istat
working
papers

N.14
2016

# Il processo di diffusione dei dati delle statistiche strutturali sulle imprese (Frame-SBS): aspetti normativi e metodologici connessi all'ampliamento del dettaglio informativo

*Carlo Boselli, Sabrina Brunetti, Mara Cammarrota, Viviana De Giorgi, Annamaria D'Urzo, Marco Ricci, Roberta Pazzini, Giovanni Seri, Giampiero Siesto e Luigi Virgili*

Istat
Istituto Nazionale
di Statistica

istat
working
papers

N.14
2016

# Il processo di diffusione dei dati delle statistiche strutturali sulle imprese (Frame-SBS): aspetti normativi e metodologici connessi all'ampliamento del dettaglio informativo

*Carlo Boselli, Sabrina Brunetti, Mara Cammarrota, Viviana De Giorgi, Annamaria D'Urzo, Marco Ricci, Roberta Pazzini, Giovanni Seri, Giampiero Siesto e Luigi Virgili*

Istat

## Istat Working Papers

Il processo di diffusione dei dati delle statistiche strutturali sulle imprese (Frame-SBS): aspetti normativi e metodologici connessi all'ampliamento del dettaglio informativo

N. 14/2016

ISBN 978-88-458-1905-6

# Il processo di diffusione dei dati delle statistiche strutturali sulle imprese (Frame-SBS): aspetti normativi e metodologici connessi all'ampliamento del dettaglio informativo[1]

*Carlo Boselli, Sabrina Brunetti, Mara Cammarrota, Viviana De Giorgi, Annamaria D'Urzo, Marco Ricci, Roberta Pazzini, Giovanni Seri, Giampiero Siesto e Luigi Virgili*

## Sommario

*La disponibilità di dati economici di fonte amministrativa e le rilevazioni sulle imprese condotte dall'Istat hanno consentito la costruzione del file di microdati Frame-SBS. Esso rappresenta la base di riferimento per produrre i dati richiesti dal Regolamento europeo n.295/2008 riguardante le statistiche strutturali sulle imprese (SBS). I dati integrati hanno permesso l'ampliamento del dettaglio informativo rilasciato. Il documento descrive la normativa di riferimento in materia di protezione dei dati personali, le scelte operative, gli aspetti metodologici e informatici e i canali attraverso i quali i dati vengono diffusi.*

**Parole chiave:** fonti amministrative, statistiche strutturali sulle imprese, tutela della riservatezza dei dati statistici, indicatori economici.

## Abstract

*The availability of administrative data for business statistics and data from business surveys carried out by the Italian National Institute of Statistics (Istat) has allowed the creation of a microdata file named Frame-SBS (Structural Business Statistics) that represents the basis for processing data in compliance with Eu Regulation no. 295/2008 on SBS. The large amount of integrated data has led to increase the detail level of the disseminated information. The present paper describes the reference legislation on statistical data confidentiality, operational decisions, methodological and IT aspects and dissemination channels.*

**Keywords:** administrative data, structural business statistics, statistical data confidentiality, economic indicators.

[1] Gli articoli pubblicati impegnano esclusivamente gli Autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat. Il lavoro è frutto dell'attività di tutti gli autori, in particolare sono da attribuire a Carlo Boselli il Paragrafo 6.1, a Sabrina Brunetti i Paragrafi 5.2 e 5.3, a Mara Cammarrota il Capitolo 1 e il Paragrafo 6.2, a Viviana De Giorgi i capitoli 2 e 3 e i programmi di calcolo degli indicatori descritti nel Capitolo 3, a Annamaria D'Urzo l'Introduzione del Capitolo 5 e il Paragrafo 5.1, a Marco Ricci i paragrafi 5.4, 5.5 e 5.6, a Roberta Pazzini il Paragrafo 6.3, a Giovanni Seri l'Appendice A, a Giampiero Siesto l'Introduzione, il Capitolo 2 e le Conclusioni e a Luigi Virgili l'Introduzione, i capitoli 1, 3 e 4, l'Appendice B e le Conclusioni.

# Indice

## Introduzione

Il documento sintetizza l'attività di lavoro della task-force "Diffusione e riservatezza", avente l'obiettivo di rilasciare i dati strutturali sulle imprese secondo il Regolamento n. 295/2008 (SBS), e di valutare la possibilità di una diffusione maggiormente dettagliata attraverso il data warehouse I.Stat.

A partire dalle unità dell'Archivio Statistico sulle Imprese Attive (Asia), anno di riferimento 2012, l'Istat ha realizzato il file di microdati Frame_SBS, contenente variabili strutturali (attività economica, localizzazione territoriale, numero di addetti e dipendenti) ed economiche (fatturato, costo di acquisto di beni e servizi, costo del personale, valore aggiunto e altre variabili più di dettaglio). Queste derivano da un processo di stima applicato ai dati di più fonti amministrative, come Camere di commercio, Agenzia delle Entrate e Inps.

Nel Frame-SBS sono presenti informazioni che permettono la classificazione delle unità secondo molteplici criteri, come ad esempio l'appartenenza a gruppi di imprese e/o lo svolgimento di attività di commercio con l'estero. Con riferimento all'anno 2013 il collettivo di riferimento è composto da 4.297.482 imprese, di cui 4.286.955 (99,7%) con meno di 100 addetti e 10.527 imprese con 100 addetti ed oltre (0,3%). Le variabili delle imprese con meno di 100 addetti derivano essenzialmente da un trattamento statistico delle fonti amministrative, che tiene conto delle risultanze della Rilevazione (campionaria) sulle piccole e medie imprese e sull'esercizio di arti e professioni (PMI), utilizzata in forma strumentale per la costruzione dei modelli di imputazione delle altre variabili di dettaglio del Frame. Le informazioni relative alle imprese con 100 addetti ed oltre derivano invece dalla Rilevazione (censuaria) sul sistema dei conti delle imprese (SCI).

Nel processo di definizione dei nuovi domini di diffusione, si è prestata particolare attenzione alla combinazione dell'attività economica e delle classi di addetti, e al calcolo di indicatori sulla distribuzione delle variabili in diversi domini (quantili e indici di variabilità), compatibilmente con le esigenze dettate dal Regolamento SBS anche in tema di trattamento della confidenzialità.

Nel corso del documento, si svilupperanno i seguenti argomenti: nel Capitolo 1 le disposizioni normative di riferimento; nel Capitolo 2 il Regolamento n. 295/2008, le fonti informative utilizzate per la costruzione del Frame-SBS e gli sviluppi in corso nell'adozione del Regolamento comunitario FRIBS; nel Capitolo 3 le scelte operative volte ad ampliare il dettaglio di diffusione dei dati ed i vincoli imposti dal trattamento della riservatezza; nel Capitolo 4 gli aspetti metodologi e informatici del trattamento dei dati Frame-SBS in una logica di processo che predispone i macrodati al trattamento della riservatezza con il software generalizzato τ-Argus per arrivare a trasmettere le serie di dati all'Eurostat e a predisporre e diffondere i dati attraverso il data warehouse I.Stat dell'Istituto; nel Capitolo 5 i canali attraverso i quali l'Istituto mette a disposizione degli utenti i dati statistici in forma tabellare e di dato elementare, fornendo inoltre una descrizione del sito web dell'Istat.

## 1. Quadro normativo di riferimento in materia di protezione dei dati personali

La diffusione dei dati relativi alle statistiche strutturali sulle imprese viene realizzata in ottemperanza alle disposizioni normative previste in materia di protezione di dati personali e al Regolamento n.295/2008 (Capitolo 2).

In base all'art. 8 del Codice in materia di protezione dei dati personali (Decreto legislativo 30 giugno 2003, n. 196) è consentito diffondere[2] anche mediante pubblicazione risultati soltanto in forma aggregata ovvero secondo modalità che garantiscano la tutela della riservatezza dei rispondenti. Anche il Decreto legislativo n. 322 del 1989 (art. 9, comma 1) riporta una precisa indicazione

[2] Con il termine diffusione si intende il dare conoscenza dei dati personali a soggetti indeterminati in qualunque forma, anche mediante la loro messa a disposizione o consultazione (decreto legislativo n. 196 del 2003, art. 4, comma 1, lettera m).

con riferimento ai dati raccolti nell'ambito di rilevazioni statistiche comprese nel Programma statistico nazionale (Psn): tali dati non possono essere esternati se non in forma aggregata.

La violazione della riservatezza statistica si verifica quando, utilizzando i dati diffusi, un soggetto (intruder) ottiene informazioni riservate su un altro soggetto, l'interessato. Col termine informazioni riservate si intendono tutte le informazioni che le unità rilevate hanno interesse a non divulgare e/o che gli istituti di statistica e gli uffici del Sistema statistico nazionale (Sistan) si sono impegnati a mantenere anonime per vincoli legali (ma anche per mantenere un rapporto di fiducia con gli intervistati). Rientrano in questa definizione anche i dati sensibili3 e i dati giudiziari4, mentre non sono considerate riservate le variabili pubbliche5.

Quando l'informazione statistica da rilasciare coinvolge variabili riservate, occorre valutare se esista un rischio di violazione della riservatezza. Si distinguono due tipologie di violazione:

- di identificazione, quando un'unità statistica viene identificata (nel caso di persone fisiche gli viene attribuito un nome e cognome) e le sono associate le informazioni riservate;
- di attributo, quando le informazioni riservate sono attribuite in modo certo ad una o più unità statistiche (in genere appartenenti ad un gruppo).

Con riferimento a quanto specificato nel Codice di deontologia, il rischio di violazione della riservatezza deve essere commisurato al danno associato ad una eventuale identificazione. La scelta dei livelli di protezione dipende dalla tipologia dei dati da rilasciare, ad esempio i dati sensibili relativi a individui (come le informazioni sullo stato di salute) devono prevedere un rischio di identificazione inferiore rispetto ad altri dati riservati (come ad esempio i dati di impresa).

Il Decreto legislativo n. 322 del 1989 e il Decreto legislativo n. 196 del 2003 sanciscono la deroga al segreto statistico. Essi prevedendo l'interscambio di dati individuali all'interno del Sistan se necessari alle esigenze statistiche previste dal Psn oppure per consentire il perseguimento degli scopi istituzionali dell'ente di appartenenza. In particolare, la direttiva "Criteri e modalità per la comunicazione dei dati personali nell'ambito del Sistema statistico nazionale" del Comitato di indirizzo e coordinamento dell'informazione statistica (Comstat) (Direttiva n. 9 del 20 aprile 2004) prevede che un ente o ufficio di statistica facente parte del Sistan possa richiedere ad un altro soggetto del Sistema dati personali già acquisiti per finalità statistiche. Le forniture di dati personali corredati di identificativi è, comunque, limitata ai casi di assoluta e stretta necessità ovvero di impossibilità a raggiungere l'obiettivo prefissato senza i dati identificativi. Gli enti Sistan devono formulare la richiesta di dati attraverso il Contact Centre dell'Istat, il sistema web per l'acquisizione ed il trattamento on line delle richieste di informazioni statistiche e dei servizi di diffusione, e deve essere redatta su apposito modello, indicando in modo dettagliato le motivazioni, le finalità perseguite e la pertinenza e non eccedenza dei dati richiesti rispetto alle finalità dichiarate, nonché, qualora siano richiesti anche dati identificativi, la stretta necessità dei medesimi.

Per quanto riguarda invece i soggetti non facenti parte del Sistan, l'art. 7 del Codice in materia di protezione dei dati personali (Decreto n. 196/2003) stabilisce che sia possibile comunicare6 file di dati individuali privi di ogni riferimento che ne permetta il collegamento con gli interessati secondo modalità che rendano questi ultimi non identificabili. In tali file l'anonimità delle unità stati-

[3] I dati sensibili sono rappresentati dai dati personali idonei a rivelare l'origine razziale ed etnica, le convinzioni religiose, filosofiche o di altro genere, le opinioni politiche, l'adesione a partiti, sindacati, associazioni od organizzazioni a carattere religioso, filosofico, politico o sindacale, nonché i dati personali idonei a rivelare lo stato di salute e la vita sessuale (art. 4 del Decreto legislativo 30 giugno 2003, n. 196).

[4] I dati giudiziari sono rappresentati dai dati personali idonei a rivelare provvedimenti di cui all'articolo 3, comma 1, lettere da a) a o) e da r) a u), del D.p.r. 14 novembre 2002, n. 313, in materia di casellario giudiziale, di anagrafe delle sanzioni amministrative dipendenti da reato e dei relativi carichi pendenti, o la qualità di imputato o di indagato ai sensi degli articoli 60 e 61 del codice di procedura penale (art. 4 del Decreto legislativo 30 giugno 2003, n. 196).

[5] Le variabili pubbliche sono definite come il carattere o la combinazione di caratteri, di tipo qualitativo o quantitativo, oggetto di una rilevazione statistica che faccia riferimento ad informazioni presenti in pubblici registri, elenchi, atti, documenti o fonti conoscibili da chiunque (definizione contenuta nel Codice di deontologia) (art. 2 dell' allegato A.3 del Decreto legislativo 30 giugno 2003, n. 196).

[6] Con il termine comunicazione si intende il dare conoscenza dei dati personali a uno o più soggetti determinati diversi dall'interessato, dal rappresentante del titolare nel territorio dello Stato, dal responsabile e dagli incaricati, in qualunque forma, anche mediante la loro messa a disposizione o consultazione (Decreto legislativo n. 196 del 2003, art. 4, comma 1, lettera l).

stiche, ovviamente già prive d'identificativi diretti, viene tutelata tramite l'applicazione di diverse metodologie statistiche che rendono altamente improbabile l'identificazione indiretta delle unità statistiche.

Attualmente l'Istat affianca alla diffusione di dati aggregati, diverse tipologie di file di microdati anonimizzati:

- File standard, rilasciati a chiunque ne faccia richiesta motivata, per finalità di studio e ricerca, e sono prodotti per alcune rilevazioni dell'Istituto su individui e famiglie;
- File per la ricerca (MFR - Microdata files for research), prodotti per rilevazioni statistiche riguardanti sia individui e famiglie sia imprese; sono realizzati specificamente per esigenze di ricerca scientifica e quindi contengono un maggiore livello di dettaglio informativo rispetto ai File standard. Il rilascio di tali file è soggetto alla sussistenza di alcuni requisiti, relativi sia all'organizzazione di appartenenza sia alle caratteristiche del progetto di ricerca per le cui finalità viene richiesto il file;
- File mIcro.STAT, file ad uso pubblico scaricabili direttamente dal sito Istat. Sono ottenuti applicando (ulteriori) misure di protezione ai file per la ricerca: il contenuto informativo dei mIcro.STAT è un sottoinsieme rispetto a quello degli MFR.

I requisiti e le condizioni per il rilascio dei file variano a seconda dei soggetti che li richiedono e sono subordinati alla sottoscrizione di precisi accordi di utilizzo.

La comunicazione di dati personali a ricercatori di università o a istituti o enti di ricerca o a soci di società scientifiche a cui si applica il codice di deontologia e di buona condotta per i trattamenti di dati personali per scopi statistici e di ricerca scientifica effettuati fuori dal Sistan, di cui all'articolo 10, comma 6, del decreto legislativo 30 luglio 1999, n. 281 e successive modificazioni e integrazioni, è consentita nell'ambito di specifici laboratori costituiti da soggetti del Sistan, a condizione che:

a) i dati siano il risultato di trattamenti di cui i medesimi soggetti del Sistema statistico nazionale siano titolari;
b) i dati comunicati siano privi di dati identificativi;
c) le norme in materia di segreto statistico e di protezione dei dati personali, contenute anche nel presente codice, siano rispettate dai ricercatori che accedono al laboratorio anche sulla base di una preventiva dichiarazione di impegno;
d) l'accesso al laboratorio sia controllato e vigilato;
e) non sia consentito l'accesso ad archivi di dati diversi da quello oggetto della comunicazione;
f) siano adottate misure idonee affinché le operazioni di immissione e prelievo di dati siano inibite ai ricercatori che utilizzano il laboratorio;
g) il rilascio dei risultati delle elaborazioni effettuate dai ricercatori che utilizzano il laboratorio sia autorizzato solo dopo una preventiva verifica, da parte degli addetti al laboratorio stesso, del rispetto delle norme di cui alla lettera c).

Per rispondere alla crescente richiesta di dati elementari da parte del mondo della ricerca scientifica, è stato creato il Laboratorio per l'Analisi dei Dati ELEmentari (Laboratorio ADELE), ubicato presso la sede di Roma e le sedi territoriali dell'Istat. Il Laboratorio ADELE è un Research Data Center (RDC), cioè un luogo sicuro dove ricercatori e studiosi di università, istituti o enti di ricerca possono effettuare analisi statistiche sui microdati delle indagini dell'Istituto. Per promuovere l'ampliamento delle informazioni a livello di singola impresa, presso il Laboratorio sono anche disponibili file derivanti dall'integrazione di più fonti.

I soggetti che si recano al Laboratorio ADELE devono osservare le regole indicate nel Codice di deontologia e di buona condotta per i trattamenti di dati personali per scopi statistici e scientifici (effettuati al di fuori del Sistan) (allegato A.4 del D.lgs. 30 giugno 2003, n. 196).

Per poter accedere al Laboratorio ADELE il soggetto richiedente deve appartenere ad una università o ad altro istituto di ricerca; deve presentare un progetto indicando i dati che intende elaborare e gli obiettivi della ricerca. L'autorizzazione è firmata dal Presidente dell'Istat. Prima di accedere al Laboratorio l'utente sottoscrive un contratto che lo obbliga al mantenimento del segreto statistico. Il ricercatore conduce autonomamente le proprie elaborazioni sui dati elementari richiesti(privi di identificativi diretti e delle variabili sensibili e

giudiziarie). Alla conclusione del progetto il risultato delle elaborazioni è valutato dal personale preposto che, prima del rilascio, ne verifica la conformità alle regole di tutela della riservatezza.

Nel laboratorio non è prevista assistenza metodologica/tecnica agli utenti; il servizio è destinato ad un'utenza specializzata in grado di individuare la rilevazione statistica di interesse, utilizzare gli strumenti hardware e software messi a disposizione e interpretare i dati e le elaborazioni realizzate.

Le condizioni di utilizzo del Laboratorio, le modalità di accesso e le regole di rilascio dell'output sono condivise nelle linee essenziali tra i Paesi europei, ed incluse in un processo di armonizzazione a livello internazionale.

L'art. 7 del Codice in materia di protezione dei dati personali (Decreto legislativo 30 giugno 2003, n. 196) prevede anche un'ulteriore possibilità di comunicazione di dati da parte di soggetti del Sistan a ricercatori operanti per conto di università, altre istituzioni pubbliche e organismi aventi finalità di ricerca, nell'ambito di progetti congiunti, finalizzati anche al perseguimento di compiti istituzionali del titolare del trattamento che ha originato i dati. Tale comunicazione è possibile purché sia garantito il rispetto delle condizioni seguenti:

- i dati siano il risultato di trattamenti di cui i medesimi soggetti del sistema statistico nazionale sono titolari;
- i dati comunicati siano privi di informazioni identificative, sensibili e giudiziarie;
- la comunicazione avvenga sulla base di appositi protocolli di ricerca sottoscritti da tutti i ricercatori che partecipano al progetto. Nei protocolli siano esplicitamente previste, come vincolanti, le norme in materia di segreto statistico e di protezione dei dati personali contenute nel Codice.

Ai ricercatori è vietato effettuare trattamenti per fini diversi da quelli esplicitamente previsti dal protocollo di ricerca, di conservare i dati comunicati oltre i termini di durata del progetto, di comunicare i dati a terzi.

## 2. Regolamento sulle statistiche strutturali sulle imprese (SBS), fonti informative, serie da trasmettere e sviluppi dei regolamenti comunitari

Le statistiche strutturali sulle imprese (SBS) sono disciplinate, a partire dall'anno di riferimento 2008, dal Regolamento n. 295/2008, adottato l'11 marzo 2008 dal Consiglio e dal Parlamento, il cui obiettivo è quello di istituire un quadro comune per la raccolta, l'elaborazione, la trasmissione e la valutazione delle statistiche comunitarie sulla struttura, l'attività, la competitività delle imprese nella comunità. I regolamenti n. 250/2009 e n. 251/2009 dell'11 marzo 2009 attuano il regolamento SBS per la definizione delle variabili, il formato tecnico per la trasmissione dei dati e per le serie che devono essere prodotte e trasmesse a Eurostat.

Il regolamento SBS si sviluppa in nove annessi, per ciascuno dei quali sono indicate le serie e la disaggregazione dei dati da trasmettere. Nella tavola seguente sono indicati il campo di osservazione di ciascun annesso e l'ente che produce e trasmette i dati.

**Tavola 1 – Annessi del regolamento SBS per campo di osservazione ed ente produttore**

| Annessi | Campo di osservazione (Nace Rev.2) | Ente produttore |
|---|---|---|
| Annesso 1 | Servizi (sezioni H-J, L-N e divisione S95 | Istat |
| Annesso 2 | Industria (sezioni B-E) | Istat |
| Annesso 3 | Commercio (sezione G) | Istat |
| Annesso 4 | Costruzioni (sezione F) | Istat |
| Annesso 5 | Assicurazioni (divisione K65, escluso gruppo K653) | Ivass (ex Isvap) |

**Tavola 1 – *segue* – Annessi del regolamento SBS per campo di osservazione ed ente produttore**

| Annessi | Campo di osservazione (Nace Rev.2) | Ente produttore |
|---|---|---|
| Annesso 6 | Enti creditizi (divisione K64) | Banca d'Italia |
| Annesso 7 | Fondi pensione (gruppo K653) | Covip |
| Annesso 8 | Servizi alle imprese (alcune Nace con cadenza annuale e altre con cadenza biennale) | Istat |
| Annesso 9 | Demografia delle imprese | Istat |

Con riferimento agli annessi strettamente di pertinenza dell'Istat, i principali domini di stima degli annessi 1-4 sono i seguenti:

a) Nace Rev.2[7] a quattro cifre senza distinzione per classi di addetti;
b) Nace Rev.2 a tre cifre per classi di addetti (0-9, 10-19, 20-49, 50-249, 250+ nell'industria e costruzioni; 0-1, 2-9, 10-19, 20-49, 50-249, 250+ per il commercio e i servizi);
c) Nace Rev.2 a due cifre (tre cifre per il commercio) per regione a livello di Nuts2

Per l'annesso 8 i dati da trasmettere riguardano le imprese con 20 addetti e oltre di alcune specifiche attività economiche a cui vengono chieste informazioni sul fatturato per tipo di prodotto e per nazionalità del cliente. Alcune attività sono investigate annualmente (Nace: 582, 62, 631, 731 e 78) mentre altre ogni due anni (Nace: 691, 692 e 702 negli anni pari; Nace: 7111, 7112, 712, e 732 negli anni dispari).

Per l'annesso 9, i dati sulla demografia delle imprese sono richiesti con un dettaglio Nace Rev.2 fino alla quarta cifra, disaggregando i dati per forma giuridica e classi di dipendenti.

Riguardo gli annessi 1-4, i dati preliminari devono essere trasmessi entro 10 mesi dalla fine dell'anno di riferimento (con un dettaglio a livello di Nace Rev.2 a tre cifre) mentre i dati definitivi devono essere trasmessi al dettaglio descritto entro 18 mesi dalla fine dell'anno di riferimento.

Nella Tavola 2 sono indicate le principali serie da trasmettere, che devono essere congiuntamente trattate per la definizione della riservatezza, identificando i domini con confidenzialità primaria (ovvero con una numerosità di imprese inferiore a 3) e quelli che devono essere considerati con confidenzialità secondarie a causa della classificazione gerarchica dell'attività economica e della disaggregazione dei dati (ad esempio per classe di addetti). Della confidenzialità e del trattamento con un software specifico τ-Argus si rimanda al Capitolo 4.

**Tavola 2 – Principali serie degli annessi 1-4 e 8 da trasmettere a Eurostat per il regolamento SBS**

| Serie | Annessi di riferimento e descrizione |
|---|---|
| | **Servizi, Industria, Commercio e Costruzioni** |
| 1A 2A 3A 4A | Statistiche annuali sulle imprese (Nace a 4 cifre) |
| 1B 2B 3B 4B | Statistiche annuali sulle imprese per classe di addetti (Nace a 3 cifre)<br>*Classi di addetti: 0-1, 2-9, 10-19, 20-49, 50-249, 250+, totale per le serie 1B e 3B*<br>*Classi di addetti: 0-9, 10-19, 20-49, 50-249, 250+, totale per le serie 2B e 4B* |
| 1C 2C 3C 4C | Statistiche annuali regionali per Nuts2 (Nace a 2 cifre per Servizi, Industria e Costruzioni; Nace a 3 cifre per Commercio) |
| 1P 2P 3P 4P | Statistiche annuali preliminari sulle imprese (Nace a 3 cifre) |
| | **Servizi** |
| 1E | Statistiche annuali sulle imprese per aggregati speciali |
| | **Industria, Costruzioni** |
| 2D 4D | Statistiche annuali sulle Kau* (Nace a 4 cifre) |

* Kau (Kind Activity Unit) = unità economica omogenea

[7] Nace Rev.2 è la classificazione europea di riferimento per le attività economiche, che si sviluppa per sezione (1 lettera), divisione (2 cifre), gruppo (3 cifre) e classe (4 cifre) di attività economica . La corrispondente versione italiana sviluppata ed utilizzata dall'Istat è l'Ateco 2007 che rispetto alla Nace contiene ulteriori disaggregazioni rappresentate da categoria (5 cifre) e sottocategoria (6 cifre).

**Tavola 2 –** *segue* **– Principali serie degli annessi 1-4 e 8 da trasmettere a Eurostat per il regolamento SBS**

| Serie | Annessi di riferimento e descrizione |
|---|---|
| 2E 4E | Statistiche multiannuali sulle imprese - investmenti intangibili (Nace a 4 cifre) |
| 2F 4F | Statistiche multiannuali sulle imprese - subfornitura (Nace a 4 cifre) |
| 2G 4G | Statistiche multiannuali sulle imprese – classi di fatturato (Nace a 3 cifre) |
| | **Industria** |
| 2H, 2J | Statistiche annuali sulle imprese sulle spese di protezione ambientale disaggregate per dominio ambientale (Nace a 2 cifre) |
| 2I, 2K | Statistiche annuali sulle imprese sulle spese di protezione ambientale per classe di addetti (Nace a 2 cifre)<br>*Classi di addetti: 0-49, 50-249, 250+, totale* |
| | **Commercio** |
| 3D | Statistiche annuali sulle imprese per classi di fatturato (Nace a 3 cifre) |
| | **Costruzioni** |
| 4H | Statistiche multiannuali sulle imprese - subfornitura per classe di addetti (Nace a 3 cifre)<br>Classi di addetti: 0-9, 10-19, 20-49, 50-249, 250+, totale |
| | **Servizi alle imprese** |
| 8A | Statistiche annuali sulle imprese per attività Nace Rev.2 (62, 582, 631, 731 e 78) per tipo di prodotto offerto |
| 8B | Statistiche annuali sulle imprese per attività Nace Rev.2 (62, 582, 631, 731 e 78) per residenza del cliente |
| 8C | Statistiche annuali sulle imprese per attività Nace Rev.2 (691, 692 e 702) per tipo di prodotto offerto |
| 8D | Statistiche annuali sulle imprese per attività Nace Rev.2 (691, 692 e 702) per residenza del cliente |
| 8E | Statistiche biennali sulle imprese per attività Nace Rev.2 (732, 711 e 712) per tipo di prodotto offerto |
| 8F | Statistiche biennali sulle imprese per attività Nace Rev.2 (732, 711 e 712) per residenza del cliente |

Le variabili più significative richieste negli annessi 1-4 sono le seguenti:

11100 numero di imprese
12110 fatturato
12120 valore della produzione
12130 margine lordo sui beni destinati alla rivendita
12150 valore aggiunto al costo dei fattori
12170 margine operativo lordo
13110 acquisto complessivo di beni e servizi
13120 acquisto di merci da rivendere senza trasformazione
13210 variazione delle scorte di beni e servizi
13211 variazione delle scorte di beni e servizi destinati alla rivendita
13310 costo del personale
13320 retribuzione lorda
16150 ore lavorate
15110 investimenti lordi in beni materiali
16110 numero di persone occupate
16130 numero di dipendenti.

Le fonti informative utilizzate dall'Istat per adempiere al regolamento SBS sono state fino all'anno di riferimento 2011, la rilevazione sulle piccole e medie imprese e sull'esercizio di arti e professioni (PMI, campionaria sulle imprese con 1-99 addetti) e la rilevazione sul sistema dei conti delle imprese (SCI, totale per le imprese con 100 addetti e oltre).

A partire dall'anno di riferimento 2012 i principali aggregati sulle imprese con meno di 100 ad-

detti sono stimati attraverso l'elaborazione di un file di microdati (Frame) costruito mediante il trattamento statistico delle informazioni provenienti da diverse fonti amministrative (Bilanci civilistici delle Camere di commercio; Studi di settore, Modello Unico e Modello Irap dell'Agenzia delle entrate; Registro Annuale del Costo del lavoro per Impresa della fonte Inps-Emens). Gli aggregati di fonte Frame quindi sono ottenuti attraverso un processo di somma delle singole variabili nei domini di interesse (attività economica, classi di addetti, regione, ecc.); per contro, le stime della rilevazione campionaria PMI per le variabili non disponibili dalle fonti amministrative sono ottenute moltiplicando le variabili per un peso finale, processo che consente di ottenere dati significativi solo per i domini programmati. Alle stime ottenute attraverso Frame/PMI si aggiungono quelle tradizionale della rilevazione SCI sulle imprese con 100 addetti e oltre che consente di costruire il file di microdati Frame-SBS.

Il campo di osservazione delle rilevazioni PMI e SCI, e quindi del Frame-SBS, è più ampio di quanto previsto dal regolamento SBS, includendo anche le sezioni P (istruzione), Q (sanità e assistenza sociale), R (attività artistiche, sportive e di intrattenimento) e la divisione 96 («altre attività di servizi per la persona») i cui dati sono diffusi attraverso il data warehouse dell'Istituto I.Stat.

Sotto l'aspetto tecnico, l'elaborazione congiunta Frame/PMI/SCI porta a costruire dei dati a livello macro per dominio, da cui si innesta il processo di trattamento della confidenzialità con τ-Argus e un processo informatico che porta a costruire le serie dei diversi annessi del regolamento. Queste serie sono poi controllate attraverso il software generalizzato EBB_Tool sviluppato dall'Eurostat che definisce un controllo di qualità sui dati all'interno di ciascuna serie, fra le diverse serie di ciascun annesso e sotto l'aspetto longitudinale con un confronto delle serie da trasmettere con quelle trasmesse l'anno precedente. Il processo SBS si conclude con la trasmissione delle serie all'Eurostat attraverso il sistema Edamis.

Al fine di consentire all'Eurostat il calcolo degli aggregati dei totali a livello Ue, i dati SBS vengono trasmessi in forma esplicita ovvero senza oscurare nulla, ma con l'indicazione dei domini che, per ragioni legate all'applicazione delle regole di riservatezza, non potranno essere diffusi a livello nazionale.

Nell'ottica di conseguire gli obiettivi della comunicazione 404 del 2009 della Commissione europea sui metodi di produzione di statistiche nella Ue nel decennio 2010-2020 e del programma statistico europeo 2013-2017 nel settore delle statistiche delle imprese, l'Eurostat ha lanciato il progetto FRIBS (*Framework Regulation Integrating Business Statistics*) volto a definire un quadro normativo armonizzato per la raccolta, la trasmissione e la diffusione di statistiche europee sulla struttura, l'attività, la competitività, le transazioni globali e le performance delle imprese.

Nel progetto del regolamento FRIBS, verranno assorbiti il regolamento sulle statistiche strutturali sulle imprese (SBS, *Structural Business Statistics*), il regolamento sulle statistiche congiunturali mensili e trimestrali (STS, *Short Term Statistics*), il regolamento sulle attività estere delle imprese a controllo nazionale e sulle attività delle imprese a controllo estero residenti nel Paese (FATS, *Foreign Affiliates Trade Statistics*), il regolamento sugli investimenti diretti esteri (FDI, *Foreign Direct Investment*), il regolamento sulle statistiche sulle società dell'informazione (ISS, *Informan Society Statistics*), il regolamento sulle statistiche sulla ricerca e sviluppo (R&D) e l'innovazione, il regolamento sulle statistiche del commercio internazionale di beni (ITGS, *International Trade in Goods Statistics*) e il regolamento sulla produzione commercializzata di prodotti manifatturati (ProdCom). Fra i principali obiettivi vi sono da una parte quelli di razionalizzare il complesso quadro normativo sulle statistiche europee sulle imprese e di definire una nuova architettura di compilazione di statistiche sulle imprese (sfruttando maggiormente, e in forma integrata, le informazioni disponibili nelle fonti amministrative al fine di ridurre l'onere statistico sulle imprese) e dall'altro di migliorare la qualità delle statistiche nel settore dei servizi, sulla globalizzazione e sulla imprenditorialità. Il regolamento FRIBS è al momento nella fase di definizione del testo legale e degli atti implementativi (conclusione prevista a inizio 2016), quindi passerà attraverso le fasi di consultazione ed approvazione da parte della Commissione Europea (fine 2016), il processo di adozione del testo legale da parte del Consiglio e del Parlamento Europeo e la definizione degli atti implementativi da parte del Comitato del sistema statistico europeo (ESSC-*European Statistical System Committee*) per entrare in forza probabilmente nel corso del 2018.

Nel frattempo sono state avviate le prime discussioni su come evolvere l'attuale Regolamento SBS, orientato esclusivamente alla produzione nazionale, verso un nuovo regolamento SBS2020 più orientato agli aspetti della globalizzazione e internazionalizzazione, introducendo nuove dimensioni nella produzione di statistiche come ad esempio sul controllo delle imprese (che dovrebbe portare a fornire statistiche disaggregate a seconda se tali unità indipendenti oppure a controllo nazionale oppure a controllo estero) oppure sull'apertura al commercio estero (classificando le imprese a seconda se operano o meno con l'estero). La produzione di statistiche è quindi in continua evoluzione e attraverso la valorizzazione dei dati di fonte amministrativa si cercherà di ridurre il fastidio statistico sulle imprese e di soddisfare le crescenti esigenze informative che provengono dagli utilizzatori, dagli *stakeholders* e dai *policy makers*. Ovviamente oltre ad apportare innovazione nei processi di produzione statistica occorrerà riprogettare il complesso processo di trattamento della riservatezza complessivo dei dati ai fini della diffusione dei dati nelle nuove dimensioni che saranno ritenute necessarie alle esigenze conoscitive di certi fenomeni.

## 3. Ampliamento del dettaglio informativo

Per far fronte alla crescente richiesta di dati economici, dall'anno di riferimento 2013, l'Istat ha aumentato il contenuto informativo nella diffusione dei dati Frame-SBS. Per le cosiddette *variabili core*[8]*,* l'utilizzo di archivi amministrativi ha permesso di ottenere informazioni sufficienti per aumentare il dettaglio dei domini di diffusione. Per tali variabili le classi di addetti [0-1] e [2-9] sono state adottate anche per i settori dell'Industria e delle Costruzioni e le informazioni sono state pubblicate aggiungendo la disaggregazione Nace a 4 cifre per classi di addetti.

Per i domini a 1 lettera e a 2, 3 e 4 cifre di attività economica che comprendono almeno 50 unità della popolazione, sono state inoltre rilasciate:

- frequenze assolute delle imprese con almeno un addetto, con almeno un dipendente, con fatturato non nullo.
- quartili e deviazione standard delle seguenti variabili: fatturato per addetto, costo del lavoro per dipendente, valore aggiunto per addetto, integrazione verticale[9], competitività di costo[10]. I rapporti sono stati calcolati escludendo dalle elaborazioni i dati che presentano valore nullo nella variabile al denominatore (relativamente al costo del lavoro per dipendente sono state escluse tutte le imprese prive di dipendenti).
- valori medi dei quartili[11] e deviazione standard delle seguenti variabili: numero di addetti, numero di dipendenti, fatturato, costo del lavoro (escludendo le imprese prive di dipendenti), valore aggiunto e margine operativo lordo. La ragione per cui sono stati rilasciati valori medi in luogo dei quartili della distribuzione, risiede nel fatto che le regole di tutela della riservatezza obbligano l'Istituto a non divulgare informazioni puntuali afferenti a singole unità del collettivo di riferimento.

Il ricorso a valori medi della distribuzione (quartili o loro medie), non influenzati dalle osservazioni estreme, è stato dettato anche dal fatto che i fenomeni economici indagati mostrano andamenti fortemente asimmetrici (con asimmetria positiva).

L'indicatore di variabilità è d'ausilio nell'analisi dell'eterogeneità.

Il programma di calcolo degli indicatori è stato sviluppato in Sas. Il file risultante contenente gli indici di posizione e la deviazione standard è stato opportunamente modificato così da risultare im-

[8] Dove per variabili core si intendono: numero di imprese, numero di addetti, numero di dipendenti, ricavi delle vendite e delle prestazioni (fatturato), altri ricavi e proventi, costi per acquisti di beni e servizi, costi per acquisti di materie prime, costi per acquisti di servizi, costi per godimento beni di terzi, costi per oneri diversi di gestione, costi del personale, retribuzione lorda totale, valore aggiunto e margine operativo lordo.

[9] Rapporto tra valore aggiunto e fatturato per 100.

[10] Rapporto tra valore aggiunto per addetto e costo del lavoro per dipendente per 100.

[11] I valori medi sono ottenuti operando una microaggregazione dei 5 (6) valori centrati sui quartili, quando la numerosità della popolazione di riferimento è dispari (pari).

mediatamente disponibile in ambiente Oracle, secondo il seguente tracciato record: dominio, nome variabile, valore della variabile, flag di rilasciabilità.

I valori medi delle distribuzioni (quartili), calcolati come descritto, non sono confrontabili con altri indicatori determinati rapportando i totali di dominio; ad esempio, il rapporto tra il totale del valore aggiunto e il totale degli addetti in un determinato dominio, fornisce un valore che in termini numerici può essere anche molto distante dal valore medio o mediano calcolato sui rapporti riferiti alle singole imprese. Ciò in quanto la distribuzione del fenomeno è verosimilmente asimmetrica (con valori estremi elevati) e a livello "micro" sono esclusi i rapporti con denominatore nullo (esempio, addetti pari a zero).

## 4. Aspetti metodologici del trattamento dei dati SBS

### 4.1 Predisposizione dei dati SBS per il trattamento della riservatezza

L'obiettivo dei sistemi nazionali di statistica di diffondere i risultati delle rilevazioni al maggior dettaglio disponibile deve essere coniugato con l'obbligo di garantire la riservatezza delle unità intervistate. In Italia, il quadro normativo di riferimento è rappresentato dal D.lgs 322/89 e dal "Codice di deontologia e di buona condotta per i trattamenti di dati personali a scopi statistici e di ricerca scientifica effettuati nell'ambito del Sistema statistico nazionale"[12] (di seguito Codice di deontologia).

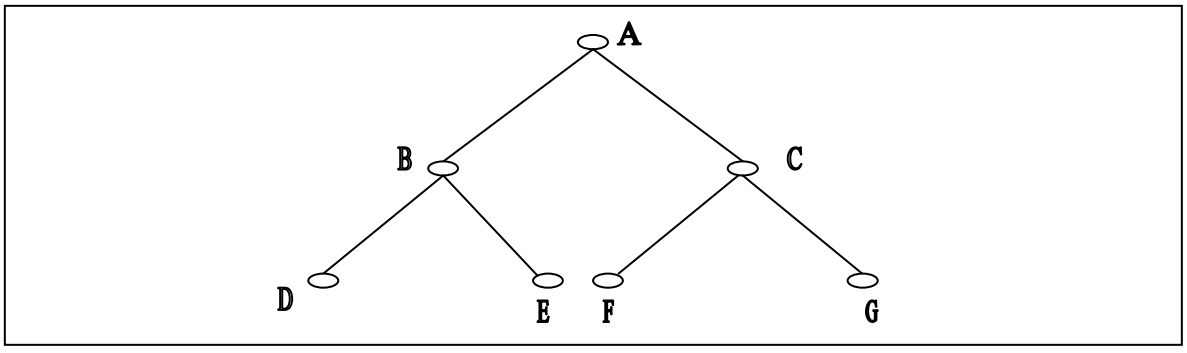
Nella fase di predisposizione dei risultati i vincoli legali si traducono in regole metodologiche e procedure statistiche volte a ridurre (entro limiti prestabiliti) il rischio d'identificazione. I canali di diffusione variano dall'accesso ai microdati tramite laboratori (fattispecie prevista dal Codice di deontologia), al rilascio di collezioni campionarie di dati elementari, alla pubblicazione di tabelle, grafici e indici. Secondo il mezzo utilizzato per la diffusione, esistono metodologie dedicate per limitare il rischio d'identificazione dei rispondenti.

Di seguito viene trattato il tema della tutela della riservatezza nel rilascio dei dati Frame-SBS, anno di riferimento 2013. Il Paragrafo 4.1.1 analizza le classificazioni gerarchiche (annidate e non annidate) utilizzate nella costruzione di dati tabellari. Successivamente viene affrontato il tema della predisposizione dei dati Frame-SBS per l'applicazione delle misure di protezione attraverso l'utilizzo dei software generalizzati disponibili.

#### *4.1.1 Classificazioni gerarchiche annidate e non annidate*

Una classificazione è detta gerarchica quando raccoglie i dati secondo una struttura ad albero del tipo "padre-figlio" come quella indicata in Figura 1:

**Figura 1 - Schema ad albero di una classificazione gerarchica annidata**



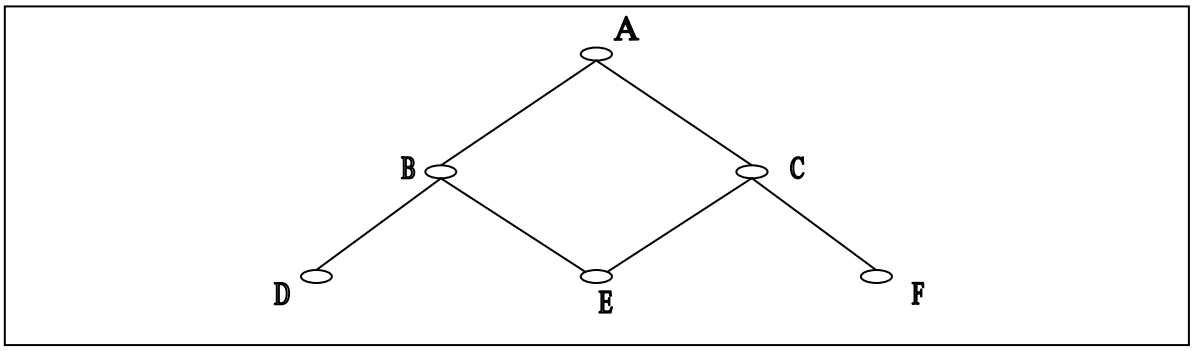

[12] Paragrafo 1.

I livelli gerarchici, corrispondenti a diversi gradi di dettaglio, possono essere rappresentati da vertici (indicati con le lettere da "B" a "G"). La distanza tra il vertice e la radice (A) è detta rango (De Wolf 2007). La Nace è un esempio di classificazione gerarchica che raggruppa le attività economiche secondo i livelli espressi da sezioni, sottosezioni, divisioni, gruppi e classi.

Una classificazione gerarchica è detta annidata (*nested*) quando le modalità sono mutuamente esclusive (Figura 1).

Analogamente, una classificazione gerarchica si dice non annidata (*non nested*) quando le modalità non sono mutuamente esclusive: almeno una modalità concorre alla composizione di due o più modalità di livello gerarchico superiore (Figura 2).

**Figura 2 - Schema ad albero di una classificazione gerarchica non annidata**




Una tabella è gerarchica se almeno una delle variabili classificatrici ha struttura gerarchica.

Riportando gli aggregati rappresentati nelle figure 1 e 2 come modalità di variabili classificatrici si ottengono le due tabelle seguenti:

Tabella 1

| A | |
|---|---|
| -B | |
| --D | |
| --E | |
| -C | |
| --F | |
| --G | |

Tabella 2

| A | |
|---|---|
| -B | |
| --D | |
| --E | |
| -C | |
| --E | |
| --F | |

I vertici rappresentano le modalità (o subtotali), mentre la radice rappresenta il totale marginale. Il numero dei trattini ("-") indica il livello gerarchico delle modalità.

Nella Tabella 2 la modalità "E" è ripetuta e l'additività non è rispettata. L'applicazione delle regole di riservatezza a una tabella così strutturata attraverso i software generalizzati disponibili non risulta possibile a meno di una riorganizzazione preventiva dei dati. Nel caso in esame è possibile scomporre la Tabella 2 in due nuove tabelle collegate[13], esaustive di tutte le modalità, come di seguito raffigurato:

Tabella 3

| A | |
|---|---|
| -B | |
| --D | |
| --E | |
| -C | |

Tabella 4

| *C* | |
|---|---|
| *-E* | |
| *-F* | |

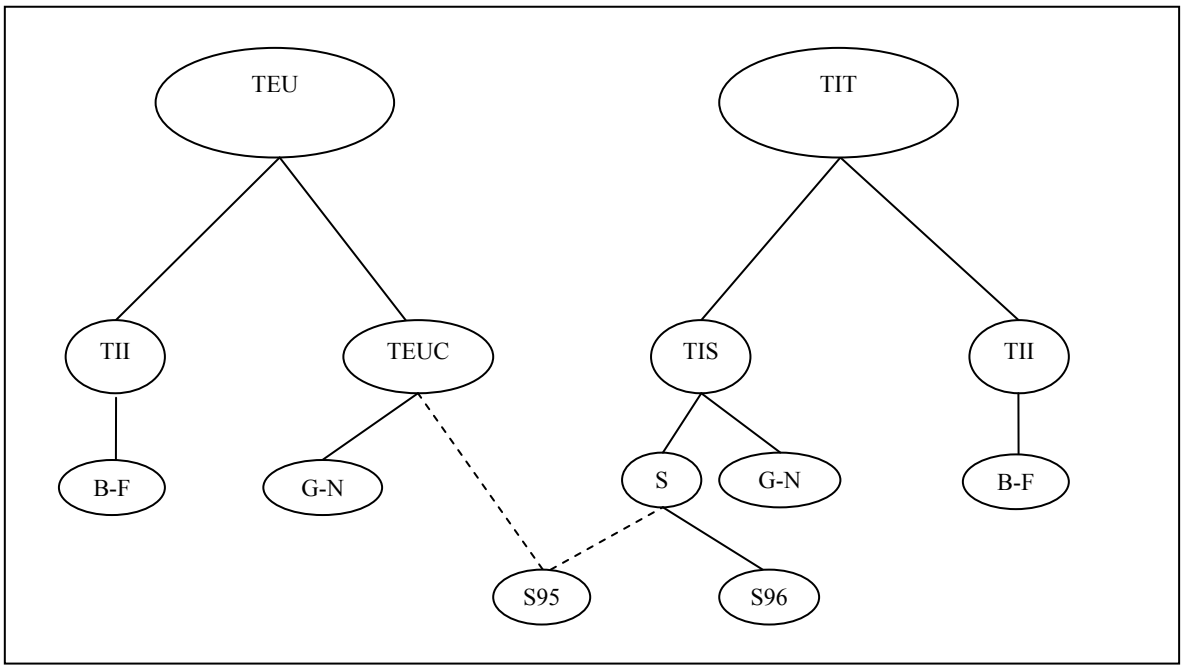
[13] Si definiscono collegate le tabelle che contengono dati relativi alle stesse variabili risposta e che condividono almeno una variabile classificatrice. Il caso più frequente è rappresentato da celle comuni, con particolare riferimento ai valori marginali.

Le tabelle 3 e 4 possono essere protette separatamente imponendo alle celle comuni (C, E) il medesimo status (o flag) di rilasciabilità.

### *4.1.2 Costruzione di classificazioni gerarchiche annidate per i dati Frame-SBS*

In Figura 3 è riportato il diagramma relativo agli aggregati TEU e TIT[14] prodotti e rilasciati dall'Istituto.

**Figura 3 - Schema gerarchico relativo a due aggregati TEU e TIT**

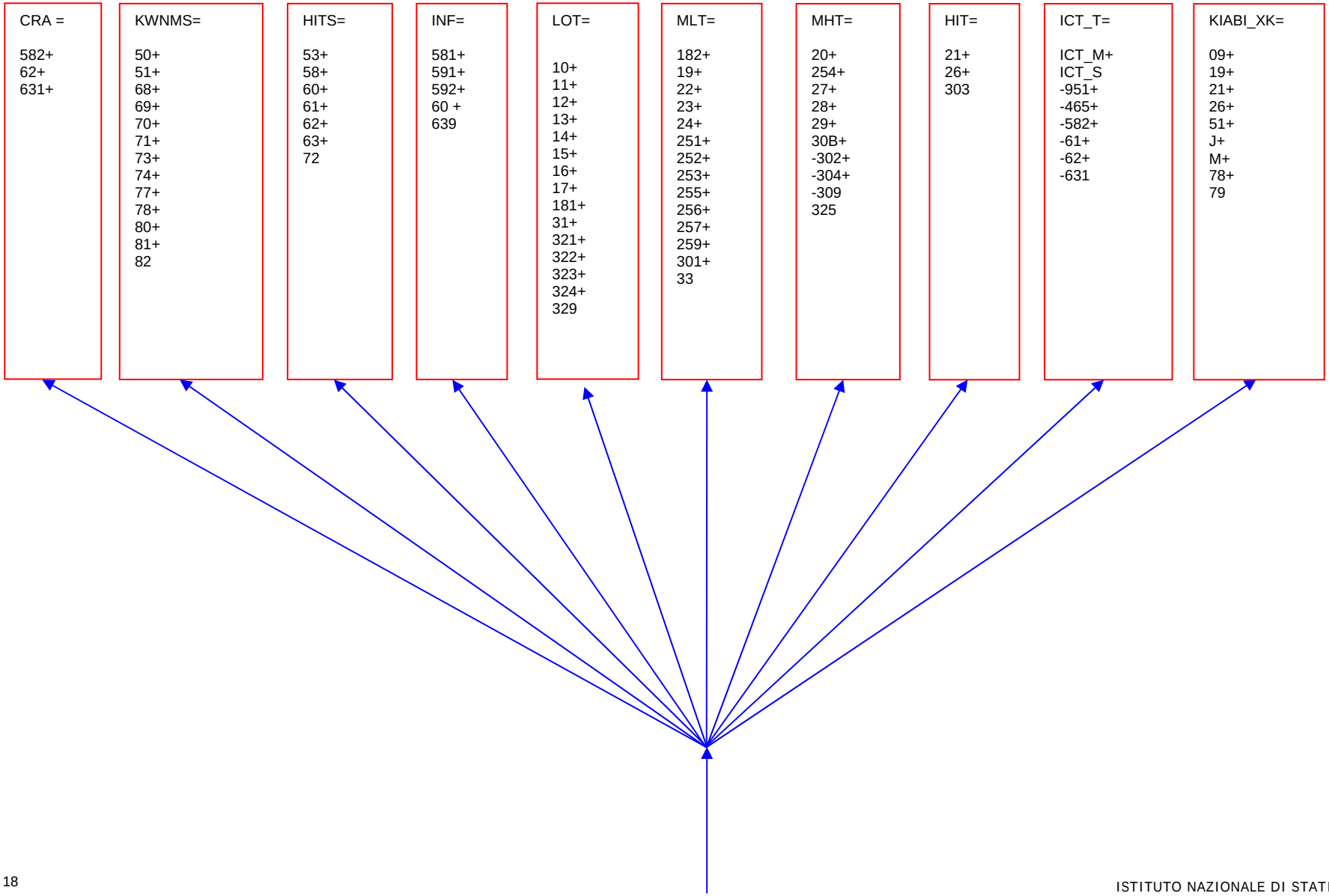


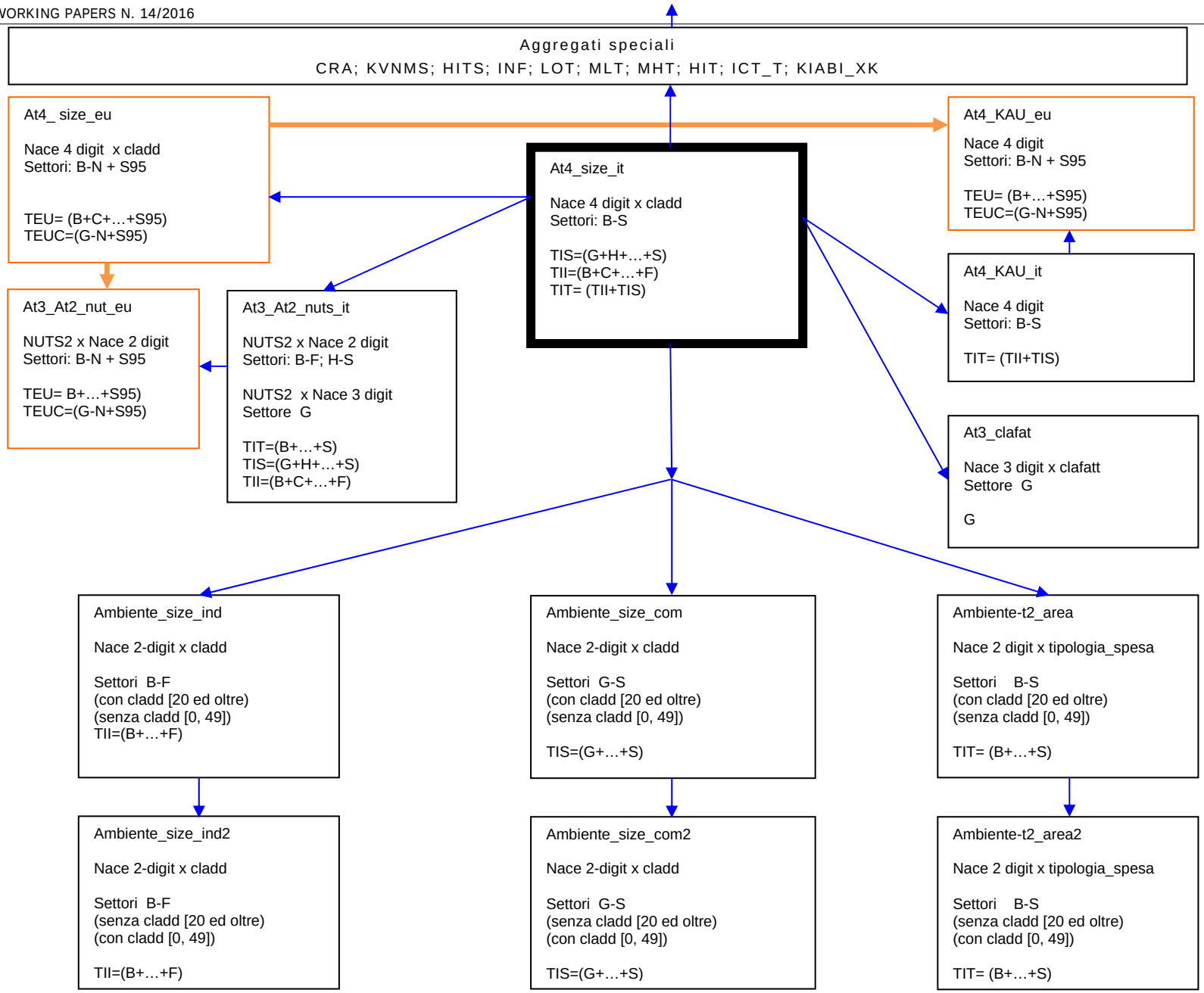

Le modalità che compongono il totale TEUC ricorrono anche tra quelle che formano l'aggregato TIS15: TEUC e TEU sono sottoinsiemi rispettivamente di TIS e TIT. Tuttavia risulta impossibile rappresentare tutti gli aggregati come modalità di una variabile di classificazione tabellare senza duplicazioni (modalità ripetute). Casi analoghi occorrono considerando altri aggregati individuati dal Regolamento europeo, con particolare riferimento agli aggregati speciali e ai dati ambientali.

Il risultato del processo di scomposizione dei dati Frame-SBS, anno di riferimento 2013, è raffigurato in Figura 4. Ogni riquadro rappresenta una tabella gerarchica con le modalità delle variabili di classificazione annidate. In ciascuno di essi, oltre alla denominazione, sono riportate: le variabili di classificazione, i settori di attività economica e i totali marginali Nace. Le frecce indicano i collegamenti determinati da celle in comune.

[14] Dove gli aggregati TIT, TEU, TEUC risultano sommando modalità della NACE come appresso specificato: TIT=TEU+P+Q+R+S96; TEU=B+C+D+E+F+TEUC; TEUC=G+H+I+J+L+M+N+S95.
[15] Dove l'aggregato TIS è definito come somma delle modalità NACE (G+H+I+J+L+M+N+O+P+Q+R+S+T+U).

**Figura 4 - Dati Frame-SBS in tabelle collegate**




Aggregati speciali
CRA; KVNMS; HITS; INF; LOT; MLT; MHT; HIT; ICT_T; KIABI_XK
At4_ size_eu
Nace 4 digit x cladd
Settori: B-N + S95
TEU= (B+C+…+S95)
TEUC=(G-N+S95)
At4_size_it
Nace 4 digit x cladd
Settori: B-S
TIS=(G+H+…+S)
TII=(B+C+…+F)
TIT= (TII+TIS)
At4_KAU_eu
Nace 4 digit
Settori: B-N + S95
TEU= (B+…+S95)
TEUC=(G-N+S95)
At4_KAU_it
Nace 4 digit
Settori: B-S
TIT= (TII+TIS)
At3_At2_nut_eu
NUTS2 x Nace 2 digit
Settori: B-N + S95
TEU= B+…+S95)
TEUC=(G-N+S95)
At3_At2_nuts_it
NUTS2 x Nace 2 digit
Settori: B-F; H-S
NUTS2 x Nace 3 digit
Settore G
TIT=(B+…+S)
TIS=(G+H+…+S)
TII=(B+C+…+F)
At3_clafat
Nace 3 digit x clafatt
Settore G
G
Ambiente_size_ind
Nace 2-digit x cladd
Settori B-F
(con cladd [20 ed oltre)
(senza cladd [0, 49])
TII=(B+…+F)
Ambiente_size_com
Nace 2-digit x cladd
Settori G-S
(con cladd [20 ed oltre)
(senza cladd [0, 49])
TIS=(G+…+S)
Ambiente-t2_area
Nace 2 digit x tipologia_spesa
Settori B-S
(con cladd [20 ed oltre)
(senza cladd [0, 49])
TIT= (B+…+S)
Ambiente_size_ind2
Nace 2-digit x cladd
Settori B-F
(senza cladd [20 ed oltre)
(con cladd [0, 49])
TII=(B+…+F)
Ambiente_size_com2
Nace 2-digit x cladd
Settori G-S
(senza cladd [20 ed oltre)
(con cladd [0, 49])
TIS=(G+…+S)
Ambiente-t2_area2
Nace 2 digit x tipologia_spesa
Settori B-S
(senza cladd [20 ed oltre)
(con cladd [0, 49])
TIT= (B+…+S)

Le tabelle in Figura 4 rappresentano un sovrainsieme dei dati da rilasciare e contengono sia gli aggregati previsti dalle serie del regolamento europeo, sia i domini utilizzati nelle diffusioni nazionali. I dettagli utilizzati nella costruzione delle tabelle sono quelli descritti nel Capitolo 2 del documento.

Le modalità che formano gli aggregati speciali permettono di tenere conto dei legami che intercorrono con la tabella *At4_size_it*. Per questa ragione nell'aggregato MHT è stata costruita *ad hoc* la modalità 30B (definita come somma delle NACE 302+304+309) riportata anche nella tabella *At4_size_it*.

Le classi di addetti previste dal Regolamento per i dati ambientali (Ambiente_size_ind e Ambiente_size_com) contengono la modalità [0-49]. Essa è non annidata rispetto alla classificazione ([0,19], [0,9], [0,1], [2,9], [10,19], [20 ed oltre), [20,49], [50,249], [250 ed oltre) ) utilizzata sia per le tabelle At4_size_it, At4_ size_eu, sia per le tabelle ambientali. È questa la ragione per cui i dati ambientali vengono ripartiti in sei tabelle (anziché in tre).

Opportune procedure descritte nel Paragrafo 4.2 permettono di proteggere l'insieme di tabelle rappresentato in Figura 4, garantendo la coerenza nei risultati finali.

## 4.2 Aspetti metodologici e tecnici nell'applicazione delle regole di riservatezza

Le tabelle rappresentano la forma più comune di diffusione dei risultati statistici. Le celle (domini) sono definite dagli incroci delle modalità delle variabili di classificazione. L'obiettivo della violazione è associare i contributi delle variabili riservate alle unità che li realizzano (ad esempio valore aggiunto e impresa). L'ipotesi riguardante le informazioni di cui dispone il soggetto che tenta la violazione è esplicitata tramite il concetto di scenario di violazione.

Per dati tabellari le variabili di classificazione possono essere considerate come chiavi identificative, mentre le variabili risposta come informazioni riservate (Capitolo 1).

Il Codice di deontologia definisce dati aggregati le "[…] combinazioni di modalità alle quali è associata una frequenza non inferiore a una soglia prestabilita, ovvero un'intensità data dalla sintesi dei valori assunti da un numero di unità statistiche pari alla suddetta soglia […]". In accordo con tale definizione, la regola di rischio adottata dall'Istat è quella della minima frequenza, o regola della soglia, secondo la quale una cella è a rischio di violazione se riferita a un numero di contributori minore di un prefissato parametro (k). Nel Codice di deontologia viene anche individuato il limite inferiore di tale parametro, stabilendo che esso non possa essere minore di tre (Appendice A).

L'adozione di misure di protezione implica una diminuzione del contenuto informativo dei dati originali. L'Istat, come la maggioranza degli istituti di statistica europei, adotta misure di protezione che riducono l'informazione rilasciata senza modificare i valori osservati: i contributi delle celle sensibili sono soppressi (o oscurati) e sostituiti con codici di rilasciabilità.

Il processo di protezione dei dati aggregati non si esaurisce con l'introduzione di valori mancanti nelle celle sensibili. Ulteriori soppressioni (soppressioni secondarie) sono necessarie per garantire che le celle a rischio (oscurate) non siano calcolate a partire dai dati rilasciati (ad esempio per differenza con i valori marginali).

L'individuazione delle soppressioni secondarie avviene minimizzando una funzione di costo. Il processo di protezione dei dati aggregati si traduce in un problema matematico di ottimizzazione risolto facendo ricorso agli algoritmi implementati nel software generalizzato τ-Argus[16] (disponibile insieme al relativo manuale alla pagina web http://neon.vb.cbs.nl/casc/tau.htm ). La complessità di calcolo aumenta in presenza di tabelle collegate. In questi casi, infatti, il tracciato delle soppressioni individuato per una tabella diventa input nella protezione delle tabelle ad essa collegate riducendo i gradi di libertà nell'individuazione della "migliore soluzione". Per tabelle collegate la distribuzione delle soppressioni dipende anche dall'ordine di protezione. Nell'ipotesi in cui tutte le modali-

[16] Gli algoritmi maggiormente utilizzati nella protezione di tabelle gerarchiche collegate (come quelle rappresentate in Figura4, Paragrafo 4.1), sono l'HiTas (o modular) e l'Optimal. Quest'ultimo può avere tempi di elaborazioni molto lunghi per tabelle con elevata complessità, in relazione al numero di livelli gerarchici, celle a rischio e al livello di protezione impostato.

tà delle variabili di classificazione risultino come aggregazioni dei dettagli più fini, la regola generale è quella di procedere dal "particolare al generale" partendo dalle tabelle più dettagliate (nella variabile di classificazione comune) e continuando fino ad arrivare a quelle meno dettagliate. Lo strumento necessario per questa procedura è il cosiddetto *history file* (o apriori file) che permette di tenere memoria delle soppressioni effettuate. Esso è ottenuto facendo ricorso ad una funzionalità di τ-Argus che da ogni tabella protetta permette di estrapolare il file da utilizzare in input nella protezione delle tabelle successive.

### *4.2.1 Applicazione delle regole di riservatezza ai dati Frame-SBS col software τ-Argus*

Nel caso dei dati Frame-SBS, lo scenario d'intrusione ipotizzato presuppone che l'*intruder* sia in grado di collocare le unità rilevate all'interno dei domini definiti dagli incroci di tutte le variabili di classificazione. Sulla base di questa assunzione la valutazione del rischio di violazione è effettuata per ogni cella da rilasciare.

La regola di rischio adottata è quella della minima frequenza (k); il parametro k è posto pari a tre e quindi sono definite a rischio le celle con un numero di contributori strettamente minore di tre.

L'algoritmo utilizzato per l'individuazione delle soppressioni secondarie è il *modularo* o *HiTaS* (De Wolf 2002). La funzione di costo è l'ammontare di valore aggiunto: le soppressioni sono determinate minimizzando il valore complessivo dei contributi oscurati.

Per la protezione dei dati Frame-SBS si utilizza il software τ-Argus che permette, attraverso una maschera, di selezionare la regola di rischio, parametrizzare il livello di protezione e selezionare l'algoritmo per l'individuazione delle soppressioni secondarie.

Con riferimento alle tabelle di Figura 4 del Paragrafo 4.1, la sequenza di protezione adottata è la seguente:

1. *At4_size_it*
2. *At4_size_eu*
3. *At3_clafat*
4. *At3_At2_nuts*
5. *At3_At2_nut_eu*
6. *At4_KAU_eu*
7. *At4_KAU_it*
8. *Aggregati speciali*
9. *Ambiente_size_ind*
10. *Ambiente_size_com*
11. *Ambiente_at2_area*
12. *Ambiente_size_ind*
13. *Ambiente_size_com*
14. *Ambiente_at2_area*

I domini più fini, Nace a quattro cifre per classi di addetti, sono i primi a essere protetti. L'*history file* risultante è utilizzato per vincolare i *flag* di rilasciabilità nelle celle comuni delle tabelle da proteggere. Le tabelle *At3_At2_nut_eu* e *At4_KAU_eu*, recepiscono, oltre agli status individuati rispettivamente nelle tabelle *At3_At2_nuts* e *At4_KAU_it*, anche quelli derivanti dalla protezione dei dati in *At4_ size_eu* che contiene i domini comuni TEU e TEUC.

La protezione dei dati Frame-SBS è ottenuta selezionando l'opzione *singleton* di τ-Argus, relativa a celle con un solo contributore. Essa garantisce che due celle soppresse che si proteggono a vicenda non siano entrambe riferite a singoli contributori (escludendo così la possibilità che un "auto-riconoscimento" comporti una diretta violazione della riservatezza).

Un unico tracciato delle soppressioni (primarie e secondarie) è adottato per tutte le variabili risposta. Il programma di audit, disponibile tra le funzionalità di τ-Argus, permette di valutare a posteriori i livelli di protezione conseguiti per le variabili che non rientrano nella definizione della funzione di costo.

Per motivi di trasparenza in I.Stat l'introduzione di valori mancanti conseguente a motivi di riservatezza è comunicata all'utenza tramite la seguente locuzione: "dato oscurato per la tutela del

segreto statistico". La cella oscurata (sia in caso di soppressione primaria sia in caso di soppressione secondaria) presenta al suo interno il flag C. Per l'invio dei dati ad Eurostat i codici di rilasciabilità prevedono la distinzione tra cella a rischio (soppressione primaria) e cella soppressa (soppressione secondaria).

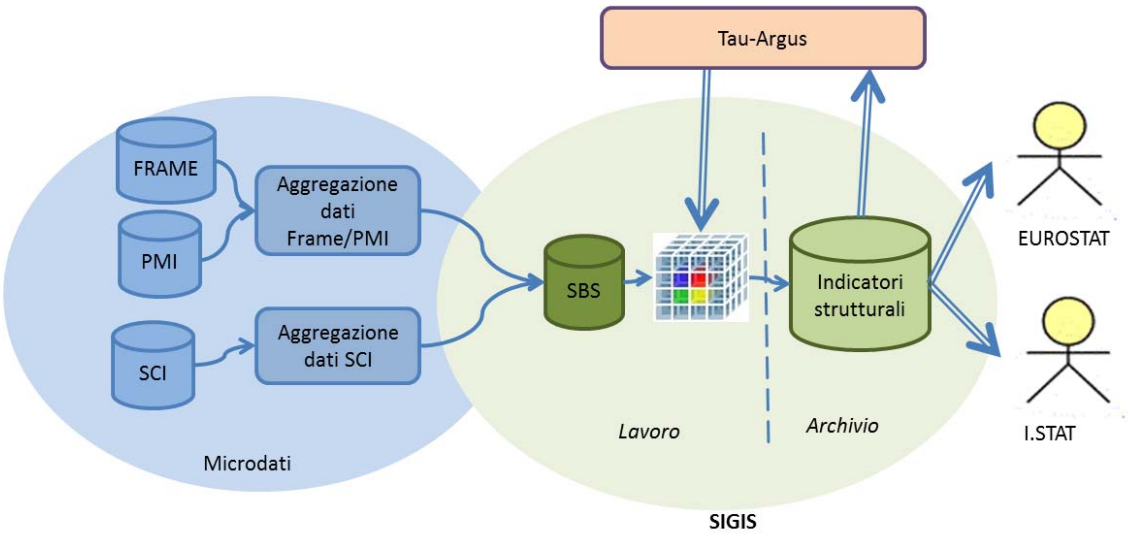
Il risultato finale del processo di protezione dei dati tramite τ-Argus è un insieme di tabelle, esaustivo dei domini da pubblicare, le cui celle contengono flag di rilasciabilità.

## 5. Aspetti informatici del trattamento dei dati SBS

L'informatizzazione della gestione dell'output di SBS è realizzata attraverso Sigis, Sistema informativo per la Gestione degli Indicatori Strutturali, che fornisce le funzioni di supporto all'attività di predisposizione degli indicatori strutturali e provvede alla loro memorizzazione su due diverse aree: *lavoro* e *archivio*.

Il processo SBS in Sigis prevede cinque distinte fasi di lavorazione, come illustrato in figura 1. La prima fase del processo è costituita dall'*aggregazione dei microdati* d'indagine dei processi produttivi Frame/PMI e SCI che compongono SBS, tale fase è implementata nell'ambiente di produzione dei dati. Nella seconda fase è effettuato il *calcolo dell'aggregato SBS*, tale fase è realizzata nell'area di *lavoro* di Sigis, comune per tutte le indagini strutturali; gli aggregati di massimo dettaglio di Frame/PMI e SCI si sommano dando luogo alla base dei dati SBS. I dati risultati di questa fase sono accessibili ai responsabili di produzione in sola lettura per elaborazioni spot e verifiche. Nella terza fase è effettuata la *predisposizione degli indicatori di diffusione SBS*, per aggregare le variabili secondo quanto previsto dai regolamenti per i diversi piani di diffusione, le operazioni realizzate in questa fase sono di *roll-up* sulle dimensioni di analisi per le diverse variabili, ovvero di aggregazione secondo la gerarchia di ciascuna dimensione partendo dal livello più fine. Nella quarta fase è realizzata l'*integrazione per la gestione della confidenzialità*, in questa fase i dati sono predisposti per le procedure gestite dal gruppo metodologico per la gestione della confidenzialità; la procedura di confidenzialità riceve da Sigis i file di dati e restituisce gli stessi con l'opportuna indicazione di confidenzialità. Nella quinta ed ultima fase è effettuata l'*estrazione degli indicatori strutturali*, in questa fase sono predisposte le informazioni per essere utilizzate per la diffusione verso Eurostat ed I.Stat.

**Figura 5 – Processo SBS in Sigis**



## 5.1 L'architettura di Sigis

Il sistema informativo gestionale indicatori strutturali, Sigis, è costituito da due ambienti:

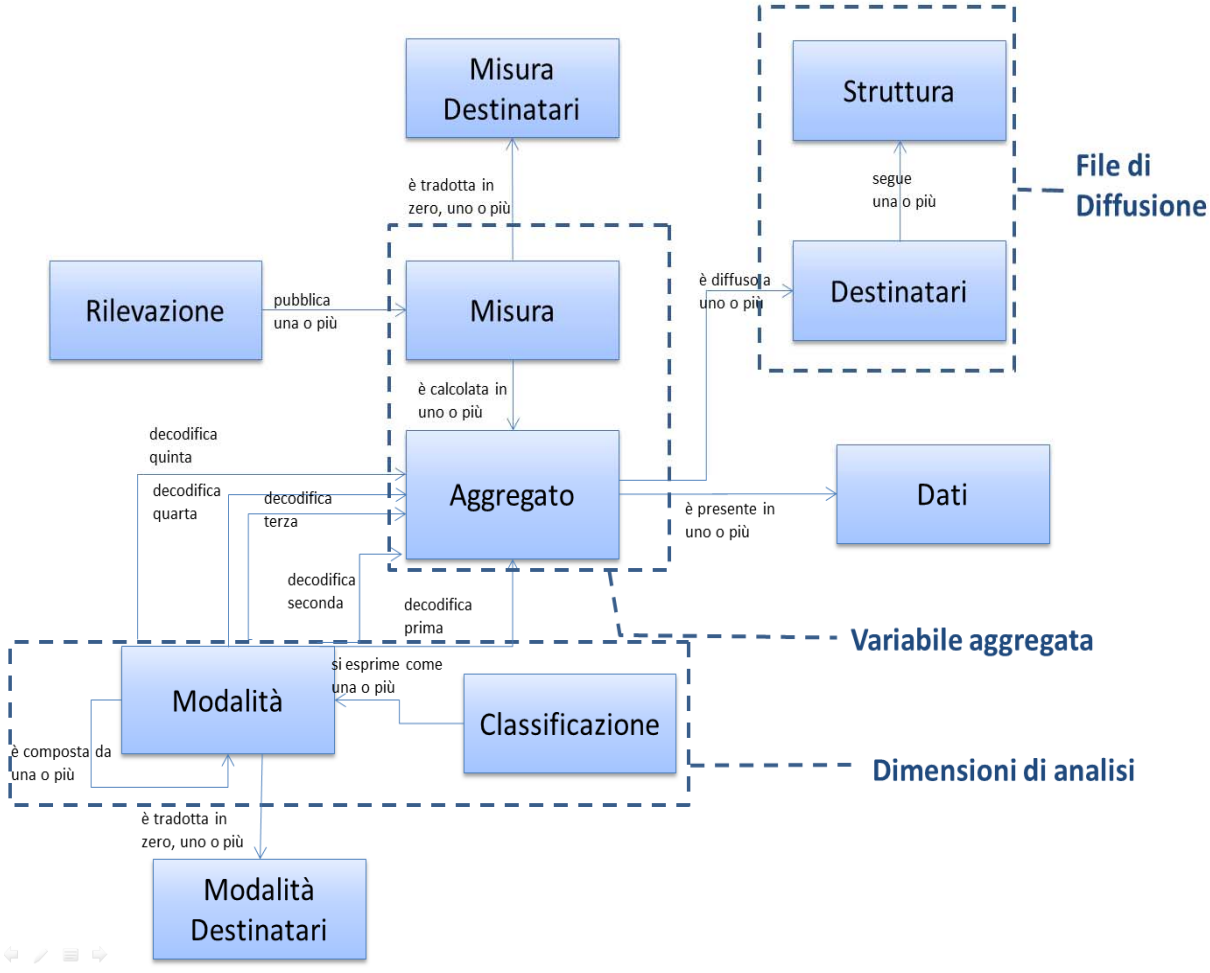
a) un area di lavoro dove sono realizzate le funzioni di caricamento dati, di roll-up per le dimensioni di analisi per ogni variabile e l'integrazione con il sistema di confidenzialità utilizzato per SBS, in particolare il sistema di confidenzialità risulta un utilizzatore del sistema quando riceve i file di dati su cui applicare le regole della riservatezza ed anche un fornitore di informazioni quando rilascia i file nei quali è assegnato lo stato di confidenzialità;

b) un area di archivio dove i dati aggregati sono memorizzati con l'indicazione di confidenzialità, a cui accedere per effettuare le estrazioni dei dati per i diversi sistemi di diffusione (I.Stat, Eurostat).

Il caricamento dei dati è implementato tramite funzionalità scritte in linguaggio Pl-Sql richiamabili tramite prodotti che interfacciano il database Oracle; le procedure di caricamento sono personalizzate per ogni rilevazione presente in Sigis.

Il modello dei dati sul quale si basa Sigis è il seguente:

**Figura 6 – Modello dei dati Sigis**




Una rilevazione diffonde una o più variabili aggregate secondo diverse dimensioni di analisi che, nel contesto in esame, arrivano ad un massimo di cinque incroci. Ciascuna voce di una dimensione può essere definita come unione di altre voci, memorizzando tali informazioni è possibile calcolarne in automatico il valore. L'aggregato può essere predisposto per uno o più file di diffusione, per ogni file di diffusione è memorizzata la tipologia della struttura, il carattere separatore

dei campi e, per ogni colonna, le indicazioni necessarie per l'individuazione del contenuto. In ogni file di output è possibile inserire diversi aggregati ed ogni aggregato può essere contenuto in diversi file, per tale motivo esiste una tabella ad hoc per memorizzare l'elenco degli aggregati per ogni file di output. Ogni variabile ed ogni modalità delle dimensioni può essere contenuta in diversi file di diffusione secondo una decodifica diversa rispetto alla codifica utilizzata per la memorizzazione nel sistema, tali informazioni sono memorizzate rispettivamente in Misura Destinatari e Modalità Destinatari. L'aggregato è caricato nei dati ogni volta che si rilasciano le informazioni per un nuovo periodo di riferimento oppure si procede ad una revisione di un periodo già diffuso; nei dati è possibile annotare ogni informazione che si ritiene opportuna, in particolare se è confidenziale o meno.

La configurazione del sistema, dove per configurazione si intende il popolamento delle tabelle dei metadati relative agli aggregati, ai file di diffusione e alle tabelle di relazioni tra essi è possibile effettuarla appena si conoscono gli aggregati per l'anno in esame anche se i dati non sono stati ancora prodotti.

La procedura di produzione file di output è generalizzata e si basa sulle informazioni presenti nel sistema per la specifica rilevazione, come parametri di input per poter produrre i file richiesti, la procedura necessita delle seguenti informazioni:

a) codice della rilevazione;
b) periodo di riferimento dei dati;
c) edizione da pubblicare;
d) tipologia di file da produrre.

La procedura una volta individuato il file da produrre ed il relativo formato, legge gli aggregati che devono essere presenti nel file quindi di essi prende il dato rispetto al periodo di riferimento e la data di edizione e produce il file da diffondere.

I file di output sono creati in una cartella del server dati e resi accessibili ai referenti di indagine su cartelle condivise in rete con Windows soltanto alle persone autorizzate.

## 5.2 L'aggregazione dei microdati del Frame e delle rilevazioni sulle piccole e medie imprese e sull'esercizio di arti e professioni (PMI) e sul sistema dei conti delle imprese (SCI)

Le aggregazioni ottenute dall'elaborazione dei microdati del Frame e dell'indagine PMI, ovvero delle imprese con meno di 100 addetti, sono prodotte dall'unità di produzione statistica e sono riportati nell'area di lavoro di Sigis nelle seguenti tabelle:

a) *Frame_imprese_at4size*, contiene le aggregazioni per Ateco (Nace) a 4 cifre delle sezioni da B a S e classe di addetti, in particolare per le sezioni da B ad F le classe di addetti fornite sono 0-9, 10-19, 20-49, 50-249 per le sezioni da G a S invece le classi di addetti sono 0-1, 2-9, 10-19, 20-49, 50-249.
b) *Frame_imprese_at4size_core*, a partire dall'edizione 2012, con l' impiego del Frame, è stato possibile introdurre un livello di dettaglio più fine anche per le attività economiche dell'Industria e delle Costruzioni (Sez. B-F), ampliando il dominio di stima dalla classe di addetti 0-9 alle classi di addetti 0-1, 2-9; pertanto è stata aggiunta tale tabella che contiene le aggregazioni per le sezioni di Ateco da B a S con le classe di addetti 0-1, 2-9, 10-19, 20-49, 50-249.

Tale dettaglio è stato applicato soltanto alle variabili principali ovvero a:

| Codice | Descrizione |
|---|---|
| 11110 | numero delle imprese |
| 11500 | altri ricavi e proventi |
| 12100 | acquisti di materie prime, sussidiarie e di consumo |
| 12110 | Fatturato |
| 12150 | valore aggiunto al costo dei fattori |
| 12170 | margine operativo lordo |
| 12200 | acquisto di servizi |
| 12300 | acquisto per godimento beni di terzi |

| | |
|---|---|
| 12900 | oneri diversi di gestione |
| 13110 | acquisto di beni e servizi |
| 13310 | costi del personale |
| 13320 | salari e stipendi |
| 16110 | numero di persone occupate |
| 16130 | numero di dipendenti |

c) *Frame_imprese_at3clfat*, contiene le aggregazioni per Ateco a 3 cifre della sola sezione G e classe di fatturato.
d) *Frame_ulregioni_at23reg*, contiene le aggregazioni per regione e Ateco a 2 cifre delle sezioni da B a S ed Ateco a 3 cifre per la sola sezione G.
e) *Frame_kau_at4*, contiene le aggregazioni delle unità funzionali per Ateco a 4 cifre delle sezioni da B a S.
f) *Frame_ambientali_at3size*, contiene le aggregazioni per Ateco a 3 cifre delle sezioni da B a S e classe di addetti e area di protezione ambientale.
g) *Frame_business8_at234size*, contiene le aggregazioni per Ateco delle sezioni J,M,N e tipo di residenza del cliente e tipo di prodotto. In tale tabella alcune attività sono presenti annualmente, in particolare 582, 62, 631, 731 e 78, mentre 691, 692 e 702 negli anni pari e 7111, 7112, 712, e 732 negli anni dispari.

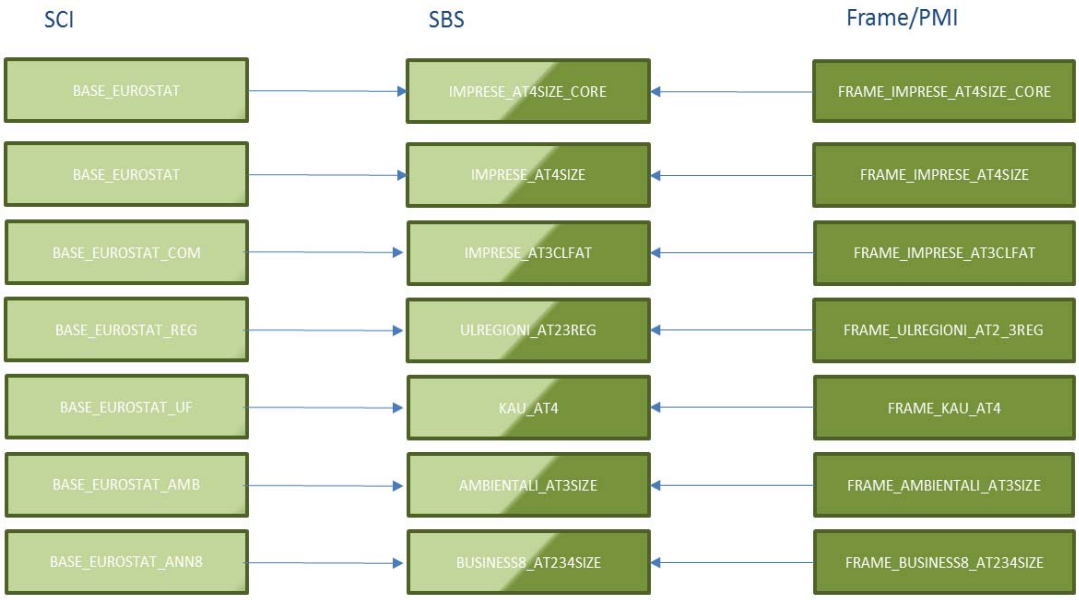
Le aggregazioni dei dati di SCI, ovvero delle imprese con 100 addetti ed oltre, sono invece calcolate nell'area di produzione dell'indagine e memorizzate nelle seguenti tabelle:

a) *Base_eurostat*, contiene le aggregazioni per Ateco a 4 cifre delle sezioni da B a S e classe di addetti.
b) *Base_eurostat_com*, contiene le aggregazioni per Ateco a 3 cifre della sola sezione G e classe di fatturato.
c) *Base_eurostat_reg*, contiene le aggregazioni per regione e Ateco a 2 cifre delle sezioni da B a S ed Ateco a 3 cifre per la sola sezione G.
d) *Base_eurostat_uf*, contiene le aggregazioni delle unità funzionali per Ateco a 4 cifre delle sezioni da B a S.
e) *Base_eurostat_amb*, contiene le aggregazioni per Ateco a 3 cifre delle sezioni da B a S e classe di addetti e area di protezione ambientale.
f) *Base_eurostat_ann8*, contiene le aggregazioni per Ateco delle sezioni J, M, N e tipo di residenza del cliente e tipo di prodotto. In tale tabella alcune attività sono presenti annualmente, in particolare 582, 62, 631, 731 e 78, mentre 691, 692 e 702 negli anni pari e 7111, 7112, 712, e 732 negli anni dispari.

## 5.3 Il calcolo degli aggregati SBS

Gli aggregati calcolati nell'ambito del sistema di produzione SCI sono trasferiti nell'area di lavoro di Sigis ed ivi uniti con gli aggregati di Frame/PMI tramite operazioni di somma sulle medesime dimensioni generando le tabelle secondo il seguente schema:

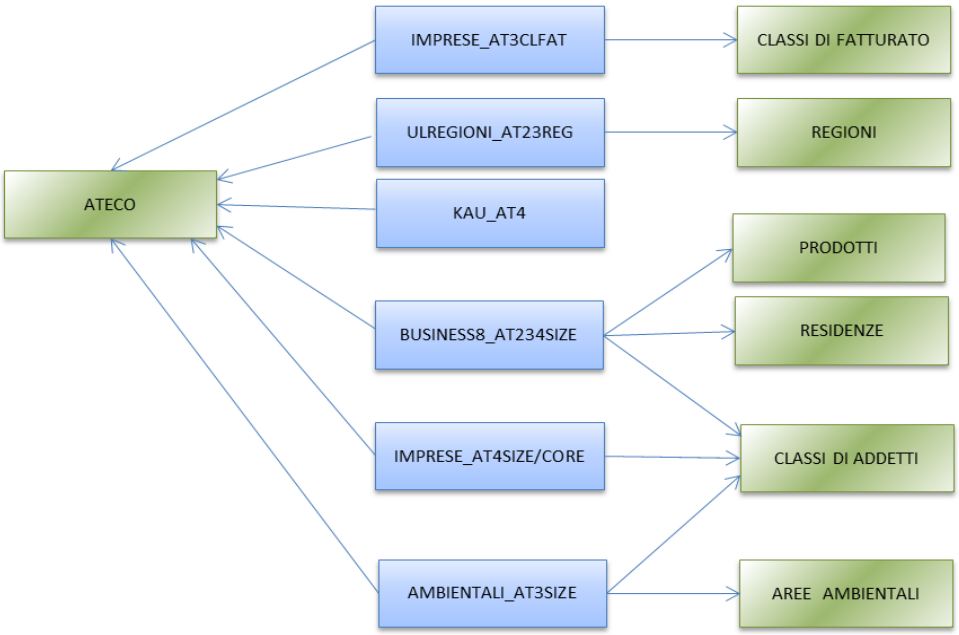
**Figura 7 – Flusso dei dati nel calcolo degli aggregati SBS**




Le tabelle cosi formate rimangono accessibili in sola lettura ai referenti di indagine per permettere verifiche ed elaborazioni particolari dovute a richieste di forniture di dati estemporanee e sono la base delle operazioni di roll-up successive.

Lo schema dei dati utilizzato è di tipo a stella ed è descritto di seguito.

**Figura 8 – Modello dimensionale dei dati SBS**




Le tabelle di colore verde rappresentano le dimensioni di analisi mentre le tabelle di colore azzurro rappresentano quelle dei dati:

a) nella tabella *Imprese_at3clfat* sono contenute le aggregazioni per Ateco e Classe di fatturato;
b) nella tabella *Ulregioni_at23reg* sono contenute le aggregazioni per Ateco e Regioni,

c) nella tabella *Kau_at4* sono contenute le variabili relative alle unità funzionali aggregate per Ateco;
d) nella tabella *Business_at234size* sono contenute le aggregazioni per Ateco, Prodotti, Residenze e Classi di addetti, dove per quest'ultima classe è presente soltanto la voce 20 e più addetti;
e) nelle tabelle *Imprese_at4size e imprese_at4size_core* sono contenute le aggregazioni per Ateco e Classe di addetti;
f) nella tabella *Ambientali_at3size* sono contenute le aggregazioni per Ateco, Aree ambientali e Classi di addetti.

## 5.4 Predisposizione degli indicatori di diffusione SBS

I dati di SBS sono letti dai file distinti a seconda dei domini di stima, cosi come descritti in precedenza, e sono uniti in un unico contenitore prima di essere inseriti in Sigis secondo la codifica necessaria. In Sigis sono state censite le variabili così come sono definite le variabili nei diversi file di origine, ma nel codice della misura si è conservata memoria del file di origine della variabile stessa secondo lo schema seguente:

1. IM4, dati aggregati per Ateco delle sezioni B-S a 4 cifre e classe di addetti provenienti da *Imprese_at4size*;
2. REG, dati aggregati per Ateco delle sezioni B-S a 2 cifre, 3 cifre per la sezione G, e regione amministrativa provenienti da *Ulregioni_at23reg*;
3. AMB, dati aggregati per Ateco a 3 cifre delle sezioni B-S, classe di addetti ed area di ambiente protetto provenienti da *Ambientali_at3size*;
4. BSP, dati aggregati per Ateco a 2/3/4 cifre delle sezioni J,M,N, per classe di addetti e per tipo di prodotto o servizio offerto provenienti da *Business8_at234size*;
5. BSR, dati aggregati per Ateco a 2/3/4 cifre delle sezioni J,M,N, per classe di addetti e per residenza del cliente provenienti da *Business8_at234size*;
6. IM3, dati aggregati per Ateco a 3 cifre e classi di fatturato per il solo settore del commercio provenienti da *Imprese_at3clfat*;
7. KAU, dati aggregati per Ateco a 4 cifre delle sezioni B-S delle unità funzionali provenienti da KAU_AT4;
8. IMC, dati aggregati per Ateco a 4 cifre e classe di addetti più dettagliata per le variabili principali di stima provenienti da *Imprese_at4size_core*.

Le dimensioni di analisi considerate per creare l'aggregato sono le seguenti:

a) Ateco 2007,
b) classe di fatturato,
c) prodotti,
d) residenza,
e) classe di addetti,
f) territorio,
g) area ambientale.

L'aggregato, cosi come sono i dati, non prevede mai l'incrocio di tutte le dimensioni, per i record del flusso IM3 la dimensione Ateco si incrocia con la dimensione della classe di fatturato, per i dati relativi ai flussi IM4, IMC e KAU la dimensione Ateco si incrocia con la dimensione della classe degli addetti, per i dati del flusso REG la dimensione Ateco si incrocia con la dimensione del territorio, per i dati del flusso AMB è presente l'incrocio Ateco, classe di addetti ed area ambientale ed per i dati del flusso BSP è presente l'incrocio Ateco, prodotti e classe di addetti, infine i record relativi al flusso BSR prevedono l'incrocio Ateco, residenza e classe di addetti.

La procedura di caricamento in Sigis, dopo aver inserito il dato effettua il calcolo di alcuni livelli di aggregazione per l'Ateco, per il territorio e per la classe di addetti, aggregazioni impostate come modalità padre e figlio nella relativa tabella, si riporta di seguito il dettaglio delle aggregazioni effettuate.

Per gli addetti sono effettuate le aggregazioni seguenti:

- la classe 0_9 a partire da 0-1, 2-9 ;
- la classe 1_19 a partire da 0_9, 10_19;
- la classe 20_ e più a partire da 20_49, 50_249, 250_;
- la classe 0_49 a partire da 0_9, 10_19, 20_49;
- il totale degli addetti a partire dalle classi 0_9, 10_19, 20_49, 50_249, 250_.

Per l'Ateco sono calcolati tutti i valori dei gruppi a partire dai valori delle quattro cifre, poi sono calcolate le divisioni a partire dai gruppi, compresi alcuni raggruppamenti particolari (ad esempio 30B, 25A, 32B) , le sezioni a partire dalle divisioni, il totale a partire dalle sezioni, infine sono calcolati degli aggregati speciali[17].

A livello territoriale si calcolano i valori regionali a partire dai valori provinciali, quindi le ripartizioni sia a cinque che a due ed infine il totale Italia a partire dalle ripartizioni a cinque.

Infine per l'inserimento in Sigis degli indicatori di distribuzione diffusi al momento su I.Stat è stata realizzata un'ulteriore procedura di caricamento; tali dati sono presenti nell'area di lavoro di Sigis ovvero in una tabella Oracle in formato verticale similare alla struttura di Sigis, la procedura di caricamento riporta i dati nelle tabelle proprie secondo la relativa codifica di misure ed aggregato, su tali dati non è effettuata alcuna aggregazione.

## 5.5 Integrazione per la gestione della confidenzialità

Il sistema di confidenzialità τ-Argus è stato considerato un destinatario particolare in quanto riceve da SIGIS dei file di dati ma fornisce anche al Sigis stesso lo stesso numero di file con l'indicazione relativa alla confidenzialità.

È stato definito per τ-Argus la tipologia di destinatario T ed i file censiti per il destinatario di tipologia T sono sono caratterizzati da un tracciato delimitato dal carattere ",", ovvero sono file di tipo "csv" e nella tabella opportuna si sono censiti tutti gli incroci che devono essere presenti per ogni file, cosi come richiesto da τ-Argus.

Per l'estrazione dei dati è utilizzata la funzionalità di produzione file di output sopra descritta con gli opportuni parametri di input e lo scambio dei dati al momento verso τ-Argus è effettuato attraverso file di tipo SAS, pertanto dopo la creazione dei file csv, è eseguita una procedura scritta in sas che converte i file secondo il formato richiesto, tale procedura accede alle tabelle di Sigis per dedurre il numero, il nome dei file prodotti e la relativa struttura.

Cosi anche i file di ritorno da τ-Argus che riportano la protezione da assegnare sono prodotti in formato sas, sono prodotti tanti file quanti file sono stati inviati, pertanto il programma SAS, che legge tali file, accede alle tabelle di Sigis per sapere il nome dei file da leggere e converte i file in file di formato csv. In questo formato i file possono essere letti da una procedura scritta ad hoc in linguaggio Pl-Sql che riporta le informazioni di tutti i file in un'unica tabella e in base a quanto riportato come confidenziale aggiorna nei dati la notazione relativa alla confidenzialità, in particolare riporta la confidenzialità assegnata per le classi di addetti e per il totale dell'area anche per tutti gli aggregati definiti per classe di addetti e per specifica area, riporta, inoltre, la confidenzialità assegnata per la sola classe di fatturato G anche per tutti gli aggregati definiti come totale del fatturato ed infine riporta la confidenzialità assegnata ad alcuni aggregati speciali (TEUC, TIT) che coprono l'intero universo agli aggregati definiti come totale.

[17] CRA = 582+631+62; HIT = 303+26+21; HITS = 61+62+63+60+72+53+58; ICT_M = 261+268+262+264+263;
ICT_S = 582+951+631+62+61+465; ICT_T = 262+264+261+268+465+62+951+582+61+631+263; INF=639+60+592+591+581;
KIABI=19+21+79+78+75+74+73+72+71+70+69+63+62+61+26+09+51+58+59+60;
KWNMS=68+73+80+77+71+81+82+51+74+50+69+78+70; LOT=11+12+13+14+15+16+17+181+31+321+322+323+10+324+329;
MHT = 325+309+304+302+254+29+28+27+20; MLT = 256+257+259+301+255+253+252+251+24+23+22+33+182+19;
TEU=L+J+N+I+95+H+B+G+D+F+C+E+M; TEUA = D+36+C+B; TEUC = 95+J+M+L+H+N+G+I; TII = F+E+D+C+B;
TIS = S+L+R+I+Q+N+G+P+H+J+M; TIT = B+E+D+L+C+F+G+H+S+Q+P+N+I+M+J+R.

### 5.6 Estrazione degli indicatori strutturali

I destinatari per la diffusione dei dati di SBS sono di due tipologie:

1. E per Eurostat;
2. I per I.Stat.

Per l'invio ad Eurostat sono stati censiti 40 file, avente sempre tracciato non fisso ma stavolta delimitato dal carattere ";" e nella tabella opportuna si sono indicati tutti gli aggregati cosi come richiesto da Eurostat per ogni file

Per la diffusione di I.Stat sono stati censiti due file, aventi sempre tracciato non fisso ma delimitato dal carattere "|", cosi come richiesto da I.Stat.

Per l'estrazione dei dati, sia per Eurostat che per I.Stat, è utilizzata la funzionalità di produzione file di output sopra descritta con gli opportuni parametri di input a seconda dell'esigenza ed i file generati sono memorizzati nella stessa directory degli altri file di output.

## 6. Diffusione e comunicazione dei dati

In linea con la legislazione vigente, la diffusione dei dati delle statistiche strutturali sulle imprese viene effettuata sia in forma aggregata, sia in forma di dati elementari. Nei paragrafi successivi si analizzano le tipologie di diffusione in forma aggregata attraverso il data warehouse I.Stat e le possibilità di comunicazione e accesso ai dati elementari. Una descrizione del sito web dell'Istat fornisce indicazioni e procedure necessarie per accedere ad ognuno di tali servizi.

### 6.1 Diffusione dei dati aggregati: tabelle e indicatori

L'ampliamento del set informativo relativo ai dati SBS, già pubblicati su I.Stat, all'interno del tema "imprese" e sotto tema "competitività", ha richiesto un intervento mirato di aggiornamento sulla filiera di produzione dei dati per I.Stat. Tale aggiornamento ha previsto diversi passaggi tra cui, la revisione del piano di spoglio SBS sulla struttura e competitività delle imprese, il conseguente aggiornamento del Sigis, l'aggiornamento del data warehouse di I.Stat e l'implementazione delle modifiche introdotte sul sito web di I.Stat.

In generale, tale attività si è articolata, da un punto di vista organizzativo, nelle quattro fasi previste dalla ruota di Deming o PDCA (Plan, Do, Check, Action). Di seguito la descrizione delle quattro fasi:

1. nella fase di "Plan", la task force ha pianificato gli aggiornamenti da introdurre alla diffusione dei dati SBS sulla struttura e competitività delle imprese, che si sostanziano essenzialmente nella produzione di nuovi indici di posizione, e in una più ampia disaggregazione delle variabili core per attività economica combinata con le classi di addetti;
2. nella fase del "Do" tali aggiornamenti sono stati tradotti in una prima modifica del piano di spoglio relativo ai dati SBS sulla struttura e competitività delle imprese, di carattere provvisorio;
3. nella fase del "Check" la modifica del piano di spoglio è stata presentata a tutti i soggetti coinvolti nella filiera di diffusione dei dati su I.stat, che comprende sia l'unità informatica di sviluppo del SIGIS, sia l'unità di sviluppo del data warehouse e del sito web di I.Stat. In questa fase sono emersi alcuni miglioramenti da apportare al piano di spoglio, inerenti l'assegnazione dei codici ai nuovi indici di posizione e alle modalità di sviluppo e visualizzazione delle query su I.Stat, che hanno portato ad un modello definitivo e condiviso del piano di spoglio;
4. nella quarta fase "Action", si è proceduto all'aggiornamento del SIGIS e all'aggiornamento del data warehouse e del sito web di I.Stat, in funzione di quanto previsto nel piano di spoglio. Il data warehouse e il sito web di I.Stat, sono stati aggiornati all'interno di un'area provvisoria non accessibile agli utenti esterni. Questa modalità di sviluppo ha consentito una fase ulteriore di test da parte del settore di produzione, sia in merito alla correttezza dei dati caricati, sia alla fruibilità del dato all'interno delle

query. Alla fase di test è seguita l'effettivo aggiornamento del data warehouse e la pubblicazione dei dati in ambiente esterno visibile a tutti gli utenti.

Dalle fasi descritte in precedenza emerge l'importanza della costruzione condivisa del piano di spoglio relativo ai dati SBS sulla struttura e competitività delle imprese. Riguardo a tale aspetto, si precisa in primo luogo che il piano di spoglio consta in un documento, che non entra nel merito della produzione del dato statistico, ma che recepisce le decisioni relative alle modalità di diffusione dello stesso ed è rivolto principalmente ai soggetti coinvolti nella filiera di diffusione dei dati, al fine di:

1. predisporre il sistema informatico di produzione dei file di dati (file in ".csv") per la pubblicazione su I.Stat (nel caso specifico il sistema informatico è Sigis);
2. predisporre il data warehouse e l'ambiente web di pubblicazione di I.Stat.

In tal senso, il piano di spoglio recepisce, in estrema sintesi, quali dati (indicati anche con il temine di "tipi dato"), vengono diffusi (ad esempio: numero di imprese, fatturato, valore aggiunto…), e allo stesso tempo le modalità di costruzione delle query, che indicano come i dati si combinano con le variabili di classificazione (ad esempio: classi di addetti, Ateco…) in fase di diffusione. Quindi nel piano di spoglio si indica quali dei tipi dato prodotti saranno pubblicati, escludendo quei dati che non ricadono nelle finalità di diffusione, e le modalità di costruzione delle query.

Nel caso specifico, nella modifica del piano di spoglio sulla struttura e competitività delle imprese sono stati introdotti i nuovi tipi dato relativi agli indici di posizione (47 nuovi tipi dato). Ad ogni tipo dato (ad esempio: primo quartile del fatturato - migliaia di euro) è stato assegnato un codice, fondamentale per l'attività di aggiornamento del data warehouse di I.Stat relativo al cubo struttura e competitività delle imprese.

Per migliorare la fruibilità del dato, i 47 nuovi tipi dato sono stati suddivisi in 6 gruppi:

1. Indicatori di distribuzione del fatturato
2. Indicatori di distribuzione del valore aggiunto
3. Indicatori di distribuzione del margine operativo lordo
4. Indicatori di distribuzione del costo del lavoro
5. Indicatori di distribuzione del numero di addetti e dipendenti
6. Indicatori di distribuzione della competitività di costo

Ciascun gruppo comprende, per la variabile oggetto di analisi, i tipi dato relativi ai principali indici di posizione e il tipo dato che indica il numero di imprese su cui è stato calcolato l'indice.

Per quanto riguarda le modalità di costruzione delle query si è deciso di realizzarne 6, ciascuna per ogni singolo gruppo di tipi dato, e di diffondere i dati per la variabile di classificazione Ateco con un livello di profondità a 4 digit. Ciascuna query, quindi, comprende i tipi dato del gruppo in testata (con una media di 7 colonne) e l'Ateco in fiancata. Se si fosse utilizzata un'unica query per tutti i 47 tipi dato si sarebbe generata una tabella che per default avrebbe avuto 47 colonne, riducendo la fruibilità del dato. D'altronde, è stato consentito all'utente di poter modificare la visualizzazione della query, in modo da far apparire i tipi dato presenti in altri gruppi; in questo modo l'utente può visualizzare contemporaneamente sia i tipi dato del gruppo, sia i tipi dato di altri gruppi (ad esempio può mettere a confronto il primo quartile del valore aggiunto con il primo quartile del fatturato nella stessa query).

Nel piano di spoglio sulla struttura e competitività delle imprese sono stati introdotte anche le modifiche relative ad una più ampia disaggregazione delle variabili core per attività economica combinata con le classi di addetti. In questo caso, non è stato necessario introdurre nuovi tipi dato, ma individuare, tra quelli già presenti sul piano di spoglio, quelli che sarebbero stati pubblicati con maggior dettaglio. Tali tipi dato (14 in tutto) sono stati raccolti in un nuovo gruppo (gruppo delle variabili core) che è stato combinato con l'Ateco con profondità a 4 digit e con la classe dimensionale. Per quanto riguarda quest'ultima classificazione, le classi dimensionali 0-1 e 2-9 addetti sono state fornite anche per le attività manifatturiere e delle costruzioni. La nuova query così costruita prevede in testata le classi dimensionali e in fiancata l'Ateco, mentre il tipo dato è selezionabile per mezzo di un menù a tendina.

L'introduzione di questa nuova query è stata oggetto di valutazione, in quanto le variabili core venivano già diffuse in altre query indicate nel piano di spoglio e presenti nello stesso ambiente di

I.Stat, ma con un livello di dettaglio inferiore per Ateco e classi di addetti. La possibilità di inglobare il maggiore dettaglio delle variabili core per Ateco e classe dimensionale all'interno delle query già esistenti, è stata infine scartata, sia per ragioni tecniche sia per la fruibilità e visibilità del dato da parte degli utenti esterni.

Infine si indica che le nuove query sono state inserite su I.Stat all'interno del tema "Imprese" e sottotema "Competitività - Statistiche nazionali sulla struttura delle imprese (dati dal 2008)". La query relativa alla più ampia disaggregazione delle variabili core per attività economica combinata con le classi di addetti, costituisce la prima query di default di questo ambiente e prende il nome di "Principali variabili per classi di Ateco e classe di addetti". Le query sugli indici di posizione ricadono all'interno di un ulteriore indice specifico di questo ambiente relativo agli "indicatori di distribuzione".

## 6.2 Comunicazione e accesso ai dati elementari

L'integrazione dei dati Frame con fonti differenti (amministrative, PMI e SCI) permette la produzione di due diversi file di microdati: "Frame-SBS - Sistema integrato di dati amministrativi e dati di indagine per la stima degli aggregati economici sulle imprese", e "TEC-FrameSBS - Struttura e performance economica delle imprese esportatrici". Entrambi i file possono essere oggetto di richiesta da parte degli enti Sistan e sono messi a disposizione della comunità scientifica presso il laboratorio ADELE. I dati Frame sono quindi archiviati nel sistema ARMIDA (ARchivio di MIcroDAti) le cui principali finalità consistono nel conservare metadati e microdati validati delle rilevazioni condotte dall'Istat, e favorire il riutilizzo dei microdati per finalità statistiche da parte di utenti esterni.

Il file Frame-SBS contiene le venti variabili seguenti:

1. Codice impresa (codice Asia)
2. Codice regione
3. Numero dipendenti
4. Numero addetti
5. Ateco a 4 cifre
6. Appartenenza a gruppi di impresa
7. Impresa artigiana
8. Classe di addetti
9. Ricavi della vendita di beni e della prestazione di servizi
10. Altri ricavi e proventi
11. Costi per materie prime, sussidiarie, di consumo e di merci
12. Costi per servizi
13. Costi per godimento di beni di terzi
14. Costi per il personale
15. Salari e stipendi (retribuzioni lorde)
16. Oneri diversi di gestione
17. Valore aggiunto
18. Margine operativo lordo
19. Esportazioni di beni (fonte: Istat-Coe)
20. Importazioni di beni (fonte: Istat-Coe)

I microdati "TEC-FrameSBS -Struttura e performance economica delle imprese esportatrici" derivano dall'integrazione di tre diverse fonti statistiche: il registro statistico delle imprese attive (Asia), il registro degli operatori che realizzano scambi con l'estero di merci (Coe) e il file Frame-SBS. Le principali variabili di interesse sono: valore aggiunto, costo del lavoro, fatturato, acquisti di beni e servizi, valore delle esportazioni (totale e valore decomposto per area geografica e raggruppamenti principali di prodotti), valore delle importazioni (totale e valore decomposto per area geografica e raggruppamenti principali di prodotti), numero di prodotti esportati e importati, numero di paesi/aree geografiche all'export e all'import.

## 6.3 Descrizione del sito web dell'Istat

Sul sito dell'Istituto i microdati hanno una sezione interamente dedicata, accessibile dall'home page alla voce di menù "Prodotti". All'interno di essa si illustrano le diverse tipologie di file prodotte dall'Istituto e per ognuna di queste le relative condizioni di rilascio e gli accordi di utilizzo.

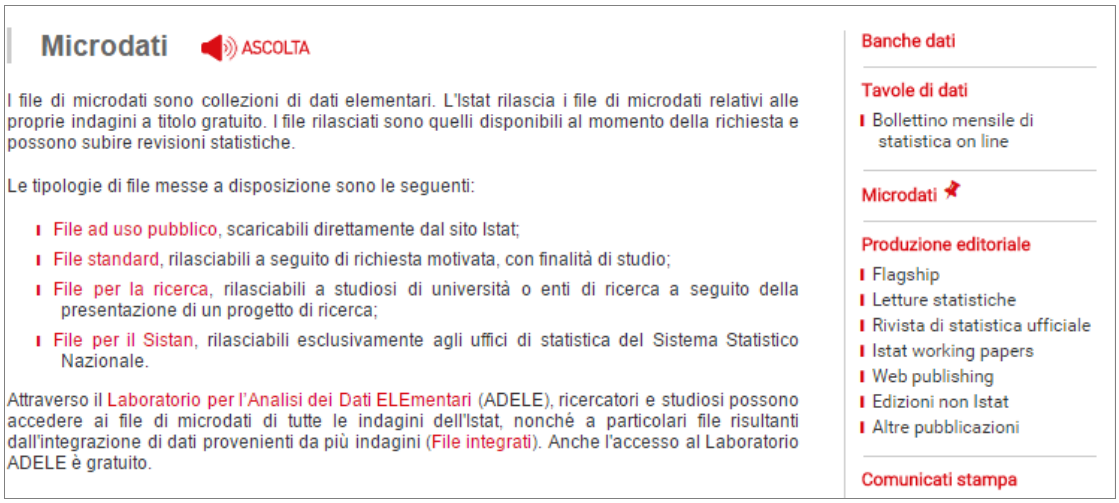
Per tale motivo si forniscono in download soltanto i dati informativi dei microdati, vale a dire i metadati con le liste delle variabili e le note metodologiche.

**Figura 9 – Collocazione della pagina "Microdati" nel menù del sito Istat**


Nell'archivio del sito web i file di microdati sono organizzati per tipologia:

- File ad uso pubblico, collezioni di dati elementari scaricabili liberamente e gratuitamente dagli utenti previa accettazione delle condizioni d'uso e registrazione o autenticazione al sito www.istat.it;
- File standard, file privi di elementi identificativi diretti rilasciati su richiesta motivata a qualsiasi utente per finalità di studio e ricerca;
- File per la ricerca, file con elevato livello di dettaglio informativo, rilasciati esclusivamente a determinati soggetti (studiosi di università o enti di ricerca) a seguito della presentazione di uno specifico progetto di ricerca;
- File per il Sistan, file per le richieste di dati elementari da parte di uffici di statistica del Sistema statistico nazionale ai fini dell'attuazione del Programma statistico nazionale e/o per l'esecuzione di trattamenti connessi all'attività istituzionale o all'ambito territoriale del richiedente;
- File integrati, particolari file accessibili tramite il Laboratorio ADELE, risultanti dall'integrazione di dati provenienti da più indagini o fonti, predisposti al fine di promuovere l'ampliamento delle informazioni a livello di singola impresa.

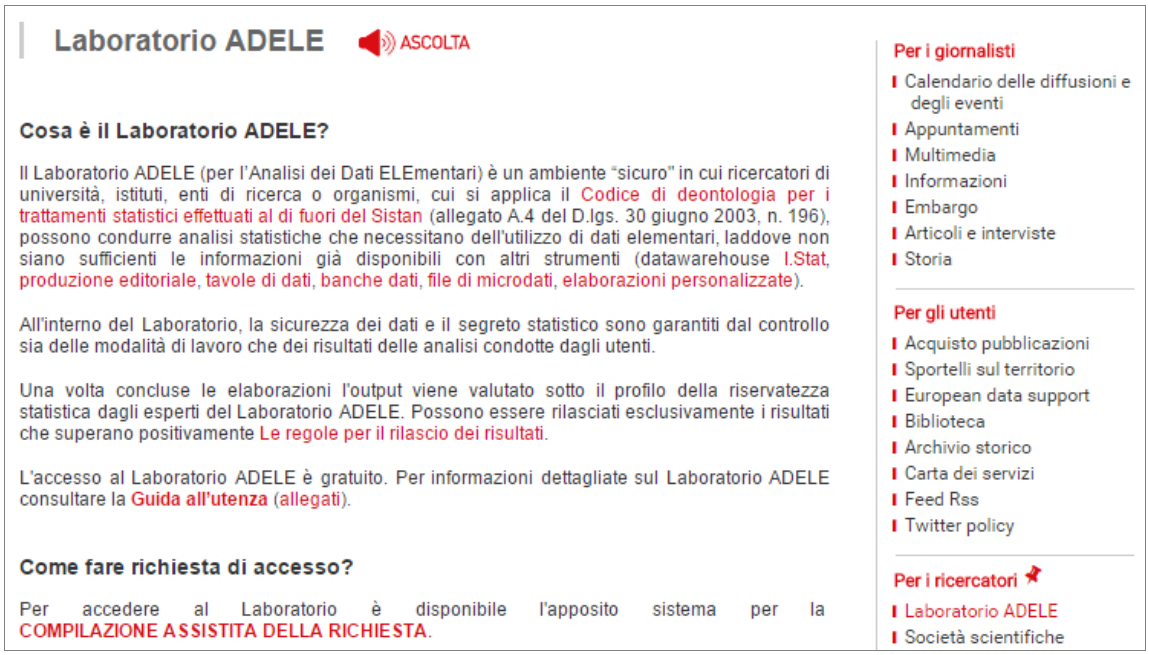
**Figura 10 – Struttura della pagina dei microdati**



In particolare nell'archivio dei "File integrati" sono presenti i metadati sulla Struttura e performance economica delle imprese esportatrici (TEC-FrameSBS), prodotta annualmente a partire dal 2013, basata sulle informazioni integrate presenti nell'archivio delle imprese esportatrici (TEC) con le principali variabili economiche ora disponibili per tutte le imprese dell'industria e dei servizi (Frame-SBS utilizzato a regime dal 2012 per la produzione di stime SBS).

I microdati del Frame-SBS (anni 2012-2013) sono invece accessibili in forma elementare attraverso il Laboratorio per l'Analisi dei Dati ELEmentari (ADELE), un ambiente in cui ricercatori, studiosi, istituti, enti di ricerca o organismi possono condurre analisi statistiche che necessitano dell'utilizzo dei dati elementari e dei file integrati, laddove non siano sufficienti le informazioni già disponibili con altri strumenti (data warehouse I.Stat, produzione editoriale, tavole di dati, banche dati, file di microdati, elaborazioni personalizzate).

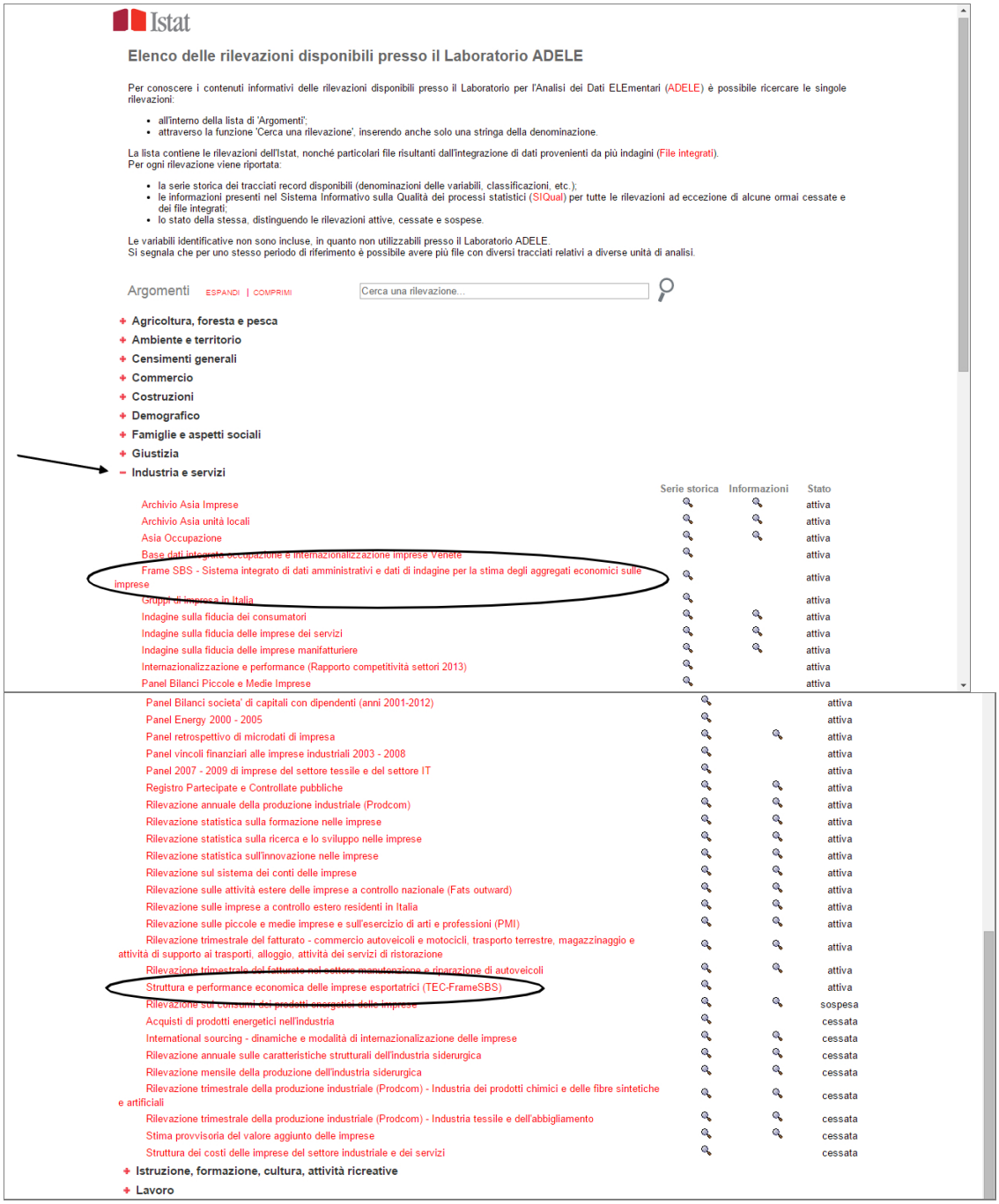
**Figura 11 – La pagina del Laboratorio ADELE**




Nel Laboratorio ADELE l'elenco delle rilevazioni Istat sono raggruppate per argomenti. Nello specifico l'argomento "Industria e servizi" contiene la lista delle variabili del Frame-SBS:

- *Frame SBS - Sistema integrato di dati amministrativi e dati di indagine per la stima degli aggregati economici sulle imprese;*
- *Struttura e performance economica delle imprese esportatrici (TEC-FrameSBS).*

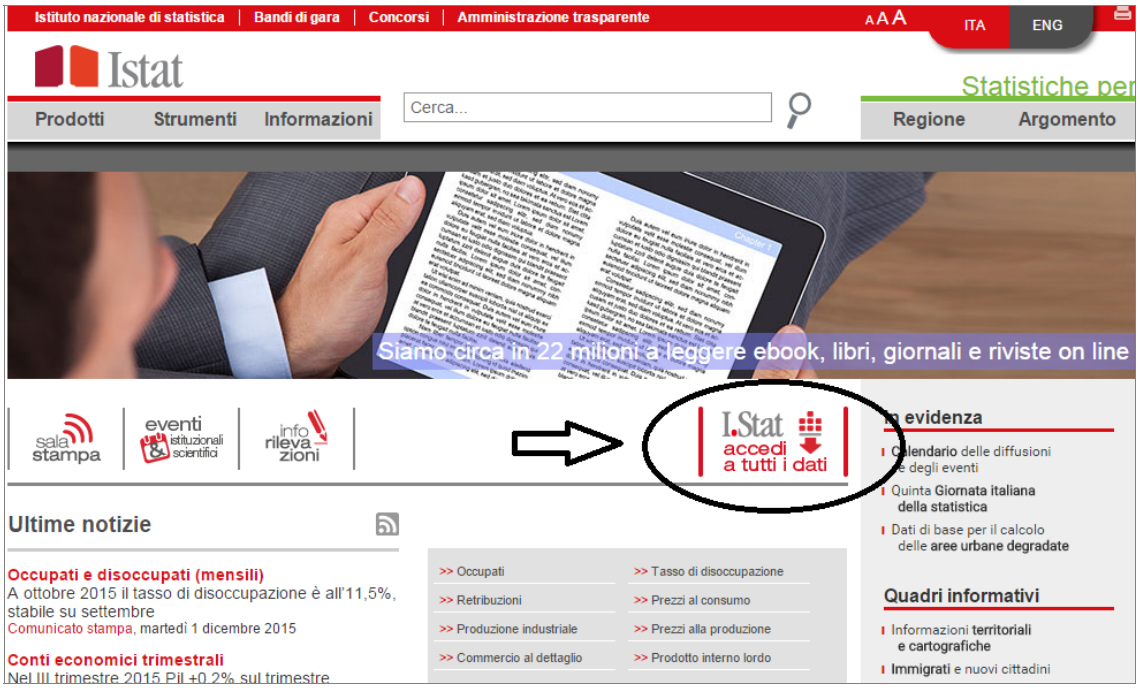
**Figura 12 - Lista delle variabili del Frame-SBS disponibili presso il Laboratorio ADELE**




Per quanto riguarda la diffusione in forma aggregata dei dati del Frame, unitamente a quelli della rilevazione PMI (per le variabili non desumibili dalle fonti amministrative) e quelli della rilevazione SCI, sul data warehouse I.Stat sono state pubblicate al tema "Imprese" le tabelle contenenti dati secondo diversi livelli di disaggregazione per gli anni di riferimento 2012 e 2013. Per gli anni precedenti le tabelle fanno riferimento alle stime della rilevazione PMI e SCI

Al data warehouse si accede tramite l'apposito banner presente nell'home page del sito istituzionale.

**Figura 13 – Banner di I.Stat presente in home page**




Attraverso la statistica report "Struttura e competitività delle imprese" si riassumono i principali risultati economici delle imprese, secondo quanto disposto dal Regolamento Ue n. 295/2008 per le statistiche strutturali (*SBS - Structural Business Statistics*). A partire dall'anno 2012, l'elaborazione utilizza il nuovo sistema informativo Frame (una base di microdati di fonte amministrativa trattati statisticamente) in combinazione con quelli della rilevazione campionaria PMI e con l'insieme dei risultati della rilevazione censuaria SCI. La statistica viene pubblicata sotto forma di comunicato stampa, corredato di tavole di dati (in forma aggregata), note metodologiche e glossario, e sul sito web è collocata all'interno del menù "Statistiche per argomento" alle voci Imprese, Industria e costruzioni e Servizi.

Circa la produzione editoriale dell'Istituto, il Rapporto sulla competitività dei settori produttivi - Anno 2014 si avvale del Frame per la stima della misura del grado di efficienza produttiva dell'impresa.

Le innovazioni metodologiche del sistema informativo Frame-SBS sono inoltre dibattute attraverso l'organizzazione di eventi, come il workshop dal titolo "Nuove informazioni statistiche per misurare la struttura e la performance delle imprese italiane" e il workshop "Micro dati per l'analisi della performance delle imprese: fonti, metodologie, fruibilità, evidenze internazionali". Nella pagina di entrambi gli eventi si possono consultare e/o scaricare le presentazioni degli interventi con i relativi abstract, disponibili anche sul social network Slideshare.

Infine tutte le informazioni riguardanti i file di microdati sono rapidamente accessibili dall'home page attraverso la *search box*, la funzione di ricerca fornita da GSA (*Google Search Appliance*), grazie alla quale gli utenti possono affinare i risultati mediante i filtri della navigazione dinamica (periodo di riferimento, tipo di documento, argomento, data di pubblicazione, ecc.).

**Figura 14 – Ricerca tramite GSA della key phrase "sbs frame" con filtro "Microdati integrati"**


## Conclusioni

Nell'ambito delle rilevazioni sulle imprese (SCI, PMI), il ricorso ad archivi amministrativi ha permesso di aumentare il contenuto informativo disponibile. La produzione del Frame-SBS per gli anni 2012 e 2013 ha consentito di ottenere stime affidabili anche per domini più fini rispetto quelli utilizzati in precedenza. Si è valutata quindi la possibilità di ampliare i dettagli informativi rilasciabili (maggiore disaggregazioni nei dati, informazioni sulla distribuzione di alcune variabili e indicatori per diversi domini di stima) anche nel rispetto della normativa vigente in materia di protezione dei dati personali.

L'analisi dei dati, sotto il profilo della tutela della riservatezza, ha confermato la possibilità di aumentare i dettagli di diffusione senza perdere in contenuto informativo rilasciato.

Partendo dai domini di diffusione previsti dal regolamento europeo sulle statistiche strutturali sulle imprese n° 295/2008 (SBS), per le *variabili core* sono state individuate le disaggregazioni di maggior interesse per le analisi economiche. Si è quindi proceduto valutando il numero di soppressioni necessario per limitare il rischio di intrusione nei nuovi domini: le classi di addetti [0-1] e [2-9] sono state adottate per tutti i settori di attività economica; si è operata la disaggregazione dei dati secondo Nace a quattro cifre e classi di addetti.

L'applicazione delle regole di riservatezza ai dati così dettagliati, attraverso il software generalizzato τ-Argus, ha permesso di individuare una valida soluzione: rispetto alle diffusioni precedenti, nonostante il maggior numero di soppressioni dovuto al maggior numero di domini di pubblicazione, il contenuto informativo pubblicato è risultato complessivamente maggiore.

Nel caso delle *variabili core*, per i domini con almeno 50 unità a livello di Sezione e Nace a 2, 3 e 4 cifre (senza classe di addetti), alle frequenze e alle intensità sono stati affiancati valori medi e di variabilità. I primi sono stati individuati tenendo in considerazione l'asimmetria distributiva dei fenomeni economici; si è proceduto quindi al calcolo dei quartili (o, per ragioni di riservatezza, a un loro valore medio), preferendo indicatori non influenzati da valori estremi. Come misura di variabilità è stata rilasciata la deviazione standard.

Il documento oltre a descrivere le specifiche del regolamento sulle statistiche strutturali e il quadro normativo di riferimento in materia di protezione dei dati personali, illustra le procedure informatiche e gli aspetti relativi alle regole per la tutela statistica della riservatezza. Nel documento sono inoltre descritte le procedure necessarie alla preparazione delle serie da trasmettere all'Eurostat, alla costruzione di tabelle interrogabili nel data warehouse I.Stat e alla predisposizione dei dati per il Laboratorio ADELE.

Un aspetto che rimane critico è la possibilità di diffondere i dati del Frame in forma più capillare secondo schemi differenti (ad esempio per imprese artigiane, appartenenti a gruppi, esportatrici, ecc.). Il limite in questo senso è rappresentato da problemi definitori, nonché dalle regole di tutela della riservatezza valutate anche in termini di contenuto informativo pubblicato: l'aumento dei dettagli rilasciati deve essere valutato rispetto al numero di soppressioni necessarie per la protezione dei dati e alla conseguente perdita di informazione.

Pertanto, al crescere delle esigenze conoscitive e al mutare dei regolamenti comunitari occorre adeguare il processo descritto e programmare le attività di diffusione secondo i nuovi domini, nel rispetto dei vincoli della riservatezza.

# Appendice A
# Nota sulla scelta dei domini SBS cui associare indici di posizione e di variabilità

Al fine di ampliare l'offerta informativa in termini di indici di posizione e di variabilità da associare alle variabili diffuse in forma aggregata tramite I.Stat alcune analisi sono state condotte per definire se e a quali condizioni un tale ampliamento fosse tangibile per gli utenti e opportuno per l'Istituto.

In particolare si è dovuto tener conto della numerosità dei singoli domini di riferimento per il rilascio delle statistiche SBS (tenendo conto dell'ampliamento previsto per gli stessi rispetto a quanto richiesto da Eurostat) e della natura stessa delle variabili trattandosi di una quantificazione diretta dei fenomeni per le variabili rilevate e di una quantificazione 'pro-capite' o di un rapporto per quelle derivate.

Per quanto riguarda il primo aspetto, evidentemente, fornire indicatori di posizione e di variabilità per domini contenenti un numero esiguo di unità può alterare i criteri adottati al fine di preservare la confidenzialità statistica aumentando le informazioni utili all'identificazione del valore osservato di una singola unità statistica. Ad esempio, fornire i valori del primo e terzo quartile di una data variabile uguali o simili tra di loro circoscrive l'osservazione associabile ad una singola unità statistica in un intervallo molto piccolo se la numerosità del dominio non è tale che sia ignota (o almeno incerta) la posizione dell'unità nella graduatoria della variabile stessa dentro il dominio. Per quanto riguarda il secondo aspetto, la circostanza ora rappresentata, può risultare meno problematica nel caso delle variabili derivate in quanto il valore di un rapporto non fornisce un'informazione precisa sul valore delle variabili che lo determinano.

Per le considerazioni fatte, i criteri per individuare le celle a rischio dal punto di vista della riservatezza statistica, e in particolare la regola della soglia (pari a 3), non risultano adeguate a garantire la confidenzialità con l'incremento dei contenuti informativi desiderati. Per analogia con le politiche adottate per il rilascio dei risultati delle elaborazioni effettuate dagli utenti presso il Laboratorio di analisi dei dati elementari dell'Istat (Laboratorio ADELE) con riferimento al tipo di indicatori previsti, è stato suggerito di considerare domini contenenti almeno 50 unità statistiche. In pratica, si è scelto di adottare la regola della soglia pari a 50 per il rilascio degli indici di posizione e variabilità associati alle statistiche SBS nei domini di riferimento.

Per verificare l'opportunità di un tale criterio sono state calcolate le numerosità dei 1809 domini totali (combinazione di Ateco fino a 3 digit e classe di addetti) ed è risultato che 569 di questi contengono meno di 50 unità (Tavola A1). Tuttavia, dei domini sotto la soglia solo 19 riguardano domini definiti da Ateco a 3 cifre (di cui 3 soggetto a confidenzialità primaria per i criteri SBS) e 5 da Ateco a 2 cifre (di cui 1 soggetto a confidenzialità primaria per i criteri SBS). Quindi, per i domini definiti esclusivamente da raggruppamenti Ateco prescindendo dalla classe dimensionale, l'applicazione di una soglia pari a 50 non altera in maniera sensibile lo schema di confidenzialità previsto dai criteri SBS (soglia pari a 3).

**Tavola A1 – Numerosità dei domini per combinazioni di Ateco e classe di addetti e numero di domini confidenziali per livelli di soglia pari a 3 e 50**

| Domini | Numerosità | Domini confidenziali per soglia | |
|---|---|---|---|
| | | pari a 50 | pari a 3 |
| Ateco 2 cifre | 77 | 5 | 1 |
| Ateco 3 cifre | 236 | 19 | 3 |
| Ateco 2 e 3 cifre e classe di addetti | 1496 | 545 | 91 |
| Totale | 1809 | 569 | 95 |

La conclusione che se ne è tratto è stata che i domini definiti dalla combinazione delle variabili

classe di addetti e Ateco sono stati esclusi dal rilascio di indici di posizione e di variabilità associati alle statistiche SBS per non fornire una informazione eccessivamente disomogenea nel complesso. D'altro canto, per i domini definiti solo dalla variabile Ateco è possibile rilasciare indici di posizione e di variabilità per le principali variabili e per alcuni indicatori in maniera (quasi del tutto) omogenea sul complesso dei domini (disaggregazioni) a 1 lettera e a 2, 3 e 4 cifre di attività economica pur applicando una soglia pari a 50 sulla numerosità degli stessi.

# Appendice B
# Regole di rischio e livelli di protezione

Nella tavola B1 sono riportate, a titolo esemplificativo, le frequenze e le celle (x1, y1) (x1, y2) (x2, y1) (x2, y2) in cui i dati sono stati soppressi. I flag A,B,C,D, sostituiscono i valori oscurati. Nell'esempio si assume che (x1, y1) sia l'unica cella a rischio e che le lettere B, C, D siano utilizzate come flag per le soppressioni secondarie.

**Tavola B1 – Un esempio di protezione dei dati**

| | **y1** | **y2** | **Tot** |
|---|---|---|---|
| **x1** | A | B | 6 |
| **x2** | C | D | 8 |
| **tot** | 4 | 10 | 14 |

Il valore a rischio, sostituito con flag A, non è individuabile con esattezza. È tuttavia possibile, utilizzando i valori marginali, definire degli intervalli che contengono i valori oscurati: i valori mancanti sono equiparabili a intervalli ricavabili con un sistema di equazioni lineari.

Nel caso ad esempio:

A+B=6
C+D=8
A+C=7
B+D=4

Le quattro equazioni riportate, unite al vincolo di non negatività dei valori (valido per alcune variabili risposta), permettono di costruire gli estremi superiore e inferiore degli intervalli di esistenza (*feasibility interval*) per tutte le celle soppresse, con particolare riferimento alla cella a rischio.

Per una cella sensibile oscurata, il rischio di violazione è considerato adeguato se il sistema di equazioni lineari, come quello descritto, definisce intervalli sufficientemente ampi attorno al valore a rischio soppresso. In caso contrario si parla di "sotto-protezione" della tabella o, se gli intervalli risultano di ampiezza nulla, di mancata protezione.

In taluni casi la parametrizzazione della regola di rischio adottata definisce i *feasibility interval*. Nella tabella seguente sono riportati i livelli di protezione superiore (*Upper protection level*, UPL) relativi ad alcune regole basate su misure di concentrazione.

**Schema B1 - Regole di rischio e livello di protezione**

| Regola di rischio | Livello di protezione superiore |
|---|---|
| Dominanza(n-k) | (100/k)(x1+x2+….xn)-X |
| Rapporto(p% ) | (p/100)x1-(X-x1-x2) |
| Priori-posteriori(p,q) | (p/q)x1-(X-x1-x2) |

Nello Schema 1, "xi" rappresenta l'i-esimo contributo, "X" il totale (ovvero la somma) di tutti i contributi; "n" è il numero di contributori rispetto al quale si valuta la regola della dominanza; "K" rappresenta la percentuale (massima) del contributo totale che può essere detenuta dai primi "n" contributori (valore soglia); "p" e "q" sono probabilità.

L'intervallo di protezione è ottenuto sommando e sottraendo dal valore vero della cella a rischio il livello superiore di protezione riportato nello Schema 1.

Non è possibile individuare un intervallo di protezione basato sulla parametrizzazione della regola della soglia. Per questa ragione alcuni autori sostengono che essa sia adeguata solo per proteggere i dati da un'identificazione esatta. È possibile tuttavia fissare a priori un livello di protezione minimo espresso in percentuale del valore della cella a rischio. Questa soluzione (implementata in τ-Argus) fornisce garanzie sull'ampiezza dell'intervallo ma non alla sua simmetria intorno al valore soppresso.

# Riferimenti bibliografici

*Codice in materia di protezione dei dati personali,* D.Lgs no. 196 of June 30, 2003, Gazzetta Ufficiale No 174, Supp. no. 123 (July 29, 2003) annex A.3 ('*Codice di deontologia e di buona condotta per i trattamenti di dati personali a scopi statistici e di ricerca scientifica effettuati nell'ambito del Sistema statistico nazionale*'). http://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2003-06-30;196!vig=.

De Wolf, P.P. 2002. "HiTaS: a heuristic approach to cell suppression in hierarchical tables." In *Inference control in statistical databases: from theory to practice*, edited by J. Domingo-Ferrer. Vol. 2316 of *Lecture Notes in Computer Science*, 74-82. Heidelberg: Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/3-540-47804-3_6.

De Wolf, P.P. 2007. "Cell suppression in a special class of linked tables. WP. 21 presented at the Joint UNECE/Eurostat work session on statistical data confidentiality, Manchester, United Kingdom, December 17-19.
http://www.unece.org/stats/documents/2007/12/confidentiality/wp.21.e.pdf.

Giessing S. 2001. "New tools for cell suppression in τ-Argus: one piece of the CASC project work draft. WP. 2 presented at the Joint UNECE/Eurostat work session on statistical data confidentiality, Skopje, The former Yugoslav Republic of Macedonia, March 14-16. http://www.unece.org/stats/documents/2001/03/confidentiality/2.e.pdf.

Hundepool A. et al. 2010. *Handbook on Statistical Disclosure Control Version 1.2*. http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf .

Virgili L., and L. Franconi. 2009. "Disclosure protection of non-nested linked tables in business statistics." WP.36 presented at the Joint UNECE/Eurostat work session on statistical data confidentiality, Bilbao, Spain, December 2-4.
http://www.istat.it/it/files/2013/12/Franconi_Virgili_wp.36.e.pdf.

Statistics Netherlands. 2008. *τ-Argus Version 3.3 User's Manual*. The Netherlands. http://neon.vb.cbs.nl/casc/Software/TauManualV3.3.pdf.