SISTEMA STATISTICO NAZIONALE
ISTITUTO NAZIONALE DI STATISTICA

L'ambiente di codifica automatica dell'Ateco 2007

Istat

2009

Metodi e Norme n. 41

I settori

AMBIENTE E TERRITORIO	*Ambiente, territorio, climatologia*
POPOLAZIONE	*Popolazione, matrimoni, nascite, decessi, flussi migratori*
SANITÀ E PREVIDENZA	*Sanità, cause di morte, assistenza, previdenza sociale*
CULTURA	*Istruzione, cultura, elezioni, musei e istituzioni similari*
FAMIGLIA E SOCIETÀ	*Comportamenti delle famiglie (salute, letture, consumi, etc.)*
PUBBLICA AMMINISTRAZIONE	*Amministrazioni pubbliche, conti delle amministrazioni locali*
GIUSTIZIA E SICUREZZA	*Giustizia civile e penale, criminalità*
CONTI ECONOMICI	*Conti economici nazionali e territoriali*
LAVORO	*Occupati, disoccupati, conflitti di lavoro, retribuzioni*
PREZZI	*Indici dei prezzi alla produzione e al consumo*
AGRICOLTURA E ZOOTECNIA	*Agricoltura, zootecnia, foreste, caccia e pesca*
INDUSTRIA E SERVIZI	*Industria, costruzioni, commercio, turismo, trasporti e comunicazioni, credito*
COMMERCIO ESTERO	*Importazioni ed esportazioni per settore e Paese*

Alla produzione editoriale collocata nei 13 settori si affiancano le pubblicazioni periodiche dell'Istituto: Annuario statistico italiano, Bollettino mensile di statistica e Compendio statistico italiano. Il Rapporto annuale dell'Istat viene inviato a tutti gli abbonati anche ad un solo settore.

SISTEMA STATISTICO NAZIONALE
ISTITUTO NAZIONALE DI STATISTICA

L'ambiente di codifica automatica dell'Ateco 2007

A cura di: Paola Vicari

Coordinamento redazionale: Enzo Venerandi

Per informazioni sul contenuto della pubblicazione
rivolgersi al Cont@ct Centre dell'Istat all'indirizzo:
https://contact.istat.it//

Eventuali rettifiche ai dati pubblicati saranno diffuse
all'indirizzo www.istat.it nella pagina di presentazione del volume

L'ambiente di codifica automatica dell'Ateco 2007

Esperienze effettuate e prospettive

Metodi e norme n. 41

ISBN 978-88-458-1629-1

Istituto nazionale di statistica
Via Cesare Balbo, 16 – Roma

Realizzazione: Istat, Servizio Editoria

Stampato nel mese di novembre 2009
presso il Centro Stampa dell'Istat
Via Tuscolana 1788 – Roma

Indice

Capitolo 7 - Procedura di consultazione su Web

Avvertenze

Nelle tavole statistiche e nei prospetti sono state adoperate le seguenti convenzioni:

Linea (-)	a) quando il fenomeno non esiste; b) quando il fenomeno esiste e viene rilevato, ma i casi non si sono verificati
Due puntini (..)	per i numeri che non raggiungono la metà della cifra dell'ordine minimo considerato
Quattro puntini (….)	quando il fenomeno esiste, ma i dati non si conoscono per qualsiasi ragione

Premessa

La nuova classificazione delle attività economiche Ateco 2007, in vigore dal 1° gennaio 2008, ha reso necessario l'adeguamento di ACTR, il software di codifica automatica in uso all'Istat già dagli anni '90. L'aggiornamento di ACTR con una classificazione molto diversa dalla precedente era già di per sé un lavoro oneroso; a questa attività se ne sono aggiunte altre che sono state gestite da un apposito gruppo di lavoro nel quale hanno collaborato strettamente la Direzione centrale registri statistici, dati amministrativi e statistiche sulla pubblica amministrazione e la Direzione centrale per le tecnologie e il supporto metodologico dell'Istat, ognuna per le specifiche competenze.

Il gruppo di lavoro ha lavorato su diversi fronti. La prima necessità era l'aggiornamento, con la nuova classificazione, della versione batch dell'applicazione che codifica le descrizioni delle attività economiche rilevate nelle indagini statistiche.

Una specifica attività è stata dedicata alla codifica di descrizioni di attività economica completamente diverse da quelle precedenti ovvero raccolte dalle dichiarazioni delle imprese presso le camere di commercio.

Considerata la diffusione e l'utilizzazione da parte di tutte le fonti amministrative della classificazione Ateco 2007, si è deciso di adattare il software di codifica progettato per la versione batch per un utilizzo di consultazione da mettere a disposizione degli utenti del sito Web dell'Istat. Con tale strumento chiunque può trovare il suo codice di attività economica descrivendo la propria attività.

Sono stati messi a punto appositi test per verificare le performance dell'applicazione sia in termini quantitativi che qualitativi per tutte le esigenze citate.

Il lavoro presentato in questo volume è frutto dell'attività del gruppo "per lo sviluppo del software ACTR per la codifica automatica dell'Ateco 2007", costituito in base a quanto deliberato dal Comitato di gestione per l'implementazione dell'Ateco 2007. Del gruppo di lavoro, coordinato da Stefania Macchia e da Angelina Ferrillo, hanno fatto parte Piero Bretti, Genoveffa Buonocore, Vincenzo Gallina, Loredana Mazza, Domenico Perrone, Valentina Talucci, Alberto Valery e Paola Vicari. Hanno collaborato inoltre, per la realizzazione dell'applicazione Web e per il progetto di integrazione ACTR e TaLTaC^2, Daniela Carbone, Cecilia Colasanti, Massimiliano Degortes, Manuela Murgia e Valeria Prigiobbe.[1]

[1] *Il lavoro è frutto dell'attività di ricerca congiunta degli autori. In ogni caso, ai soli fini dell'attribuzione, il capitolo 1 è da attribuirsi a S. Macchia, il capitolo 2 a P. Vicari, i paragrafi 3.1 e 3.4 nonché l'appendice a L. Mazza, i paragrafi 3.2 e 3.3 ed il capitolo 4 ad A. Ferrillo, i paragrafi 5.1, 5.2 e 5.4 a S. Macchia, il paragrafo 5.3 a P. Vicari, il paragrafo 6.1 a V. Talucci, i paragrafi 6.2, 6.3, 6.4 e 6.5 a M. Murgia, il paragrafo 7.1 a P. Vicari, i paragrafi 7.2 e 7.3 a D. Carbone e V. Prigiobbe ed il paragrafo 7.4 ad A. Valery.*

Capitolo 1 - La codifica automatica in Istat: il sistema ACTR

1.1 Il sistema ACTR

Fino a pochi anni fa, l'attività di codifica delle variabili rilevate a testo libero veniva fatta manualmente, pur essendo molto costosa in termini di tempo e non garantendo la standardizzazione del processo. Per questi motivi, nel 1998 si è deciso di automatizzare il processo di codifica ed è stato sperimentato il sistema ACTR v.3 (*Automatic Coding by Text Recognition*), progettato e commercializzato da Statistics Canada. È stato scelto ACTR perché è generalizzato, ossia indipendente dalla lingua e dalla classificazione di riferimento ed è già utilizzato da diversi uffici di statistica (Tourigny et Moloney, 1995).

ACTR si basa sulla metodologia originariamente sviluppata dal Census Bureau (Hellerman, 1982), ma utilizza degli algoritmi messi a punto da Statistics Canada (Wenzowski, 1988).

L'attività di codifica è preceduta da una fase di standardizzazione dei testi, chiamata *parsing*, che fornisce 14 diverse funzioni, quali la rimozione dei caratteri ininfluenti, delle parole inutili, di suffissi/prefissi, l'individuazione di sinonimi eccetera. Il *parsing* ha la finalità di rimuovere tutte le varianti grammaticali e sintattiche, in modo da rendere uguali due descrizioni con lo stesso contenuto semantico originariamente diverse.

Come sulla risposta da codificare, il *parsing* viene effettuato anche sulle descrizioni del dizionario, cosiddetto *reference file*. I testi così trattati vengono confrontati tra di loro e, se si realizza un match perfetto, *direct match*, viene assegnato un unico codice, altrimenti il software utilizza un algoritmo per individuare nel dizionario i testi più simili a quello da codificare, *indirect match*. Operativamente, utilizza un apposito algoritmo per misurare la similarità tra i testi (Macchia et al., 2007); tale misura (S) assume valori compresi nell'intervallo [0,10] i cui estremi corrispondono ad un abbinamento testuale nullo (S=0) o ad un abbinamento perfetto (S=10). Il sottoinsieme dei testi del *reference* 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 (1) 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 (S_1) e successivo (S_2).

I possibili risultati del sistema ACTR si suddividono quindi in:

(1)
$S_1 > Smax$ e $(S_1 - S_2) > \Delta S$ (codice unico) (1a)
$S_1 > Smax$ e $(S_1 - S_2) \leq \Delta S$ (codici multipli) (1b)
$Smin < S_1 \leq Smax$ (codici possibili) (1c)
$S_1 \leq Smin$ (casi falliti) (1d)

Se è soddisfatta la condizione (1a) la voce del dizionario a punteggio massimo (S_1) 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 nei casi (1b) e (1c).

I valori dei parametri soglia sono fissati dall'utente in funzione dei suoi obiettivi di qualità: valori alti elevano la precisione (percentuale di codici unici corretti) dei risultati a scapito del tasso di codifica (percentuale di codici unici assegnati); quindi la scelta dei valori ottimali si gioca sul bilanciamento tra questi due aspetti della qualità dei risultati.

L'impostazione di tali parametri costituisce un elemento della cosiddetta strategia di codifica.

Come già accennato, ACTR è un software generalizzato, indipendente quindi dalla lingua e dalla classificazione, il che significa che l'implementazione dell'ambiente di codifica è a carico dell'utente che deve adattare il sistema alla lingua ed alla classificazione che vuole utilizzare e verificarne le performance.

La costruzione del cosiddetto dizionario per la codifica (*reference file*) è l'impegno più gravoso in quanto la sua dimensione e la sua qualità hanno un impatto diretto sui risultati della codifica. Questa attività si estrinseca nella rielaborazione dei testi del manuale della classificazione ufficiale così da avvicinarli al modo di esprimersi dei rispondenti; tale attività viene effettuata in una serie di passi, il più importante dei quali è l'integrazione di queste descrizioni con quelle derivanti dalle risposte empiriche precodificate raccolte in indagini che rilevano lo stesso fenomeno. Altrettanto importante è l'implementazione nell'ambiente di codifica delle regole classificatorie funzionali all'interpretazione delle descrizioni fornite dai rispondenti in modo da far lavorare il sistema di codifica con una logica simile a quella umana.

1.2 Le applicazioni di codifica automatica sviluppate in Istat

Sono state implementate in Istat diverse applicazioni di codifica afferenti a diverse classificazioni. Le più importanti sono relative alle variabili elencate di seguito e sono state utilizzate in numerose indagini, tra le quali i Censimenti del 2001 (Censimento della Popolazione e Censimento dell'Industria):

- Professione;
- Attività economica;
- Titolo di studio;
- Stato estero/Cittadinanza;
- Comune.

I risultati ottenuti sono stati valutati tramite due indicatori (cfr. cap. 5):

- *Recall rate* (tasso di codifica): percentuale di codici unici assegnati automaticamente;
- *Precision rate* (tasso di accuratezza): percentuale di codici unici corretti assegnati automaticamente.

Coerentemente con questi due indicatori, le performance delle applicazioni citate sono state più che soddisfacenti e coerenti con quelle ottenute da altri Istituti di Statistica (De Angelis et al., 2000).

Per quanto riguarda l'attività economica, i risultati ottenuti sono stati sempre superiori nelle indagini sulle imprese, rispetto a quelle sulle famiglie, come può vedersi nella tavola 1.1. Ciò è imputabile al fatto che, mentre il concetto di settore di attività economica è familiare per i rispondenti alle indagini sulle imprese, non lo è altrettanto per quelli delle indagini sulle famiglie.

Tavola 1.1 - Risultati di codifica delle attività economiche

RILEVAZIONI	Testi da codificare	*Recall rate*	*Precision rate*
Censimento intermedio dell'Industria	1.793	58,8	91,0
Indagine di qualità/Censimento della popolazione 1991	6.288	54,5	85,0
I Indagine pilota sulle forze di lavoro	-	43,5	85,0
I Indagine pilota/Censimento della Popolazione 2001	-	51,2	93,7
II Indagine pilota/Censimento della Popolazione 2001	-	51,9	90,0
Censimento della Popolazione 2001 (Questionario Convivenze)	-	53,6	92,3
Censimento dell'Industria 2001	1.130.693	80,7	-

1.3 Obiettivi del gruppo di lavoro

Con il rilascio della nuova classificazione delle attività economiche Ateco 2007, si è ritenuto necessario adeguare l'ambiente di codifica. Questa attività, come approfonditamente esposto nel cap. 3, non è assolutamente banale, viste le innovazioni profonde della nuova classificazione e le dimensioni dell'ambiente di codifica in termini di numerosità di descrizioni contenute nel *reference file*.

Questo lavoro è stato assegnato ad un gruppo di lavoro appositamente costituito, cui è stato attribuito un mandato più ampio, ossia quello di progettare ed implementare un'applicazione di codifica in grado di rispondere a tre tipologie di esigenze:

1. quella tradizionale, finora già gestita, di codificare descrizioni di attività economiche rilevate nell'ambito di indagini statistiche;
2. quella di codificare descrizioni provenienti da archivi esterni di interesse per l'Istat;
3. quella di fornire agli utenti del sito Web dell'Istat una funzione di consultazione dell'ambiente di codifica finalizzata all'individuazione del proprio codice di attività economica.

In particolare, la necessità di codificare descrizioni provenienti da archivi esterni è soprattutto conseguente all'esigenza di transcodificare l'archivio delle imprese (secondo l'Ateco 2007), utilizzando diverse fonti informative, tra le quali l'archivio fornito dalle Camere di commercio (cfr. cap. 6).

Delle tre esigenze, sopra citate, solo la prima rispecchia pienamente la funzione per cui il sistema di codifica è stato progettato, ossia trattare descrizioni non eccessivamente lunghe e caratterizzate da una struttura rispetto alla quale viene coerentemente costruito il *reference file.* Descrizioni con queste caratteristiche sono quelle rilevate in un'indagine statistica, in quanto nei questionari la formulazione del quesito impatta direttamente sulla struttura della risposta attesa ed inoltre vengono abitualmente fornite istruzioni ben definite su come formulare la risposta. Non altrettanto può dirsi delle altre due esigenze, che sono infatti caratterizzate da descrizioni abbastanza imprevedibili nella loro formulazione, in quanto fornite in modo assolutamente libero, senza alcuno schema di riferimento. Per di più, mentre nella codifica in batch di un archivio si deve tendere a massimizzare il tasso di codifica, nel caso della funzione di consultazione on-line, può essere preferibile fornire all'utente una serie di opzioni tra le quali egli stesso possa scegliere quella più attinente alla propria attività economica.

L'ambiente di codifica, quindi, è stato progettato in modo da tener conto di tutti questi aspetti, per esempio, è stato integrato con altri software per il trattamento dei testi per gestire le descrizioni di archivi esterni (cfr. cap. 6) e sono state impostate soglie, diverse da quelle abitualmente utilizzate nel batch, per la funzione di consultazione Web (cfr. cap. 7).

Sono inoltre stati progettati e realizzati appositi test per verificare le performance dell'applicazione sia in termini quantitativi che qualitativi in funzione di tutte le esigenze citate (cfr. cap. 5).

Capitolo 2 - La nuova classificazione Ateco 2007

La nuova classificazione delle attività economiche, Ateco 2007, è in vigore dal 1 gennaio 2008. L'Ateco 2007, versione nazionale della nuova classificazione delle attività economiche Nace rev. 2,[2] è profondamente diversa dalla precedente Ateco 2002 in quanto la nuova classificazione europea Nace rev. 2 è il risultato di una revisione completa discussa a livello internazionale. La Nace rev. 2, a sua volta, è infatti la versione europea della nuova Isic rev. 4, definita in ambito Onu, e alla quale si allineano tutti i Paesi del mondo. La nuova Nace e la nuova Ateco consentono di ottenere finalmente dati comparabili a livello internazionale.

Già negli anni novanta si era provveduto a rendere comparabili le classificazioni delle attività economiche e dei prodotti; la classificazione di riferimento era la Isic rev. 3 dell'Onu. In questo panorama di convergenza internazionale, la Naics - la classificazione adottata nei paesi del nord America (Canada, Stati Uniti e Messico) - aveva una corrispondenza molto limitata con la classificazione internazionale definita dall'Onu. La nuova Isic rev. 4 costituisce invece il riferimento per tutte le altre classificazioni nazionali esistenti adottate dal 2007 in poi, inclusi i Paesi che precedentemente utilizzavano la Naics o la Anzsic (adottata in Australia e Nuova Zelanda).

L'altro elemento fondamentale da tener presente è che, a parte la lieve revisione operata nel 2002 con la Nace rev. 1.1, la classificazione delle attività economiche era sostanzialmente la stessa dal 1991 e risultava estremamente inadeguata a riflettere l'attuale stato dell'economia e i cambiamenti nel mondo produttivo intervenuti da circa vent'anni a questa parte.

La sinergia di questi due elementi - convergenza internazionale e realtà economica cambiata - hanno fatto sì che la nuova classificazione fosse molto diversa dalla precedente. Ciò rappresenta un onere per l'implementazione nelle nuove statistiche ma un grande vantaggio sia in termini di confrontabilità dei dati a livello internazionale sia in termini di dati statistici più conformi alla realtà economica.

Il processo di revisione si è svolto a livello internazionale - per oltre cinque anni - interpellando, oltre ai singoli Istituti di Statistica, le associazioni industriali e di categoria e la Banca Centrale Europea. Le prime consultazioni hanno riguardato le regole di classificazione che sono rimaste immutate rispetto a quelle adottate nella classificazione precedente; successivamente è stata approvata la struttura e infine le note esplicative.

Anche a livello nazionale ci si è allineati all'esigenza di una classificazione unica; il Comitato[3] che ha definito la nuova Ateco 2007, coordinato dall'Istat, ha provveduto a seguire sia i lavori internazionali sia a definire la classificazione nazionale. Il Comitato è pervenuto a una versione unica della classificazione non solo nella definizione della struttura ma anche nell'interpretazione dei singoli punti. Ad un certo punto dei lavori ci si è resi conto che, per soddisfare le esigenze di tutti e avere allo stesso tempo una classificazione unica, era necessario definire anche una VI cifra. Le seste cifre interessano solo una parte della classificazione infatti, su 918 categorie, solo 150 si dividono in due o più sotto-categorie dando vita a 456 sotto-categorie che, sommate a quelle che restano uguali alla quinta cifra aggiungendo solo uno zero finale, determinano 1.224 seste cifre.

Era evidente che, di fronte a un cambiamento così importante della classificazione, ci si dovesse organizzare sia all'interno dell'Istituto per produrre le nuove statistiche sia per aiutare

[2] La Nace rev. 2 è stata approvata con Regolamento della Commissione n. 1893/2006 pubblicato su Official Journal del 30 dicembre 2006.

[3] Al Comitato hanno partecipato: Camere di commercio e Unioncamere, Agenzia delle Entrate, Inps, Ministeri, Confindustria Industria e Servizi, Banca d'Italia, Associazioni di categoria eccetera.

gli utenti nell'adottare la nuova classificazione. Dal primo gennaio 2008, infatti tutti gli Enti (Agenzia delle Entrate, Inail, Inps e Camere di commercio) hanno dovuto adottare la nuova classificazione. Ciò significa che già da febbraio 2008 (con la prima dichiarazione dell'anno prevista dall'Agenzia delle Entrate) i cittadini si sono trovati a dover scegliere il nuovo codice Ateco a 6 cifre.

Per quanto riguarda l'Istat, che produce dati fino a un dettaglio massimo di 5 cifre, il Regolamento della Commissione prevede l'obbligo di produrre il Registro delle imprese con la nuova classificazione delle attività economiche dal 2008. Le statistiche congiunturali dovranno fornire dati con la nuova classificazione a partire dall'inizio del 2009; mentre le statistiche strutturali dovranno fornire dati definitivi a Eurostat, relativi all'anno 2008, in doppia classificazione (Ateco 2002 e Ateco 2007) nel 2010. La Contabilità nazionale si allineerà per ultima - nel 2011 - alla nuova classificazione Ateco 2007, per il periodo di riferimento 2008-2011.

Sia per dare un ulteriore ausilio agli utenti sia per supportare la transcodifica del Registro delle imprese dall'Ateco 2002 all'Ateco 2007, si è deciso di aggiornare il software ACTR con la nuova classificazione Ateco 2007. Considerata la mole dei cambiamenti, la messa appunto di ACTR per Ateco 2007 non può considerarsi un semplice aggiornamento ma quasi una messa a punto *ex novo*.

ACTR nella sua versione batch è un prodotto consolidato per l'Istat e sperimentato in diverse indagini sulle imprese, soprattutto per i Censimenti. Il progetto di metterne a punto una versione on-line a disposizione degli utenti esterni era stato concepito da tempo. La nuova classificazione delle attività economiche è stata l'occasione giusta per realizzare la versione on-line del prodotto.

2.1 La nuova Ateco 2007: cambiamenti nella struttura e nei contenuti

Sebbene alcuni criteri di costruzione della classificazione, nonché la formulazione delle note esplicative, siano stati revisionati, le caratteristiche generali della classificazione Ateco sono rimaste invariate. Sono stati introdotti nuovi concetti ai livelli più alti della classificazione e sono stati creati nuovi dettagli per riflettere le diverse tipologie di attività produttive e le nuove industrie emergenti. Allo stesso tempo si è cercato di mantenere invariata la struttura della classificazione in tutte le aree che non richiedevano esplicitamente un cambiamento derivante dall'introduzione di nuovi concetti.

Il dettaglio della classificazione è aumentato sostanzialmente (le classi sono aumentate da 514 a 615 e, di conseguenza, le categorie della versione italiana da 883 a 918). Per quanto concerne le attività di produzione di servizi, questo aumento è visibile a tutti i livelli, incluso il più alto; per altre attività, quali ad esempio l'Agricoltura, il maggior dettaglio riguarda principalmente il livello più basso della classificazione.

La tavola seguente presenta i principali cambiamenti strutturali, in termini numerici, tra Ateco 2002 e Ateco 2007.

Tavola 2.1 - Differenze fra Ateco 2002 e Ateco 2007

	Ateco 2002	Ateco 2007	Differenza
Sezioni	17	21	+4
Divisioni	62	88	+26
Gruppi	224	272	+48
Classi	514	615	+101
Categorie	883	918	+ 35
Sotto-categorie	-	1.224	+1.224

Da questa tavola si può intuire che la nuova classificazione è molto più dettagliata; in realtà la questione è più complessa: esistono sezioni di attività completamente nuove che raccolgono "pezzi" di attività già esistenti e descrivono inoltre attività precedentemente non rilevate.

Se si divide la stessa tavola tra una parte relativa al solo settore Manifatturiero e una relativa alle altre attività, si nota che il Manifatturiero si contrae a favore di un ampia crescita delle altre attività concentrate soprattutto nei servizi (nel loro senso più ampio) che, già a livello di due cifre, vengono identificati da 25 divisioni in più (cfr. tavola 2.2).

Tavola 2.2 - Differenze tra Ateco 2002 e Ateco 2007. Settore Manifatturiero separato dagli altri settori

	Ateco 2002	Ateco 2007	Differenza
	Attività manifatturiere		
Sezioni	1	1	-
Divisioni	23	24	+1
Gruppi	103	95	-8
Classi	242	230	-12
Categorie	347	317	-30
Sotto-categorie	-	415	+415
	Altre attività		
Sezioni	16	20	+4
Divisioni	39	64	+25
Gruppi	121	177	+56
Classi	272	385	+113
Categorie	536	601	+65
Sotto-categorie	-	809	+809

Al livello più alto della classificazione, alcune sezioni possono essere facilmente comparate con la versione precedente della classificazione. Tuttavia, l'introduzione di alcuni concetti nuovi a livello di sezione, ad esempio la sezione "Informazione e Comunicazione" o il raggruppamento delle attività legate al Riciclaggio (sezione E), non consente di effettuare facilmente un confronto generale tra l'Ateco 2007 e la versione precedente.

Più in dettaglio: nella nuova sezione J "Servizi di informazione e Comunicazione" confluiscono le attività editoriali precedentemente identificate dal gruppo 22.1 del settore Manifatturiero e ora raggruppate nella divisione 58.

La nuova sezione E "Fornitura di acqua; Reti fognarie; Attività di gestione dei rifiuti e risanamento" presenta un problema analogo. Anche qui un'intera divisione dell'Ateco 2002 - "Recupero e preparazione per il riciclaggio" - appartenente al Manifatturiero ne fuoriesce e, raccogliendo un'altra divisione proveniente dai servizi e dettagliando maggiormente tutte le attività, diventa una nuova sezione che rispecchia delle attività particolarmente significative nell'attuale realtà economica.

Un cambiamento molto rilevante e complesso da gestire è la creazione di una divisione riservata alle attività di "Riparazione, Manutenzione e Installazione" che diventa l'ultima divisione del settore Manifatturiero. Poiché tale classificazione è stata concepita come utilizzabile e utilizzata da tutti i Paesi del mondo e poiché tali attività sono molto diffuse nei paesi in via di sviluppo, si è ritenuto necessario dedicare a questa attività una divisione specifica.

Dal complesso processo di revisione/convergenza iniziato nel 2001 e conclusosi alla fine del 2006, emerge una classificazione più moderna ma profondamente diversa dalla precedente.

La corrispondenza tra due diverse versioni di una classificazione avviene per mezzo di tabelle di corrispondenza tra la classificazione vecchia e quella nuova (Ateco 2002 - Ateco 2007) e viceversa (Ateco 2007 - Ateco 2002).

Per avere un'idea dell'impatto dei cambiamenti sulle statistiche ufficiali a seguito dell'implementazione della Ateco 2007, risulta utile esaminare i diversi tipi di corrispondenza tra Ateco 2002 e Ateco 2007:

- corrispondenza 1 a 1: le classi di Ateco 2002 corrispondono esattamente ad una classe Ateco 2007 e viceversa;
- corrispondenza N a 1: due o più classi di Ateco 2002 corrispondono ad una classe in Ateco 2007;
- corrispondenza 1 a N: una classe di Ateco 2002 è suddivisa in due o più classi di Ateco 2007.

A livello di 4 cifre (le classi) i codici che non si transcodificano automaticamente ma che si dividono in due o più codici nuovi sono circa il 45 per cento (corrispondenza 1 - N). Poiché l'Ateco definisce una quinta cifra nazionale più dettagliata, in alcuni casi nuove quarte cifre europee equivalgono alle quinte cifre della precedente versione italiana (Ateco 2002); ciò ha fatto sì che i codici a cinque cifre del Registro delle imprese, che transcodificavano in due o più codici, fossero pari a circa il 35 per cento.

2.2 I maggiori cambiamenti nella classificazione delle attività economiche

Le sezioni della classificazione Ateco 2002 per Agricoltura e Pesca sono state unite. Tuttavia, il dettaglio di questa nuova sezione A ("Agricoltura, silvicoltura e pesca") è stato sostanzialmente incrementato, in risposta alle continue richieste di maggior dettaglio della classificazione Isic, dovute in larga misura al fatto che l'agricoltura è una componente fondamentale della struttura economica di molti paesi in via di sviluppo.

Sono state create nuove divisioni delle attività manifatturiere per rappresentare industrie nuove o esistenti che hanno aumentato la propria rilevanza economica o sociale, come ad esempio la divisione 21 ("Fabbricazione di prodotti farmaceutici di base e preparati farmaceutici") e la divisione 26 ("Fabbricazione di computer e prodotti di elettronica e ottica"). La ragione della differenziazione di quest'ultima divisione dalla precedente divisione 30 ("Fabbricazione di macchine per ufficio, di elaboratori e sistemi informatici") di Ateco 2002 è nell'avere uno strumento migliore per le statistiche sulle attività ad alto contenuto tecnologico. Altre divisioni nuove, come le divisioni 11 ("Produzione di bevande") e 31 ("Fabbricazione di mobili") sono nate dalla scissione di divisioni preesistenti, innalzando quindi i relativi componenti dal livello di gruppo al livello di divisione.

La maggior parte delle altre divisioni della sezione C della Ateco 2002 ("Attività manifatturiere") è rimasta immutata, eccetto le divisioni 22 ("Editoria, stampa e riproduzione di supporti registrati") e 37 ("Recupero e preparazione per il riciclaggio"), parti sostanziali delle quali sono state destinate ad altre sezioni.

La riparazione e l'installazione di macchine ed apparecchiature, prima classificata all'interno della fabbricazione del tipo di apparecchiatura corrispondente, è stata inserita nella divisione 33 ("Riparazione, manutenzione ed installazione di macchine ed apparecchiature"). Tutte le attività di riparazione specializzata sono adesso classificabili separatamente.

È stata creata una nuova sezione E ("Fornitura di acqua; reti fognarie, attività di trattamento dei rifiuti e risanamento") che racchiude le attività relative alle "misure igienico sanitarie" della divisione 90 della Ateco 2002, della divisione 41 "raccolta, depurazione e distribuzione d'acqua" e le attività di "recupero materiali", che corrisponde sostanzialmente alla divisione 37 di Ateco

2002. Questa sezione raggruppa adesso attività d'interesse per le politiche comuni, ma è stata anche adattata in base all'effettiva organizzazione di queste attività in diversi paesi. Il dettaglio delle attività descritte è stato sostanzialmente incrementato.

La sezione F ("Costruzioni") risulta nella classificazione Ateco 2007 complessivamente molto più dettagliata; è stato introdotto il concetto di "lavori di costruzione specializzati", in sostituzione della struttura della versione precedente, sostanzialmente basata sul processo di costruzione.

La riparazione di beni personali e per la casa è stata eliminata dalla sezione G ("Commercio all'ingrosso e al dettaglio; riparazione di autoveicoli e motocicli") della classificazione Ateco 2002. Tuttavia, è stata mantenuta la classificazione delle attività di "commercio e riparazione di autoveicoli e motocicli" nella divisione 45 di Ateco 2007 (corrispondente alla divisione 50 in Ateco 2002) per motivi di continuità e comparabilità.

Il dettaglio della sezione I ("Servizi di alloggio e di ristorazione") è stato incrementato per riflettere la differente natura e la specializzazione delle attività descritte.

È stata creata una nuova sezione J ("Servizi di informazione e comunicazione"), che riunisce le attività di fabbricazione e distribuzione di prodotti culturali e informativi, la fornitura di mezzi di trasmissione e distribuzione di tali prodotti, nonché di dati o comunicazioni, le attività delle tecnologie di informazione, l'elaborazione elettronica dei dati e altre attività dei servizi d'informazione. Le componenti principali di questa sezione sono le attività di editoria, inclusa l'edizione di software (divisione 58), le attività di produzione cinematografica e registrazioni musicali e sonore (divisione 59), le attività di programmazione e trasmissione radiofonica e televisiva (divisione 60), le telecomunicazioni (divisione 61), le attività delle tecnologie di informazione (divisione 62) e altre attività dei servizi di informazione (divisione 63). Queste attività erano comprese nella classificazione Ateco 2002 nelle sezioni D ("Attività manifatturiere"), I ("Trasporti, magazzinaggio e comunicazioni"), K ("Attività immobiliari, noleggio e servizi alle imprese") e O ("Altri servizi pubblici, sociali e personali"); la nuova struttura ha quindi un forte impatto di comparabilità con la versione Ateco precedente. Tuttavia, questa nuova versione delle attività dei "Servizi di informazione e comunicazione" fornisce un approccio più coerente rispetto alla versione precedente della classificazione Ateco, essendo basata sulla tipologia delle attività esercitate.

La sezione "Attività immobiliari, noleggio e servizi alle imprese" della classificazione Ateco 2002 è stata suddivisa in tre sezioni distinte nella Ateco 2007. Le attività immobiliari sono state incluse in una sezione a parte (sezione L) data la dimensione e l'importanza di tale settore nel Sistema dei Conti Nazionali. Le attività rimanenti sono state raggruppate nella sezione M ("Attività professionali, scientifiche e tecniche"), che racchiude attività che richiedono un elevato livello di formazione e che rendono disponibili agli utenti conoscenze e capacità specialistiche, nonché nella sezione N ("Attività amministrative e di servizi di supporto"), che comprende attività di supporto alle operazioni commerciali in generale, non incentrate sul trasferimento di conoscenze specialistiche. L'informatica e attività connesse (Ateco 2002, divisione 72) non fanno più parte di questa sezione. La riparazione di computer è stata raggruppata con la riparazione di beni per la casa nella sezione S, mentre l'edizione di software e le attività di informatica sono state raggruppate nella nuova sezione J.

L'ambito dell'istruzione (sezione P) è stato modificato al fine di includere esplicitamente la formazione specialistica a livello sportivo, culturale e altro, nonché i relativi servizi specialistici di supporto.

La sezione Q ("Sanità e assistenza sociale") presenta adesso un maggior dettaglio rispetto alla versione precedente della classificazione Ateco, grazie alla creazione di tre divisioni al posto di una. È stato inoltre ristretto l'ambito che, includendo solo le prestazioni sanitarie per esseri umani, fornisce uno strumento migliore per misurare questa componente importante della sfera

economica. Di conseguenza, le attività veterinarie sono state eliminate da questa sezione e inserite, a livello di divisione, nella sezione M ("Attività professionali, scientifiche e tecniche").

Parti sostanziali della sezione O di Ateco 2002 ("Altri servizi pubblici, sociali e personali") sono state spostate nelle sezioni E ("Fornitura di acqua; reti fognarie, attività di gestione dei rifiuti e risanamento") e J ("Servizi di informazione e comunicazione") di Ateco 2007, come già segnalato. Le attività rimanenti sono state raggruppate in due nuove sezioni: "Attività artistiche, di intrattenimento e divertimento" (sezione R) e "Altre attività di servizi" (sezione S). Di conseguenza, attività quali creazioni artistiche, attività di biblioteche e sale da gioco sono state portate a livello di divisione. La riparazione di computer e di beni per uso personale e per la casa è stata inserita in questa nuova sezione S.

Capitolo 3 - Attività realizzate per la revisione dell'ambiente di codifica automatica

3.1 Aggiornamento del dizionario

Come già accennato (cfr. cap. 1) l'implementazione dell'ambiente di codifica è completamente a carico dell'utente che deve adattare il sistema alla lingua ed alla classificazione che si vuole utilizzare, partendo dalla costruzione del *reference file* che rappresenta l'attività più gravosa. Si ricorda, infatti, che la sua dimensione e la sua qualità hanno un impatto diretto sui risultati della codifica.

Per procedere all'adeguamento dell'ambiente di codifica automatica dell'Ateco 2002, in vista della nuova classificazione delle Attività economiche 2007, è stato necessario prendere conoscenza innanzi tutto della base informativa di partenza ovvero del dizionario informatizzato (*reference file*) e dei file di *parsing* utilizzati nel contesto di codifica 2002. L'applicazione utilizzava un dizionario costituito da 30.745 record provenienti:

- dalla rielaborazione dei testi del manuale della classificazione ufficiale relativa all'Ateco 1991 (2002, S. Macchia et al.);
- dalla rielaborazione dei testi del manuale della classificazione ufficiale relativa all'Ateco 2002;
- dalle risposte empiriche precodificate rilevate in indagini che rilevano lo stesso fenomeno (cfr. cap. 1).

Il dizionario era così strutturato:

- colonna 4-10 codice numerico classificazione Ateco 2002
- colonna 15-18 flag identificativo
- colonna 20-220 descrizione attività economica.

Si precisa che il flag identificativo è stato pensato da un precedente Gruppo di lavoro che aveva predisposto la base informativa Ateco 1991 al fine dell'adozione della codifica automatica nei Censimenti del 2000. Tale flag aveva fini esclusivamente documentativi, per facilitare cioè il lavoro di eventuali successivi aggiornamenti del dizionario. Il flag indica, infatti, la fonte dei testi precodifcati inseriti. Nella tavola che segue sono indicate le dimensioni dettagliate del *reference.*

Tavola 3.1 - Dimensioni *reference* Ateco 2002

	Numero record
Testi della classificazione ufficiale	9.460
Empiriche totali	21.285
di cui empiriche (e2) (a)	3.760
Dizionario totale Ateco 2002	30.745

(a) Individuano le empiriche che hanno cambiato codice nella fase di aggiornamento per l'Ateco 2002

Come detto sopra, prima di procedere all'approntamento del nuovo contesto di codifica relativo all'Ateco 2007, si è ritenuto necessario prendere atto delle variazioni intercorse dalla chiusura dei lavori del precedente Gruppo di lavoro (rilascio applicazione Ateco 1991) finalizzate alla messa a punto dell'applicazione Ateco 2002. Per partire, quindi, da una base

informativa corretta sia dal punto di vista della coerenza della codifica che delle trasformazioni effettuate nei file di *parsing* si è reso necessario effettuare un'analisi accurata:

- sulle empiriche (e2) (3.760 record) transcodificate manualmente in base alla tavola di raccordo Ateco 1991 - Ateco 2002 (codici 1 - N);
- sui relativi file di *parsing* utilizzati nel contesto di codifica Ateco 2002 (cfr. paragrafo 3.2).

A seguito di queste attività, l'applicazione ha subito una serie di aggiornamenti che hanno portato, tra l'altro, al rilascio di un *reference* file costituito da 30.740 descrizioni.

Messa a punto la base informativa di partenza relativa all'Ateco 2002, si è proceduto alla costruzione dell'ambiente di codifica Ateco 2007 sulla base delle tabelle di transcodifica.

Vista la complessità della classificazione e la delicatezza dell'operazione di transcodifica, il lavoro è stato approntato per step successivi ovvero:

- transcodifica codici Ateco 2002 in codice Ateco 2007 in base alle tabelle di raccordo;
- analisi da parte degli esperti della classificazione dei casi dubbi o incoerenti di transcodifica;
- controllo dei sinonimi inseriti nei file di *parsing* per verificare la coerenza della transcodifica da Ateco 2002 a Ateco 2007;
- effettuazione di test su campioni ricavati da un file precodificato delle Camere di commercio, opportunamente estratti in riferimento alle divisioni già transcodificate;
- analisi di qualità sui file degli output prodotti dai test.

Il primo, il secondo e terzo step sono stati realizzati per macro settori, ovvero man mano che la classificazione Ateco 2007 con le relative tabelle di transcodifica erano disponibili. Queste attività sono state estremamente onerose in virtù dell'elevato numero di descrizioni empiriche da riesaminare che, come è noto, hanno un impatto molto forte sulle percentuali di successo dell'applicazione di codifica.

Gli ultimi due step, infine, costituiscono una fase essenziale per verificare il tasso di codifica e quello di accuratezza (*precision rate*), ma anche per procedere ad arricchire ed addestrare ulteriormente l'ambiente di codifica. Le descrizione non codificate, infatti, qualora avessero avuto un contenuto informativo sufficiente per l'attribuzione di un codice, sono state inserite nel dizionario informatizzato. Di contro si è provveduto ad apportare le correzioni necessarie all'ambiente di codifica emerse a seguito dell'analisi dei testi codificati in modo non corretto.

Queste attività sono state realizzate ricorsivamente finché non sono stati raggiunti tassi di codifica e di accuratezza sufficientemente elevati, in linea con le applicazioni utilizzate in precedenti indagini, (cfr. cap. 1).

3.2 Struttura del dizionario transcodificato

Il dizionario, messo a punto a seguito delle attività descritte nel precedente capitolo, ha la seguente struttura:

- colonna 1-7 codice numerico Ateco 2007
- colonna 13-19 codice numerico Ateco 2002
- colonna 27-30 flag identificativo
- colonna 32-232 descrizione attività economica

I codici numerici Ateco 2007 sono associati alle voci della Classificazione ufficiale delle Attività Economiche Ateco 2007 derivata dalla Nace rev. 2.

Il flag identificativo individua la provenienza della descrizione associata al codice Ateco transcodificato in base all'Ateco 2007.

La struttura è rimasta invariata rispetto al dizionario costruito per l'Ateco 1991 e l'Ateco 2002; si sono lasciati i titoli, le note esplicative e le descrizioni empiriche dell'Ateco 1991 che hanno mantenuto il flag identificativo rispettivamente (a1), (b1) e (ei).

Con il precedente aggiornamento dell'ambiente di codifica all'Ateco 2002 i titoli, le note esplicative e le empiriche nuove o modificate rispetto all'Ateco 1991 sono caratterizzati rispettivamente dal flag identificativo (a2), (b2), (e2) e (c2), quest'ultimo individua le empiriche provenienti dalla Cpa 2004[4] (Ferrillo, Ottobre 2004).

Con l'aggiornamento dell'ambiente di codifica in base all'Ateco 2007, i flag identificativi delle descrizioni di attività economiche del dizionario informatizzato, diventano:

Tavola 3.2 - Flag identificativi delle descrizioni

Sigla 1991	Sigla 2002	Sigla 2007
(a1)	(a2)	(a7)
(b1)	(b2)	(b7)
(ei)	(e2)	(e7)
	(c2)	(c2)
(e0)	(e0)	(e0)
(ep)	(ep)	(ep)
(pi)	(pi)	(pi)
(na)	(na)	(na)
(nc)	(n2)	(n7)

(a*) (b*) (e*) sono state lasciate invariate.

I flag identificativi elencati si riferiscono:
(e0): ai codici numerici e alla descrizione dell'attività dei titoli dell'Ateco 2007 a livello di categoria;
(ep): alle empiriche che provengono dalla fase di validazione del Censimento della Popolazione (CP);
(pi): alle descrizioni provenienti dal confronto tra la classificazione Ateco e Prodcom;
(na): alle empiriche che provengono dalla Nace rev. 1.1 emendata e armonizzata con le altre nomenclature;
(nc): alle empiriche con descrizioni troppo generiche che non consentono cioè l'attribuzione di un codice corretto, pertanto risultano non codificabili;
(n2): alle empiriche, provenienti dall'aggiornamento Ateco 2002, con descrizioni troppo generiche che non consentono cioè l'attribuzione di un codice corretto, pertanto risultano non codificabili;
(a*) (b*) (e*):

- in parte dalla duplicazione delle descrizioni della divisione 14 (Confezione di articoli di abbigliamento; confezione di articoli in pelle e pelliccia) con la sostituzione della parola produzione al posto di confezione;
- in parte dalla duplicazione della categoria 45.20.4 (Riparazione e sostituzione di pneumatici per autoveicoli) con la sostituzione della parola "pneumatici" con "gomme per auto", e 45.3 (commercio di parti ed accessori di autoveicoli) con la sostituzione della parola "gomme" con "pneumatici";
- alcune (b*) provengono anche dalla duplicazione di alcune voci della divisione 01 (Coltivazioni agricole e produzione di prodotti animali, caccia e servizi connessi) con la sostituzione della parola "produzione" al posto di "coltivazione" prevista dalla classificazione;
- alcune (e*) provengono anche dalla duplicazione di alcune voci della categoria 56.30.0 (bar e caffè) con la sostituzione della parola caffè al posto di bar.

[4] Regolamento (Cee) n. 3696/93 del Consiglio, del 29 ottobre 1993, relativo alla classificazione statistica dei prodotti associata alle attività nella Comunità economica europea, GU n. L 342 del 31 dicembre 1993. L'associazione con le attività economiche, sulla base di tale nomenclatura, dà alla Cpa una struttura che a tutti i livelli è simmetrica a quella della Nace rev. 1.1.

3.3 Dimensione e caratteristiche del dizionario informatizzato e dei file di *parsing*

Per fornire elementi di valutazione circa la complessità dell'applicazione di codifica Ateco 2007 e del suo aggiornamento rispetto alle applicazioni delle precedenti versioni della classificazione, se ne descrivono in questo paragrafo le caratteristiche sia in termini di dimensioni del dizionario elaborabile che dei file di *parsing.*

Per facilità di comprensione si sintetizzano nella seguente tavola le funzioni espletate da ciascun file di *parsing,* riportando qualche esempio attinente alle attività economiche (per un maggiore dettaglio, si rimanda all'Appendice).

Tavola 3.3 - File di *parsing*

Fase 1 pre-trattamento	
Questa fase prevede una serie di funzioni tra le quali, quella gestita dal file WCHR.	
WCHR	Serve per definire tutti i caratteri validi (lettere dell'alfabeto, eventualmente numeri, altri caratteri) e per uniformare la grafia (dal minuscolo al maiuscolo) FILE INDISPENSABILE.
Fase 2 trattamento delle stringhe – *Phrase processing*	
DCLS	cancella tutto ciò che è compreso tra incisi esempio [ ] tutto ciò che è contenuto tra parentesi viene eliminato.
DSTR	cancella le stringhe ritenute inutili (esempio NELL').
RSTR	sostituzione di stringhe (esempio c/contoterzi = CONTOTERZO).
Una volta effettuate le funzioni corrispondenti ai tre file di *parsing* descritti, ACTR suddivide la stringa in parole.	
Fase 3 trattamento delle parole – *Word processing*	
RWRD	consente la gestione dei sinonimi, degli errori di ortografia e delle parole ininfluenti (esempio MINUTO = DETTAGLIO), sostituzione con *blank* di parole ininfluenti, articoli eccetera; Col 1 PAROLA DA SOSTITUIRE, Col 31 PRIMA PAROLA SOSTITUITA, Col 61 SECONDA PAROLA SOSTITUITA esempio VITICOLTORE= COLTIVAZIONE VITE.
DWRD	consente la gestione di sinonimi a livello di coppie di parole (esempio ABBIGLIAMENTO FIRMATO = ABBIGLIAMENTO; ABBIGLIAMENTO MATRIMONIO = ABBIGLIAMENTO ADULTO); Col 1 PRIMA PAROLA DA TRASFORMARE, Col 31 SECONDA PAROLA DA TRASFORMARE, Col 61 PRIMA PAROLA SOSTITUITA, Col 91 SECONDA PAROLA SOSTITUITA (esempio Col 1 PROD, Col 31 VENDITA, =Col 61 PRODUZIONE, Col 91 COMMERCIO).
HWRD	gestione parole separate dal trattino (esempio BABY-SITTER = BABYSITTER) Col 1 PAROLA CON TRATTINO DA SOSTITUIRE, Col 61 PRIMA PAROLA SOSTITUITA.
IWRD	Definisce *"caratteri non ammessi"*, in modo da sottrarre dalle successive elaborazioni tutte le parole che li contengono (esempio se si dichiarano qui i NUMERI non ammessi, ROMA2 DIVENTA BLANK.
EXCP	gestione parole che non devono subire trasformazioni (esempio BORSA = BORSA;) Col 1 BORSA, Col 31 BORSA.
PRFX	gestione dei prefissi da eliminare Col 1 (eliminazione solo se la parola troncata rimane di 4 lettere)
SUFX	gestione dei suffissi da eliminare (esempio GELATERIA = GELATER, ORALE = ORAL) eliminazione solo se la parola troncata rimane di 4 lettere.
MCHR	definizione di caratteri doppi o tripli da rimuovere (esempio ABBIGLIAMENTO = ABIGLIAMENTO)
SORT	Inserendo questa funzione, le parole rimanenti a seguito dei processi già descritti vengono ordinate, secondo un ordine alfabetico ascendente.

La dimensione del dizionario informatizzato per la codifica automatica dell'Ateco è cresciuta negli anni, anche grazie alla maggiore complessità delle successive versioni della classificazione. Per l'Ateco 1991 il dizionario complessivo ammontava a 29.931 testi, per l'Ateco 2002 a 30.744 testi e per l'Ateco 2007 a 33.768 testi.

Le tabelle seguenti mostrano la tipologia dettagliata del dizionario riferito all'Ateco 2007 e la struttura dei file di *parsing* nelle tre distinte versioni.

Tavola 3.4 - Descrizioni per tipologia del dizionario aggiornato all'Ateco 2007

SIGLA	N. record	Percentuale
(a*)	27	0,1
(a1)	1.222	3,6
(a2)	938	2,8
(a7)	726	2,1
(b*)	138	0,4
(b1)	4.581	13,6
(b2)	1.206	3,6
(b7)	1.901	5,6
(c2)	13	..
(e*)	253	0,7
(e0)	918	2,7
(e2)	3.577	10,6
(e7)	1.688	5,0
(ei)	15.134	44,8
(ep)	403	1,2
(n2)	11	..
(na)	10	..
(nc)	54	0,2
(pi)	968	2,9
Totale	**33.768**	**100,0**

Tavola 3.5 - Dimensione file di *parsing* utilizzati per la codifica automatica di Ateco 1991, 2002 e 2007

FILE PARSING	Ateco 1991 Numero record	Ateco 2002 Numero record	Ateco 2007 Numero record
WCHR	93	96	96
DCLS	2	2	2
DSTR	8	8	28
RSTR	1.067	1.070	1.098
RWRD	10.049	10.519	11.481
DWRD	2.752	2.893	3.107
EXCP	98	148	1.001
SUFX	17	17	17

3.4 Verifica della coerenza del *parsing*

Il sistema ACTR, al momento del caricamento del database (cfr. cap.1), effettua solo un controllo finalizzato a scartare i record "non validi", ossia quelli che, a seguito del *parsing*:

- hanno entrambi la stessa descrizione;
- hanno lo stesso codice ed eventualmente lo stesso filtro.[5]

Non viene effettuato cioè alcun controllo di coerenza sui file di *parsing;* tale attività resta pertanto completamente a carico dell'utente. Poiché si è ritenuto che la gestione di questi controlli sia piuttosto pesante, considerando sia la complessità della classificazione che le dimensioni dei file dei sinonimi, si è provveduto, durante i lavori del già citato precedente Gruppo di lavoro, a sviluppare un software che evidenzi eventuali incoerenze, lasciandone la risoluzione al gestore dell'applicazione. Prima di procedere all'aggiornamento del dizionario per l'Ateco 2007, si è ritenuto opportuno, pertanto, sottomettere i file di *parsing* predisposti per il contesto di codifica relativo all'Ateco 2002 ai controlli di coerenza implementati nella citata procedura.

I file sottomessi ai controlli sono:

- RSTR *Replacement String* (*Phrase processing* del *parsing);*
- RWRD *Replacement Word (Word processing* del *parsing);*
- DWRD *Double Word (Word processing* del *parsing).*

I controlli sono effettuati internamente a ciascun file (un file su se stesso) e tra un file e l'altro, mantenendo come ipotesi di partenza uno *strategy file* **(**cfr. allegato 1) che esegua i processi di *parsing* secondo l'ordine RSTR → RWRD → DWRD. Il software, sviluppato in Visual Basic, è dotato di un'apposita interfaccia che consente l'elaborazione di tre fasi di controlli.

La **prima fase** effettua una serie di controlli singolarmente su ciascun file. I passaggi realizzati sono i seguenti:

1. ordinamento del file, secondo i seguenti criteri:
   - RSTR → ordinamento alfabetico sulla seconda colonna
   - RWRD → ordinamento alfabetico sulla seconda colonna, quindi sulla terza
   - DWRD → ordinamento alfabetico sulla terza colonna, quindi sulla quarta;
2. eliminazione delle schede duplicate;
3. eliminazione delle schede nelle quali i termini delle colonne di destra sono uguali a quelli delle colonne di sinistra (trasformazione A → A).

La **seconda fase** è finalizzata a mettere in luce diversi tipi di incoerenze all'interno di ciascuno dei file citati ed a memorizzarle su appositi file Excel. In particolare si procede:

1. all'individuazione delle trasformazioni incoerenti, quali ad esempio:
   - trasformazione di A → B
   - trasformazione di A → C
2. all'individuazione delle trasformazioni ricorsive, quali ad esempio:
   - trasformazione di A → B
   - trasformazione di B → C

oppure

   -trasformazione di A → Z B
   -trasformazione di B → C D

[5] Tra le funzionalità più avanzate del sistema ACTR c'è l'uso dei campi filtro. Il campo filtro non è obbligatorio, ma, se utilizzato, consente all'utente di definire sottoinsiemi di dizionario in cui restringere la ricerca da parte del sistema ai fini dell'assegnazione del codice. Per esempio, rifacendosi al caso dell'attività economica, il dizionario potrebbe essere suddiviso in tante categorie, cui corrisponderebbero altrettanti valori del filtro, corrispondenti ai rami principali della classificazione (agricoltura, estrazione, industria, commercio eccetera).

Con la **terza fase**, infine, vengono realizzati i controlli tra il file delle RWRD e quello delle DWRD. In particolare si procede:

1. all'individuazione di coincidenze tra la prima colonna del file RWRD e le prime due colonne del file DWRD, ad esempio:
   - trasformazione in RWRD di A → Z
   - quindi trasformazione in DWRD di A B → C D
   - oppure trasformazione in DWRD di B A → C D
2. all'individuazione di coincidenze tra la prima colonna del file RWRD e le seconde due colonne del file DWRD, quali ad esempio:
   - trasformazione in RWRD di A → B
   - quindi trasformazione in DWRD di C D → A B
   - oppure trasformazione in DWRD di C D → B A

Si fa presente, comunque, che quest'ultimo tipo di coincidenze non costituisce necessariamente un errore, quindi non è indispensabile un intervento correttivo, ma può essere sufficiente verificare che non si tratti di una svista, ma di una trasformazione voluta.

Ci si è avvalsi della procedura sopra descritta per effettuare l'analisi di coerenza dei file di *parsing* utilizzati nel contesto di codifica preparato per l'Ateco 2002. I file di partenza nel contesto di codifica predisposto per l'Ateco 2002 e sottomessi ai controlli di coerenza erano così formati:

- File RSTR 1.070 record
- File RWRD 10.519 record
- File DWRD 2.893 record.

Dal passaggio del programma di controllo sui file sopra elencati procedendo per le fasi descritte nella premessa, si sono ottenuti i seguenti risultati:

1. **Prima Fase** si è semplicemente effettuato l'ordinamento dei file per colonne, secondo i criteri sopra illustrati, e si è proceduto all'eliminazione delle schede doppie;
2. **Seconda Fase** sono state prodotte le seguenti schede:
   - RSTR nessuna scheda da controllare
   - RWRD 53 schede da controllare
   - DWRD 62 schede da controllare;
3. **Terza Fase** sono state prodotte le seguenti schede:
   - RWRDiDWRDi 34 schede da controllare
   - RWRDiDWRDo 15 schede da controllare.

Per meglio capire il tipo di incoerenze che sono state sanate all'interno dei file di *parsing* si riportano alcuni esempi.

Esempio 1- Incoerenze all'interno del file delle RWRD (Fase II)

INPUT	OUTPUT
ACQUISTO	COMMERCIO
AVQUSTO	ACQUISTO
INPUT	OUTPUT
CALZETTERIA	CALZE
CALZETTER	CALZETTERIA
CALZETTERIE	CALZETTERIA
CALZETT	CALZETTERIA
CALZET	CALZETTERIA

Esempio 2 - Incoerenze tra i file delle RWRD e quello delle DWRD (III Fase)

TIPO	INPUT	OUTPUT/INPUT	OUTPUT	OUTPUT
R	ABBIBLIAMENTO	ABBIGLIAMENTO		
D	ABBIBLIAMENTO	BAMBINO	ABBIGLIAMENTO	BAMBINO

Per eliminare le incompatibilità emerse, è stato necessario effettuare alcune modifiche all'interno dei file stessi. Ciò ha comportato:

1. l'eliminazione di alcune schede;
2. la sostituzione di sinonimi e l'inserimento di nuovi;
3. l'inserimento di nuove schede.

Dopo questi tipi di interventi si è ritenuto opportuno effettuare un ulteriore passaggio di controllo sui file di *parsing* modificati ottenendo i seguenti risultati:

1. **Prima Fase** si è semplicemente effettuato l'ordinamento dei file per colonne;
2. **Seconda Fase** sono state prodotte le seguenti schede:
   - RSTR nessuna scheda da controllare
   - RWRD 11 schede da controllare
   - DWRD nessuna scheda da controllare;
3. **Terza Fase** sono state prodotte le seguenti schede:
   - RWRDiDWRDi nessuna scheda da controllare
   - RWRDiDWRDo 15 schede da controllare.

Si fa presente che la presenza in taluni file di ulteriori schede da controllare non costituisce necessariamente un errore. Non è stato pertanto indispensabile un intervento correttivo, in quanto è stato verificato che non si trattava di una svista, bensì di una trasformazione voluta per consentire una migliore lettura del testo di output come mostra l'esempio 3.

Esempio 3 - Incoerenze tra i file delle RWRD e quello delle DWRD (III Fase)

TIPO	INPUT	OUTPUT/INPUT	OUTPUT	OUTPUT
RWRD	APPARTAMENTI	EDIFICIO		
DWRD	LOC	EDIFICIO	LOCAZIONE	APPARTAMENTI

La stessa procedura è stata di seguito utilizzata per l'analisi di coerenza dei file di *parsing* per il contesto di codifica Ateco 2007.

Capitolo 4 - I criteri classificatori adottati

Nel corso della revisione del *parsing* e del dizionario per la classificazione delle attività economiche, per far funzionare il software di codifica automatica ACTR nel contesto italiano e per la classificazione specifica, sono state discusse e decise alcune regole di classificazione da applicare alle descrizioni delle attività economiche provenienti da utenti di vario tipo: aziende rispondenti a censimenti ed indagini, descrizioni dell'attività economica raccolte dalle Camere di commercio eccetera. Per la costruzione del *parsing* si sono usati dei criteri generali e dei criteri particolari peculiari per certe situazioni.

I criteri generali sono stati risolti nell'ambito dei file di *parsing,* quelli particolari con l'inserimento nel *reference* di apposite empiriche. Nel seguito si riportano quelle in uso dopo la revisione dell'applicazione per la nuova Ateco 2007.

4.1 Criteri generali

Nei file di *parsing*, in particolare in quelli relativi alla definizione dei sinonimi, è stato possibile inserire diverse dizioni di uso comune equiparandole alle corrispondenti presenti nella "Classificazione delle attività economiche", ad esempio:

- **Vendita = Commercio**
- **Fabbricazione = Produzione**

Si è cercato poi di prevedere tutte le possibili abbreviazioni univoche come per esempio:

- **Lavoraz. = Lavorazione**
- **Abbigl. = Abbigliamento**
- **Bicicl. = Bicicletta**

Prevalenze in caso di doppie attività economiche

La casistica delle doppie attività esercitate dai rispondenti alle indagini è emersa analizzando le risposte ai Censimenti e alle indagini; queste per essere codificate richiedono la definizione di appositi criteri guida ispirati, nella gran parte dei casi, a far prevalere quella che fornisce maggior reddito.

La complessità della problematica ha tuttavia fatto sì che anche i cosiddetti "criteri generali" presentino delle eccezioni che sono state, di volta in volta, risolte con l'inserimento di apposite empiriche.

Doppie attività: "criteri generali"

Si esaminano nel seguito le coppie di attività nell'ambito delle quali è stata stabilita la prevalenza (>)

- **Coltivazione** rispetto a **raccolta**
  Coltivazione > raccolta.
- **Coltivazione** rispetto a **commercio**
  Coltivazione > commercio.
- **Produzione/lavorazione** rispetto a **commercio**
  Produzione > commercio.
  Fabbricazione > commercio.
  Trasformazione > commercio.
  Lavorazione > commercio.
  Macellazione > commercio.

Per trasformazione > commercio vale la seguente eccezione:

l'attività della *stagionatura* fa parte della fase di trasformazione di un certo prodotto. Tuttavia, quando viene considerata insieme all'attività del *Commercio,* predomina quest'ultimo, nonostante la prevalenza contraria espressa sopra. Negli esempi che seguono viene mostrato in quali casi l'attività prevalente risulta essere il commercio all'ingrosso:

46.33.1 Stagionatura e commercializzazione latticini.
46.32.2 Commercio ingrosso salumi, stagionatura prosciutti.

- **Produzione** rispetto a **noleggio**
  Produzione > noleggio.
- **Produzione** rispetto a **riparazione**
  Produzione > riparazione.
- **Produzione** rispetto a **manutenzione**
  Produzione > manutenzione.
- **Produzione** rispetto a **installazione**
  Produzione > installazione.
- **Fabbricazione /produzione/costruzione** rispetto a **progettazione**
  Fabbricazione > progettazione.
  Costruzione > progettazione.
- **Produzione** rispetto a **consulenza**
  Produzione > consulenza.
- **Commercio** rispetto a **riparazione**
  Commercio > riparazione.

Valgono le seguenti eccezioni, per le quali l'attività prevalente è la riparazione:

33.19.0 Riparazione vendita pallets.
45.20.4 Commercio riparazione sostituzione di pneumatici.

- **Commercio** rispetto a **manutenzione**
  Commercio > manutenzione.
- **Commercio** rispetto a **installazione**
  Commercio > installazione.
- **Ingrosso** rispetto a **riparazione/manutenzione/installazione**
  Ingrosso > riparazione/manutenzione/installazione.
- **Dettaglio** rispetto a **riparazione/manutenzione/installazione**
  Dettaglio > riparazione/manutenzione/installazione.

4.2 Criteri adottati nelle singole divisioni

Divisione 01 (Coltivazioni agricole e produzione di prodotti animali, caccia e servizi connessi)

Avendo constatato che i rispondenti spesso usano come sinonimo di *coltivazione* la parola *produzione*, sono state duplicate le descrizioni che contengono la parola *coltivazione* con *produzione* esempio:

Coltivazione di frumento duro.

Produzione di frumento duro.

Coltivazione e raccolta di prodotti agricoli va classificata nella categoria dove si coltiva il prodotto:

Raccolta di prodotti agricoli conto terzi va classificata nella categoria 01.61.0.

Nell'ambito della divisione si sono utilizzati i seguenti criteri particolari:

- Le prime lavorazioni dei prodotti agricoli (quelle necessarie per preparare i prodotti per i mercati primari) vanno classificate nella divisione 01;
- La lavorazione e trasformazione dei prodotti agricoli è invece intesa come trasformazione e pertanto va classificata nella divisione 10 (Industrie alimentari).

Si sono stabiliti poi i seguenti codici non completi (a due e tre cifre):

01 (e2) Agricoltore
01 (e2) Agricoltore imprenditore
01 (a2) Agricoltura
01 (a2) Agricoltura e relativi servizi
01 (a2) Agricoltura, caccia e relativi servizi
01 (e2) Attività agricola coltivazioni miste
01 (e2) Attività agricola
01 (e2) Azienda agricola di prodotti agricoli
01 (a2) Coltivazione agricola
01 (e2) Coltivazione campi
01 (e7) Coltivazione del fondo
01 (e7) Coltivazione e conduzione di terreno
01 (e2) Coltivazione tabacco - vigneto - pomodori
01 (e2) Conduzione aziende agricole
01 (e7) Conduzione poderi agricoli
01 (e2) Contadino
01 (e7) Gestione dell'azienda agricola
01 (e7) Industria agricola
01 (a2) Orto-colture specializzate vivaistiche e sementiere in piena aria
01 (b2) Produzione di semi per fiori, frutta e ortaggi in piena aria
01 (e2) Produzione ortofrutticola
01 (e7) Coltivazione ortaggi e olivicoltura
01 (e7) Coltivazioni floricole piante ornamentali
01.1 (e2) Azienda agricola coltivazione di ortofloricole in piena aria
01.1 (e2) Azienda agricola coltivazione di ortofloricole in serra
01.1 (e2) Azienda agricola coltivazione di ortovivaistica in piena aria
01.1 (e2) Azienda agricola di ortofloricole in piena aria
01.1 (e2) Azienda agricola di ortofloricole in serra
01.1 (e2) Azienda agricola di ortovivaistica in piena aria
01.1 (b1) Coltivazione di cereali (compreso il riso)
01.1 (a2) Coltivazione di ortaggi, specialità orticole, fiori e prodotti di vivai
01.1 (a2) Coltivazione di specialità orticole e fiori in piena aria
01.1 (a2) Coltivazione di specialità orticole e fiori in serra
01.1 (a2) Coltivazione di specialità orticole e prodotti di vivaio in piena aria
01.1 (a2) Coltivazione mista di ortaggi e fiori in piena aria
01.1 (a2) Coltivazione mista di ortaggi e fiori in serra
01.1 (a2) Coltivazione mista di ortaggi e prodotti di vivaio in piena aria
01.1 (a2) Coltivazione mista di ortaggi, specialità orticole, fiori e prodotti di vivaio in piena aria
01.1 (a2) Coltivazione mista di ortaggi, specialità orticole, fiori e prodotti di vivaio in serra
01.1 (a2) Coltivazioni agricole; orticoltura, floricoltura
01.1 (e2) Coltivazioni ortofrutticole
01.1 (e2) Produzione di ortovivaistica in piena aria
01.1 (e2) Sericoltura

01.2 (ep) Azienda agricola coltivazione di olivo e frutta
01.2 (ep) Azienda agricola coltivazione di vite e frutta
01.2 (ep) Azienda agricola coltivazione di vite e olivo
01.2 (ep) Azienda agricola di olivo e frutta
01.2 (ep) Azienda agricola di vite e frutta
01.2 (ep) Azienda agricola di vite e olivo
01.2 (a2) Coltivazione di frutta, frutta a guscio, prodotti destinati alla preparazione di bevande e spezie
01.2 (a1) Coltivazione mista vitivinicola, olivicola e frutticola
01.2 (b1) Colture frutticole diverse, coltivazione di prodotti destinati alla preparazione di bevande e spezie
01.2 (b1) Colture miste olivicole e frutticole
01.2 (b1) Colture miste vitivinicole e frutticole
01.2 (b1) Colture miste vitivinicole e olivicole
01.2 (ei) Olivicoltura e viticoltura
01.2 (ep) Produzione di olivo e frutta
01.2 (ep) Produzione di vite e frutta
01.2 (ep) Produzione di vite e olivo
01.2 (ep) Azienda agricola frutticola
01.2 (b1) Coltivazione di frutta
01.2 (a1) Coltivazioni frutticole diverse
01.2 (e7) Coltivazione di alberi da frutta
01.2 (ep) Frutteto
01.4 (ei) Allevamento bovini equini
01.4 (a1) Allevamento di bovini
01.4 (a2) Allevamento di ovini, caprini, equini
01.4 (ep) Azienda agricola allevamento bovini
01.4 (ep) Azienda agricola produzione di latte e di carne bovina
01.4 (e2) Zootecnia

La divisione 01 contiene 848 empiriche.

Divisione 06 (Estrazione di petrolio grezzo e di gas naturale)

Si sono stabiliti i seguenti codici non completi (a due cifre):

06 (e2) Industria mineraria idrocarburi

La divisione 06 contiene 31 empiriche.

Divisione 10 (Industrie alimentari)

Regola generale:

Produzione/lavorazione > commercio, noleggio, riparazione, manutenzione, installazione, progettazione, consulenza.

Per le attività che riguardano la produzione di più prodotti si è stabilito negli specifici casi che:

Produzione pane e pasticceria > produzione pane

Si sono stabiliti i seguenti codici non completi (a due, tre e quattro cifre):

10	(e2)	Produzione alimenti
10.61	(ei)	Produzione semole e sfarinati
10.61	(a2)	Lavorazione delle granaglie e cereali
10.61	(e2)	Produzione di farina
10.61	(e2)	Produzione di semola
10.61	(e2)	Molitura cereali e vendita
10.61	(e2)	Molitura e commercio ingrosso cereali
10.61	(e2)	Commercio e macinazione cereali
10.61	(e2)	Industria molitoria
10.61	(e2)	Industria molitoria c/t trasformazione cereali
10.61	(e2)	Industria molitoria cereali
10.61	(e2)	Lavorazione di cereali
10.61	(e2)	Lavorazione e commercio materie prime cereali
10.61	(e2)	Molitura cereali e commercio minuto ingrosso prodotti agricoltura
10.61	(e2)	Molitura dei cereali lavorazioni industriali e commercio altri prodotti agricoli
10.61	(e2)	Molitura di cereali e lavorazioni industriali
10.61	(e2)	Molitura di cereali per farina
10.61	(e2)	Molitura di cereali per semole
10.61	(e2)	Molitura per semolini
10.61	(e2)	Produzione di semole e semolini di cereali
10.61	(e2)	Stoccaggio e trasformazione di cereali
10.7	(e2)	Azienda dolciaria
10.7	(a2)	Fabbricazione di altri prodotti amidacei alimentari
10.7	(e2)	Industria dolciaria
10.7	(e2)	Preparazioni dolciarie
10.8	(b2)	Produzione di zuppe, minestre e brodi, cibi precotti
10.8	(b2)	Produzione minestre e brodi, cibi precotti
10.8	(e2)	Produzione pasta fresca piatti pronti sughi tutto surgelato

La divisione 10 contiene 1.534 empiriche.

Divisione 14 (Confezione di articoli di abbigliamento; confezione di articoli in pelle e pelliccia)

Sono state duplicate le descrizioni che contenevano la parola *confezione*, sostituendola con *produzione*. Ad esempio nella categoria:

18.14.0 Confezione di biancheria personale.

Si è aggiunta anche la voce:

18.14.0 Produzione di biancheria personale.

La confezione di maglieria e abbigliamento si classifica nella categoria 14.13.1, facendo prevalere l'abbigliamento.

La divisione contiene 883 empiriche.

Divisione 16 (Industria del legno e dei prodotti in legno e sughero (esclusi i mobili); fabbricazione di articoli in paglia e materiali da intreccio)

I falegnami in genere fabbricano mobili o infissi; poiché molti si identificano con la descrizione generica di *falegname* si è stabilito che:

- se fabbricano infissi e mobili vengono classificati con la prevalenza degli infissi;
- se indicano *falegname* o *falegnameria* generica vengono classificati nella categoria 16.23.1 come gli esempi che seguono:

Falegnameria fabbricazione di mobili e infissi *16.23.1*
Laboratorio di falegnameria *16.23.1*
Falegname *16.23.1*

Mentre:

Falegnameria mobili su misura *31.09.1*

La divisione contiene 469 empiriche.

Divisione 17 (Fabbricazione di carta e di prodotti di carta)

Cartotecnica e litografia si classifica nella categoria 17.23.0 (prevale la produzione di prodotti cartotecnici).

La divisione contiene 288 empiriche.

Divisione 18 (Stampa e riproduzione di supporti registrati)

Cartotecnica e legatoria si classifica nella categoria 18.14.0 (prevale l'attività di legatoria).

Cartotecnica editoriale si classifica nella categoria 18.14.0 (prevale l'attività di legatoria).

Si sono stabiliti i seguenti codici non completi (a tre cifre):

18.1	(a1)	Altri servizi connessi alla stampa
18.1	(na)	Attività ausiliarie connesse alla stampa
18.1	(a2)	Lavorazioni ausiliarie connesse alla stampa
18.1	(e7)	Stampa quotidiani e periodici

La divisione contiene 404 empiriche.

Divisione 22 (Fabbricazione di articoli in gomma e materie plastiche)

Si sono stabiliti i seguenti codici non completi (a tre cifre):

22.2	(e2)	Assemblaggi materie plastiche
22.2	(e2)	Lavorazione di materie plastiche
22.2	(e2)	Trasformazione materie plastiche

La divisione contiene 436 empiriche.

Divisione 23 (Fabbricazione di altri prodotti della lavorazione di minerali non metalliferi)

Produzione di conglomerati cementizi si classifica nella categoria 23.63.0.

Produzione di conglomerati bituminosi si classifica nella categoria 23.99.0.

Produzione di emulsioni di catrame per conglomerati bituminosi si classifica nella categoria 19.20.4.

Si sono stabiliti i seguenti codici non completi (a tre cifre):

23.1	(a2)	Fabbricazione di prodotti in vetro
23.1	(a2)	Fabbricazione di vetro
23.1	(e2)	Impresa artigiana lavorazione del vetro
23.1	(e7)	Lavorazione e commercializzazione prodotti in vetro

La divisione contiene 779 empiriche.

Divisione 25 (Fabbricazione di prodotti in metallo, esclusi macchinari e attrezzature)

Produzione di targhe ed insegne stradali in metallo si classifica nella categoria 25.99.9, mentre le stesse, se sono luminose, vanno nella categoria 27.90.0.

Fabbricazione di griglie e reti di fili di ferro o di acciaio saldati si classifica nella categoria 25.93.1.

La divisione contiene 1360 empiriche.

Divisione 26 (Fabbricazione di computer e prodotti di elettronica e ottica; apparecchi elettromedicali, apparecchi di misurazione e di orologi)

Si sono stabiliti i seguenti codici non completi (a due e quattro cifre):

26	(a2)	Fabbricazione di apparecchi medicali, di apparecchi di precisione, di strumenti ottici e di orologi
26	(a2)	Fabbricazione di apparecchi radiotelevisivi
26	(a2)	Fabbricazione di apparecchi radiotelevisivi e di apparecchiature per le comunicazioni
26	(a2)	Fabbricazione di apparecchiature per le comunicazioni
26	(ei)	Fabbricazione di impianti elettronici radiotelevisivi
26	(a2)	Fabbricazione di strumenti ottici e di orologi
26.51	(ei)	Costruzione apparecchiature di controllo e regolazione
26.51	(ei)	Costruzione apparecchiature di misura controllo regolazione
26.51	(a1)	Costruzione di apparecchi di misura comprese parti staccate e accessori
26.51	(b1)	Costruzione di apparecchi di misura elettrici e elettronici (comprese parti staccate o accessori)
26.51	(a2)	Fabbricazione apparecchi misura elettrici ed elettronici; strumenti di precisione
26.51	(b1)	Fabbricazione di accessori di apparecchi di misura elettrici
26.51	(b1)	Fabbricazione di accessori di apparecchi di misura elettronici
26.51	(b1)	Fabbricazione di altre macchine di misurazione
26.51	(b1)	Fabbricazione di altri apparecchi collaudo
26.51	(b1)	Fabbricazione di altri apparecchi misurazione
26.51	(b1)	Fabbricazione di altri strumenti di collaudo
26.51	(b1)	Fabbricazione di altri strumenti di misurazione
26.51	(b1)	Fabbricazione di analizzatori di spettro
26.51	(b1)	Fabbricazione di apparecchi di misura elettrici e elettronici

26.51	(a2)	Fabbricazione di apparecchi di misura elettrici ed elettronici comprese parti
26.51	(b1)	Fabbricazione di apparecchiature per controllare grandezze elettriche
26.51	(b1)	Fabbricazione di apparecchiature per controllare grandezze non elettriche
26.51	(b1)	Fabbricazione di apparecchiature per misurare grandezze elettriche
26.51	(b1)	Fabbricazione di apparecchiature per misurare grandezze non elettriche
26.51	(b1)	Fabbricazione di apparecchiature per regolare motori di veicoli
26.51	(b1)	Fabbricazione di parti staccate di apparecchi di misura elettrici
26.51	(b1)	Fabbricazione di parti staccate di apparecchi di misura elettronici
26.51	(b1)	Fabbricazione di parti staccate e accessori di apparecchi di misura, controllo e regolazione
26.51	(a2)	Fabbricazione di strumenti e apparecchi di misurazione (escluse le apparecchiature di controllo dei processi industriali)
26.51	(a2)	Fabbricazione di strumenti e apparecchi di misurazione, controllo, prova, navigazione e simili, escluse le apparecchiature di controllo dei processi industriali
26.51	(a2)	Fabbricazione di strumenti e apparecchi di prova (escluse le apparecchiature di controllo dei processi industriali)
26.51	(b1)	Fabbricazione di termometri
26.51	(ei)	Fabbricazione sistemi di controllo e misura elettronici
26.51	(ei)	Produzione e commercializzazione apparecchiature e strumentazione elettronica misura
26.51	(ei)	Produzione vendita strumenti misura
26.51	(ei)	Strumentazione elettronica di misura di parametri fisici e chimici
26.70	(a2)	Fabbricazione di strumenti ottici e di attrezzature fotografiche

La divisione contiene 799 empiriche.

Divisione 28 (Fabbricazione di macchinari ed apparecchiature n.c.a.)

La rigenerazione/riciclaggio dei toner, cartucce, nastri stampa si classifica nella categoria 28.23.0.

Si sono stabiliti i seguenti codici non completi (a due cifre):

28	(e2)	Azienda metalmeccanica
28	(e2)	Costruzione di macchinari
28	(e2)	Fabbricazione macchine manutenzione
28	(e2)	Industria metalmeccanica

La divisione contiene 1.469 empiriche.

Divisione 30 (Fabbricazione di altri mezzi di trasporto)

Quando si parla di barche senza nessuna specifica si sottintende che siano da diporto.

Si sono stabiliti i seguenti codici non completi (a due, tre e quattro cifre):

30	(a2)	Fabbricazione di altri mezzi di trasporto
30.11	(a2)	Costruzioni navali e riparazioni di navi
30.9	(a2)	Fabbricazione di altri mezzi di trasporto n.c.a.
30.92	(a2)	Fabbricazione di biciclette e accessori

La divisione contiene 337 empiriche.

Divisione 31 (Fabbricazione di mobili)

Fabbricazione di mobili per l'arredamento si sottintende per la casa, pertanto si classifica nella categoria 31.09.1.

Per ciò che riguarda invece la doppia attività di:

Ebanista restauratore si classifica nella categoria 95.24.0, stabilendo la prevalenza dell'attività di restauro.

Si sono stabiliti i seguenti codici non completi (a tre e quattro cifre):

31.0	(a2)	Fabbricazione di mobili
31.0	(pi)	Produzione di mobili di plastica
31.01	(e2)	Arredamenti su misura gelaterie bar
31.01	(a2)	Fabbricazione di mobili per uffici e negozi
31.01	(e2)	Fabbricazione di mobili per ufficio
31.01	(e2)	Montaggio arredamenti per ufficio
31.01	(e2)	Produzione mobili e sedie ufficio
31.09	(a2)	Fabbricazione di altri mobili

La divisione contiene 524 empiriche.

Divisione 32 (Altre industrie manifatturiere)

Laboratorio di odontotecnico si classifica nella categoria 32.50.2.

Laboratorio oftalmico si classifica nella categoria 32.50.4.

Si sono stabiliti i seguenti codici non completi (a due, tre e quattro cifre):

32	(a2)	Altre industrie manifatturiere
32.12	(a2)	Fabbricazione di oggetti di gioielleria e articoli annessi n.c.a.
32.9	(a2)	Altre industrie manifatturiere n.c.a.
32.99	(ei)	Lavorazione penne a sfera oggetti pubblicitari e articoli regalo

La divisione contiene 729 empiriche.

Divisione 33 (Riparazione, manutenzione ed installazione di macchine ed apparecchiature)

Regole generali:

Produzione > riparazione/manutenzione/installazione.
Commercio > riparazione/manutenzione/installazione.
Dettaglio > riparazione/manutenzione/installazione.
Ingrosso > riparazione/manutenzione/installazione.

Si sono stabiliti i seguenti codici non completi (a quattro cifre):

33.12	(e2)	Manutenzione macchinari industriali
33.12	(e7)	Manutenzione macchinari, impianti ed attrezzature industriali

La divisione contiene 691 empiriche.

Divisione 35 (Fornitura di energia elettrica, gas, vapore e aria condizionata)

Ente nazionale energia elettrica si classifica nella categoria 35.11.0.

Azienda gas municipale si classifica nella categoria 35.22.0.

Gestione centrali di teleriscaldamento si classifica nella categoria 35.30.0.

Si sono stabiliti i seguenti codici non completi (a due e tre cifre):

35	(e2)	Energia elettrica acqua
35	(a2)	Produzione di energia elettrica, di gas, di vapore e acqua calda

35.1 (a2) Distribuzione e commercio di energia elettrica
35.2 (a2) Distribuzione e commercio di combustibili gassosi mediante condotte

La divisione contiene 126 empiriche.

Divisione 36 (Raccolta, trattamento e fornitura di acqua)

Gestione acquedotti si classifica nella categoria 36.00.0.
Gestione acque irrigue si classifica nella categoria 36.00.0.
La divisione contiene 63 empiriche.

Divisione 38 (Attività di raccolta, trattamento e smaltimento dei rifiuti; recupero dei materiali)

Autodemolizioni e commercio ricambi usati si classifica nella categoria 46.77.1.
Commercio ricambi usati per autovetture si classifica nella categoria 46.77.1.
Cantieri di demolizioni di navi si classifica nella categoria 38.31.2.
Cernita stracci si classifica nella categoria 38.32.3.
Riciclaggio rifiuti si classifica nella categoria 38.32.3.
Si sono stabiliti i seguenti codici non completi (a due, tre e quattro cifre):

38 (a1) Raccolta e smaltimento dei rifiuti solidi
38 (ei) Raccolta e smaltimento dei rifiuti solidi e liquidi
38 (ei) Raccolta e smaltimento dei rifiuti solidi e liquidi industriali
38 (ei) Raccolta e smaltimento rifiuti
38 (ei) Raccolta e smaltimento rifiuti ospedali
38 (ei) Raccolta e trasporto rifiuti speciali
38 (ei) Raccolta e trattamento rifiuti
38 (ei) Raccolta smaltimento dei rifiuti solidi più autotrasporto
38 (ei) Raccolta smaltimento rifiuti solidi e speciali
38 (ei) Raccolta smaltimento rifiuti speciali
38 (ei) Raccolta trasporto commercio rifiuti
38 (ei) Raccolta, trasporto e smaltimento rifiuti
38 (a2) Recupero e preparazione per il riciclaggio
38.1 (b1) Raccolta di immondizie e rifiuti di origine industriale
38.1 (ei) Raccolta rifiuti
38.1 (ei) Raccolta rifiuti solidi
38.1 (ei) Raccolta rifiuti solidi c/t
38.1 (ei) Raccolta trasporto rifiuti
38.1 (ei) Trasferimento rifiuti
38.1 (b1) Trasporto dei rifiuti
38.1 (ei) Trasporto rifiuti c/t
38.2 (b1) Conferimento a discariche dei rifiuti
38.2 (b1) Immersione o interramento dei rifiuti
38.2 (ei) Impianto industriale smaltimento rifiuti
38.2 (ei) Impianto stoccaggio rifiuti
38.2 (ei) Impianto trattamento e smaltimento rifiuti solidi
38.2 (ei) Incenerimento e smaltimento rifiuti
38.2 (ei) Incenerimento rifiuti

38.2	(ei)	Livellamento copertura discarica
38.2	(ei)	Messa in discarica dei rifiuti
38.2	(ei)	Servizi smaltimento rifiuti
38.2	(ei)	Servizi smaltimento rifiuti pulizia
38.2	(ei)	Smaltimento di rifiuti
38.2	(b1)	Smaltimento mediante incenerimento o altri procedimenti
38.2	(ei)	Smaltimento rifiuti c/t
38.2	(e2)	Smaltimento rifiuti solidi urbani e pericolosi
38.2	(ei)	Sotterramento rifiuti
38.2	(ei)	Stoccaggio provvisorio rifiuti
38.2	(ei)	Stoccaggio rifiuti
38.2	(ei)	Stoccaggio rifiuti assimilabili
38.2	(ei)	Trasporto e smaltimento rifiuti
38.2	(ei)	Trasporto e stoccaggio rifiuti industriali
38.2	(ei)	Trasporto stoccaggio rifiuti
38.2	(ei)	Trasporto stoccaggio rifiuti pulizie varie
38.2	(ei)	Trattamento rifiuti
38.2	(ei)	Trattamento rifiuti industriali
38.2	(ei)	Trattamento rifiuti solidi e fanghi
38.3	(e2)	Recupero imballi
38.3	(ei)	Recupero rifiuti di bordo
38.3	(ei)	Smaltimento riciclaggio rifiuti
38.32	(b2)	Lavorazione di cascami non metallici per, trasformarli in nuove materie prime
38.32	(b2)	Lavorazione di oggetti non metallici, non usati per, trasformarli in nuove materie prime.
38.32	(b2)	Lavorazione di oggetti non metallici, usati o meno, per, trasformarli in nuove materie prime.
38.32	(b2)	Lavorazione di oggetti non metallici, usati per trasformarli in nuove materie prime
38.32	(b2)	Lavorazione di rottami non metallici per, trasformarli in nuove materie prime
38.32	(b2)	Preparazione per il riciclaggio di cascami non metallici
38.32	(b2)	Preparazione per il riciclaggio di rottami non metallici
38.32	(a2)	Recupero e preparazione per il riciclaggio di cascami e rottami non metallici
38.32	(b2)	Recupero per il riciclaggio di cascami non metallici
38.32	(b2)	Recupero per il riciclaggio di rottami non metallici

La divisione contiene 308 empiriche.

Divisione 39 (Attività di risanamento e altri servizi di gestione dei rifiuti)

Bonifiche ambientali si classifica nella categoria 39.00.0.
Bonifica residuati bellici si classifica nella categoria 39.00.0.
Bonifica di terreni inquinati si classifica nella categoria 39.00.0.
La divisione contiene 25 empiriche.

Divisione 41 (Costruzione di edifici)

Costruzioni edili senza nessuna specifica si classifica nella categoria 41.20.0.
Cooperative edilizie e costruzione di alloggi per soci si classifica nella categoria 41.10.0.
Demolizione e ricostruzione di edifici si classifica nella categoria 41.20.0.
La divisione contiene 269 empiriche.

Divisione 42 (Ingegneria civile)

Si sono stabiliti i seguenti codici non completi (a due e tre cifre):

42	(b2)	Costruzione di opere di ingegneria civile
42	(e2)	Costruzioni di ingegneria civile
42	(a2)	Genio civile
42	(a2)	Lavori di ingegneria civile
42	(e2)	Lavori generali di costruzione di ingegneria civile
42.2	(b2)	Costruzione di condotte e linee di comunicazione elettriche urbane
42.2	(e2)	Costruzione e manutenzione elettrodotti e gasdotti
42.2	(e2)	Riparazione di condotte e linee di comunicazione elettriche urbane

La divisione contiene 179 empiriche.

Divisione 43 (Lavori di costruzione specializzati)

Lavori di isolamento edile ed impermeabilizzazione si classifica nella categoria 43.29.0.
Carpentiere edile si classifica nella categoria 43.99.0.

Dall'analisi delle descrizioni provenienti delle Camere di commercio si è riscontrato che molte imprese sono state registrate con la descrizione di attività economica "*attività lavori edili*" (524 casi) ed "*attività edilizia*" (361 casi). Per queste due descrizioni è stato condotto uno studio per verificare quale codice Ateco è stato attribuito con gli Studi di Settore, alle imprese presenti nell'archivio delle imprese attive Asia, e poiché nell'80,3 per cento e nel 61,8 per cento è stato attribuito il codice 43.39.0, si è deciso di inserire:

a) attività lavori edili con codice categoria 43.39.0;

b) attività edilizia con codice n.c. essendo questa descrizione molto simile ad *edilizia* (descrizione molto generica) già presente nel dizionario con codice n.c.

Si sono stabiliti i seguenti codici non completi (a tre cifre):

43.1	(a1)	Demolizione di edifici e sistemazione del terreno
43.1	(ei)	Scavi sbancamenti demolizioni
43.1	(ei)	Scavo demolizione movimento terra
43.2	(a2)	Installazione dei servizi in un fabbricato

La divisione contiene 698 empiriche.

Divisione 45 (Commercio all'ingrosso e al dettaglio e riparazione di autoveicoli e motocicli)

Autofficina prevale rispetto al soccorso stradale e si classifica nella categoria 45.20.1.
Autofficina (autoriparazioni) e vendita ricambi si classifica nella categoria 45.20.1.
Autofficina e autorimessa si classifica nella categoria 45.20.1.
Autocarrozzeria e soccorso stradale si classifica nella categoria 45.20.2.
Concessionaria auto ricambi officina si classifica nella categoria 45.11.0.

Riparazione delle auto e riparazione mezzi agricoli si classifica nella categoria 45.20.1.

Riparazione e commercio ingrosso pneumatici si classifica nella categoria 45.31.0.

La vendita di motocicli e di biciclette si classifica nella categoria 45.40.1.

La vendita dei motocicli e di accessori si classifica nella categoria 45.40.1.

Per la *riparazione di pneumatici* quando non è specificato per auto e/o per moto si sottintende per auto.

La divisione contiene 526 empiriche.

Divisione 46 (Commercio all'ingrosso, escluso quello di autoveicoli e di motocicli)

Regola generale:

Commercio all'ingrosso è prevalente sul *commercio al dettaglio e su manutenzione / riparazione / installazione.*

Commercializzazione è intesa come commercio all'ingrosso.

Intermediari di commercio con deposito è inteso come intermediario.

Attenzione:

Non esiste il commercio all'ingrosso effettuato con distributori automatici.

Non esiste il commercio all'ingrosso di libri usati.

Commercio all'ingrosso senza specifica di prodotto si classifica nella divisione 46.

Intermediari di commercio, agente di commercio, rappresentante di commercio, procacciatore di commercio senza specifica di prodotto si classificano nel gruppo 46.1.

Agenzia di commercio senza specifica del prodotto commercializzato si classifica nel gruppo 46.1.

Deposito con rappresentanze farmaceutiche, deposito rappresentanza medicinali si classificano nella categoria 46.18.3.

La gastronomia all'ingrosso si classifica nella categoria 46.38.3.

Commercio ingrosso lattiero caseari salumi si classifica nella categoria 46.33.1.

Commercio ingrosso arredamenti (senza specifica "per ufficio e/o per la casa") si classifica nella categoria 46.47.1.

Il consorzio agrario, per convenzione consolidata e riscontrata, si classifica nella categoria 46.75.0 data la prevalenza nella loro attività di vendita all'ingrosso di fertilizzanti, diserbanti, fitofarmaci.

L'ingrosso di carte elettroniche, schede telefoniche anche prepagate si classificano nella categoria 46.49.9.

Commercio ingrosso e ricarica estintori si classifica nella categoria 46.69.9.

Per taluni prodotti non è stato necessario specificare il tipo di commercio, ingrosso o dettaglio, perché difficilmente commerciabili al dettaglio; essi sono:

Nella categoria 46.21.1: *commercio settore risicolo.*

Nella categoria 46.21.2: *alimenti per animali da allevamento, mangimi, mangimi cereali, paglia fieno, prodotti per l'agricoltura sementi, prodotti zootecnici.*

Nella categoria 46.23.0: *bovini suini vivi, bestiame, suini vivi.*

Nella categoria 46.24.1: *Import export pelli finite e pellame, pelli grezze.*

Nella categoria 46.31.2: *etichettamento e vendita derivati pomodoro.*

Nella categoria 46.32.2: *prosciutti.*

Nella categoria 46.33.1: *derivati del latte burro, stagionatura confezionamento formaggio grana e prodotti alimentari, distribuzione uova e ovoprodotti, stagionatura e commercio gorgonzola.*

Nella categoria 46.33.2: *olio, olio conferito dai soci, olio vegetale.*

Nella categoria 46.34.1: *commercio e imbottigliamento vini, commercio vini non prodotti, esportazione vini.*

Nella categoria 46.34.2: *acqua oligominerale, acque gasate, concessionario di bevande, distribuzione bevande.*

Nella categoria 46.35.0: *magazzino vendita generi monopolio di stato.*

Nella categoria 46.38.2: *importazione di olio di fegato di merluzzo.*

Nella categoria 46.38.9: *sale (per uso strade, alimentazione animali eccetera), materie prime alimentari per bar pasticceria, distribuzione di pasta fresca.*

Nella categoria 46.39.2: *distribuzione prodotti alimentari.*

Nella categoria 46.41.1: *tessuto grezzo, tessuti stock, pellame vegetale.*

Nella categoria 46.42.1: *esportazione articoli abbigliamento.*

Nella categoria 46.42.2: *compravendita pellicce.*

Nella categoria 46.42.4: *import export calzature.*

Nella categoria 46.44.2: *forniture alberghiere di porcellane e vetrerie.*

Nella categoria 46.46.1: *distribuzione rappresentante prodotti farmaceutici, gestione distribuzione intermedia farmaci, importazione prodotti medicinali sieri vaccini emoderivati.*

Nella categoria 46.46.3: *gas tecnici puri ossigeno e medicinali, lastre radiografiche, leghe dentarie e materiale odontoiatrico, prodotti ospedalieri, distribuzione dispositivi apparecchi medicali, fornitura strumenti medicali.*

Nella categoria 46.47.1: *importazione e esposizione mobili.*

Nella categoria 46.48.0: *forniture orologeria, pietre per oreficeria.*

Nella categoria 46.49.1: *carta.*

Nella categoria 46.49.5: *articoli per equitazione.*

Nella categoria 46.51.0: *hardware, sistemi telematici, vendita e assistenza registratori di cassa.*

Nella categoria 46.52.0: *apparati elettronici, componenti elettronici, strumentazioni e componenti elettrici e elettronici in genere, apparati telecomunicazioni, prodotti per telecomunicazioni.*

Nella categoria 46.62.0: *macchine lavorazione legno, macchine utensili, macchine fresatrici, macchine lavorazione legno e alluminio.*

Nella categoria 46.63.0: *gru pedane caricatrici, macchine edili.*

Nella categoria 46.64.0: *macchine e accessori per calzaturifici, macchine per conceria.*

Nella categoria 46.65.0: *arredi per pubblici esercizi, arredamenti negozi.*

Nella categoria 46.66.0: *scaffalature terminali sistemi presenze.*

Nella categoria 46.69.1: *materiale ferroviario, materiale rotabile ferroviario, ricambi rotabili ferroviari.*

Nella categoria 46.69.2: *automatismi elettrici, isolanti termoelettrici, macchine attrezzature elettroniche e ottiche, sistemi fibre ottiche, lettori banda magnetica, quadri elettrici, semiconduttori e microsistemi, elettroforniture.*

Nella categoria 46.69.9: *apparecchiature pneumatici, attrezzature accessori per industrie materie plastiche, attrezzature industriali, attrezzature per garage, bilance affettatrici, carrelli elevatori manutenzione attrezzature per magazzino e interni, cuscinetti a sfere e affini, cuscinetti articoli tecnici, impianti attrezzature macchinari, commercio installazione macchine per la gastronomia, macchinari per l'enologia, macchine industriali, macchine per il trattamento della carta, macchine per industria grafica, macchine per marmi, macchine pulizia, ricambi industriali, commercio riparazione bilance affettatrici attrezzature per la trasformazione alimentare, utensili per fornaci, veicoli carrelli elevatori, vendita assistenza sistemi per l'automazione industriale, attrezzature panifici panetterie, bilance misuratori fiscali, attrezzature per oleifici, vendita assistenza attrezzature professionali per la ristorazione e il*

commercio, vendita assistenza carrelli elevatori, impianti refrigeranti, strumentazione per rilevamento inquinamento atmosferico, strumenti di pesatura, materiale antincendio antinfortunistica, macchine e apparecchi di sollevamento e movimentazione.

Nella categoria 46.71.0: *gas kerosene.*

Nella categoria 46.72.1: *acciai, materiali ferrosi, metalli ferrosi, metalli ferrosi semilavorati, prodotti siderurgici.*

Nella categoria 46.72.2: *barre bronzo, metalli non ferrosi, metalli preziosi e semilavorati, semilavorati e rottami di alluminio.*

Nella categoria 46.73.1: *legname, legnami esteri, legnami travi ferro, prefabbricati in legno, trucioli legno.*

Nella categoria 46.73.2: *blocchi marmo bianco venato, blocchi marmo Carrara, calce idrata, canale tubi rame, laterizi, marmi e pietre, marmi graniti pietre, marmi graniti pietre lapidei onici travertini, pietrame, sabbia ghiaia inerti, conglomerati bituminosi, inerti conglomerati calcestruzzo, accessori e materie prime per la produzione di isolante, grassello di calce.*

Nella categoria 46.74.2: *impianti riscaldamento e condizionamento, fornitura per impianti termici.*

Nella categoria 46.75.0: *concimi, fertilizzanti fitofarmaci, materie prime per uso farmaceutico e cosmetico, prodotti chimici industria, prodotti per l'agricoltura fertilizzanti, riattivazione carboni attivi.*

Nella categoria 46.76.1: *materie prime tessili.*

Nella categoria 46.76.2: *prodotti chimici petrolchimici e materie plastiche, granuli termoplastici.*

Nella categoria 46.76.3: *imballaggi o parte di essi, imballi in legno e non.*

Nella categoria 46.77.1: *rottami ferrosi e non ferrosi, recupero ferrosi, rottami ferrosi e non, rottami metallici, rottami parti meccaniche, rottami autotrasporti, demolitore auto usate, rottami auto.*

Nella categoria 46.77.2: *scarti lavorazione pelle, cernita cartaccia e commercio materiali di recupero, biomasse, recupero carta da macero imballaggi in cartone metallo legno e plastica, materiali vari recupero non metallici, recupero smaltimento rifiuti speciali ferro carta legno, stracci, rottami non ferrosi.*

Nella categoria 46.90.0: *Cash and carry.*

Si sono stabiliti i seguenti codici non completi (a due, tre e quattro cifre):

46	(e2)	Commercio all'ingrosso
46	(b1)	Commercio all'ingrosso di articoli per fotografia, cinematografia, ottica e di strumenti scientifici
46	(b2)	Commercio all'ingrosso di cd e dvd vergini o registrati
46	(a2)	Commercio all'ingrosso di supporti, vergini o registrati, audio, video, informatici (dischi, nastri e altri supporti)
46	(a2)	Commercio all'ingrosso e intermediari del commercio (esclusi autoveicoli e motocicli)
46	(e2)	Commercio all'ingrosso importazione e commercio
46	(e2)	Importazione e commercio
46.1	(e2)	Agente di commercio
46.1	(b2)	Attività di commercializzazione dei commissionari e delle cooperative
46.1	(e7)	Agente di affari in mediazione
46.1	(e7)	Intermediario di commercio specializzato
46.1	(e2)	Agente e rappresentante di commercio

46.1 (ei) Agente commercio senza deposito
46.1 (ei) Agente con deposito
46.1 (ei) Agente e rappresentante
46.1 (ei) Agenti di commercio con deposito
46.1 (ei) Agenzia commercio
46.1 (ei) Intermediazione commercio ingrosso
46.1 (ei) Procacciatore d'affari
46.1 (ei) Servizi intermediazione commerciale
46.1 (ei) Subagente professionista commercio
46.18 (a2) Intermediari del commercio specializzato di prodotti particolari n.c.a.
46.18 (e7) Intermediario del commercio prodotti particolari
46.2 (a2) Commercio all'ingrosso di materie prime agricole e di animali vivi
46.2 (e2) Commercio ingrosso di materie utili all'agricoltura
46.2 (e2) Commercio ingrosso di prodotti utili all'agricoltura
46.3 (e2) Commercio all'ingrosso di salumeria e formaggi
46.31 (e2) Commercio ingrosso funghi freschi e secchi
46.32 (a2) Commercio all'ingrosso di carni e di prodotti di salumeria
46.33 (a2) Commercio all'ingrosso di prodotti lattiero-caseari, uova, oli e grassi commestibili
46.34 (a2) Commercio all'ingrosso di bevande alcoliche e altre bevande
46.38 (a2) Commercio all'ingrosso di altri prodotti alimentari, inclusi pesci, crostacei e molluschi
46.38 (e2) Commercio ingrosso prodotti della pesca
46.4 (a2) Commercio all'ingrosso di altri beni di consumo finale
46.41 (a2) Commercio all'ingrosso di prodotti tessili
46.41 (e2) Commercio prodotti tessili
46.41 (e2) Import export prodotti tessili
46.42 (a2) Commercio all'ingrosso di abbigliamento e calzature
46.43 (a2) Commercio all'ingrosso di elettrodomestici, di apparecchi radiotelevisivi e telefonici e altra elettronica di consumo
46.44 (a2) Commercio all'ingrosso di articoli di porcellana e di vetro, di carte da parati e prodotti per la pulizia
46.46 (a2) Commercio all'ingrosso di prodotti farmaceutici (compresi strumenti e apparecchi sanitari)
46.46 (ei) Commercio ingrosso medicinali articoli medico chirurgici
46.47 (a2) Commercio all'ingrosso di altri prodotti per uso domestico
46.6 (a2) Commercio all'ingrosso di macchinari e attrezzature
46.7 (a2) Commercio all'ingrosso di prodotti intermedi non agricoli, di rottami e cascami
46.7 (b1) Commercio all'ingrosso specializzato non classificato in una delle categorie precedenti
46.7 (e7) Commercio all'ingrosso di altri prodotti
46.72 (a2) Commercio all'ingrosso di metalli
46.72 (a2) Commercio all'ingrosso di metalli e di minerali metalliferi
46.72 (a2) Commercio di metalli all'ingrosso
46.73 (a2) Commercio all'ingrosso di legname e di materiali da costruzione, vetro piano, vernici e colori
46.77 (a2) Commercio all'ingrosso di rottami e cascami
46.77 (a2) Commercio di rottami e cascami all'ingrosso

La divisione contiene 2.853 empiriche.

Divisione 47 (Commercio al dettaglio, escluso quello di autoveicoli e di motocicli)

Regola generale:

Commercio al dettaglio è prevalente sulla *manutenzione, riparazione e installazione.*

Il negozio, il negoziante, il dettagliante e la rivendita sono intesi come *commercio al dettaglio.*

Commercio dettaglio fisso è inteso come *commercio al dettaglio.*

Commercio al pubblico, commercio al minuto, punto vendita si trasformano in *commercio al dettaglio.*

Commercio dettaglio senza ulteriore specifica si classifica nella divisione 47.

Produzione di pane e dettaglio alimentari si classifica nella categoria 10.71.1, è prevalente la produzione di pane.

Dettaglio di pane e alimentari si classifica nella categoria 47.11.4, è prevalente la vendita degli alimentari.

Dettaglio solo pane si classifica nella categoria 47.24.1.

Commercio di calzature, pelletteria valigeria si classifica nella categoria 46.42.4 se ingrosso e 47.72.1 se dettaglio (la vendita delle calzature è prevalente).

La vendita di antiquariato è prevalente *su cose usate e oggetti usati* e si classifica nella categoria 47.79.2.

Vendita di libri senza ulteriore specifica si classifica nella categoria 47.61.0.

Vendita dettaglio di cartolibreria e articoli da regalo si classifica nella categoria 47.61.0.

Commercio al dettaglio cartolibreria cancelleria si classifica nella categoria 47.61.0.

Edicola e *cartolibreria* si classifica nella categoria 47.62.1 (è prevalente l'edicola).

Commercio al dettaglio giornali giocattoli cartoleria si classifica nella categoria 47.62.1.

La vendita di generi di monopolio è prevalente su, *cartoleria, cartolibreria, ricevitoria lotto e totocalcio, mercerie e bazar, articoli da regalo e profumeria.*

La vendita di generi di monopolio ed edicola si classifica invece nella categoria 47.62.1.

La vendita di porchetta sottintende la ristorazione ambulante e si classifica nella categoria 56.10.4.

Commercio e noleggio videocassette, cd, dvd si classifica nella categoria 77.22.0.

La vendita di attrezzature e animali domestici è prevalente *sulla vendita minuto di articoli da pesca* e si classifica nella categoria 47.76.2.

Diversamente dal commercio all'ingrosso, dove sono state inserite *empiriche* nella classe corrispondente dell'ingrosso associate solo alla parola commercio con prodotti specifici, per il commercio al dettaglio sono state inserite, nei file di *parsing,* le seguenti trasformazioni che hanno permesso di considerare sempre come *dettaglio* il commercio associato a taluni prodotti o attività:

Commercio macelleria = dettaglio macelleria.

Commercio merceria = dettaglio merceria.

Commercio chincaglieria = dettaglio chincaglieria.

Per il *commercio ambulante:*

Commercio ambulante senza altre specifiche si classifica nella categoria generica 47.89.0.

Per il *commercio diretto:*

Commercio porta a porta = commercio diretto.

Si sono stabiliti i seguenti codici non completi (a due, tre e quattro cifre):

47	(a2)	Case d'asta al dettaglio e vendite all'asta via internet
47	(e2)	Commercio al dettaglio
47	(a2)	Commercio al dettaglio (escluso quello di autoveicoli e di motocicli)
47	(e2)	Commercio al dettaglio biancheria

47	(a2)	Commercio al dettaglio di articoli sportivi, biciclette, armi e munizioni, di articoli per il tempo libero
47	(a2)	Commercio al dettaglio di elettrodomestici, apparecchi radio, televisori, lettori e registratori di dischi e nastri
47	(e2)	Commercio al dettaglio sede fissa
47	(a2)	Commercio dettaglio elettrodomestici, apparecchi radio, televisori, lettori e registratori
47	(e7)	Attività principale vendita al minuto
47.1	(a2)	Commercio al dettaglio in esercizi non specializzati
47.11	(a2)	Commercio al dettaglio in esercizi non specializzati con prevalenza di prodotti alimentari e bevande
47.19	(a2)	Commercio al dettaglio in esercizi non specializzati con prevalenza di prodotti non alimentari
47.2	(a2)	Commercio al dettaglio di prodotti alimentari, bevande e tabacco in esercizi specializzati
47.2	(ei)	Commercio dettaglio alimentari cotti
47.2	(ei)	Commercio dettaglio alimenti preconfezionati
47.29	(a2)	Altro commercio al dettaglio di prodotti alimentari, bevande e tabacco in esercizi specializzati
47.51	(a2)	Commercio al dettaglio di tessili
47.51	(e2)	Commercio dettaglio prodotti tessili
47.59	(a2)	Commercio al dettaglio di mobili e di articoli d'illuminazione
47.59	(a2)	Commercio al dettaglio di mobili, di articoli per l'illuminazione e articoli per la casa n.c.a.
47.6	(a2)	Commercio al dettaglio di libri, giornali, riviste e articoli di cartoleria
47.7	(a2)	Commercio al dettaglio di altri prodotti in esercizi specializzati
47.7	(e2)	Negozio artigianato
47.72	(a2)	Commercio al dettaglio di calzature e articoli in cuoio
47.73	(a2)	Commercio al dettaglio di prodotti farmaceutici, medicali, cosmetici e di articoli di profumeria
47.9	(a2)	Commercio al dettaglio al di fuori dei negozi
47.91	(a2)	Commercio al dettaglio per corrispondenza
47.91	(a2)	Commercio al dettaglio per corrispondenza e per televisione di prodotti non alimentari
47.91	(a2)	Commercio al dettaglio per corrispondenza, telefono, televisione di prodotti alimentari
47.99	(a2)	Commercio al dettaglio effettuato in altre forme al di fuori dei negozi
47.99	(a2)	Commercio al dettaglio effettuato in forme al di fuori dei negozi

La divisione contiene 2.393 empiriche.

Divisione 49 (Trasporto terrestre e trasporto mediante condotte)

Si sono stabiliti i seguenti codici non completi (a due, tre e quattro cifre):

49	(e2)	Ente trasporti
49	(e2)	Trasporti strada
49	(a2)	Trasporti terrestri
49	(a2)	Trasporti terrestri; trasporti mediante condotte
49.3	(a2)	Trasporti terrestri di passeggeri

49.50 (a1) Gestione delle centrali di spinta dislocate lungo la rete delle condotte
49.50 (a2) Trasporti mediante condotte
49.50 (b2) Trasporto mediante condotte di altri prodotti
49.50 (b2) Trasporto mediante condotte di slurry

La divisione contiene 345 empiriche.

Nella divisione 50 (Trasporto marittimo e per vie d'acqua)

Si sono stabiliti i seguenti codici non completi (a due cifre):

50 (a1) Trasporti lagunari
50 (a2) Trasporti marittimi e per vie d'acqua
50 (ei) Trasporti per vie d'acqua compresi i trasporti lagunari
50 (a1) Trasporti per vie d'acqua interne
50 (b1) Trasporti per vie d'acqua interne (compresi i trasporti lagunari)
50 (e7) Trasporto marittimo

La divisione contiene 59 empiriche.

Divisione 51 (Trasporto aereo)

Si sono stabiliti i seguenti codici non completi (a due cifre):

51 (e2) Azienda trasporti aerei
51 (a2) Trasporti aerei
51 (e2) Trasporto aereo passeggeri e merci

La divisione contiene 32 empiriche.

Divisione 52 (Magazzinaggio e attività di supporto ai trasporti)

Autorimessa > soccorso ACI si classifica nella categoria 52.21.5.

Il facchinaggio inteso come manovalanza, carico scarico senza ulteriore specifica si classifica nella categoria 52.24.4.

Le attività di pilotaggio e ancoraggio all'interno del porto si classificano nella categoria 52.22.0.

Si sono stabiliti i seguenti codici non completi (a due, tre e quattro cifre):

52 (a2) Attività di supporto e ausiliarie dei trasporti; attività delle agenzie di viaggio
52 (a2) Movimentazione merci e magazzinaggio
52.10 (ei) Immagazzinaggio
52.2 (a2) Altre attività connesse ai trasporti
52.24 (a2) Movimentazione merci
52.29 (a2) Attività delle altre agenzie di trasporto

La divisione contiene 539 empiriche.

Divisione 55 (Alloggio)

Per le attività doppie che riguardano "Alberghi" si è stabilito che:
Albergo > ristorante.
Attività alberghiera > attività termale.

Hotel > villaggio.
La divisione contiene 195 empiriche.

Divisione 56 (Attività dei servizi di ristorazione)

Per le attività doppie che riguardano *i servizi di ristorazione* si è stabilito che:
Ristorante > bar.
Mensa > catering.
Ristorazione senza nessuna specifica si classifica nel gruppo 56.1.
Per il catering:
Preparazione confezione distribuzione pasti catering si classifica nel gruppo 56.2.
Catering senza nessuna specifica si classifica nel gruppo 56.2.
Per le attività doppie che riguardano il *Bar* si è stabilito che :
Bar > gelateria.
Bar > pasticceria.
Bar > pizzeria a taglio.
Bar > paninoteca.
Bar > osteria.
Bar > birreria.
Bar > tabacchi.
Bar > night club.
Bar > discoteca.
Bar > sala da ballo.
Le stesse prevalenze valgono per il *caffè* inteso come bar.
Si sono stabiliti i seguenti codici non completi (a due, tre e quattro cifre):

56	(a7)	Servizi di ristorazione
56.1	(e2)	Attività di ristorazione
56.1	(e2)	Impresa di ristorazione
56.1	(e2)	Ristorazione
56.2	(ei)	Catering
56.2	(ei)	Preparazione confezione distribuzione pasti catering
56.2	(ei)	Prestazione servizi di ristorazione c/t catering
56.2	(ei)	Ristorazione commerciale catering
56.2	(ei)	Servizi di catering
56.2	(ei)	Servizi ristorazione c/t catering
56.29	(a2)	Mense e fornitura di pasti preparati

La divisione contiene 195 empiriche.

Divisione 58 (Attività editoriali)

L'edizione di libri prevale sull'edizione di riviste, periodici e altre pubblicazioni:
Edizione libri riviste periodici: 58.11.0.
Edizioni libri riviste opuscoli altre pubblicazioni: 58.11.0.
L'edizione di giornali prevale sui libri:
Edizione giornali e libri: 58.13.0.
Si sono stabiliti i seguenti codici non completi (a due e tre cifre):

58	(ei)	Attività editoriali

58	(e7)	Editoria
58.1	(e7)	Edizione quotidiani e periodici

La divisione contiene 132 empiriche.

Divisione 59 (Attività di produzione cinematografica, di video e di programmi televisivi, di registrazioni musicali e sonore)

Attore doppiatore si classifica nella categoria 90.01.0. (prevale l'attività di attore).
Doppiatore si classifica nella categoria 59.12.0.
Si sono stabiliti i seguenti codici non completi (a tre cifre):

59.1	(ei)	Edizione e distribuzione video cartoni e documentari
59.1	(a2)	Produzioni e distribuzioni cinematografiche e di video

La divisione contiene 213 empiriche.

Divisione 60 (Attività di programmazione e trasmissione)

Si sono stabiliti i seguenti codici non completi (a due cifre):

60	(a1)	Attività radiotelevisive
60	(ei)	Attività radiotelevisive e produzione programmi
60	(ei)	Attività radiotelevisione
60	(ei)	Attività radiotelevisiva e produzione programmi televisivi
60	(e2)	Attività radiotelevisiva locale
60	(ei)	Diffusione programmi radio tv
60	(ei)	Diffusione radio televisiva
60	(ei)	Emittente radio televisiva privata
60	(ei)	Emittente radiotelevisiva
60	(ei)	Esercizio emittente radiotelevisione
60	(ei)	Gestione emittenti radio televisive c/p e c/t
60	(ei)	Produzione di programmi radio e tv altre attività radiotelevisive
60	(b1)	Produzione di programmi radiofonici e televisivi
60	(b1)	Produzione di programmi radiofonici e televisivi sia in diretta sia registrati su nastro o su altro supporto
60	(ei)	Produzione programmi gestione rete radio televisiva
60	(ei)	Radiotelevisione privata
60	(ei)	Realizzazione programmi radiotelevisivi
60	(ei)	Servizio emittenza radiotelevisiva
60	(ei)	Stazione radio televisiva
60	(ei)	Stazione radio tv telediffusioni
60	(ei)	Trasmissione diffusione programmi radio televisivi

La divisione contiene 64 empiriche.

Divisione 61 (Telecomunicazioni)

Si sono stabiliti i seguenti codici non completi (a due cifre):

61	(a2)	Fornitura di accesso a internet (provider)
61	(e2)	Gestione provider

61	(e2)	Internet provider
61	(e2)	Internet service provider
61	(e2)	Operatore telecomunicazioni soluzioni connettività internet
61	(a2)	Telecomunicazioni
61	(e7)	Servizi di interconnessione internet

La divisione contiene 93 empiriche.

Divisione 62 (Produzione di software, consulenza informatica e attività connesse)

La consulenza informatica prevale sulla *produzione* software e si classifica nella categoria 62.02.0.

La gestione realizzazione software per pagine Web si classifica nella categoria 62.01.0.

Si è convenuto di considerare per *sistemi informatici* anche l'hardware pertanto, per esempio:

Consulenza tecnica di sistemi informatici si classifica nella categoria 62.02.0.

Si sono stabiliti i seguenti codici non completi (a due cifre):

62	(e2)	Comunicazione e informatica

La divisione contiene 170 empiriche.

Divisione 63 (Attività dei servizi d'informazione e altri servizi informatici)

Hosting siti Web si classifica nella categoria 63.11.3.

Gestione pagine internet si classifica nella categoria 63.11.3.

Servizi registrazione domini si classifica nella categoria 63.11.3.

La divisione contiene 224 empiriche.

Divisione 64 (Attività di servizi finanziari, escluse le assicurazioni e i fondi pensione)

Si sono stabiliti i seguenti codici non completi (a due, tre e quattro cifre) come:

64	(a2)	Intermediazione monetaria e finanziaria (escluse le assicurazioni e i fondi pensione)
64.1	(a2)	Intermediazione monetaria
64.19	(a2)	Altre intermediazioni monetarie
64.9	(a2)	Altre intermediazioni finanziarie
64.9	(e2)	Finanziamenti

La divisione contiene 172 empiriche.

Divisione 65 (Assicurazioni, riassicurazioni e fondi pensione, escluse le assicurazioni sociali obbligatorie)

Si sono stabiliti i seguenti codici non completi (a due e quattro cifre):

65	(a2)	Assicurazioni e fondi pensione (escluse le assicurazioni sociali obbligatorie)
65	(e2)	Compagnia di assicurazione
65	(e2)	Impresa di assicurazioni
65	(e2)	Istituto di assicurazione

65.30 (e2) Ente di assicurazione pensioni
65.30 (b2) Erogazione dei redditi da pensione
65.30 (a2) Fondi pensione
65.30 (e2) Fondi pensione non obbligatoria

La divisione contiene 40 empiriche.

Divisione 66 (Attività ausiliarie dei servizi finanziari e delle attività assicurative)

Si sono stabiliti i seguenti codici non completi (a tre e quattro cifre):

66.1 (a2) Attività ausiliarie della intermediazione finanziaria (escluse le assicurazioni e i fondi pensione)
66.19 (a2) Attività ausiliarie della intermediazione finanziaria n.c.a.
66.19 (b2) Attività ausiliare della intermediazione finanziaria
66.2 (e7) Consulenza assicurativa automobilistica
66.2 (e7) Attività degli agenti periti e liquidatori indipendenti assicurazioni
66.2 (e7) Agenti periti e liquidatori delle assicurazioni

La divisione contiene 162 empiriche.

Divisione 68 (Attività immobiliari)

L'attività immobiliare generica senza specificare se su beni di proprietà, leasing o per conto terzi, si è stabilito di considerala sempre come una attività per conto terzi.

Attività immobiliari senza ulteriori specifiche si classifica nella categoria 68.31.0.

Cooperative edilizie e costruzione di alloggi per soci si classifica nella categoria 41.10.0.

Si sono stabiliti i seguenti codici non completi (a due cifre):

68 (ei) Gestione centri commerciali
68 (ei) Gestione negozi centri commerciali

La divisione contiene 204 empiriche.

Divisione 71 (Attività degli studi di architettura e d'ingegneria; collaudi ed analisi tecniche)

Si sono stabiliti i seguenti codici non completi (a tre e quattro cifre):

71.1 (ei) Progettazione edilizia
71.1 (a2) Studi di architettura e di ingegneria
71.1 (ei) Studio architettura urbanistica ingegneria
71.1 (ei) Studio consulenza nel campo ingegneria architettura
71.1 (ei) Studio tecnico architetti e geometri
71.1 (ei) Sviluppo progetti settori ingegneria architettura
71.1 (e7) Progettazione edilizia residenziale
71.12 (e7) Progettistica e direzione lavori

La divisione contiene 418 empiriche.

Divisione 72 (Ricerca scientifica e sviluppo)

Biologo marino si classifica nella categoria della ricerca 72.19.0 in quanto l'attività solo tecnica propria della categoria 74.90.9 è più rara.

Si sono stabiliti i seguenti codici non completi (a tre cifre):

72.1	(e2)	Ricerca applicata e sviluppo sperimentale
72.1	(a2)	Ricerca e sviluppo
72.1	(e2)	Ricerca e sviluppo sperimentale
72.1	(ei)	Ricerca scientifica

La divisione contiene 105 empiriche.

Divisione 73 (Pubblicità e ricerche di mercato)

Produzione di targhe ed insegne stradali pubblicitarie si classifica nella categoria 73.11.0.

La divisione contiene 279 empiriche.

Divisione 74 (Altre attività professionali, scientifiche e tecniche)

Si sono stabiliti i seguenti codici non completi (a quattro cifre):

74.20	(a2)	Attività inerenti alla fotografia

La divisione contiene 451 empiriche.

Divisione 77 (Attività di noleggio e leasing operativo)

Attività di noleggio e commercio nel caso di *videocassette e cd* predomina l'attività di noleggio e si classifica nella categoria 77.22.0.

Attività di gestione e noleggio nel caso di *giochi d'intrattenimento* predomina il noleggio e si classifica nella categoria 77.22.0.

Si sono stabiliti i seguenti codici non completi (a due cifre):

77	(a2)	Noleggio di macchinari e attrezzature senza operatore e di beni per uso personale e domestico
77	(a7)	Leasing operativo

La divisione contiene 352 empiriche.

Divisione 78 (Attività di ricerca, selezione, fornitura di personale)

Si sono stabiliti i seguenti codici non completi (a due cifre):

78	(e2)	Somministrazione lavoro

La divisione contiene 51 empiriche.

Divisione 79 (Attività dei servizi delle agenzie di viaggio, dei tour operator e servizi di prenotazione e attività connesse)

Si sono stabiliti i seguenti codici non completi (a tre cifre):

79.1	(a7)	Agenzia di viaggi tour operator
79.1	(a1)	Attività delle agenzie di viaggio e degli operatori turistici

La divisione contiene 81 empiriche.

Divisione 81 (Attività di servizi per edifici e paesaggio)

La pulizia e il facchinaggio si classifica nella categoria 81.29.9.

Si sono stabiliti i seguenti codici non completi (a tre cifre):

81.2	(a1)	Servizi di pulizia
81.2	(e7)	Attività di pulizie
81.2	(e7)	Impresa di pulizie e disinfezione
81.2	(ei)	Agenzia di pulizie
81.2	(ei)	Azienda servizi pulizia
81.2	(ep)	Cooperativa prestazione servizi di pulizia
81.2	(ei)	Impresa artigiana di pulizia

La divisione contiene 208 empiriche.

Divisione 84 (Amministrazione pubblica e difesa; assicurazione sociale obbligatoria)

Si sono stabiliti i seguenti codici non completi (a due, tre e quattro cifre):

84	(e2)	Amministrazione statale
84	(e2)	Ente statale
84	(e2)	Ente-pubblica amministrazione
84	(a2)	Pubblica amministrazione e difesa
84	(a2)	Pubblica amministrazione e difesa; assicurazione sociale obbligatoria
84.1	(a2)	Amministrazione pubblica, politica economica e sociale
84.11	(a2)	Attività generali della pubblica amministrazione
84.12	(a2)	Attività della pubblica amministrazione rivolta alla regolamentazione sanitaria, di regolamentazione dell'istruzione, di regolamentazione dei servizi culturali
84.13	(a2)	Attività della pubblica amministrazione rivolta alla regolamentazione delle attività economiche
84.2	(a2)	Servizi della pubblica amministrazione forniti alla intera collettività

La divisione contiene 343 empiriche.

Divisione 85 (Istruzione)

La scuola di musica senza nessuna specifica si classifica nella categoria 85.52.0.

L'asilo sia pubblico che privato è stato considerato come scuola materna e si classifica nella categoria 85.10.0.

L'asilo nido invece si classifica nella categoria 88.32.0.

Per le attività doppie che riguardano l'Istruzione, si è stabilito che:

Asili nido e scuole materne si classifica nella categoria 85.10.0.

Scuola elementare e media si classifica nella categoria 85.31.1.

Scuola materna, elementare e media si classifica nella categoria 85.31.1.

Istruzione primaria e secondaria primo grado si classifica nella categoria 85.31.1.

Si sono stabiliti i seguenti codici non completi (a due, tre e quattro cifre):

85	(e2)	Dipendente statale insegnante
85	(e2)	Formazione scolastica
85	(e2)	Insegnamento (scuola)
85	(e2)	Istituto di istruzione

85	(e2)	Istituto privato di istruzione
85	(e2)	Istituto scolastico
85	(a2)	Istruzione
85	(e2)	Istruzione di minori
85	(e2)	Scuola
85	(e2)	Scuola parificata
85	(e2)	Scuola privata
85	(e2)	Scuola privata laica
85	(e2)	Scuola statale
85.3	(ei)	Docente di ruolo scuola superiore
85.3	(ei)	Docente scuola media superiore
85.3	(ei)	Docente scuola superiore
85.3	(e2)	Gestione scuola istruzione secondaria
85.3	(ei)	Gestione scuola media inferiore superiore parificata
85.3	(e2)	Istituto collegiale privato scuola secondaria
85.3	(ep)	Istituto scolastico superiore
85.3	(e2)	Istruzione e formazione secondaria
85.3	(a2)	Istruzione secondaria
85.3	(a2)	Istruzione secondaria di formazione generale
85.3	(b1)	Istruzione secondaria di secondo grado
85.3	(b1)	Istruzione secondaria di secondo grado: licei e istituti che rilasciano diplomi di maturità
85.3	(e2)	Istruzione secondaria superiore
85.3	(a1)	Licei e istituti che rilasciano diplomi di maturità
85.3	(ei)	Scuola media superiore
85.3	(e2)	Scuola privata di istruzione secondaria
85.3	(ei)	Scuola privata superiore
85.3	(ei)	Scuola superiore
85.31	(e7)	Istruzione primaria e secondaria
85.4	(e2)	Istituto superiore di formazione manageriale

La divisione contiene 357 empiriche.

Divisione 86 (Assistenza sanitaria)

Il medico generico si classifica nella categoria 86.21.0.

Il medico condotto si classifica nella categoria 86.21.0.

Il presidio sanitario privato si classifica nella categoria 86.22.0.

Il medico psicologo si classifica nella categoria 86.22.0.

Lo psicologo professionista si classifica nella categoria 86.90.3.

I servizi di assistenza agli anziani ed ammalati si classificano nella categoria 87.10.0 in quanto si sottintende l'assistenza infermieristica residenziale.

Si sono stabiliti i seguenti codici non completi (a due, tre e quattro cifre):

86	(b7)	Assistenza dove la componente medica assume un carattere prevalente
86	(e2)	Cure mediche
86	(e7)	Attività sanitaria
86.10	(ei)	Servizio ospedaliero
86.10	(ei)	Attività ospedaliera
86.10	(ei)	Attività sanitaria ospedaliera

86.10 (ep) Ente ospedaliero
86.2 (a2) Servizi degli studi medici
86.2 (e7) Ambulatorio medico
86.90 (a2) Altre istituzioni sanitarie senza ricovero
86.90 (a2) Altri servizi sanitari

La divisione contiene 511 empiriche.

Divisione 87 (Servizi di assistenza sociale residenziale)

Si sono stabiliti i seguenti codici non completi (a due cifre):

87 (ei) Assistente sociale residenziale
87 (a1) Assistenza sociale residenziale

La divisione contiene 126 empiriche.

Divisione 88 (Assistenza sociale non residenziale)

Si sono stabiliti i seguenti codici non completi (a due cifre):

88 (ei) Servizi assistenza domiciliare

La divisione contiene 82 empiriche.

Divisione 90 (Attività creative, artistiche e di intrattenimento)

Il restauratore d'arte generico si classifica nella categoria 90.03.0.

Si sono stabiliti i seguenti codici non completi (a tre cifre):

90.0 (ei) Attività teatrale

La divisione contiene 308 empiriche.

Divisione 91 (Attività di biblioteche, archivi, musei ed altre attività culturali)

Si sono stabiliti i seguenti codici non completi (a due, tre e quattro cifre):

91.0 (b1) Attività dei musei e conservazione dei luoghi e dei monumenti storici
91.0 (a2) Attività di biblioteche, archivi, musei e altre attività culturali
91.0 (a2) Gestione di musei e del patrimonio culturale

La divisione contiene 88 empiriche.

Divisione 93 (Attività sportive, di intrattenimento e di divertimento)

Si sono stabiliti i seguenti codici non completi (a tre e quattro cifre):

93.1 (e2) Attività sportive tra soci
93.1 (a2) Attività sportive
93.11 (a2) Gestione di stadi e altri impianti sportivi
93.2 (e7) Attività ricreative

La divisione contiene 316 empiriche.

Divisione 94 (Attività di organizzazioni associative)

Si sono stabiliti i seguenti codici non completi (a tre e quattro cifre):

94.1	(a2)	Attività di organizzazioni economiche, di titolari di impresa, professionali
94.1	(a2)	Organizzazioni economiche, titolari di impresa, professionali
94.99	(e2)	Associazione

La divisione contiene 291 empiriche.

Divisione 95 (Riparazione di computer e di beni per uso personale e per la casa)

Si sono stabiliti i seguenti codici non completi (a tre cifre):

95.2	(a2)	Riparazione di beni di consumo personali e per la casa

La divisione contiene 350 empiriche.

Divisione 96 (Altre attività di servizi per la persona)

L'attività di parrucchiere è predominante sul *solarium* e si classifica nella categoria 96.02.0.

Si sono stabiliti i seguenti codici non completi (a quattro cifre):

96.01	(a2)	Servizi di lavanderia, pulitura a secco e tintura di articoli tessili e pellicce
96.01	(e7)	Lavanderia
96.01	(e7)	Servizi delle lavanderie

La divisione contiene 345 empiriche.

4.3 Informazioni di carattere generale

Nell'ambito poi delle varie divisioni permangono queste informazioni di carattere generale:

- Se non è specificato il metallo si presume che si tratti di ferro.
- Estrusione è un processo di lavorazione per preparare tubi, lastre, barre, profilati.
- Per le barche si sottintende che siano da diporto.
- Oli non commestibili si intendono gli oli vegetali come quelli di sansa finali.
- Lavorazione stocco deve essere inteso come lavorazione dello stoccafisso, baccalà.
- Addolcitori acqua sono apparecchi anticalcare.
- Ovoprodotti sono albumina di uova, tuorli congelati, essiccati, freschi liquidi (particolarmente nell'industria alimentare); uova di volatili essiccate conservate eccetera.
- Le motoseghe vengono considerate accessori e utensili agricoli.
- Gli articoli igienico-sanitari sono bagni e docce intesi come materiali da costruzione.
- Rottami sono più che altro metallici.
- Sistema operativo = software.
- Quando si parla di carburante per combustione è sottinteso per uso domestico.
- Quando si parla di progettazione e realizzazione (produzione) predomina la realizzazione.
- Engineering = ingegneria integrata.
- Guardia giurata = sorveglianza privata.

4.4 Trattamento delle risposte non significative

L'analisi degli output di ACTR ha dimostrato una casistica di risposte prive di un contenuto sufficiente per l'attribuzione di un codice corretto, che tuttavia venivano associate a codici

particolari e veicolati negli Unici erroneamente (l'errore di codifica può derivare dal fatto che sono spesso risposte molto brevi, in genere, rappresentate da singole parole, alle quali l'algoritmo per l'attribuzione dei pesi alle parole assegna un valore elevato).

Per ovviare a tale problematica, queste risposte sono state inserite nel dizionario associate alla sigla n.c., che sta per "non codificabile":

n.c	(e2)	Attività edilizia
n.c.	(e2)	Edilizia
n.c.	(nc)	Abbigliamento
n.c.	(ep)	Alleanza nazionale
n.c.	(a2)	Altre attività di tipo professionale e imprenditoriale n.c.a.
n.c.	(a2)	Altre attività di tipo professionale n.c.a.
n.c.	(a2)	Altre attività professionali e imprenditoriali
n.c.	(e2)	Altre attività n.c.a.
n.c.	(e2)	Altre attività professionali
n.c.	(n2)	Altri servizi
n.c.	(nc)	Altro fibra
n.c.	(nc)	Amministrazione
n.c.	(ep)	Arredi sanitari
n.c.	(ei)	Artigiano
n.c.	(nc)	Attività autonoma
n.c.	(nc)	Attività non specializzata
n.c.	(n2)	Attività di servizi
n.c.	(ei)	Attività esclusiva
n.c.	(ep)	Azienda addetti alle vendite
n.c.	(ep)	Azienda commerciale
n.c.	(ep)	Azienda impianti
n.c.	(ep)	Azienda speciale multiservizi
n.c.	(nc)	C/p c/t
n.c.	(nc)	Caccia pesca
n.c.	(n2)	Carpenteria
n.c.	(nc)	Cessata in liquidazione
n.c.	(nc)	Civile
n.c.	(n2)	Commercio
n.c.	(nc)	Commercio e confezione di propria produzione
n.c.	(nc)	Commercio parti ed accessori
n.c.	(nc)	Confezione
n.c.	(nc)	Confezione industriale
n.c.	(nc)	Confezione riparazione
n.c.	(nc)	Confezioni
n.c.	(n2)	Confezioni conto terzi
n.c.	(nc)	Consorzio
n.c.	(n2)	Consulenza
n.c.	(nc)	Conto
n.c.	(ei)	Conto lavoro per terzo
n.c.	(nc)	Conto socio
n.c.	(n2)	Conto terzi
n.c.	(e2)	Cooperativa di servizi
n.c.	(nc)	Cooperativa sociale

n.c.	(nc)	Costruzioni
n.c.	(nc)	Direzione e amministrazione
n.c.	(a2)	Estrazione di petrolio greggio e di gas naturale; servizi connessi all'estrazione del petrolio e di gas naturale, esclusa la prospezione
n.c.	(a2)	Fabbricazione di coke, raffinerie di petrolio, trattamento dei combustibili nucleari
n.c.	(a2)	Fabbricazione di macchine per ufficio, di elaboratori e sistemi informatici
n.c.	(a2)	Fabbricazione di mobili; altre industrie manifatturiere
n.c.	(ep)	Famiglia
n.c.	(nc)	Fiduciarie
n.c.	(nc)	Fiore
n.c.	(nc)	Impiegato statale
n.c.	(nc)	In liquidazione
n.c.	(n2)	Indipendente
n.c.	(nc)	Intimo
n.c.	(nc)	Intrecciati
n.c.	(nc)	Lavorazione autonoma
n.c.	(nc)	Lavorazione in proprio c/terzi
n.c.	(nc)	Lucidatura
n.c.	(nc)	Lucidatura c/t
n.c.	(nc)	Manutenzione e riparazione apparecchiature
n.c.	(nc)	Miscelazione
n.c.	(nc)	Monopolio
n.c.	(nc)	Montaggi meccanici
n.c.	(nc)	Montaggio
n.c.	(nc)	Ora in liquidazione
n.c.	(ep)	Pavimenti industriali
n.c.	(nc)	Pensionato
n.c.	(n2)	Prestazione di servizi
n.c.	(nc)	Prestazioni conto terzi
n.c.	(nc)	Produzione
n.c.	(nc)	Produzione e lavorazione
n.c.	(nc)	Produzione e lavorazione c/t
n.c.	(nc)	Produzione e lavorazione c/t parti
n.c.	(nc)	Produzione e servizi
n.c.	(n2)	Progettazione
n.c.	(n2)	Rappresentante
n.c.	(nc)	Riparazione
n.c.	(nc)	Riparazione c/t
n.c.	(nc)	Sede
n.c.	(ei)	Servizi
n.c.	(e2)	Servizi alle imprese
n.c.	(ei)	Servizi svolti per conto soci
n.c.	(nc)	Servizio automobilistico
n.c.	(ep)	Società di distribuzione
n.c.	(nc)	Società in liquidazione
n.c.	(nc)	Società in liquidazione volontaria
n.c.	(e2)	Socio lavoratore

n.c.	(ei)	Studio
n.c.	(ei)	Studio tecnico
n.c.	(ei)	Studio tecnico progettazione
n.c.	(ep)	Terzo settore
n.c.	(nc)	Uffici commerciali
n.c.	(nc)	Ufficio
n.c.	(nc)	Ufficio di rappresentanza
n.c.	(ep)	Ufficio tecnico
n.c.	(nc)	Vendita e amministrazione
n.c.	(ep)	Vendita ed assistenza tecnica
n.c.	(e7)	Attività di consulenza
n.c.	(e7)	Locazioni
n.c.	(e7)	Pubblico

Capitolo 5 - Analisi di qualità dei risultati dell'applicazione di codifica

La qualità dei risultati dei sistemi di codifica automatica, come già accennato nel cap.1, si misura in termini di due parametri:

- Il tasso di codifica o efficacia (*recall rate*) → percentuale dei testi codificati univocamente sul totale dei testi da codificare;
- Il tasso di accuratezza (*precision rate*) → percentuale dei testi codificati in modo univoco correttamente sul totale dei testi codificati.

La qualità delle applicazioni di codifica automatica precedentemente implementate era stata sempre misurata su campioni di descrizioni abbastanza omogenee tra loro, derivanti da indagini Istat (2005 Cuccia e altri, 2002 Macchia e altri). Dai risultati ottenuti (cfr. tavola 1.1, capitolo 1) si evince come i tassi di codifica siano tendenzialmente più elevati per le indagini sulle imprese, il che è spiegato dal fatto che il concetto di "attività economica", senz'altro noto per le imprese, non è altrettanto chiaro per le famiglie.

L'ambiente di codifica automatica Ateco 2007 sarà invece utilizzato su diverse tipologie di testi, il che ha posto delle problematiche particolari nella progettazione del sistema di controllo della qualità. In particolare, come già citato nel paragrafo 1.3, l'applicazione sarà utilizzata per tre diverse finalità:

- codificare le descrizioni delle attività economiche fornite dai rispondenti ai questionari delle indagini Istat (siano esse sulle imprese o sulle famiglie);
- codificare le descrizioni delle attività economiche delle imprese, provenienti da archivi esterni (quali l'archivio delle Camere di commercio);
- fornire agli utenti del sito Web dell'Istat una funzione di consultazione dell'ambiente di codifica automatica, finalizzato all'individuazione dell'attività economica espletata dalle imprese.

Le descrizioni fornite dagli utilizzatori di queste tre funzioni sono indubbiamente non omogenee, infatti:

- le risposte fornite alle indagini Istat sono tendenzialmente sintetiche e sono fornite sulla base delle istruzioni per la compilazione dei questionari;
- le descrizioni provenienti da archivi esterni sono invece spesso discorsive e contengono elementi non significativi ai fini dell'individuazione dell'attività economica;
- relativamente agli utenti del sito Web, l'eterogeneità dei soggetti interessati non consente di delineare una tipologia di testi da trattare.

Per questi motivi, l'analisi di qualità è stata effettuata progettando tre test, diversi sia per la metodologia adottata per la valutazione sia per le caratteristiche dei campioni oggetto di analisi.

5.1 I test: sui dati censuari

Al fine di misurare le performance dell'applicazione su dati rilevati in indagini statistiche, sono stati utilizzati i testi del Censimento dell'industria 2001. Data l'elevata numerosità delle descrizioni da trattare, è stata adottata una metodologia pensata in modo da ottimizzare il processo di analisi dei risultati, così da non esaminare più volte testi tra di loro molto simili (D'Orazio, Macchia 2002).

Il primo passo previsto secondo questa metodologia consiste nell'individuare nell'insieme delle descrizioni quelle tra di loro effettivamente differenti; a tal fine, il file da trattare viene sottoposto ad un processo di *parsing* semplificato con il quale si eliminano articoli, punteggiatura e poco altro. A ciascuna descrizione viene quindi associata la frequenza con la quale è stata rilevata. Poi, sulla base dell'analisi delle frequenze, si definiscono delle classi di frequenza da utilizzarsi per la stratificazione in base a cui estrarre il campione.

Nel caso in esame, l'universo era costituito da 1.130.693 descrizioni che, a seguito del *parsing* ridotto, corrispondevano a 228.738 testi tra di loro diversi. Questi ultimi sono stati suddivisi in 13 classi di frequenza, dalle quali è stato estratto un campione tale che il tasso di codifica atteso (π) per le diverse classi fosse lo stesso, mentre l'errore campionario sulle stime decrescesse al crescere dell'ampiezza delle classi, in modo da far sì che per le classi più ampie fossero analizzate pressoché tutte le descrizioni (il Δ della tavola è l'ampiezza dell'intervallo di confidenza).

Come può vedersi dalla tavola 5.1, è stato quindi estratto un campione di 3.191 descrizioni.

Tavola 5.1 - Determinazione del campione in base alla precisione che si intende ottenere, basata sulla freq. del num. di testi unici (std.) per classe

Popolazione						Campione						
Classe (freq.)		N_i	Pr. (π)	$\pm\Delta$	Testi standar-dizzati (Unici)	n_0	opt(n)	n=arrotonda (opt(n))	%	Testi standardizzati		
										$\sum_i freq/N_i$	testi ponderati	%
1°	1	169.420	0,75	0,040	169.420	234,3750	234,051	234	0,14	1	234	0,14
2°	2	55.790	0,75	0,040	27.895	234,3750	233,395	233	0,42	2	466	0,84
3°	3-8	92.401	0,75	0,035	20.906	306,1224	305,112	305	0,33	4,95	1.510	1,63
4°	9-25	91.828	0,75	0,035	6.475	306,1224	305,105	305	0,33	15,62	4.764	5,19
5°	26-50	65.636	0,75	0,030	1.854	416,6667	414,038	414	0,63	36,78	15.227	23,20
6°	51-90	60.638	0,75	0,030	902	416,6667	413,823	414	0,68	69,36	28.715	47,35
7°	91-130	38.941	0,75	0,030	360	416,6667	412,256	360	0,92	109,43	38.941	100,00
8°	131-180	38.105	0,75	0,030	247	416,6667	412,16	247	0,65	155,52	38.105	100,00
9°	181-300	56.913	0,75	0,020	246	937,5000	922,307	246	0,43	236,03	56.913	100,00
10°	301-430	46.462	0,75	0,020	128	937,5000	918,957	128	0,28	366,60	46.462	100,00
11°	431-730	78.602	0,75	0,010	139	3750,0000	3579,24	139	0,18	577,92	78.602	100,00
12°	731-1410	83.200	0,75	0,010	83	3750,0000	3588,27	83	0,10	1.043,48	83.200	100,00
13°	>= 1411	252.726	0,75	0,010	83	3750,0000	3695,17	83	0,03	4.253,26	252.726	100,00
	Totale	**1.130.662**			**228.738**			**3.191**	**0,28**		**645.865**	**57,12**

Il campione è stato sottoposto a codifica ottenendo i risultati riportati nella tavola 5.2.

Tavola 5.2 - Risultati della codifica automatica su campione del Censimento

RECORD	Valori assoluti	Valori percentuali
Unici	2.504	78,47
Multipli	97	3,04
Possibili	545	17,08
Falliti	45	1,41
Totale	**3.191**	**100,00**

L'efficacia ottenuta su questo campione è indubbiamente più che soddisfacente, tanto più se si analizza per classe di frequenza. Come si può vedere, infatti, mentre gli Unici si distribuiscono su tutte le classi, comprese quelle corrispondenti a valori più elevati, non si rilevano falliti corrispondenti alle classi superiori a 180 occorrenze.

Inoltre, il 71,25 per cento degli Unici (corrispondenti a 1.784 match) hanno punteggio uguale a 10, ossia sono frutto di un abbinamento diretto con testi presenti nel dizionario e di questi ultimi oltre il 53 per cento corrispondono alle classi con occorrenze superiori a 91, il che testimonia un arricchimento della base informativa in linea con il modo di esprimersi dei rispondenti.

Tavola 5.3 - Risultati della codifica automatica per classi di frequenza e punteggio

CLASSI DI FREQUENZA	Unici				Falliti	
	Valori assoluti	Valori percentuali sul totale degli Unici	Match diretti (Punteggio = 10)		Valori percentuali	Valori percentuali sul totale dei falliti
			Valori assoluti	% sul totale dei match diretti		
1	116	4,63	29	1,63	14	31,11
2	131	5,23	49	2,75	3	6,67
3-8	203	8,11	100	5,61	8	17,78
9-25	224	8,95	153	8,58	4	8,89
26-50	331	13,22	231	12,95	6	13,33
51-90	352	14,06	261	14,63	6	13,33
91-130	304	12,14	249	13,96	3	6,67
131-180	218	8,71	183	10,26	1	2,22
181-300	219	8,75	180	10,09		
301-430	116	4,63	101	5,66		
431-730	131	5,23	110	6,17		
731-1.410	80	3,19	70	3,92		
$\geq$ 1.411	79	3,15	68	3,81		
Totale	**2.504**	**100,00**	**1.784**	**100,00**	**45**	**100,00**

Passando all'analisi di qualità, come può vedersi dalla tavola 5.4, il 95 per cento dei codici assegnati sono corretti; dei non corretti, la parte più significativa è quella dei non codificabili, ossia di descrizioni troppo generiche, oppure che esprimono più di un'attività economica, a cui però il sistema ha associato un singolo codice.

Una ulteriore analisi ha dimostrato inoltre come in ciascuna classe di frequenza, la ripartizione tra corretti e non corretti (errati e non codificabili) sia abbastanza uniforme; un lieve decremento della percentuale di corretti si rileva tuttavia nelle classi intermedie: 9-25, 26-50 e 51-90.

Prendendo in esame il punteggio, il 98 per cento dei codificati con punteggio 10 sono corretti, a fronte dell'85 per cento di quelli con punteggio 8.

Gli errati ed i non codificabili si concentrano invece tra i codificati con punteggio 8 (rispettivamente il 72,20 per cento ed il 67,44 per cento); un innalzamento della qualità sarebbe ottenibile con uno spostamento della soglia minima verso l'alto, tuttavia, rappresentando i codificati con punteggio 8 il 22,36 per cento ed essendo comunque elevata l'accuratezza sul totale dei codificati, non si ritiene opportuno procedere in tal senso.

Tavola 5.4 - Accuratezza della codifica automatica su campione del Censimento

RISULTATI DELLA CODIFICA AUTOMATICA		Match indiretti		Match diretti	Totale
		Punteggio = 8	Punteggio = 9	Punteggio = 10	
Corretto	Frequenza	476	156	1750	2382
	Percentuale	19,01	6,23	69,89	95,13
	% riga	19,98	6,55	73,47	
	% colonna	85,00	97.50	98,09	
Errato	Frequenza	26	3	7	36
	Percentuale	1,04	0,12	0,28	1,44
	% riga	72,22	8.33	19,44	
	% colonna	4,64	1,88	0,39	
Non codificabile	Frequenza	58	1	27	86
	Percentuale	2,32	0,04	1,08	3,43
	% riga	67,44	1,16	31,40	
	% colonna	10,36	0,63	1,51	
Totale	Frequenza	560	160	1784	2504
	Percentuale	22,36	6,39	71,25	100,00

5.2 II test: sull'archivio delle Camere di commercio

Nell'ottica di valutare le performance dell'applicazione su dati provenienti da fonti esterne all'Istituto, quindi non rilevati nell'ambito di indagini statistiche, sono stati utilizzati i testi forniti dalle Camere di commercio; in particolare, dell'archivio completo, le cui caratteristiche sono descritte nel cap. 6, sono state trattate per l'analisi di qualità soltanto le descrizioni più corte di 200 byte.

Come specificato di seguito, questi testi sono stati sottoposti a codifica automatica tramite ACTR, ottenendo un tasso di codifica del 61 per cento, che rappresenta di per sé un ottimo risultato.

Come già accennato, queste descrizioni si caratterizzano per il fatto di essere spesso più discorsive rispetto a quelle rilevate con i questionari di indagine; infatti, nel momento in cui le imprese si iscrivono alle Camere di commercio tendono a fornire una molteplicità di elementi non sempre pertinenti rispetto all'attività economica espletata nel contingente, per esempio dilungandosi su particolari inerenti la costituzione dell'impresa, le attività espletate in periodi precedenti eccetera.

La peculiarità di queste descrizioni implica che sia abbastanza raro che più descrizioni siano uguali tra di loro, quindi non si è ritenuto opportuno adottare per questo test la stessa strategia di campionamento utilizzata nel primo. D'altro canto, per un sottoinsieme di queste descrizioni più corte di 200 byte si disponeva di codici Ateco a cinque cifre assegnati tramite gli Studi di Settore.[6] Si è quindi ritenuto opportuno sfruttare questa informazione, utilizzandola come termine di confronto; non essendoci tuttavia alcuna garanzia circa la qualità di questi codici, la metodologia scelta è stata quella di sottoporre al codificatore esperto soltanto le descrizioni per le quali non si verificasse una completa coincidenza tra il codice assegnato da ACTR e quello già presente nell'archivio, in quanto si è assunto che se due soggetti diversi avevano assegnato lo stesso codice alla stessa descrizione, avvalendosi di diverse metodologie, si poteva concludere che il codice fosse esatto.

[6] Gli Studi di Settore sono un'indagine condotta annualmente dall'Agenzia delle Entrate sulle piccole e medie imprese. L'Istat li utilizza da diversi anni e ha avuto modo di verificarne l'elevata qualità dell'informazione contenuta dovuta a diversi elementi: a) l'obbligo di risposta implicito nell'indagine condotta dall'Agenzia delle Entrate, b) il dettaglio molto spinto delle variabili indagate, c) l'estrema vicinanza tra le variabili e le metodologie utilizzate nei questionari con quelle rilevate dall'Istat. Gli Studi di Settore sono stati la principale fonte utilizzata per la riclassificazione puntuale delle imprese presenti in Asia e coprono almeno il 70 % delle imprese contenute nell'Archivio.

L'analisi ha quindi riguardato il sottoinsieme di descrizioni dell'archivio codificate da ACTR, per le quali era già disponibile il codice a cinque cifre assegnato tramite gli Studi di Settore.

L'attribuzione dei codici da parte di ACTR a queste descrizioni è riportata nella seguente tavola.

Tavola 5.5 - Codici ACTR assegnati alle descrizioni che già disponevano del codice a 5 cifre assegnato con gli Studi di Settore

TIPOLOGIA DI CODICI ASSEGNATI	Numero di record
Codici a 2 cifre	441
Codici a 3 cifre	2.719
Codici a 4 cifre	178
Codici a 5 cifre	80.779
Totale	**84.117**

Per l'analisi delle non coincidenze ci si è quindi avvalsi degli 80.779 record cui anche ACTR ha assegnato un codice a cinque cifre.

Su questi è stato effettuato il confronto tra i due codici, individuando i diversi livelli di dettaglio delle coincidenze. La tavola 5.6 riporta i risultati di questa analisi.

Tavola 5.6 - Confronto codici assegnati da ACTR e dagli Studi di Settore

TIPOLOGIA DI CODICI ASSEGNATI	Valori assoluti	Valori percentuali
Uguali a 5 cifre	54.141	67
Uguali le prime 4 cifre	2.948	4
Uguali le prime 3 cifre	4.858	6
Uguali le prime 2 cifre	5.953	7
Uguale solo la 1° cifra	6.926	9
Completamente diversi	5.953	7
Totale	**80.779**	**100**

Come può vedersi, il fatto che il 67 per cento dei codici siano completamente coincidenti rappresenta di per sé un buon risultato in termini di qualità.

Il bacino di descrizioni rispetto alle quali concentrare l'analisi di qualità è quindi costituito da 26.638 record nei quali i codici divergono.

Tuttavia, come sarà descritto dettagliatamente nell'ambito del progetto di riclassificazione dell'archivio delle Camere di commercio (par. 6.4), è stata adottata una strategia secondo la quale venivano associati agli Unici anche quei Multipli e Possibili che avevano individuato nel dizionario descrizioni simili a quella da codificare, alle quali corrispondeva però sempre lo stesso codice. È chiaro che, visto il punteggio proprio dei Multipli e dei Possibili, il livello di similarità tra il testo da codificare e quelli abbinati del dizionario era inferiore, per cui anche l'attendibilità del codice non garantiva gli stessi livelli qualitativi; poiché tale strategia era esclusivamente funzionale al progetto citato, ma non condivisa da altre applicazioni di codifica automatica, è stato deciso di non considerare questa tipologia di codificati per l'analisi di qualità, per cui il bacino di record cui attingere è risultato essere di 17.746 record.

Per l'analisi di qualità è stato quindi estratto da questo archivio un campione di 4 mila descrizioni, scelto proporzionalmente alle classi di non coincidenza, ammettendo un margine di errore dello ±0.01 per cento.

Tavola 5.7 - Confronto codici assegnati da ACTR e dagli Studi di Settore (Unici al netto dei Multipli e Possibili con lo stesso codice) - Campione per il controllo di qualità

TIPOLOGIA CODICI ASSEGNATI	Valori assoluti	Valori percentuali	Campione per controllo qualità
Codici ACTR - campione coincidenti per le prime 4 cifre	2.306	13,0	520
Codici ACTR - campione coincidenti per le prime 3 cifre	3.042	17,1	686
Codici ACTR - campione coincidenti per le prime 2 cifre	4.185	23,6	943
Codici ACTR - campione coincidenti per la prima cifra	5.040	28,4	1.136
Codici ACTR - campione Non coincidenti	3.173	17,9	715
Totale	**17.746**	**100,0**	**4.000**

Tale campione è stato sottoposto all'analisi di codificatori esperti che hanno dichiarato, apponendo appositi flag, per quali descrizioni fosse corretto il codice di ACTR (A), per quali fosse corretto il codice pre-esistente (C), per quali fossero entrambi errati (E) ed, infine, quali fossero dubbi (D).

I risultati di questa analisi sono riportati nella seguente tavola.

Tavola 5.8 - Analisi qualità campione Cciaa

TIPOLOGIE	Campione per controllo qualità	A		C		E		D	
		n	%	n	%	n	%	n	%
Coincidenti per le prime 4 cifre	520	489	*94*	23	*4*	8	*2*	-	-
Coincidenti per le prime 3 cifre	686	592	*86*	87	*13*	7	*1*	-	-
Coincidenti per le prime 2 cifre	943	921	*98*	8	*1*	1	*0*	13	*1*
Coincidenti per la prima cifre	1.136	351	*31*	119	*10*	666	*59*	-	-
Non coincidenti	715	569	*80*	99	*14*	10	*1*	37	*5*
Totale	**4.000**	**2.921**		**336**		**692**		**50**	

Come può vedersi, anche per questo campione di testi, l'accuratezza è elevata (tra l'80 per cento ed il 94 per cento) per tutte le classi di analisi, a meno di quella relativa ai codici coincidenti solo per la prima cifra. Il fatto che il sistema sia in grado di assicurare livelli di qualità soddisfacenti, nonostante tratti testi non forniti secondo indicazioni precise (come avviene per le risposte ai questionari di indagine), ma fallisca pesantemente soltanto per un sottoinsieme molto ristretto di questi, potrebbe sembrare strano.

Tuttavia, dall'esame puntuale delle descrizioni emerge che questa classe è popolata in modo preponderante da testi molto generici ed afferenti al settore dell'edilizia. La combinazione di questi due fattori ha fatto sì che entrambi i codici assegnati con le due diverse metodologie fossero corretti a livello di prima cifra, ma non lo fossero per i restanti, in quanto la coerenza rispetto al testo avrebbe imposto un codice non al massimo dettaglio, vista la genericità della descrizione. Il motivo per cui in questa casistica si concentrano testi del settore dell'edilizia deriva soprattutto dal fatto che questa classe ha subito una forte ristrutturazione nella classificazione Ateco e questi risultati hanno messo in luce che, evidentemente, ulteriori interventi in tal senso devono essere effettuati nell'ambiente di codifica.

5.3 III test: le indagini speciali

A causa dei profondi cambiamenti intervenuti nella nuova classificazione delle attività economiche Ateco 2007, si è ritenuto opportuno, nel corso del 2007, condurre delle indagini su settori specifici di attività economica.

I settori prescelti sono stati:

- Informazione e Comunicazione – Ict (sezione J);
- Attività degli studi di architettura e d'ingegneria; collaudi e analisi tecniche (divisione 71);
- Ricerca e sviluppo sperimentale nel campo delle scienze naturali e dell'ingegneria - R&S (gruppo 72.1);
- Attività di design specializzate (gruppo 74.1);
- Servizi integrati di gestione agli edifici – Pulizie (divisione 81);
- Altre attività professionali, scientifiche e tecniche n.c.a.; attività di supporto per le funzioni d'ufficio; servizi di supporto alle imprese n.c.a. (gruppi: 74.9, 82.1 e 82.9).

Sono stati inviati 6 questionari distinti ad un campione di imprese che risultavano svolgere le precedenti attività in base alla variabile attività economica presente nel Archivio statistico delle imprese attive Asia. Al fine di ridurre al minimo il disturbo statistico i questionari sono stati progettati nella maniera più semplice possibile.

Tavola 5.9 - Campione indagine per tipologia di settore economico

SETTORE ECONOMICO	Valori assoluti	Valori pericentuali
Ict	15.276	33,0
Ricerca & Sviluppo	2.042	4,4
Ingegneria e architettura	5.053	10,9
Pulizie	7.366	15,9
Design	2.801	6,1
Servizi	13.689	29,6
Totale	**46.227**	**100,0**

Il questionario era articolato in poche sezioni: la prima riguardava lo stato di attività dell'impresa; nella seconda sezione si richiedeva la descrizione dell'attività economica; la terza sezione elencava le singole attività economiche e chiedeva alle imprese di indicare la percentuale di fatturato realizzata in ognuna delle attività svolte, in tal modo sarebbe stato poi possibile attribuire il codice di attività svolto in maniera univoca.

Alla fine del questionario, con l'eccezione del questionario relativo alla Ricerca e Sviluppo, veniva chiesto di descrivere un'altra attività se l'impresa non si riconosceva in una di quelle elencate nel questionario.

Nel Registro delle imprese sono presenti più di 4 milioni e 300 mila imprese; più del 90 per cento di queste sono di piccola o piccolissima dimensione, di conseguenza è impossibile controllarle una ad una o essere certi del codice di attività che viene loro attribuito con delle procedure automatizzate che, a loro volta, elaborano l'informazione proveniente dagli archivi amministrativi. Può accadere che l'impresa inserita nel campione, in ragione dell'attività economica svolta, non svolga effettivamente quell'attività e che sia necessario attribuirgli un'attività diversa da quelle previste.

Le indagini sono state indirizzate ad un campione di piccole imprese (circa 45 mila unità) nei suddetti settori di attività economica; il settore della Ricerca e Sviluppo (gruppo 72.1) è stato coinvolto interamente nell'indagine in quanto non eccessivamente numeroso. Per gli altri settori

ci si è orientati a intervistare tutte le imprese della classe dimensionale tra 10 e 19 addetti e un campione di quelle fino a 9 addetti. Poiché i settori dell'Ict e quelli delle Pulizie erano i più complessi da riclassificare, in quanto molto più dettagliati che nella classificazione precedente e con delle attività completamente nuove, si è deciso di intervistare tutte le imprese della classe dimensionale da 10 a 49 addetti, campionando solo la parte inferiore a 10 addetti.

Il tasso di risposta all'indagine è stato di circa il 30 per cento, come riportato nella seguente tavola.

Tavola 5.10 - Tasso di risposta per i diversi settori

SETTORE ECONOMICO	Valori percentuali
Ict	32,8
Ricerca & Sviluppo	30,6
Ingegneria e architettura	27,7
Pulizie	26,2
Design	26,4
Servizi	30,0
Totale	**29,9**

Il fine principale delle indagini era svolgere una rilevazione su attività economiche completamente nuove o molto più dettagliate rispetto alla precedente versione della classificazione delle attività economiche in modo da poter attribuire un codice Ateco 2007 corretto anche alle imprese di piccola dimensione.

Il secondo fine era raccogliere definizioni di attività da parte delle imprese e arricchire il dizionario ACTR con nuove empiriche. In particolare, si volevano ottenere definizioni di attività nuove, specialmente per il settore Ict e per il settore dei Servizi, per poter essere in grado di classificare le imprese in base alla descrizione della loro attività da ricondurre all'Ateco 2007. I questionari compilati hanno consentito di raccogliere circa 14 mila descrizioni libere delle attività svolte dalle imprese: circa 5 mila su Ict, 600 su Ricerca e Sviluppo, 1.400 su studi di Ingegneria e architettura, circa 2 mila su Pulizie, circa 750 su Design, 4.100 su altri Servizi.

Per poter attribuire un codice Ateco a cinque cifre sono state analizzate le risposte delle imprese alle sezioni del questionario dove veniva richiesto di indicare la percentuale di fatturato realizzata con le diverse attività. Sul totale dei rispondenti, è stato possibile attribuire un codice Ateco a cinque cifre a circa il 52 per cento (flag = 1).

Al fine di realizzare un test di qualità, è stata presa in considerazione solo questa sotto popolazione corrispondente a poco più di 7.000 casi. Per quanto riguarda gli altri questionari compilati, nel 33 per cento dei casi non è stato possibile attribuire un codice Ateco a cinque cifre analizzando le risposte fornite dalle imprese; per il restante 15 per cento ACTR è riuscito ad attribuire un codice in base alle descrizioni libere fornite dalle imprese. Il questionario che ha avuto il maggior tasso di risposta è stato quello del settore Ict (cfr. tab. 5.10); per tale settore la percentuale di risposte attribuite (flag = 1) è stata il 47,6 per cento, il questionario dei servizi ha raggiunto il 20,1 per cento di risposte codificabili; in base all'analisi delle risposte fornite dagli intervistati si sono potuti attribuire pochi codici Ateco a cinque cifre ai questionari relativi al Design e alla Ricerca e Sviluppo.

Tavola 5.11 - Distribuzione imprese per modalità di attribuzione Ateco 2007

MODALITÀ DI ATTRIBUZIONE	Flag	Valori assoluti	Valori percentuali
Non attribuzione Ateco 2007	0	4.568	33,1
Ateco attribuita da risposte a indagine	1	7.135	51,6
Ateco attribuite da ACTR (per i multipli ed i possibili è stata scelta la prima Ateco attribuita da ACTR con punteggio più alto)	2	2.117	15,3
Totale		**13.820**	**100,0**

La sotto popolazione con flag = 1 è stata sottomessa ad ACTR che ha dato un tasso di codifica del 44,5 per cento. Tale risultato è basso, se comparato alle altre performance ottenute con ACTR, ma più che giustificato dal fatto che tratta settori nuovi e definizioni nuove per le quali il dizionario di ACTR era ancora fortemente carente.

Per quanto riguarda il grado di precisione di questo terzo test, sono stati analizzati dai codificatori esperti 1.148 record per verificarne la correttezza. Per misurare il grado di precisione di ACTR si è deciso di sottoporre a verifica una specifica sotto-popolazione; i 7.135 casi con flag = 1 sono stati confrontati con le Ateco a 5 cifre attribuite con ACTR; si è deciso di esaminare i casi in cui il codice attribuito dall'indagine e quello attribuito con ACTR non corrispondeva a livello di V cifra. I risultati dell'analisi dei codificatori sono riportati nella tavola 5.12.

I codificatori hanno stabilito che: nell'88,2 per cento dei casi, il codice attribuito da ACTR era corretto, nel 3 per cento dei casi i codici attribuiti da indagine e da Ateco erano entrambi errati, nel 2,4 per cento dei casi era corretto il codice attribuito dall'indagine e nel 6,4 per cento dei casi il testo era troppo generico oppure elencava più di un'attività.

Tavola 5.12 - Test di precisione sulle risposte codificate

MODALITÀ ATTRIBUZIONE	Valori assoluti	Valori percentuali
Codice Ateco attribuito da ACTR corretto	1.013	88,2
Codici Ateco errati entrambi	34	3,0
Codice Ateco attribuita da risposte a indagine corretto	27	2,4
Testo non codificabile (testo generico o testo con più attività dichiarate)	74	6,4
Totale	**1.148**	**100,0**

Nonostante il tasso di codifica non sia eccessivamente elevato, il tasso di precisione è molto buono; considerate le premesse e il fatto di muoversi su terreni quasi inesplorati si può considerare il tasso di codifica del 44,5 per cento più che soddisfacente. Lo scopo finale delle indagini era proprio raccogliere testi nuovi per arricchire il dizionario e migliorare la qualità di ACTR in termini di performance; in particolare, era importante raccogliere testi scritti dalle imprese per descrivere le attività del settore dei servizi e di tutte quelle attività di cui non si sospetta l'esistenza in un settore che assume un peso sempre maggiore nei paesi sviluppati.

Nell'ambito delle descrizioni raccolte sono stati analizzati soprattutto i falliti e i multipli prodotti da ACTR soprattutto dai questionari relativi all'Ict e agli altri Servizi. L'analisi per creare nuovi testi nel dizionario non è stata ancora completata; al momento sono stati inseriti 600 nuovi testi.

5.4 Aggiornamento dell'ambiente di codifica in funzione dei risultati dei test di qualità

I test descritti nei paragrafi precedenti hanno avuto una doppia valenza: non soltanto quella di verificare le performance delle applicazioni di codifica su diverse tipologie di dati, ma anche quella di analizzarne i risultati per aggiornare le basi informative utilizzate dal sistema ACTR.

Tale aggiornamento è stato di due tipologie: correttive, laddove dai test era emerso un errore nell'attribuzione del codice (appartengono a questa tipologia di interventi la modifica di descrizioni contenute nel dizionario, così come di regole di *parsing*), e di arricchimento del dizionario, tramite l'analisi dei testi non codificati da ACTR in cui si rilevava un contenuto informativo sufficiente per l'individuazione di un codice.

Vista la mole e la complessità di tutti questi interventi sull'ambiente di codifica, si è ritenuto opportuno verificarne gli effetti per essere certi della tenuta e magari dell'innalzamento sia del *recall rate*, che del *precision rate.* A tal fine, è stato effettuato un ulteriore passaggio di codifica di tutti e tre i campioni sull'ambiente così modificato e, per ciascuno di essi, sono stati messi a confronto i risultati ottenuti con quelli del primo passaggio di codifica. Sono state quindi individuate diverse casistiche di risultati, alcune delle quali da sottoporre ad una nuova analisi da parte dei codificatori esperti.

Si fornisce la decodifica delle sigle di seguito utilizzate:

UV → gli Unici prodotti dal primo passaggio di codifica;

UN → gli Unici prodotti dal secondo passaggio di codifica (quello effettuato a seguito degli interventi di correzione ed arricchimento);

- C → Codice corretto.

Le casistiche riguardano rispettivamente l'intersezione degli insiemi ($UV \bigcap UN$) e la non intersezione ($UV \bigcap UN = \varnothing$).

In particolare, nell'ambito dell'intersezione degli insiemi $UV \bigcap UN$:

se codice UV = C, allora:

Casistica 1 → Codice UN = codice UV (insieme da non sottoporre ad ulteriore analisi)

Casistica 2 → Codice UN ≠ codice UV (da sottoporre ad analisi dei codificatori esperti)

se codice UV ≠ C, allora

Casistica 3 → Codice UN = codice UV (da sottoporre ad analisi dei codificatori esperti, perché senz'altro errato)

Casistica 4 → Codice UN ≠ codice UV (da sottoporre ad analisi dei codificatori esperti per verificare se il nuovo codice è corretto)

Relativamente alla non intersezione degli insiemi $UV \bigcap UN = \varnothing$:

Casistica 5 → Se codice UV = C (da analizzare perché il fatto che con il nuovo ambiente di codifica non si abbia più un Unico costituisce un malfunzionamento della nuova applicazione);

Casistica 6 → Se codice UV ≠ C (da analizzare per verificare se è corretto che con il nuovo ambiente non si abbia più un Unico oppure se ci si sarebbe attesi un nuovo Unico con un codice diverso dal vecchio);

Casistica 7 → UN (devono essere comunque analizzati per verificare la correttezza del nuovo codice assegnato).

Nonostante le molteplici casistiche, la numerosità dei testi da analizzare non si è dimostrata così rilevante, per cui è stato possibile procedere all'analisi da parte dei codificatori esperti e quindi all'ulteriore messa a punto dell'applicazione di codifica.

Capitolo 6 - Procedura per il trattamento di testi lunghi e ridondanti (archivio Cciaa)

Come già accennato nei capitoli precedenti, si è voluta testare la possibilità di aggiornare il registro Istat delle imprese codificando, sulla base della classificazione Ateco 2007, l'informazione proveniente da archivi esterni all'Istituto. La particolarità di questi archivi, come descritto in precedenza, sta nel fatto che l'informazione in essi contenuta viene raccolta secondo regole e principi che non rispecchiano quelli adottati dall'Istat per la rilevazione dell'informazione qualitativa contenuta nelle domande a testo libero riportate nei questionari di indagine.

In particolare, l'archivio utilizzato è stato quello delle Camere di commercio (Cciaa) contenente la descrizione dell'attività economica svolta da ogni singola impresa. I testi delle Cciaa non potevano però essere sottoposti direttamente ad ACTR per la codifica automatica in quanto:

- erano troppo lunghi in assoluto e in relazione a quelli raccolti in indagini statistiche; inoltre, oltre il 50 per cento di testi eccedeva i 200 byte, ossia la massima lunghezza gestibile da ACTR;
- contenevano un numero elevato di errori di ortografia;
- contenevano informazioni ridondanti e prive di significato ai fini della codifica, ossia dell'attribuzione di un codice di attività economica.

La ridondanza e la non significatività (ai fini della codifica automatica) dei testi delle Cciaa deriva dalla non adozione di regole e principi per la raccolta dell'informazione testuale, come sopra accennato. L'attribuzione "corretta" di un codice ad un testo, sia essa in modalità manuale che automatizzata, richiede infatti che il testo raccolto sia *i)* sintetico, *ii)* univocamente interpretabile e *iii)* ricco di significato. Queste tre caratteristiche o regole vengono osservate durante la raccolta di risposte a domande aperte istruendo il rilevatore, attraverso apposita formazione, o inserendo istruzioni e regole per la compilazione delle domande nel caso di questionari auto-compilati. Poiché la raccolta dell'informazione testuale da parte delle Cciaa non ha l'obiettivo della codifica o dell'interpretazione statistica dell'informazione, va da sé che i testi raccolti non possano essere sottoposti a processo di codifica se non previo trattamento.

È stato quindi pensato di realizzare una procedura che, prima di sottoporre i testi a codifica, ne effettuasse un trattamento attraverso qualche metodo di analisi testuale che fosse in grado di estrarre dal testo l'informazione saliente rendendola, in termini di struttura e contenuto, tale da poter essere sottoposta a codifica automatica. Ovviamente tutto questo senza alterare il significato del testo originario.

La procedura realizzata si articola in due step logici: nel primo step avviene il trattamento dei testi al fine di semplificarli attraverso l'individuazione e successiva eliminazione delle parti ridondanti o assolutamente non significative per l'attribuzione di un codice; nel secondo step il testo pre-trattato viene sottoposto a codifica automatica.

Per la realizzazione delle due fasi della procedura sono stati utilizzati i software TaLTaC^2, Sas e Visual Basic per il primo step e ACTR per il secondo.

Nel seguito di questo capitolo vengono descritti i passi della procedura ed i risultati ottenuti nel secondo step di codifica attraverso i due usuali indicatori, ossia il *recall* and il *precision rate*. Prima di questo è però necessario descrivere il software TaLTaC^2 che gioca un ruolo fondamentale nel primo passo della procedura.

6.1 Il software TaLTaC^2

TaLTaC^2 è un software per l'analisi automatica dei testi che lavora nella duplice logica di *Text Analysis* (TA) e di *Text Mining* (TM): la prima è da considerarsi la vera e propria analisi automatizzata dei testi, la seconda è un'applicazione specifica della prima, in particolare si riferisce a tutte le tecnologie informatiche necessarie a gestire una grossa mole di dati e di informazioni.

È opportuno dare alcune indicazioni terminologiche condivise per il prosieguo del paragrafo, una sorta di glossario di riferimento.

Si definisce un corpus di testi una qualsiasi raccolta di dati testuali fra loro confrontabili (documenti, verbali, domande aperte, resoconti di focus group, interviste qualitative, raccolta di lettere o canzoni eccetera). Una volta definita la tipologia di testi da analizzare (documenti scritti, resoconti di interviste, forum telematici), è necessario curarne l'organizzazione interna e la trascrizione, prestando attenzione ad alcuni requisiti, determinanti per la significatività dei risultati. Un requisito fondamentale è che i testi di cui si compone ogni corpus siano comparabili tra loro, per struttura, dimensioni, autore o destinatari.

L'unità elementare che costituisce il corpus è la parola detta anche forma grafica.

Le forme grafiche corrispondono all'insieme di caratteri compresi tra due spazi; generalmente coincidono con le parole che compongono il corpus, ma possono anche rappresentare ad esempio, l'unificazione di un insieme di forme dotato di significato specifico che, come si vedrà in seguito, prendono il nome di segmenti ripetuti.

Le occorrenze, che sono la prima misura statistica che viene operata all'interno del corpus, non sono altro che un conteggio del numero di volte che una singola parola (forma grafica) compare nel testo.

Il vocabolario è l'insieme di parole diverse che vanno a costituire il corpus in analisi. Può essere espresso a livello di forme grafiche (le parole così come compaiono nel corpus), o per lemmi, cioè le forme presenti nei dizionari.

La dimensione di un corpus non è altro che il numero totale di occorrenze rispetto alle quali è possibile definire tre tipologie di corpus: piccolo quando il numero totale di occorrenze è <=15 mila; medio quando è compreso tra 15 mila e 45 mila; medio-grande quando è compreso tra le 45 mila e 100 mila o addirittura supera le 100 mila occorrenze. Non ci sono difficoltà dal punto di vista dell'ampiezza dei documenti da analizzare, infatti la dimensione del file testuale può essere vastissima (milioni di parole) come nel caso oggetto di studio.

Attraverso l'uso di questo software è possibile effettuare analisi quantitative su almeno due livelli: uno in cui l'unità di analisi è la singola parola; l'altro in cui l'unità di analisi sono i frammenti, detti anche unità di contesto.

I frammenti non sono altro che testi corti (abstract, risposte a domande aperte di un questionario, bibliografie o, come nel nostro caso, la definizione estesa di attività economica che ogni impresa fornisce alle Camere di commercio) raccolti in una collezione di documenti che vanno a costituire il corpus di dati testuali finale da analizzare.

L'ultima definizione che è necessario fornire è quella di segmento ripetuto, si tratta di sequenze di parole nel testo che rimandano a un significato più preciso rispetto alla semplice parola (ad esempio le singole parole "ragione" e "sociale" hanno un significato particolare quando vengono studiate sequenzialmente, formando così il segmento "ragione_sociale").

Il software nasce dai risultati di ricerche svolte presso le Università degli Studi di Salerno e di Roma "La Sapienza" nel corso degli anni novanta, coordinate da Sergio Bolasco.[7]

Un percorso tipico di analisi automatizzata del testo segue almeno quattro step: a) il primo e più delicato è proprio quello della predisposizione del testo; b) il secondo è quello dell'analisi lessicale; c) il terzo è quello dell'estrazione dell'informazione; d) l'ultimo è quello dell'analisi testuale. Ciascuno di questi punti della filiera conoscitiva ha al suo interno diverse procedure ma non è questa la sede più indicata per approfondirle. Molto sinteticamente, la prima fase, quella di predisposizione, consta fondamentalmente di una corretta lettura e scansione o acquisizione, su diversi supporti informatici, del testo da studiare in funzione dell'unità di analisi scelta. Come si vedrà poco più avanti, questa prima fase viene assistita dal software attraverso una procedura che prende il nome di normalizzazione, cui è stato sottoposto anche il corpus delle Camere di commercio oggetto di studio. La seconda fase, quella cioè dell'analisi lessicale, può essere definita un'analisi verticale; essa fornisce una rappresentazione del corpus dovuta allo studio del suo vocabolario ossia del suo linguaggio. L'analisi ruota dunque intorno allo studio delle singole parole per come si presentano all'interno del testo, in funzione delle loro occorrenze, della loro natura sintattica eccetera. Una fase di centrale importanza è quella dell'estrazione dell'informazione statistica che caratterizza diversi momenti dell'analisi e si avvale di molteplici funzioni proprie del software, tra le più rilevanti vale la pena ricordare l'estrazione di un linguaggio peculiare in base al calcolo di un preciso test statistico (Tfidft),[8] il calcolo dello scarto standardizzato per il confronto con lessici esterni, il calcolo di sequenze di parole caratteristiche in base ad un indice relativo come nel caso dei segmenti ripetuti. Infine vi è l'analisi testuale che riguarda tutte le operazioni rivolte direttamente al corpus, in grado cioè di fornire una rappresentazione puntuale del testo, ad esempio per concetti, per dimensioni del discorso o per categorie o per contesti d'uso (detti concordanze).

Rispetto alle fasi sopra citate i maggiori punti di forza di TaLTaC^2 sono, in particolare, la procedura di normalizzazione; il calcolo e l'estrazione dei segmenti ripetuti significativi; l'effettuazione del confronto con risorse linguistiche esterne; se necessario l'attribuzione di un tag grammaticale (un'etichettatura grammaticale); l'operare una ricostruzione del testo; infine l'esecuzione di procedure complesse di *text mining.*

Di seguito sono descritte alcune delle citate procedure implementate in TaLTaC^2 che sono, peraltro, state utilizzate per il trattamento dell'archivio Cciaa.

Per fare chiarezza, è necessario prima di tutto definire il testo come una successione di caratteri distinti in separatori e caratteri alfabeto; le unità d'analisi (forme grafiche) sono quindi definite come successioni di caratteri dell'alfabeto comprese tra separatori (o spazi).

La normalizzazione, agisce sull'insieme dei caratteri alfabeto eliminando possibili fonti di sdoppiamento del dato. Ad esempio, abbassando le maiuscole non rilevanti (la, La), oppure uniformando la grafia dei nomi propri, delle sigle, dei numeri e delle date che presentano una forte variabilità. Questa procedura si avvale di: a) algoritmi, b) liste, c) parametri opzionali. La riduzione degli spazi multipli e la trasformazione dei doppi apici in virgolette sono le uniche operazioni obbligatorie (non deselezionabili) in quanto fondamentali per il corretto funzionamento delle procedure successive. La fase basata sugli algoritmi permette:

[7] Ordinario di Statistica presso il Dipartimento di studi geoeconomici, linguistici, statistici e storici per l'analisi regionale della Sapienza (il lavoro è frutto della collaborazione di ricercatori e colleghi di varie università italiane e francesi).

[8] L'indice Tfidft nasce nell'ambito della disciplina dell'Information Retrieval come strumento per ordinare i risultati di una ricerca in base alla pertinenza tra i documenti estratti e le *keywords* di ricerca. L'indice rappresenta un peso attribuito ad ogni parola sulla base della sua occorrenza e della sua distribuzione all'interno della collezione dei documenti, ed è questo peso che viene preso in considerazione per effettuare l'ordinamento dei risultati. Questo test viene formalizzato nel seguente modo $w = tf*logN/n$ dove tf è la frequenza del termine in ciascun documento, n il numero di documenti contenenti quel termine e N il numero totale dei documenti del corpus. Questo indice pondera le parole in funzione della loro rilevanza.

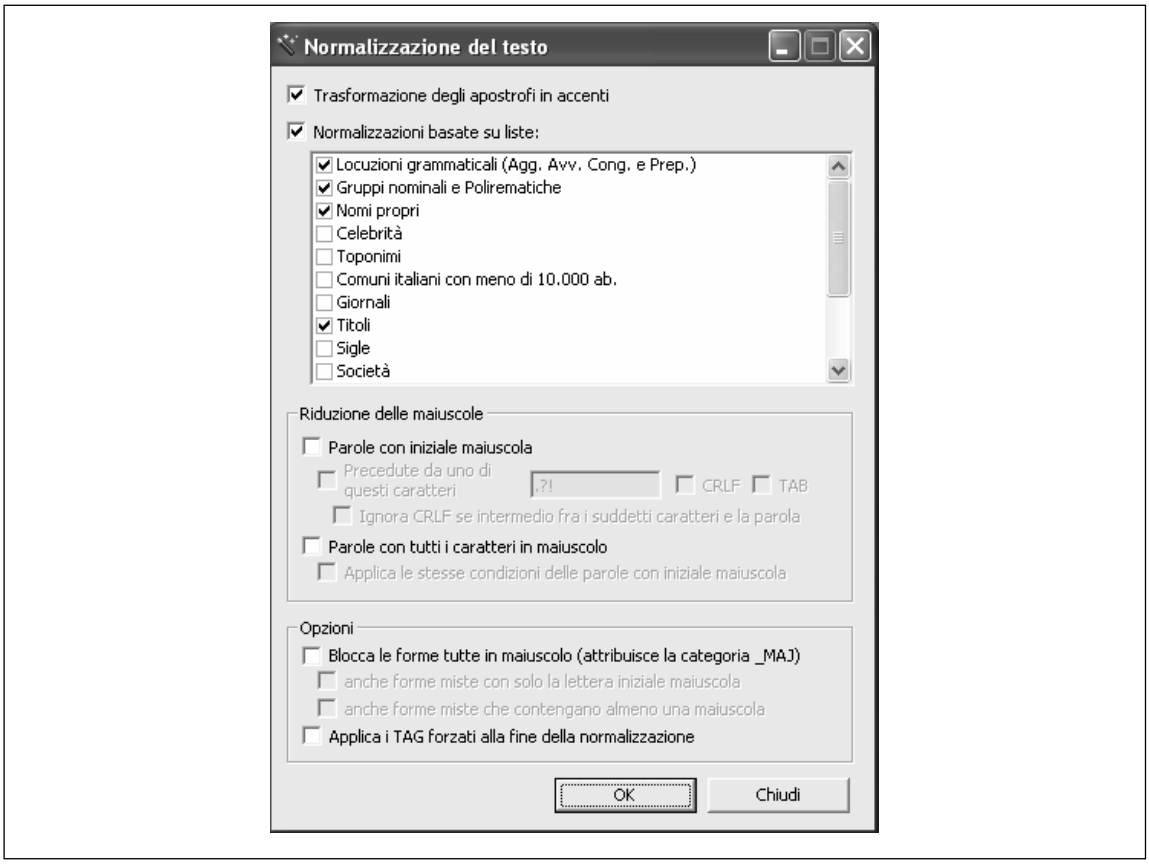
- la riduzione degli spazi multipli e la trasformazione dei doppi apici in virgolette;
- l'aggiunta dello spazio dopo l'apostrofo e la trasformazione degli apostrofi in accenti;
- la normalizzazione delle percentuali;
- la normalizzazione delle date;
- la normalizzazione dei numeri.

La fase basata su liste permette di etichettare parole e/o sequenze di parole, la cui specificità andrebbe perduta, con l'obiettivo di:

- Etichettare univocamente le forme e le sequenze ambigue utilizzando le differenze Maiuscolo/minuscolo. Nomi propri: *Rosa* può essere sostantivo e aggettivo se non si tiene conto della maiuscola;
- Uniformare la grafia delle forme presenti nel corpus (toponimi o celebrità scritti in minuscolo: *venezia*, *andreotti* eccetera);
- Riduzione delle maiuscole.

Di seguito la schermata di TaLTaC^2 per la scelta dei parametri per la normalizzazione del testo.

Figura 6.1 - Normalizzazione del testo

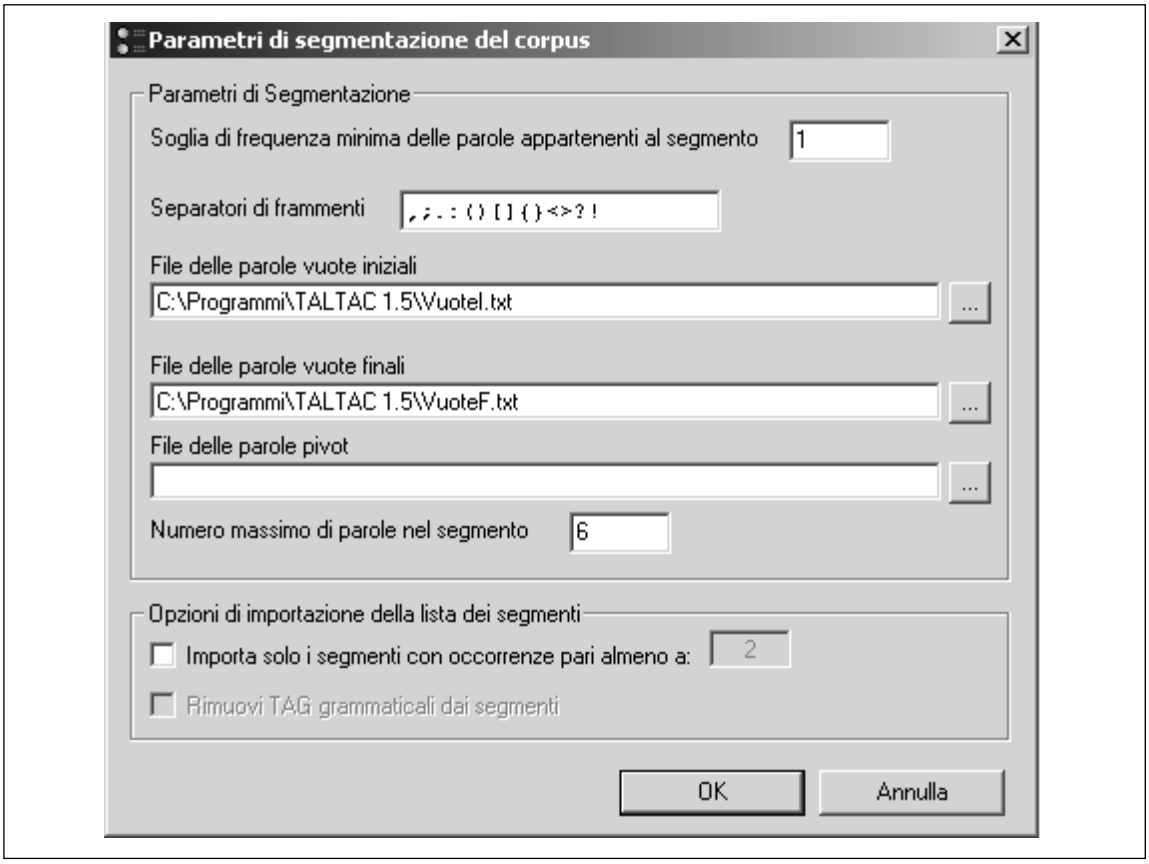


La seconda procedura che è stata eseguita con TaLTaC^2 è quella del calcolo e dell'estrazione dei segmenti ripetuti. Sulla definizione di segmento ripetuto si rimanda al paragrafo 6.2; qui ci si limita ad illustrare come vengono estratti i segmenti e quali test statistici sono calcolati per studiarne la rappresentatività all'interno di un corpus. In TaLTaC^2 il numero di segmenti individuati dalla procedura di selezione automatica dipende da diversi parametri: 1) la

soglia di frequenza minima (numero di occorrenze) delle parole appartenenti al segmento; 2) il numero massimo delle parole che compongono il segmento; 3) la soglia di frequenza dei segmenti che si desidera importare.

Di seguito la schermata di TaLTaC[2] per la selezione dei parametri per la segmentazione del testo.

Figura 6.2 - Segmentazione del testo



Successivamente all'individuazione dei segmenti ripetuti è possibile calcolare un indice (nel seguito indice IS) con cui valutare la loro rilevanza nel corpus.

Dove:

$$IS = \left[\sum_{i=1}^{l} \frac{f_{segm}}{f_{fg_i}} \right] * P$$

l = lunghezza segmento in parole
f_{segm} = frequenza segmento
f_{fgi} = frequenza forma grafica i nel testo
P = numero parole piene presenti nel segmento

Tale indice rappresenta il grado di assorbimento del segmento rispetto alle parole che lo costituiscono. Ad esempio, se il segmento <base_contrattuale> ha occorrenza 9 e le parole <base> e <contrattuale> hanno rispettivamente occorrenza 10 e 9 possiamo dire che il segmento in questione assorbe il 90 per cento delle occorrenze della parola <base> e il 100 per cento delle occorrenze della parola <contrattuale>, pertanto l'indice vale 0,9.

Questa informazione consente di concludere che è poco informativo trattare le due forme separatamente, perché nel corpus si parla quasi esclusivamente di <base contrattuale> e non ad

esempio di <base navale> o di <rapporto contrattuale>. Ovviamente, un segmento sarà da considerarsi tanto più rilevante quanto più forte sarà il suo grado di assorbimento delle parole componenti.

Il calcolo e l'estrazione dei segmenti ripetuti nell'analisi delle attività economiche dichiarate dalle imprese Cciaa è risultato centrale e verrà descritto più approfonditamente nel paragrafo che segue.

Un'altra procedura di TaLTaC^2 che è stata sfruttata nel suo utilizzo per il processo integrato finalizzato alla riclassificazione dei codici di attività economica è rappresentata dal confronto con un lessico di frequenza. Si tratta in particolare di risorse, sia di tipo statistico sia di tipo linguistico, integrate fra loro e personalizzabili dall'utente; esse consentono livelli di analisi differenti: da un lato si parla di analisi del testo e di analisi lessicale e dall'altro di recupero ed estrazione d'informazione, secondo i principi del *data mining* e del *text mining*.

Queste tecniche offrono molto in termini di analisi dei dati testuali; è bene precisare però che le procedure della statistica testuale non si limitano semplicemente a "contare le parole", anche se, per corpus di vaste dimensioni, la semplice distribuzione delle parole più ricorrenti o più frequenti è spesso di per sé altamente rilevante. Le strategie di analisi danno la possibilità di "navigare" nel testo per approfondirne i contenuti; è possibile applicare tecniche di analisi statistica multivariata proiettando le parole su di un piano fattoriale, arrivando infine alla determinazione di profili lessicali specifici, grazie al confronto tra alcune parti e la totalità del corpus.

Il confronto con un lessico di frequenza consente di estrarre l'informazione peculiare del corpus, utilizzando le risorse statistico-linguistiche esogene di TaLTaC^2.

Questa funzione permette di mettere a confronto un campo di una lista qualsiasi con il campo omologo di un lessico o lista di frequenza, generando una tavola in cui figureranno, a seconda del tipo di confronto selezionato (per criteri di uguaglianza o di differenza), i soli record comuni, i soli record originali di una delle due liste, tutti i record di entrambe le liste, comuni ed originali.

Il primo tipo di confronto agisce sulla frequenza relativa con cui le parole compaiono nel lessico di frequenza e nel testo in analisi e fornisce una misura di significatività (scarto standardizzato),[9] che indica la misura della sovra o sotto-rappresentazione della forma nel testo. Naturalmente tanto più lo scarto ha un valore elevato tanto più la forma può essere considerata peculiare e quindi caratterizzante il testo (cfr. Bolasco, 1999, pag. 223).

Il calcolo dello scarto standardizzato consente di individuare le forme con i maggiori scarti d'uso in valore assoluto, vale a dire le parole chiave del testo che, risultando rappresentate in misura significativa rispetto a quanto lo sono generalmente nel linguaggio di riferimento, sono da considerarsi come le più significative del corpus in esame.

Sono statisticamente significative le forme il cui scarto presenta un valore maggiore di 3,84, valore del χ^2 con 1 grado di libertà e p-value=0,05.

Si può sintetizzare che TaLTaC^2 è composto da un insieme di strumenti che consentono lo studio di qualsiasi tipo di dati di natura testuale, raccolti in forma di collezione di testi come un unico corpus, utilizzando le tecniche della "statistica testuale". Questo approccio consente di studiare informazioni non strutturate presenti in una base documentale di ampie dimensioni

[9] Confrontando la frequenza relativa con cui la parola compare nel nostro vocabolario e nel lessico di frequenza utilizzato è possibile individuare le parole sovra-rappresentate nel nostro testo, attraverso il calcolo dello scarto standardizzato: $z_i = \frac{f_i - f_i^*}{\sqrt{f_i^*}}$

(centinaia o migliaia di pagine, o file anche di 100 MB), unitamente a informazioni strutturate (variabili quantitative o qualitative) contenute in un database ad essa associato.

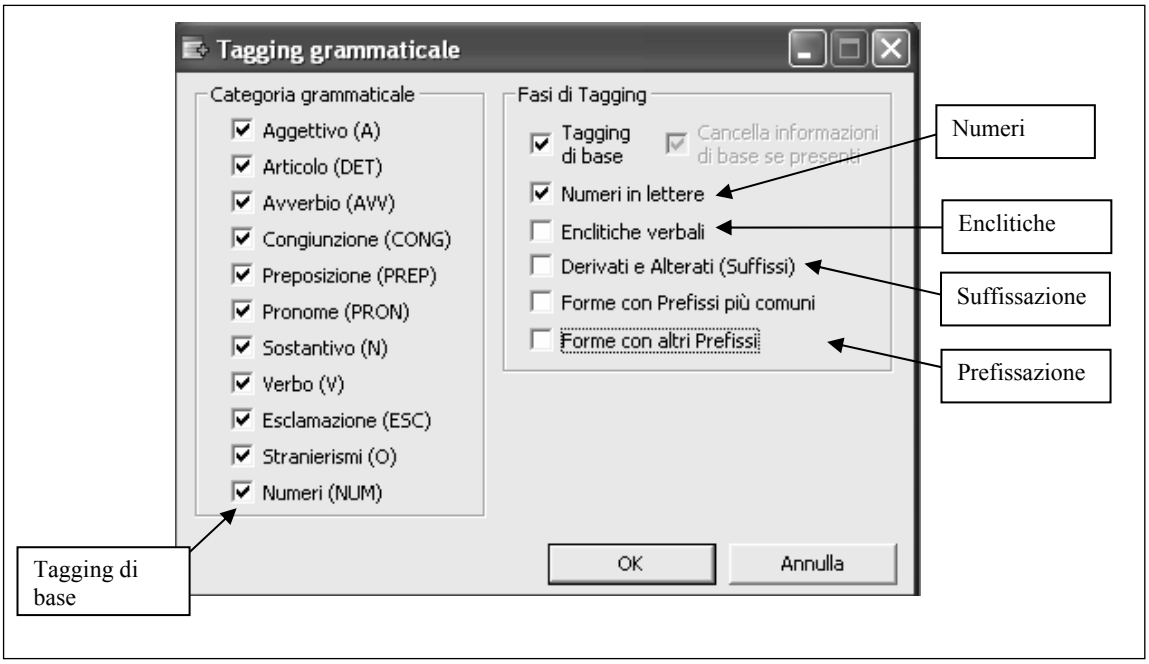
Un'altra ricchezza del programma, che però non è servita per la riclassificazione delle attività economiche delle imprese, è il riconoscimento grammaticale delle forme grafiche.

Attraverso il Tagging grammaticale si confronta il proprio vocabolario con il DizTaLTaC^2 (che è una risorsa statistica interna), costituito da circa 74 mila lemmi, pari a oltre 530 mila forme flesse, per etichettare grammaticalmente le forme grafiche presenti nel vocabolario considerato.

La procedura di Tagging non effettua un riconoscimento completo delle forme del corpus perché per questo scopo sarebbe necessario eliminare le ambiguità semantiche da tutte le forme grafiche presenti in un corpus (Stato con la maiuscola è la nazione, stato con la minuscola è il participio passato del verbo essere, ci si trova dunque di fronte ad un'ambiguità nella scelta del tag tra un Nome o un Verbo), ma si limita ad attribuire alle forme non ambigue la categoria e il lemma.

Di seguito la schermata di TaLTaC^2 per il Tagging grammaticale.

Figura 6.3 - Tagging grammaticale



Nel caso di ambiguità il programma visualizza le categorie grammaticali possibili: sarà il ricercatore, se necessario, ad attribuire manualmente la categoria più opportuna.

Un altro aspetto importante è che TaLTaC^2 è predisposto sia nell'input che nell'output per l'utilizzo di altri software di *text analysis* e *text mining*.

In generale, l'analisi svolta in TaLTaC^2 permette di selezionare ed estrarre l'informazione più significativa dal corpus di testi analizzato (linguaggio peculiare, linguaggio rilevante, linguaggio specifico) e di operare secondo i principi del *text mining* mediante ricerche per parole chiave o per concetti.

I risultati ottenuti in TaLTaC^2 possono interagire direttamente con altri software linguistici (Tree Tagger, Nooj-Intex, Lexical Studio) e statistici (Spad, Spss, Sas).

Ciascun documento del corpus può essere diviso in sezioni, sulle quali operare separatamente analisi di recupero ed estrazione dell'informazione. A ciascun documento si possono associare numerose informazioni strutturate, da mettere in relazione con le informazioni testuali.

TaLTaC^2 produce in output vari tipi di matrici:

- frammenti per forme (documenti per parole); in questa matrice possono essere associate anche le variabili strutturate disponibili a priori e le variabili ricavate a posteriori dall'analisi testuale o dal *text mining* in TaLTaC^2;
- forme per testi (parole per parti), contenente i profili di frequenza lessicali secondo le partizioni prescelte nonché le annotazioni relative alle varie unità selezionate (linguaggio peculiare, rilevante, specifico), oltre alle annotazioni grammaticali e semantiche;
- co-occorrenze (parole per parole), contenente il numero di volte in cui due parole si associano, all'interno di un intervallo predefinito di testo.

6.2 Prima fase procedura di trattamento e codifica testi: identificazione e cancellazione delle informazioni ridondanti

Da un'analisi visiva di alcune descrizioni delle Cciaa era emerso che le parti ridondanti del testo erano singole parole - aggettivi, verbi o nomi non inerenti il settore di attività economica - oppure sequenze di parole. Ad esempio, descrizioni tipiche delle Cciaa erano: "la società è stata fondata dal 1995 con lo scopo di produrre scarpe" oppure "la società è un'associazione non-profit con l'obbiettivo di produrre e vendere giocattoli fatti a mano". In questi testi molto lunghi, l'unica informazione necessaria all'attribuzione di un codice Ateco è "produrre scarpe" per la prima e "produrre e vendere giocattoli fatti a mano" per la seconda. A seguito di questo, si è pensato di dovere creare sia una lista di parole che una di sequenze di parole da dover eliminare dal file delle Cciaa. Per l'identificazione di queste parti di testo è stato utilizzato il software TaLTaC^2 mentre per la successiva cancellazione è stata sviluppata un procedura in Visual Basic.

Il file delle Cciaa e il dizionario dell'Ateco sono stati importati in ambiente TaLTaC^2 divenendo i due corpus di lavoro e sono stati sottoposti alla fase di normalizzazione che consente di pre-trattare il testo, eliminando le possibili fonti di sdoppiamento del dato (ad esempio: abbassare le maiuscole non rilevanti in modo che tutte le parole del vocabolario si presentano con lettere minuscole). Dopo la normalizzazione, il corpus o vocabolario (insieme di forme grafiche) delle Cciaa contava 35.892 forme grafiche e quello dell'Ateco 10.667.

Successivamente, da ognuno dei due corpus, sono stati estratti i "segmenti ripetuti" che, nel caso oggetto di studio, rappresentano la sequenze di parole ridondanti. In ambiente TaLTaC^2 questi sono definiti come "una sequenza di forme grafiche di lunghezza n compresa tra due separatori forti che definiscono i limiti dello spezzone di testo nell'ambito del quale vengono estratti i segmenti".

Per l'individuazione dei segmenti occorre fornire al software alcuni parametri. Quelli utilizzati per questa applicazione sono stati:

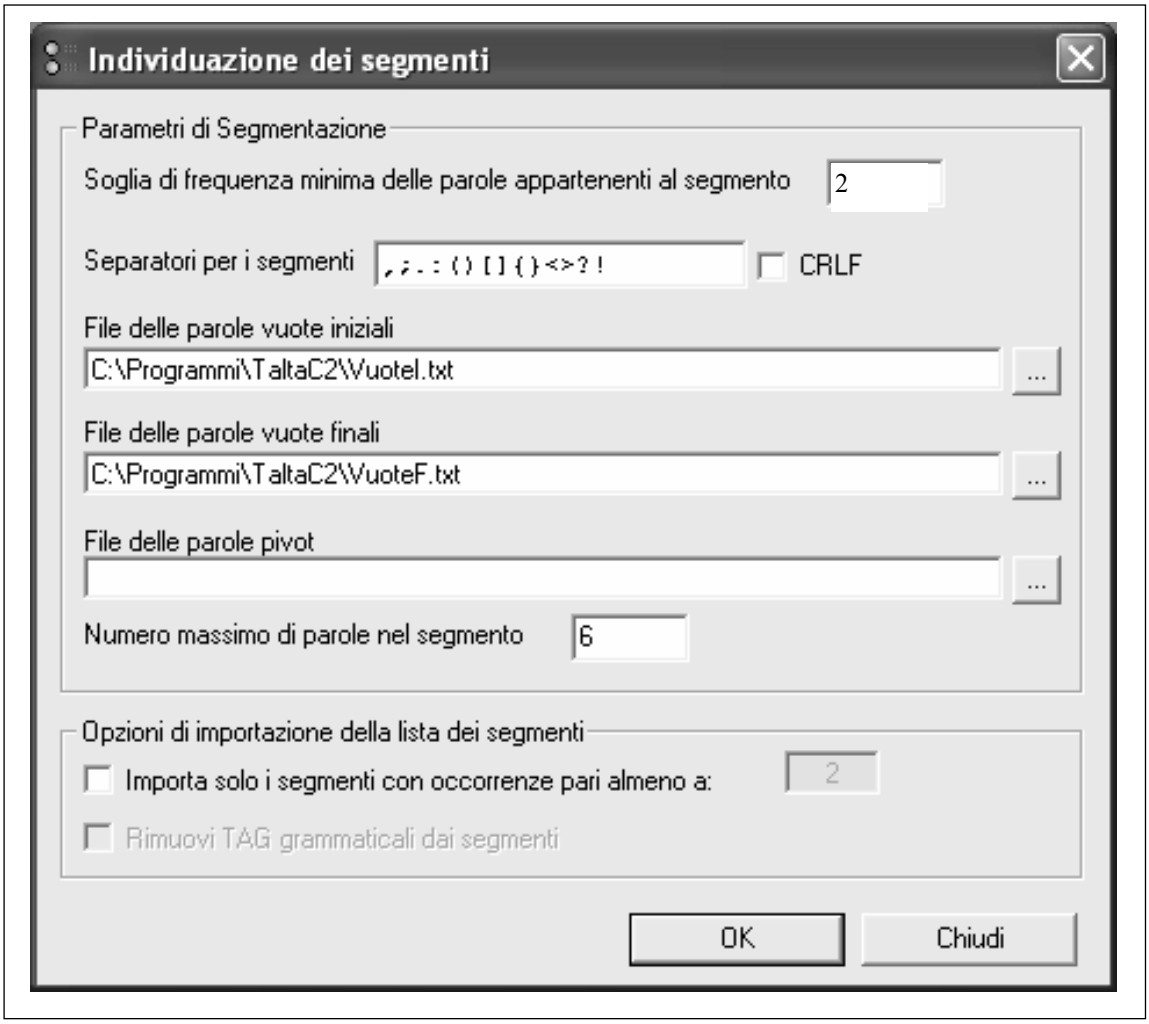
- il numero massimo di parole nel segmento o, equivalentemente, il numero di forme grafiche che il segmento contiene;
- la soglia di frequenza minima delle parole appartenenti al segmento ossia la frequenza di ciascuna forma grafica nel corpus.

Dopo alcune prove empiriche, i valori da assegnare ai parametri sono stati rispettivamente 6 e 2. Quest'ultimo valore è stato scelto appositamente molto basso per poter individuare la

maggior varietà di segmenti ridondanti da cancellare dal file delle Cciaa al fine di rendere i testi il più corti possibile e quindi utilizzabili dal sistema di codifica automatica.

Di seguito la finestra di TaLTaC^2 per l'individuazione dei segmenti ripetuti.

Figura 6.4 - Individuazione dei segmenti



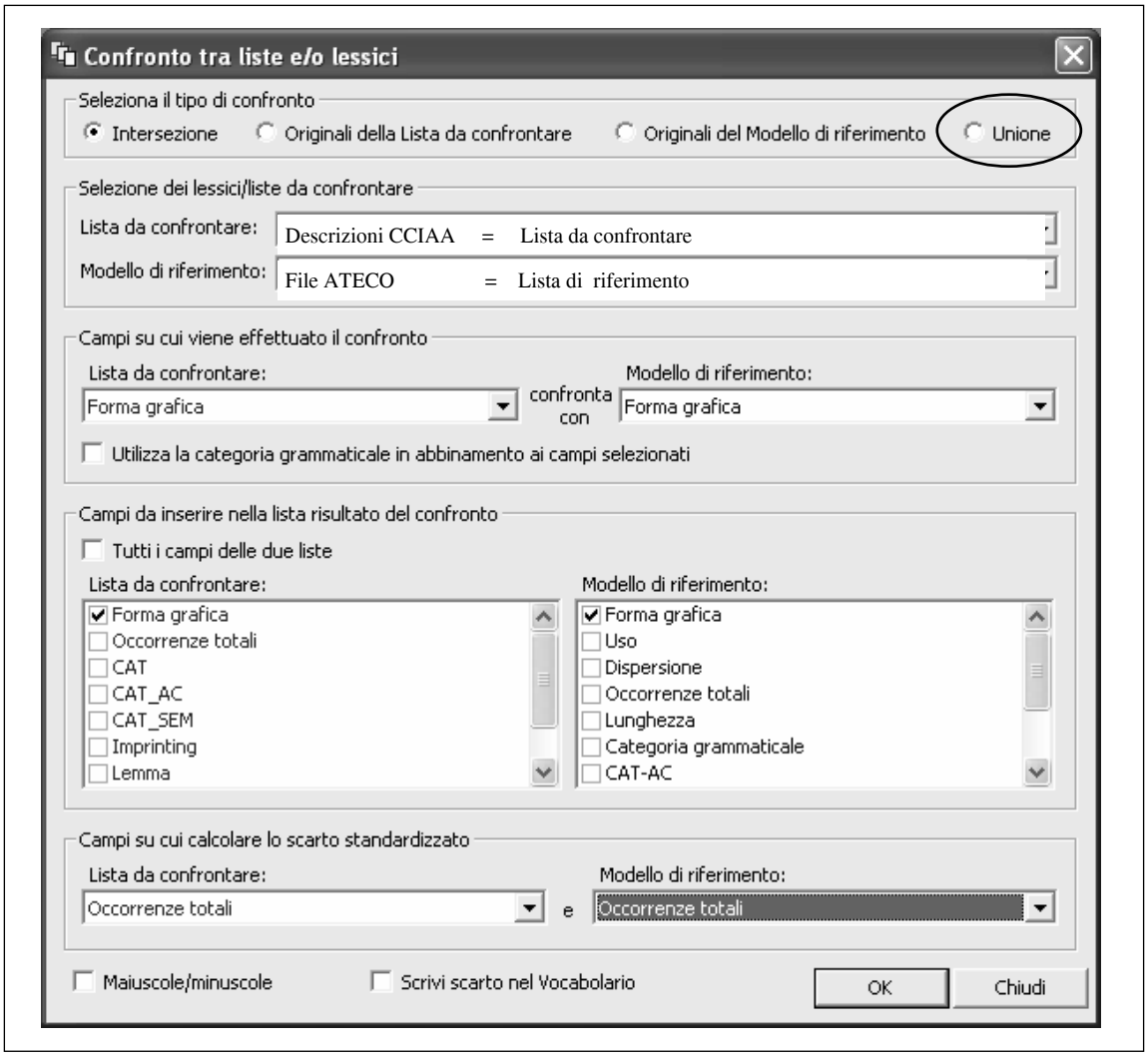
Dal file delle Cciaa sono stati estratti 36.401 segmenti e dal quello Ateco 6.245. Le due liste di segmenti sono state confrontate utilizzando un'apposita funzione di TaLTaC^2 ossia "l'unione tra liste" che effettua il confronto tra stringhe secondo un criterio di uguaglianza. Il file Ateco e quindi la relativa lista è stato considerato il corpus riferimento, ossia contenente l'intero set di testi rilevanti per il settore di attività economica. Tra i segmenti risultati diversi, quelli presenti solo nell'archivio delle Cciaa (e che avevano una frequenza superiore a 50), sono stati considerati tipici di questa fonte di informazione e quindi cancellati: in totale sono stati eliminati 2.108 segmenti ripetuti dal file delle Cciaa.

Il passo successivo è stato quello di individuare le parole ripetute. Con una procedura del tutto analoga a quella appena descritta, sono state estratte le parole (forme grafiche) dai due

archivi. Le liste sono state messe a confronto, sempre con la funzione "unione tra liste" di TaLTaC^2, da cui sono emerse 6.661 parole comuni e circa 22 mila parole diverse. Per queste ultime è stata necessaria un'analisi puntuale (in funzione del numero di occorrenze con cui comparivano nel testo) per non includere nell'insieme delle parole da cancellare quelle non eliminabili in quanto sinonimi o declinazioni di parole del dizionario Ateco, oppure perché affette da errori ortografici che le rendevano fittiziamente diverse da quelle del file Ateco. La lista finale di 19.142 parole ridondanti è stata cancellata dal file delle Cciaa sempre utilizzando un programma sviluppato in Visual Basic.

Di seguito è riportata la finestra di TaLTaC^2 per il confronto tra liste. In particolare questa fa riferimento al confronto fra parole come si evince dal riquadro Lista da confrontare dove è biffata la casella relativa a Forma grafica.

Figura 6.5 - Confronto tra liste



Un importante chiarimento. Nella finestra compare anche la possibilità di utilizzare "l'intersezione" come tipologia di confronto. Questa funzione, che sarebbe stata la scelta

ottimale ai fini dell'applicazione, è stata però introdotta nel software successivamente a questo lavoro e quindi si è dovuto utilizzare la funzione di unione, disponibile al momento.

6.3 Seconda fase procedura di trattamento e codifica testi: codifica dei testi trattati

Dopo la fase di cancellazione, i testi sono stati sottoposti a codifica automatica attraverso il sistema ACTR. In particolare, cancellazione dei testi ridondanti e codifica sono avvenute attraverso più fasi consecutive. Infatti, da alcune sperimentazioni, si era notato che la cancellazione delle singole parole comprometteva a volte la leggibilità dei testi di input in quanto ne alterava la struttura grammaticale, mentre questo non accadeva per la cancellazione delle sequenze di parole (i segmenti ripetuti). Si è pensato allora di procedere alla cancellazione e codifica dei testi secondo questi due passi:

- nel primo passaggio sono stati cancellati da tutti i testi i segmenti ripetuti; i testi così trattati sono stati sottoposti a codifica automatica;
- successivamente, solo dalle descrizioni non codificate nel primo passaggio sono state cancellate le parole ridondanti e i testi così elaborati sottoposti a codifica.

Inoltre, l'archivio della Cciaa è stato suddiviso in due sottogruppi, uno contenente i testi lunghi fino a 200 byte e l'altro con testi di lunghezza superiore a questo valore che rappresenta la lunghezza massima gestibile da ACTR.

La procedura di cancellazione e codifica ha seguito due iter diversi per i due file:

- Per quello con testi più corti di 200 byte si è provveduto a codificare direttamente e, sui non codificati, a cancellare i segmenti. Dato l'ottimo risultato della codifica, come vedremo in seguito, non è stata effettuata la successiva cancellazione delle parole perché il guadagno in termini di record codificati (*recall rate*) non sarebbe stato bilanciato dalla perdita in chiarezza delle descrizioni, che è un elemento fondamentale per l'analisi di qualità dei codici assegnati.
- Per l'archivio con testi di lunghezza superiore ai 200 byte, si è invece proceduto alla cancellazione sia dei segmenti che delle parole ripetute, secondo la sequenza prima indicata, ma previa suddivisione del testo in quattro parti in funzione del punto o del punto e virgola, per avere più probabilità di trattare singole frasi di senso compiuto. Questo approccio implicava la possibilità di assegnare più di un codice Ateco ad una stessa azienda, lasciando a successivi studi l'individuazione e l'attribuzione di un unico codice relativo all'attività prevalente. Del resto l'attribuzione di codici multipli era connaturale alle descrizioni delle Cciaa che, essendo relative all'oggetto sociale delle società, contenevano effettivamente la descrizione relativa a più settori economici nell'ambito dei quali l'azienda avrebbe potuto svolgere la propria attività. Prima, comunque, di implementare la procedura sull'universo delle descrizioni delle Cciaa è stato effettuato un test su un campione di 60 mila record (metà più lunghi di 200 byte e metà più corti) per testare la bontà della procedura sia relativamente alla sequenza dei passi di cancellazione e codifica dei testi, sia in merito alla suddivisone in più parti delle descrizioni più lunghe di 200 byte.

La tavola che segue riporta i risultati del test della procedura, mostrando i tassi di codifica prima e dopo la cancellazione dei testi.

Tavola 6.1 - *Recall rate* prima e dopo la cancellazione dei segmenti ripetuti

TEST SU CAMPIONE DI 60.000 RECORD	Numero di aziende codificate (Valori assoluti)	*Recall rate* (Valori percentuali)
Campione di descrizioni più corte di 200 byte		
Unici prima della cancellazione dei segmenti ripetuti	12.486	41,6
Unici dopo la cancellazione dei segmenti ripetuti	18.772	62,4
Campione di descrizioni più lunghe di 200 byte		
Unici prima della cancellazione dei segmenti ripetuti	-	-
Unici dopo la cancellazione dei segmenti ripetuti	11.844	39,5

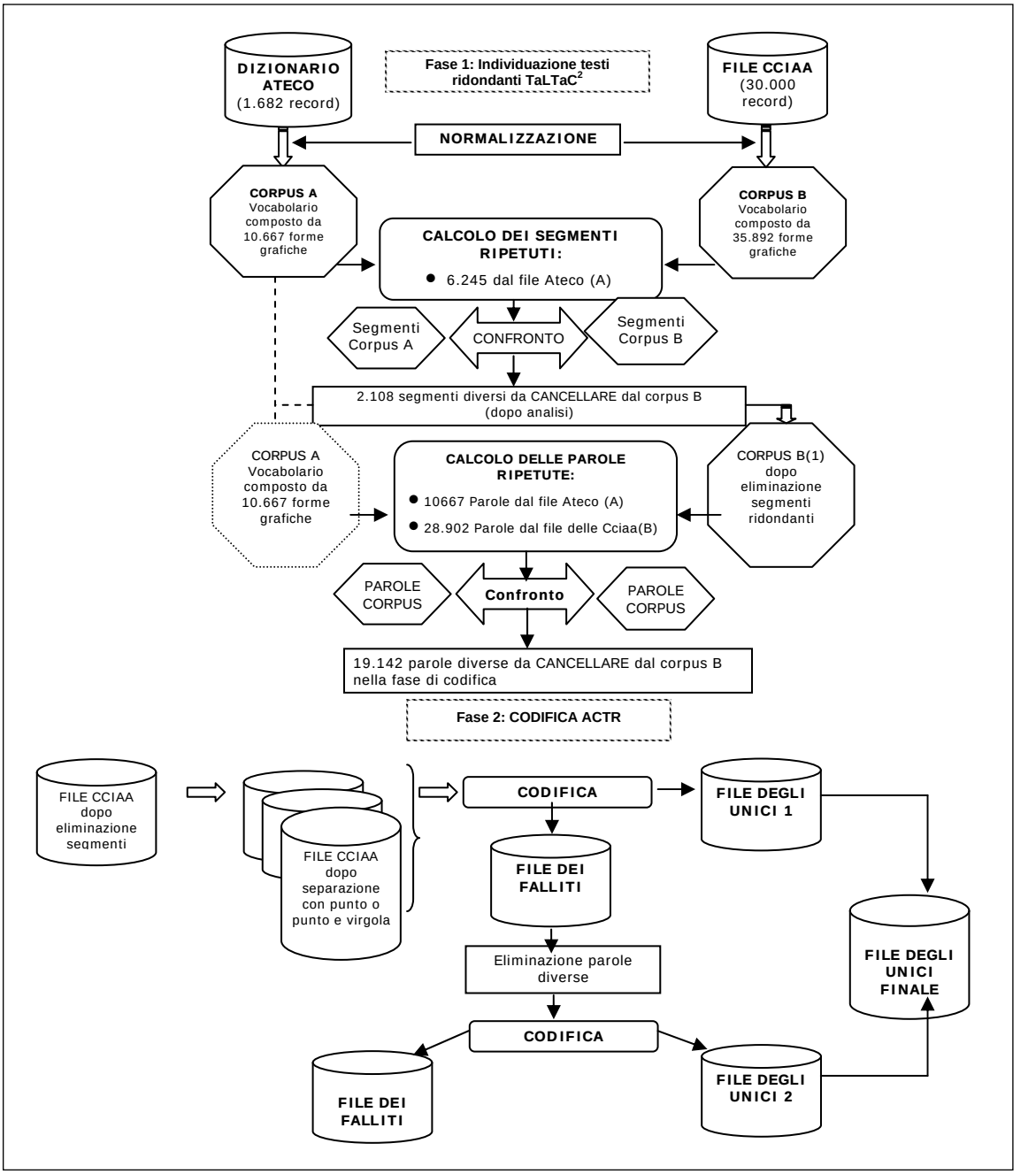
Si può notare come la cancellazione dei segmenti ripetuti abbia avuto un impatto positivo sul *recall rate* che, nel caso dei testi più corti, è cresciuto di 20 punti percentuali raggiungendo un livello più alto di quanto ottenuto in precedenti esperienze di codifica. Per questo motivo è stato deciso, come già accennato, di non procedere con la cancellazione delle parole. Per quanto riguarda i testi più lunghi, il pre-trattamento ne ha addirittura consentito la codifica ottenendo un buon tasso di codifica del 40 per cento circa. Per questi testi si è proceduto al secondo passaggio di codifica dopo la cancellazione delle parole ridondanti, come riportato nella tavola seguente:

Tavola 6.2 - *Recall rate* sui testi più lunghi dopo la cancellazione delle parole ridondanti

CAMPIONE DI DESCRIZIONI PIÙ LUNGHE DI 200 BYTE	Numero di aziende codificate	*Recall rate*
Unici prima della cancellazione delle parole ripetute	11.844	39,5
Unici dopo la cancellazione delle parole ripetute	13.917	46,4

Anche in questo caso si è registrato un incremento del *recall rate* (di circa 7 punti percentuali) che, sebbene più contenuto del precedente, in quanto parte delle parole da togliere erano già contenute nei segmenti eliminati, ha permesso di considerare la procedura efficace e di applicarla all'intero universo dei dati delle Cciaa sui quali è stata valutata la qualità dell'applicazione come sarà descritto nel paragrafo seguente.

Di seguito viene riportato uno schema grafico della procedura di codifica integrata.



6.4 La qualità del processo integrato di codifica automatica

La valutazione qualitativa della procedura di codifica intergrata in termini di tasso di codifica e di precisione ha l'obiettivo di valutare se sia possibile estendere la codifica automatica anche ad archivi diversi da quelli generalmente utilizzati in Istituto ossia provenienti da indagini statistiche. In questo modo sarebbe possibile estendere il trattamento dei dati anche a fonti che

per loro natura non possono entrare direttamente in un processo di codifica automatica (come già detto nel corso dei capitoli precedenti), allargando la fonte di informazione a cui attingere per l'aggiornamento dei dizionari alla base delle classificazioni.

La procedura di codifica automatica integrata è stata applicata ai testi delle Cciaa rimasti da codificare.[10] Questi ammontano a 1.685.037 descrizioni delle quali 817.528 più corte di 200 byte e le rimanenti, 864.509, più lunghe. La procedura è stata valutata sia in termini quantitativi che qualitativi, utilizzando:

- il *recall rate* per determinare la quantità di codici unici assegnati all'intero universo delle descrizioni delle Cciaa;
- il *precision rate* per valutare la qualità della codifica analizzando un campione di testi delle Cciaa.

I tassi di codifica ottenuti sono molto buoni, come si può osservare dai dati riportati nella seguente tavola.

Tavola 6.3 - *Recall rate* del processo integrato di codifica automatica

DESCRIZIONI DELLE CCIAA	Numero di record	*Recall rate* = % di codici unici
più corte di 200 byte	817.528	61
più lunghe di 200 byte	864.509	36
Totale	**1.682.037**	**48 in media**

È molto importante precisare che per questa applicazione il significato di *recall rate* è stato ampliato in quanto include oltre agli unici, classificati come tali da ACTR, anche quei multipli e quei possibili ai quali l'algoritmo di codifica ha associato descrizioni diverse, ma afferenti sempre allo stesso codice. Il motivo di questa scelta è insito nell'obiettivo del presente lavoro che era quello di codificare il maggior numero possibile di descrizioni, ben sapendo che ad una maggiore quantità di codici "unici" assegnati in questo modo sarebbe potuto corrispondere un minor livello di qualità (i multipli e i possibili hanno un punteggio di codifica più basso degli unici). Nel seguito della trattazione verranno usati i seguenti simboli per distinguere le diverse tipologie di unici:

- UU indicherà gli unici veri e propri, ossia quelli derivanti direttamente da ACTR;
- UM indicherà gli unici provenienti dai multipli;
- UP sarà riferito agli unici derivanti dai possibili.

Il *recall rate* descritto nella precedente tavola viene quindi decomposto nel seguente modo.

Tavola 6.4 - Decomposizione del recall rate tra i codici UU, UM e UP

DESCRIZIONI DELLE CCIAA	Unici per le descrizioni più corte di 200 byte		Unici per le descrizioni più lunghe di 200 byte	
	Valori assoluti	Valori percentuali	Valori assoluti	Valori percentuali
Totale Unici	501.650	61	307.538	36
Unici (UU)	355.001	43	126.689	15
Unici da possibili e da multipli (UP-UM)	147.649	18	180.849	21

[10] Il file originario delle Cciaa ammontava a circa 6 milioni di record. La maggioranza dei testi è stata codificata attraverso gli Studi di Settore, lasciando alla codifica automatica la parte restante dei testi a cui non era stato possibile attribuire un codice.

Se avessimo considerato solo gli UU, il *recall rate* sarebbe stato molto più basso, specialmente per le descrizioni lunghe, mentre sarebbe stato ancora consistente per quelle più corte di 200 byte. Visto l'obiettivo di una codifica estesa, sono stati quindi inclusi anche i codici UM e UP tenendo sotto controllo il livello di qualità.

Un approfondimento meritano i testi più lunghi di 200 byte. Come già accennato, questi testi sono stati spezzati in sub-descrizioni in funzione della presenza del punto o del punto e virgola nel testo, prima di essere sottoposti a codifica automatica. Ogni singola porzione di testo è stata codificata ammettendo la possibilità di attribuire ad un'unica azienda anche più di un codice Ateco. Da qui la necessità di dover stabilire a posteriori, attraverso gli Studi di Settore, quale fosse l'attività prevalente dell'azienda rispetto alla quale essere classificata univocamente.

Le due seguenti tabelle mostrano il *recall rate* per descrizione e la presenza congiunta di codici unici nella varie descrizioni.

La prima tavola mostra dei *recall rate* abbastanza elevati ad eccezione della prima descrizione. Questa differenza dipende dalla numerosità dei testi da codificare che, essendo molto più alta per il primo gruppo, contiene una maggior varietà di descrizioni tra le quali anche quelle impossibili da codificare.

Tavola 6.5 - *Recall rate* delle singole descrizioni

DESCRIZIONI	Numero di descrizioni	Unici Valori assoluti	*Recall rate* Valori percentuali
Prima descrizione (D1)	863.472	211.369	24
Seconda descrizione (D2)	348.846	149.485	43
Terza descrizione (D3)	75.685	37.797	50
Quarta descrizione (D4)	16.808	8.730	52
Quinta descrizione (D5)	4.173	2.170	52
Totale	**1.308.984**	**409.551**	**100**

La seconda tavola mostra la "sovrapposizione" di codici unici per azienda. Si può osservare che per la gran parte delle aziende, ossia il 73,5 per cento, l'attività economica è contenuta in una sola descrizione che, per circa la metà dei casi (45,7 per cento), è la prima.

Tavola 6.6 - Sovrapposizione di codici unici per azienda

UNICI	Valori assoluti	Valori percentuali	
Unici in tutte e 5 le descrizioni	504	0,2	
Unici in 4 descrizioni	2.774	0,9	
Unici in 3 descrizioni	13.088	4,3	
Unici in 2 descrizioni	65.168	21,2	
Unici in solo 1 descrizione	226.004	73,5	*di cui:*
Totale	**307.538**	**100,0**	45,7% 1ª descrizione
			25,4% 2ª descrizione
			2,2% 3ª descrizione
			0,1% 4ª descrizione
			0,0% 5ª descrizione

Quindi, considerando i buoni tassi di codifica e la bassa incidenza di casi di sovrapposizione di codici unici per azienda, che non rende troppo complicata la determinazione dell'attività prevalente attraverso gli studi di settore, si è deciso di utilizzare tutti i codici unici associati alla singola azienda per una migliore individuazione del settore economico di appartenenza.

Per valutare la qualità di codici unici assegnati si è utilizzato, come già detto, il *precision rate*. Per determinare questo indice è stato necessario analizzare manualmente i testi delle descrizioni codificate come uniche: esperti della codifica dell'Ateco hanno controllato la corrispondenza o meno tra il testo di input ed il codice assegnato. Al termine di questo lavoro è stato possibile determinare la porzione di codici unici corretti e la parte di quelli errati, individuando per questi ultimi le possibili fonti di errore. In particolare, è stato possibile classificare i codici errati in funzione di due tipologie di errori, ossia:

1) codici erroneamente assegnati: dipendono dalla carenza informativa del dizionario e da regole di parsing sbagliate;

2) descrizioni impossibili da codificare: si verificano nel caso in cui la descrizione è generica o comunque non ha un significato tale da consentirne una codifica univoca per cui un codificatore manuale non avrebbe assegnato alcun codice. Per il sistema automatico invece, la presenza di parole particolari (parole con un alto peso) "attrae" il testo verso dei codici unici che in questo modo vengono assegnati erroneamente.

Data la complessità di questa fase di valutazione della qualità e considerando il fatto che per i testi lunghi occorreva determinare il *precision rate* per ciascuna descrizione, si è deciso di calcolare l'indice non sull'intero universo dei testi, ma su di un campione casuale. In particolare sono stati utilizzati due campioni casuali, uno di 2 mila casi per i testi lunghi e uno di mille unità per quelli più corti. L'uso di una maggiore numerosità campionaria per i testi lunghi è legato al minore ammontare di codici unici (circa il 50 per cento in meno rispetto ai testi corti) e quindi alla necessità di ottenere un numero significativo di casi da analizzare. Le due numerosità campionarie garantiscono, rispettivamente, un errore campionario sulle stime pari rispettivamente a ± 2 per cento e ± 3 per cento calcolato sotto le ipotesti di popolazioni finite, P = *recall rate*, Q = 1 - P e $\alpha = 0{,}05$.

La formula utilizzate per calcolare l'errore è la seguente:

$$e = 1{,}96 * \sqrt{\frac{p*q}{n} \Big/ \frac{N-n}{N-1}}$$

dove:
p = recall rate, q = 1 - p
N = dimensione universo
n = dimensione del campione

I due campioni sono stati sottoposti al sistema di codifica integrata fornendo i seguenti risultati.

Tavola 6.7 - Campione di testi corti: tipologia di unici e *recall rate*

TIPOLOGIE	Valori assoluti	Valori percentuali
Dimensione campionaria	1.000	
Falliti	329	33
Codificati (Unici) = *recall rate*	671	67
Tipologia di codici unici:		
Unici UU	465	69
Unici da UP-UM	206	31
Totale	671	100

Tavola 6.8 - Campione di testi lunghi: tipologia di unici e *recall rate*

TIPOLOGIE	Descrizioni (a)									
	Totale		1ª descrizione		2ª descrizione		3ª descrizione		4ª descrizione	
	v.a.	%	v.a.	%	v.a.	%	v.a.	%	v.a.	%
Dimensione campionaria	2.000	100	2.000	100	793	100	178	100	38	100
Falliti (b)	1.283	64	1.536	67	426	54	90	51	16	42
Codificati (Unici) = *recall rate*	717	36	464	23	367	46	88	49	22	58
Tipologia di codici Unici:										
Codici UU			157	34	199	54	57	65	18	82
Codici UP-UM			307	66	168	46	31	35	4	18
Totale			464	100	367	100	88	100	22	100

(a) La 5ª descrizione non è stata usata in quanto conteneva pochissimi casi (soltanto 11)
(b) I Falliti contengono anche le descrizioni vuote

Il *recall rate* campionario è simile a quello calcolato sull'intero universo (vedi tavola 6.7). La presenza di codici UP-UM incrementa il *recall rate* complessivo e questo è vero soprattutto per la prima descrizione dei testi lunghi dove questi codici rappresentano il 66 per cento degli unici.

Questo risultato può comunque essere accettato in quanto non influenza troppo negativamente il livello di qualità. Se si osserva infatti la tavola che segue, è possibile notare che i *precision rate* sono molto alti a prescindere dal tipo di codici unici:

Tavola 6.9 - *Precision rate* dei testi più corti e più lunghi di 200 byte

TESTI	Testi più corti di 200 byte		Testi più lunghi di 200 byte							
			1ª descrizione		2ª descrizione		3ª descrizione		4ª descrizione	
	v.a.	%	v.a.	%	v.a.	%	v.a.	%	v.a.	%
Unici	671	100	464	100	367	100	88	100	22	100
non corretti	22	3	57	12	64	17	8	9	4	18
corretti = *Precision rate*	649	97	407	88	303	83	80	91	18	82

In particolare:

- i *precision rate* sono più alti dell'80 per cento per tutte le descrizioni. Il livello di qualità più alto si registra per le descrizioni corte, come era del resto da aspettarsi, in quanto l'essere corti li rende più simili ai testi derivanti dalle domande aperte usate nelle indagini statistiche e quindi più adatti ad una codifica automatica;
- tra i testi lunghi il miglior *precision rate* si ottiene per la prima descrizione a dimostrazione del fatto che i codici UP-UM non influenzano negativamente la qualità.

6.4.1 L'utilizzo della procedura integrata in altri contesti

Per capire se questo nuovo sistema di codifica, basato sull'integrazione di metodi di codifica e metodi di analisi testuale, possa essere utilizzato in altre situazioni dove i testi da codificare hanno lunghezze eccessive o contenuti poco adatti ad un processo di codifica, sono state fatte ulteriori analisi sui testi lunghi che vanno a confrontare il *recall* e il *precision rate* per le singole descrizioni. In particolare ci siamo soffermati sulle prime due descrizioni che, a differenza delle altre, avevano una numerosità elevata.

Dalla tavola 6.9 del precedente paragrafo, osserviamo che la seconda descrizione ha un *precision rate* dell'83 per cento ossia più basso di quello della prima descrizione pari all'88 per

cento. E questo nonostante la seconda descrizione contenga una proporzione di codici UU più elevata (54 per cento contro il 34 per cento della prima, cfr. tavola 6.8).

Per capire il perché di questo abbiamo analizzato il "tipo di errore" commesso nella codifica di entrambe le descrizioni, notando che nella seconda c'è una maggiore incidenza di "descrizioni impossibili da codificare", come riportato nella tavola di seguito:

Tavola 6.10 - Incidenza delle diverse tipologie di errore sulla codifica

TIPOLOGIE DI ERRORE	1ª descrizione		2ª descrizione	
	Valori assoluti	Valori percentuali	Valori percentuali	Valori assoluti
Codici erroneamente assegnati	47	83	26	41
Descrizioni impossibili da codificare	10	17	38	59
Totale di codici non corretti	**57**	**100**	**64**	**100**

Questo fatto, insieme alla maggiore incidenza di codici UP-UM per la seconda descrizione, ci dice che, con l'approccio integrato, l'errore sulla seconda descrizione è dovuto all'associazione di unici fittizi (UP-UM) a descrizioni generiche o poco significative. Questo è una conferma del fatto che il sistema non associa i codici UU a questo tipo di descrizioni e che, quindi, l'ambiente di codifica è ben costruito.

È stato osservato, inoltre, che la qualità delle due descrizioni cambia a seconda se siano entrambi presenti o se ciascuna rappresenti l'unica descrizione per l'azienda:

- solo prima descrizione: *precision rate* = 88 per cento
- solo seconda descrizione: *precision rate* = 83 per cento
- sia prima che seconda descrizione: *precision rate* della prima = 85 per cento, *precision rate* della seconda = 78 per cento.

Questo è un risultato ovvio, in quanto, il fatto di essere entrambe presenti significa che la descrizione originaria è molto lunga e di conseguenza più difficile da codificare. Comunque, da quanto detto potremmo dedurre che, in caso di codici multipli per azienda, potrebbe essere conveniente utilizzare solo una descrizione, ossia la prima, per stabilire il settore di attività economica.

Questo ultimo risultato trova conferma nei dati riportati nella seguente tavola che confronta la correttezza delle prima e della seconda descrizione quando sono entrambe codificate per la stessa azienda. Anche se i casi analizzati sono solo 137, la tavola mostra che la prima descrizione ha una frequenza più alta di codici corretti (21 contro 13). Comunque il risultato più importante che si evince da questo ultimo prospetto è che quando entrambe le descrizioni sono codificate il codice assegnato è corretto (97 casi su 137) e questo a dimostrazione dell'efficienza e dell'accuratezza della nuova procedura integrata di codifica automatica.

Tavola 6.11 - *Precision rate* dei codici attribuiti ad entrambe le descrizioni

1ª DESCRIZIONE	2ª descrizione		
	Codici corretti	Codici non corretti	Totale
Codici corretti	97	21	118
Codici non corretti	13	6	19
Totale	**110**	**27**	**137**

6.5 Conclusioni

La procedura di codifica integrata rappresenta un ottimo strumento innovativo in quanto offre risultati positivi sia in termini di efficienza sia di accuratezza dei codici attributi che hanno *recall* e *precision rate* molto elevati.

Questo ci permette di concludere che la procedura può essere utilizzata anche in altre situazioni dove i testi da codificare non sono a priori adatti per un sistema di codifica automatica. Infatti, grazie all'integrazione di un software per la codifica automatica, ACTR, ed un software per l'analisi testuale, TaLTaC^2, è stato possibile trattare l'informazione, cancellando le ridondanze e le parole inutili, senza alterare il contenuto e il significato del testo.

Questo nuovo approccio rende quindi possibile estendere la codifica anche a nuove fonti di informazione diverse da quelle provenienti dalle indagini aumentando la possibilità di aggiornare le basi informative degli ambienti di codifica.

Capitolo 7 - Procedura di consultazione su Web

7.1 La necessità di individuare l'Ateco su Web

Il progetto di mettere a disposizione di tutti gli utenti lo strumento di codifica automatica è stato concepito nel 2004 dopo aver aggiornato il sistema ACTR con la classificazione Ateco 2002. Lo strumento ormai era stato testato nella sua forma batch in molte indagini e ne era stata apprezzata l'utilità sia per la classificazione delle attività economiche sia per altre classificazioni. Già con la nuova versione del 2002 si era visto che la pagina del sito dedicata all'Ateco era molto consultata e quindi sembrava opportuno mettere a disposizione degli utenti quanti più strumenti possibili. Soprattutto, dopo ogni aggiornamento della classificazione, anche i semplici cittadini hanno bisogno di trovare una risposta ai loro quesiti.

Come già detto nel capitolo 2, per la prima volta la classificazione Ateco è unica e utilizzata da tutti gli Enti interessati (Agenzia delle Entrate, Camere di commercio, Inail e Inps). In precedenza, l'Agenzia delle Entrate pubblicava su Gazzetta Ufficiale l'Atecofin che differiva dall'Ateco per alcuni maggiori dettagli individuati dalle lettere dell'alfabeto al posto della quinta cifra numerica; le Camere di commercio applicavano l'Atecori nella quale erano presenti diversi ulteriori dettagli, rispetto all'Ateco, individuati dalle seste cifre numeriche. In quest'ultimo caso anche l'interpretazione della classificazione non era esattamente la stessa; ad esempio le Camere di commercio classificavano le pizzerie a taglio nel Manifatturiero mentre l'Istat le classificava nella Ristorazione. Uno dei compiti del sotto Comitato Ateco[11] era definire l'interpretazione comune della classificazione che, in alcuni punti, poteva risultare ambigua o non chiara, anche per problemi di traduzione dall'originale inglese.

I primi dati statistici con la nuova classificazione saranno diffusi nel 2009 ma, per gli adempimenti fiscali, essa è già in vigore dal 1 gennaio 2008. Il sito dell'Agenzia delle Entrate è collegato direttamente alla pagina web dell'Istat per quanto riguarda l'Ateco 2007; diventava pertanto prioritario per l'Istat fornire più strumenti possibili per la consultazione dell'Ateco. La pagina relativa alla classificazione Ateco risultava essere quella più consultata dopo quella relativa ai prezzi.

I primi adempimenti fiscali dell'anno cadevano a fine febbraio; in quel periodo la casella di posta elettronica dedicata all'Ateco ha ricevuto molti quesiti. ACTR su Web è stato realizzato anche pensando all'alleggerimento del lavoro quotidiano per gli esperti codificatori dell'Istat; infatti, dopo la messa in linea dell'applicazione, gli interrogativi inviati all'apposita casella di posta elettronica *ateco2007@istat.it* sono notevolmente diminuiti.

La realizzazione di ACTR su Web comporta alcune importanti ricadute per gli aspetti relativi alla classificazione Ateco ma anche per le altre classificazioni e per l'Istituto nel suo insieme.

Le descrizioni inserite dagli utenti all'atto dell'immissione della richiesta on-line possono essere utilizzate, a seguito di una specifica analisi e trattamento da parte delle figure competenti, per incrementare il dizionario ACTR che già sottostà all'applicazione.

L'utilità di una tale funzione su Web può andare oltre la specifica esigenza espressa per l'Ateco. In merito alle rilevazioni sulle imprese, per esempio, nei questionari di alcune indagini è

[11] Il sotto Comitato è stato costituito all'interno del Comitato Ateco con il mandato di discutere la classificazione e di definire la versione italiana della classificazione. Affinché le riunioni fossero proficue, il sotto Comitato era costituito da un gruppo ristretto di esperti di Istat, Unioncamere, Agenzia delle Entrate (Studi di Settore) e Inps più eventuali esperti dei singoli settori convocati di volta in volta.

riportata per ciascuna azienda l'attività economica risultante dagli archivi delle imprese e si richiede agli intervistati di aggiornarla se non corretta, oppure se variata. In presenza di tale funzione, nei questionari stessi si può riportare il riferimento al sito Web cui accedere per descrivere l'attività economica espletata e verificarne il codice corrispondente.

Anche per le altre classificazioni, inoltre, può essere utile, per diverse tipologie di utenti, rendere disponibile una funzione di trattamento on-line; tali figure possono essere:

- i rilevatori comunali ai quali a volte si richiede di codificare alcune risposte testuali;
- gli enti del Sistan che si avvalgono delle classificazioni utilizzate dall'Istat;
- altri soggetti specifici, quali, ad esempio, i medici per le Cause di morte.

L'altro obiettivo fondamentale di questa applicazione, per l'Ateco e per le altre classificazioni, è che i dizionari di queste possono essere aggiornati sulla base di un ritorno controllato delle richieste più significative effettuate dagli utenti. Questa attività richiede una gestione, non eccessivamente onerosa, ma costante di ciò che perviene dall'esterno e rappresenta forse l'implicazione più interessante offerta dal progetto, in quanto, una volta realizzata, consentirebbe di tener conto di tutto un complesso di informazioni aggiuntive sulla cui base operare aggiornamenti ad integrazione della classificazione standard delle attività economiche, necessità che si presenta con una cadenza interna più frequente delle versioni ufficiali delle classificazioni stesse.

Nel momento in cui si è deciso che era opportuno procedere alla realizzazione di ACTR per gli utenti esterni, si è valutato che adattare il prodotto ad un utilizzo su sito Web richiedeva tempi di realizzazione piuttosto brevi. L'altro elemento chiave per decidere la messa on-line dell'applicazione era relativo alla valutazione del numero di query che potevano giungere contemporaneamente al server dell'Istat e alla capacità di questo di gestirle in un lasso ragionevole di tempo ovvero pochi secondi di attesa per l'utente. Come già detto, ci si aspettava un numero molto elevato di contatti da parte degli utenti esterni. I test di carico effettuati per verificare questo aspetto (cfr. par. 7.3.6) hanno avuto esito positivo e quindi si è proceduto alla realizzazione definitiva del prodotto.

7.2 Caratteristiche dell'applicazione di consultazione su Web

Dal punto di vista contenutistico non ci sarebbe differenza tra i risultati ottenibili nel far elaborare in batch da ACTR un dataset di descrizioni di attività economiche e consultare on-line ACTR su Web perché individui un codice corrispondente alla descrizione fornita.

Indubbiamente, dal punto di vista tecnico, il mettere a disposizione su Web questa funzione di consultazione implica le complessità di gestire più accessi contemporanei e di non poter utilizzare direttamente l'interfaccia grafica di ACTR (cfr. par. 7.3).

Ci sono tuttavia differenze importanti tra l'applicazione batch e quella Web che sono state implementate per soddisfare due tipologie di esigenze:

- mentre la finalità principale di un'applicazione batch è di massimizzare i codici univoci assegnati automaticamente, un utente che consulta il sito Web per individuare il codice Ateco corrispondente all'attività economica da lui espletata può trarre vantaggio dal poter analizzare diverse descrizioni di attività (corrispondenti a diversi codici) affini a quella espletata e di selezionare quindi, tra queste, quella a lui più attinente;
- mentre nel codificare i dati di un'indagine, può essere di utilità anche individuare un codice non al massimo dettaglio, qualora la descrizione fornita sia generica (ad esempio perché i dati vengono pubblicati ad un certo livello di dettaglio, oppure perché ci si può avvalere di ulteriori informazioni disponibili sul questionario di rilevazione per completare un codice generico), un utente che consulta il sito Web per individuare

il codice Ateco corrispondente all'attività economica da lui espletata ha bisogno necessariamente del codice al massimo livello di dettaglio.

Per queste due motivazioni sono state introdotte le seguenti varianti all'applicazione su Web che influiscono sui risultati, differenziandola dall'applicazione batch:

- sono stati modificati i parametri soglia utilizzati da ACTR per misurare la similarità tra i testi ed individuare il codice, in modo da aumentare le possibilità che, laddove la descrizione fornita non realizzi un *direct match*, il sistema proponga un ventaglio di codici possibili piuttosto che uno solo corrispondente alla descrizione più simile a quella fornita;
- è stato elevato a 7 il set di codici con le corrispondenti descrizioni proposte dal sistema (il parametro utilizzato nel batch è usualmente 5);
- non è stato messo in linea il dizionario completo utilizzato da ACTR per l'applicazione batch, ma un estratto di quest'ultimo contente esclusivamente le descrizioni corrispondenti ai codici a cinque cifre;
- questo dizionario è stato collegato ad una tavola, contente le descrizioni corrispondenti ai codici a sei cifre, tale che l'utente, una volta individuato il codice a cinque cifre pertinente alla propria attività, possa visualizzare anche i codici a sei cifre (con le relative descrizioni) corrispondenti a quest'ultimo.

Il fatto di inibire l'attribuzione di un codice che non sia al massimo dettaglio è stato inoltre gestito non soltanto fornendo all'utente indicazioni specifiche su come descrivere la propria attività economica (ad esempio non fornire descrizioni generiche, non utilizzare abbreviazioni che potrebbero risultare ambigue per il sistema eccetera), ma anche tramite una messaggistica di errore che, in caso di mancata attribuzione del codice, rimandasse a suggerimenti su come esplicitare meglio la propria attività.

Inoltre, è stata introdotta un'ulteriore funzione in questa applicazione ossia la memorizzazione delle query effettuate dagli utenti per consentire di valutare la qualità dei risultati e per aggiornare sistematicamente la base informativa utilizzata dal sistema.

In pratica, con cadenza settimanale, il dataset contenente le query effettuate viene inviato agli esperti della classificazione; questi ultimi lo sottopongono ad un passaggio di codifica batch (che utilizza lo stesso dizionario e gli stessi parametri soglia dell'applicazione Web), ne analizzano i risultati ed effettuano quindi, con la stessa cadenza, gli interventi di manutenzione della base informativa, sia in termini di correzione di eventuali errori che di arricchimento del dizionario.

7.3 Il funzionamento dell'applicazione su Web

7.3.1 Descrizione dei flussi

L'applicazione Web è costituita da pagine Html dinamiche scritte in Php e consta essenzialmente di due parti.

La **prima parte** comprende le pagine Web che si trovano sul server Samo1 (S.O. Linux) che ospita il sito dell'Istituto ed in particolare sono:

- la pagina di presentazione dello strumento (figura 7.1) in cui vengono acquisite le query dell'utente;
- la pagina di pubblicazione dei risultati ottenuti con il sistema ACTR. Questo sistema è configurato in modo da fornire in output un elenco di descrizioni delle possibili attività (da 1 a 7 item); l'utente deve poi scegliere da quest'elenco la descrizione che ritiene più vicina alla propria attività economica e selezionare il bottone di conferma (figura

7.2). Nel caso in cui alla query dell'utente non corrisponda nessuna attività economica, il sistema ACTR fornisce un messaggio di ricerca fallita (figura 7.5);

- la pagina di pubblicazione dei risultati finali che fornisce il codice a 6 cifre e il relativo titolo ufficiale della classificazione (figura 7.3);
- la pagina statica contenente i suggerimenti per un corretto utilizzo dello strumento (figura 7.4).

Figura 7.1 - Pagina iniziale dell'applicazione

La query utente è la descrizione sintetica (massimo 200 caratteri) dell'attività economica che viene successivamente elaborata dal software ACTR per la codifica automatica del testo.

Figura 7.2 - Risultato della query utente

La funzione di "nuova ricerca" ricarica la pagina iniziale eliminando tutti i dati digitati precedentemente dall'utente per avviare quindi una nuova sessione di ricerca.

Figura 7.3 - Output finale della ricerca: codifica ufficiale delle attività economiche Ateco 2007

In questa pagina è stata implementata la funzionalità di navigazione da quest'ultima pagina di output a quella intermedia per permettere all'utente di poter fare una diversa selezione dall'elenco delle classificazioni proposte da ACTR.

Figura 7.4 - Istruzione per la ricerca dell'attività economica

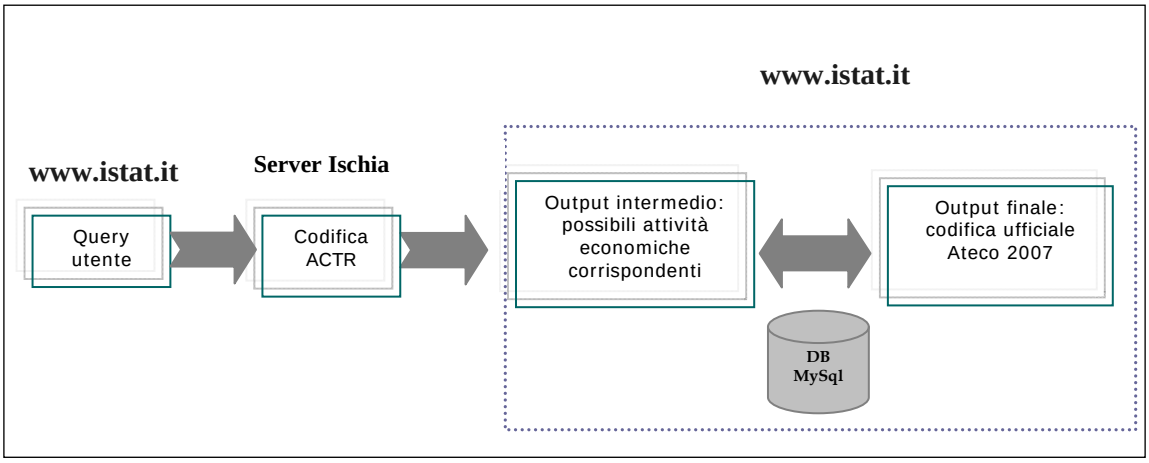
Nella pagina Suggerimenti viene data una spiegazione più estesa di come utilizzare questa applicazione.

Figura 7.5 - Ricerca fallita

Nel caso l'utente digiti una stringa errata viene visualizzata questa pagina che fornisce delle indicazioni per effettuare una corretta ricerca.

La **seconda parte** dell'applicazione comprende i programmi in Php che si trovano sul server Ischia su cui gira il software ACTR. Questi programmi vengono chiamati in http dalle pagine del sito e si preoccupano di costruire gli input per il software ACTR, eseguire il software, recuperare gli output, formattarli e farli pubblicare nella pagina mostrata nella figura 7.2. Si noti che il software ACTR, come descritto nel successivo paragrafo, è tale che non può accettare più richieste in parallelo, per cui è stato necessario serializzare le richieste utente provenienti dalla pagina Web principale.

Il flusso dell'intera applicazione è schematizzato come segue:



Nell'ultima fase, l'applicazione prende il codice ACTR (di cinque cifre) selezionato dall'utente e lo confronta con i dati memorizzati sulla tavola "ACTRATECO" del data base mysql del sito e va a prendere tutti quei codici Ateco 2007 (di sei cifre) che hanno le prime cinque cifre uguali. Il risultato di questa select è visualizzato in output (figura 7.3) fornendo all'utente oltre ai codici anche i titoli e le note esplicative.

7.3.2 Caricamento codici Ateco nel database

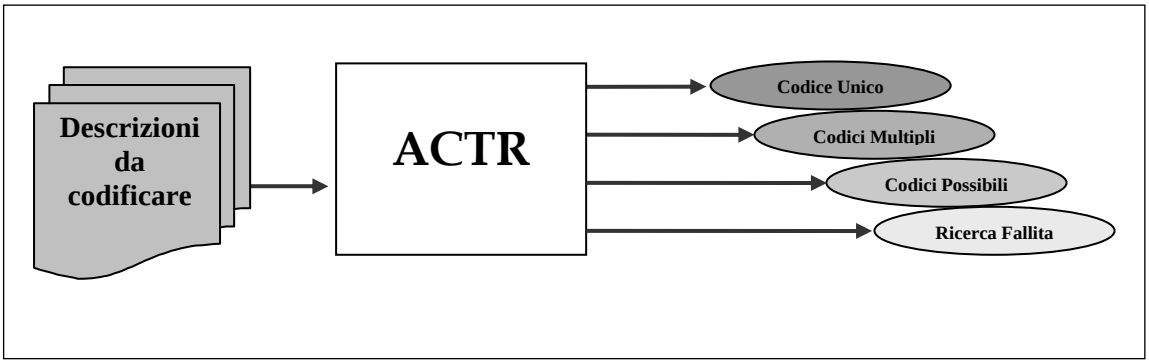
I codici della classificazione delle attività economiche Ateco 2007 vengono memorizzati nella Tavola "ACTRATECO" del database mysql del sito mediante un programma in Php. I dati vengono letti da un file Excel ed inseriti nel database.

È previsto un aggiornamento sia dei codici, sia delle relative descrizioni con periodicità annuale.

7.3.3 Integrazione del software ACTR nell'applicazione Web

L'applicativo ACTR legge il file di input che viene creato dinamicamente dall'applicazione Php con le query utente, elabora questo file utilizzando opportune regole di *parsing*, trova il/i codici e la relativa descrizione che corrispondono alla query digitata dall'utente. I risultati vengono scritti alternativamente in uno dei quattro file Ascii, a seconda del tipo di corrispondenza trovata (cfr. par. 1.1).

L'applicazione Web legge i file forniti da ACTR, li decodifica e visualizza sulla pagina con un numero massimo di 7 righe nel caso di risultati multipli o possibili. Nel caso invece ACTR non trovi nessuna associazione viene fornito un opportuno messaggio (figura 7.5).



7.3.4 Gestione dello storico delle query utente

Tutte le query digitate dagli utenti vengono scritte in un file memorizzato nel server windows; periodicamente, in modalità automatica, il file viene inviato via e-mail ad una persona designata (esperto della classificazione) che si occupa dell'analisi di questo file. L'analisi del file permette un arricchimento ed affinamento del dizionario di riferimento di ACTR.

Per automatizzare questo processo è stato creato un demone in VBscript (ossia un processo sempre attivo) che con una fissata periodicità:

- verifica l'esistenza del file da inviare;
- crea e invia la e-mail con il file allegato;
- rinomina il file inviato aggiungendo nel nome l'informazione della data di invio.

Questi file così rinominati vengono storicizzati per un periodo di tempo opportuno e poi cancellati in modo automatico da un altro demone.

7.3.5 Aggiornamento ambiente applicativo ACTR

L'ambiente applicativo sul server ACTR-Ateco deve essere periodicamente aggiornato dagli esperti Istat della classificazione. In particolare vengono di volta in volta creati nuovi file (il dizionario, il file di configurazione eccetera) contenenti nuove regole di *parsing* che devono essere sostituiti ai relativi file sul server. A questo scopo è stata sviluppata un'applicazione Web in Php per fare l'uploader dei file sul server. L'accesso a questa pagina è controllato da un'autenticazione mediante username e password, sfruttando il servizio di autenticazione di rete del Web Server su cui viene eseguito, ovvero Apache. Verificate le credenziali dell'utente viene aperta una sessione e l'utente viene indirizzato verso la pagina protetta (figura 7.6).

Figura 7.6 - File Uploader

In questa maschera vengono selezionati i cinque file necessari per l'aggiornamento dell'applicativo ACTR. Tali file sono stati precedentemente preparati in una cartella predefinita. La funzionalità di uploader trasferisce infine i file sul server in una cartella di appoggio.

Infine, l'aggiornamento effettivo dell'ambiente di produzione viene fatto automaticamente da un programma VBscript che controlla, decomprime, sostituisce i suddetti file. L'esecuzione dello script è programmata tutti i giorni in orari opportuni per non creare disservizio. Alla fine di questo processo di aggiornamento viene mandata una e-mail contenente le informazioni sull'esito delle operazioni.

7.3.6 Test di carico

Dal momento che, come descritto nel par. 7.1, prima di esporre l'applicazione era difficile stimare il numero di accessi simultanei, è stato deciso di effettuare dei test di carico per verificare le prestazioni dell'applicazione e l'eventuale necessità di una diversa soluzione applicativa. Di seguito sono riportati i risultati ottenuti durante questi test di carico utilizzando il prodotto di simulazione JMeter.

I test sono stati effettuati simulando un numero da 10 fino a mille richieste contemporanee di ricerca di una stringa. Si tenga presente che i risultati ottenuti sono soggetti al traffico esistente sulla rete durante il periodo di simulazione.

I risultati dei test sono riassunti nelle tavole seguenti. I dati riportati in particolare sono:

- il tempo massimo, minimo e medio di risposta di un campione in millisecondi;

- il *throughput*, che esprime la quantità di dati trasmessi in una unità di tempo (dati/sec.);
- la percentuale di errore nell'esecuzione dei processi.

Tavola 7.1 - Risultati dei test di carico

NUMERO DI PROCESSI CONTEMPORANEI	10	50	100	200	300	500	600	650	700	1.000
Tempo massimo di risposta (ms)	2.217	10.878	21.889	47.779	68.150	113.234	134.139	145.208	156.942	188.977
Tempo minimo di risposta (ms)	251	227	285	305	264	279	227	252	214	-
Tempo medio di risposta (ms)	1.232	5.531	11.086	25.651	35.358	56.571	67.187	72.462	78.687	92.881
Throughput (dati/sec.)	4,5	4,6	4,6	4,2	4,4	4,4	4,5	4,5	4,5	5,3
% di errore	-	-	-	-	-	-	-	-	-	23,80

Un dato interessante da notare è la percentuale di errore. Come si vede nella tavola 7.1 fino ad un massimo di 700 richieste contemporanee la percentuale di errore si mantiene nulla; con richieste contemporanee superiori a 700 alcuni processi vanno in errore. Di seguito i grafici mostrano l'andamento dei tempi di risposta in funzione del carico sull'applicazione.

Grafico 7.1 - Tempo medio di risposta

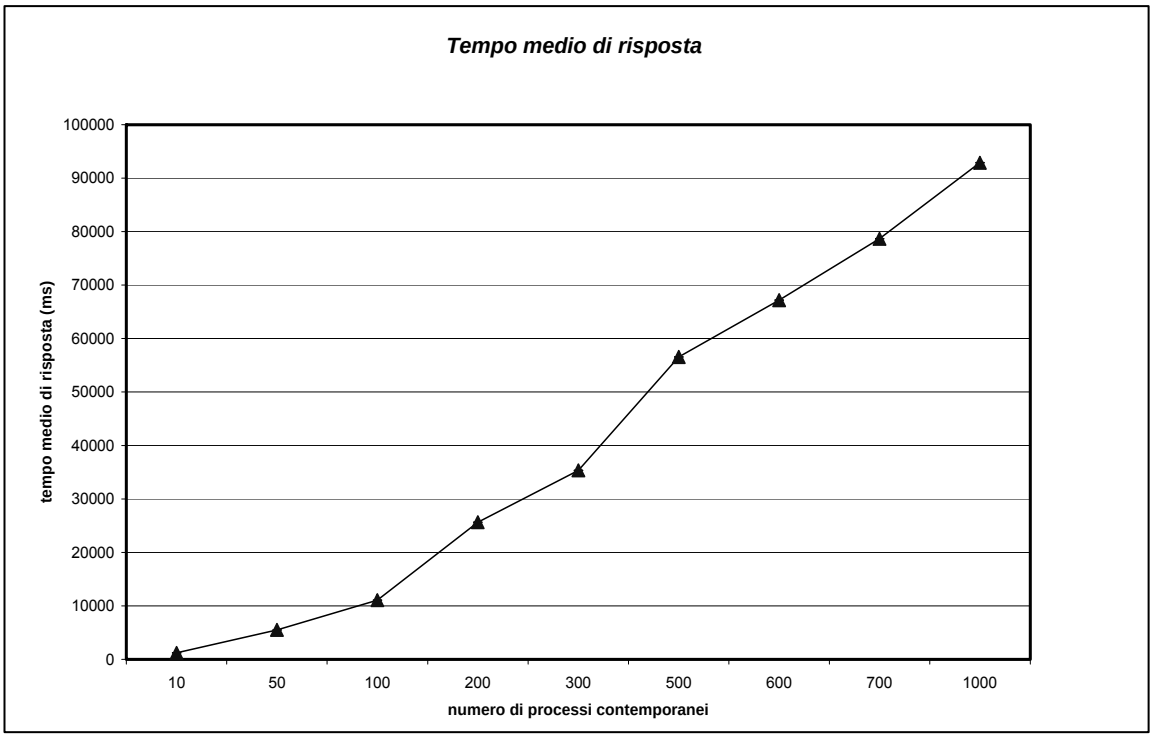


Grafico 7.2 - Tempo minimo di risposta

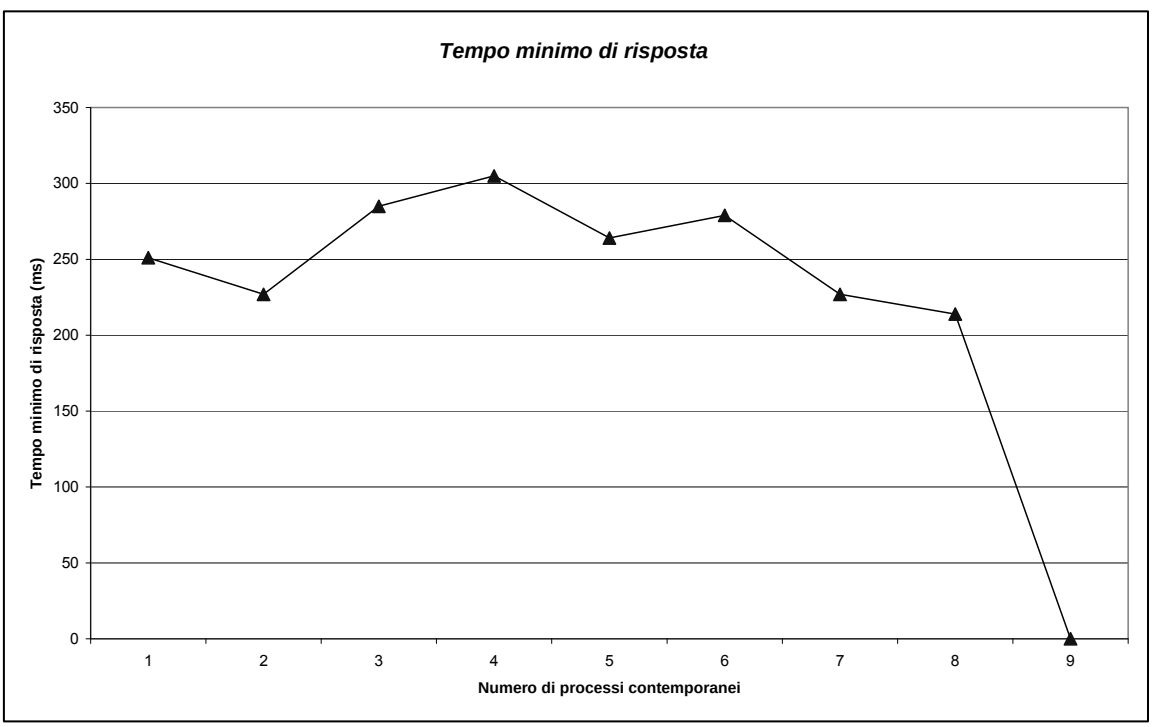
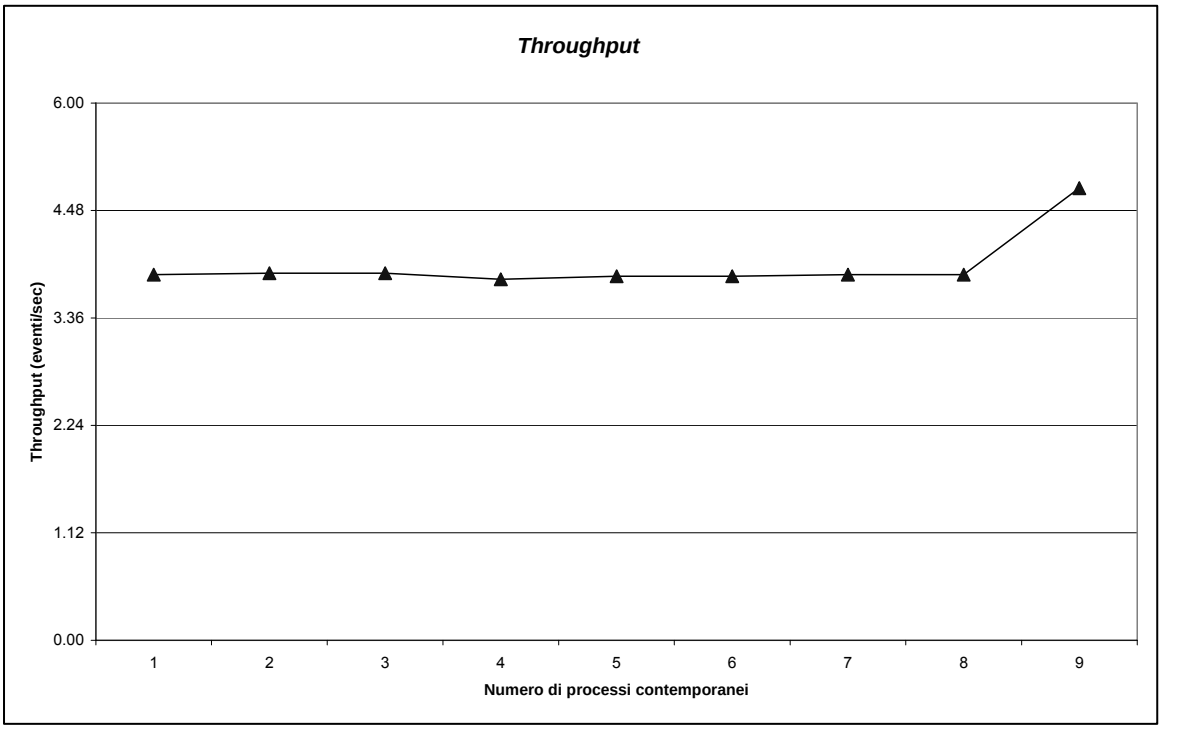


Grafico 7.3 - Throughput



Riassumendo, per un numero di processi contemporanei compreso tra 10 e 700, i valori misurati presentano la seguente variabilità:

Tempo massimo di risposta	2 sec - 3 min
Tempo minimo di risposta	0,2 sec - 0,3 sec
Tempo medio di risposta	1,2 sec - 1,3 min.

Visti questi risultati non si è ritenuto necessario apportare modifiche all'applicazione in quanto, pur non essendo in grado di stimare a priori con un elevato livello di precisione la quantità di accessi degli utenti, le performance dimostrate con questi test sono state considerate più che soddisfacenti.

7.4 Monitoraggio dell'applicazione

L'applicazione finora descritta è stata messa in produzione alla fine di maggio c.a. e, da allora, il monitoraggio è stato effettuato analizzando con cadenza settimanale il file delle query effettuate dagli utenti (la sintesi del monitoraggio è riportata nella tavola 7.2).

Tavola 7.2 - Sintesi del monitoraggio dell'applicazione Web

DATA	N. query	Percentuale codifica	N. testi vuoti	Dimensione dizionario a 5 cifre
06/06/08	8478	43	25	32.750
09/06/08	9225	43	38	32.750
16/06/08	6290	43	18	32.881
20/06/08	10.386	41	31	32.962
27/06/08	10.327	42	30	32.965
07/07/08	10.535	43	30	32.965
14/07/08	10.925	43	32	33.146
21/07/08	10.304	43	34	33.146
28/07/08	8.813	40	19	33.146
04/08/08	8.044	40	21	33.146
11/08/08	5.514	41	16	33.146
18/08/08	1.803	46	6	33.146
25/08/08	3.069	42	9	33.146
01/09/08	4.579	43	15	33.146
08/09/08	8.632	39	37	33.146
15/09/08	9.233	40	26	33.146
22/09/08	8.690	40	21	33.146
29/09/08	13.091	43	34	33.137
06/10/08	11.268	42	31	33.137
13/10/08	10.000	42	38	33.137
20/10/08	10.075	42	31	33.137

La prima osservazione da fare riguarda l'entità degli accessi che risulta decisamente elevata, attestandosi nelle ultime settimane intorno a 10 mila query.

Relativamente al tasso di codifica, il *recall rate*, nel suo insieme (circa 43 per cento), è indubbiamente inferiore a quello ottenuto dall'applicazione nel suo utilizzo per codificare i dati rilevati nelle indagini statistiche, tuttavia è stato considerato soddisfacente per diversi ordini di motivi, quali:

- il fatto che la percentuale di Unici ottenuti sottoponendo il file delle query a codifica batch sia del 43 per cento non significa che sia questa la percentuale di casi in cui gli utenti hanno individuato il codice corrispondente alla propria attività; infatti, come già accennato, sono stati assestati i parametri soglia per elevare la probabilità di proposizione di multipli nel caso di *indirect match*, il che significa che in moltissimi casi è stata data agli utenti la possibilità di selezionare il codice corrispondente alla propria attività tra quelli proposti dal sistema, soddisfacendo quindi la propria richiesta;
- la base informativa messa in linea (*reference file*), come già specificato, è costituita da un sottoinsieme del dizionario utilizzato dall'applicazione tradizionale, contenendo soltanto le descrizioni corrispondenti ai codici a cinque cifre;

- le descrizioni fornite dagli utenti hanno caratteristiche molto diverse da quelle fornite nelle indagini, per alcuni aspetti assimilabili a quelle delle Camere di commercio, a meno del fatto che sono tendenzialmente brevi; infatti, nonostante le istruzioni fornite, sono spesso destrutturate e con scarso contenuto informativo (evidentemente, l'utente che inserisce on-line una descrizione tende a porre meno attenzione alle istruzioni rispetto a quanto non faccia quando risponde ad un questionario di indagine, forse anche perché sa di poter effettuare più di un tentativo per ottenere una risposta).

Infatti, dall'analisi effettuata sui falliti, sono emerse diverse tipologie di casi non codificabili, tra le quali si citano di seguito le più frequenti:

- presenza di errori ortografici più disparati;
- inserimento del vecchio codice Ateco, come se la funzione messa in linea avesse la finalità di realizzare una transcodifica tra due tabelle;
- esplicitazione di una singola parola, priva del contenuto informativo per l'individuazione di un codice (ad esempio *abbigliamento*, senza specificare se si tratta di produzione o vendita, *edilizia* oppure *industria* eccetera);
- fornitura della descrizione della propria professione che, nella maggior parte dei casi, non è riconducibile ad un settore univoco di attività economica (ad esempio *impresario*, *operatore personal computer*, *revisore contabile* eccetera);
- descrizioni assolutamente carenti in termini di informatività (ad esempio *persona fisica*, *libero professionista* eccetera).

Alcune delle problematiche citate sono in fase di risoluzione, infatti già si è provveduto ad inibire la digitazione di numeri ed è in fase di studio la possibilità di anteporre all'esecuzione di ACTR un passaggio di correzione ortografica, tale da migliorare le descrizioni da sottoporre a codifica.

Tuttavia, come può vedersi dalla tavola, l'esame dell'insieme di casi codificati ha consentito l'arricchimento del dizionario utilizzato dalla codifica, il che, in prospettiva, comporterà dei miglioramenti anche nelle applicazioni batch.

Appendice

Specifiche tecniche *Parsing* Ateco 2007

Premessa di carattere generale

Come già detto nel capitolo 1 di questo documento, la fase di standardizzazione dei testi, detta *parsing*, ha lo scopo di eliminare dal testo la variabilità grammaticale o sintattica che non incide sul significato semantico ma soltanto sulla forma e che pertanto è irrilevante ai fini dell'abbinamento con le voci del dizionario della classificazione. La fase di standardizzazione 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* (mappatura dei caratteri, eliminazione di parole ininfluenti, definizione di sinonimi per parole singole o per gruppi di parole, rimozione di suffissi, prefissi eccetera), consente una notevole flessibilità e possibilità di personalizzazione del processo di standardizzazione.

Le fasi del processo di standardizzazione si possono suddividere in cinque parti:

1. **Pre-processing** (trattamento dei caratteri) Strimm trimming (pulizia dei caratteri). Consente la standardizzazione del testo tramite l'eliminazione di caratteri estranei, quali spazi multipli, caratteri di tabulazione eccetera. In particolare:
   - AUTOTRIM consente la cancellazione dei caratteri spuri, quali doppi blank, tabulazioni, andata a capo
   - TRIMLEF consente la cancellazione dei blank a sinistra del simbolo specificato ad esempio)]}-/\!?,
   - TRIMRIGHT consente la cancellazione dei blank a destra del simbolo specificato ad esempio ([{-/$
   - WCHR (Word Characters) consente la definizione dei caratteri validi di riferimento e la relativa traduzione ad esempio da minuscolo in maiuscolo ovvero la mappatura dei caratteri che compongono la parola (solitamente si passa da minuscolo a maiuscolo, si eliminano le vocali accentate).
2. **Phrase processing** (trattamento delle stringhe)
   - DCLS (Deletion Clauses) consente la cancellazione di tutto ciò che è compreso tra incisi o clausole ad esempio [ ]
   - DSTR (Deletion Strings) consente la cancellazione di stringhe ritenute inutili nella fase di codifica, ad esempio *nell'*
   - RSTR (Replacement Strings) consente la sostituzione di stringhe ad esempio c/terzo→*contoterzo*.
3. **Separazione in parole**
   - WCHR (Word Characters). La separazione della stringa in parole avviene in questa fase, facendo riferimento al WCHR, in quanto i caratteri non contenuti in tale lista sono considerati come delimitatori delle parole.
4. **Word Processing** (trattamento delle parole)
   - RWRD (Replacement Words) consente la vera e propria gestione dei sinonimi: sostituzione di parole in altre parole ad esempio *minuto→dettaglio*, oppure di parole in una coppia di parole ad esempio *autotrasporto→ auto trasporto,* così come la

sostituzione con blank di parole ininfluenti, ad esempio preposizioni, articoli eccetera.

- DWRD (Double Words) realizza la gestione dei sinonimi a livello di coppie di parole in sequenza. Consente infatti la sostituzione di due parole, che in sequenza acquistano un significato preciso, o con una parola o con due; ad esempio *abbigliamento firmato → abbigliamento* oppure *abbigliamento matrimonio → abbigliamento adulto*
- HWRD (Hyphenated Words) consente il trattamento di parole separate dal trattino ad esempio *baby-sitter → babisitter*
- IWRD (Illegal Words) consente la cancellazione dei caratteri illegali o spuri all'interno delle parole ad esempio i numeri
- EXCP (Excepion Words) consente la definizione di parole che non debbono essere sottoposte all'eliminazione di suffissi, prefissi, ad esempio *borsa →borsa*
- MCHR (Multiple characters) consente la definizione di caratteri che se doppi o tripli all'interno di una parola vengono ridotti ad uno, ad esempio *abbigliamento → abigliamento* può pertanto limitare l'effetto di errori d'ortografia
- PRFX (Prefixes) consente di definire dei prefissi da eliminare; viene effettuato solo se la parola troncata rimane almeno di 4 lettere
- SUFX (Suffixes) consente l'eliminazione dei suffissi delle parole ad esempio *oa, oe,oi;* viene effettuato solo se la parola troncata rimane almeno di 4 lettere.

5. **Post processing** (ulteriore trattamento delle parole)
   - RDUP (Remove Duplicates) consente la rimozione di parole doppie ossia duplicate
   - SORT (Ordinamento delle parole) la scelta se utilizzare questa funzione è comunque dettata dal tipo di contesto.

Nei processi sopra descritti l'unico obbligatorio è il WCHR. Il file risulta essere necessario perché in esso vengono specificati i caratteri che costituiscono le parole oggetto del *parsing*.

L'ordine con cui eseguire i processi nell'ambito di ciascuna fase, viene elencato nel *Parsing Strategy File*. La strategia varia ovviamente in funzione del contesto applicativo di codifica. I processi inclusi nella strategia sono individuati da parole chiave (*Process Keyword*); a ciascun processo corrisponde un *parsing data file*, denominato con una parola chiave dei dati (*Filedata Keyword*). L'ordine dei processi non è casuale: la sequenza scelta deve garantire la coerenza logica della successione degli effetti delle trasformazioni realizzate. L'unico processo che deve sempre occupare una posizione precisa è quello definito *Exception Words*. Infatti, questa funzione, impedendo le trasformazioni relative alle lettere multiple, ai suffissi e ai prefissi, richiede di essere indicata prima che queste siano già avvenute, quindi in linea generale è bene, nella strategia di *parsing*, anteporre sempre il processo che definisce le parole eccezione (EXCP) a quello dei SUFX, dei MCHR e PRFX.

Il *parsing* sviluppato per il contesto di codifica relativo all'Ateco 2007

Nella stesura dei file di *parsing* si sono seguiti i seguenti criteri di carattere generale:

1. Nell'ambito del **Phrase processing** ovvero del trattamento che riguarda le stringhe sono state risolte: abbreviazioni problematiche, preposizioni semplici che risultavano determinanti ai fini dell'attribuzione del codice, la congiunzione "ed", l'articolo indeterminativo "un" e la gestione di alcune sigle.

Le abbreviazioni problematiche ovvero quelle che sottointendono parole rare, prima che queste vengano esplose nelle RWRD nelle parole più frequenti ad esempio *Inf. in alluminio → Infissi in alluminio*. La trasformazione più frequente *Inf=Informatico* verrà di seguito risolta nelle RWRD.

Si è ricorso all'uso delle RSTR nel caso di preposizioni semplici che risultavano determinanti al fine dell'attribuzione del codice e che quindi non devono essere soppresse. L'alternativa di considerarle sempre parole determinanti ai fini della codifica, ovvero di non sostituirle con blank nella RWRD, è stata scartata, perché si è ritenuto che ciò avrebbe pesato negativamente sul tasso di codifica. Per esempio nel caso della preposizione "*in"* abbinata a "*tessuto"* c'era la necessità di differenziare i due testi "Commercio dettaglio abbigliamento *in tessuto*" e "Commercio al dettaglio *di tessuti* per abbigliamento". A tal fine nelle RSTR *in tessuto→intessuto* è stata cioè resa parola unica (ciò è stato fatto anche per *in stoffa→ intessuto).*

Relativamente alla congiunzione *"e"* è stato adottato un trattamento diverso a seconda che sia espressa *"e"* oppure *"ed"*. La prima, infatti, è stata sostituita con blank nelle RWRD, la seconda invece, potendo rappresentare anche l'abbreviazione di "*edilizia"*, ha subito un doppio trattamento:

Per l'uso della *"preposizione ed"* davanti alla parole che iniziano per vocale sono state realizzate nelle RSTR le seguenti trasformazioni:

ed a	*. e a*
ed e	*. e e*
ed i	*. e i*
ed o	*. e o*
ed u	*. e u*

Queste schede hanno permesso di uniformare, nel caso della congiunzione con parole che iniziano per vocale, la E con la ED. Successivamente, infatti, nello step delle RWRD la E verrà sostituita con blank.

Anche per l'articolo indeterminativo *"un"* si è reso necessario prevedere più schede, per uniformarlo alla lettera *"u"*che nel passaggio successivo delle RWRD viene sostituita con blank.

Sono state, per esempio, inserite le seguenti trasformazioni

un b	*.u b*
un c	*. u c*

L'uso delle RSTR si rileva ancora molto utile nella gestione delle sigle come ad esempio *P.S. → Poliziastato* oppure *V.F. → Vigile fuoco.* Per le sigle molto brevi e problematiche per esempio *N.U. → Nettezza Urbana* bisogna ricordare di lasciare un blank prima di scrivere la stringa nella prima colonna per evitare che la trasformazione sia effettuata su altre sigle che la contengono ad esempio *O.N.U.* che sarebbe stato trasformato in *→ O.Nettezza urbana,* se non fosse stato lasciato il blank prima della stringa.

Nelle RSTR è molto importante anche l'ordine di successione delle varie trasformazioni da eseguire. Bisogna, pertanto, fare in modo che nella successione delle trasformazioni, le schede più brevi, contenute in un'altra scheda più lunga a sua volta oggetto di trasformazione, vengano sottomesse dopo quelle più lunghe. Chiarendo con un esempio, nella trasformazione che segue è stato stabilito il seguente ordine

c/lavorazioni → .contolavorazione.
c/lavorazione → .contolavorazione.
c/lavorazion → .contolavorazione.
c/lavoraz → .contolavorazione.
c/lav → .contolavorazione.
c/l → .contolavorazione.

Infatti se *c/lav → .contolavorazione* fosse stata la prima dell'elenco, *c/lavorazione* sarebbe stata trasformata in *contolavorazione. orazione.*

Si ricorda che l'uso delle RSTR, oltre che essere molto laborioso, non sempre può risultare risolutivo, in quanto rimane assolutamente legato alla stringa prevista che invece può presentare

una variabilità elevatissima.

2. Nell'ambito del **Word Processing** ovvero del trattamento delle parole sono state risolte le problematiche elencate di seguito.

Tutte le parole che subiscono una trasformazione nelle DWRD sono state convertite al maschile singolare nella fase delle RWRD come per esempio *abrasiva → abrasivo.* Ciò è stato ritenuto necessario per permettere di gestire nelle DWRD un numero minore di sinonimi e quindi di schede. È stato necessario prevedere alcune eccezioni:

- sostantivi che assumono un significato diverso se usati al femminile/maschile, plurale/singolare, come per esempio *maniche → manica*, mentre *manici → manico*
- le *gomme* non sono state trasformate nel singolare *gomma* perché, dall'esame dei file utilizzati per l'addestramento di ACTR, è emerso che il plurale della parola viene spesso usato come sinonimo di *pneumatici*. Non è stato creato neanche il sinonimo *gomme = pneumatici,* e si è convenuto invece che, quando si inserisce un'empirica che tratta questo prodotto, deve essere inserita sia l'empirica con *gomme* che quella corrispondente con *pneumatici.*

In caso di parole che esprimono categorie (ad esempio *cereali*), sono stati definiti sinonimi nelle RWRD, laddove tutte le specificazioni che in italiano fanno parte della stessa categoria vengono trattate dall'Ateco nelle stesse categorie ad esempio *avena, granturco, mais, orzo, saggina, segale, sorgo → cereale.*

Non è stato possibile invece ricondurre alla categoria degli *elettrodomestici* i prodotti che ne fanno parte, perché la classificazione Ateco distingue per esempio in categorie diverse la produzione di *frigoriferi per uso domestico* da quelli per *uso industriale.*

Abbreviazioni problematiche, ovvero quelle a cui non corrisponde univocamente una singola parola. Dall'esame del file utilizzato per l'addestramento, è stato stabilito di esplicitare l'abbreviazione nello step delle RWRD, soltanto qualora questa sottintenda una parola piuttosto che un'altra in una percentuale di casi molto elevata (criterio della "massima frequenza"). Per esempio, è molto più frequente che:

inf → informatico piuttosto che *Inf → Infisso*

Oppure:

bib → bevanda (bibita, quindi bevanda) piuttosto che *Bib → Biblioteca.*

La gestione delle abbreviazioni che sottintendono parole più rare è stata risolta come detto al punto 1 nelle RSTR.

Le abbreviazioni problematiche a cui non è stato possibile applicare il criterio della "massima frequenza" sono state risolte nelle DWRD, esplicitandole a seconda delle parole con cui sono abbinate. Il troncamento *Cons,* per esempio, può essere ricondotto sia a *Consulenza* che a *Conservazione.* Il trattamento di tali abbreviazioni è stato rimandato, come già detto, nella fase delle DWRD, utilizzando più schede che hanno permesso di trattare le parole in sequenza ad esempio:

cons carne → conservazione carne
cons finanza → consulenza finanza.

Altri troncamenti che hanno comportato un lavoro complesso e non sempre, però, definitivamente risolutivo, vista la grande variabilità di risposta, sono stati sia l'abbreviazione *acq,* che può essere per esempio ricondotta ad *Acquisto, Acqua o Acquedotto,* che l'abbreviazione *conf* che può essere ricondotta sia a *Confezionamento* che a *Confezioni.* Il criterio utilizzato anche in quest'ultimi casi è stato quello di riportare il significato dell'abbreviazione ad una o all'altra in funzione della seconda parola, utilizzando sempre più schede nelle DWRD.

Abbreviazioni che contengono il punto internamente alla parola come per esempio *ambul.te* (che sta per *ambulante)* sono state risolte, quando possibile, nelle DWRD perché l'uso della trasformazione nelle RSTR (necessario invece per quelle che terminano per "*la"* e "*le"*) avrebbe potuto generare trasformazioni sbagliate.

Non è stato possibile eliminare la congiunzione "*ed"*, come invece è stata eliminata la congiunzione *"e",* in quanto spesso "*ed"* è utilizzato come abbreviazione di *edilizia.* Per risolvere questa abbreviazione *"ed"* con *edilizia,* si è proceduto ad inserire nel file di *parsing*, relativo alle DWRD, le possibili combinazioni di parole usate con l'abbreviazione di edilizia come ad esempio:

cantiere	*ed*	*cantiere*	*edilizia*
demolizioni	*ed*	*demolizioni*	*edilizia*
scavo	*ed*	*scavo*	*edilizia*
artigiano	*ed*	*artigiano*	*edilizia*
attività	*ed*	*attività*	*edilizia*
ristrutturazione	*ed*	*ristrutturazione*	*edilizia*
cooperativa	*ed*	*cooperativa*	*edilizia*

Per le parole singole che sottintendono la produzione di un prodotto si è scissa la parola nella *Produzione + Prodotto* nelle RWRD, per esempio *Calzaturificio → Produzione Calzatura, Prosciuttificio → Produzione Prosciutto*

Un'eccezione a questo criterio generale è rappresentata dalle parole *"Panificio e sinonimi"* che sono stati trasformati in una parola unica *"Produzionepane".* Ciò è stato necessario per gestire la codifica dell'attività di *"produzione forni per panifici"* che, in assenza di tale eccezione, sarebbe stata trasformata in "*produzione forno produzione pane"* (quindi soppressa la parola doppia *"produzione")* e attribuita al codice relativo al *"forno produzione di pane".* Per la parola *produzionepane* sono state previste poi nelle DWRD le stesse prevalenze adottate per le doppie attività di *produzione* e *commercio, dettaglio e ingrosso* (cfr. cap. 4).

La parola *panetteria* invece, in quanto può essere usata indifferentemente sia per indicare la *produzione del pane* che la *vendita del pane,* non ha subito nessuna trasformazione.

Per le parole singole che sottintendono già una attività, diversa dalla produzione, si è operato trasformando per esempio *Autotrasportatore → Auto Trasporto;* si è scissa cioè la parola nell'*Attività + Mezzo* nelle RWRD.

Un caso particolare è rappresentato da *Automercato,* sia esso scritto unito, separato o con il trattino. Questo è stato riportato alla parola unica *Commercioauto.* Ciò ha permesso di poter gestire come coppie di parole, al livello di DWRD, le regole di prevalenza rispetto per esempio alla vendita di *Autoricambi.*

Per le parole che possono presentarsi unite, separate dal trattino o doppie, per esempio quelle che indicano un mezzo di trasporto come *Auto Carro*, *Auto Articolato,* per cercare di uniformarne il trattamento, si sono riportate, nelle DWRD, ad un'unica parola *Autocarro.* Lo stesso procedimento si è avuto per il *Super mercato → Supermercato (*comprese le relative abbreviazioni).

Sono state unificate nelle DWRD parole che singolarmente potevano generare falsi match, mentre accoppiate portano necessariamente ad abbinamenti esatti. Ad esempio *Polizia Stato → Poliziastato, Pubblica Sicurezza → Poliziastato.*

Si è tentato di evitare ridondanze, pertanto i casi di parole doppie come per esempio *Commercio Ingrosso → Ingrosso* e *Commercio Dettaglio → Dettaglio* sono stati risolti nelle DWRD, come si nota, con un'unica parola. Per il *Commercio Ambulante* non si è proceduto così perché la parola ambulante non è esclusiva del settore del commercio.

Un lavoro maggiore e un po' più complesso hanno invece richiesto le doppie attività con la definizione della rispettiva prevalenza (cfr. cap. 4). Su indicazione degli esperti della classificazione sono state inserite nei file di *parsing* regole di prevalenza di attività come ad esempio:

Produzione > Commercio
Produzione > Ingrosso
Produzione > Dettaglio.

Tali regole, a meno dei sinonimi tipo *Vendita → Commercio* e *Fabbricazione → Produzione* (gestiti nelle RWRD), sono state tutte inserite nelle DWRD e pertanto sono considerate *regole generali* che vengono sempre rispettate nel passaggio di codifica automatica. Per far sì che tali regole siano sempre eseguite, indipendentemente dalla sequenza delle parole, sono state previste per ogni regola due schede ad esempio:

Produzione Commercio → Produzione
Commercio Produzione → Produzione.

I casi eccezione rispetto alle regole generali sono stati risolti nelle RSTR, prima cioè che le parole vengano trasformate nelle DWRD secondo le prevalenze sopra indicate. Ad esempio nel caso del *Commercio e riparazione di pneumatici* (purché non all'ingrosso) prevale l'attività di *riparazione,* al contrario della regola generale. Quindi si è deciso di trasformare nelle RSTR le coppie di parole *"commercio pneumatici"* e *"riparazione pneumatici"* in parole uniche, (*commerciopneumatici e riparazionepneumatici)* in modo che successivamente, nelle DWRD, *commercio riparazionepneumatici → riparazionepneumatici*. Si è cercato di prevedere nelle RSTR tutta la casistica possibile essendo, come già detto, la trasformazione vincolata esclusivamente alla stringa. Si sono, volutamente, tenute distinte le parole *Gomma* e *Gomme* (intese nel senso di pneumatici). Quest'ultima parola è stata anche inserita tra le parole eccezione (EXCP).

Nelle RWRD sono state anche gestite numerose parole che sono state ritenute ininfluenti ai fini della codifica come per esempio *Affine, Annesso…..* che sono state riportate a blank, come le preposizioni semplici e articolate.

Nelle EXCP sono state gestite numerose parole che, per il loro significato non unico, potevano comportare dei match non corretti e non univoci come per esempio *Confezioni* che volutamente è rimasta distinta dalla *Confezione → Confezionamento* equiparata pertanto all'*Abbigliamento.*

La Strategia utilizzata nel contesto di codifica dell'Ateco 2007

Come già detto nella premessa, nel *Parsing strategy file* è possibile specificare quali fasi della standardizzazione debbono essere eseguite ed in quale ordine. Nell'applicazione sviluppata per l'Ateco 2007, la strategia utilizzata prevede i seguenti processi così ordinati come mostra la tavola:

Tavola 1- Processi utilizzati nella strategia di parsing 2007

PhraseProcess1	DCLS
PhraseProcess2	DSTR
PhraseProcess3	RSTR
WordProcess1	RWRD
WordProcess2	DWRD
WordProcess3	DWRD
WordProcess4	DWRD
WordProcess5	EXCP
WordProcess6	SUFX
WordProcess7	
WordProcess8	
PostProcess1	RDUP
PostProcess2	SORT
Autotrim	Yes
TrimLeft	)]}-/\!?,.
TrimRight	([{-/$
Hyphen	-

Analizzando i processi esplicitati nella Tavola si vede che il file delle HWRD (Hyphenated Words), che consente il trattamento di parole separate dal trattino, non è stato inserito. È bastato, infatti, non inserire il carattere *hyphen* (-) nel file WCHR (Word Characters) dove viene eseguita la traduzione dei caratteri. Nella fase di separazione in parole, le parole scritte col trattino vengono trattate, pertanto, come due parole singole. Ne consegue che è stato sufficiente far diventare parola unica le parole originariamente scritte con trattino nel file DWRD, tenendo presente eventuali trasformazioni già avvenute a livello di singola parola nel RWRD.

Altra peculiarità di questa applicazione è stata quella di far girare tre volte il processo delle DWRD (Double words). Ciò si è reso necessario per consentire una trasformazione corretta di descrizioni che contenevano al loro interno clausole di esclusione e che non potevano essere abolite con il processo DCLS, ma che, d'altro canto, se non trattate, avrebbero rischiato di generare match non corretti. Un esempio per chiarire, è l'attività economica: *"estrazione di minerali metallici non ferrosi, (escluso i minerali di uranio e di torio)"*. Era necessario, pur mantenendo la clausola di esclusione, far sì che le singole parole "minerali", "uranio" e "torio" non generassero match con altre attività che includevano l'estrazione di minerali non ferrosi.

Con questa strategia di *parsing*, la stringa originaria ha subito le seguenti trasformazioni:

Original Text:	estrazione di minerali metallici non ferrosi, (escluso i minerali di uranio e di torio)
String trimming	estrazione di minerali metallici non ferrosi, (escluso i minerali di uranio e di torio)
Word Characters (Translation)	estrazione di minerali metallici non ferrosi, (escluso i minerali di uranio e di torio)
Deletion Clauses	estrazione di minerali metallici non ferrosi, (escluso i minerali di uranio e di torio)
Deletion Strings	estrazione di minerali metallici non ferrosi, (escluso i minerali di uranio e di torio)
Replacement Strings	estrazione di minerali metallici non ferrosi, (escluso i minerali di uranio e di torio)
Word Characters (Elimination)	estrazione di minerali metallici non ferrosi escluso i minerali di uranio e di torio
Replacement Words	estrazione minerale metallo non ferro non minerale uranio torio
Double Words	estrazione minerale metallo nonferro non uranio torio
Double Words	estrazione minerale metallo nonferro nonuranio torio
Double Words	estrazione minerale metallo nonferro nonuranio nontorio
Exception Words	estrazione minerale metallo nonferro nonuranio nontorio
Suffixes	estrazion mineral metall nonferr nonuran nontor
Duplicate Word Removal	estrazion mineral metall nonferr nonuran nontor
Word sortine	estrazion metall mineral nonferr nontor nonuran
Parsed Text:	estrazion metall mineral nonferr nontor nonuran

Dalla tavola si evince ancora che i file MCHR (Multiple characters) e PRFX (Prefixes) non sono stati utilizzati nella strategia per la codifica della variabile Ateco 2007 in quanto non ritenuti utili per questo contesto.

Bibliografia

Bolasco S. *L'analisi multidimensionale dei dati,* Roma, Carocci ed., pag. 179-248, 1999.

Cuccia F., De Angelis S., Laureti A., Macchia S., Mastroluca S., Perrone D. *"La codifica delle variabili testuali nel Censimento generale della popolazione",* collana Documenti Istat (n. 1/2005).

De Angelis R. et al. "Applicazioni sperimentali della codifica automatica: analisi di qualità e confronto con la codifica manuale", *Rivista di statistica ufficiale - Quaderni di ricerca Istat* n. 1 (2000).

Eurostat. *Nace rev. 2. Introductory Guidelines,* a cura della divisione Statistical governance, quality and evaluation, Statistical office of european communities, 2007.

D'Orazio M., Macchia S. "A system to monitor the quality of automated coding of textual answers to open questions", *Research In Official Statistics ROS*, n. 2 (2002).

Ferrillo A. Codifica automatica della variabile Ateco - Metodologia per l'aggiornamento dell'ambiente di codifica ACTR rispetto alla classificazione delle attività economiche Ateco 2002, documento interno Istat, Ottobre 2004.

Knaus R. "Methods and problems in coding natural language survey data", *Journal of Official Statistics*, vol. 1 (1987): 45-67.

Lyberg L., Dean P. "Automated Coding of Survey Responses: an international review", in *Conference of European Statisticians, Work session on Statistical Data Editing*, Washington DC, 1992.

Macchia S. et al. *Sperimentazione, implementazione e gestione dell'ambiente di codifica automatica della classificazione delle Attività Economiche*, collana Documenti Istat (n. 2/2002).

Macchia S. et al. Metodi e software per la codifica automatica e assistita dei dati, collana Tecniche e strumenti Istat (n. 4/2007).

Mazza L., Murgia M. "Efficiency and accuracy of an innovative automated coding system based on different approaches of textual analysis", 7th International Conference on Social Science Methodology – Università Federico II, Napoli (1-5 settembre 2008).

Regolamento Ce n. 1893/2006 del Parlamento Europeo e del Consiglio del 20 dicembre 2006 pubblicato sulla Gazzetta ufficiale Ce L 393 del 30 dicembre 2006.

Regolamento (Cee) n. 3696/93 del Consiglio, del 29 ottobre 1993, relativo alla classificazione statistica dei prodotti associata alle attività nella Comunità economica europea, G.U. n. L 342 del 31 dicembre 1993.

Tourigny, J.Y., Moloney J. "The 1991 Canadian Census of Population experience with automated coding", in *United Nations Statistical Commission*, *Statistical Data Editing,* 2, 1995.

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

Metodi e Norme - Nuova serie - Volumi pubblicati

Anno 2000

6. *L'indice del costo della vita valevole ai fini dell'applicazione della scala mobile delle retribuzioni. Dalle origini alla cessazione (1945-97)*
7. *Le nuove stime dei consumi finali delle famiglie secondo il Sistema Europeo dei Conti SEC95*

Anno 2001

8. *La nuova indagine sulle cause di morte. La codifica automatica, il bridge coding e altri elementi innovativi*
9. *Il settore delle costruzioni in contabilità nazionale. I nuovi standard europei dettati dal SEC95*
10. *Indagini sociali telefoniche. Metodologia ed esperienze della statistica ufficiale*
11. *Elenco dei comuni al 31 maggio 2001* 🖫
12. *Classificazione delle professioni* O

Anno 2002

13. *Le statistiche culturali in Europa*
14. *Gli investimenti lordi di contabilità nazionale dopo la revisione: nota metodologica*
15. *Panel Europeo sulle famiglie*

Anno 2003

16. *Metodi statistici per il record linkage*
17. *Metodologia e organizzazione dell'indagine multiscopo sulla domanda turistica "Viaggi e vacanze"*
18. *Classificazione delle attività economiche. Ateco 2002*

Anno 2004

19. *Inventario sulle fonti e metodi di calcolo per le valutazioni a prezzi costanti - Italia*
20. *Metodologia e tecniche di tutela della riservatezza nel rilascio di informazione statistica*
21. *Metodologia di stima degli aggregati di contabilità nazionale a prezzi correnti*
22. *Numeri indici dei prezzi alla produzione dei prodotti industriali venduti sul mercato interno - Base 2000=100*

Anno 2005

23. *I conti economici nazionali per settore istituzionale: le nuove stime secondo il Sec 95* O
24. *La rete di intervistatori Capi dell'Istat per la conduzione dell'indagine continua sulle Forze di Lavoro*
25. *Il monitoraggio del processo e la stima dell'errore nelle indagini telefoniche*
26. *Classificazione delle forme giuridiche delle unità legali*

Anno 2006

27. *Gli stranieri nella rilevazione continua sulle forze di lavoro*
28. *L'indagine campionaria sulle nascite: obiettivi, metodologia e organizzazione*
29. *Rilevazione mensile sull'occupazione, gli orari di lavoro e le retribuzioni nelle grandi imprese*
30. *La classificazione Istat dei titoli di studio italiani. Anno 2003* O
31. *Il sistema di indagini sociali multiscopo. Contenuti e metodologia delle indagini*
32. *La rilevazione sulle forze di lavoro: contenuti, metodologie, organizzazione*
33. *Il calcolo della spesa pubblica per la protezione dell'ambiente – Linee guida per riclassificare i rendiconti delle amministrazioni pubbliche*

Anno 2007

34. *Come si progetta il monitoraggio del lavoro sul campo di un'indagine sulle famiglie* O
35. *Istruzioni integrative per l'applicazione della Icd-10 nella codifica delle cause di morte* O

Anno 2008

36. *La progettazione e lo sviluppo informatico del sistema CAPI sulle forze di lavoro*
37. *L'indagine europea sui redditi e le condizioni di vita delle famiglie (Eu-Silc)*

Anno 2009

38. *Integrazione di dati campionari Eu-Silc con dati di fonte amministrativa*
39. *La misura della povertà assoluta*
40. *Classificazione delle attività economiche. Ateco 2007 - Derivata dalla Nace Rev. 2*
41. *L'ambiente di codifica automatica dell'Ateco 2007 - Esperienze effettuate e prospettive*

🖫 dati forniti su floppy disk
O dati forniti su cd-rom

Istat

Produzione editoriale e altri servizi

Le pubblicazioni a carattere generale

Annuario statistico italiano 2009
pp. XXIV+860+1 cd-rom; € 50,00
ISBN 978-88-458-1618-5

Bollettino mensile di statistica
pp. 116 circa; € 15,00
ISSN 0021-3136

Compendio statistico italiano 2008
Italian Statistical Abstract 2008
pp. 368; € 15,00
ISBN 978-88-458-1608-6

Rapporto annuale.
La situazione del Paese nel 2008
pp. XVI+412; € 30,00
ISBN 978-88-458-1617-8
ISSN 1594-3135

Rivista di statistica ufficiale
n. 2-3/2007
pp. 90; € 10,00
ISSN 1828-1982

Le novità editoriali a carattere tematico

AMBIENTE E TERRITORIO

Atlante di geografia statistica e amministrativa (*)
Edizione 2009
pp. 268+1 cd-rom; € 30,00
ISBN 978-88-458-1609-3

Atlante statistico territoriale delle infrastrutture
Indicatori statistici, n. 6, edizione 2008
pp. 272+1 cd-rom; € 28,00
ISBN 978-88-458-1580-5

Statistiche ambientali
Annuari, n. 10, edizione 2008
pp. 618+1 cd-rom; € 50,00
ISBN 978-88-458-1591-1

POPOLAZIONE

Evoluzione e nuove tendenze dell'instabilità coniugale (*)
Argomenti, n. 34, edizione 2008
pp. 164; € 18,00
ISBN 978-88-458-1582-9

Popolazione e movimento anagrafico dei comuni
anno 2005
Annuari, n. 18, edizione 2008
pp. 236+1 cd-rom; € 28,00
ISBN 978-88-458-1578-2

SANITÀ E PREVIDENZA

I bilanci consuntivi degli enti previdenziali (*)
anno 2007
Informazioni, n. 3, edizione 2009
pp. 104+1 cd-rom; € 22,00
ISBN 978-88-458-1625-3

Statistiche della previdenza e dell'assistenza sociale (*)
I - I trattamenti pensionistici
anno 2006
Annuari, n. 11, edizione 2008
pp. 132+1 cd-rom; € 20,00
ISBN 978-88-458-1607-9

Statistiche della previdenza e dell'assistenza sociale (*)
II - I beneficiari delle prestazioni pensionistiche - Anno 2006
Annuari, n. 12, edizione 2009
pp. 124+1 cd-rom; € 22,00
ISBN 978-88-458-1616-1

CULTURA

Spettacoli, musica e altre attività del tempo libero (*)
anno 2006
Informazioni, n. 6, edizione 2008
pp. 228+1 cd-rom; € 28,00
ISBN 978-88-458-1599-7

Statistiche culturali
anno 2007
Annuari, n. 47, edizione 2009
pp. 164+1 cd-rom; € 25,00
ISBN 978-88-458-1622-2

L'uso dei media e del cellulare in Italia (*)
anno 2006
Informazioni, n. 2, edizione 2008
pp. 292+1 cd-rom; € 28,00
ISBN 978-88-458-1579-9

FAMIGLIA E SOCIETÀ

Conciliare lavoro e famiglia (*)
Una sfida quotidiana
Argomenti, n. 33, edizione 2008
pp. 264; € 22,00
ISBN 978-88-458-1573-7

I consumi delle famiglie
anno 2007
Annuari, n. 14, edizione 2009
pp. 176+1 cd-rom; € 25,00
ISBN 978-88-458-1621-5

Evoluzione e nuove tendenze dell'instabilità coniugale (*)
Argomenti, n. 34, edizione 2008
pp. 164; € 18,00
ISBN 978-88-458-1582-9

Le pubblicazioni con (*) sono riportate in più settori editoriali

L'indagine europea sui redditi e le condizioni di vita delle famiglie (Eu-Silc)
Metodi e norme, n. 37, edizione 2008
pp. 188; € 18,00
ISBN 978-88-458-1596-6

Integrazione di dati campionari Eu-Silc con dati di fonte amministrativa
Metodi e norme, n. 38, edizione 2009
pp. 122; € 17,00
ISBN 978-88-458-1612-3

La misura della povertà assoluta
Metodi e norme, n. 39, edizione 2009
pp. 98; € 15,00
ISBN 978-88-458-1613-0

Spettacoli, musica e altre attività del tempo libero (*)
anno 2006
Informazioni, n. 6, edizione 2008
pp. 228+1 cd-rom; € 28,00
ISBN 978-88-458-1599-7

Gli stranieri nel mercato del lavoro (*)
I dati della rilevazione sulle forze di lavoro in un'ottica individuale e familiare
Argomenti, n. 36, edizione 2008
pp. 158; € 18,00
ISBN 978-88-458-1605-5

Time Use in Daily Life
A Multidisciplinary Approach to the Time Use's Analysis
Argomenti, n. 35, edizione 2008
pp. 332; € 30,00
ISBN 978-88-458-1587-4

L'uso dei media e del cellulare in Italia (*)
anno 2006
Informazioni, n. 2, edizione 2008
pp. 292+1 cd-rom; € 28,00
ISBN 978-88-458-1579-9

I viaggi in Italia e all'estero nel 2006 (*)
Informazioni, n. 2, edizione 2009
pp. 96+1 cd-rom; € 17,00
ISBN 978-88-458-1620-8

La vita quotidiana nel 2007
Informazioni, n. 10, edizione 2008
pp. 248+1 cd-rom; € 30,00
ISBN 978-88-458-1606-2

PUBBLICA AMMINISTRAZIONE

Atlante di geografia statistica e amministrativa (*)
Edizione 2009
pp. 268+1 cd-rom; € 30,00
ISBN 978-88-458-1609-3

I bilanci consuntivi degli enti previdenziali (*)
anno 2007
Informazioni, n. 3, edizione 2009
pp. 104+1 cd-rom; € 22,00
ISBN 978-88-458-1625-3

Le cooperative sociali in Italia
anno 2005
Informazioni, n. 4, edizione 2008
pp. 144+1 cd-rom; € 22,00
ISBN 978-88-458-1588-1

Finanza locale: entrate e spese dei bilanci consuntivi (comuni, province e regioni)
anno 2005
Annuari, n. 12, edizione 2008
pp. 128+1 cd-rom; € 20,00
ISBN 978-88-458-1593-5

Le fondazioni in Italia
anno 2005
Informazioni, n. 1, edizione 2009
pp. 150; € 25,00
ISBN 978-88-458-1611-6

Statistiche della previdenza e dell'assistenza sociale (*)
I - I trattamenti pensionistici - Anno 2006
Annuari, n. 11, edizione 2008
pp. 132+1 cd-rom; € 20,00
ISBN 978-88-458-1607-9

Statistiche della previdenza e dell'assistenza sociale (*)
II - I beneficiari delle prestazioni pensionistiche - Anno 2006
Annuari, n. 12, edizione 2009
pp. 124+1 cd-rom; € 22,00
ISBN 978-88-458-1616-1

GIUSTIZIA E SICUREZZA

L'attività notarile
Dieci anni della nuova indagine 1997-2006
Informazioni, n. 4, edizione 2009
pp.66+1 cd-rom; € 17,00
ISBN 978-88-458-1626-0

Evoluzione e nuove tendenze dell'instabilità coniugale (*)
Argomenti, n. 34, edizione 2008
pp. 164; € 18,00
ISBN 978-88-458-1582-9

CONTI ECONOMICI

Contabilità nazionale Conti economici nazionali Anni 1996-2007
Annuari, n. 12, edizione 2009
pp. 336+1 cd-rom; € 35,00
ISBN 978-88-458-1615-4

Valore aggiunto ai prezzi di base dell'agricoltura per regione
anni 2002-2007
Informazioni, n. 9, edizione 2008
pp. 200+1 cd-rom; € 23,00
ISBN 978-88-458-1602-4

LAVORO

Classificazione delle attività economiche - Ateco 2007 (*)
Derivata dalla Nace Rev. 2
Metodi e norme, n. 40, edizione 2009
pp. 656; € 43,00
ISBN 978-88-458-1614-7

Conciliare lavoro e famiglia (*)
Una sfida quotidiana
Argomenti, n. 33, edizione 2008
pp. 264; € 22,00
ISBN 978-88-458-1573-7

Forze di lavoro - Media 2007
Annuari, n. 13, edizione 2008
pp. 216+1 cd-rom; € 28,00
ISBN 978-88-458-1604-8

Lavoro e retribuzioni
anni 2005-2006
Annuari, n. 9, edizione 2009
pp. 200+1 cd-rom; € 25,00
ISBN 978-88-458-1610-9

La progettazione e lo sviluppo informatico del sistema Capi sulle forze di lavoro
Metodi e norme, n. 36, edizione 2008
pp. 100; € 15,00
ISBN 978-88-458-1594-2

Statistiche della previdenza e dell'assistenza sociale (*)
I - I trattamenti pensionistici - Anno 2006
Annuari, n. 11, edizione 2008
pp. 132+1 cd-rom; € 20,00
ISBN 978-88-458-1607-9

Statistiche della previdenza e dell'assistenza sociale (*)
II - I beneficiari delle prestazioni pensionistiche - Anno 2006
Annuari, n. 12, edizione 2009
pp. 124+1 cd-rom; € 22,00
ISBN 978-88-458-1616-1

Le pubblicazioni con (*) sono riportate in più settori editoriali

Gli stranieri
nel mercato del lavoro (*)
I dati della rilevazione sulle forze di lavoro in un'ottica individuale e familiare
Argomenti, n. 36, edizione 2008
pp. 158; € 18,00
ISBN 978-88-458-1605-5

PREZZI

Il valore della moneta in Italia dal 1861 al 2007
Informazioni, n. 8, edizione 2008
pp. 170; € 18,00
ISBN 978-88-458-1601-7

AGRICOLTURA E ZOOTECNIA

Le Statistiche agricole verso il Censimento del 2010: valutazioni e prospettive
Atti del Convegno
ottobre 2006
pp. 456; € 33,00
ISBN 978-88-458-1592-8

INDUSTRIA E SERVIZI

Classificazione delle attività economiche - Ateco 2007 (*)
Derivata dalla Nace Rev. 2
Metodi e norme, n. 40, edizione 2009
pp. 656; € 43,00
ISBN 978-88-458-1614-7

Statistiche dei trasporti
anno 2004
Annuari, n. 5, edizione 2007
pp. 280; € 22,00
ISBN 978-88-458-1545-4

Statistiche sull'innovazione nelle imprese
anni 2002-2004
Informazioni, n. 1, edizione 2008
pp. 192; € 18,00
ISBN 978-88-458-1577-5

I viaggi in Italia e all'estero nel 2006 (*)
Informazioni, n. 2, edizione 2009
pp. 96+1 cd-rom; € 17,00
ISBN 978-88-458-1620-8

COMMERCIO ESTERO

Commercio estero e attività internazionali delle imprese
Annuario Istat-ICE 2008
1. Merci, servizi, investimenti diretti
2. Paesi, settori, regioni
L'Italia nell'economia internazionale
Rapporto ICE 2008-2009
Sintesi del Rapporto ICE 2008-2009
Annuari, n. 11, edizione 2009
pp. 360+432+344+48 + 1 cd-rom
€ 100,00 *(in cofanetto)*
ISBN 978-88-458-1623-9

Altri prodotti e servizi

ABBONAMENTI E PRENOTAZIONI 2010

L'offerta per l'acquisizione automatica delle pubblicazioni edite dall'Istat nel 2010 si articola in due modalità: abbonamenti e prenotazioni.
Il sistema degli abbonamenti prevede due tipologie "Generale" e "Tutti i settori".
L'abbonamento all'area "Generale" comprende l'Annuario statistico italiano, gli 11 fascicoli del Bollettino mensile di statistica, il Rapporto annuale e il Compendio statistico italiano nella versione bilingue.
L'abbonamento "Tutti i settori" comprende l'invio di tutta la produzione editoriale 2010 ad esclusione dei volumi appartenenti alle collane *Tecniche e strumenti*, *Essays*, *Quaderni del Mipa* e *Censimenti.*
Gli utenti interessati alla produzione editoriale relativa a singoli settori potranno attivare **una prenotazione** dei volumi. In tal modo riceveranno le pubblicazioni non appena queste si renderanno disponibili e, per ogni invio, riceveranno una fattura con uno sconto del 20% sul prezzo di copertina e non verranno applicate le spese di spedizione.

I coupon sono anche scaricabili dal sito
www.istat.it/servizi/abbonamenti

Ulteriori informazioni possono essere richieste a:

Istat
Direzione centrale comunicazione
ed editoria - EDI/D

Via Cesare Balbo, 16 - 00184 ROMA
Tel. 06.4673.3278-3280-3267 - Fax 06.4673.3477

e-mail: editoria.acquisti@istat.it

WWW.ISTAT.IT

Nel sito Internet è possibile informarsi sulla produzione editoriale più recente, richiedere prodotti e servizi offerti dall'Istat, leggere e prelevare i comunicati stampa, accedere alle banche dati, collegarsi con altri siti nazionali e internazionali.

CATALOGO ON LINE

Dalla home page del sito Internet è possibile collegarsi con il catalogo on line, che contiene l'elenco completo delle pubblicazioni edite dall'Istat a partire dall'anno 2000. Attraverso questo utile strumento è possibile effettuare la ricerca del volume per titolo, per settore, per collana, per anno di edizione e per codice ISBN. Ogni pubblicazione è presentata attraverso una scheda che riporta, oltre alle caratteristiche tecniche, anche una breve descrizione del prodotto. Molti dei volumi presenti in questo catalogo sono scaricabili gratuitamente.

CONT@CT CENTRE

Dal sito Internet è possibile ricevere informazioni su dati e pubblicazioni Istat, avere assistenza nella ricerca delle statistiche ufficiali europee e supporto nella individuazione delle metodologie e classificazioni ufficiali comunitarie (Eurostat). Solo dopo essersi registrati compilando l'apposito *form* è possibile richiedere i seguenti servizi: certificazioni prezzi e retribuzioni, dati elementari per uffici Sistan, collezioni campionarie di dati elementari (file standard), dati censuari e cartografici, abbonamenti e dati del commercio estero, ricerche storiche e bibliografiche, elaborazioni personalizzate. Inoltre ai giornalisti è dedicata un'area speciale per rispondere alle richieste di dati, pubblicazioni e approfondimenti su particolari tematiche.

Acquisto pubblicazioni

Istat

Inviare questo modulo via **fax** al numero **06.4673.3477** oppure spedire in **busta chiusa** a:
Istituto Nazionale di Statistica, DCCE, Commercializzazione dei prodotti
Via Cesare Balbo, 16 – 00184 Roma

Per ulteriori informazioni telefonare al numero 06 4673.3267

Desidero ricevere le seguenti pubblicazioni

Titolo	Codice ISBN	Prezzo

Per un totale di euro _____

Eventuale sconto (1) euro _____

Contributo spese di spedizione euro _____ **6,00**_____

Importo da pagare euro _____

(1) il Sistan, gli Enti pubblici, le Biblioteche e le Università usufruiscono di uno **sconto del 10%** se acquistano direttamente dall'Istat. Per tutti gli utenti che acquistano oltre 20 volumi è previsto uno **sconto del 20%**.

DATI PER LA FATTURAZIONE

Ente/Cognome e Nome _____

Referente _____

Cod.fiscale* | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | P.IVA* | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ |

Indirizzo _____ Cap _____ Città _____

Prov._____ tel._____ fax _____ e-mail_____

* è necessario inserire sia il codice fiscale sia la partita IVA

DESTINATARIO DEI PRODOTTI (se diverso da quanto indicato nei dati per la fatturazione)

Ente/Cognome e Nome _____

Indirizzo _____ Cap _____ Città _____

Prov._____ tel._____ fax _____ e-mail_____

MODALITÀ DI PAGAMENTO. L'importo dovrà essere versato dall'acquirente, dopo il ricevimento della fattura, sul c/c postale n. 619007, oppure con bonifico bancario c/o la Banca Nazionale del Lavoro, indicando con chiarezza il numero, la data della fattura e il codice cliente. Per i versamenti tramite bonifico bancario le coordinate sono: c/c n. 218050, ABI 01005.8, CAB 03382.9; via swift: B.N.L.I. IT RR, codice CIN K, codice anagrafico 63999228/j; IBAN IT64K0100503382000000218050.

INFORMATIVA - I dati da Lei forniti saranno utilizzati esclusivamente per l'esecuzione dell'ordine e per l'invio, da parte dell'Istat, di promozioni commerciali, senza alcun impegno da parte Sua. Il trattamento dei dati avverrà nell'assoluto rispetto del d.lgs. 196/2003, esclusivamente ad opera dei dipendenti dell'Istituto incaricati. Il titolare dei dati è l'Istituto nazionale di statistica, Via Cesare Balbo n. 16, 00184 Roma; responsabile del trattamento dei dati è il Direttore centrale comunicazione ed editoria, anche per quanto riguarda l'esercizio dei diritti dell'interessato di cui all'articolo 7 del d.lgs. n. 196/2003. In qualsiasi momento potrà far modificare o cancellare i Suoi dati indirizzando la richiesta a Istat, DCCE, Commercializzazione dei prodotti, Via Cesare Balbo n. 16, 00184 Roma, oppure via e-mail all'indirizzo editoria.acquisti@istat.it, o inviando un fax al numero 064673.3477.

Data _____ Firma _____ **PV10**

Abbonamenti 2010

Istat

Inviare questo modulo via **fax** al numero **06.4673.3477** oppure spedire in **busta chiusa** a:
Istituto Nazionale di Statistica, DCCE, Commercializzazione dei prodotti
Via Cesare Balbo, 16 – 00184 Roma

Per ulteriori informazioni telefonare ai numeri 06 4673.3278-3280-3267

Desidero sottoscrivere i seguenti abbonamenti per l'anno 2010	**ITALIA**	**ESTERO**
Generale (Bollettino mensile di statistica, Annuario statistico italiano, Rapporto annuale e Compendio statistico italiano)	☐ euro 180,00	☐ euro 200,00
Tutti i settori (escluso Censimenti)	☐ euro 700,00	☐ euro 800,00
Eventuale sconto [1]		
Importo da pagare		

[1] il Sistan, gli Enti pubblici, le Biblioteche e le Università usufruiscono di uno **sconto del 10%** soltanto se sottoscrivono l'abbonamento direttamente con l'Istat.

DATI PER LA FATTURAZIONE

Ente/Cognome e Nome _____

Referente _____

Cod.fiscale* | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | P.IVA* | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ |

Indirizzo _____ Cap _____ Città _____

Prov._____ tel._____ fax _____ e-mail_____

* è necessario inserire sia il codice fiscale sia la partita IVA

DESTINATARIO DEI PRODOTTI (se diverso da quanto indicato nei dati per la fatturazione)

Ente/Cognome e Nome _____

Indirizzo _____ Cap _____ Città _____

Prov._____ tel._____ fax _____ e-mail_____

DESTINATARIO DELLA FATTURA (se diverso da quanto indicato nei dati per la fatturazione)

Ente/Cognome e Nome _____

Indirizzo _____ Cap _____ Città _____

Prov._____ tel._____ fax _____ e-mail_____

MODALITÀ DI PAGAMENTO. L'importo dovrà essere versato dall'acquirente, dopo il ricevimento della fattura, sul c/c postale n. 619007, oppure con bonifico bancario c/o la Banca Nazionale del Lavoro, indicando con chiarezza il numero, la data della fattura e il codice cliente. Per i versamenti tramite bonifico bancario le coordinate sono: c/c n. 218050, ABI 01005.8, CAB 03382.9; via swift: B.N.L.I. IT RR, codice CIN K, codice anagrafico 63999228/j; IBAN IT64K0100503382000000218050.

INFORMATIVA - I dati da Lei forniti saranno utilizzati esclusivamente per l'esecuzione dell'ordine e per l'invio, da parte dell'Istat, di promozioni commerciali, senza alcun impegno da parte Sua. Il trattamento dei dati avverrà nell'assoluto rispetto del d.lgs. 196/2003, esclusivamente ad opera dei dipendenti dell'Istituto incaricati. Il titolare dei dati è l'Istituto nazionale di statistica, Via Cesare Balbo n. 16, 00184 Roma; responsabile del trattamento dei dati è il Direttore centrale comunicazione ed editoria, anche per quanto riguarda l'esercizio dei diritti dell'interessato di cui all'articolo 7 del d.lgs. n. 196/2003. In qualsiasi momento potrà far modificare o cancellare i Suoi dati indirizzando la richiesta a Istat, DCCE, Commercializzazione dei prodotti, Via Cesare Balbo n. 16, 00184 Roma, oppure via e-mail all'indirizzo editoria.acquisti@istat.it, o inviando un fax al numero 064673.3477.

Data _____ Firma _____ **PV10**

Prenotazioni per settore 2010

Istat

Inviare questo modulo via **fax** al numero **06.4673.3477** oppure spedire in **busta chiusa** a:
Istituto Nazionale di Statistica, DCCE, Commercializzazione dei prodotti
Via Cesare Balbo, 16 – 00184 Roma

Per ulteriori informazioni telefonare ai numeri 06 4673.3278-3280-3267

Desidero ricevere **una copia** di tutti i volumi **edizione 2010** relativi al/ai settore/i sottoindicato/i non appena si renderanno disponibili. Attraverso tale sottoscrizione verrà garantito uno sconto di prenotazione sul prezzo di copertina pari al **20%** e non verranno applicate spese di spedizione. Per ogni invio verrà emessa e inviata una fattura.

- Ambiente e territorio ..... ☐
- Popolazione (escluso Censimenti) ..... ☐
- Sanità e previdenza ..... ☐
- Cultura ..... ☐
- Famiglia e società ..... ☐
- Pubblica amministrazione ..... ☐
- Giustizia e sicurezza ..... ☐
- Conti economici ..... ☐
- Lavoro ..... ☐
- Prezzi ..... ☐
- Agricoltura e zootecnia (escluso Censimenti) ..... ☐
- Industria e servizi (escluso Censimenti) ..... ☐
- Commercio estero ..... ☐

DATI PER LA FATTURAZIONE

Ente/Cognome e Nome _____

Referente _____

Cod.fiscale* | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | P.IVA* | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ | _ |

Indirizzo _____ Cap _____ Città _____

Prov._____ tel._____ fax _____ e-mail_____

* è necessario inserire sia il codice fiscale sia la partita IVA

DESTINATARIO DEI PRODOTTI (se diverso da quanto indicato nei dati per la fatturazione)

Ente/Cognome e Nome _____

Indirizzo _____ Cap _____ Città _____

Prov._____ tel._____ fax _____ e-mail_____

MODALITÀ DI PAGAMENTO. L'importo dovrà essere versato dall'acquirente, dopo il ricevimento della fattura, sul c/c postale n. 619007, oppure con bonifico bancario c/o la Banca Nazionale del Lavoro, indicando con chiarezza il numero, la data della fattura e il codice cliente. Per i versamenti tramite bonifico bancario le coordinate sono: c/c n. 218050, ABI 01005.8, CAB 03382.9; via swift: B.N.L.I. IT RR, codice CIN K, codice anagrafico 63999228/j; IBAN IT64K0100503382000000218050.

INFORMATIVA - I dati da Lei forniti saranno utilizzati esclusivamente per l'esecuzione dell'ordine e per l'invio, da parte dell'Istat, di promozioni commerciali, senza alcun impegno da parte Sua. Il trattamento dei dati avverrà nell'assoluto rispetto del d.lgs. 196/2003, esclusivamente ad opera dei dipendenti dell'Istituto incaricati. Il titolare dei dati è l'Istituto nazionale di statistica, Via Cesare Balbo n. 16, 00184 Roma; responsabile del trattamento dei dati è il Direttore centrale comunicazione ed editoria, anche per quanto riguarda l'esercizio dei diritti dell'interessato di cui all'articolo 7 del d.lgs. n. 196/2003. In qualsiasi momento potrà far modificare o cancellare i Suoi dati indirizzando la richiesta a Istat, DCCE, Commercializzazione dei prodotti, Via Cesare Balbo n. 16, 00184 Roma, oppure via e-mail all'indirizzo editoria.acquisti@istat.it, o inviando un fax al numero 064673.3477.

Data _____ Firma _____ **PV10**

I Centri di informazione statistica

PIÙ INFORMAZIONI. PIÙ VICINE A VOI.

Per darvi più servizi e per esservi più vicino l'Istat ha aperto al pubblico una rete di Centri d'informazione statistica che copre l'intero territorio nazionale. Oltre alla vendita di prodotti informatici e pubblicazioni, i Centri rilasciano certificati sull'indice dei prezzi, offrono informazioni tramite collegamenti con le banche dati del Sistema statistico nazionale (Sistan) e dell'Eurostat (Ufficio di statistica della Comunità europea), forniscono elaborazioni statistiche "su misura" ed assistono i laureandi nella ricerca e selezione dei dati.
Presso i Centri d'informazione statistica, semplici cittadini, studenti, ricercatori, imprese e operatori della pubblica amministrazione troveranno assistenza qualificata e un facile accesso ai dati di cui hanno bisogno. D'ora in poi sarà più facile conoscere l'Istat e sarà più facile per tutti gli italiani conoscere l'Italia. Per gli orari di apertura al pubblico consultare il sito www.istat.it nella pagina "Servizi".

ANCONA Via Castelfidardo, 4
Telefono 071/5013011
Fax 071/5013085

BARI Piazza Aldo Moro, 61
Telefono 080/5789317
Fax 080/5789335

BOLOGNA Galleria Cavour, 9
Telefono 051/6566111
Fax 051/6566185-182

BOLZANO Via Canonico M. Gamper,1
Telefono 0471/418400
Fax 0471/418419

CAGLIARI Via Firenze, 17
Telefono 070/34998700-1
Fax 070/34998732-3

CAMPOBASSO Via G. Mazzini, 129
Telefono 0874/604854-8
Fax 0874/604885-6

CATANZARO Viale Pio X, 116
Telefono 0961/507629
Fax 0961/741240

FIRENZE Lungarno C. Colombo, 54
Telefono 055/6237711
Fax 055/6237735

GENOVA Via San Vincenzo, 4
Telefono 010/584970
Fax 010/5849742

MILANO Via Porlezza, 12
Telefono 02/806132214
Fax 02/806132205

NAPOLI Via G. Verdi, 18
Telefono 081/4930190
Fax 081/4930185

PALERMO Via G. B. Vaccarini, 1
Telefono 091/6751811
Fax 091/6751836

PERUGIA Via Cesare Balbo, 1
Telefono 075/5826411
Fax 075/5826484

PESCARA Via Caduta del Forte, 34
Telefono 085/44120511-2
Fax 085/4216516

POTENZA Via del Popolo, 4
Telefono 0971/377261
Fax 0971/36866

ROMA Via Cesare Balbo, 11/a
Telefono 06/46733102
Fax 06/46733101

TORINO Via Alessandro Volta, 3
Telefono 011/5166758-64-67
Fax 011/535800

TRENTO Via Brennero, 316
Telefono 0461/497801
Fax 0461/497813

TRIESTE Via Cesare Battisti, 18
Telefono 040/6702558
Fax 040/6702599

VENEZIA-MESTRE Corso del Popolo, 23
Telefono 041/5070811
Fax 041/5070835

La biblioteca centrale

È la più ricca biblioteca italiana in materia di discipline statistiche e affini. Il suo patrimonio, composto da oltre 500.000 volumi e 2.700 periodici in corso, comprende fonti statistiche e socio-economiche, studi metodologici, pubblicazioni periodiche degli Istituti nazionali di statistica di tutto il mondo, degli Enti internazionali e dei principali Enti e Istituti italiani ed esteri. È collegata con le principali banche dati nazionali ed estere. Il catalogo informatizzato della biblioteca è liberamente consultabile in rete sul sito Web dell'Istat alla voce Biblioteca (www.istat.it).
Oltre all'assistenza qualificata che è resa all'utenza in sede, è attivo un servizio di ricerche bibliografiche e di dati statistici a distanza, con l'invio dei risultati per posta o via fax, cui i cittadini, gli studenti, i ricercatori e le imprese possono accedere.
È a disposizione dell'utenza una sala di consultazione al secondo piano.

ROMA Via Cesare Balbo, 16 - secondo piano -Telefono 06/4673.2380 Fax 06/4673.2617

https://contact.istat.it//

Orario: da lunedì a giovedì 9.00 - 16.00 venerdì 9.00 - 14.00

Lavoro / Labour

Industria e servizi / Industry and Services

L'ambiente di codifica automatica dell'Ateco 2007

La nuova classificazione delle attività economiche Ateco 2007, in vigore dal 1° gennaio 2008, ha reso necessario l'adeguamento di ACTR, il software di codifica automatica in uso all'Istat già dagli anni Novanta.
Questo volume, frutto dell'attività di un apposito gruppo di lavoro, descrive l'attività svolta per lo sviluppo del software ACTR per la codifica automatica dell'Ateco 2007 a partire dall'aggiornamento, reso necessario con la nuova classificazione, dell'applicazione che codifica le descrizioni delle attività economiche rilevate nelle indagini statistiche. Considerata la diffusione e l'utilizzazione da parte di tutte le fonti amministrative della classificazione Ateco 2007, il software di codifica è stato inoltre messo a disposizione degli utenti del sito Web dell'Istat. Con tale strumento a disposizione, dunque, chiunque può risalire al codice di attività economica descrivendo la propria attività.

Automated Coding Environment of Ateco 2007

Ateco 2007, the new classification of economic activities, active since January 1st 2008, made an update of ACTR necessary, the system for automated coding used at Istat since the 90s.
This book is the result of the activity of a working group and describes the procedures developed to implement the ACTR application for the automated coding of Ateco 2007. It started with updating the coding environment to process the descriptions of economic activities collected in the statistical surveys, according to the new classification.
Due to the fact that all administrative sources make use of the Ateco classification, the software has been made available for users on Istat's website. With this tool users can identify the economic activity code by inserting their own activity in free words.

ISBN 978-88-458-1629-1

1M012009041000000

€ 17,00