettura logica del sistema

Il sistema è basato su un'architettura di tipo *client/server* basata sui dati con un DBRMS come contenitore centrale dei dati elaborati e da elaborare ed il *software* ACTR come motore d'elaborazione.

Al centro del sistema si trova l'archivio dei dati, consultato dagli operatori e da dove ciascun operatore può estrarre e lavorare localmente ogni singolo record. Al fine di velocizzare l'attività di codifica manuale i dati presenti nel contenitore sono opportunamente filtrati: agli operatori di codifica é stato consentito l'accesso alle sole stringhe con effettiva ambiguità di codifica. L'attività di filtraggio dei dati corrisponde alla fase *batch* di lavorazione durante la quale il software ACTR è utilizzato su blocchi di dati allo scopo di popolare e predisporre i contenitori sorgente distinti per tipologie d'esito ("Unico", "Multiplo", "Possibile", "Fallito"). Ciascuna elaborazione *batch* è eseguita in modo centralizzato sistematicamente su tutte le stringhe non nulle per tutti i progetti di codifica dell'indagine, contemporaneamente su uno o più comuni di una provincia in funzione della sequenza degli arrivi provinciali di dati al sistema di ricezione del censimento.

[8] Per progetto di codifica si intende l'attività di codifica associata ad ogni singola variabile

Dal punto di vista tecnologico, il sistema si basa sul *DBRMS Oracle* in ambiente Unix SP, un NT4 Windows Server e diversi PC NT4 Windows *clients*. Nel *DBRMS* sono contenuti i microdati del censimento da codificare, i dati codificati ed i meta dati per la gestione dei progetti di codifica. Nel Windows Server è istallato il motore di codifica utilizzato per tutte le elaborazioni automatiche centralizzate. Le diverse componenti c*lient* utilizzano il modello *fat client*. Tutte le componenti informatiche sono interconnesse dall'infrastruttura di rete locale all'ISTAT tra le differenti sedi.



Schema 2: Architettura del sistema

3.2 L'architettura applicativa

Il sistema è sviluppato secondo un modello a tre strati: strato di presentazione, strato logiche applicative e strato di accesso ai dati. Lo strato di presentazione è costituito dalle componenti d'interfaccia utente con funzioni d'autenticazione, d'amministrazione e di lavorazione. Lo strato logico implementa le diverse funzioni e la logica dei progetti di codifica. Lo strato di accesso ai dati consente l'interfacciamento tra differenti data base e consente l'integrazione tra i differenti ambienti d'elaborazione.

La fase interattiva è realizzata con un'interfaccia utente, sviluppata con *Oracle Forms*, e consente una navigazione all'interno dell'ambiente, strutturata secondo una sequenza gerarchica fortemente semplificata; le informazioni di supporto alla attività di codifica sono sempre evidenziate al fine di minimizzare l'attività mnemonica dell'operatore.
La richiesta del record da elaborare è gestita all'interno dell'ambiente Oracle, che consente lo scambio di informazioni tra il client e il server. La parallelizzazione della lavorazione è gestita con il *locking* dei record ed è a garanzia della consistenza delle informazioni in lavorazione distribuite su più operatori.
In caso di interruzione dell'attività, l'uscita dal sistema consente di rimettere il record non codificato in lavorazione, diversamente, l'utente può prenotare e mantenere in stato di prenotazione la stringa di codifica. L'elaborazione on line utilizza le risorse locali del client, per minimizzare i tempi d'attesa degli operatori di codifica ad ogni richiesta di codifica on line. In questo modo è stato eliminato il ritardo di rete e l'eventuale ritardo di servizio dell'ACTR server.

La logica di lavorazione è definita nei metadati del sistema dove ogni variabile di codifica è associata alla propria strategia e al corrispondente progetto di codifica. La strategia di codifica *batch* può basarsi su uno o due passi di elaborazione automatica, a seconda della mancanza o della presenza di una condizione "filtro". La sequenza di lavorazione è tale che per ogni passo sono valutati gli esiti della codifica e solo i fallimenti sono riproposti all'eventuale successivo passo di elaborazione automatica. Più variabili di codifica possono appartenere allo stesso progetto di codifica ed ogni progetto di codifica ha definiti i propri archivi dei dizionari e di *parsing*. Ad esempio le variabili di codifica che rappresentano il "Comune di nascita", il "Comune di dimora abituale", ecc. rientrano nel progetto di codifica del Comune. Ogni variabile di codifica definisce un ambiente di elaborazione con le informazioni di supporto

all'attività, distinte a seconda delle diverse funzioni di uso. Nel caso della codifica del "Comune di nascita", la stringa da codificare è referenziata attraverso la tipologia di variabile di codifica (tipo C) insieme al microdato e la sigla della colonna che contiene l'informazione. Una gerarchia relaziona la variabile di codifica alle corrispondenti variabili associate, quali: di imputazione (tipo I), ovvero l'identificativo del campo in cui inserire l'esito della codifica; di "filtro" (tipo F), come ad es. la sigla della provincia di nascita; di ausilio (tipo A), che costituiscono le informazioni di supporto alla decisione quali ad es. la Professione e l'Attività Economica per la codifica del Titolo di studio. Nei metadati di sistema sono archiviate le informazioni relative ai dizionari di codifica ed alle loro versioni per ogni progetto di codifica consentendo, ad ogni connessione di un operatore per ogni progetto di codifica, l'aggiornamento dei metadati locali.
Gli esiti dell'attività di codifica automatica e manuale sono inseriti nelle strutture di *output* d'elaborazione che si distinguono in "Unici" e "da codificare". Il primo gruppo contiene tutte le informazioni relative agli esiti delle codifiche effettuate con successo ("Unici"), mentre il secondo archivia la storia dell'attività di codifica *batch* per ogni singola stringa di codifica con associato il corrispondente risultato ("Multiplo", "Possibile", "Fallito") e lo *score* attribuito dal ACTR nel relativo passo di elaborazione.
L'attività di codifica manuale trasferisce il record dal contenitore "da codificare" al contenitore degli "Unici" inserendo l'identificativo dell'operatore che ha effettuata la codifica. Per ogni singola codifica effettuata manualmente è memorizzata la data di inserimento della stringa nel sistema e la data dell'effettiva codifica. Queste informazioni consentono di mantenere aggiornato il data warehouse per il monitoraggio delle attività di codifica effettuate.

Lo strato di accesso ai dati di lavorazione si interfaccia sia con i microdati rilevati che con il motore di codifica. L'integrazione tra l'ambiente Unix su cui risiedono i dati del DB Oracle e l'ambiente Windows su cui gira l'applicazione ACTR è realizzata attraverso una interfaccia dati con funzioni di estrazione, trasformazione e caricamento dal DB verso windows e vice versa, secondo predefinite sequenze di procedura. Lo scambio di informazioni tra basi diverse basi dati Oracle avviene via *db link*.
Per ogni elaborazione *batch*, lo strato logico esegue il servizio di richiesta di nuove stringhe confrontando le informazioni relative ai nuovi invii con i dati gia codificati; successivamente le stringhe da codificare sono predisposte ed inserite negli input *files* all'interno del window server, nelle corrispondenti directory di progetto per ogni variabile di codifica. La fase di trasferimento inversa esegue il passo di procedura per il trasferimento degli *output files* presenti nell'ambiente windows alle corrispondenti strutture di *output* del DB.
All'interno dell'ambiente windows, a livello *client*, l'elaborazione ACTR on line consente di sottoporre una richiesta di elaborazione locale per un singolo record; in tale fase il record viene inserto nell'input file corrispondente alla scelta del progetto di codifica e l'esito della codifica è visualizzato dall'interfaccia utente condividendo con il software ACTR le strutture di output.

3.3 I ruoli nell'applicazione

La fase interattiva prevede due differenti ruoli nell'applicazione: l'*operatore logistico*, di tipo amministrativo, e l'*operatore di codifica* che effettua l'attività di codifica manuale delle stringhe.
L'operatore logistico svolge la sua attività di codifica su due livelli d'intervento: il livello gestionale e il livello di supporto alla codifica. Il livello gestionale consente la definizione dei metadati di sistema, la gestione delle utenze e l'alimentazione delle nuove stringhe da codificare. Il livello di supporto fornisce gli strumenti di monitoraggio dell'attività degli operatori ed interagisce con gli operatori condividendo i casi dubbi.
Nell'ambito del livello gestionale l'operatore logistico si occupa operativamente dell'esecuzione delle procedure *batch* sui dati di ciascuna provincia che risulta accettata dal sistema di ricezione del censimento. L'operatore logistico segue

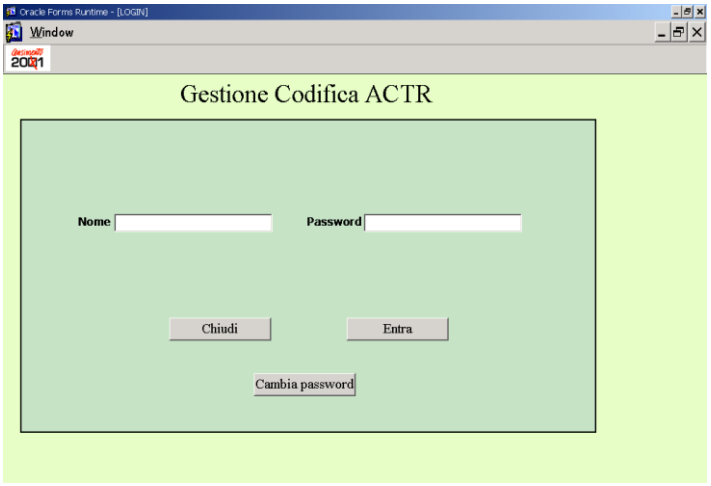
l'evoluzione dell'elaborazione *batch* al fine di valutare la qualità degli esiti per ogni passo di lavorazione per le variabili di codifica. Al termine d'ogni elaborazione *batch* su un predefinito territorio, il sistema fornisce all'operatore logistico gli indicatori sugli esiti dei diversi risultati per ogni variabile di codifica. L'ottimizzazione del processo di elaborazione *batch* è realizzato attraverso la gestione "a caldo" dell'aggiornamento dei dizionari, migliorando le performance di codifica *batch* e riducendo di conseguenza il lavoro interattivo.
L'attività di supporto alla codifica si avvale del monitoraggio della produzione, attraverso una reportistica standard, con l'obiettivo di valutare complessivamente l'attività nel suo insieme o distinta per ogni operatore.
La gestione delle utenze, oltre a consentire il monitoraggio sull'attività e garantire la sicurezza dei dati, è utilizzata per ottimizzare il processo di produzione, consentendo una specializzazione degli operatori sulle stringhe per tipologie omogenee rispetto ai dizionari di codifica attraverso l'associazione operativa di ciascun operatore di codifica ad uno o più progetti di codifica.

L'operatore di codifica esegue l'attività interattiva assegnando il codice in tutti i casi che risultano ambigui al termine dell'elaborazione *batch*. Ad ogni richiesta di una nuova elaborazione da effettuare, l'applicazione fornisce la stringa da elaborare insieme agli esiti delle elaborazioni *batch,* consentendo eventualmente all'operatore di recuperare rapidamente il valore di codifica tra uno dei risultati forniti dall'applicazione. Oltre all'assegnazione immediata del codice può rielaborare stringhe codificate oppure "prenotare" la lavorazione di un record in caso di dubbio. La funzione di "prenotazione" consente la condivisione della stringa di codifica tra l'operatore logistico e l'operatore di codifica.
La funzione di elaborazione on line consente di interrogare il dizionario oggetto della codifica o un qualsiasi altro dizionario tra i diversi progetti di codifica. La lavorazione di ciascuna stringa può avvenire in modo progressivo, attraverso una ricerca filtrata per codice o secondo una parola chiave che caratterizza codifiche affini. A supporto della scelta da effettuare per l'attribuzione del codice l'operatore può visualizzare contemporaneamente le informazioni ausiliare della medesima unità di rilevazione e le informazioni relative agli esiti *batch*. L'elaborazione locale di ACTR, modalità *on line*, gestisce il potenziale disallineamento del sistema distribuito con un controllo di versione tra la componente server e quella di ciascun *client* nell'istante di accesso al sistema.

4 L'interfaccia Oracle Form e le funzioni a supporto della codifica *computer-assisted*

L'interfaccia utente di supporto alla codifica è stata progettata tenendo in considerazione il ruolo di "*Operatore Logistico*" e di "*Operatore di codifica*". Nelle sua fase di avvio, a prescindere dal ruolo ricoperto, l'interfaccia (Figura 1) richiede la digitazione della *login* (*Nome*) e della Password.

Figura 1: Interfaccia di login.



4.1 L'interfaccia Operatore Logistico

Se il ruolo dell'utente è quello di "Operatore Logistico", l'interfaccia (Figura 2) consente la scelta tra le funzioni logistiche o di operatore ACTR[9]. Questa schermata non viene visualizzata nel caso in cui l'utente sia riconosciuto solo come operatore di codifica.

Figura 2: Interfaccia iniziale dell'operatore Logistico.

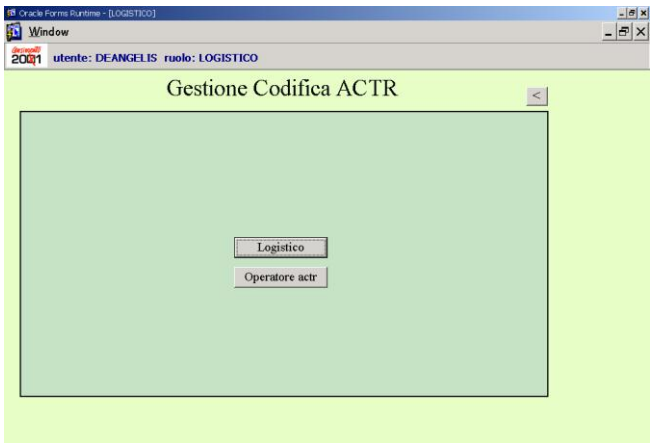
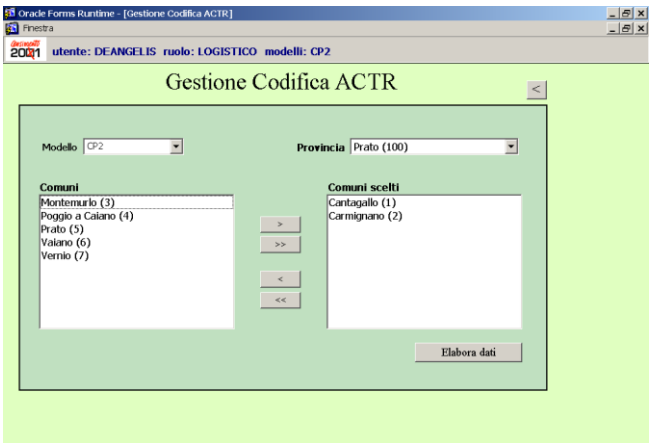


Figura 3: Interfaccia iniziale per elaborazione *batch*.



Il pulsante *Logistico* consente l'accesso alla schermata di Figura 3 e richiede di selezionare, attraverso la tendina in alto

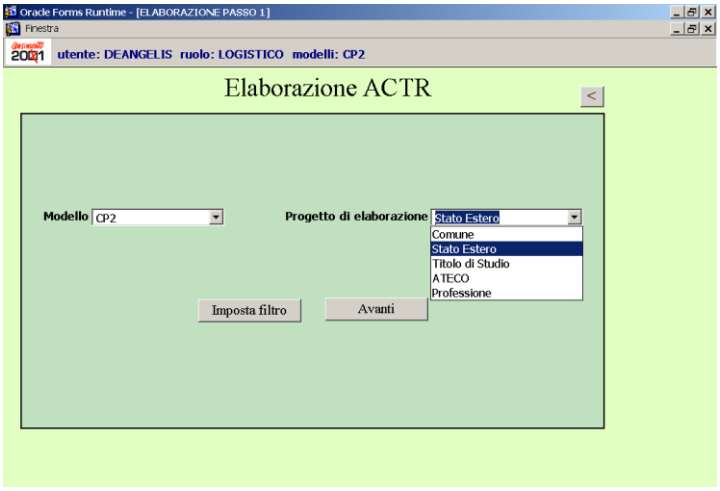
[9] Per l'interfaccia dell'operatore di codifica cfr. § 4.2.

a sinistra, il *Modello* di rilevazione ISTAT (ad esempio CP.1 o CP.2)[10]. Eseguita tale azione l'operatore dovrà selezionare il territorio partendo dalla scelta di una provincia[11]. In tal modo nel riquadro in basso sulla sinistra comparirà la lista dei comuni della provincia selezionata disponibili per l'esecuzione delle procedure *batch*. L'operatore logistico selezionerà i comuni che intende sottoporre a procedura, trascinandoli nel riquadro dei comuni scelti, ed avvia la procedura *Elabora dati*. Il sistema consente di valutare l'esito della procedura confermando, o eventualmente annullando, l'operazione eseguita.

4.2 L'interfaccia Operatore di codifica

Compito dell'operatore codifica è quello di risolvere tutti i casi non codificati automaticamente (compito che comunque anche l'operatore logistico può svolgere - Figura 2).

Figura 4: L'interfaccia per l'inizio dell'attività di codifica



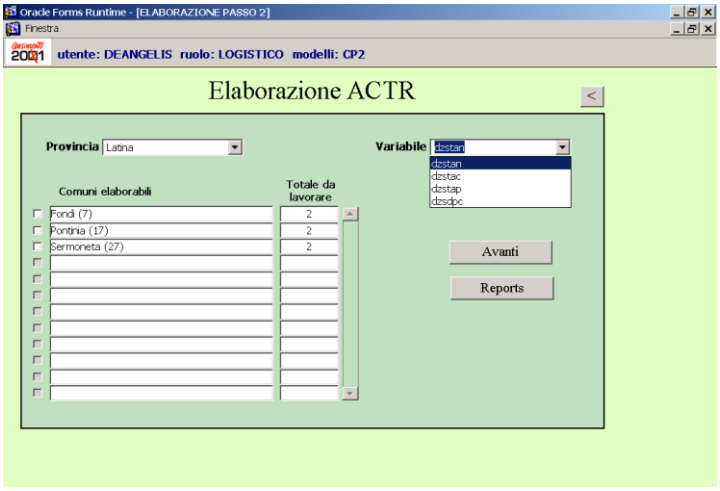
La schermata che si attiva (Figura 4)[12] richiede di selezionare un modello di rilevazione ed un "*PROGETTO DI ELABORAZIONE*" (Comune, Stato Estero, Titolo di Studio, ecc.).

Di seguito (Figura 5) verrà selezionato il territorio (*PROVINCIA*) e, attraverso la tendina in alto a destra (*VARIABILE*), la variabile afferente al "PROGETTO DI ELABORAZIONE" precedentemente selezionato. Dopo aver evaso queste due operazioni preliminari, nella schermata verranno visualizzati, rispettivamente, nella tabella posta a sinistra, i comuni disponibili per la lavorazione (*COMUNI ELABORABILI*) e, a destra, il volume dei casi da trattare relativo ad ogni comune (*TOTALE DA LAVORARE*) che è scalabile in relazione alla lavorazione effettuata.
Se, ad esempio, si sceglierà di lavorare il comune di Fondi (2 casi da trattare relativi alla variabile Stato Estero di nascita), nel momento in cui viene effettuata una codifica il "*TOTALE DA LAVORARE*" relativo a tale Comune diventerà 1 (uno). Elaborata l'ultima codifica, il comune di Fondi non comparirà più nella lista dei "*COMUNI ELABORABILI*". Evasa totalmente per ogni singolo comune elaborabile l'attività di codifica, anche la provincia nella

[10] Compito della Direzione Centrale Censimento della Popolazione, Territorio e Ambiente era curare la codifica delle variabili testuali contenute all'interno dei modelli di rilevazione delle Convivenze (CP.2) e di un residuo delle stringhe contenute nei modelli di rilevazione delle Famiglie (CP.1) non codificate in outsourcing. Per ulteriori approfondimenti cfr. §2.1.
[11] La non visualizzazione, nella tendina "PROVINCIA", di alcune province sta ad indicare che nel territorio di riferimento non ci sono più comuni da sottoporre a procedura *batch.*
[12] Alla schermata di Figura 4 l'operatore di codifica, riconosciuto per il suo ruolo, perviene dopo aver digitato la "LogOn" e la "Password" dalla schermata riportata in Figura 1. L'operatore di codifica, invece, vi perviene cliccando il tasto "Operatore ACTR" dalla Figura 2.

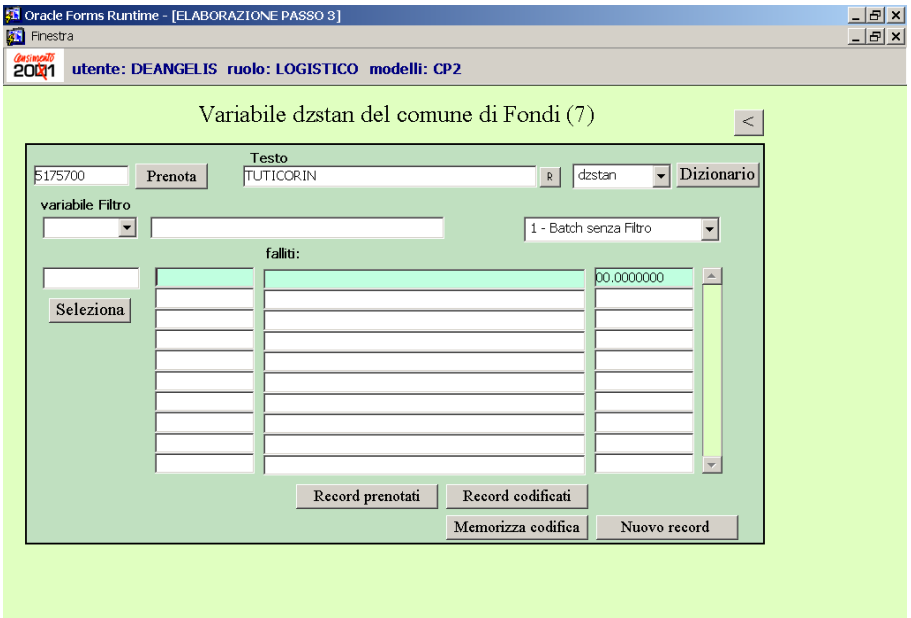
tendina in alto a sinistra (*PROVINCIA*) non verrà più visualizzata[13]. La stessa schermata consente anche di poter accedere ad alcune funzioni di reportistica che si attivano attraverso il pulsante "Reports" di cui tratteremo nel prosieguo.

Figura 5: Ricerca delle stringhe di codifica per territorio



La schermata successiva (Figura 6) è quella che consente di operare la codifica manuale di una o più stringhe precedentemente non riconosciute in automatico dal *software* ACTR.

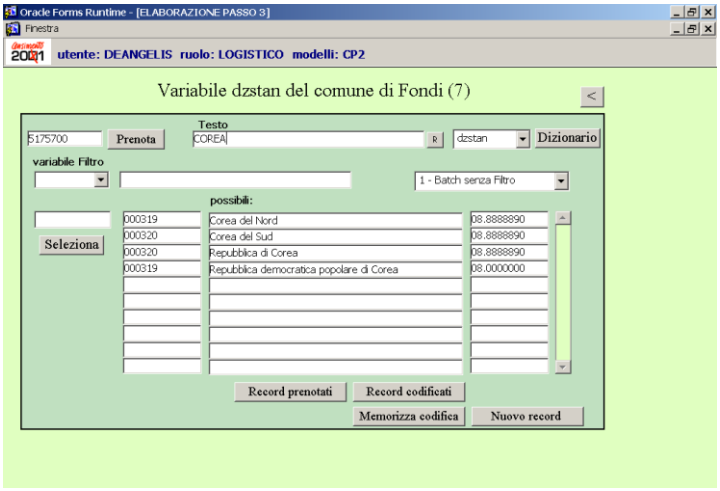
Figura 6: Interfaccia di codifica



[13] Può accadere, comunque, che nonostante non ci siano più comuni elaborabili la provincia selezionata rimanga ancora in visualizzazione a causa della funzione di "Prenotazione" prevista nell'interfaccia che verrà illustrata più avanti.

Nell'esempio che segue, inerente la variabile "Stato Estero di nascita", vediamo che la prima dizione riportata dal rispondente (*TESTO*) corrisponde a "TUTIRICON". L'esito del primo passaggio del *software* ACTR (*1- BATCH SENZA FILTRO*) conferma che è un "*Fallito*" poiché nel dizionario non è stata rintracciata alcuna voce afferente allo Stato Estero "Tutiricon". Se, invece, il testo riportato dal rispondente fosse stato "COREA", l'esito del passaggio *batch* sarebbe stato un "*Possibile*" (Figura 6.1) in quanto nel dizionario sono presenti diverse voci relative allo Stato della Corea: Corea del nord e Corea del sud. In un caso del genere, nella tabella centrale verrebbero mostrate una serie di possibili alternative con, nella prima colonna i codici dello Stato Estero, nella parte centrale le dizioni a cui i codici si riferiscono e nella colonna di destra il punteggio di *match*.

Figura 6.1: Interfaccia, il caso di un possibile



L'operatore di codifica trovandosi in presenza della dizione "TUTIRICON" potrà, tuttavia, prioritariamente, azionare la tendina in alto a sinistra "*VARIABILE FILTRO"* per verificare se vi siano informazioni utili atte a risolvere la complessità del caso presentato[14]; in questo caso, nello Stato Estero di cittadinanza (DZSTAC) e nello Stato Estero di dimora abituale nel 2000" (DZSDPC) il rispondente ha riportato la dizione "INDIA" (Figura 6.2 e 6.3).

Figura 6.2: Interfaccia di codifica, esito del DZSTAC

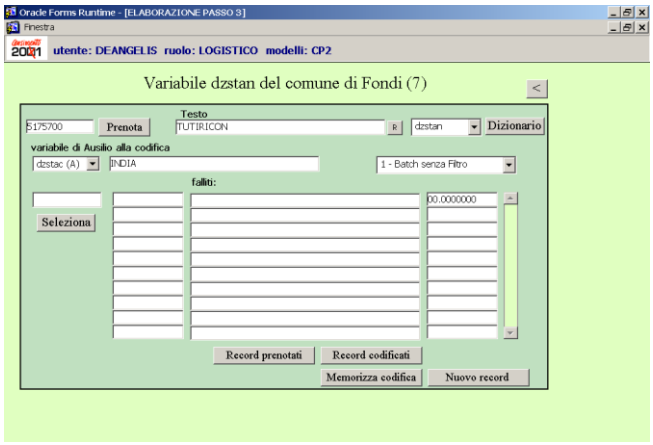
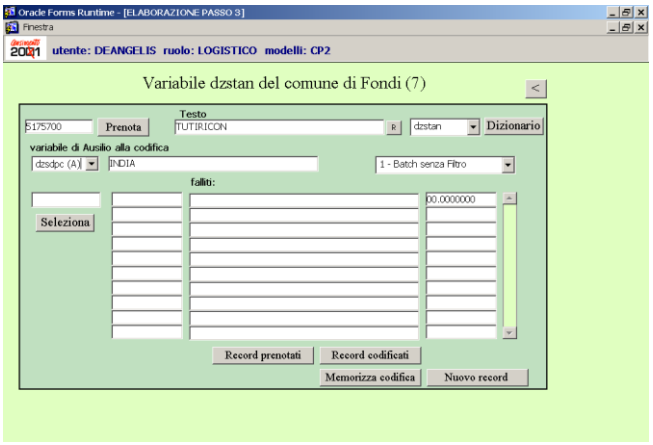


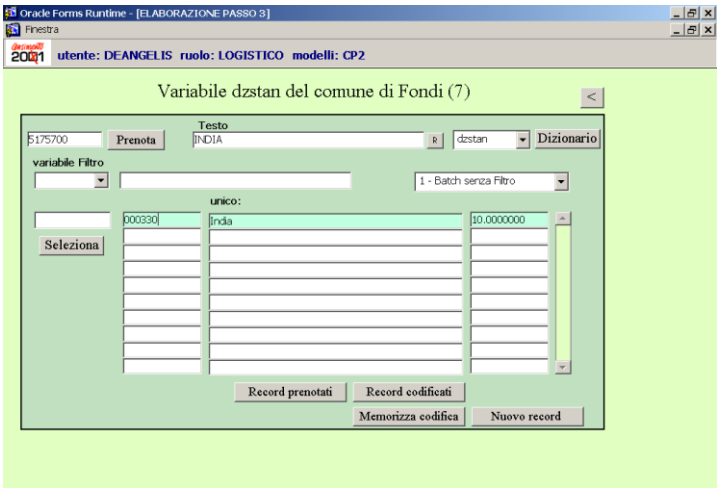
Figura 6.3: Interfaccia di codifica, esito del DZSDPC



[14] Ricordiamo, infatti, che nella fase di progettazione del sistema sono state previste alcune forme di ausilio alla codifica "*on line*". Nella tendina VARIABILI FILTRO sono visualizzate alcune indicazioni specificate dai rispondenti in corrispondenza di variabili strettamente connesse a quella oggetto di codifica o anche le risposte fornite ai quesiti pre-codificati. Per ulteriori chiarimenti cfr. §2.4

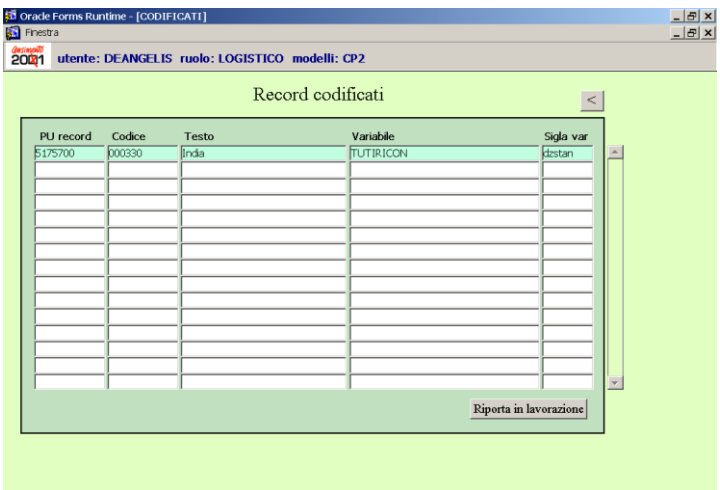
Attraverso queste prime informazioni a sua disposizione, l'operatore può verificare, attraverso ricerca mirata anche tramite internet, se in India esiste una città chiamata "TUTIRICON"; in caso positivo si digiterà nel campo *"TESTO"* la dizione INDIA e immediatamente dopo potrà essere azionato il dizionario attivando il tasto in alto a destra ("dizionario").

Figura 6.4: Interfaccia di codifica, la codifica India



Nella tabella centrale (Figura 6.4) comparirà quindi, come "*Unico*", il codice "000330" afferente allo Stato Estero India che l'operatore dovrà selezionare e, di seguito, memorizzare attraverso il tasto "Memorizza codifica". In qualsiasi momento l'operatore è in grado di visualizzare le codifiche effettuate relativamente al comune che sta lavorando ("Record codificati"). In tal caso, nella schermata che si attiva (Figura 6.5) verranno visualizzate le informazioni inerenti il progressivo unico riferito al singolo record (*PU RECORD*), il codice dello Stato Estero attribuito (*CODICE*), il testo di *match* contenuto nel dizionario a cui il codice si riferisce (*TESTO)*, il testo originario riportato dal rispondente (*VARIABILE*), e la sigla della variabile oggetto di lavorazione (*SIGLA VAR*).

Figura 6.5: Visualizzazione sintetica dei record lavorati



L'operatore può servirsi di tale funzionalità sia per verificare come sono stati risolti in precedenza casi che nel corso della lavorazione possono presentarsi nuovamente, sia per riportare in lavorazione alcune stringhe precedentemente codificate. Quest'ultima possibilità avviene posizionandosi con il cursore sulla riga contenente la stringa da riportare in lavorazione e attivando il tasto "Riporta in lavorazione" posto in basso a destra.

4.3 La codifica delle stringhe senza attribuzione di codice e la funzione di "prenotazione"

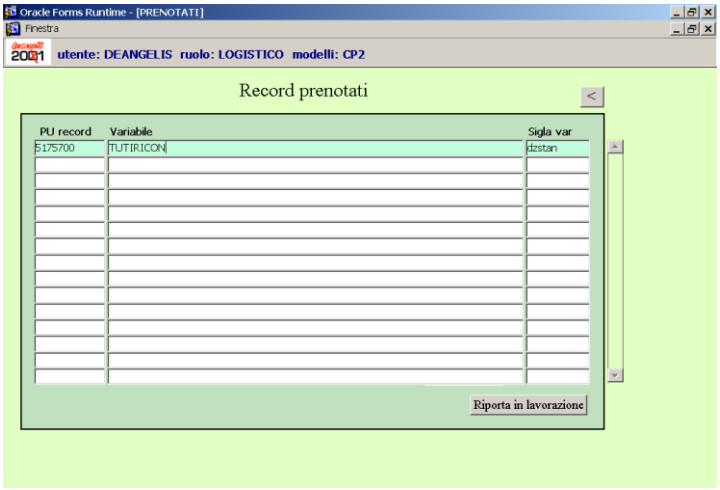
Le alternative alla gestione del caso esaminato fino ad ora, a cui è stato possibile attribuire un codice grazie anche alle variabili di ausilio alla codifica, potevano consigliare altri due percorsi: la codifica della stringa senza attribuzione di un codice o la "Prenotazione" della stringa stessa. Considerando ancora la schermata di Figura 6, soffermiamoci brevemente sulla prima possibilità. Ipotizziamo, ad esempio, che l'operatore di codifica rispetto ad una dizione ostica non abbia alcuna informazione di ausilio basata sulla selezione delle *"VARIABILI FILTRO"*; in tal caso l'operatore potrebbe decidere di procedere ugualmente alla codifica della stringa senza attribuire ad essa un codice. Il procedimento avviene semplicemente attivando il tasto "Memorizza codifica". In questo caso la schermata di visualizzazione delle codifiche effettuate, attraverso il tasto "Record codificati", non riporta, a differenza di quanto appena visto in Figura 6.5, il Codice e il Testo ma, anche in una situazione del genere, è comunque possibile riportare in lavorazione la stringa elaborata.

L'operatore può altresì sospendere la lavorazione di un record prenotandolo. La funzione di prenotazione consente di ritardare la lavorazione di una stringa ad un momento successivo (tasto "Prenota" in alto a sinistra[15]). Anche in questo caso l'operatore può visualizzare, attraverso il tasto "Record prenotati", le prenotazioni effettuate relativamente al comune che sta lavorando. Nella schermata che si attiva (Figura 6.6) vengono visualizzate le seguenti informazioni: il progressivo unico riferito al singolo *record* (PU RECORD), il testo originario riportato dal rispondente (VARIABILE), e la sigla della variabile oggetto di lavorazione (DZSTAN).

Dopo aver effettuato opportuni approfondimenti sull'argomento o contattato codificatori esperti nella materia di interesse, l'operatore di codifica ha la possibilità di riportare in lavorazione la stringa precedentemente prenotata assegnando il codice ritenuto più idoneo; in caso contrario l'unica alternativa possibile sarà rappresentata dalla codifica senza attribuzione di codice secondo le procedure precedentemente esaminate.

[15] I record prenotati sono visibili all'interno del comune di lavorazione "*COMUNI ELABORABILI"* e vengono contabilizzati come volume complessivo di lavorazione "*TOTALE DA LAVORARE".* Cfr. Figura 5.

Figura 6.6: Visualizzazione sintetica dei record prenotati



4.4 La gestione di casi particolari

Durante la fase di lavorazione "*computer assisted*" ci si è resi conto che in un numero non trascurabile di casi i rispondenti riportavano la dizione di un comune (di nascita o di dimora abitale nel 2000) nel campo riservato alla specifica dello Stato Estero e viceversa[16]. Nella Figura 6.7 è riportato il quesito relativo al "Luogo di nascita" contenuto nel modello di rilevazione Istat CP.1 (Foglio di famiglia)

Figura 6.7: CP.1, quesito sul luogo di nascita

Come si può osservare, per il quesito relativo al luogo di nascita erano previste tre modalità di risposta. Nel caso di biffatura della modalità 2 (in altro comune italiano) o 3 (all'estero) i rispondenti dovevano indicare rispettivamente il Comune o lo Stato Estero di nascita; poteva accadere tuttavia che nonostante la persona avesse indicato di essere nata "in altro comune italiano", riportasse la specifica dello stesso nelle caselle riservate allo Stato Estero e viceversa.

[16] La stessa situazione, anche se meno frequente, è stata riscontrata per il quesito sulla Professione e il settore di Attività Economica.

A tale proposito, nell'interfaccia per l'operatore di codifica, è stata prevista la possibilità di consultare simultaneamente i diversi dizionari. In Figura 6.8 è riportato un caso di specifica di un comune (*TESTO)* nel campo riservato allo Stato Estero di nascita[17].

Figura 6.8: Presenza di un Comune nel campo Stato Estero

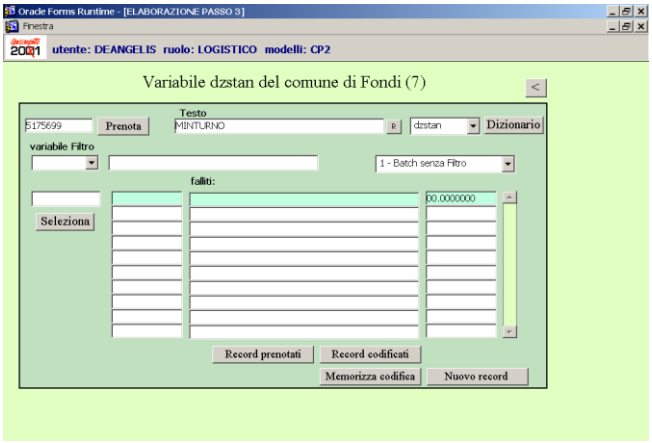
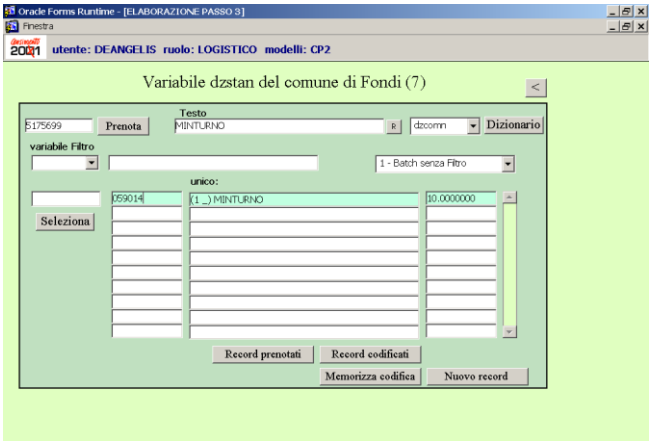


Figura 6.9: Codifica del Comune nel campo Stato Estero



Per poter operare correttamente la codifica l'operatore dovrà selezionare il dizionario relativo al "Comune di nascita" (DZCOMN) ed effettuare la codifica seguendo lo stesso percorso già illustrato in precedenza (Figura 6.9).

4.5 La codifica delle stringhe attraverso il motore di ricerca per parole chiave

In questo paragrafo verrà trattata la funzione di ricerca e selezione delle stringhe da codificare che contengano una determinata parola "chiave". Tale funzionalità si attiva per mezzo del tasto "Imposta filtro" posto in basso a sinistra (Figura 4)[18]. Verrà poi selezionata (Figura 6.10) la variabile di lavorazione in relazione a cui identificare la parola che dovrà essere digitata nel campo *"FILTRO"*. Opzionale è la scelta territoriale che l'operatore di codifica attiverà solo se vuole limitare la ricerca ad un preciso contesto provinciale. Nel caso non venga specificata l'indicazione della provincia, la ricerca verrà effettuata su tutto il territorio nazionale.

Figura 6.10: Ricerca delle stringhe con la funzione filtro

[17] Minturno è un comune della provincia di Latina.

[18] La funzione è attiva solo dopo aver operato la scelta del "MODELLO" e del "PROGETTO DI ELABORAZIONE" come richiesto dalla schermata di Figura 4.

Nelle schermate di Figura 6.11-6.13 sono riportati i tre casi trovati nella provincia di Latina, rispetto allo Stato Estero di nascita (DZSTAN) che contengono la parola "ARGE". Il numero complessivo di casi rintracciati, in base alle specifiche impostate nella Figura 6.10, si trova in basso a sinistra (Record trovati 3) mentre il territorio comunale afferente alla provincia selezionata (Latina) è nella parte alta della schermata (Variabile dzstan comune di Formia provincia di Latina). Per codificare o prenotare le stringhe estratte, per rivedere le codifiche o le prenotazioni effettuate, si segue lo stesso percorso già illustrato nei paragrafi precedenti. La funzione di ricerca per *keyword* porta l'enorme vantaggio di codifica sequenziale di tutte le descrizioni contenenti una determinata parola chiave con conseguente risparmio di tempo e un basso rischio di errore in termini di attribuzione di codici diversi a stringhe analoghe; tuttavia, come ha dimostrato l'esperienza di lavorazione, tale funzionalità si adatta meglio a situazioni in cui risulta limitata la variabilità semantica delle parole.

Figura 6.11: Esito della ricerca con filtro, caso 1

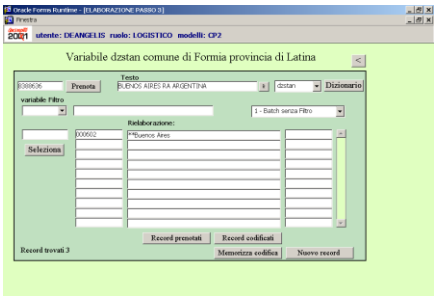


Figura 6.12: Esito della ricerca con filtro, caso 2

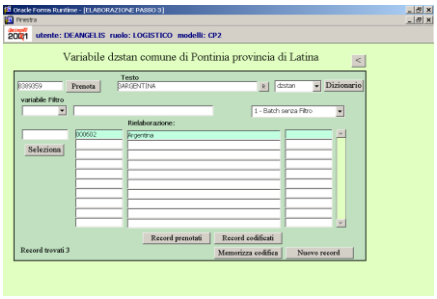
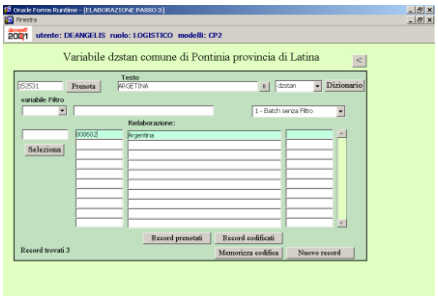


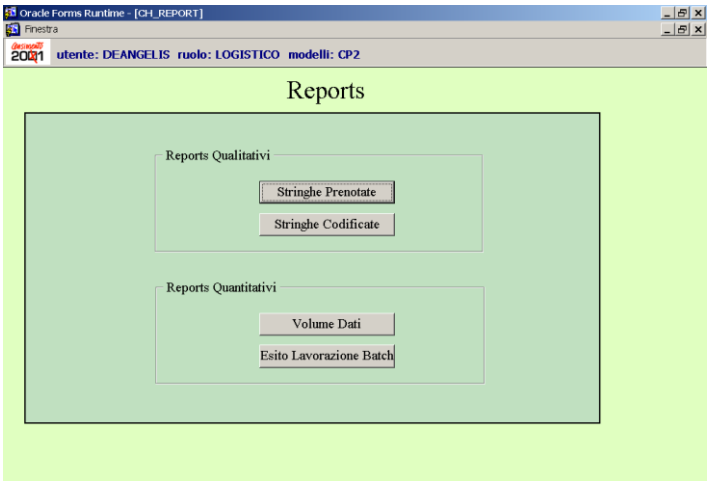
Figura 6.13: Esito della ricerca con filtro, caso 3



4.6 Le funzioni di reportistica

La reportistica è stata suddivisa in due grandi tronconi: i report qualitativi e quantitativi.

Figura 7: Interfaccia di accesso alla reportistica

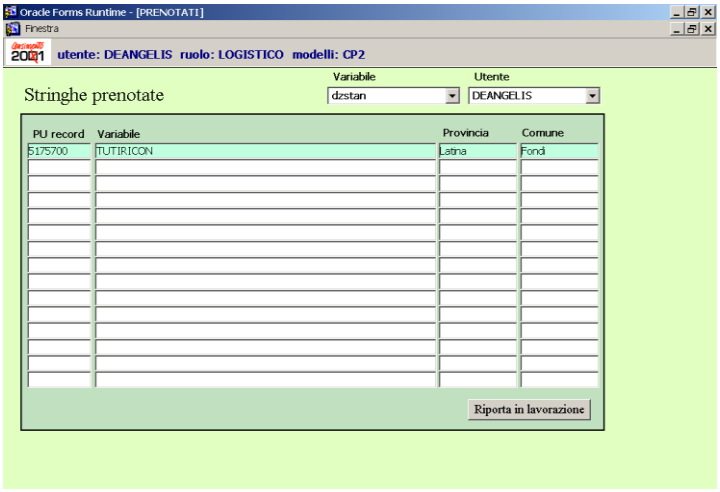


E' opportuno precisare che l'accesso alla reportistica non è fruibile per tutti gli utenti in maniera indifferenziata. L'utente riconosciuto[19] come Operatore Logistico può utilizzare indifferentemente tutte le funzionalità di reportistica

[19] Il riconoscimento del ruolo attribuito all'utente avviene automaticamente nel momento in cui nelle schermata di FIGURA 1 si digita la "Logon".

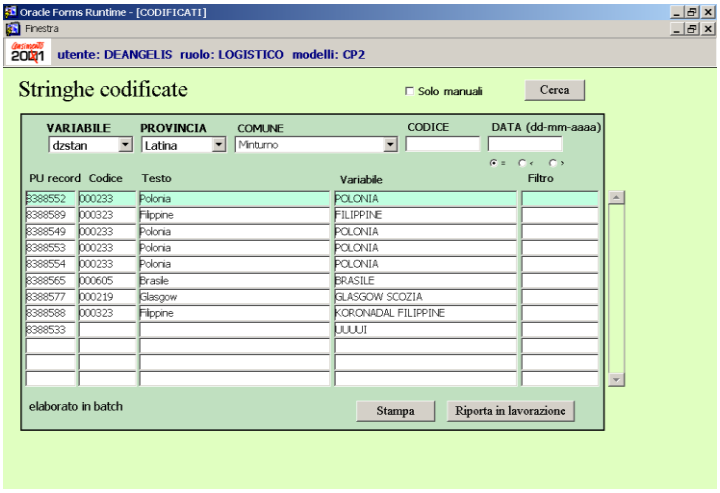
offerte dal sistema mentre l'Operatore di Codifica può consultare esclusivamente la reportistica qualitativa ed in particolare quella riguardante le codifiche effettuate (STRINGHE CODIFICATE) di cui tratteremo a breve.
Per quanto riguarda i report qualitativi, premendo il tasto "*STRINGHE PRENOTATE"* (Figura 7) l'Operatore Logistico può esaminare, rispetto ad una variabile e ad un operatore di codifica, le stringhe in stato di prenotazione.

Figura 7.1: Interfaccia delle stringhe prenotate



E' possibile così visualizzate (Figura 7.1) le prenotazioni effettuate con le seguenti informazioni: il progressivo unico riferito al singolo record (*PU RECORD*), il testo originario riportato dal rispondente (*VARIABILE*), la provincia (*PROVINCIA*) ed il comune (*COMUNE*) in cui è avvenuta la prenotazione. Si potrà anche riportare in lavorazione la stringa precedentemente prenotata ("Riporta in lavorazione").
Attraverso il tasto "*STRINGHE CODIFICATE"* (Figura 7) si accede, invece, alla reportistica relativa alle codifiche effettuate sia in automatico che dagli Operatori di codifica. La schermata che si attiva (Figura 7.2) richiede prioritariamente di operare la scelta di una variabile[20] (*VARIABILE*) e di un contesto territoriale provinciale (*PROVINCIA*) e comunale (*COMUNE*) in cui è avvenuta la codifica.

Figura 7.2: Interfaccia delle stringhe codificate



[20] La scelta di una variabile è vincolata alla selezione del *"PROGETTO DI ELABORAZIONE"*precedentemente operata nella schermata di FIGURA 4.

Di seguito potranno essere visualizzate le codifiche effettuate con le seguenti informazioni: il progressivo unico riferito al singolo record (*PU RECORD*), il codice attribuito (*CODICE*), il testo di *match* contenuto nel dizionario a cui il codice si riferisce (*TESTO)*, il testo originario riportato dal rispondente (*VARIABILE*), il filtro (*FILTRO*) collegato alla variabile[21]. Nella Figura 7.2 sono riportate le codifiche effettuate, rispetto alla variabile Stato Estero di nascita (DZSTAN), nel comune di Minturno in provincia di Latina. Trattasi di nove codifiche[22] di cui sei operate in automatico e tre dagli operatori di codifica. Il sistema, oltre a fornire indicazioni inerenti l'operatore che ha effettuato la codifica, offre anche la possibilità di effettuare ricerche filtrate: per autore della codifica, per codice attribuito e per data di codifica.

Figura 7.3: Ricerca per operatore

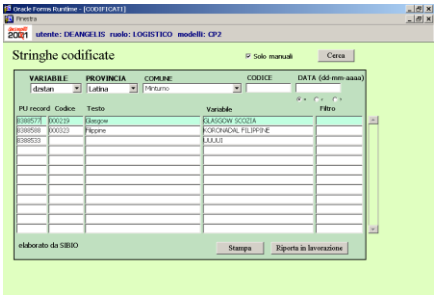


Figura 7.4: Ricerca per codice

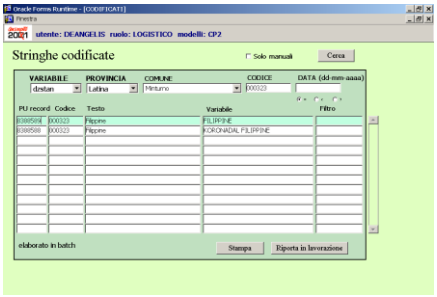
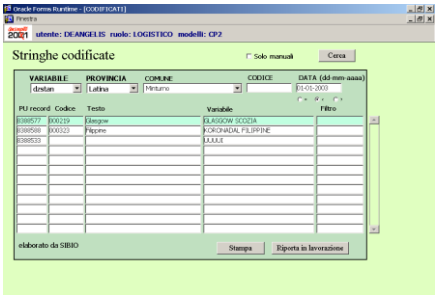


Figura 7.5: Ricerca per data



Nella Figura 7.3 è riportato l'esito di una ricerca delle sole codifiche effettuate dagli operatori di codifica, attraverso la selezione della casella "*SOLO MANUALI*". Se si ha la necessità di effettuare una ricerca per codice (Figura 7.4), sarà necessario digitare nel campo "*CODICE*" il codice da ricercare. E' opportuno evidenziare che la ricerca per codice può essere effettuata simultaneamente con la ricerca per autore manuale della codifica. Infine si può effettuare la ricerca per data di codifica (Figura 7.5) semplicemente digitando nel campo "*DATA*" il giorno, mese ed anno in cui la codifica è stata effettuata. Anche questo filtro di ricerca funziona simultaneamente con quello dell'autore manuale della codifica. L'interfaccia offre infine la possibilità di stampare le ricerche effettuate (STAMPA) nonché di riportare in lavorazione le stringhe codificate (RIPORTA IN LAVORAZIONE).

Come precedentemente detto, tale reportistica è la sola alla quale possono accedere anche gli Operatori di codifica. La differenza sostanziale di funzionalità che la schermata offre, a seconda del ruolo assegnato, consiste nel fatto che l'operatore di codifica può visualizzare le stringhe lavorate esclusivamente da lui, mentre la schermata dell'Operatore logistico permette di tenere sotto controllo tutte le codifiche effettuate dai singoli operatori o in automatico attraverso ACTR. Ciò implica che l'operatore di codifica può utilizzare i filtri di ricerca per codice e per data di codifica mentre non risulta abilitata la ricerca per autore.

Premendo sul tasto "*VOLUME DATI"* l'Operatore Logistico può accedere alla reportistica relativa all'attività di codifica evasa e da evadere (report quantitativi). La schermata che si attiva (Figura 7.6) richiede prioritariamente di operare la scelta di un modello di rilevazione utilizzato per censire le persone residenti in famiglia (CP.1) o in convivenza (CP.2).

[21] Nel caso della variabile "Stato estero" non esistono filtri e pertanto nella colonna *"FILTRO"* non verrà visualizzata alcuna informazione. Se avessimo selezionato la variabile 'Comune di nascita' (DZCOMN), il filtro sarebbe stato rappresentato dalle sigle delle province italiane. Per ulteriori chiarimenti cfr. § 2.3.

[22] Nell'ultima codifica in basso non sono riportati il *"CODICE"* ed il *"TESTO"*. Si tratta in questo caso di una codifica senza attribuzione di un codice. Sull'argomento cfr. il § 4.3.

Figura 7.6: Interfaccia di accesso per modello

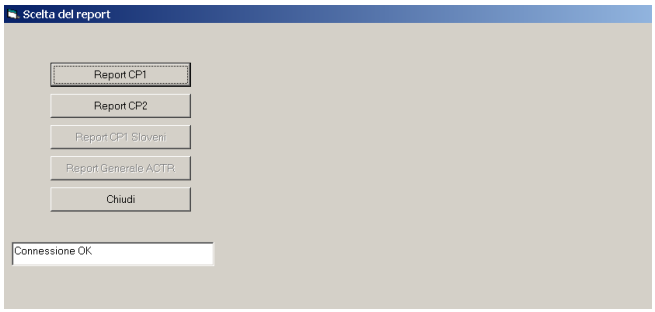
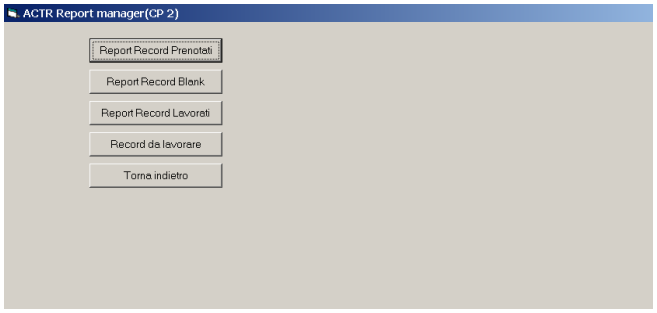
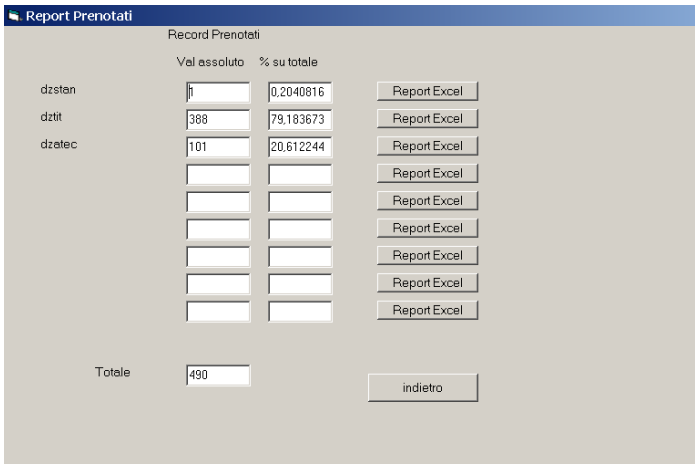


Figura 7.7: Interfaccia per l'accesso alle diverse funzioni di reportistica



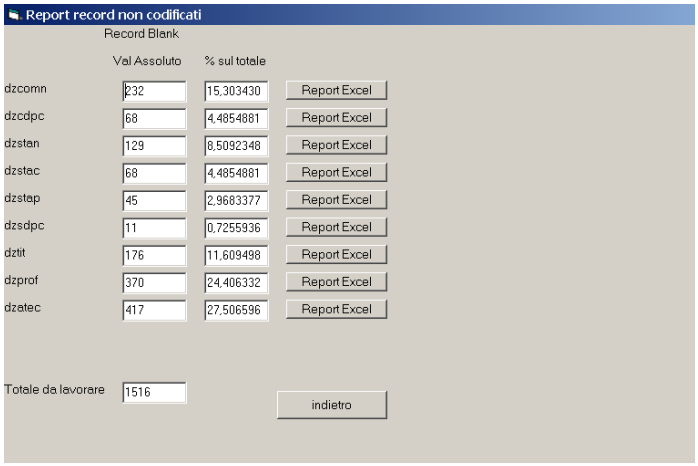
L'operatore logistico può accedere a diverse opzioni di reportistica (Figura 7.7); ad esempio, attivando il tasto *"REPORT RECORD PRENOTATI"* si possono analizzare il numero di casi in stato di prenotazione (Figura 7.8) ed esportare in formato Excel il relativo listato.

Figura 7.8: Report record in stato di prenotazione



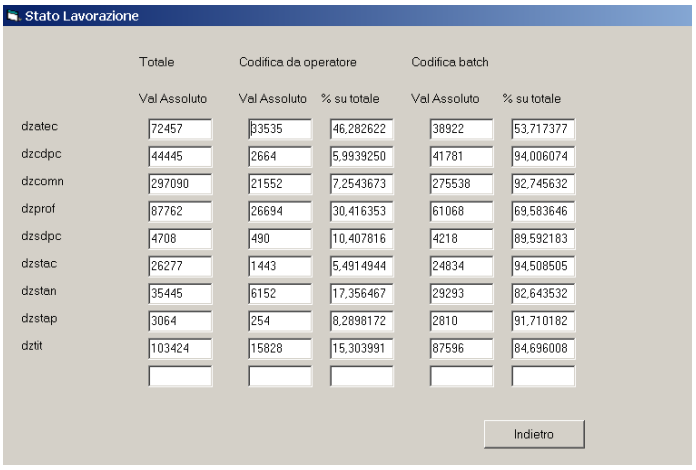
Se si ha necessità di analizzare il numero di casi a cui l'operatore di codifica non è riuscito ad attribuire un codice alle stringhe (Figura 7.9), potrà essere consultato il *"REPORT RECORD BLANK";* anche in questo caso è possibile esportare in formato Excel il relativo listato.

Figura 7.9: Report record blank



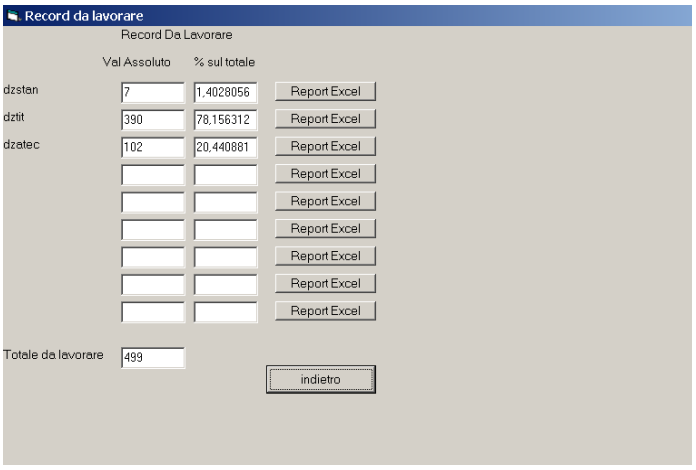
L'opzione *"REPORT RECORD LAVORATI"* offre, invece, un quadro di sintesi, per singola variabile, delle codifiche effettuate sia attraverso il riconoscimento automatico che dagli operatori di codifica (Figura 7.10)[23].

Figura 7.10: Report record lavorati



Infine si può anche accedere, tramite il tasto *"RECORD DA LAVORARE"*, all'analisi delle stringhe che ancora non sono state lavorate con la possibilità di esportare il relativo listato (Figura 7.11)[24].

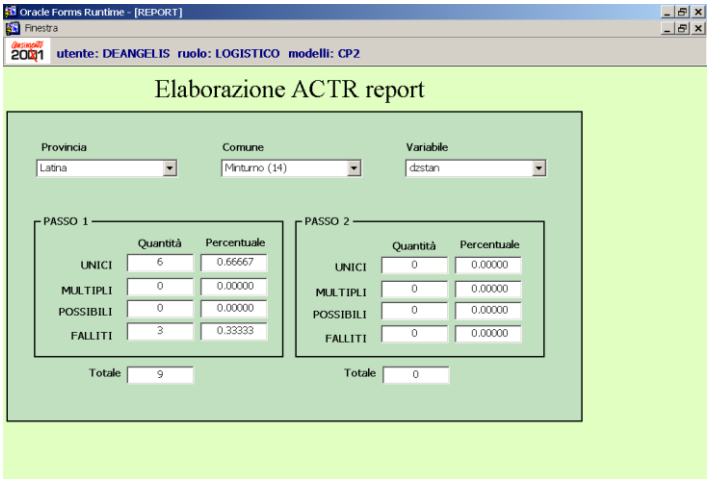
Figura 7.11: Report record da lavorare



Attraverso il tasto "*ESITO LAVORAZIONE BATCH"* dalla schermata di Figura 7, l'Operatore Logistico può accedere alla reportistica relativa all'attività di codifica evasa in automatico.

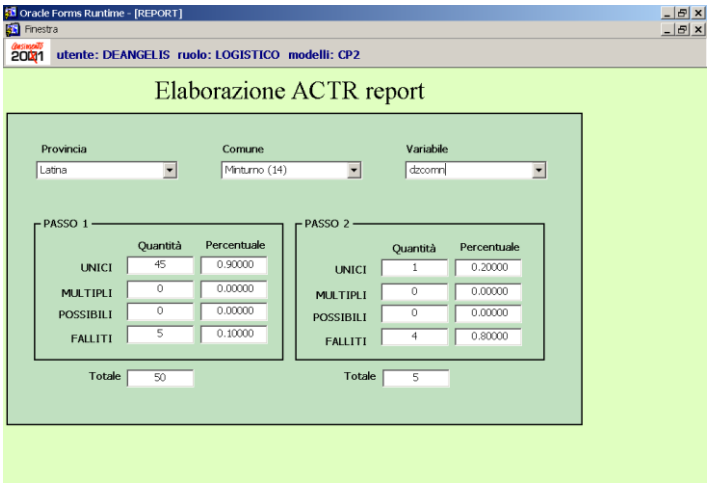
[23] Nel conteggio sono inclusi anche i casi di codifiche effettuate senza attribuzione del codice.
[24] Nel conteggio sono inclusi anche le stringhe in stato di prenotazione.

Figura 7.12: Esito *batch* DZSTAN



Vediamo un esempio (Figura 7.12). Rispetto al comune di Minturno in provincia di Latina sono state sottoposte a procedura *batch* nove stringhe relative alla variabile "Stato Estero di nascita" (DZSTAN). L'esito del processo, riportato nel riquadro *PASSO1* è così riassumibile: sei stringhe sono state riconosciute dal *software* ACTR come "UNICI" e codificate in automatico mentre tre sono risultate "FALLITE". Nella schermata è anche presente un riquadro *PASSO2*. Ricordiamo, infatti, che per alcune variabili, quali ad esempio il 'Comune di nascita' (DZCOMN), il 'Comune di dimora abituale nel 2000' (DZCDPC) e il "Titolo di studio" (DZTIT), sono stati progettati due passaggi della procedura *batch. N*el primo passo il *software* ACTR analizza le stringhe tenendo conto del filtro che nel caso dei "Comuni" è rappresentato dalle sigle delle province italiane; i testi in relazione ai quali le codifiche che risultassero "MULTIPLE" o "POSSIBILI" o "FALLITE" verranno sottoposti ad un secondo passaggio *batch*, ovvero il *PASSO2*, nel quale il filtro non verrà tenuto in considerazione. Ad esempio, nella schermata di Figura 7.13 è riportato l'esito della lavorazione *batch* relativo alla variabile 'Comune di nascita' (DZCOMN).

Figura 7.13: Esito *batch* DZCOMN



Nel complesso sono state sottoposte a procedura di riconoscimento automatico 50 stringhe: nel *PASSO1*, *batch* con filtro, sono state codificate come "UNICI" ben 45 delle suddette, mentre solo in cinque casi l'esito è stato un "FALLITO". Nel *PASSO2*, *batch* senza filtro, le cinque stringhe fallite sono state ancora una volta processate e l'esito è risultato il seguente: in un caso è stato possibile attribuire il codice, mentre i restanti quattro casi sono rimasti irrisolti. I file multipli, possibili e falliti del primo o secondo passo (lì dove previsto) necessiteranno della valutazione da parte degli operatori di codifica.

5 I risultati ottenuti: analisi quantitativa e qualitativa

5.1 Le codifiche effettuate in outsourcing sui testi acquisiti tramite lettura ottica

Come già specificato nel capitolo 2, le variabili testuali Comune, Stato Estero e Titolo di studio contenute nei modelli di rilevazione delle persone in famiglia (mod. Istat CP.1), sono state codificate in *outsourcing*, a cura del consorzio di ditte che ha curato anche l'acquisizione dei dati relativi agli stessi modelli tramite lettura ottica.
In totale la Elsag, capogruppo del consorzio aggiudicatario della gara di appalto, ha provveduto a codificare oltre 58 milioni di stringhe ed in particolare:

Tabella 5.1 – Codifiche effettuate in outsourcing

Variabili	*Comune*	*Stato Estero*	*Titolo di studio*	*Totale*
Testi codificati	40.922.421	3.935.689	13.251.889	58.109.999

Le percentuali di attribuzione del codice così come i livelli di accuratezza sono stati verificati, per ogni singolo invio da parte della Elsag, da una seconda ditta esterna incaricata di certificare che ogni lotto di informazioni rispettasse, sia in termini quantitativi che qualitativi, tutti i parametri stabiliti in sede contrattuale. In fase di monitoraggio, in caso di mancato rispetto di anche uno solo dei requisiti stabiliti dall'Istat, si è provveduto a rispedire alla ditta il set di dati inadempienti rispetto alle condizioni prefissate, affinché la stessa provvedesse al perfezionamento necessario per l'accettazione del pacchetto da parte del committente.

Benché a causa degli elevati costi da sostenere[25], la Elsag non fosse stata formalmente incaricata di procedere anche con l'attribuzione dei codici alle variabili Professione e Attività Economica rilevate nei Fogli di famiglia, tuttavia la stessa (a cui erano stati comunque forniti il dizionario delle Attività Economiche perfezionato in Istat e la Classificazione delle Professioni 2001), come previsto nella documentazione tecnica economica presentata in sede di gara, ha prodotto 4.343.542 codifiche per la prima variabile su un totale di 20.837.427 descrizioni lette otticamente (20.84%) e 3.322.721 codifiche per la seconda, pari al 17.24% del totale delle stringhe (19.268.202) acquisite anch'esse tramite lettura ottica. Nonostante da una prima analisi effettuata la percentuale di attribuzione dei codici sembri quantomeno distribuirsi in maniera abbastanza uniforme su tutte le 103 province italiane, resta tuttavia ancora da verificare l'affidabilità in termini qualitativi dei codici in questione dal momento che, non essendo tale attività inserita tra quelle da evadere contrattualmente, non è stato perfezionato alcun tipo di monitoraggio sul set di dati in questione da parte della ditta incaricata di certificare la qualità del lavoro espletato in *outsourcing.* Qualora il livello di accuratezza di tali codici risulti in linea con gli standard qualitativi già ottenuti a seguito di codifiche precedentemente espletate in materia di Professione e Attività Economica (ad esempio con quelli raggiunti dalla codifica delle stesse variabili rilevate nell'ambito delle Convivenze in occasione sempre del Censimento 2001[26]), questi potranno comunque essere considerati nella fase di progettazione del campione da estrarre per la diffusione dei dati al massimo dettaglio relativo alle variabili economiche di interesse.

[25] Cfr. paragrafo 2.1
[26] Cfr.paragrafo 5.4

5.2 La codifica in house dei Fogli di famiglia

Il 14° Censimento della popolazione ha coinvolto circa 21 milioni di famiglie, 57 milioni di individui, 25 milioni di abitazioni, 100.000 rilevatori, 10.000 coordinatori comunali, 8.100 comuni, 1.000 coordinatori provinciali e 103 camere di commercio. E' facile intuire che numeri di tale entità sottintendono un impianto organizzativo decisamente sostenuto e, per quanto tale, non privo di eccezioni e difficoltà di cui si è dovuto tener conto sia durante la progettazione della rilevazione che in corso d'opera. Ad esempio, è stato necessario realizzare modelli di rilevazione in lingua slovena per alcune minoranze linguistiche residenti al confine, che, a fronte dell'analisi dei costi da sostenere, sono stati acquisiti non tramite lettura ottica ma attraverso il tradizionale *data entry* e per i quali, a seguito della traduzione in lingua italiana effettuata, si è provveduto all'interno dell'Istituto per la codifica delle variabili testuali.

In relazione ai Fogli di famiglia sloveni sono state codificate 12.704 variabili testuali ed in particolare:

Tabella 5.2 – Risultati della codifica delle variabili testuali contenute nei modelli di rilevazione CP.1 sloveni

VARIABILI	*CODIFICHE BATCH*	*RECALL RATE DI ACTR*	*CODIFICHE MANUALI*	*PERCENTUALE CODIFICHE MANUALI*	*TOTALE CODIFICHE EFFETTUATE*
Comune di nascita	3.618	85,7	603	14,3	4.221
Comune di dimora abituale precedente	91	84,3	17	15,7	108
Comune abituale di studio/lavoro	2.145	83,9	413	16,1	2.558
Stato estero di nascita	1.362	93,7	91	6,3	1.453
Stato estero di cittadinanza	134	81,2	30	18,8	164
Stato estero di cittadinanza precedente	670	87,0	100	13,0	770
Stato estero di dimora abituale precedente	36	85,7	6	14,3	42
Stato estero abituale di studio e lavoro	73	88,0	10	12,0	83
Titolo di studio	2.804	84,3	500	15,7	3.304
Totale	**10.933**	**85,9**	**1.770**	**14,1**	**12.703**

La performance di ACTR, nel caso dei modelli sloveni, per la variabile Comune non va mai oltre l'85,7%, valore che si colloca al di sotto delle percentuali di codici attribuiti in *batch* raggiunte in media nella fase di codifica sia dei modelli CP.1 integrativi che dei Fogli di convivenza (cfr. Tabella 5.3 e Tabella 5.5.). Tale risultato è presumibilmente da attribuirsi al "disagio" causato dal dover trattare descrizioni non originali ma frutto di una traduzione che, in alcuni casi, può aver comportato la composizione di testi particolarmente lontani dalle dizioni ufficiali, dai sinonimi e dalle empiriche presenti all'interno dei dizionari. Tra le codifiche effettuate dagli operatori manuali solo in 32 casi (1,8% dei testi codificati *on line*, 0,25% sul totale dei testi trattati) alla descrizione fornita dal rispondente non è stato possibile assegnare alcun codice. Da sottolineare, inoltre, che nell'ambito dei modelli CP.1 sloveni (così come in quelli integrativi), per quanto detto nel paragrafo 2.1, non sono state codificate le variabili Professione e Attività Economica, mentre sono state oggetto di codifica il 'Comune abituale di studio o lavoro' e lo 'Stato Estero di studio o lavoro' che, invece, non sono rientrate nell'attività di codifica del modelli CP.2 dal momento che, in relazione alle convivenze, in occasione del Censimento della Popolazione 2001, non è stata effettuata la rilevazione degli spostamenti sistematici per motivi di studio o lavoro.

A causa dell'articolata rete di rilevazione sul territorio messa a punto per la corretta realizzazione dell'evento censuario, si è dovuto tener conto di alcune esigenze dei comuni, particolarmente gravati dai numerosi compiti connessi con la fase di consegna, di raccolta e di revisione dei modelli, nonché con il confronto dei dati censuari con le risultanze anagrafiche stabilito per legge. Nell'ambito di tali esigenze è stato necessario, ad esempio, gestire l'eventualità che alcuni comuni inviassero in ritardo alcuni modelli CP.1 ad integrazione di quelli già trasmessi alla ditta per la lettura ottica per la rilevazione di individui precedentemente sfuggiti alla conta. Tali modelli sono stati quindi registrati manualmente e codificati *in house* attraverso il sistema di codifica automatica e *computer-assisted* con i seguenti risultati:

Tabella 5.3 – Risultati della codifica delle variabili testuali contenute nei modelli di rilevazione CP.1 integrativi

VARIABILI	***CODIFICHE BATCH***	***RECALL RATE DI ACTR***	***CODIFICHE MANUALI***	***PERCENTUALE CODIFICHE MANUALI***	***TOTALE CODIFICHE EFFETTUATE***
Comune di nascita	44.990	94,4	2.669	5,6	47.659
Comune di dimora abituale precedente	3.552	92,9	271	7,1	3.823
Comune abituale di studio/lavoro	9.737	92,9	741	7,1	10.478
Stato estero di nascita	8.671	86,4	1.362	13,6	10.033
Stato estero di cittadinanza	6.658	95,1	342	4,9	7.000
Stato estero di cittadinanza precedente	1.249	92,7	99	7,3	1.348
Stato estero di dimora abituale precedente	1.150	91,9	101	8,1	1.251
Stato estero abituale di studio e lavoro	149	79,3	39	20,7	188
Titolo di studio	27.617	90,0	2.987	10,0	30.604
Totale	**103.773**	**92,3**	**8.611**	**7,7**	**112.384**

Come per i Fogli di famiglia sloveni, anche in questo caso non sono state codificate le variabili Professione e Attività Economica, bensì il Comune e lo Stato Estero relativo agli spostamenti pendolari per motivi di studio o di lavoro. A differenza del caso precedente, la percentuale di attribuzione in *batch* dei codici alla variabile Comune per i CP.1 integrativi non è mai scesa al di sotto del 92.9%. Nella fase *computer-assisted* non è stato possibile codificare solo 291 testi (3.4% dei testi codificati *on line*, 0.26% del totale delle descrizioni processate).

Verificata la buona qualità dei testi letti otticamente ed il buon livello di precisione dei codici assegnati in automatico attraverso ACTR, sono stati processati all'interno della Direzione Centrale Censimenti della Popolazione, Territorio e Ambiente anche le descrizioni relative alle variabili Comune e Stato Estero dei modelli CP.1 non codificate in *outsourcing*[27] con i risultati riportati nella seguente tabella.

[27] cfrt. Par.2.1

Tabella 5.4 – Recall rate di ACTR sulle variabili testuali lette otticamente

Variabili	***Codifiche batch***	***Recall rate di Actr***	***Totale Testi Processati***
Comune	255.484	35,2	726.692
Stato Estero	87.943	41,7	211.126
Totale	343.427	36,6	937.818

L'abbattimento delle percentuali di testi codificati in automatico tramite ACTR deriva dal fatto di aver processato descrizioni acquisite tramite lettura ottica non sottoposte a procedure di videocorrezione.

*5.3 I risultati ottenuti sui dati delle convivenze, analisi **quantitativa***

L'attività di codifica delle variabili testuali rilevate all'interno dei Fogli di convivenza (mod.Istat CP.2) ha rappresentato l'obiettivo primario in funzione del quale all'interno della Direzione Centrale del Censimento della Popolazione, Territorio e Ambiente è stato realizzato il sistema per la codifica automatica e *computer-assisted* delle stringhe alfabetiche. A seguito del Censimento 2001 sono stati rilevati 401.723 individui abitualmente dimoranti in ospedali, istituti penitenziari, istituti assistenziali, case di riposo per anziani, convivenze ecclesiastiche e militari, ecc., in relazione ai quali sono state acquisite 674.174 descrizioni afferenti alle variabili Comune, Stato Estero, Titolo di studio, Professione e Attività Economica. I testi, acquisiti tramite il tradizionale *data entry*, sono stati tutti processati all'interno del sistema di codifica con i seguenti risultati:

Tabella 5.5 –Risultati della codifica delle variabili testuali contenute nei modelli di rilevazione CP.2

VARIABILI	***CODIFICHE BATCH***	***RECALL RATE DI ACTR***	***CODIFICHE MANUALI***	***PERCENTUALE CODIFICHE MANUALI***	***TOTALE CODIFICHE EFFETTUATE***
Comune di nascita	275.538	92,7	21.552	7,3	297.090
Comune di dimora abituale precedente	41.781	94,0	2.664	6,0	44.445
Stato estero di nascita	29.293	82,6	6.159	17,4	35.452
Stato estero di cittadinanza	24.834	94,5	1.443	5,5	26.277
Stato estero di cittadinanza precedente	2.810	91,7	254	8,3	3.064
Stato estero di dimora abituale precedente	4.218	89,6	490	10,4	4.708
Titolo di studio	87.596	84,4	15.828	15,6	103.424
Professione	61.068	69,6	26.189	30,04	87.257
Attività economica	38.922	53,6	33.535	46,4	72.457
Totale	**566.060**	**83,8**	**108.114**	**16.2**	**674.174**

Delle 674.174 descrizioni 566.060, ovvero l'83,8%, sono state codificate in *batch* mentre nel 16,2% dei casi (108.114) è stato necessario l'intervento degli operatori manuali. Tra i testi trattati nella fase *computer-assisted* solo nell'1,3% dei casi (0,2 per cento del totale delle descrizioni) non è stato possibile selezionare alcun codice.

In termini quantitativi, ACTR ha garantito livelli di attribuzione del codice in automatico prossimi a quelli raggiunti in occasione dei test effettuati sui dati delle due indagini pilota pre-censuarie [28]. Si fa presente che, sia in occasione del Censimento 2001 che dei test effettuati sui dati delle due indagini pilota, sono stati selezionati in automatico solamente i codici riconosciuti da ACTR come "match unici"; non sono stati altresì previsti algoritmi per la selezione *batch* del codice tra quelli che il software di codifica automatica ha inserito tra i "match multipli" e i "match possibili".

Tabella 5.6 – Confronto performance ACTR

Variabili	*Rilevazione corrente Forze di Lavoro – Indagine pilota*	*Prima Indagine Pilota – Censimento 2001*	*Seconda Indagine Pilota – Censimento 2001*	*Censimento 2001*
Comune			94,5	93,4[29]
Stato estero			83,2	89,6[30]
Titolo di studio		75,7	87,0	84,4
Professione	66,7	65,5	68,8	69,6
Attività Economica	43,5	51,2	51,9	53,6

Rispetto alle sperimentazioni effettuate sui dati delle Forze di Lavoro e delle due indagini pre-censuarie, la performance di ACTR su Professione e Attività Economica è sensibilmente migliorata passando per la prima da un *recall rate* minimo del 65,5% registrato sui testi rilevati nel 2000 al 69,6% raggiunto con il Censimento e, per la seconda, dal 43,5% ottenuto sui dati delle Forze di Lavoro al 53,6%. E' altresì leggermente peggiorata l'efficacia di ACTR in relazione alla variabile Comune. I test effettuati a seguito della seconda indagine pilota avevano assicurato, infatti, l'attribuzione del 94,5% dei codici in automatico a fronte del 93,4% ottenuto nell'ambito della codifica dei dati censuari. Il lieve decremento della percentuale di codifica automatica raggiunta potrebbe essere dovuta al fatto che il software era stato precedentemente utilizzato soltanto una volta e su un campione abbastanza esiguo che non presentava quindi una variabilità di testi elevata, quanto lo è, invece, quella dei 341.535 testi codificati in questa occasione.

In relazione alla variabile Titolo di studio, l'efficacia di ACTR è migliorata rispetto a quella registrata in occasione della prima indagine pilota (75,7%), ma è diminuita se rapportata alle percentuali di codifica *batch* raggiunte con la pilota del 2000 (87%). Se è vero da un lato che, rispetto alla seconda indagine pilota, non solo è stata perfezionata una nuova classificazione in materia di titoli di studio conseguiti nel nostro Paese, ma è stato anche implementato un dizionario contenente oltre 3.000 voci con una conseguente elevata probabilità di aumentare la *performance* del software, è altresì possibile che la presenza di titoli di studio tra quelli di cui specificare la tipologia quali i "diplomi non universitari post-secondari" particolarmente soggetti a descrizioni "variegate" associata alla particolare categoria di individui (residenti in Convivenza) di cui sono stati processati i dati (basti pensare ai religiosi che all'interno del modello di rilevazione, nello spazio dedicato alla specifica del titolo di studio più elevato conseguito, hanno descritto corsi di studio spesso non assimilabili a quelli attualmente contemplati nel sistema di istruzione italiano) possa aver compromesso, anche se minimamente, la funzionalità del *software*. Il Titolo di studio, peraltro, è la variabile in relazione alla quale ACTR, anche in termini di accuratezza, ha fatto registrare un sensibile decremento rispetto alle esperienze passate[31].

[28] La prima indagine pilota è stata effettuata a ottobre 1998, la seconda a ottobre 2000
[29] Media delle percentuali di codifiche effettuate in automatico relative alla variabile Comune
[30] Media delle percentuali di codifiche effettuate in automatico relative alla variabile Stato estero
[31] Cfr.paragrafo 1.3

Per quanto riguarda, invece, lo Stato Estero, l'efficacia di ACTR è migliorata di oltre sei punti percentuali passando dall'83,2% dei testi codificati in *batch* in occasione della seconda indagine pilota all'89,6% raggiunto con la codifica delle stringhe rilevate nel 2001 attraverso i modelli CP.2. L'implementazione del dizionario[32] basato sulla classificazione degli Stati Esteri 2001 redatta all'interno dell'Istat con alcuni nomi di capitali, i codici ISO a una, due o tre lettere, gli aggettivi maschile e femminile di cittadinanza, denominazioni di uso locale quale Al Magrib per Marocco o denominazioni non più in uso quale ex Zaire per Repubblica Democratica del Congo o ex Ceylon per Sri Lanka ha fattivamente contribuito in termini di *performance* del software di codifica.

Passando alla fase "*computer assisted*", come si può osservare nella tabella 5.5, sono state codificate 108.114 stringhe di cui 24.216 relative alla variabile Comune, 8.346 inerenti lo Stato Estero, 15.828 il Titolo di studio, 26.189 la Professione e 33.535 l'Attività Economica. Hanno partecipato all'attività di codifica delle variabili testuali contenute nei modelli di rilevazione delle Convivenze cinque operatori manuali della Direzione Centrale del Censimento della Popolazione, Territorio e Ambiente. I cinque operatori (di cui tre laureati, uno munito di diploma universitario di statistica ed uno a pochi esami dal diploma di laurea) sono stati formati innanzitutto sulla natura delle variabili da codificare e sulla struttura delle classificazioni ufficiali afferenti a ciascuna di esse nonché sugli *step* da seguire per la corretta attribuzione del codice (prevedendo ad esempio per il Titolo di studio una prima consultazione "*on line*" dei dizionari subordinata all'indicazione del "filtro" da parte del rispondente, una seconda senza alcun tipo di vincolo, nonché la consultazione rapida delle altre risposte fornite dallo stesso individuo all'interno del questionario e così via[33]). Gli stessi hanno altresì fattivamente collaborato sia nella fase "*training on the job*", fornendo preziose indicazioni per il miglioramento in corso d'opera dell'intero sistema di codifica approntato anche per garantire la corretta e rapida soluzione dei casi che, non codificati in *batch*, rappresentavano la casistica più complessa in termini di attribuzione dei codici, che per l'implementazione dei vari dizionari, comunicando di volta in volta le empiriche più significative e ricorrenti da inserire nei *reference file*, previa analisi e valutazione dei testi da parte dei responsabili dell'attività di codifica.

Nella tabella che segue si riporta una valutazione dei giorni uomo che sono stati necessari per portare a termine l'attività di codifica manuale delle descrizioni alfabetiche rilevate all'interno dei Fogli di Convivenza sulla base della stima del tempo medio necessario per l'attribuzione del codice a ciascuna stringa:

[32] Cfr.paragrafo 1.2
[33] Cfr.paragrafo 2.3

Tavola 5.7 - Stima dei tempi necessari per la codifica computer-assisted delle variabili alfabetiche

Variabili	*Numero di campi codificati nella fase computer-assisted*	*Stima del tempo necessario per la codifica computer-assisted di ciascun testo*	*Stima del numero di campi che si codificano in un'ora*	*Stima del numero di campi che si codificano in una giornata di lavoro*[34]	*Stima dei giorni necessari per la codifica manuale delle variabili testuali rilevate nei Fogli di Convivenza*
Comune	24.216	30 secondi	120	600	40
Stato estero	8.346	45 secondi	80	400	20
Titolo di studio	15.828	75 secondi	48	240	66
Professione	26.189	75 secondi	48	240	109
Attività Economica	33.535	90 secondi	40	200	168

Si fa presente che le stime sopra riportate e l'elevato standard di qualità del lavoro effettuato risentono non solo del fatto che gli operatori manuali che hanno partecipato all'avventura censuaria si identificano in soggetti particolarmente preparati e muniti di un titolo di studio piuttosto elevato, ma anche del fatto che gli stessi, lavorando a stretto contatto sia tra di loro che con i diretti responsabili di tutta la procedura hanno potuto facilmente gestire in tempo reale qualsiasi sorta di problematica di natura tecnica, ovvero connessa con malfunzionamenti di tipo informatico del sistema, o di tipo concettuale attraverso scambi di suggerimenti e consigli per la corretta e celere risoluzione dei casi più ostici.

Naturalmente una distribuzione sul territorio più complessa del carico di lavoro ed il coinvolgimento di personale più numeroso e soprattutto con un livello di istruzione inferiore potrebbe inevitabilmente comportare un innalzamento dei tempi medi di evasione dell'attività di codifica e, soprattutto, una diminuzione del livello di accuratezza relativo all'attribuzione dei codici. Tuttavia, anche in caso di una organizzazione più articolata dell'attività di codifica, un'attenta fase di addestramento del personale coinvolto in termini di

- formazione sulle classificazioni ufficiali e sulla struttura dei dizionari
- formazione relativa agli *step* procedurali da seguire per l'assegnazione del codice
- addestramento per il corretto utilizzo del sistema di codifica perfezionato per l'avventura censuaria
- implementazione dinamica in corso d'opera dei dizionari

potrebbero comunque garantire degli standard accettabili, sia in relazione alla qualità del dato che al tempo necessario per la conclusione dei lavori.

[34] E' stata calcolata una media, ipotizzando 5 ore di lavoro nette al giorno

*5.4 I risultati ottenuti sui dati delle convivenze, analisi **qualitativa***

Come descritto nei paragrafi precedenti, il processo di codifica sui dati delle convivenze è stato organizzato in due step: la codifica automatica con ACTR e la codifica assistita di quanto non è stato possibile codificare automaticamente.
In funzione dell'architettura implementata per la codifica assistita (cfr. cap. 2 e 3) che consentiva un pieno controllo delle operazioni ed era stato preceduto da una formazione dei codificatori manuali curata approfonditamente sia dagli esperti delle classificazioni che dagli esperti dei sistemi di codifica, si è ritenuto che non fosse necessario sottoporre al controllo di qualità i risultati di questa fase, ma soltanto quelli della codifica automatica.
Per questi ultimi, è stata adottata la metodologia descritta nel paragrafo 1.5. Il primo passaggio è stato quindi quello di effettuare, sui testi di ciascuna variabile, il "*parsing grezzo*" così da individuare i testi "*diversi"*, con le relative frequenze. Sono state quindi definite le classi di frequenza ed estratti i campioni da sottoporre a controllo di qualità.
Come può vedersi dalla tabella 5.8, il numero di testi effettivamente *"diversi"* rispetto a quelli rilevati e codificati cala drasticamente. In particolare, poi, per la variabile Stato Estero non si è ritenuto necessario estrarre un campione perché la numerosità dei testi *"diversi"* a seguito del *parsing* si è rivelata talmente contenuta che è stato possibile sottoporre a validazione tutto l'universo di riferimento.
E' interessante, peraltro, rilevare la variabilità linguistica dimostrata dai rispondenti nei quesiti di ciascuna variabile, che ha indubbiamente impattato sia sull'efficacia che sull'accuratezza. Il numero di testi effettivamente diversi evidenzia la variabilità linguistica per ciascuna variabile: pur non tenendo infatti conto della numerosità dei codici prevista da ciascuna delle classificazioni, è chiaro che i quesiti di Professione ed Attività Economica sono quelli maggiormente problematici da questo punto di vista. Questa affermazione è ulteriormente avvalorata dal fatto che queste due classificazioni contemplano un numero di codici contenuto, nettamente inferiore, per esempio, rispetto a quelli della classificazione dei comuni.

Tabella 5.8 Risultati della fase di 'parsing grezzo'

Variabile	***Numero di testi originari***[35]	***Numero di Testi 'diversi'***
Professione	61.068	3.270
Attività Economica	37.957	3.028
Titolo di studio	86.433	5.595
Stato Estero	60.850	751
Comune	310.884	12.295

Proseguendo con il secondo step della metodologia (definizione delle classi di frequenza ed estrazione dei campioni), si è quindi ottenuto un abbattimento dei testi da sottoporre all'analisi dei codificatori manuali che ha consentito di realizzare questa attività con un'équipe molto ridotta in tempi molto brevi.[36]

[35] Il numero di testi codificati automaticamente relativi alle variabili Comune, Stato estero, Titolo di studio e Attività Economica in questa colonna è lievemente differente da quello della tabella .5.5 perché l'analisi di qualità è stata realizzata dopo la revisione quantitativa dei dati censuari.
[36] Si precisa che per l'analisi qualitativa ci si è avvalsi della supervisione di Loredana Mazza, in qualità di esperta delle classificazioni e delle relative applicazioni di codifica automatica.

Tabella 5.9 ***Variabile Professione:*** *numerosità campionaria*

Classe di frequenza	***Testi 'diversi'***		***Numerosità campionaria***	***Frazione di campionamento***
	Numero.	***Incidenza dello strato***		
1	1.950	3,19%	139	7,13%
2 - 4	692	2,93%	175	25,29%
5 - 15	337	4,60%	160	47,48%
16 - 70	183	10,66%	127	69,40%
71 - 249	70	14,32%	38	100,00%
oltre 249	38	64,30%	70	100,00%
Totale	3.270		**709**	

Tabella 5.10 ***Attività Economica:*** *numerosità campionaria*

Classe di frequenza	***Testi 'diversi'***		***Numerosità campionaria***	***Frazione di campionamento***
	Numero.	***Incidenza dello strato***		
1	1.859	4,91%	139	7,48%
2 - 4	631	4,32%	171	27,10%
5 - 15	324	6,84%	157	48,46%
16 - 70	147	12,44%	109	74,15%
71 - 499	55	25,90%	55	100,00%
oltre 499	12	45,59%	12	100,00%
Totale	3.028		**643**	

Tabella 5.11 ***Titolo di studio:*** *numerosità campionaria*

Classe di frequenza	***Testi 'diversi'***		***Numerosità Campionaria***	***Frazione di campionamento***	***Numerosità campionaria in funzione del filtro***
	Numero.	***Incidenza dello strato***			
1	3.631	1,53%	220	7,48%	220
2 - 5	1.169	11,56%	243	27,10%	383
6 - 39	555	12,60%	238	48,46%	710
40 - 99	122	11,85%	94	74,15%	424
100 - 499	85	16,67%	85	100,00%	534
oltre 499	33	29,04%	33	100,00%	332
Totale	5.595		**913**		2.603

Tabella 5.12 ***Variabile Comune:*** *numerosità campionaria*

Classe di frequenza	***Testi 'diversi'***		***Numerosità Campionaria***	***Frazione di campionamento***	***Numerosità campionaria in funzione del filtro***
	Numero.	***Incidenza dello strato***			
1 - 2	3.732	1,53%	144	3,82%	150
3 - 15	4.700	11,56%	223	4,78%	344
16 - 30	1.792	12,60%	261	14,92%	544
31 - 50	948	11,85%	231	25,28%	564
51 - 110	717	16,67%	264	36,16%	794
111 - 297	298	16,75%	199	65,97%	791
oltre 298	108	29,04%	108	100,00%	519
Totale	12.295		**1.430**		3.706

Le variabili Titolo di Studio e Comune presentavano una caratteristica peculiare rispetto alle altre variabili: l'assegnazione del codice doveva tener conto della risposta al quesito pre-codificato rispettivamente sul livello e durata del titolo di studio e della sigla della provincia. Per questo motivo, anche l'analisi di qualità è stata effettuata considerando i valori di queste variabili "filtro". A tal fine, testi uguali tra di loro, ma ai quali corrispondevano filtri diversi sono stati validati separatamente. E' questo il significato della colonna *'Numerosità Campionaria in funzione del filtro'* riportata nelle tabelle 5.11 e 5.12.

I campioni così estratti sono stati quindi sottoposti all'analisi dei codificatori esperti; i risultati sono riportati nelle tabelle seguenti.

Tabella 5.13 Precision rate: sintesi su tutte le variabili

Variabile	***Precision***	***Precision Ponderata rispetto ai Testi originari***
Professione	93,23	99,47
Attività Economica	90,00	92,31
Titolo di studio	92,12	98,32
Stato Estero	99,99	99.99
Comune	98,81	99,87

Tabella 5.14 ***Variabile Professione:*** *precision rate*

Classe di frequenza	***Precision rate rispetto ai Testi 'diversi'***	***Precision rate rispetto ai Testi originari***
1	85,61	85,61
2 - 4	92,57	93,46
5 - 15	93,13	92,07
16 - 70	96,85	97,40
71 - 249	100,00	100,00
oltre 249	100,00	100,00

Tabella 5.15 ***Variabile Attività Economica:*** *precision rate*

Classe di frequenza	***Precision rate rispetto ai Testi 'diversi'***	***Precision rate rispetto ai Testi originari***
1	84,17	84,17
2 - 4	92,98	92,14
5 - 15	88,54	88,41
16 - 70	89,91	90,18
71 – 499	98,18	98,20
oltre 499	91,67	89,81

Tabella 5.16 ***Variabile Titolo di studio:*** *precision rate*

Classe di frequenza	***Precision rate rispetto ai Testi 'diversi'***	***Precision rate rispetto ai Testi originari***
1	86,82	86,82
2 – 5	84,07	85,04
6 – 39	93,23	95,14
40 – 99	92,65	95,95
100 – 499	94,19	98,61
oltre 499	98,49	99,02

Tabella 5.17 ***Variabile Comune:*** *precision rate*

Classe di frequenza	***Precision rate rispetto ai Testi 'diversi'***	***Precision rate rispetto ai Testi originari***
1 – 2	98,00	98,40
3 – 15	98,84	99,72
16 – 30	98,53	99,18
31 – 50	97,70	99,32
51 – 110	99,62	99,68
111 - 297	99,24	99,94
oltre 298	98,65	99,99

Tabella 5.18 ***Variabile Stato estero/Cittadinanza:*** *precision rate*

Classe di frequenza	***Precision rate rispetto ai Testi 'diversi'***	***Precision rate rispetto ai Testi originari***
1	85,61	85,61
2 – 4	92,57	93,46
5 – 15	93,13	92,07
16 – 100	96,85	97,40
101 – 499	100,00	100,00
oltre 499	100,00	100,00

Come si può rilevare, l'accuratezza è sempre coerente con i risultati ottenuti nel corso delle precedenti esperienze e, in alcuni casi superiore.

Per l'Attività Economica, l'accuratezza sarebbe più elevata (97,71%) se calcolata al netto di un errore nell'attribuzione del codice ad una risposta molto frequente ("Servizio d'istituto"), dovuta ad una carenza del dizionario elaborabile. Questo errore è stato riscontrato in corso d'opera durante il monitoraggio del processo di codifica; si è provveduto quindi ad effettuare l'integrazione del dizionario, così da non perseguire nell'errore, ed è stata individuata l'azione correttiva da contemplare in fase di check.

In generale, poi, l'accuratezza è sempre più elevata per i testi appartenenti alle classi di frequenza più elevate, il che garantisce rispetto alla possibilità di apprezzabili distorsioni delle stime attribuibili ad un processo automatizzato. Ciò è evidenziabile nelle tabelle sopra riportate, dalle quali emerge che l'unica eccezione è costituita dalla variabile Attività Economica, per la quale si è verificato il citato errore inerente la risposta testuale "Servizio d'istituto".

E' interessante anche analizzare i motivi che hanno indotto all'individuazione di codici errati. Le principali cause di errore sono:

- errori di ortografia
- carenze dei dizionari elaborabili
- risposte ambigue e/o non esaustive
- risposte non codificabili.

Il primo fattore ha inciso nel caso di risposte molto brevi (una o due parole), oppure nelle quali l'errore ortografico era presente nella parola più significante. Si precisa però che, laddove è stato possibile, a seconda della disponibilità di una casistica significativa, sono stati introdotti nei dizionari testi affetti dai più frequenti errori ortografici (ciò è stato effettuato, per esempio, per la variabile Attività Economica, per la quale, nel corso dell'esperienza di codifica dei dati del censimento intermedio dell'industria, sono stati recepiti gli errori maggiormente frequenti).

Le carenze dei dizionari hanno potuto far sì che 'modi di dire' non contemplati potessero realizzare match impropri, per via di qualche parola in comune, con descrizioni che contenutisticamente non erano inerenti. Laddove è stato possibile evidenziare il problema in corso d'opera, i dizionari sono stati aggiornati immediatamente, altrimenti gli arricchimenti sono stati effettuati ex post, al fine di garantire un maggior grado di accuratezza per le applicazioni future.

Le risposte ambigue sono quelle per le quali, se la codifica fosse effettuata in corso di intervista, il rilevatore chiederebbe delucidazioni. Nel caso di codifica posteriore alla rilevazione, il codificatore manuale avrebbe dovuto avvalersi di eventuali altre variabili disponibili sul modello, oppure non attribuire il codice per lasciare che venga

imputato in fase di controllo e correzione. L'errore del sistema è quindi stato quello di attribuire un codice piuttosto che lasciare il testo tra i non codificati; anche in questo caso ciò può essere dipeso da match impropri con testi simili.

Le risposte non codificabili sono invece quelle che non hanno un contenuto associabile a nessun codice della classificazione di riferimento; casi tipici sono per esempio quelli inerenti la Professione dove il rispondente ha dichiarato di essere "pensionato" o "disoccupato", invece di non rispondere al quesito, così come specificato nelle regole di compilazione del questionario. Per ovviare ai match impropri, relativamente alla casistica di risposte non pertinenti rilevata nel corso di precedenti esperienze, sono stati inseriti nei dizionari testi associati ad un codice fittizio: "nc = non codificabile", in modo da consentirne un trattamento ad hoc in fase di controllo e correzione. Questa soluzione è stata adottata per le applicazioni di Professione ed Attività Economica; non è stato tuttavia possibile prevedere tutta la casistica di risposte non codificabili.

Volendo fornire un'indicazione quantitativa sulle tipologie di errore riscontrate, gli esperti delle classificazioni hanno classificato come "E" sia gli errori dovuti a errori di ortografia, che quelli dovuti a carenze di dizionario, che a risposte non codificabili, mentre hanno distinto come "D" (dubbi) gli errori attribuibili a risposte ambigue.

Si riporta di seguito una tavola con le casistiche di errore, quantificate in valore assoluto, riscontrate per ciascuna variabile.

Tabella 5.19 Errori per tipologia in valore assoluto

Tipologia di errore	***Professione***	***Attività Economica***	***Titolo di Studio***	***Stato Estero***	***Comune***
'E'	19	36	150	4	6
Di cui err. ortografici	6	1	35	4	1
'D'	29	29	55	2	38

In sintesi, ci si può ritenere soddisfatti della qualità ottenuta dalla codifica automatica.

La variabile più problematica è stata l'Attività Economica, ma, come già detto, ciò non è dipeso dalla complessità della classificazione e dalla variabilità linguistica dimostrata dai rispondenti, perché questi problemi sono comuni alla Professione, per la quale i risultati sono stati migliori sia in termini di efficacia che di accuratezza; la difficoltà di questa variabile è infatti a monte, nella comprensione del quesito da parte del rispondente, quando si tratta di famiglie o individui e non di imprese.

Migliorabili sono inoltre i risultati sul Titolo di studio; nella costruzione degli ambienti applicativi si è stati molto rigorosi rispetto all'utilizzo di termini "corretti" rispetto a quelli ritenuti "definitori" nella classificazione ufficiale; si è verificato, però, che i rispondenti non sono altrettanto rigorosi nell'utilizzo di un termine piuttosto che un altro. Per esempio, i rispondenti hanno spesso selezionato la biffatura corrispondente ai a titoli di studio che contemplano una durata di meno di cinque anni ed utilizzato il termine *"diploma"* (corrispondente a titoli di studio che contemplano una durata di cinque anni), laddove sarebbe stato corretto utilizzare il termine *"qualifica"* (corrispondente a titoli di studio che contemplano una durata di meno di cinque anni). Il sistema, quindi, nel primo passaggio di codifica che utilizza il campo filtro, non ha trovato una corrispondenza di filtro/testo nel dizionario informatizzato; invece nel secondo passaggio di codifica ha prodotto un match unico all'interno però dei titoli di studio che prevedono la parola *diploma/maturità*, ovvero tra i titoli di studio che contemplano una durata di cinque anni. Ciò ha comportato ben 36 errori in valore assoluto, che, ponderati, sono pari a 436 stringhe errate; sarebbe dunque opportuno arricchire il dizionario contemplando più spesso la parola *"diploma"*, in virtù del fatto che il quesito pre-codificato guiderebbe comunque sul livello e la durata del titolo di studio, a prescindere dal vocabolo utilizzato.

Un altro arricchimento da effettuare su questa applicazione sarebbe quello che consentirebbe di recepire i cambiamenti nel tempo dei titoli di studio; molti di questi infatti sono stati negli ultimi anni collocati in corrispondenza di un più elevato livello di istruzione (si pensi ai titoli "infermieristici" che possono attualmente essere conseguiti post-diploma, mentre prima richiedevano un livello di istruzione inferiore). Per come è stata realizzato l'ambiente applicativo, i titoli di studio rientranti in questa casistica sono stati tutti codificati nel livello superiore, a prescindere dal "filtro" selezionato dal rispondente, in base al quale non è stato realizzato alcun match nel primo passaggio di codifica.
Nell'ottica comunque di continuare ad utilizzare queste applicazioni per le indagini che rilevano questi quesiti a testo libero, l'attività di integrazione dei dizionari per tutte le variabili dovrà proseguire utilizzando i risultati ottenuti di volta in volta, al fine garantire performance sempre migliori e di rimanere aggiornati rispetto all'evoluzione delle classificazioni di riferimento.

6 Conclusioni e prospettive

Dall'esperienza effettuata in occasione del Censimento Generale della Popolazione 2001 inerente il trattamento delle variabili a testo libero si può indubbiamente trarre un giudizio positivo, in quanto la codifica automatica ha comportato numerosi vantaggi rispetto alle modalità di lavoro adottate nelle precedenti rilevazioni censuarie, quali: una riduzione in termini di tempo e di risorse umane dedicate all'attività di codifica, la standardizzazione del processo, la possibilità di monitorare i risultati in corso d'opera (sia dal punto di vista quantitativo che qualitativo) e quindi una più elevata qualità degli stessi.
Un altro vantaggio consiste nel fatto che il grande investimento attuato per la costruzione degli ambienti applicativi di ciascuna delle variabili trattate non è stato funzionale esclusivamente a questo Censimento, ma ha prodotto un patrimonio fruibile per tutte le ulteriori indagini che rilevino le medesime variabili. Inoltre, proprio l'utilizzo di queste applicazioni sui dati del Censimento della Popolazione ha comportato un arricchimento dei dizionari soprattutto in termini di modi di esprimersi dei rispondenti che garantirà *performance* dei software di codifica automatica sempre migliori.
Una riflessione analoga può essere, infine, fatta per il sistema informatizzato implementato per la gestione della codifica realizzata all'interno dell'Istat: questa architettura, infatti, pur essendo stata disegnata in funzione delle caratteristiche strutturali dei dati rilevati nel Censimento 2001 è adattabile e quindi riproducibile per altre indagini e di conseguenza può costituire un punto di partenza per la standardizzazione del processo.

Bibliografia

ACTRv3 User Guide (1998), Statistics Canada

Appel M. and Hellerman E. (1983). "Census Bureau experience with Automated Industry and Occupation Coding". In American Statistical Association, *Proceedings of Section on Survey Research Methods*, pages 32-40.

Balestrino R., Reale A. (2000) Il profilo dell'errore nell'acquisizione dei dati di indagine con strumenti di lettura ottica. Società Italiana di Statistica. XL Riunione Scientifica. 247-250. Firenze, 26-28 aprile 2000

Blaise Developer's Guide (1999), Statistics Netherlands

Cochran, W. G., "Sampling Techniques", 3rd edition, Wiley, New York, 1977.

De Angelis R., Macchia S. and Mazza L. (2000), Applicazioni sperimentali della codifica automatica: analisi di qualità e confronto con la codifica manuale, Istat Quaderni di ricerca – Rivista di statistica Ufficiale, 1, 29-54

Hellermann, E., 'Overview of the Hellerman I&O Coding System'. US Bureau of the Census internal paper, Washington, 1982.

Istat, (1991), "Classificazione delle attività economiche", Metodi e norme. Serie C – n.11

Istat "Classificazione delle Professioni, Metodi e norme– nuova serie n.12 edizione 2001

Lyberg L. e Dean P. (1992), Automated Coding of Survey Responses: an international review, Conference of European Statisticians, Work session on Statistical Data Editing, Washington D.C.

Macchia S., D'Orazio M. (2002), A system to monitor the quality of automated coding of textual answers to open questions, Research in Official Statistics (ROS), n. 2 2002, pp. 7-21

Macchia S., Mastroluca S.,Reale A., (2001) 'Planning the quality of the automatic coding process for the next Italian General Population Census', Q2001 International Conference on Quality in Official Statistics, Stockholm, May 14-15, 2001

Macchia S., Mastroluca S. (2004) 'The automatic coding process in the 2001 Italian General Population Census: efficacy and quality', Q2004 European Conference on Quality and Methodology in Official Statistics, Mainz, May 24-26, 2004

Tourigny, J. Y. and Moloney, J., 'The 1991 Canadian Census of Population experience with automated coding', *Paper presented at United Nations Statistical Commission on Statistical Data Editing*, 1995.

Wenzowski M.J. (1988), ACTR – A Generalised Automated Coding System. Survey Methodology, vol. 14: 299-308.