Progettazione di una procedura informatica generalizzata per la sperimentazione del metodo ***Microstrat*** di coordinamento della selezione delle imprese soggette a rilevazioni nella realtà Istat.

di

Marco Broccoli, Roberto Di Giuseppe, Daniela Pagliuca

Indice

Introduzione

Nell'ambito del "Gruppo di lavoro interdirezionale per la definizione di una metodologia di coordinamento dei campioni delle indagini sulle imprese al fine di ottimizzare le procedure di selezione e di ridurre il carico sui rispondenti" (nel seguito "gruppo di lavoro"), costituito nel 2004, all'unità *MTS-F Software generalizzati per la produzione statistica* è stata affidata la progettazione e lo sviluppo di una o più procedure generalizzate per il coordinamento dei campioni. Il presente lavoro è stato prodotto dall'unità *MTS-F* per documentare i passi seguiti nello sviluppo della procedura informatica generalizzata implementata per sperimentare il metodo *Microstrat* (Riviere P., 2001), seguendo le linee guida che il suddetto gruppo di lavoro ha delineato durante lo svolgimento delle proprie attività. In particolare viene documentata una prima ipotesi operativa, adattabile ad indagini che estraggono i dati sulle imprese dall'archivio *ASIA*, simulando l'applicazione del metodo *Microstrat* alla realtà dell'Istituto secondo un contesto sperimentale ben definito. Questa procedura rappresenta un primo importante passo ed è stata sviluppata in modo flessibile, al fine di poter essere implementata per successive fasi di analisi, secondo contesti sperimentali più generali, e per adeguarsi a differenti contesti evolutivi dal punto di vista organizzativo.

Con riferimento alla *Direttiva 2004* "Metodi e procedure per il coordinamento dei campioni delle rilevazioni sulle imprese" il presente lavoro rappresenta la documentazione della *Attività* "Progettazione di una o più procedure informatiche generalizzate per l'applicazione del metodo di coordinamento della selezione delle imprese soggette a rilevazioni".

1. Una possibile soluzione generale del problema del coordinamento dei campioni in Istat

Il problema della selezione coordinata dei campioni in Istat potrebbe risolversi, da un punto di vista operativo, con l'implementazione di una procedura software generalizzata adattabile a tutte le indagini che estraggono i dati sulle imprese dall'archivio *ASIA*.

L'implementazione di un software generalizzato per la produzione statistica, allo stato attuale, non può non osservarsi se non calato in un sistema più complesso; in altri termini non è più possibile occuparsi della sola generalizzazione di un software, ma è necessario studiare i flussi che provengono da diversi archivi, in un'ottica di sistema informativo integrato.

Nel seguito si descrive un sistema generale comprendente l'implementazione di un software di *Selezione coordinata*, schematizzato secondo quanto raffigurato in *figura 1*.

Nello schema descritto in *figura 1*, l'archivio *ASCO* (*A*rchivio *S*elezione *CO*ordinata) è quello che contiene le informazioni necessarie all'applicazione del metodo scelto per la selezione coordinata (comprese le relazioni con gli archivi *ASIA* e *SIDI*, già realizzati ed utilizzati in Istat).

E' importante sottolineare che si potrebbe pensare ad *ASCO* sia come sottosistema di *ASIA*, che come archivio indipendente. Per completezza di esposizione nel seguito si descrive l'ipotesi più strutturata, comprendente i flussi informativi provenienti da e verso i due archivi, considerandoli separatamente. Nulla cambia se l'archivio *ASIA* dovesse contenere in se le informazioni di *ASCO* (ed in particolare le tabelle descritte nel successivo *paragrafo 6*); ovviamente lo schema illustrato in *figura* 1 ne risulterebbe semplificato, in quanto non dovrebbe includere alcun flusso o processo intermedio atto ad attivare la *Vista-ASIA*.

Con riferimento alla detta *figura 1* vale che:

- La funzione "*Procedure di aggiornamento ASIA***"** è relativa ad aggiornamenti dell'archivio *ASIA* da parte della struttura che gestisce l'archivio, secondo procedure esterne al sistema che si sta rappresentando (nello schema abbiamo evidenziato il generico "*Fonti di aggiornamento*" presupponendo procedure, documentazione etc. non di interesse per l'analisi in questione). Nello schema appare anche un flusso proveniente dalla funzione "*Gestione dati errati e non rispondenti*", che riguarda eventuali informazioni sulle imprese pervenute alle indagini su dati errati (Ateco, Ragione sociale, indirizzo, …) che potrebbe essere utile memorizzare in *ASIA* (questo flusso, come si specificherà nel seguito, non entra nel processo alla base dell'applicazione generalizzata – "*Selezione coordinata*", in quanto la selezione avviene secondo un piano campionario definito e il coordinamento delle unità si riferisce alla lista teorica dei rispondenti).
- La funzione "*Estrazione periodica Vista-Asia*" è necessaria per tener conto delle modifiche intervenute in *ASIA* e considerare quindi in ogni momento la situazione più aggiornata. In realtà, alla base di questa funzione si è ipotizzato che l'archivio *ASCO* possa essere legato ad *ASIA* tramite una vista logica dei dati (che avverà mediante una query, memorizzata come parametro all'interno del sistema di selezione coordinata). Si possono considerare altre modalità: ad esempio tramite un popolamento delle tabelle necessarie al funzionamento dell'applicazione, con cadenza periodica (ad esempio mensile, trimestrale o annuale).

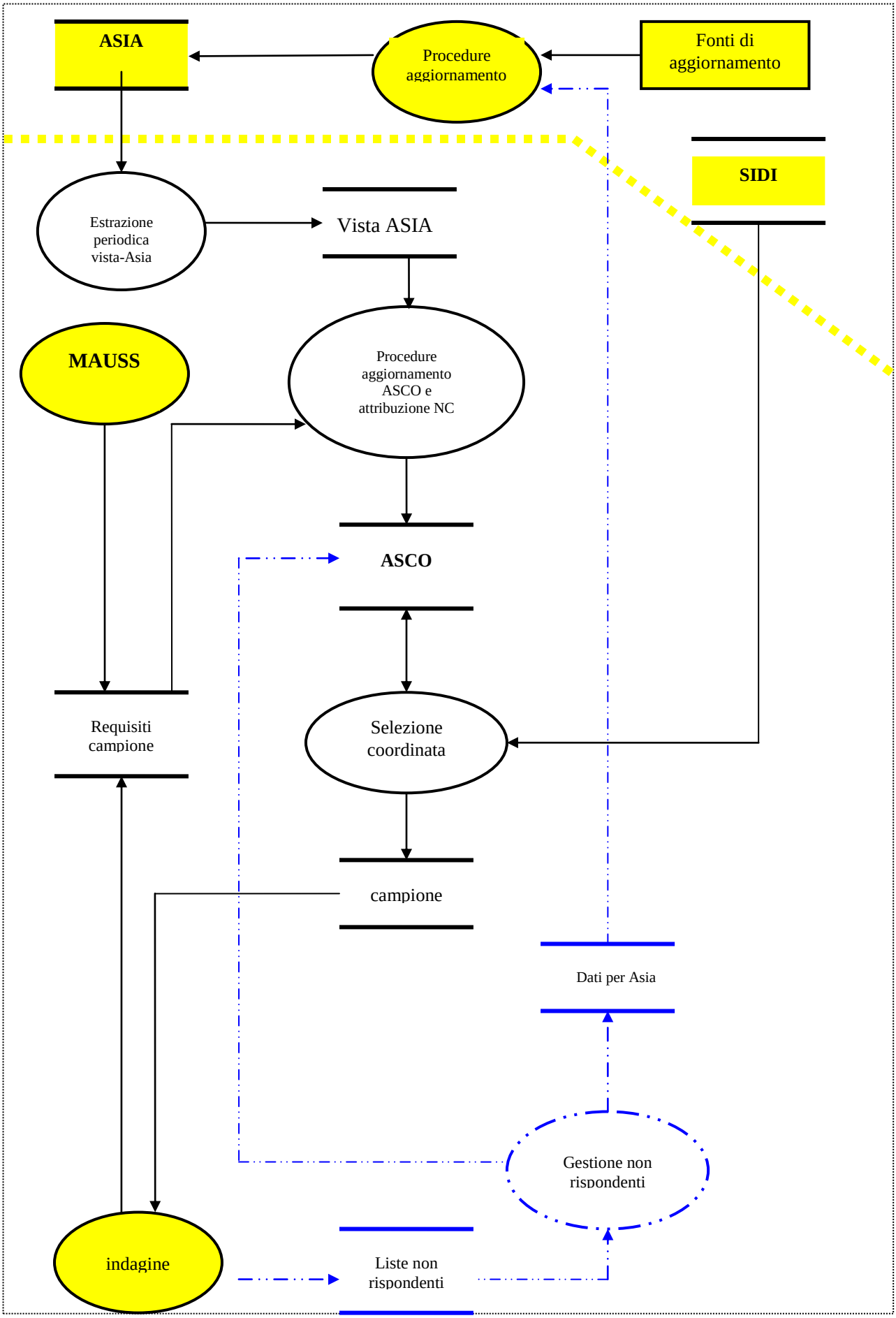
  Tale processo, come già scritto all'inizio del paragrafo, potrebbe risultare superfluo, ove si decidesse di considerare *ASCO* come un sottoinsieme di *ASIA*.

  Nel caso invece si decida di attivare una funzione come questa di estrazione periodica, sarà necessario analizzare approfonditamente alcune questioni, quali quelle riportati nel seguito:

- è necessario decidere la periodicità di tale funzione;
- è opportuno conoscere le condizioni logiche per l'estrazione delle imprese "*attive*";
- occorre gestire le imprese considerando la divisione in unità locali, le plurilocalizzate etc.

- La funzione "*Procedura aggiornamento ASCO e Attribuzione NC*" riguarda l'aggiornamento dell'archivio *ASCO* sulla base delle informazioni correnti di *ASIA* e la gestione dell'attribuzione dei Numeri Casuali (NC) o comunque di un altro identificativo che consenta di gestire il controllo del coordinamento delle unità; ovviamente tale funzione dipende dal metodo di selezione coordinata. Occorre qui definire le attribuzioni successive dei numeri casuali, o degli identificativi, alle imprese "nuove" e la cancellazione "logica" di quelle da considerare *cessate.*

Figura 1


ASIA
Procedure aggiornamento
Fonti di aggiornamento
SIDI
Estrazione periodica vista-Asia
Vista ASIA
MAUSS
Procedure aggiornamento ASCO e attribuzione NC
ASCO
Requisiti campione
Selezione coordinata
campione
Dati per Asia
Gestione non rispondenti
indagine
Liste non rispondenti

- La funzione "*Selezione coordinata*" è la procedura informatica vera e propria alla base del processo di selezione coordinata, ovvero l'implementazione dell'algoritmo relativo al metodo scelto. A seconda della metodologia che si sceglie, l'algoritmo alla base della selezione varia e, ad esempio, potrebbe prevedere il calcolo della molestia statistica[1] o meno.

- La funzione "*Gestione non rispondenti*" è opzionale in quanto non fa parte del vero e proprio processo di selezione. Consentirebbe però di tenere conto di eventuali variazioni (di Ateco, indirizzo, etc.) individuate durante il processo di realizzazione delle indagini e di riportarle nell'archivio *ASIA;* e permetterebbe inoltre di tenere memoria di particolari situazioni riguardanti le unità che non rispondono o che rispondono in modo errato ai questionari (funzione utile per i responsabili di indagine). Come sopra scritto, nel coordinamento della selezione delle imprese non si deve tenere conto di queste informazioni nel rispetto del piano di campionamento ed il flusso che parte da questo processo e arriva ai dati memorizzati in *ASCO* è solo informativo.

ARCHIVI LOGICI:

- *ASIA*: archivio dati amministrativi che memorizza informazioni sulle imprese.
- *SIDI*: memorizza informazioni sulle indagini, compresa eventualmente la molestia statistica.
- *VISTA ASIA*: vista aggiornata di *ASIA* contenente tutte le informazioni sulle imprese necessarie alla estrazione del campione.
- *ASCO* (*A*rchivio *S*elezione *CO*ordinata): conterrà tutte le informazioni necessarie all'applicazione del metodo scelto per la selezione coordinata, comprese le relazioni con gli archivi *ASIA* e *SIDI*.
- *REQUISITI CAMPIONE:* conterrà informazioni sulla popolazione di interesse, la stratificazione, i piani di rotazione… (potrebbe essere parte di *SIDI*); contiene eventualmente le informazioni provenienti dall'uso del software generalizzato MAUSS (Di Giuseppe R., Giaquinto P., Pagliuca D., 2004) o da altre procedure di allocazione campionaria utilizzate nelle indagini.
- *CAMPIONE*: conterrà il campione estratto, per ciascuna indagine e tempo specificati.
- *LISTE NON RISPONDENTI*: informazioni riguardanti le unità non rispondenti o rispondenti in modo errato.

- *DATI PER ASIA*: informazioni riguardanti informazioni sui non rispondenti, dati errati o obsoleti presenti in *ASIA*.

L'obiettivo di realizzare un prodotto integrato con altri sistemi informativi presenti in ISTAT spinge verso una organizzazione dei dati secondo uno schema relazionale.
Altri attori sicuramente (*ASIA*) o probabilmente (*SIDI*) coinvolti nel progetto di selezione coordinata hanno organizzato le loro informazioni in basi dati relazionali, utilizzando l'RDBMS[2] ORACLE. Tali considerazioni porterebbero verso una base dati ORACLE gestita da una applicazione client-server nella quale anche parte della componente dati potrebbe essere distribuita o replicata, in funzione della metodologia applicata.

1.1 Alcune informazioni aggiuntive sugli archivi

A riguardo di tali archivi, si può aggiungere quanto segue:

- ASIA -

Il coinvolgimento dell'archivio *ASIA* deve essere analizzato con precisione. L'ipotesi illustrata nello schema prevede una estrazione periodica delle sole informazioni necessarie alla selezione coordinata, atta al popolamento della lista campionaria, consentendo un successivo ricollegamento con *ASIA* per attingere ad altre informazioni (ragione sociale, indirizzo…) necessarie ad altri scopi (invio del questionario,…).
I punti da definire sono soprattutto relativi a definire il "*quando*" e il "*come*" operare tale estrazione, esplicitando i concetti di "*nuova impresa*" ed "*impresa cessata*" e le modalità di trattamento delle imprese plurilocalizzate.

Come sopra scritto, è importante sottolineare che si potrebbe pensare ad *ASCO* come ad un sottosistema di *ASIA*, così come ad un sottoinsieme indipendente.
Nel caso in cui *ASCO* sia un sottosistema di *ASIA*, le informazioni di cui necessita solo il processo di *Selezione Coordinata* saranno organizzate in tabelle relazionali, fisicamente separate ma logicamente integrate con quelle di *ASIA*.
Al contrario, tutte le informazioni già presenti in *ASIA* non saranno replicate - al fine di evitare ridondanza di dati e possibili disallineamenti - ma saranno rese disponibili per il sottosistema

[1] Molestia statistica: carico di lavoro che ciascuna impresa deve sostenere per la compilazione del questionario relativo ad ogni indagine per cui viene selezionata.
[2] RDBMS – Relational Data Base Management System

ASCO, che dovrà essere strettamente integrato in *ASIA* mediante relazioni da generare durante la fase di definizione del Database.

Eventuali "ritorni" verso *ASIA*, come ad esempio le informazioni concernenti i campioni estratti, potranno essere gestiti automaticamente dalla procedura informatica della *Selezione Coordinata*.

Quanto detto presuppone ovviamente una analisi congiunta *ASIA-ASCO* per mettere a punto i criteri di estrazione da *ASIA*, la definizione dei vincoli di integrità tra le tabelle *ASIA/ASCO* e le informazioni che la procedura di *Selezione Coordinata* deve passare ad *ASIA*.

SIDI

Meno vincolante è il coinvolgimento di *SIDI*, in quanto alcune informazioni necessarie alla selezione coordinata (ad esempio la popolazione di interesse, la dimensione campionaria, la stratificazione, i piani di rotazione, i valori di molestia statistica) potrebbero essere presenti nel Sistema di documentazione delle indagini. Al momento non si è verificato se le informazioni che interessano al problema in analisi siano già contenute in *SIDI* o se sia previsto l'inserimento in un prossimo futuro; in tal caso diverrebbe fondamentale che tutte le indagini che entrano nel coordinamento dei campioni siano documentate completamente e correttamente in *SIDI*. Se invece si scegliesse di prescindere da *SIDI*, tutte le informazioni relative alle indagini afferiranno in una entità di proprietà del sistema di selezione coordinata, che sarà gestita dai responsabili delle indagini.

2. Una procedura informatica per applicare il metodo *Microstrat* alla selezione delle imprese soggette a rilevazioni nella realtà Istat

Una problematica ampia come quella descritta nel paragrafo precedente è da affrontare con estrema attenzione, in quanto ha un impatto notevole sulla realtà organizzativa dell'Istituto.

Un primo importante passo per risolvere la questione del coordinamento dei campioni è stato quello di procedere con la valutazione delle metodologie alla base del problema per selezionare quel particolare metodo che meglio si adatta alla realtà, sperimentandolo tramite una procedura informatica sviluppata per simulare l'applicazione del metodo.

La questione iniziale che il gruppo di lavoro ha dovuto affrontare è stata quella di valutare se prediligere o meno un sistema che considera la *molestia* causata alle imprese. Da una prima analisi, basandosi su quanto emerge dalla letteratura in materia, il metodo *Microstrat* è risultato il più adeguato, il che ha portato a procedere ad una valutazione dell'applicazione di questo metodo alla realtà dell'Istituto (Riviere P., 2001). Si è dunque proceduto nell'implementazione del metodo *Microstrat* e nell'organizzazione di una simulazione i cui risultati sono stati analizzati e comparati con quelli ottenuti applicando altri metodi di selezione dei campioni, in particolare il campionamento casuale semplice stratificato e il metodo di Jales (Ohlsson E., 1995).

Il metodo *Microstrat* presuppone un grosso sforzo implementativo, in quanto è necessario conservare una grande mole di informazioni delle indagini passate e ciò è vero per ciascuna delle indagini per le quali è necessario selezionare le unità campionarie. L'algoritmo che ne deriva è molto "pesante" in termini di tempi di elaborazione.

Per i suddetti motivi si è proceduto ad attuare una simulazione molto accurata, che ha previsto una analisi profonda e si è demandato ad una fase successiva lo studio di software generalizzato integrato con altri sistemi dell'Istituto.

La sperimentazione si è basata su alcune ipotesi di partenza.

Si è considerato esclusivamente la situazione di indagini disgiunte rispetto alle imprese oggetto di rilevazione (coordinamento globale negativo); si è deciso di escludere dall'insieme delle indagini della sperimentazione sia le rilevazioni censuarie che le rilevazioni campionarie senza selezione probabilistica, che non influiscono nel coordinamento delle imprese, così come si sono escluse alle indagini con parte del campione costituita da panel ruotato.

Infine, il valore della molestia statistica viene considerato uguale per tutte le indagini.

Per ciò che concerne la popolazione di riferimento, è costituita dall'archivio *ASIA* 2001, come si approfondirà nel successivo *paragrafo 4*.

Si è perciò proceduto con:

- la raccolta e l'analisi dei dati delle indagini, comprensiva dell'analisi delle informazioni memorizzate in *ASIA* e dei programmi attualmente utilizzati dalle indagini per selezionare le imprese (questo lavoro di selezione e preparazione dei dati è stato anche utile per l'applicazione degli altri metodi di selezione);
- la definizione di un percorso per la realizzazione della sperimentazione del metodo *Microstrat* e dunque per l'implementazione della procedura informatica adeguata a simulare la selezione delle imprese;
- lo studio dell'algoritmo alla base del metodo *Microstrat*, implementato nella procedura informatica;
- l'analisi e realizzazione di una base di dati, disegnata opportunamente per realizzare la sperimentazione del metodo.

Nei prossimi paragrafi vengono descritte alcune informazioni, separatamente per i punti di cui sopra; vengono infine illustrati alcuni risultati sui tempi delle elaborazioni.

3. La raccolta dei dati per la sperimentazione del metodo *Microstrat*

La sperimentazione del metodo ha previsto lo studio di otto indagini e la successiva selezione di alcune tra queste, adeguate per effettuare una simulazione.

In dettaglio l'analisi è partita dalla raccolta di informazioni relativamente agli anni 2001-2003 e considerando le seguenti indagini[3]:

- Rilevazione mensile delle vendite al dettaglio (IST-00151)
- Rilevazione trimestrale del fatturato e dell'occupazione delle imprese del commercio all'ingrosso e intermediari del commercio (IST-00948)
- Rilevazione trimestrale del fatturato e dell'occupazione nel settore dell'informatica
- Indagine trimestrale su posti vacanti ed ore lavorate (IST-01381)
- Rilevazione annuale della produzione industriale (Prodcom) (IST-00070)
- Piccole e medie imprese e esercizio di arti e professioni (Pmi) (IST-00954)

[3] La denominazione utilizzata per le indagini e il relativo codice seguono l'ultimo elenco disponibile in www.sistan.it: "Elenco delle rilevazioni rientranti nel programma statistico nazionale 2004-2006, che comportano obbligo di risposta da parte dei soggetti privati, a norma dell'art. 7 del decreto legislativo 6 settembre 1989, n. 322".

- Struttura del costo del lavoro (IST-00714)
- Rilevazione statistica sulle tecnologie dell'informazione, della comunicazione e competitivita' delle imprese (IST-01175).

La raccolta è stata effettuata dai responsabili delle indagini seguendo le indicazioni del gruppo di lavoro, riportate nel seguito.
Sono stati richiesti, con riferimento al periodo e alle indagini sopra scritte, i dati memorizzati in file, formato Excel o Ascii o SAS, relativamente alle seguenti informazioni:

- I codici degli strati utilizzati nell'indagine; assieme a tali codici sono state richieste le modalità delle variabili presenti su *ASIA* che li identificano;
- *nh* Le numerosità campionarie relative agli strati utilizzati (derivanti, ad esempio, da software *MAUSS*);
- *Nh* La numerosità delle unità presenti in quello strato nella popolazione;
- Il dominio di stima, o i domini di stima nel caso l'indagine ne preveda più di uno.

Ovviamente anche per i domini di stima, come per lo strato, è necessario esplicitare le modalità delle variabili presenti su *ASIA* che identificano i codici per ciascuno dei domini utilizzati.

La tavola 1 che segue è stata inviata ai responsabili delle indagini, come esempio delle informazioni richieste:

Tavola 1 - Esempio di dati richiesti

Strato	regione	Ateco2	cladd	Dom1	Ateco2	regione	Dom2	Ateco2	Cladd	N	Camp
0110I1	01	10	I1	1001	10	01	10I1	10	I1	1	1
0113I1	01	13	I1	1301	13	01	13I1	13	I1	2	2
0114I1	01	14	I1	1401	14	01	14I1	14	I1	93	33
0114I2	01	14	I2	1401	14	01	14I2	14	I2	10	3
0114I3	01	14	I3	1401	14	01	14I3	14	I3	3	2
0114I1	01	14	I1	1401	14	01	14I1	14	I1	3	2
0114I1	01	14	I1	1401	14	01	14I1	14	I1	3	2
0114I1	01	14	I1	1401	14	01	14I1	14	I1	109	16
0114I2	01	14	I2	1401	14	01	14I2	14	I2	36	2
0114I3	01	14	I3	1401	14	01	14I3	14	I3	11	4
0114I4	01	14	I4	1401	14	01	14I4	14	I4	3	3
0114I1	01	14	I1	1401	14	01	14I1	14	I1	8	2
0114I2	01	14	I2	1401	14	01	14I2	14	I2	2	2
0114I1	01	14	I1	1401	14	01	14I1	14	I1	1	1
0114I1	01	14	I1	1401	14	01	14I1	14	I1	23	15
0114I2	01	14	I2	1401	14	01	14I2	14	I2	6	5
0114I3	01	14	I3	1401	14	01	14I3	14	I3	4	2
0115I1	01	15	I1	1501	15	01	15I1	15	I1	121	13
0115I2	01	15	I2	1501	15	01	15I2	15	I2	26	15
0115I3	01	15	I3	1501	15	01	15I3	15	I3	10	3
0115I4	01	15	I4	1501	15	01	15I4	15	I4	3	2

E’ stato anche richiesto - ove gli strati siano definiti mediante variabili la cui codifica è costruita *ad hoc* per l’indagine, ad esempio classe addetti – di specificare chiaramente la codifica, e dunque le modalità di tali variabili.

E’ stata inoltre richiesta la popolazione di riferimento dell’indagine (e le variabili che la definiscono ad esempio, imprese con Ateco2=xx).

E’ stato infine richiesto per ogni indagine a quale anno di *ASIA* si riferisce e quale classificazione Ateco ha utilizzato.

Le informazioni di cui sopra sono state raccolte per procedere ad una simulazione delle estrazioni campionarie basandosi sulle allocazioni campionarie effettivamente adoperate nel corso dei tre anni analizzati.
Le informazioni ricevute sono sintetizzate nelle due tabelle successive:

Tavola 2a – Sintesi delle informazioni ricevute da parte delle indagini

ID_Indagine	Titolo	Descrizione	Periodicità	Servizio	Anno Rilevazione	Periodo Riferimento Indagine	ASIA	ATECO
1	PRODCOM	Produzione Industriale	Annuale	SSI	2003		2001	2002
2	PMI	Piccole e medie imprese	Annuale	SSI	2003		2000	2002
					2002		2001	1991
					2001		2000	1991
3	ICT	Commercio elettronico	Annuale	SSI	2003		2000	2002
					2002		1999	1991
4	CL	Rilevazione sul costo del lavoro	Quadriennale	SSI	2000		1998	1991
5	VD	Vendite al dettaglio	Annuale	SCO	2002		2000	2002
					2004		2002	2002
6	CI	Commercio all'ingrosso	Trimestrale	SCO	2004	2004	2001	2002
					2003	2003	2000	2002
					2002	2002	2000	2002
					2002	2001	1999	2002
					2002	2000	1999	2002
7	SI	Servizi informatici	Trimestrale	SCO	2002			2002
					2001			2002
					2000			2002
8	IPV	Indagine posti vacanti	Trimestrale	SCO			2002	2002

Tavola 2b – Sintesi delle informazioni ricevute da parte delle indagini

ID_Indagine	Titolo	Popolazione di riferimento	Note
1	PRODCOM	INDUSTRIA: sezioni C (ad esclusione della sottosezione CA); D (ad esclusione della sottosezione DF) quindi dalla divisione 13 alla 36 ad esclusione della divisione 23.	
		Imprese attive con almeno 3 addetti	
2	PMI	INDUSTRIA, sezioni C, D, E, F : 1-9, 10-19, 20-49, 50-99	
		COMMERCIO sezione G : 1, 2-4, 5-9, 10-19, 20-49, 50-99	
		SERVIZI ALBERGHIERI E ALLE IMPRESE, sezioni H, I, K, J (divisione 67) : 1-4, 5-9, 10-19, 20-49, 50-99	
		ALTRI SERVIZI, M, N, O (divisioni 90, 92, 93) : 1-9, 10-19, 20-49, 50-99	
3	ICT	Almeno 10 addetti	Le imprese con almeno 250 addetti sono censite. La conversione da ATECO91 a ATECO2002 viene eseguita da programma SAS prima dell'estrazione.
		imprese con almeno 10 addetti dei settori ATECO91 a livello di sezione da D a K,	Le imprese con almeno 250 addetti sono censite
4	CL	Sezioni ATECO91 C-K con almeno 10 addetti	
5	VD	Divisione 52 ATECO	Risulta compatibile la portabilità delle classificazioni ATECO 91 e 2002 per la popolazione di riferimento
		Divisione 52 ATECO	Risulta compatibile la portabilità delle classificazioni ATECO 91 e 2002 per la popolazione di riferimento
6	CI	Divisione 51 ATECO 2002	
		Divisione 51 ATECO 2002	Relaz. 1:1 tra ATECO 91 e 02 per le voci di classificazione usate come popolazione di riferimento (516-517['91] -> 518-519['02])
		Divisione 51 ATECO 2002	Relaz. 1:1 tra ATECO 91 e 02 per le voci di classificazione usate come popolazione di riferimento (516-517['91] -> 518-519['02])
		Divisione 51 ATECO 2002	Relaz. 1:1 tra ATECO 91 e 02 per le voci di classificazione usate come popolazione di riferimento (516-517['91] -> 518-519['02])
		Divisione 51 ATECO 2002	Relaz. 1:1 tra ATECO 91 e 02 per le voci di classificazione usate come popolazione di riferimento (516-517['91] -> 518-519['02])
7	SI	(72100-72600) ATECO	Risulta compatibile la portabilità delle classificazioni ATECO 91 e 2002 per la popolazione di riferimento
		(72100-72600) ATECO	Risulta compatibile la portabilità delle classificazioni ATECO 91 e 2002 per la popolazione di riferimento
		(72100-72600) ATECO	Risulta compatibile la portabilità delle classificazioni ATECO 91 e 2002 per la popolazione di riferimento
8	IPV	ATECO 2002 compreso tra '100' e '749' con addetti totali medi di almeno 10 unità	

Le indagini effettivamente selezionate per la simulazione sono le seguenti: Prodcom, Piccole e medie imprese, Commercio elettronico, Rilevazione del costo del lavoro, Indagine dei posti vacanti.

Le prime quattro sono afferenti al servizio delle statistiche strutturali e l'ultima di competenza del servizio delle statistiche congiunturali; tre indagini, sempre specifiche delle statistiche congiunturali (Vendite al dettaglio, Commercio all'ingrosso e Servizi informatici), sono state eliminate dalla sperimentazione, in quanto presupponevano un fattore di sovrapposizione tra le unità campionarie riferite a diverse estrazioni. Come scritto nel paragrafo 2, in questa prima fase di sperimentazione si è deciso di non considerare questo tipo di indagini.

Le cinque indagini selezionate si riferiscono comunque a differenti popolazioni di riferimento ed a differenti numerosità degli strati. In tal modo si può comunque verificare il metodo in analisi secondo una simulazione che garantisca una certa variabilità in termini di caratteristiche delle indagini.

Ciò è descritto nella *tavola 3*:

Tavola 3 – Informazioni relative alle indagini usate per la simulazione

DescrizioneIndagine	DescrizionePeriodicita	Popolazione di riferimento	Numero strati
PRODCOM	Annuale	252.082	2.200
Piccole e medie imprese	Annuale	3.978.612	25.129
Commercio elettronico	Annuale	195.450	1.966
Rilevazione costo del lavoro	Quadriennale	187.948	1.664
Indagine posti vacanti	Trimestrale	188.093	57

4. Il percorso della sperimentazione

La sperimentazione del metodo *Microstrat* è stata effettuata sulla base di un contesto ben definito, come descritto nel *paragrafo 2,* e sulla base di una serie di considerazioni: nel seguito si riportano i punti salienti che hanno portato a definire il percorso di realizzazione della simulazione.

1) L'analisi dei dati delle indagini e la definizione dei dati per la simulazione

I dati raccolti sono stati inviati dai servizi responsabili delle indagini secondo la classificazione utilizzata dall'indagine stessa al momento dello svolgimento della fase di allocazione campionaria.

Il primo problema che si è affrontato nell'analisi dei dati riferiti ai diversi anni, è stato quello del *cambiamento delle classificazioni*, determinante per quanto concerne l'archivio *ASIA*.

Fino ad *ASIA*-2000 la classificazione delle attività economiche si riferiva all'Ateco91; in *ASIA*-2001 sono presenti sia l'Ateco91 che l'Ateco2002; da *ASIA*-2002 è presente solo l'Ateco2002.
Le variazioni maggiori riguardano l'Ateco a 5 cifre; già per l'Ateco a 4 cifre tali variazioni diminuiscono sensibilmente (le informazioni ricevute in proposito, ci hanno indicato circa 600.000 variazioni per quanto riguarda l'Ateco a 5 cifre e dalle 20.000 alle 30.000 per l'Ateco a 4 cifre); ci sono comunque, anche se pochissime, variazioni anche per quello che riguarda i codici Ateco a 2 e 3 cifre.

Durante il periodo in esame (2001-2003) abbiamo riscontrato i seguenti casi:

- Indagini riferite ad un certo anno che usano *ASIA*2000[4] o *ASIA*2001 (ed anche precedenti) e la classificazione Ateco91
- Indagini riferite ad un certo anno che usano *ASIA*2000, *ASIA*2001, ma la classificazione Ateco91 è stata tradotta poi in quella Ateco2002
- Indagini riferite ad un certo anno che usano *ASIA*2002 e la classificazione Ateco2002
- Indagini riferite ad un certo anno che in base alla loro popolazione di riferimento risultano indipendenti dalla versione dell'Ateco.

Le possibilità che il gruppo di lavoro ha considerato per la sperimentazione sono nel seguito illustrate:

[4] ASIA2000 è l'archivio statistico delle imprese attive all'anno 2000

- Procedere considerando solo un anno per ciascuna indagine, la cui popolazione di riferimento si debba basare solo su una specifica classificazione. Si vuole dunque considerare un solo archivio *ASIA* di riferimento e si simula il fluire del tempo ripetendo i dati delle indagini. Tale scelta permette di verificare il metodo, anche se la simulazione si sviluppa basandosi su una condizione non reale;
- Procedere applicando il metodo ai dati così come sono stati raccolti, quindi non variando le classificazioni utilizzate. Tale scelta permette di verificare alcune condizioni che potrebbero realmente verificarsi in un contesto reale, ma non consente di verificare al meglio il metodo, in quanto l'esecuzione del processo e l'analisi sperimentale risentirebbero del fenomeno del cambiamento dello strato, dovuto alla classificazione variata e non invece a variazioni nelle caratteristiche delle imprese stesse.

Si è deciso di optare per la prima soluzione, considerando che l'obiettivo della sperimentazione è quello di verificare il metodo.
In questa prima fase della sperimentazione si è dunque ipotizzata assenza di movimento demografico delle imprese.

La popolazione di riferimento presa in considerazione è stata *ASIA*2001, che contiene sia la codifica Ateco91 che Ateco2002. Nella sperimentazione si è tenuto conto della sola classificazione Ateco2002.

La seconda questione affrontata, è stata quella di dover *standardizzare la definizione degli strati*, diversa per ciascuna indagine.

Sono state impostate delle condizioni (condizioni di *where*) che definiscono le varie stratificazioni a partire dalla classificazione Ateco2002. Tali condizioni sono memorizzate nella base dei dati.
L'applicazione a regime garantirà la possibilità per l'utente di costruire la propria stratificazione tramite una funzione interattiva, costruita allo scopo. Al momento però, non avendo la possibilità e la necessità di sviluppare una funzione di *utility* come questa, piuttosto complessa e che dunque richiede del tempo per essere implementata, è stato necessario, basandosi sullo studio dei programmi di selezione utilizzati nelle indagini, preparare i dati fissando le condizioni per identificare la stratificazione in modo standard per tutte le indagini.
Tali condizioni sono state verificate e convalidate dal gruppo di lavoro, in modo da poter procedere con la simulazione.

2) La definizione della situazione di partenza per la simulazione

L'algoritmo alla base del metodo *Microstrat* è stato implementato per simulare le estrazioni campionarie seguendo un preciso calendario, come illustrato in *tavola 4*.

Tavola 4 – La simulazione: date e estrazioni

Sequenza Estrazione	Data Estrazione	Identificativo indagine	Indagine	Peridicità	Estrazione
1	2002 02 01	1	Produzione PRODCOM	Annuale	1
2	2002 02 15	8	Indagine posti vacanti	Trimestrale	1
3	2002 03 01	2	Piccole e medie imprese	Annuale	1
4	2002 05 02	3	Commercio elettronico	Annuale	1
5	2003 02 01	1	PRODCOM	Annuale	2
6	2003 02 15	8	Indagine posti vacanti	Trimestrale	2
7	2003 03 01	2	Piccole e medie imprese	Annuale	2
8	2003 05 02	3	Commercio elettronico	Annuale	2
9	2004 02 01	1	PRODCOM	Annuale	3
10	2004 02 15	8	Indagine posti vacanti	Trimestrale	3
11	2004 03 01	2	Piccole e medie imprese	Annuale	3
12	2004 05 02	3	Commercio elettronico	Annuale	3
13	2004 06 01	4	Rilevazione sul costo del lavoro	Quadriennale	1

Nel successivo *paragrafo 5* vengono specificati i passi dell'algoritmo.

In questa sede si vuole sottolineare che la simulazione è partita considerando un punto di partenza, ovvero il verificarsi di una certa indagine.

Fissando il punto di inizio è stato possibile simulare il succedersi delle diverse estrazioni durante un periodo prefissato.

La base dati (si veda *paragrafo 6*) è stata opportunamente disegnata per sviluppare una procedura informatica definita secondo un contesto sperimentale precisamente delineato.

Ovviamente, nel caso il metodo Microstrat venisse adottato dall'Istituto e si volesse considerare la situazione pregressa già dalla prima estrazione dei dati, si dovrà prevedere la revisione sia della base dati (ricostruzione di tutte le estrazioni dei due anni precedenti) che dell'algoritmo, per tener conto di questa ricostruzione a posteriori.

3) La procedura implementata ed il contesto applicativo attuale

La procedura informatica sviluppata per implementare il metodo *Microstrat* nel contesto sperimentale definito è un programma, la cui esecuzione avviene in modalità *batch*; non prevede al momento che sia memorizzato il carico statistico cumulato[5] ma solo quello "unitario" da associare alla selezione stessa; l'esecuzione permetterà poi il calcolo in tempo reale del carico statistico cumulato che dovrà effettuarsi considerando finestre di intervalli temporali che variano nel tempo a seconda della data di effettuazione dell'indagine.
L'implementazione di alcune funzioni, anche abbastanza complesse, quale quelle che permettono di definire la stratificazione delle indagini in modo interattivo a partire dalle variabili di *ASIA*, è demandata ad una fase successiva a quella attuale, relativa all'analisi del metodo di coordinamento.

Si demanda dunque ad una fase successiva sia lo sviluppo di una applicazione *user-friendly*, che la definizione della base dati che realmente sarà utilizzabile a regime, così come l'implementazione dell'algoritmo definitivo, che dipende da scelte ad oggi non ritenute necessarie per questa prima fase.
Come sopra scritto, la procedura è stata implementata utilizzando una base dati relazionale. Sarà possibile solo in una fase successiva decidere il linguaggio di programmazione definitivo da utilizzare e l'architettura dei dati adeguata allo sviluppo di una procedura generalizzata integrata; nello sviluppo della procedura si è per ora utilizzato VISUAL BASIC con supporto dati ACCESS-ODBC-ORACLE.

[5] Carico statistico cumulato: carico di lavoro complessivo che ciascuna impresa ha sostenuto per tutte le indagini per cui è stata selezionata nel corso della propria vita.

5. Il metodo e l'algoritmo

La procedura è stata sviluppata nell'ottica di garantire uno strumento operativo per la sperimentazione.

Ricapitolando quanto sopra scritto, l'implementazione dell'algoritmo ha perciò considerato i seguenti punti:

- si è considerato solo il caso di coordinamento *globale* negativo;
- per la determinazione del carico statistico cumulato (CSC) si assume che per ogni unità selezionata nel campione il carico sia incrementato di una quantità pari ad uno, indipendentemente dall'indagine di riferimento e non differenziando tale incremento se una indagine prevede più somministrazioni di un questionario ad una stessa impresa;
- si riduce la sperimentazione a cinque indagini, prendendo in considerazione i dati relativi ad una sola estrazione (popolazione di riferimento *ASIA2001*) e replicandoli nel triennio considerato;
- si parte dall'ipotesi che la situazione iniziale parta con l'effettuarsi di una certa indagine, senza considerare il passato.

Nel seguito si riportano i passi dell'algoritmo alla base del metodo *Microstrat*.

Escludendo il caso particolare della prima indagine, dalla seconda in poi si prevede l'esecuzione di diversi passi per ogni selezione; è infatti necessario tenere conto di tutti i possibili microstrati di intersezione che vengono a definirsi rispetto:

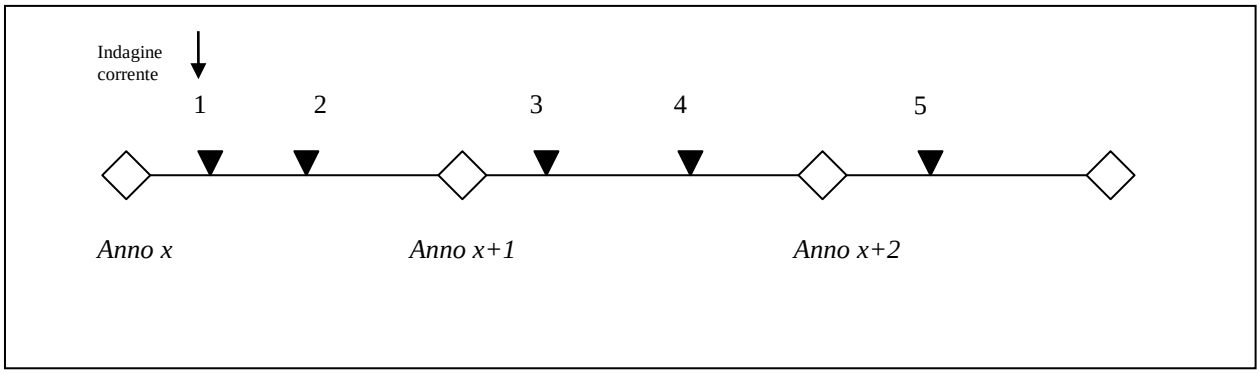
a) alla precedente indagine, ovvero alla estrazione di una qualsiasi altra indagine che temporalmente si è verificata precedentemente a quella corrente (*si considera in questo caso la stratificazione precedente come fosse una particolare microstratificazione)*;

b) a tutte le indagini dell'anno precedente *(in questo caso entrano nella formazione dei microstrati tutte le stratificazioni di tutte le indagini effettuate in un arco di tempo pari ai 12 mesi precedenti rispetto a quella corrente)*;

c) a tutte le indagini dei due anni precedenti *(in questo caso entrano nella formazione dei microstrati tutte le stratificazioni di tutte le indagini effettuate in un arco di tempo pari ai 24 mesi precedenti rispetto a quella corrente).*

d) se una indagine si riferisce ad insiemi di microstrati disgiunti rispetto alla precedente, ciò non deve comportare variazioni nell'esecuzione dell'algoritmo (ovvero si passa a considerare direttamente la finestra di un anno).

Passo 0 iniziale – Si considera la prima indagine della sperimentazione : Indagine corrente=1 (caso particolare in cui non si considera alcuna microstratificazione)



- Si attribuisce ad ogni impresa un numero casuale

- Si ordinano i dati:

1. Sulla base degli strati dell'indagine corrente (si hanno così le imprese divise a seconda della stratificazione della indagine 1);
2. All'interno di ogni strato per numero casuale crescente.

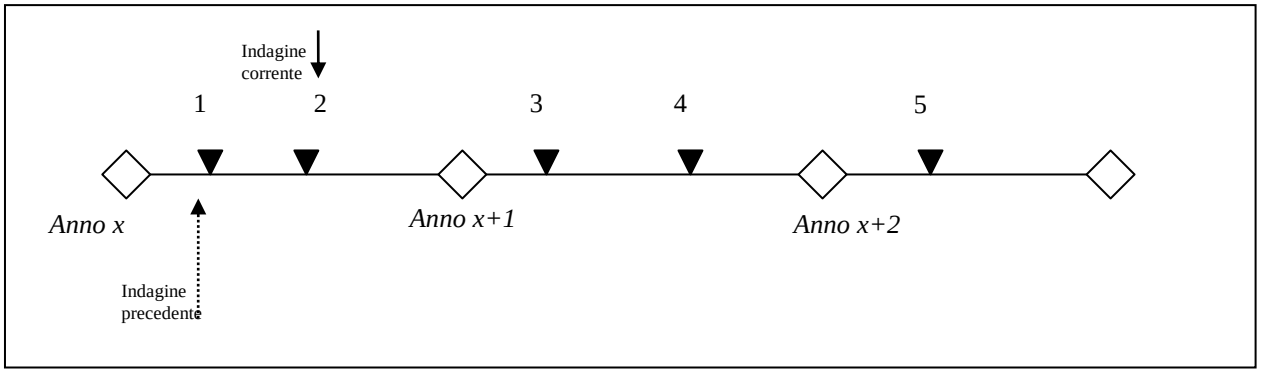
- Si **selezionano** i dati secondo l'allocazione definita per l'indagine corrente 1.

Si selezionano i numeri casuali più bassi per ciascuno strato.

- Si aggiorna il carico statistico delle imprese selezionate.

Ciò significa - alla prima indagine - associare un carico pari ad 1 alle imprese selezionate.

Passo 1 - Si considera la seconda indagine della sperimentazione : Indagine corrente = 2
(si ha una sola indagine precedente e si considera la stratificazione dell'indagine precedente come un microstrato)



- Si ordinano i dati:

1. Sulla base degli strati dell'indagine precedente (si hanno così le imprese divise a seconda della stratificazione della indagine precedente 1);
2. All'interno di ciascuno strato sulla base del carico statistico in ordine crescente (a questo stadio può valere solo 0 e 1), in modo che il carico più alto sia alla fine dello strato.

- Si permutano i numeri casuali, in modo che per ogni strato dell'indagine precedente, il più piccolo numero sia assegnato a quello con carico statistico più basso (a questo stadio può valere solo 0 e 1).

- Si ordinano nuovamente i dati, questa volta :

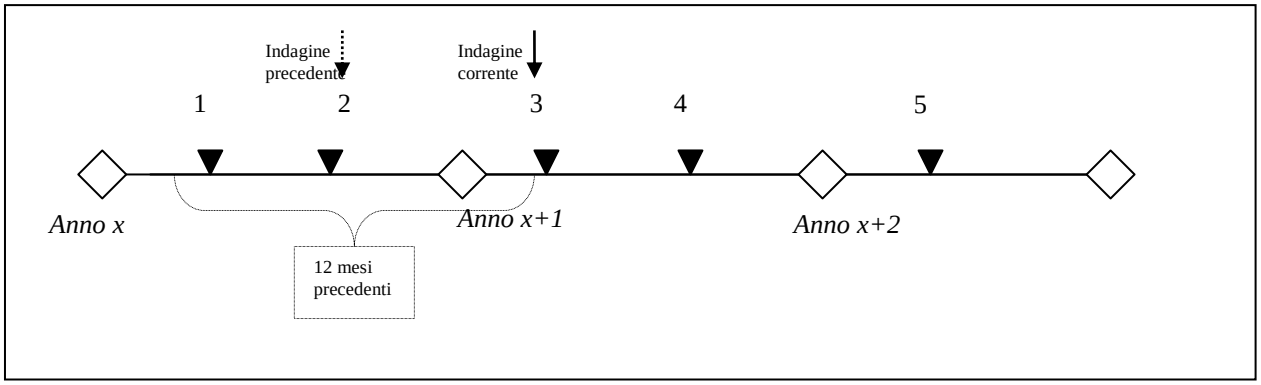
1. Sulla base degli strati dell'indagine corrente (si hanno così le imprese divise a seconda della stratificazione dell'indagine corrente 2, riunendo dunque quelle che precedentemente entravano in strati diversi);
2. All'interno di ciascuno strato dell'indagine corrente 2, sulla base del numero casuale in ordine crescente.

- A questo punto si **selezionano** le unità secondo l'allocazione definita per l'indagine, scegliendo le imprese a cui sono associati i numeri casuali più bassi per ciascuno strato.

- Si aggiorna il carico statistico delle imprese selezionate

Ciò significa – dalla seconda indagine – sommare un valore pari ad 1 al valore del carico che le imprese selezionate avevano precedentemente (0 o 1 a questo stadio).

Passo 3 - Si considera la terza indagine della sperimentazione : Indagine corrente = 3 (dalla terza indagine si formano i veri e propri microstrati).



Si considera prima l'indagine precedente 2, poi è necessario considerare tutte le indagini effettuate nell'arco dei 12 mesi precedenti (1 e 2 contemporaneamente) e successivamente quelle effettuate nei 24 mesi precedenti (in questo caso non esistono).
L'algoritmo eseguirà i seguenti passi:

- Si ordinano i dati:
1. Sulla base degli strati dell'indagine precedente (si hanno le imprese divise a seconda della stratificazione 2)
2. All'interno di ciascuno strato sulla base del carico in ordine crescente, in modo che il carico più alto sia alla fine dello strato.

- Si permutano i numeri casuali, in modo che per ogni strato dell'indagine precedente, il più piccolo numero sia assegnato a quello con carico più basso.

<u>Non si selezionano i dati ma si procede con gli ordinamenti.</u>

- Si ordinano nuovamente i dati :
1. Sulla base dei microstrati che si formano considerando l'intersezione tra gli strati di tutte le indagini che entrano nella finestra dei 12 mesi precedenti rispetto all'indagine corrente 3 (in questo caso solo le indagini 1 e 2) si hanno le imprese divise a seconda della microstratificazione formata sulla base delle imprese 1 e 2.
2. All'interno di ciascun microstrato sulla base del numero casuale in ordine crescente.

- Si permutano i numeri casuali, in modo che per ogni microstrato il più piccolo numero sia assegnato a quello con carico più basso.

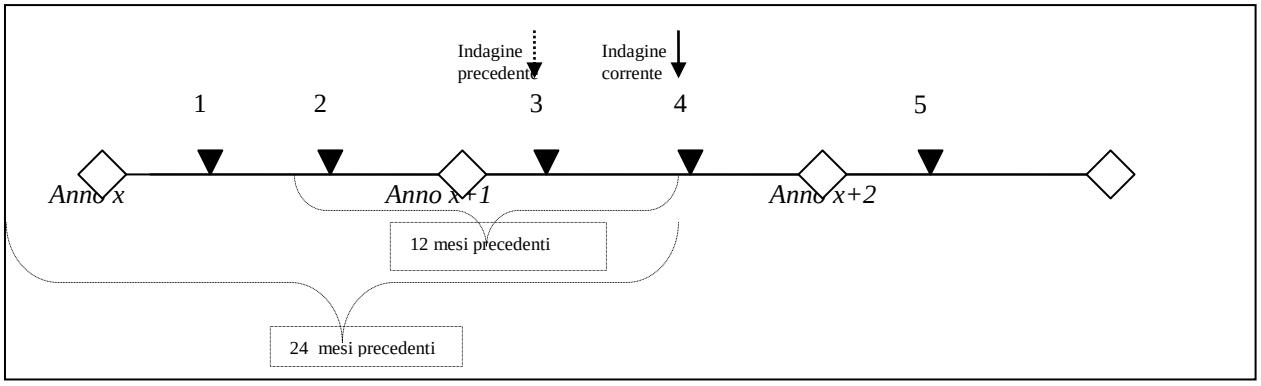
- Si ordinano nuovamente i dati, questa volta :

1. Sulla base degli strati dell'indagine corrente (si hanno così le imprese divise a seconda della stratificazione dell'indagine corrente 3, riunendo dunque quelle che precedentemente entravano in microstrati diversi)
2. All'interno di ciascuno strato dell'indagine corrente 3, sulla base del numero casuale in ordine crescente

- A questo punto si **selezionano** i dati secondo l'allocazione definita per l'indagine, scegliendo le imprese a cui sono associati i numeri casuali più bassi per ciascuno strato.

Passo 4

- **Si considera la quarta indagine della sperimentazione : Indagine corrente = 4**



L'algoritmo in questo caso dovrà considerare prima l'indagine precedente 3, poi le indagini effettuate nell'arco dei 12 mesi precedenti (la 2 e la 3) e successivamente quelle effettuate nei 24 mesi precedenti (la 1 la 2 e la 3).

In dettaglio:

- Si ordinano i dati:

1. Sulla base degli strati dell'indagine precedente (si hanno le imprese divise a seconda della stratificazione 3);

2. All’interno di ciascuno strato sulla base del carico in ordine crescente, in modo che il carico più alto sia alla fine dello strato.

- Si permutano i numeri casuali, in modo che per ogni strato dell’indagine precedente, il più piccolo numero sia assegnato a quello con carico più basso.

Non si selezionano i dati ma si procede con gli ordinamenti.

- Si ordinano nuovamente i dati :
1. Sulla base dei microstrati che si formano considerando l’intersezione tra gli strati di tutte le indagini che entrano nella finestra dei 12 mesi precedenti rispetto all’indagine corrente 4 (in questo caso le indagini 2 e 3) si hanno le imprese divise a seconda della microstratificazione formata sulla base delle imprese 2 e 3;
2. All’interno di ciascun microstrato sulla base del numero casuale in ordine crescente.

- Si permutano i numeri casuali, in modo che per ogni microstrato il più piccolo numero sia assegnato a quello con carico più basso.

Ancora una volta non si selezionano i dati ma si procede con gli ordinamenti.

- Si ordinano nuovamente i dati :
1. Sulla base dei microstrati che si formano considerando l’intersezione tra gli strati di tutte le indagini che entrano nella finestra dei 24 mesi precedenti rispetto all’indagine corrente 4 (in questo caso le indagini 1, 2 e 3) si hanno le imprese divise a seconda della microstratificazione formata sulla base delle imprese 1, 2 e 3;
2. All’interno di ciascun microstrato sulla base del numero casuale in ordine crescente.

- Si permutano i numeri casuali, in modo che per ogni microstrato il più piccolo numero sia assegnato a quello con carico più basso.

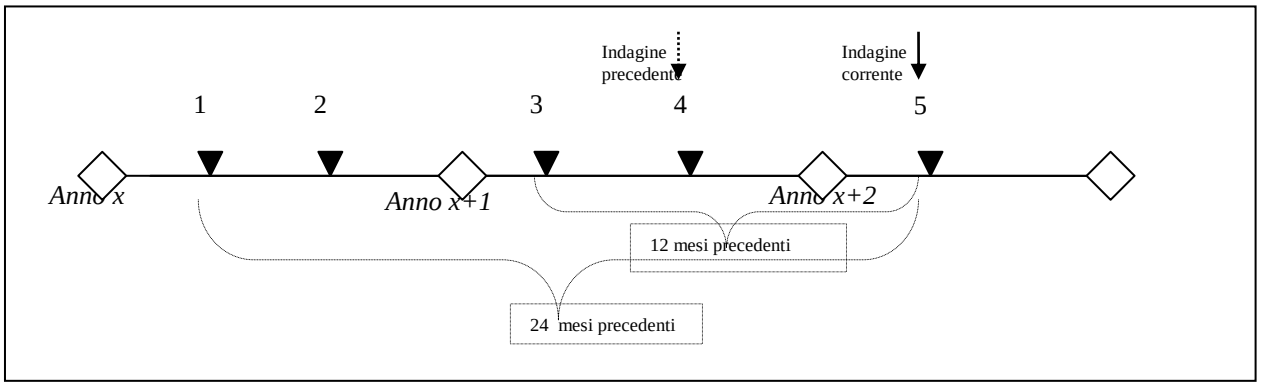
- Si ordinano nuovamente i dati, questa volta :
1. Sulla base degli strati dell'indagine corrente (si hanno così le imprese divise a seconda della stratificazione dell’indagine corrente 4, riunendo dunque quelle che precedentemente entravano in microstrati diversi);

2. All’interno di ciascuno strato dell’indagine corrente 4, sulla base del numero casuale in ordine crescente.

- A questo punto si **selezionano** i dati secondo l’allocazione definita per l’indagine, scegliendo le imprese a cui sono associati i numeri casuali più bassi per ciascuno strato.

Passo 5

- **Si considera la quarta indagine della sperimentazione : Indagine corrente = 5**



L’algoritmo in questo caso dovrà considerare prima l’indagine precedente 4, poi le indagini effettuate nell’arco dei 12 mesi precedenti (la 3 e la 4) e successivamente quelle effettuate nei 24 mesi precedenti (la 1, 2 la 3 e la 4).
In dettaglio:

- Si ordinano i dati:
1. Sulla base degli strati dell'indagine precedente (si hanno le imprese divise a seconda della stratificazione 4);
2. All’interno di ciascuno strato sulla base del carico in ordine crescente, in modo che il carico più alto sia alla fine dello strato.

- Si permutano i numeri casuali, in modo che per ogni strato dell’indagine precedente, il più piccolo numero sia assegnato a quello con carico più basso.

<u>Non si selezionano i dati ma si procede con gli ordinamenti.</u>

- Si ordinano nuovamente i dati :

1. Sulla base dei microstrati che si formano considerando l'intersezione tra gli starti di tutte le indagini che entrano nella finestra dei 12 mesi precedenti rispetto all'indagine corrente 4 (in questo caso le indagini 3 e 4) si hanno le imprese divise a seconda della microstratificazione formata sulla base delle imprese 3 e 4;
2. All'interno di ciascun microstrato sulla base del numero casuale in ordine crescente.

- Si permutano i numeri casuali, in modo che per ogni microstrato il più piccolo numero sia assegnato a quello con carico più basso.

<u>Ancora una volta non si selezionano i dati ma si procede con gli ordinamenti.</u>

- Si ordinano nuovamente i dati:

1. Sulla base dei microstrati che si formano considerando l'intersezione tra gli starti di tutte le indagini che entrano nella finestra dei 24 mesi precedenti rispetto all'indagine corrente 5 (in questo caso le indagini 1, 2, 3 e 4) si hanno le imprese divise a seconda della microstratificazione formata sulla base delle imprese 1, 2, 3 e 4;
2. All'interno di ciascun microstrato sulla base del numero casuale in ordine crescente.

- Si permutano i numeri casuali, in modo che per ogni microstrato il più piccolo numero sia assegnato a quello con carico più basso.

- Si ordinano nuovamente i dati, questa volta:

1. Sulla base degli strati dell'indagine corrente (si hanno così le imprese divise a seconda della stratificazione dell'indagine corrente 5, riunendo dunque quelle che precedentemente entravano in microstrati diversi);
2. All'interno di ciascuno strato dell'indagine corrente 5, sulla base del numero casuale in ordine crescente.

- A questo punto si **selezionano** i dati secondo l'allocazione definita per l'indagine, scegliendo le imprese a cui sono associati i numeri casuale più bassi per ciascuno strato.

6. Lo schema relazionale alla base della procedura informatica

In questo paragrafo vengono presentate le tabelle dello schema riportato in *figura 2*.

Definiamo nell'ambito di uno schema relazionale tutte le tabelle che permetteranno al sistema informativo integrato di gestire tutte le funzioni applicative che tale progetto richiede, partendo dai dati che descrivono la realtà d'interesse.

Figura 2. Schema fisico relazionale sottostante il progetto di selezione coordinata

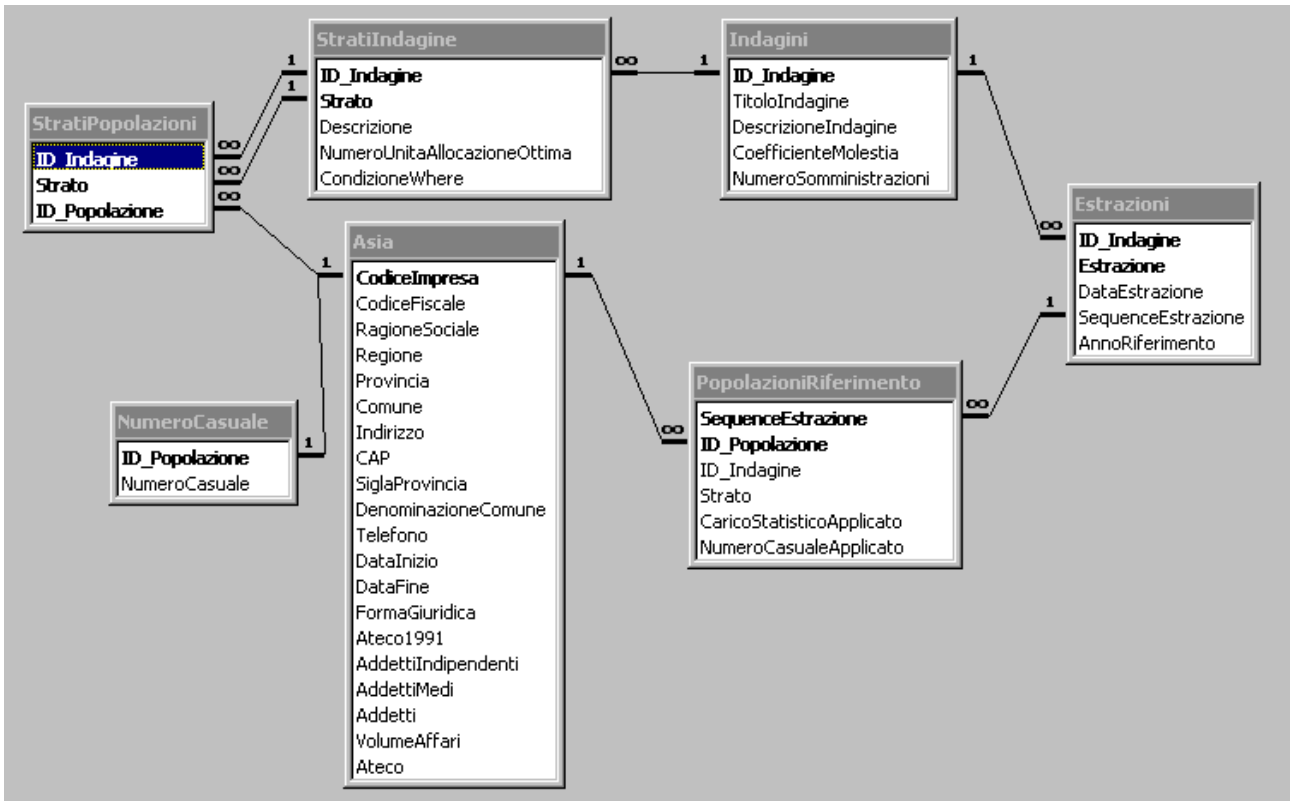


Tabella "Indagini"

Ogni indagine inserita nel progetto di selezione coordinata viene rappresentata da una istanza nella tabella "Indagini".

Chiave primaria(ID_Indagine): rappresenta l'**identificativo dell'indagine** nel progetto. Tale codice, nell'ambito di una codifica univoca ISTAT e in una ottica di futura integrazione con gli altri sistemi informativi di carattere gestionale, documentativi e di diffusione, potrebbe essere l'identificativo dell'indagine nel PSN o in *SIDI*.

Attributi:

- **Titolo** – Nome per esteso dell'indagine
- **Descrizione** – Descrizione dettagliata della rilevazione
- **CoefficienteMolestia** – Rappresenta un valore che definisce il peso statistico per la compilazione del modello da parte del rispondente.
- **NumeroSomministrazioni** – Rappresenta il numero delle somministrazioni del modello statistico alle unità selezionate.

Relazioni :

- Ogni indagine ha una o più estrazioni del campione.
- Ogni indagine ha uno o più strati campionari.

Tabella "Estrazioni"

Ogni selezione coordinata di un campione afferente ad una indagine viene rappresentata da una istanza nella tabella "Estrazioni".

Chiave primaria(ID_Indagine, Estrazione): rappresenta l'identificativo univoco della selezione composta dall'identificativo dell'indagine e da un progressivo dell'estrazione.

Attributi:

- **DataEstrazione** – Riferimento temporale dell'estrazione del campione
- **SequenceEstrazione** – Contatore progressivo temporale dell'estrazione nell'ambito dell'intero progetto.
- **AnnoRiferimento** - Anno di riferimento della versione di popolazione (*ASIA*).

Relazioni : Ad ogni estrazione vengono selezionate più unità campionarie

Tabella "*ASIA*"

La tabella "*ASIA*" rappresenta una vista dell'archivio *ASIA* (universo delle imprese).

Chiave primaria(CodiceImpresa): Codice identificativo dell'unità (o impresa) nella popolazione (*ASIA*).

Attributi:
CodiceFiscale – RagioneSociale – Regione – Provincia – Comune – Indirizzo – Cap – SiglaProvincia – DenominazioneComune – Telefono – DataInizio – DataFine – FormaGiuridica – Ateco1991 – AddettiIndipendenti – AddettiMedi – Addetti – VolumeAffari – Ateco.

Relazioni :

- Ogni unità della tabella *ASIA* deve avere un solo NumeroCasuale
- Ogni unità della tabella *ASIA* può avere più ricorrenze nella tabella PopolazioneRiferimento.

Tabella "NumeroCasuale"

Ad ogni unità della popolazione viene assegnato uno "score" come carico di molestia statistica associata.

Chiave primaria(ID_Popolazione): Codice identificativo dell'unità (o impresa) nella popolazione (*ASIA*).

Attributi: **NumeroCasuale** – Rappresenta l'ultimo numero casuale assegnato ad ogni unità risultante dalle permutazioni effettuate dal progetto.

Relazioni : Ogni unità nella popolazione deve avere un "NumeroCasuale".

Tabella "StratiIndagine"

Ad ogni versione dell'indagine vengono associate tante tuple quanti sono gli strati definiti nel disegno campionario.

Chiave primaria(ID_Indagine, Strato): rappresenta l'identificativo univoco dello strato definito nel disegno campionario d'indagine. Esso è composto dall'identificativo dell'indagine e da un progressivo dello strato nell'ambito dell'indagine in questione.

Attributi:

- **Descrizione** – Descrizione dell'insieme che definisce il singolo strato
- **NumeroUnitaAllocazioneOttima** – Indica la frequenza assoluta delle unità da selezionare nella popolazione in quello strato per formare il campione.
- **CondizioneWhere** – Indica la stringa SQL che permette di identificare l'insieme logico delle unità nella popolazione al fine della selezione delle unità campionarie nello strato. Ad esempio se si volesse determinare l'insieme logico delle imprese della ripartizione geografica sud con attività economica prevalente con classificazione Ateco a 2 cifre pari a 91 con più di 99 addetti ciò corrisponderebbe a :"Ripartizione='Sud' And Ateco='91' And Addetti>99"

Relazioni : Ogni indagine deve avere uno o più StratiIndagine.

Tabella "PopolazioneRiferimento"

Rappresenta la tabella di lavoro che viene utilizzata per l'estrazione delle unità del campione. In essa verranno popolate un numero d'istanze pari alla somma delle unità nella popolazione di tutti gli strati dell'indagine in fase di elaborazione.

Chiave primaria(ID_Popolazione, SequenceEstrazione): rappresenta l'identificativo univoco dell'unità nella popolazione per ogni singola estrazione. Esso è composto dall'identificativo della popolazione e dal progressivo delle estrazioni

.

Attributi:

- **Strato** – Identificativo dello strato dell'indagine cui appartiene l'unità.
- **CaricoStatisticoApplicato** – Rappresenta il valore del carico statistico applicato al seguito dell'estrazione dell'unità**.**
- **ID_Indagine** – Rappresenta l'identificativo dell'Indagine cui è riferita la unità
- **NumeroCasualeApplicato** – Numero casuale assegnato all'unità al termine delle permutazioni.

Relazioni :

- Ogni estrazione ha almeno una unità in PopolazioneRiferimento.
- Ogni unità presente in PopolazioneRiferimento deve essere presente in *ASIA*.

Tabella "StratiPopolazioni"

Rappresenta la tabella di lavoro che viene utilizzata per attribuire ad ogni impresa lo strato relativo ad ogni indagine in accordo con le condizioni di Where delle tabella StratiIndagini.

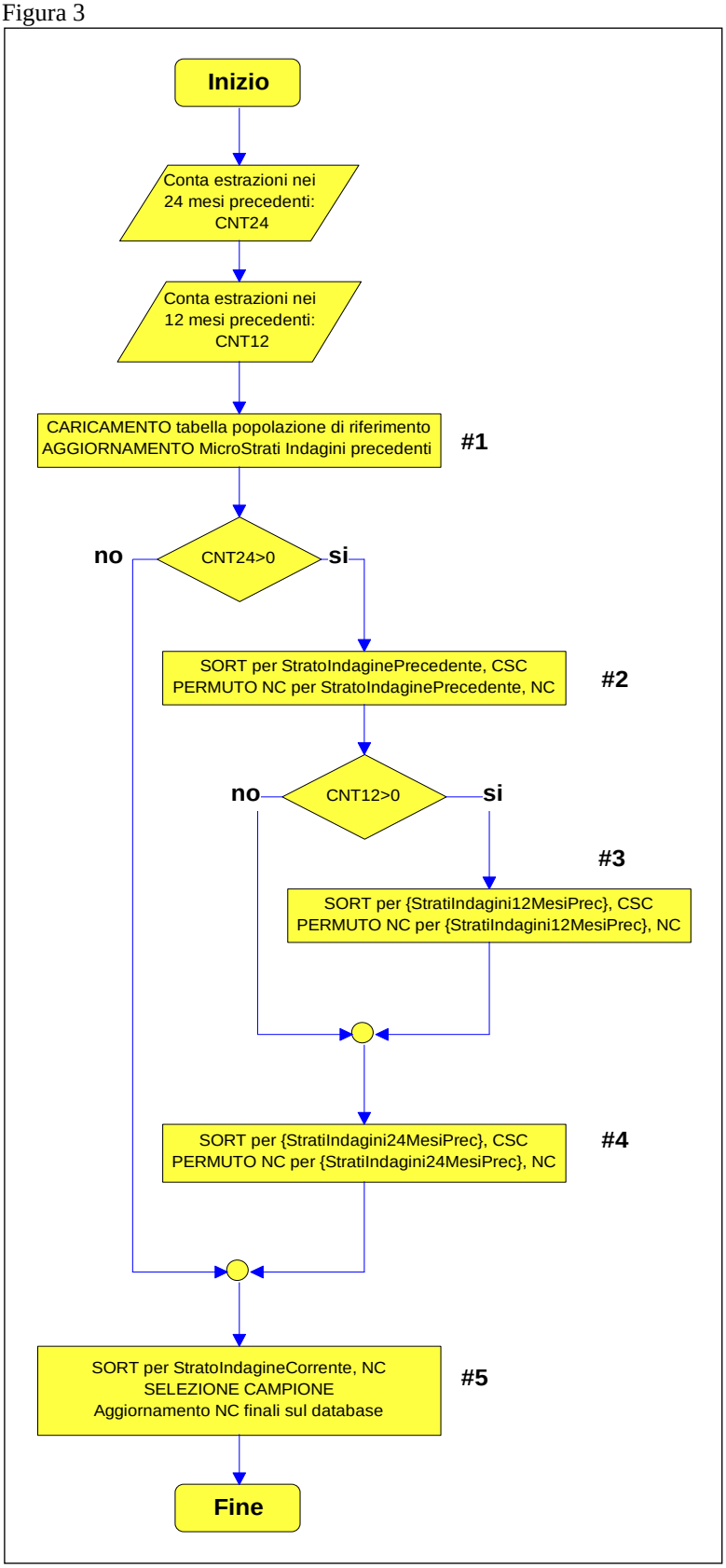
Chiave primaria(ID_Indagine, Strato, ID_Popolazione): rappresenta l'identificativo univoco della terna Indagine – Impresa – Strato.

Relazioni :

- Ogni StratiIndagine ha una o più unità in StratiPopolazioni.
- Ogni Impresa ha una o più unità in StratiPopolazioni.

7. I risultati della simulazione: riflessioni sui tempi di elaborazione

Figura 3



La realizzazione di un software di simulazione, data la complessità dell'algoritmo, ha evidenziato alcune criticità in particolar modo per l'indagine più corposa: la PMI. Infatti, con una popolazione di riferimento di quasi quattro milioni d'imprese e un'elevata stratificazione (25129 strati), determina un aumento delle microstratificazioni da considerare e, di conseguenza, i tempi di elaborazione risultano molto alti.

La simulazione è stata realizzata in ambito locale. I tempi presentati nel prospetto finale di questo documento (si veda *tavola 5*) non prevedono dunque la valutazione di eventuali rallentamenti dovuti all'utilizzo della rete.

In un'ottica di realizzazione di applicazione TP (in tempo reale) su rete geografica, pur non essendo in grado di valutare al momento il "ritardo" in termini di prestazione sulla presente architettura ISTAT, è possibile affermare che vi saranno tempi più lunghi d'estrazione del campione.

Basandosi solo su queste prime informazioni, non si può dire in generale che la validità del metodo risulti inficiata dai tempi di elaborazione. Infatti, avendo la maggior parte delle indagini campionarie dell'Istituto una popolazione di riferimento di circa 200.000 unità, la selezione del campione potrà essere effettuata in un tempo (accettabile) di circa un'ora.

E' pure da evidenziare che nella realizzazione dell'applicazione finale si cercherà di ottimizzare il processo di tutte quelle operazioni di I/O (Input/Output) critiche, definendo, per quanto possibile, strutture più adeguate a contenere i tempi elaborativi e definendo una serie di indici secondari che accelerino l'interrogazione del database.

Nel seguito del paragrafo si riportano i dettagli dei tempi di elaborazione, partendo dall'indagine Prodcom.
Tale indagine ha una numerosità, in termini di popolazione di riferimento, di 252082 imprese suddivise in 2200 strati; alla prima estrazione, non avendo precedenti campioni estratti, la procedura effettua la selezione del campione in soli 31 minuti. Al secondo anno di simulazione i microstrati derivanti dall'intersezione con le estrazioni effettuate dalla altre indagini diventano 12480, ciò provoca un tempo d'estrazione del campione di 56 minuti. Al terzo anno, dove le estrazioni da coordinare diventano 8, il tempo sale ad un'ora e 19 minuti.

Passando all'indagine sui posti vacanti, la popolazione di riferimento è di 188093 unità, con una stratificazione in termini di numerosità molto bassa: solo 57 strati.
Alla sua prima estrazione, coordinando con la prima estrazione della Prodcom, l'intersezione dei microstrati è di 2381. La determinazione del campione avviene in tempi del tutto accettabili, ovvero in poco più di 24 minuti. Nelle epoche successive le microstratificazioni crescono a 14934 e al crescere delle indagini da coordinare, i tempi al secondo e terzo anno sono rispettivamente di 47 e 62 minuti.

Come già accennato, l'indagine PMI ha i tempi più alti riscontrati nella selezione del campione. La sua popolazione pari a 3978612 imprese suddivise in 25129 strati, provoca tempi che crescono esponenzialmente al passare delle epoche d'estrazione, per effetto del coordinamento delle indagini che hanno effettuato la selezione del campione nei 24 mesi precedenti. Infatti per la prima estrazione si ha bisogno di 11 ore e 55 minuti tenendo conto anche delle 26894 intersezioni in termini di microstrati; per passare poi al secondo anno (coordinando 6 estrazioni) a un tempo pari a 15 ore e 42 minuti; e per finire al terzo anno con un tempo di 18 ore d'elaborazione per selezionare un campione coordinato a 8 indagini estratte nei 24 mesi precedenti.

L’indagine del commercio elettronico con la sua popolazione di 195450 unità, una stratificazione pari a 1966 riesce a effettuare le estrazioni in tempi accettabili. Nelle tre epoche d’estrazioni si passa da 36 a 60 a 82 minuti.

Infine per quanto riguarda la rilevazione sul costo del lavoro - unica a periodicità quadriennale nell’ambito delle indagini inserite nel calendario di simulazione e con una popolazione di riferimento di 187948 imprese e 1664 strati – al momento in cui questa viene coordinata con le 8 indagini estratte nei 24 mesi precedenti, si creano 15104 microstratificazioni: il tempo finale per effettuare la selezione è di un’ora e 13 minuti.
Nella tavola che segue viene schematizzato quanto detto in precedenza.

Tavola 5

Tempo	1	2	3	4	5	6
Indagine	ProdCom	PV	PMI	CE	ProdCom	PV
Data di simulazione	01/02/2002	15/02/2002	01/03/2002	02/05/2002	01/02/2003	15/02/2003
Popolazione di riferimento	252.082	188.093	3.978.612	195.450	252.082	188.093
Strati Indagine	2.200	57	25.129	1.966	2.200	57
Microstrati in totale	2.200	2.381	26.894	15.917	12.480	14.934
Estrazioni nei 24 mesi precedenti	-	1	2	3	4	5
Estrazioni nei 12 mesi precedenti	-	1	2	3	4	4
Caricamento dati (#1) in hh:mm:ss	0.01.44	0.00.22	0.23.02	0.02.24	0.02.32	0.01.42
Elab. Ind. Prec. (#2)	-	0.03.26	1.36.03	0.03.48	0.05.29	0.03.52
Elab. Ind. 12 mesi (#3)	-	0.03.40	1.30.47	0.04.07	0.05.30	0.03.47
Elab. Ind. 24 mesi (#4)	-	-	-	-	-	0.03.43
Sort e Selez. Campione etc.. (#5)	0.29.18	0.17.15	8.25.57	0.26.12	0.42.52	0.34.05
Tempo Totale dell'estrazione	**0.31.02**	**0.24.43**	**11.55.49**	**0.36.31**	**0.56.23**	**0.47.09**

Tempo	7	8	9	10	11	12	13
Indagine	PMI	CE	ProdCom	PV	PMI	CE	RCL
Data di simulazione	01/03/2003	02/05/2003	01/02/2004	15/02/2004	01/03/2004	02/05/2004	01/06/2004
Popolazione di riferimento	3.978.612	195.450	252.082	188.093	3.978.612	195.450	187.948
Strati Indagine	25.129	1.966	2.200	57	25.129	1.966	1.664
Microstrati in totale	26.894	15.917	12.480	14.934	26.894	15.917	15.104
Estrazioni nei 24 mesi precedenti	6	7	8	8	8	8	8
Estrazioni nei 12 mesi precedenti	4	4	4	4	4	4	4
Caricamento dati (#1) in hh:mm:ss	0.35.11	0.09.15	0.09.33	0.07.37	1.49.11	0.11.50	0.11.26
Elab. Ind. Prec. (#2)	1.43.11	0.04.05	0.05.19	0.03.29	1.47.41	0.04.08	0.03.56
Elab. Ind. 12 mesi (#3)	1.43.46	0.04.12	0.05.34	0.04.07	1.52.56	0.04.14	0.04.04
Elab. Ind. 24 mesi (#4)	1.42.45	0.04.09	0.05.24	0.04.03	1.49.19	0.04.12	0.03.57
Sort e Selez. Campione etc.. (#5)	9.57.24	0.39.04	0.53.55	0.43.12	10.42.03	0.57.36	0.50.28
Tempo Totale dell'estrazione	**15.42.17**	**1.00.45**	**1.19.45**	**1.02.28**	**18.01.10**	**1.22.00**	**1.13.51**

8. Conclusioni

Il gruppo di lavoro ha valutato una prima soluzione operativa per risolvere il problema del coordinamento dei campioni, implementando una procedura software generalizzata adattabile a tutte le indagini che estraggono i dati sulle imprese dall'archivio *ASIA*. In particolare è stata implementata una procedura informatica per simulare l'applicazione del metodo *Microstrat* alla realtà dell'Istituto, secondo un contesto sperimentale ben definito.

La procedura implementata rappresenta un primo importante passo ed è stata sviluppata in modo flessibile al fine di essere utilizzata per successive fasi di analisi, secondo contesti sperimentali più generali, e per adeguarsi a differenti contesti evolutivi dal punto di vista organizzativo.

Una problematica ampia come questa va affrontata con estrema attenzione, in quanto ha un impatto notevole sulla realtà organizzativa dell'Istituto e l'implementazione definitiva di un software generalizzato deve essere analizzata calandosi in un sistema più complesso e studiando i flussi che provengono da diversi archivi, in un'ottica di sistema informativo integrato.

L'ipotesi illustrata in questo documento prevede che lo sviluppo definitivo di una procedura informatica si basi su una struttura dati contenente il patrimonio informativo necessario all'applicazione del metodo, che potrebbe vedersi come parte di *ASIA* o come base dati indipendente. E' stato anche accennata l'ipotesi del coinvolgimento *SIDI*, in quanto alcune informazioni necessarie alla selezione coordinata potrebbero essere in futuro gestite dal Sistema di documentazione delle indagini. L'obiettivo di realizzare un prodotto integrato con altri Sistemi Informativi presenti dell'Istituto spinge verso una organizzazione dei dati secondo uno schema relazionale, in quanto gli altri archivi coinvolti nel progetto di selezione coordinata hanno organizzato le loro informazioni in basi dati relazionali, utilizzando l'*RDBMS ORACLE*.

Bibliografia

Di Giuseppe R., Giaquinto P., Pagliuca D. (2004), MAUSS (Multivariate Allocation of Units in Sampling Surveys): un software generalizzato per risolvere il problema dell'allocazione campionaria nelle indagini Istat, *Contributi Istat N. 7/2004.*

Ohlsson E. (1995) Coordination of samples using permanent random numbers, in Cox B.G., Binder D.A., Chinappa B.N., Christianson A., Colledge M.J., Kott P.S. (eds.), *Business Survey Methods, Willey, New York*

Riviere P. (2001), Coordinating samples using the microstrata methodology, *Proceedings of Statistics Canada Symposium 2001, "Achieving data quality in a statistical agency: a methodological perspective".*