istat
working
papers

N. 4
2013

Uno schema standardizzato per il trattamento statistico di un archivio amministrativo

Antonio Bernardi, Fulvia Cerroni e Viviana De Giorgi

Istat
Istituto Nazionale
di Statistica

istat
working
papers

N. 4
2013

Uno schema standardizzato per il trattamento statistico di un archivio amministrativo

Antonio Bernardi, Fulvia Cerroni e Viviana De Giorgi

Istat

Comitato scientifico

Giorgio Alleva, Emanuele Baldacci, Francesco Billari
Tommaso Di Fonzo, Andrea Mancini, Roberto Monducci
Fabrizio Onida, Linda Laura Sabbadini, Antonio Schizzerotto

Comitato di redazione

Alessandro Brunetti, Patrizia Cacioli, Marco Fortini
Romina Fraboni, Stefania Rossetti, Daniela Rossi
Maria Pia Sorvillo

Segreteria tecnica

Maria Silvia Cardacino, Laura Peci, Marinella Pepe, Gilda Sonetti

Istat Working Papers

Uno schema standardizzato per il trattamento statistico di un archivio amministrativo

N. 4/2013

ISBN 978-88-458-1749-6

Istituto nazionale di statistica
Servizio Editoria
Via Cesare Balbo, 16 – Roma

Uno schema standardizzato per il trattamento statistico di un archivio amministrativo

Antonio Bernardi, Fulvia Cerroni e Viviana De Giorgi

Sommario

Il Servizio Acquisizione, gestione e diffusione delle fonti amministrative (DAM), facente parte della Direzione centrale registri statistici, dati amministrativi e statistiche sulla Pubblica Amministrazione (DCAR), ha istituzionalmente assegnato il compito di sviluppare "metodologie per l'analisi di qualità e il trattamento di dati amministrativi". Il Servizio, pertanto, ha sviluppato alcune metodologie standard da adottare per verificare la qualità dei dati tratti da archivi amministrativi, al fine di fornire delle regole da seguire per l'utilizzo nei processi di produzione statistica di tale tipologia di dati che, sempre di più, si sta incrementando. In questo documento si presentano i risultati ottenuti applicando le indicate metodologie all'archivio degli Studi di Settore, gestito dall'Agenzia delle Entrate. La caratteristica principale di tale fonte è la raccolta di informazioni su una molteplicità di aspetti aziendali, strutturali, occupazionali, fiscali e contabili riferiti a circa 4 milioni di piccole e medie imprese italiane. La scelta di questo archivio è, infatti, dovuta sia alle sue indicate potenzialità informative, sia alla comparabilità dei suoi dati con le variabili richieste dal regolamento comunitario sulle statistiche strutturali. E sebbene tutte le procedure descritte in questo documento siano state sviluppate e applicate sui dati contenuti negli Studi di Settore, esse potranno essere estese a fonti simili, dopo aver ovviamente realizzato gli opportuni adattamenti. Il presente documento è così articolato: un primo paragrafo con le caratteristiche generali del processo di trattamento statistico di archivi amministrativi; un secondo, articolato in sottoparagrafi, con gli schemi delle procedure statistiche a essi applicabili; un terzo, con alcuni risultati ottenuti dall'archivio degli Studi di Settore; seguono in chiusura alcune osservazioni.

Parole chiave: validazione di dati amministrativi per usi statistici, variabili di benchmark, dati economico-fiscali

Abstract

This paper describes some standard methodologies for assessing the quality of administrative data, and provides some rules to be followed for using them in statistical production processes. It presents the results of applying such methods on a source provided to Istat by the Italian Tax Authority, "Studi di Settore" (Sector Studies), a fiscal mechanism used in Italy to determine the presumed income taxpayers should declare. The main feature of this source is the collection of information on a variety of aspects of about 4 million small and medium-sized Italian enterprises: structure, employment, taxation, accounting items. The choice of this source is due to both its informative potentiality and the comparability of its data with the variables required by the Eu regulation on structural business statistics. Though all the procedures described in this document have been developed and applied to Sector Studies, they may be extended to other administrative sources. This paper is structured as follows: the first paragraph describes the standard methodologies for the statistical treatment of administrative archives; the second paragraph depicts the diagrams of statistical procedures; the last paragraph shows some results obtained from Sector Studies.

Keywords: assessing administrative data for statistical uses, benchmark variables, accounting data

Indice

1. Caratteristiche generali del processo di validazione

Da alcuni anni l'Istat ha intrapreso la strada dell'utilizzo di dati provenienti dagli archivi amministrativi delle pubbliche amministrazioni e da soggetti privati, garantendone le connessioni e la coerenza con i dati rilevati dall'Istat stesso attraverso le indagini statistiche condotte direttamente sulle imprese, sulle famiglie e le istituzioni. L'utilizzo di tali dati richiede preliminarmente una serie di elaborazioni che, nel loro complesso, costituiscono quello che viene comunemente definito *processo di validazione statistica di un archivio amministrativo*.

Come noto, l'impiego dei dati in discorso va inquadrato nell'ambito della strategia di contenimento della spesa pubblica che, per gli enti istituzionalmente preposti alla realizzazione di statistiche pubbliche, poggia principalmente su un utilizzo sempre diffuso di informazioni già in possesso della Pubblica Amministrazione, cioè sull'utilizzo di prodotti già realizzati dal settore pubblico. Questa impostazione consente infatti un più efficiente utilizzo delle risorse disponibili per la produzione statistica ufficiale, poiché riduce la necessità di ricorrere a indagini ad hoc per rilevare dati già esistenti, creando le condizioni per un efficace interscambio delle informazioni registrate nei sistemi informativi costituiti dalle diverse pubbliche amministrazioni.[1]

L'adozione di questo criterio di fondo appare molto promettente; ne è la prova il fatto che anche gli altri Istituti nazionali di statistica dell'UE stanno percorrendo la stessa strada. La sua applicazione, tuttavia, presenta alcuni problemi. Il più importante di questi è che la raccolta e il trattamento dei dati amministrativi va al di là delle possibilità di controllo da parte degli Istituti nazionali di statistica, per cui elementi fondamentali come le definizioni delle unità di rilevazione, delle variabili di classificazione e delle unità di analisi derivano da norme amministrative e molto spesso non corrispondono a quelle adottate dalla nomenclatura statistica, la cui applicazione è invece richiesta agli Istituti nazionali di statistica (Wallgren e Wallgren, 2007). Questa circostanza comporta anche la necessità di un notevole sforzo da compiere per definire in maniera chiara la fruibilità statistica dei dati amministrativi (Bakker, 2009; ESC, 2007; Van der Laan, 2000). Un'ulteriore criticità è che per lo svolgimento delle sue attività l'Istat sarà sempre più dipendente dalla possibilità di disporre di dati raccolti da altri soggetti e, pertanto, l'utilizzo di tali dati richiederà preliminarmente una serie di elaborazioni che, nel loro complesso, costituiscono quello che viene comunemente definito *processo di validazione statistica di un archivio amministrativo*.

Da quanto premesso emerge quanto sia importante definire una procedura che permetta di determinare, in maniera sistematica e standardizzata, e quindi verificabile, la qualità complessiva dell'archivio amministrativo che si intende utilizzare per scopi statistici, consentendo di portare a compimento, quando possibile, il processo di validazione statistica sopra menzionato. Nel caso che la loro qualità risulti adeguata, i dati amministrativi potranno essere messi a disposizione delle strutture organizzative dell'Istituto interessate, per essere utilizzati a fini statistici.

1.1 L'attività del Servizio DAM

In questa nota sono descritte e documentate le esperienze fin qui maturate dal *Servizio acquisizione gestione e diffusione delle fonti amministrative* (DAM) con il processo di validazione delle variabili di tipo economico-contabile contenute nei quadri F e G degli *Studi di Settore* (d'ora in poi Sds). Tali sperimentazioni, iniziate a fine 2007 e progressivamente ridefinite e calibrate (Bernardi et al., 2008) costituiscono un primo importante passo del processo di validazione attraverso il quale è possibile trasformare un archivio amministrativo in un archivio statistico (Bernardi et al., 2010). Questa attività ha permesso di mettere a disposizione di altre strutture dell'Istituto (la Direzione centrale della contabilità nazionale e la Direzione centrale delle Statistiche strutturali sulle imprese) l'archivio Sds valida-

[1] Le motivazioni che spingono gli istituti di statistica ufficiale a utilizzare gli archivi amministrativi, nella fattispecie quelli fiscali (Eurostat, 1999), per scopi statistici al fine di integrare e/o sostituire parti di indagini, riguardano soprattutto la possibilità di ridurre l'onere statistico sulle imprese e di reperire informazioni aggiornate utili ai fini statistici già disponibili presso le amministrazioni per altri scopi. Capita, infatti, che le imprese, oltre a essere obbligate a compilare i moduli amministrativi, ricevano più questionari di indagine, che chiedono informazioni tra loro simili e talvolta riconducibili a quelle già dichiarate, per esempio, alle autorità fiscali.

to (in alcune sue componenti) per essere utilizzato per fini statistici. In particolare l'archivio fornito è stato utilizzato per l'integrazione dei dati dell'indagine sulle piccole e medie imprese e sull'esercizio di arti e professioni (Pmi), attraverso la specifica procedura di recupero delle mancate risposte totali dell'anno 2008 (Casciano et al., 2011b), applicata dopo aver constatato i riscontri positivi di una sperimentazione effettuata sui dati dell'anno 2007 (Casciano et al., 2010).

L'ultimo aspetto da segnalare riguarda l'estensibilità ad altre fonti del metodo di analisi applicato agli Sds. Concluso tutto il ciclo delle sperimentazioni considerate in questa prima fase di ricerca di metodologie standard per il trattamento e la validazione statistica di dati amministrativi, si può infatti affermare che la metodologia messa a punto e descritta nella presente nota, pur essendo stata elaborata attraverso un training su dati aventi specifiche caratteristiche, è generalizzabile a una qualunque fonte amministrativa (Cerroni e De Giorgi, 2010). Tale conclusione deriva dalle modalità che sono state seguite nella realizzazione di questi percorsi sperimentali, che hanno permesso di definire una procedura standard che può essere ripetuta, con le opportune integrazioni, su altri archivi amministrativi.

1.2 Utilizzo di dati amministrativi

La chiave d'uso degli archivi amministrativi consiste nell'analisi degli stessi al fine di riuscire a costruire regole che permettano di passare dalle informazioni amministrative a quelle statistiche. L'analisi preliminare della fonte amministrativa e l'individuazione di una o più fonti di controllo (benchmark), unitamente al processo di validazione delle informazioni contenute nella fonte e al rilascio di metadati e documentazione, costituiscono le principali fasi per un giudizio sulle possibilità di utilizzo a fini statistici della fonte. Inoltre sono la premessa per una sua validazione e un eventuale impiego per integrare e/o sostituire i dati rilevati dalle indagini statistiche correnti svolte direttamente sulle imprese, sulle famiglie e sulle istituzioni. Tale impiego, oltre a incrementare la base informativa utilizzabile per la produzione statistica, consente di migliorare la qualità e la tempestività dei dati rilevati con indagini *ad hoc* o anche aumentare il dettaglio con cui le informazioni sono rese disponibili per gli utenti. A questi vantaggi si aggiunge una tendenziale riduzione, o quanto meno un contenimento, dell'onere statistico sui rispondenti in ragione di maggiori possibilità di ridurre la dimensione delle indagini e, quindi, i costi che un istituto di statistica deve sostenere per le proprie attività.

L'analisi, ovviamente, oltre a puntare a ottenere possibili vantaggi, ha anche alcune conseguenze ed effetti collaterali che vanno comunque presi in considerazione. In particolare, la non partecipazione dell'istituto di statistica al processo di produzione del dato amministrativo potrebbe comportare la mancanza di controllo alla fonte, che potrebbe dare luogo alla presenza di valori mancanti, di *outlier* e di problemi nei dati. Ciò sembra indicare la necessità di considerare una collaborazione fra gli esperti dell'ente fornitore del dato e l'istituto di statistica, per raccordare le diverse definizioni e classificazioni delle unità e delle variabili, o quanto meno per trovare le regole di riconciliazione delle stesse e, obiettivo finale, propendere verso una strategia comune. La finalità dell'analisi generale della fonte amministrativa, preliminare a un qualunque trattamento e analisi dei contenuti, consiste nel dare un giudizio, anche in forma sintetica, sulla possibilità d'uso della stessa a fini statistici. Tale fase serve a dare indicazioni più precise su quali parti (variabili e/o record) sono direttamente utilizzabili e sui tempi di disponibilità della fonte al fine di capire se effettivamente utilizzarle significhi un guadagno in termini di riduzione dell'onere statistico e di rispetto della tempestività del dato di indagine e se esistono, a tale riguardo, i necessari presupposti.

In generale, gli elementi da indagare di un archivio amministrativo, pur dipendendo dagli usi che ci si prefigge di realizzare con lo stesso, riguardano l'individuazione della popolazione di riferimento, delle unità di rilevazione e dell'insieme di valori possibili per ciascuna variabile, il grado di copertura della fonte in termini sia di unità sia di variabili, la portabilità del formato dei dati, le classificazioni e disaggregazioni di variabili, le classificazioni delle unità e la presenza di una o più fonti di benchmark, un aspetto evidentemente molto importante nella successiva fase di validazione dei contenuti informativi della fonte. Le fonti di benchmark possono essere rappresentate sia da informazioni note su aggregati di dati elementari sia da fonti statistiche (indagini) che permettono anche il confronto puntuale sulle singole unità di analisi delle informazioni. Il processo attuale di ac-

quisizione di un archivio amministrativo - che si avvia allorché un Servizio dell'Istituto ne ha fatto richiesta e ha ottenuto l'assenso da parte degli organi preposti – è riassunto nel seguente prospetto, articolato in tre fasi, ciascuna consistente in operazioni di verifica e/o di elaborazione.

Prospetto 1 - **Fasi del processo di verifica ed elaborazioni per la trasformazione di un archivio amministrativo in un registro statistico**

	1.	**Analisi dell'archivio amministrativo e dei rapporti tra l'istituto e l'ente amministrativo detentore dell'archivio**
Pertinenza	1.1	Rilevanza dell'archivio
	1.2	Accordi tra l'istituto e l'ente amministrativo
Accuratezza	1.3	Controlli di qualità effettuati alla fonte
Puntualità	1.4	Puntualità del dato in riferimento agli obiettivi dell'indagine
Accessibilità / Tempestività	1.5	Aspetti sulla consegna, aggiornamento e messa a regine dell'archivio
Comparabilità	1.6	Mantenimento della qualità nel dato
	2.	**Analisi dei metadati dell'archivio**
Pertinenza	2.1	Rilevanza della popolazione di riferimento
Accessibilità	2.2	Esistenza di codici identificativi univoci
Coerenza	2.3	Corrispondenza delle unità (concetti statistici e amministrativi) e delle definizioni (definizioni, classificazioni)
Comparabilità	2.4	Mantenimento della qualità del metadato nel tempo
	3.	**Validazione delle variabili dell'archivio**
Pertinenza	3.1	Numero di variabili da validare
Coerenza	3.2	Tipo di variabili / esistenza di benchmark
Accuratezza	3.3	Copertura della popolazione statistica
Comparabilità	3.4	Comparabilità nel tempo

Nel prospetto 1, a sinistra di ciascuna verifica o elaborazione è indicato il principio di qualità sottostante agli indicatori annessi.

Le operazioni di verifica e/o di elaborazione danno luogo a indicatori di qualità che riassumono i risultati ottenuti e che possono assumere per motivi di praticità tre modalità: valutazione positiva (+), in quanto è stato possibile accertare una buona qualità per l'operazione in esame ed effettuarla nei tempi previsti con le risorse assegnate; valutazione parzialmente positiva (+/-), essendo stato necessario impiegare oneri aggiuntivi di tempi e/o di risorse per concludere l'operazione in esame; valutazione negativa (-), in quanto non è stato possibile portare a termine l'operazione in esame.[2]

Con tale schema è allora possibile tracciare una sintesi, sia pure qualitativa, di tutto il processo di validazione di ogni singolo archivio disponibile facendo ad esempio un semplice conteggio dei segni positivi, negativi e delle valutazioni intermedie per ciascuna delle tre precedenti fasi. Questo approccio ha avuto sviluppi similari in letteratura (Daas, 2009) ed è attualmente oggetto di studi teorici volti a individuare delle rappresentazioni, aventi dimensionalità *ridotta*, che forniscono un quadro complessivo delle relazioni delle variabili amministrative e degli indicatori di qualità annessi (Bernardi, 2011).

Può essere senz'altro utile tracciare una relazione sulla qualità statistica (Cerroni e De Giorgi, 2010), secondo gli standard definiti in ambito internazionale (Eurostat, 2003), e che darà indicazioni più precise sulle potenzialità della fonte, anche se spesso per gli archivi amministrativi non si riesce a dare una precisa valutazione ad alcune caratteristiche della stessa qualità. Così, valutata anche la qualità statistica, se non ci fosse pertinenza dei suoi contenuti, oppure l'accuratezza o la tempistica della fonte non fossero soddisfacenti, non si avrebbe motivo di realizzare le analisi suc-

[2] Non è detto che un eventuale esito negativo di qualcuna delle valutazioni effettuate sia sufficiente a impedire l'utilizzabilità dell'archivio amministrativo a fini statistici: sta al ricercatore che utilizzerà l'archivio il compito di decidere se ciò possa compromettere l'uso dello stesso archivio.

cessive: la fonte, almeno in termini della tempestività e puntualità del dato o dell'attinenza alle analisi e agli usi che se ne vogliono fare, non garantirebbe la stessa qualità di una fonte statistica.

Il processo di valutazione dei contenuti della fonte è la fase che richiede più tempo e i cui risultati saranno validi come linee guida nell'utilizzo delle specifiche informazioni analizzate. L'obiettivo è quello di verificare la completezza e la qualità dei dati al fine di tracciare un giudizio sulla loro utilizzabilità in sostituzione o integrazione di indagini statistiche.

Il processo di valutazione del contenuto informativo deve mirare a individuare le caratteristiche delle informazioni disponibili e a verificare la possibilità di definire collegamenti tra queste e le nomenclature statistiche. Da una parte, nell'archivio amministrativo ci sono informazioni registrate in base a definizioni legate direttamente a leggi (per esempio fiscali), a provvedimenti amministrativi e a norme giuridiche, dall'altra nella fonte statistica le informazioni seguono le definizioni dei regolamenti statistici ufficiali. Le difficoltà maggiori si incontrano nel confrontare e raccordare definizioni di diversa natura e ciò richiede professionalità diverse da quelle dello statistico, con una più approfondita conoscenza delle normative. Deve essere poi considerato che spesso le variabili amministrative costituiscono solo un sottoinsieme delle variabili definite dai regolamenti e la scelta di una o dell'altra fonte amministrativa si basa, quindi, anche sul numero di variabili comuni tra la fonte amministrativa e l'indagine. A questo proposito si ritiene di estrema importanza la collaborazione tra gli esperti dell'Istituto di statistica e gli esperti che definiscono e gestiscono la fonte amministrativa.

Il percorso di valutazione della fonte viene determinato dagli obiettivi associati all'uso della stessa e cioè: 1) pervenire a una serie di informazioni, verificate e documentate di rispondenza alle nomenclature statistiche delle definizioni e classificazioni, di copertura delle stesse fonti rispetto agli universi statistici, di completezza delle informazioni associate alle variabili da utilizzare, di qualità delle fonti con riferimento agli aspetti di coerenza interna dei dati, definita con regole; 2) rendere disponibile un archivio come input di un processo statistico.

1.2.1 Analisi dell'archivio amministrativo e dei rapporti tra l'istituto e l'ente detentore

La prima fase consiste in una valutazione generale dell'archivio amministrativo e inizia verificando i termini dell'accordo formale tra l'Istituto e l'Ente titolare dell'archivio amministrativo, per quanto riguarda la possibilità di definire una collaborazione finalizzata a utilizzare nel migliore modo possibile l'archivio stesso. Si valuta poi la rilevanza che in prospettiva l'archivio amministrativo può assumere per l'ampliamento del panorama informativo dell'Istituto e se collateralmente esso possa portare una riduzione delle indagini correnti (minore carico statistico). È pure di rilievo osservare se vi sia un'accettabile sincronizzazione temporale tra la disponibilità dell'archivio amministrativo e le esigenze conoscitive dell'Istituto e se l'ente amministrativo, detentore dell'archivio, esegua delle operazioni di verifiche sulla qualità del medesimo. Infine, si dovrà accertare se tutte le suddette caratteristiche siano stabili nel tempo. In particolare la comparabilità interviene come principio chiave all'interno di ciascun livello di analisi, poichè riguarda sia i rapporti con l'ente detentore, sia i metadati che i dati. Per i dati amministrativi, soprattutto per quelli sottoposti a vincoli legislativi, spesso questa caratteristica non può essere assicurata per lunghi periodi di tempo.

1.2.2 Analisi dei metadati dell'archivio amministrativo

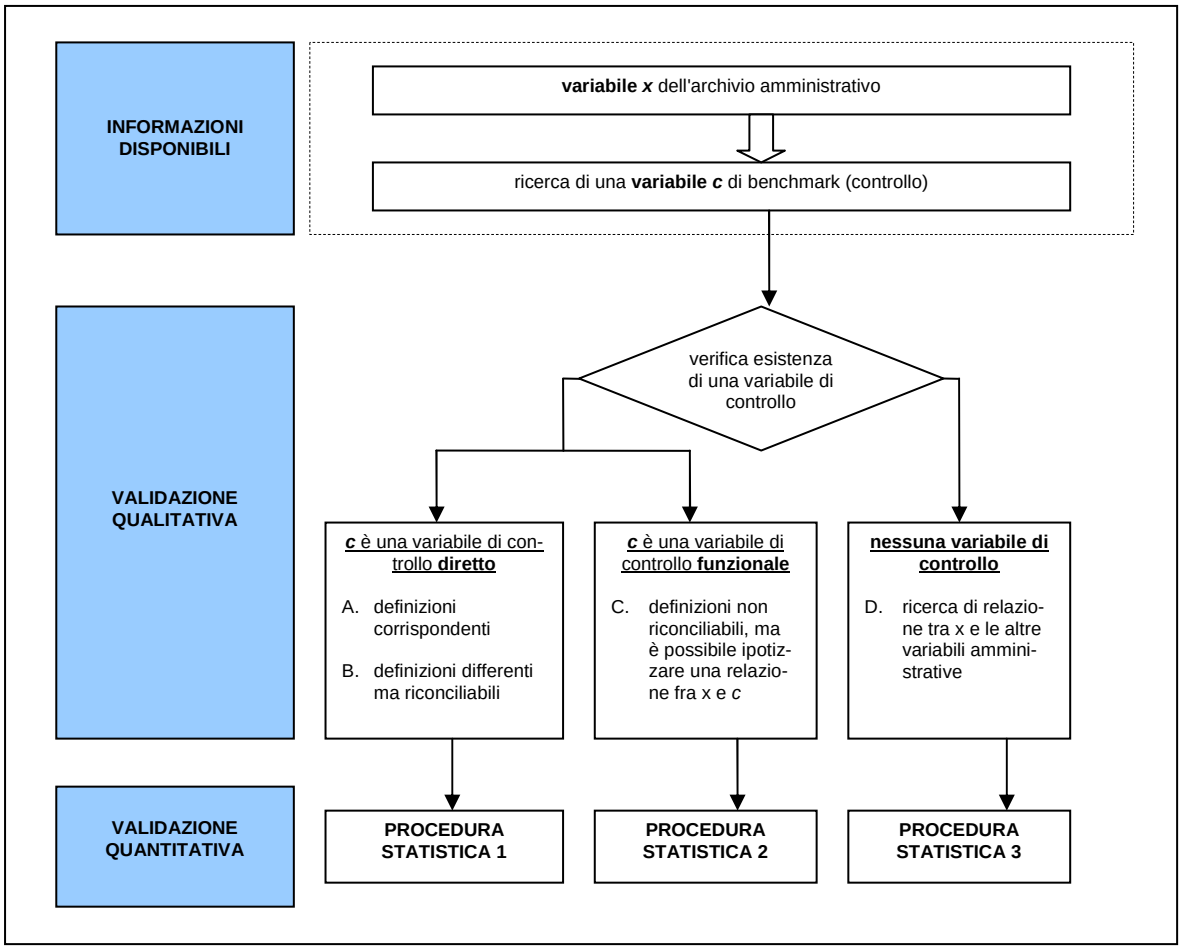
La seconda fase dell'analisi consiste nella verifica dei metadati associati all'archivio amministrativo. Si osserverà quindi se vi sia l'esistenza di chiavi di collegamento identiche a quelle impiegate dall'Istituto e se siano chiare e corrispondenti anche le definizioni delle unità di rilevazione, delle variabili di classificazione e delle variabili di analisi. Poi si verificherà quante variabili possano teoricamente essere sottoposte al processo di validazione rispetto al totale delle variabili dell'archivio e, qualora l'utilizzo dell'archivio è a regime già da alcuni anni, se vi sia una stabilità dei metadati nel tempo. Anche in questa seconda fase è possibile cercare una sintesi del processo di validazione tramite un conteggio associato a ciascuno dei singoli indicatori.

1.2.3 Verifiche sulle variabili dell'archivio amministrativo

La terza fase dell'analisi consiste nell'integrazione di un archivio amministrativo attraverso un complesso di operazioni e di verifiche, riassunte con il termine validazioni, eseguite sulle singole variabili. Si inizia esaminando se la singola variabile abbia, oppure no, una variabile di controllo diretta, vale a dire una variabile proveniente da indagini collaudate e corrispondente sotto il profilo delle definizioni, cioè misurante lo stesso fenomeno della variabile amministrativa sotto studio. Si potranno in tal caso confrontare le misurazioni fornite dalle due variabili per valutare che a definizioni analoghe corrispondano valori compatibili se non addirittura identici. Se la variabile di controllo diretto non esiste, si esplora la possibilità di trovare una relazione con altre variabili, esterne come nel caso precedente all'archivio amministrativo ma comunque "certificate", che dovrebbero essere funzionalmente collegate con la variabile dell'archivio amministrativo in esame, cercando di stimare la relazione in oggetto. Infine, se non esistono variabili di controllo (diretto o funzionale), si valuterà la qualità della variabile amministrativa in esame cercando un collegamento con altre variabili interne all'archivio amministrativo idonee a supportarne un'adeguata interpretazione.

Questo processo può essere schematizzabile come nel successivo prospetto 2.

Prospetto 2 - Fasi del processo di validazione di una variabile di archivio amministrativo



Come emerge dal precedente schema, la validazione della variabile amministrativa in esame, indicata con x, avverrà con metodi dipendenti dall'esistenza o meno di una variabile di controllo, indicata con *c*, e dalla successiva articolazione nei quattro casi distinti indicati nel prospetto 2 con le lettere dalla A alla D.

Il caso A non si verifica spesso, richiedendo una precisa corrispondenza delle definizioni originarie che è assai improbabile, essendo le due indagini (amministrativa e statistica) generate indipendentemente l'una dall'altra.

Nel caso B, più frequente del precedente, ci si trova allorché le definizioni possono essere raccordate. Per esempio, la *x* e la *c* sono aggregati rimodellabili con una diversa combinazione delle componenti elementari vale a dire sia la variabile *x* formata dalla somma di *n+1* variabili mentre la variabile *c* sia formata dalla somma delle prime n variabili di *x*, per cui l'esclusione da *x* della componente in più la potrebbe parificare a *c*.[3]

Nel caso C, presente quando non esiste una variabile di controllo con definizioni corrispondenti o riconducibili alla *x*, poiché emerge la presenza di una variabile *c* connessa alla *x*, si può tentare una validazione di tipo funzionale diversa dalle precedenti.

Nell'ultimo scenario, il caso D, si prende atto che non esiste alcuna variabile di controllo per *x*, né diretta né funzionale, e si avvia l'analisi di qualità per la medesima ponendola in relazione con una o più variabili tratte dal medesimo archivio amministrativo. In tal caso si può parlare del passaggio da un'analisi di coerenza esterna, in cui si confronta la *x* con una variabile *c* che non appartiene all'archivio amministrativo sotto esame, a una di coerenza interna nella quale il confronto e la validazione si attuano con variabili dello stesso archivio.

Va esplicitato allora che il concetto di validazione non è invariante, ma si modifica a seconda della procedura applicata: mentre la validazione nei casi A e B si basa sul confronto puntuale delle osservazioni di *x* e di c, nel caso C la variabile *x* si potrà ritenere validate e quindi integrabile qualora il suo comportamento complessivo rispetto a *c* presenti dinamiche funzionali logiche e coerenti. Nel caso D la validazione sarà ammissibile solo se emergeranno schemi relazionali di *x* con altre variabili dell'archivio amministrativo di tipo coerente e logico.

In base alle situazioni da A a D verificatesi, si avvierà una corrispondente procedura statistica.

Per i casi A e B si avvierà la *procedura statistica 1*, adatta ogni qual volta vi sia una ragionevole presunzione che la *x* corrisponda, eventualmente tramite operazioni di raccordo, alla *c*. Essa consisterà nel visionare le eventuali differenze tra *x* e *c* attraverso una serie di indici di locazione, di scala e di forma distributiva. Nel caso di esito affermativo di corrispondenza di *x* con *c*, i valori di *x* abbinati con *c* forniranno un range di ammissibilità per tutti gli altri valori di *x* che non fosse stato possibile collegare a *c*. È questo il caso avvenuto con le variabili degli Sds e le variabili di controllo tratte dall'indagine sulle piccole e medie imprese e sull'esercizio di arti e professioni (d'ora in poi Pmi). Le prime sono frutto di una rilevazione globale, le seconde sono campionarie, per cui in concreto solo una parte delle imprese Sds si sono abbinate all'indagine campionaria Pmi mentre le altre imprese degli Sds sono ovviamente rimaste non abbinabili. Nel caso di esiti insoddisfacenti nell'instaurare un legame diretto tra *x* e *c*, la supposta variabile di controllo non verrà più trattata come tale e potrebbe, invece, essere impiegata come una variabile collegabile funzionalmente alla prima, ricadendo così nella fattispecie C, qui di seguito esposta.

Per il caso C si attiverà la *procedura statistica 2*, consistente in una serie di analisi esplorative, di *data mining* e regressive, volte a far emergere la relazione funzionale della *x* con la *c*. In questo ambito si dovrebbe giungere alla costruzione di una fascia di confidenza entro cui selezionare i casi di *x* compatibili con *c*, applicabile poi ai casi di *x* eventualmente non abbinati.

Nel caso D si avvierà la *procedura statistica 3*, che prevede analisi multivariate mirate a far emergere i legami tra la *x* e le altre variabili dell'archivio amministrativo, previa individuazione ed esclusione dei dati *outlier*. Nel caso che emergano legami ragionevoli con altre variabili dell'archivio amministrativo si potrà ritenere che la *x* abbia una buona qualità e possa quindi essere integrata.

Occorre infine aggiungere due avvertenze di carattere generale sui criteri seguiti nell'esecuzione delle tre precedenti procedure. Primo, poiché esse sono state applicate a insiemi di dati molto nume-

[3] In tali circostanze si dovrà fare attenzione e verificare se il confronto con le nuove definizioni, modificate in quanto decurtate di una componente, conservi il significato del precedente paragone che lo si voleva basare su tutte le voci, osservando quindi se la componente esclusa da x per trovare una corrispondenza funzionale con c non abbia un peso dominante rispetto a tutti gli altri.

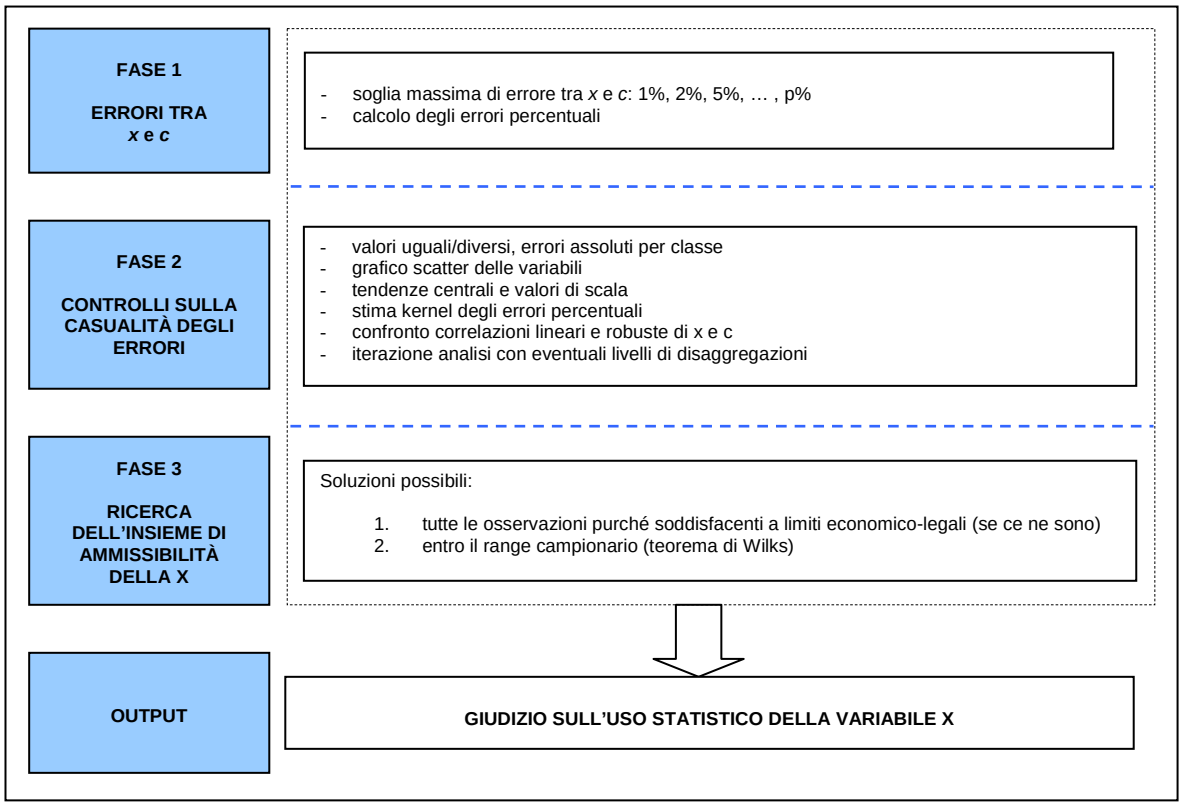
rosi che sono, come noto, di lettura a volte difficile, è parso opportuno in alcune circostanze ripetere l'analisi dei dati con tecniche alternative capaci di fornire interpretazioni complementari e confermative, mentre negli altri casi di più facile lettura sono stati eseguiti solo i passi d'analisi strettamente necessari, tralasciando quindi altri possibili percorsi d'indagine. Secondo, non si può escludere che dalle operazioni eseguite siano emerse situazioni intermedie, nelle quali il riscontro dei dati non permetta cioè di distinguere nettamente quale tra le quattro fattispecie (A, B, C, D) sia quella prevalente.

2. Le procedure statistiche

2.1 La procedura statistica 1

Il procedimento statistico in esame si avvia qualora si suppone di essere nei casi A o B prima descritti e si articola nei passi seguenti:

Prospetto 3 - Fasi della procedura statistica 1 (schema 1)



Qui si suppone, come da casi A o B, che le variabili *x* e *c* abbiano definizioni corrispondenti o riconducibili. Si inizia calcolando, sul sottoinsieme di *x* abbinato con *c*, il numero di osservazioni eguali e differenti, che già fornisce una prima importante valutazione. Poi si ipotizza ammissibile uno scarto di *x* rispetto a *c* massimo per esempio del $\pm p\%$, dove p può dipendere da diversi fattori (tipo di variabili, tipo d'archivio, finalità dell'integrazione dei dati, ecc.). Si fissa una griglia di scarti con classi $\pm 1\%$, $\pm 2\%$, $\pm 2\%$, ... ,$\pm p\%$, oltre $\pm p\%$, e si calcolano le frequenze relative. Questo permette di esaminare sia le frequenze ricadenti entro tali classi sia di selezionare i casi non rispettanti il vincolo $\left|\frac{x-c}{c}\right| \cdot 100 < p\%$ verificando che essi rappresentino una quota minimale e che abbiano una qualche caratteristica che li faccia apparire come casuali.

Infatti, se è vera l'ipotesi di corrispondenza delle definizioni, in linea generale si dovrebbe osservare che:

1. le osservazioni con scarti che ne suggeriscano l'esclusione saranno una minoranza, con netta prevalenza di quelle caratterizzate da scarti piccoli sulle altre;[4]
2. uno *scatter* tra *x* e *c*, non sempre tuttavia nitidamente osservabile a causa dell'elevato numero di punti da rappresentare, dovrebbe presentare una nuvola di punti non lontani da una retta a 45° gradi parimenti oscillanti sopra e sotto la stessa;
3. nel complesso le *misure di locazione e di scala* di *x* e *c* dovrebbero essere molto simili;
4. l'analisi della correlazione lineare tra x e c, comparandola con l'analoga ottenuta con il coefficiente di correlazione di Spearman, dovrebbe avere valori vicini a 1;
5. la distribuzione degli scarti relativi $\left(\frac{x-c}{c}\right)$ si dovrebbe caratterizzare per una *stima kernel*[5] unimodale, simmetrica attorno allo zero e con marcata curtosi. I precedenti passi possono, se necessario, essere iterati in base a variabili di classificazione, quali l'area geografica, gruppi di attività economica, ecc..

Nel caso emerga un'effettiva corrispondenza quantitativa tra *c* e x, per cui il sottoinsieme di dati di x abbinati ed esaminati risultino validati, si può pensare di gestire il secondo sottoinsieme dei dati di x, la parte non collegabile alla c, in uno dei due seguenti modi.

Primo, si possono ritenere validati tutti i dati di tale sottoinsieme se esistono le seguenti circostanze:

a) che la fonte amministrativa sia molto autorevole;[6]
b) che l'esito positivo delle verifiche osservato sul primo sottoinsieme sia un indice di buona qualità complessiva dei dati;
c) che la parte non abbinata presenti indici di locazione e di scala simili alla parte abbinata;
d) che siano soddisfatti eventuali limiti economico-legali.[7.]

Secondo, si può trovare un *range* ammissibile per i valori della *x* non abbinabili con *c* dal confronto fatto sulla parte di *x* collegata e, poi, usare tale soglia per selezionare i casi della parte di *x* non abbinata. Quest'ultimo sviluppo può essere fatto grazie a un teorema di Wilks secondo cui, a prescindere dalla forma distributiva della popolazione e in presenza di numerosità elevata delle osservazioni, il campo di variazione, definito dal massimo meno il minimo campionario, stima in modo consistente il corrispondente campo di variazione della popolazione, per cui al crescere della numerosità campionaria tende a uno la probabilità che i valori del suddetto *range* campionario corrispondano a quelli dell'intera popolazione (Zuliani, 1964). In altri termini, poiché le osservazioni degli Sds si abbinano con tutte le corrispondenti di Pmi, ed essendo questa un'indagine campionaria ne consegue che anche il sottoinsieme degli Sds abbinati con Pmi può essere considerato un ulteriore campione casuale delle osservazioni delle piccole e medie imprese. Ne segue, allora, che il campo di variazione individuato sui dati abbinati può essere usato per selezionare i dati ammissibili nella restante parte degli Sds non abbinati a Pmi.[8]

---

[4] In subordine al punto 1, può essere ancora accettabile l'ipotesi di corrispondenza delle definizioni quando le osservazioni abbiano una quota non piccola di casi differenti tra *x* e *c* ma la misura degli scarti per tali casi sia in gran parte vicino allo zero. Rispetto quindi alla situazione ideale inizialmente descritta, in cui le osservazioni con scarti non nulli sono una piccola minoranza anche se alcuni degli scarti possono essere elevati, si può pure osservare che le osservazioni con scarti non nulli sono in numero relativamente maggiore, ma praticamente con tutti gli scarti di livello trascurabile (per esempio sotto l'aspetto economico). In quest'ultimo caso potrebbe essere utile anche qualificare ulteriormente gli scarti, come per esempio tentare un adattamento a una curva normale (con media nulla e varianza da calcolare), anche se va tenuto presente che, come noto, in presenza d'elevata numerosità i test di adattamento ricevono una potenza eccessiva, cioè in pratica inducono a rifiutare l'ipotesi H_0 di normalità anche in presenza di piccole discrepanze.

[5] La stima *kernel* è una rappresentazione perequata (*smoothed*) dell'istogramma che rende più facile verificarne le caratteristiche principali quali l'uni/pluri/modalità, la simmetria, l'esistenza di gruppi, eccetera.

[6] Ogni fonte amministrativa è ovviamente autorevole alla stessa stregue di tutte le altre, ma nella pratica si osservano fonti maggiormente coerenti e compatibili rispetto ad altre.

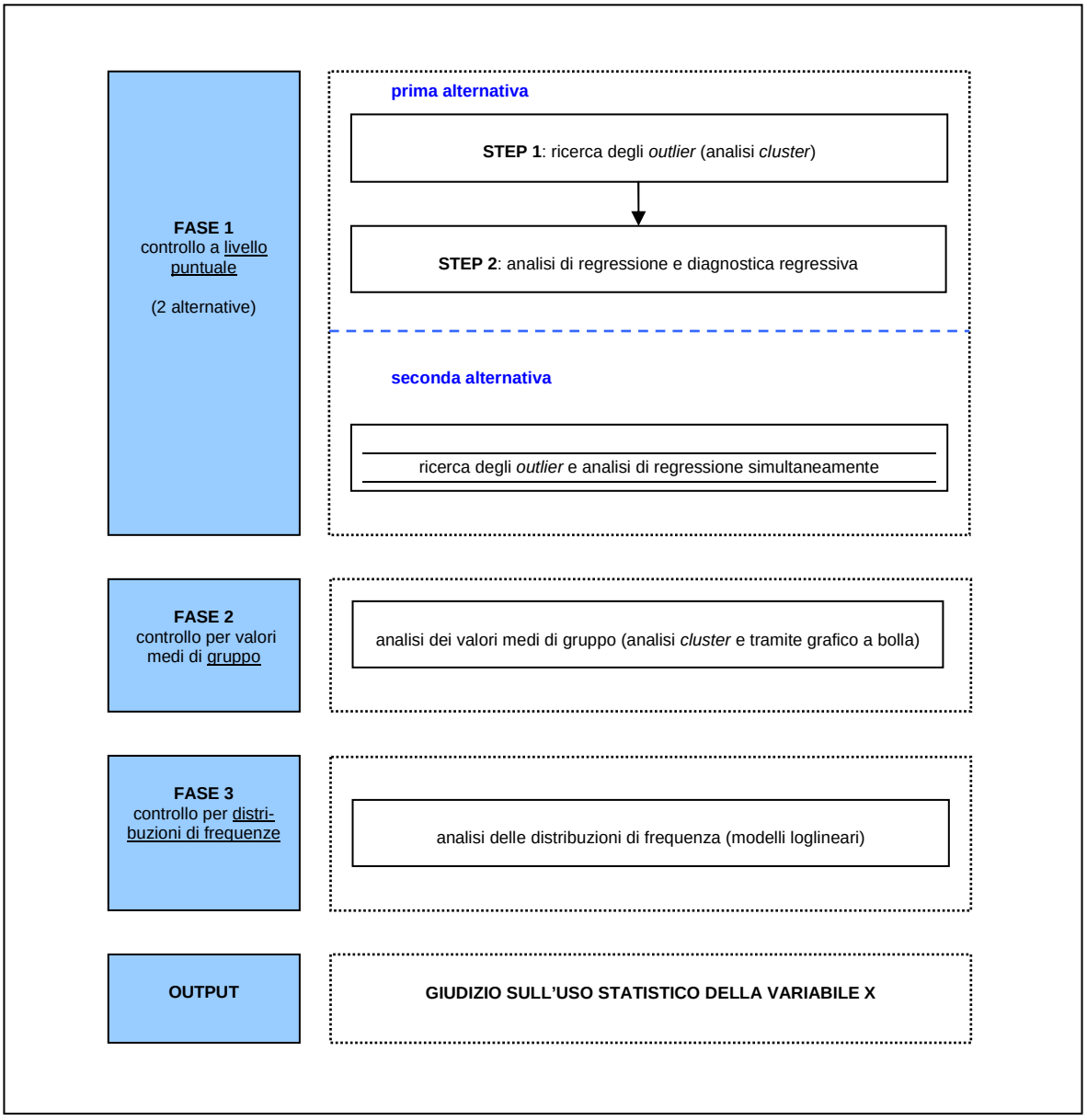
[7] Per esempio si selezionano solo le unità con costi di produzione maggiori o uguali a zero, oppure si selezionano solo le imprese il cui fatturato non ecceda i limiti di ammissibilità degli Sds.

[8] Si suppone inoltre, cautelativamente, che i due sottoinsiemi in cui viene scissa la *x* abbiano comunque simili i principali valori di locazione e di scala.

2.2 La procedura statistica 2

Il procedimento statistico in esame è da avviare qualora si supponga di essere nel caso *c* (Prospetto 2) e si articola nei passi indicati qui di seguito.

Prospetto 4: Fasi della procedura statistica 2 (schema 2)



Lo schema presenta diversi possibili modi, qui definiti fasi, di analizzare la relazione tra *x* e *c*. Si osserva che la fase 1, che ha l'obiettivo di eseguire tale analisi a livello puntuale, può essere condotta seguendo due alternative di lavoro basate su modelli regressivi descritti nel paragrafo successivo. Qualora la fase 1 fornisse risultati non chiari, si potrà passare alle fasi 2 o 3. La fase 2, che è consigliata in letteratura in caso di grandi quantità di dati, cerca una rappresentazione della relazione di *x* con *c* tramite una sintesi grafica basata sulla tendenza delle medie dei gruppi di *x* e di *c* individuate dall'analisi *cluster*. La fase 3, infine, riduce ulteriormente la pretesa conoscitiva sul legame tra *x* e *c* e opera non più sui dati delle due variabili ma sulle loro distribuzioni di frequenza in tavola quadrata, e verifica con modelli loglineari se sia presente una chiara associazione.

2.2.1 Fase 1, controllo a livello puntuale

Questa fase prevede due alternative di lavoro. La prima alternativa opera in due passi nel primo dei quali si stima la relazione tra *x* e *c* dopo aver escluso gli *outlier* che potrebbero oscurarla, mentre la seconda alternativa svolge lo stesso compito della prima ma in un unico passo. La prima alternativa appare quindi meno soddisfacente e sarà impiegata solo nel caso di inapplicabilità della seconda.

2.2.1.1 Prima alternativa

La prima alternativa viene eseguita in due passi distinti (step 1 e step 2). Si inizia con la ricerca dei dati anomali e dati estremi, distinguendo con i primi i dati che emergono come estranei, cioè i dati che non dovrebbero essere inclusi nell'insieme in esame, e con i secondi, comunemente definiti *outlier*, i dati che appartengono all'insieme di definizione ma sono, dal punto di vista numerico, molto distanti dagli altri.

Circa il primo aspetto (dati anomali, che quindi hanno problemi legati alla natura dei dati), la loro individuazione può essere eseguita verificando il rispetto dei criteri legali che definiscono le unità sottoposte alla rilevazione. Per esempio, se in un archivio relativo alle imprese in forma giuridica societaria fosse presente un'impresa di persone, essa sarebbe un dato anomalo; oppure, se negli Sds apparisse un'impresa con fatturato eccedente il limite di rilevazione imposto per legge - attualmente € 7.500.000 - essa rappresenterebbe, nuovamente, un dato anomalo.

Circa il secondo aspetto (dati estremi, quindi con problemi legati alla struttura dei dati), esso può essere studiato con una distribuzione grafica delle *x* e delle *c* osservando se esistano punti separati nettamente dagli altri. Qualora però il grafico non sia leggibile chiaramente, come spesso avviene in presenza di scatter con elevate numerosità, si può ricorrere a una procedura di analisi dei gruppi che si basa sull'idea che se i dati estremi devono essere distinti dagli altri essi devono di conseguenza essere allocati in *cluster* isolati, cioè distanti dagli altri, con pochi elementi ciascuno, quindi a bassa frequenza, e di dimensioni relativamente contenute, e cioè con raggio limitato. Per tali finalità analitiche, l'analisi dei gruppi verrà svolta con quello tra i due seguenti metodi che fornirà i risultati più chiari:

- A. procedura di raggruppamento k-medie in due stadi (SAS, 2009);
- B. procedura di raggruppamento *cluster single-linkage* (Everitt et al., 2001).

Il metodo A è eseguito in due stadi. Nel primo stadio si esegue un'assegnazione delle unità ai gruppi con una sola iterazione e si rilevano i *cluster* isolati per mezzo della loro distanza dal *cluster* più vicino, che dovrà essere elevata, della loro frequenza, che dovrà essere bassa, e della loro ampiezza, che dovrà essere limitata. Per conferma della bontà del procedimento si calcola sul dataset avente come righe i *cluster* rilevati e come colonne le tre suddette variabili (distanza, frequenza e ampiezza) la matrice delle correlazioni, che dovrà mostrare con chiarezza una correlazione positiva tra frequenza e ampiezza (un *cluster* più ha elementi più è ampio), una correlazione negativa tra distanza e ampiezza (se l'ampiezza sale si riducono le distanze tra *cluster*) e sempre una correlazione negativa tra distanza e frequenza (se la frequenza sale i *cluster* tendono a dilatarsi e si avvicinano). Nel secondo stadio si effettua un'analisi k-medie ove però vengono forniti come centroidi iniziali quelli non appartenenti ai *cluster* isolati prima trovati e imponendo che i *cluster* rilevati non abbiano la loro ampiezza (raggio) superiore a un valore opportunamente desunto dalla fase precedente. Si osserva che grazie al primo accorgimento -l'assegnazione di centroidi regolari come avvio del procedimento di *cluster*ing - si converge meglio verso l'individuazione dei *cluster* che si potrebbero definire regolari (nel senso di non contenenti dati estremi). Con il secondo accorgimento - il vincolo sull'ampiezza massima ammissibile per un *cluster* - si evita invece che i *cluster* isolati, e che tali devono restare, vengano accidentalmente ricompresi in un *cluster* regolare.

Il metodo B è basato sul raggruppamento con il criterio del "vicino più vicino" e, a differenza del metodo precedente, anziché ricorrere alla procedura iterata cerca di raggiungere gli stessi scopi nel seguente modo: i dati vengono congiunti in *cluster* scegliendo come misura dei legami tra due gruppi la distanza intercorrente tra i loro due punti interni più vicini (*nearest neighbour*), in modo che il procedimento termini relegando i punti isolati in *cluster* distanti dagli altri. Tale procedura

sembra avere il pregio di evitare sia i due stadi della k-medie sia l'impostazione del raggio del *cluster*, che è comunque un'operazione non sempre semplice, ma può essere di difficile lettura la sua evidenza grafica specie se si tratta di affrontare dataset con alta numerosità. Le osservazioni segnalate da entrambe le procedure vengono considerate estreme e scartate dalle analisi successive.

Con queste due procedure si conclude la prima parte della prima alternativa (step 1), cioè l'identificazione dei dati estremi. Essi pertanto faranno parte di *cluster* isolati e separati dagli altri.

Si passa poi a stimare la relazione tra i valori di *x* abbinati e i corrispondenti di *c* tramite regressione della *x* sulla *c* e successivo esame della coerenza del coefficiente della variabile *c* in segno, valore e significatività.

Si applica infine alla relazione stimata una diagnostica regressiva, che può prevedere alcuni dei seguenti esami (Cameron and Trivedi, 2009, p. 91-97):

1. *residual diagnostic plot*, in cui il grafico sintetizza i risultati del modello regressivo riportando i valori adattati sulle ascisse e i valori residui sulle ordinate; va ricordato che però distribuzioni di questo genere sono di difficile lettura in presenza di un'elevata numerosità di punti;
2. *verifica di eventuali osservazioni influenzali* di *x*, dove si tratterà quindi di esaminare il *leverage* della variabile dipendente *x* a livello puntuale;
3. *test sulla forma funzionale assegnata alla relazione* tra *x* e *c* attraverso la trasformazione di Box-Cox;
4. *test di specificazione di variabili omesse*, adattando modelli con potenze di *x* e alcune variabili dummy per considerare particolari attività economica, aree geografiche, ecc. che vengono sottoposte al test di Wald;
5. *test per verificare la forma funzionale della media condizionale*, vale a dire il *Ramsey reset test* che si considera complementare al test delle variabili omesse di Wald;
6. *test di eteroschedasticità, cioè* si esegue il test di Cook-Weisberg per verificare eventuali modifiche negli intervalli di confidenza dei coefficienti;
7. *test omnibus*, cioè il *test information matrix* (IM) che esamina congiuntamente la simmetria, la curtosi e l'eteroschedasticità degli errori.

Con la precedente diagnostica regressiva si concludono le operazioni dell'alternativa 1. Supponendo che i casi di x inizialmente esclusi siano relativamente pochi, si potrà o ritenere validati anche i casi di x non abbinati purché rispettino le condizioni a-d di cui alla procedura 1 oppure procedere a definire una fascia di confidenza per i valori di *x* abbinati con cui poi selezionare i valori di x appartenenti all'insieme non abbinato.

2.2.1.2 Seconda alternativa

A differenza dell'alternativa 1 descritta nel paragrafo precedente, qui si stima la relazione tra *x* e *c* congiuntamente all'individuazione dei dati anomali. Si inizia con un'ottica esplorativa, eseguendo una *regressione m-band* e/o una *regressione lowess* per valutare la forma che potrebbe avere la relazione tra *x* e *c*. Poi si stima la relazione usando una regressione quantile e/o a una *regressione robusta di tipo M*.

La *regressione m-band* è un modello di regressione non parametrica nel senso che non produce un'equazione regressiva esplicita. È soprattutto uno strumento molto utile a scopi esplorativi che consente di esplorare i dati per investigare possibili forme della relazione non lineare tra *x* e *c* con il vantaggio di non richiedere al ricercatore di specificare una relazione funzionale in anticipo. In breve, ordinato in senso crescente il dataset in base al regressore (qui la c), si specifica in quanti sottoinsiemi si vuole frazionare il medesimo per esempio, se si hanno 1.000 punti coordinati e si specifica 10, questo equivale a prendere i primi 100 punti coordinati, poi i punti da 101 a 200, …, infine i punti da 901 a 1.000) e per ciascuno di questi 10 sottoinsiemi si calcola la mediana (indice di locazione robusto) di *c* e di *x* e riportando in un diagramma a punti le 10 coppie coordinate (x_1 e c_1, x_2 e c_2, …, x_{10} e c_{10}) unite da una spezzata e si osserva la spezzata contestualmente con i 1.000 punti originari.

La regressione *lowess* (*LOcally WEigthed Scatterplot Smoothing*) è una variazione della comune regressione OLS del vettore *x* sul vettore *c* in quanto sostituisce nel primo vettore alla singola osserva-

zione x_i una media dei valori di un dato numero di osservazioni precedenti e seguenti la x_i ponderandole con pesi crescenti al crescere dell'errore annesso (valore osservato meno valore adattato). Anche in questo caso si tratta di una regressione non parametrica che, come la precedente m-band, non stima i parametri della relazione ma fornisce un'idea della tendenza della relazione a livello locale.

La *regressione quantile* è una regressione di tipo robusto che ha diversi aspetti interessanti, tra i quali si ricorda che è semiparametrica - nel senso che produce stime dei coefficienti e non richiede ipotesi sulla distribuzione parametrica degli errori salvo che siano i.i.d. (incorrelati e identicamente distribuiti) e che consente di verificare l'adattamento tra *x* e *c* non solo sull'intera distribuzione ma anche su specifici quantili della variabile dipendente (Cameron e Trivedi, 2009, p. 206), permettendo così di monitorare i tratti in cui il legame tra *x* e *c* si deteriora.

La *regressione robusta di tipo M* è anch'essa semiparametrica e può risultare ancora più adatta alle finalità qui in esame essendo ritenuta maggiormente resistente ai dati estremi. Nel contesto in oggetto la regressione viene impostata con la variabile *c* supposta esente da errori, in quanto variabile di controllo, mentre errori possono invece essere presenti sulla variabile x. Si inizia con una regressione OLS applicata su tutte le osservazioni aventi la statistica D di Cook minore di 1, essendo ritenute tutte le altre osservazioni troppo influenti e quindi escluse. Poi sono calcolati dei pesi per ciascuna osservazione della *x* tramite una finestra di Huber (dove più un'osservazione è male adattata più riceve pesi piccoli, e viceversa) e si calcola la conseguente regressione ponderata iterata, cui segue una nuova finestra di pesi fino a ottenere stime convergenti che nell'iterazione finale sono in grado di fornire un'efficienza delle stime dei coefficienti pari al 95% degli OLS. I risultati emersi in questa procedura possono poi essere posti a confronto con quelli osservati durante l'esecuzione della procedura 1, poiché contenute differenze nei parametri stimati implicano che la presenza di dati estremi è modesta (Hamilton, 2009, p. 256-257).

2.2.2 Fase 2, controllo per valori medi di gruppo

Nell'ipotesi che la *x* e la *c* non mostrino alcun chiaro collegamento tramite le tecniche descritte in fase 1, si può pensare di valutare se sia possibile rinvenire una qualche semplice relazione a livello di medie di gruppo. A tal fine si ricerca una sintesi dei dati con la procedura k-medie che in letteratura è suggerita soprattutto in presenza di grosse moli di dati da trattare (Agresti, 2002, p. 180; Khattre e Naik , 2000, p. 300). Si applica pertanto tale procedura alle variabili *x* e *c* ottenendo una loro rappresentazione per esempio in m gruppi, ciascuno contenente n_i osservazioni. Per ciascun gruppo si calcola la media delle osservazioni incluse di *x* e di *c*, ottenendo in tal modo m coppie di punti (x_1,c_1), (x_2,c_2),…, (x_m,c_m), definiti centroidi, e poi si esamina tramite un grafico a bolle[9] se tali centroidi si allineino per esempio su una retta. Qualora così emergesse si avrebbe l'informazione che le variabili *x* e *c*, pur non avendo un legame ravvicinato in una ottica puntuale, sono tuttavia compatibili in termini dei loro valori medi.

Per la validazione della parte non abbinata si può procedere come per la fase 1 e cioè si possono ritenere validati tutti i casi (purché siano rispettate le condizioni a-d della procedura 1) oppure si può approfondire ulteriormente la questione e verificare che la parte non abbinata presenti raggruppamenti simili a quella abbinata.

2.2.3 Fase 3, controllo per distribuzioni di frequenza

Un ulteriore modo di ricercare una relazione tra *x* e *c*, che si può attivare qualora non sia emerso dai precedenti tentativi un'idea chiara sul loro collegamento, è quello di procedere a un'analisi dei dati sintetizzandoli in frequenze, tramite discretizzazione delle variabili. Tale metodo, consigliato in letteratura (Basilevsky, 1993), si basa sull'idea che gli errori di misura presenti nei dati abbiano minori effetti se dai dati elementari (serie statistiche) si passa all'analisi della distribuzioni di frequenza da essi ottenibili (seriazioni statistiche). Viene individuato, a questo scopo, il numero ottimale delle

[9] Con aree proporzionali alla numerosità dei gruppi.

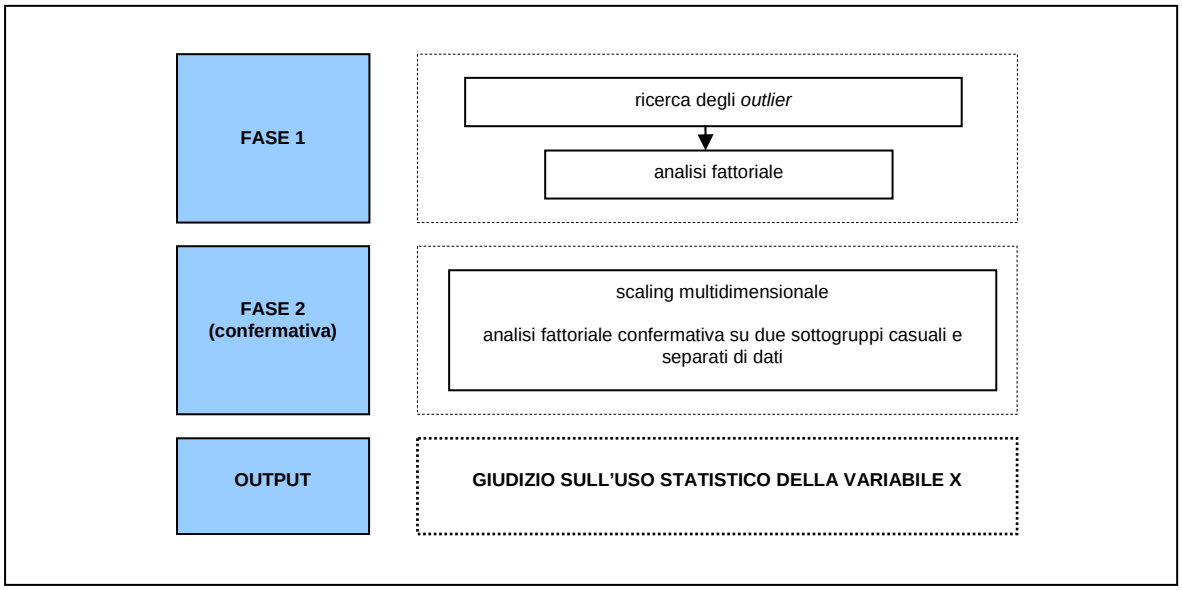
classi calcolato sulla variabile *x* con un criterio automatico (Piccolo, 2000) o con il criterio delle classi equifrequenti a griglia fine, non mascheranti, cioè, eventuali dettagli della distribuzione (classi contenenti ognuna per esempio il 5% dei casi) e viene verificata la concordanza suggerita dai due criteri. Gli estremi delle classi così individuate sulla *c* vengono applicati alla variabile *x*. Viene poi effettuata l'analisi esplorativa della tabella a doppia entrata definita dalle classi di *x* e di *c*, tramite calcolo del peso dei casi collocati nella diagonale principale e nelle parti triangolari basse e alte sul totale dei casi, seguito dal calcolo del coefficiente *K* di Cohen.[10] All'analisi esplorativa può seguire infine un ulteriore approfondimento con la stima di specifici modelli log-lineari quali:

1. il *modello QIM* (*Quasi Indipendence Model*) in cui si ipotizza che le frequenze siano distribuite casualmente nelle celle della tavola di contingenza al di fuori della diagonale principale (Allison, 1999, p. 44). Se il modello passa la fase delle verifiche (Agresti, 2002; Cameron e Trivedi, 2009) si può concludere che i dati si abbinano soprattutto sulla diagonale principale e che pertanto non c'è altra dinamica presente all'interno della tavola di contingenza;
2. il *modello UL*, che verifica se è vero che l'associazione tra *x* e *c* diminuisca all'allontanarsi dalla diagonale principale.

2.3 La procedura statistica 3

La procedura in esame è alla base del processo di analisi e validazione di una variabile quantitativa in assenza di una variabile *c* di controllo, e si articola nei seguenti passi:

Prospetto 5 - Fasi della procedura statistica 3 (schema 3)



In generale, si osserva che nella presente situazione, venendo a mancare per *x* una variabile di controllo con cui effettuare la valutazione della sua qualità, l'unica strada percorribile appare quella di relazionare la *x* alle altre variabili dello stesso archivio. Si avrà quindi in questo caso una validazione che, in caso di esito positivo, informerà che la *x* è ragionevolmente collegata ad altre variabili interne dell'archivio amministrativo e possiede quindi il requisito della coerenza interna. La preventiva detenzione dei casi estremi consisterà nel segnalare come *outlier* i punti con alti punteggi sulle componenti principali mentre sui dati non estremi inclusi nelle analisi successive si potrà cercare di ricondurre i legami emersi ai concetti di complementarietà e di succedaneità.

[10] Introdotto nel 1960, il coefficiente kappa esprime una misura del grado di concordanza tra due variabili di una tabella quadrata, e raffronta gli elementi concordanti collocati sulla diagonale principale e gli elementi concordanti che si avrebbero in caso di indipendenza delle variabili stesse.

In particolare, lo schema in esame si articola in 4 fasi:

1. ricerca delle variabili teoricamente collegabili;
2. verifica e individuazione della presenza di dati estremi;
3. analisi fattoriale;
4. un'analisi di conferma se necessario con il *multidimensional scaling*;
   - in alternativa al multidimensional scaling in caso di elevate numerosità di dati, un'analisi fattoriale su due partizioni di dati.

Si inizia selezionando sulla base di criteri teorici le variabili potenzialmente collegate, perseguendo il criterio di considerare un numero di variabili di partenza il più ampio possibile.

Dopo una analisi preliminare svolta confrontando le statistiche di scarto quadratico medio (sqm) e la mediana degli scarti delle differenze in valore assoluto dalla mediana (mad) si trova un primo riscontro alla presenza di dati estremi.

Poi si prosegue con la ricerca di dati estremi attraverso il metodo basato su delle elaborazioni delle componenti principali (le statistiche Hisq e Disq più avanti illustrate) e poichè l'elevata numerosità non consente di leggere agevolmente tali statistiche, esse vengono esaminate attraverso una sintesi grafica ovvero con un grafico box plot. Si concluderà questa prima fase valutando quante siano le osservazioni selezionate e se esse presentino schemi non casuali.

Si passa, quindi, all'analisi fattoriale la quale, come noto, permette di trovare una correlazione a livello di gruppi di variabili, rappresentando quindi l'estensione multivariata dell'analisi delle correlazioni, ove invece le variabili vengono confrontate a coppie. Come noto, con l'analisi fattoriale si trovano (se esistono) le variabili latenti e cioè le variabili che attirano attorno a loro gruppi di altre variabili, fornendo così una spiegazione sul perché le variabili incluse nel gruppo in esame siano tra loro correlate. L'analisi, eseguita almeno inizialmente con le opzioni standard più diffuse (criteri di selezione degli autovalori di Kaiser, rotazioni *varimax*, analisi della matrice residua, ecc.) si conclude esaminando i collegamenti delle variabili attirate con la variabile latente attirante e quindi anche con un'interpretazione di quest'ultima. Nel caso di un archivio amministrativo, potrebbe essere possibile per esempio riscontrare che un gruppo di k variabili gravita su una precisa variabile latente, che le k variabili sono legate da correlazioni positive e che l'interpretazione fornita della variabile latente consenta di spiegare ulteriormente perché le variabili sono tra loro correlate positivamente (relazione di complementarietà). Si può ipotizzare di inserire nell'analisi fattoriale alcuni recenti sviluppi teorici, quali l'applicazione con variabili indicatrici e con forme d'interazione tra le variabili.[11]

Un'applicazione del moderno *multidimensional scaling* permette di confermare (o di mettere in dubbio) quanto emerso con la procedura fattoriale e, quindi, nel caso di concordanza dei risultati è possibile affermare che le variabili dell'archivio amministrativo esaminato hanno superato una verifica di coerenza interna e in tal senso possono essere dichiarate validate.

In caso di elevata numerosità dei casi una soluzione alternativa al *multidimensional scaling* è partizionare in modo casuale i dati in due sottoinsiemi di uguale numerosità ed eseguire una analisi fattoriale di tipo esplorativo sul primo sottoinsieme e un'analisi di tipo confirmatorio sul secondo sottoinsieme.

[11] Sono disponibili sul software Stata nella versione 11 del luglio 2009.

3. Applicazione agli Studi di settore

3.1 Descrizione delle fonti e analisi dei metadati

3.1.1 La rilevazione sulle piccole e medie imprese e sull'esercizio di arti e professioni

La rilevazione Pmi ha come campo di osservazione le imprese con 1-99 addetti appartenenti ai settori di attività economica industriale, commerciale e dei servizi alle imprese ed alle famiglie e risponde alle esigenze richieste dal regolamento comunitario sulle statistiche strutturali (SBS).

La rilevazione è campionaria e le informazioni sono raccolte attraverso la compilazione di un questionario molto complesso utile a soddisfare sia il regolamento strutturale sulle imprese SBS sia le esigenze della Contabilità Nazionale. Il questionario si basa su 7-9 pagine da compilare, di cui le prime quattro rappresentano le informazioni base della rilevazione Pmi e richiedono la compilazione di dati quantitativi per il conto economico, l'occupazione, il costo del lavoro del personale dipendente, il personale esterno alle imprese con i relativi costi, gli investimenti effettuati nell'esercizio per categoria di beni, le spese di protezione dell'ambiente legate all'introduzione di tecnologie pulite, ed altre variabili economiche. Le altre pagine rappresentano, invece, informazioni multiscopo di natura prettamente qualitativa finalizzate a rilevare particolari aspetti emergenti dell'attività d'impresa.

Il disegno di campionamento utilizzato è di tipo casuale stratificato e la lista delle unità campionate è estratta dall'archivio Archivio statistico delle imprese attive (Asia), costruito sulla base dell'integrazione di varie fonti, di carattere sia amministrativo sia statistico. Il disegno di campionamento adottato è a uno stadio stratificato, con selezione delle unità con probabilità uguali; gli strati sono definiti dalla concatenazione delle modalità delle variabili *Regione*, *Attività economica* e *Classe di addetti*. Per maggiori dettagli sulla metodologia dell'indagine si può consultare Istat, *Conti economici delle imprese. Anno 2003*. Per gli ultimi dati disponibili si può fare riferimento a Istat, *Struttura e Competitività del Sistema delle Imprese Industriali e dei Servizi. Anno 2008*.

3.1.2 L'archivio amministrativo degli Studi di settore

Gli Sds sono uno strumento che l'Agenzia delle entrate utilizza per rilevare i parametri fondamentali di liberi professionisti, lavoratori autonomi e imprese, sotto l'aspetto sia strutturale sia economico, al fine di valutare la loro capacità di generare ricavi. La disciplina degli Sds è stata introdotta nel 1993 con D.L. 30 agosto 1993, n. 331, convertito dalla legge 29 ottobre 1993 n. 427). L'obiettivo della procedura è quello di stimare il ricavo presunto del singolo contribuente come funzione delle variabili strutturali (che sono un mix di variabili di scala e qualitative, di input e output, di capitale e di processo) da porre a confronto con quello dichiarato. Attraverso una procedura di *cluster*ing l'Agenzia delle entrate individua i gruppi di contribuenti omogenei rispetto alle variabili strutturali rilevate e mediante l'analisi discriminante associa ciascun contribuente a uno o più *cluster* con determinate probabilità di appartenenza.

Gli Sds rappresentano una delle fonti amministrative di maggiore interesse per le analisi statistiche, soprattutto per la fascia delle piccole e medie imprese che svolgono attività economiche nel campo industriale, commerciale e dei servizi e dei professionisti: sono, infatti, obbligati a compilare tale modulistica tutti i soggetti che realizzano ricavi non superiori alla soglia di 7.500.000 euro.

I questionari sugli Sds sono specifici per ciascuna attività economica esercitata e sono suddivisi nelle seguenti sezioni:

- Quadro A – Personale addetto all'attività
- Quadro B – Unità locale destinata all'esercizio dell'attività
- Quadro C – Modalità di svolgimento dell'attività
- Quadro D – Elementi specifici dell'attività
- Quadro E – Beni strumentali
- Quadro F – Elementi contabili (imprese)
- Quadro G – Elementi contabili (professionisti)
- Quadro X – Altre informazioni rilevanti ai fini dell'applicazione degli Studi di settore
- Quadro Z – Dati complementari.

I quadri C e D sono tipici del settore manifatturiero.

Il quadro F, e in misura minore il quadro G, richiedono la compilazione di dati quantitativi contabili comparabili e/o raccordabili con quelli richiesti dalla IV direttiva comunitaria per il conto economico, con dettagli articolati per le diverse voci.

Per quanto riguarda gli usi correnti in Istat delle informazioni contenute negli Sds, essi vengono già utilizzati per l'attribuzione dell'attività economica prevalente e per l'imputazione della variabile Volume di affari delle imprese di Asia. Essi sono inoltre rilevanti per le stime degli aggregati di Contabilità Nazionale, per le quali è in corso una sperimentazione di utilizzo effettivo, per l'analisi degli aspetti economici delle piccole e medie imprese italiane nel Rapporto annuale dell'Istat (Istat, 2009; Istat, 2010) e per l'integrazione dei dati del campione dell'indagine Pmi.

Escludendo le cause di inapplicabilità e di esclusione, gli Sds riflettono in modo accurato la popolazione di riferimento rappresentata dalle piccole e medie imprese e i dati validati per usi statistici interni sono disponibili con una tempestività di circa 18 mesi dalla fine del periodo di riferimento. La loro accessibilità e chiarezza sono soddisfatte dalla possibilità da parte degli utilizzatori interni di accedere ai dati, ai metadati e alla documentazione; per gli utenti esterni è previsto l'uso dei soli risultati di specifiche analisi. La comparabilità nel tempo è garantita dal costante controllo dei cambiamenti nella normativa e legislazione, mediante collaborazione tra l'Istat e l'Agenzia delle Entrate. Le variabili del quadro contabile degli Sds sono coerenti con le stime dell'indagine Pmi e, a livello definitorio, anche con gli aggregati di Contabilità Nazionale, per i quali è ipotizzabile una stima del valore aggiunto, della struttura dei costi, dei margini commerciali e delle diverse tipologie di indipendenti.

Le informazioni descritte nelle righe precedenti possono essere sintetizzate nel prospetto 6 seguendo le linee generali del prospetto 1.

Prospetto 6 - Fasi di verifica ed elaborazione degli Sds in Istat

	1.	**Analisi dell'archivio amministrativo e dei rapporti tra l'istituto e l'ente amministrativo detentore dell'archivio**	
Pertinenza	1.1	Rilevanza dell'archivio	+
	1.2	Accordi tra l'istituto e l'ente amministrativo	+
Accuratezza	1.3	Controlli di qualità effettuati alla fonte	+/-
Puntualità	1.4	Puntualità del dato in riferimento agli obiettivi dell'indagine	+
Accessibilità / Tempestività	1.5	Aspetti sulla consegna, aggiornamento e messa a regine dell'archivio	+
Comparabilità	1.6	Mantenimento della qualità nel dato	+/-
	2.	**Analisi dei metadati dell'archivio**	+
Pertinenza	2.1	Rilevanza della popolazione di riferimento	+
Accessibilità	2.2	Esistenza di codici identificativi univoci	+
Coerenza	2.3	Corrispondenza delle unità (concetti statistici e amministrativi) e delle definizioni (definizioni, classificazioni)	+/-
Comparabilità	2.4	Mantenimento della qualità del metadato nel tempo	+
	3.	**Validazione delle variabili dell'archivio**	+
Pertinenza	3.1	Numero di variabili da validare	+
Coerenza	3.2	Tipo di variabili / esistenza di benchmark	+
Accuratezza	3.3	Copertura della popolazione statistica	+
Comparabilità	3.4	Comparabilità nel tempo	+/-

Nel prospetto 6 si evidenzia come la valutazione è positiva per la maggior parte delle fasi di verifica/elaborazione. Le uniche fasi che presentano una valutazione non completamente positiva sono: (1.3) l'accuratezza nei controlli di qualità effettuati dall'ente detentore della fonte in cui l'Istat non interviene se non a posteriori una volta acquisito l'archivio; (1.6) la comparabilità della qualità dell'archivio nel tempo realizzata dall'ente detentore, (2.3) la corrispondenza delle unità e delle definizioni nei metadati e (3.4) la comparabilità nel tempo dei valori delle variabili. Le ultime tre fasi sono soggette a modifiche a seguito delle variazioni nella normativa e legislazione vigente.

Un aspetto decisivo per garantire la continuità dei flussi informativi e la loro comparabilità nel tempo è rappresentato dalla gestione dei rapporti di collaborazione con l'ente detentore dell'archivio amministrativo finalizzata ad avere un ruolo attivo nella fase di progettazione del questionario, al fine di limitare le conseguenze delle eventuali modifiche nella normativa di riferimento e, quindi, di continuare a garantire nel tempo la comparabilità tra le statistiche prodotte dall'Istituto.

A titolo di esempio, si può fare riferimento alla questione relativa alla comparabilità della variabile *costo del personale*, e quindi della variabile di calcolo *valore aggiunto*. Nello specifico, nel questionario riferito all'anno di imposta 2006 si è rilevata un'incongruenza di definizione della variabile *spese per il personale dipendente* degli Sds rispetto all'omologa definizione presente nell'indagine Pmi: nel quadro F degli Sds, infatti, si rilevava in una stessa voce non solo le *spese per lavoro dipendente* ma anche le *spese per lavoratori coordinati e continuativi* (co.co.co.), le quali nell'indagine rappresentano invece costi per consulenze, e quindi sono annoverate nelle spese per servizi. L'Istat, attraverso il Servizio DAM, ha collaborato alla nuova stesura del quadro contabile del questionario al fine di creare i presupposti per un allineamento delle definizioni degli aggregati da rilevare (*costi del personale* e *costi per servizi*): l'operazione si è conclusa con la creazione di una sottovoce relativa alle *spese per co.co.co* nella variabile principale così da poterla scorporare e aggregarla così ai *costi per servizi*. Tale modifica ha consentito anche un miglior allineamento della variabile valore aggiunto (Cerroni e De Giorgi, 2008a).

3.1.3 La copertura dell'archivio degli studi di settore rispetto ad Asia

L'archivio Sds non è completamente esaustivo della popolazione scelta come riferimento, rappresentata dall'*Archivio Statistico delle Imprese Attive* (Asia), che contiene tutte le unità economiche italiane che nell'anno di riferimento esercitano arti e professioni nelle attività industriali, commerciali e dei servizi alle imprese e alle famiglie.

Come precisato nel paragrafo precedente vi sono delle imprese escluse dagli Sds, ovvero imprese che non hanno l'obbligo di compilare lo studio di settore. In questo paragrafo si tenta di calcolare la completezza dell'archivio Sds rispetto all'archivio Asia e di capire i motivi della non sovrapposizione delle due fonti. La successiva Tavola 1 riporta in percentuale la copertura.

Tavola 1 - Percentuale di copertura degli insiemi di riferimento di Sds e Asia. Anno 2007

		Asia		
		assente	*presente*	
Studi di Settore	*assente*	-	16,4	
	presente	6,8	76,8	
				100,0

Fonte: nostre elaborazioni su dati Istat

Come si evince dalla tavola precedente la percentuale di sovrapposizione tra gli insiemi di riferimento di Sds e Asia è del 76,8%. I restanti 16,4% e 6,8% rappresentano rispettivamente la percentuale di unità presenti in Asia ma assenti negli Sds e la percentuale di unità presenti negli Sds e assenti in Asia.

In particolare, il 16,4% di imprese che sono in Asia, ma non ritroviamo negli Sds è rappresentato da: 1) imprese che hanno un fatturato maggiore della soglia rilevata dagli Sds, che è di 7,5 milioni di euro e che non sono quindi non soggette alla compilazione dello studio (15,5%); 2) imprese che esercitano un'attività economica che non è sottoposta a studio di settore oppure imprese escluse dallo studio o a cui lo studio non è applicabile (0,6%); 3) errori dovuti a discordanze nell'attribuzione dell'attività economica in una delle due fonti ed errori dovuti al mancato accoppiamento del codice fiscale (0,3%). A parte la residuale quota dello 0,3% sono, dunque, imprese che sistematicamente mancano negli Sds.

Viceversa, il 6,8% di imprese che sono negli Sds ma non ritroviamo in Asia è rappresentato da:

1. unità con fatturato dell'ordine di poche migliaia di euro, che potrebbe aver rappresentato per Asia un segnale di inattività (4,7%);
2. unità fuori campo di osservazione di Asia o per attività economica o per forma giuridica (1,5%);
3. errori (0,6%).

Si precisa che dall'anno di imposta 2008, il regime fiscale ha introdotto i cosiddetti contribuenti minimi, la cui caratteristica principale, oltre ad altre di carattere strutturale, è rappresentata da ricavi non superiori ai 30 mila euro. Nell'anno 2008 erano circa 500 mila, nel 2009 circa 700 mila. L'informazione di contribuente minimo è stata utilizzata per determinare lo stato di attività di Asia a partire dall'anno 2009. Ciò vuol dire che negli Sds anno 2009 una quota maggiore di imprese presenti in Asia sono assenti negli Sds.

Nell'ottica di un utilizzo integrato di fonti amministrative le informazioni contenute nella dichiarazione dei contribuenti minimi per le imprese con ricavi minimi e nei bilanci civilistici delle società di capitali assieme alla fonte studi di settore rappresentano la quasi totalità delle imprese.[12]

In questo contesto, per valutare l'utilizzo degli Sds[13] per le piccole e medie imprese, si è ritenuto opportuno restringere l'analisi al solo insieme delle imprese di Asia con meno di 100 addetti e con attività economica prevalente compatibile con l'indagine. La sovrapposizione tra insieme Asia di piccole e medie imprese e gli Sds nell'anno 2007 si attesta sull'83%, maggiore quindi di circa il 5% rispetto all'analisi precedente senza restrizioni. La mancata copertura è da ricercare sempre nelle stesse cause strutturali dell'archivio Sds.

3.2 Applicazione della procedura 1 alla variabile ricavi degli Studi di settore

Come esempio dell'applicazione della procedura 1 si presenta il caso della variabile ricavi contenuta nel quadro F – anno 2007. La variabile degli Sds corrispondente ai ricavi è la somma di due variabili:

- Ricavi di cui alle lettere a) e b) dell'art. 85, comma 1, del TUIR (indicata con a)
- Ricavi derivanti dalla vendita di generi soggetti ad aggio o ricavo fisso (indicata con b)

Nell'indagine Pmi i ricavi sono rappresentati dalla variabile Ricavi delle vendite e delle prestazioni (indicata con c).

Siamo dunque nel caso B, cioè nel caso in cui le definizioni sono differenti ma riconciliabili attraverso un'operazione di somma, in questo caso solo delle variabili degli Sds.

Si analizzerà allora la somma $x = a + b$, variabile di analisi, mediante l'uso della variabile di controllo diretto *c*. L'uso del codice impresa presente nell'archivio Asia come variabile di aggancio tra le due fonti rende possibile l'abbinamento uno a uno dei record. Arrotondando i valori alle migliaia di euro si ottengono 28.156 abbinamenti dei quali il 77% sono uguali, mentre il restante 23%, che in termini di ammontare complessivo della variabile ricavo rappresenta un 28% circa, ha valori diversi nelle due fonti. Uno su cinque dei record abbinati hanno dunque valori non concordanti, e si tratta di ricavi generalmente maggiori. Si fissa allora una soglia massima di errore pari al 15% (differenza massima accettabile): gli scarti oltre tale soglia sono all'incirca il 5% dei valori (26.597 record), pertanto la percentuale complessiva delle differenze comprese entro la soglia del 15%, tra i valori della stessa variabile nelle due fonti, è pari al 95%. Se si volesse utilizzare una soglia più restrittiva, per esempio il 5%, l'analoga percentuale sarebbe comunque molto alta (92% circa).

Un'analisi più approfondita sui soli valori differenti[14] attraverso la stima kernel ci dice che gli scarti sono concentrati attorno allo 0 con forte curtosi, indice di dati sostanzialmente poco distanti,

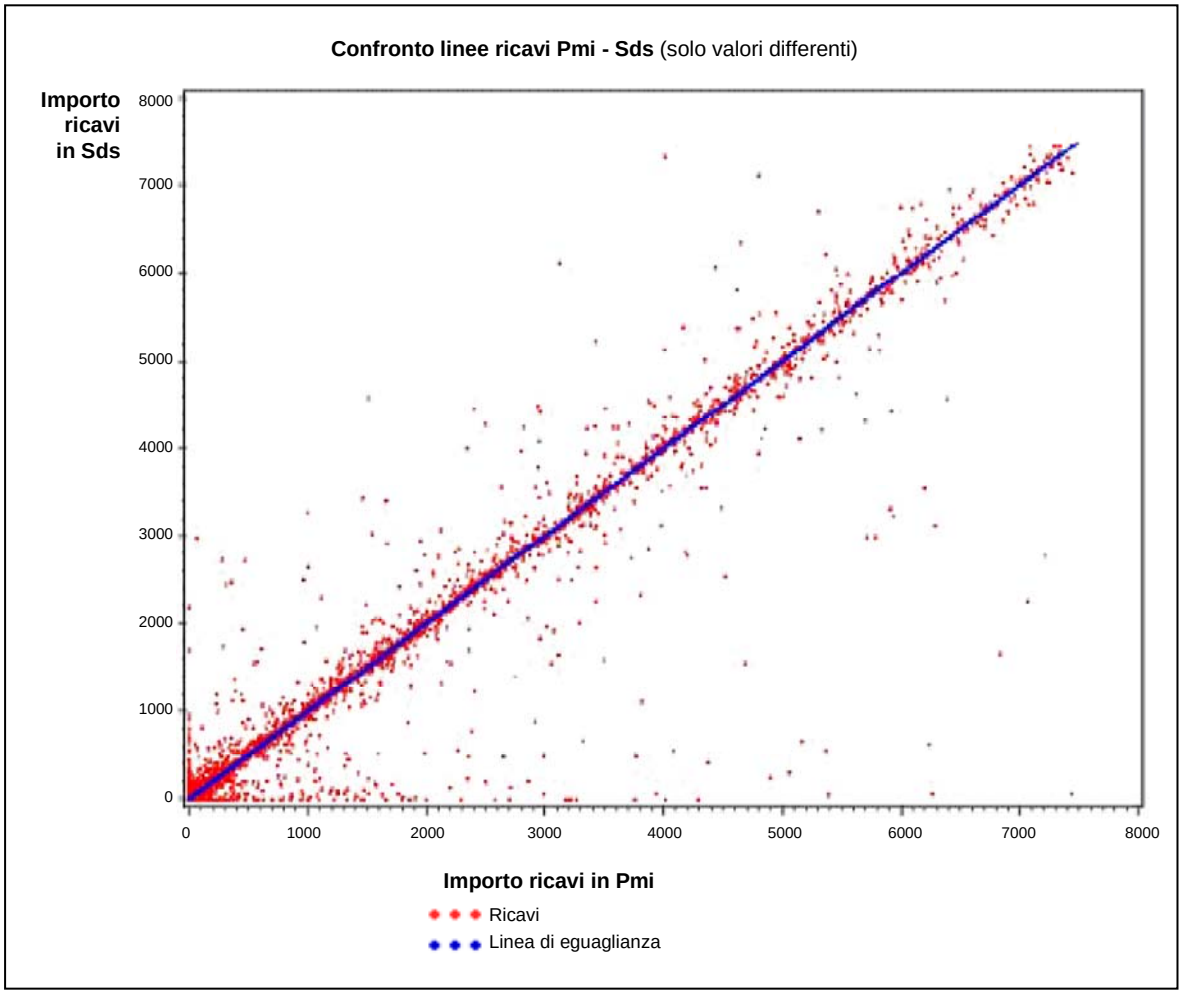
[12] Per una valutazione della copertura di Pmi si rimanda a Casciano et al. (2011b); per un riferimento all'uso di fonti integrate si rimanda a Cerroni et al. (2011).

[13] Eventualmente integrati con le poche informazioni dei contribuenti minimi.

[14] Il 23% dei 28.156 casi esaminati.

e del fatto che vi sono alcuni singoli valori di *c* nettamente inferiori ai corrispondenti di *x* che generano una forte asimmetria negativa. Un grafico scatter, sempre dei soli valori differenti, conferma tale caratteristica della distribuzione, come evidenziato in figura 1.

Figura 1 - Grafico scatter dei ricavi differenti



Ulteriori confronti e analisi basati sugli indici di locazione e di scala mostrano sia sui dati nel complesso sia sui soli dati differenti come le due distribuzioni siano molto simili.

Queste analisi confermano l'ipotesi dell'uguaglianza delle definizioni della variabile di analisi e della variabile di controllo.

Per l'utilizzo dei dati è necessario individuare l'insieme dei valori di *x* compatibili con *c* distinguendo prima di tutto l'uso che se ne dovrà fare. Premesso che a causa dell'elevata numerosità caratterizzante i dati in esame quando si porranno a confronto indici statistici si distinguerà, come indicato in letteratura, tra dati significativamente differenti da un punto di vista statistico e dati significativamente differenti da un punto di vista economico,[15] occorre distinguere se l'uso dei dati dovrà avvenire a livello micro oppure macro.

Se si ha l'obiettivo di trattare informazioni a livello puntuale, cioè di singola impresa, allora sulla base delle precedenti evidenze non sembra ammissibile ritenere che tutti i dati degli Sds

[15] Questa è una conseguenza del fatto che la potenza dei test statistici, cioè la capacità di rifiutare l'ipotesi nulla di eguaglianza delle statistiche poste a confronto, cresce con la numerosità dei dati e, nei casi qui in esame, è proprio quello che si verifica.

siano ritenuti compatibili con quelli di Pmi. Si potrà però pensare di impiegare evidentemente i 21.755 dati risultati eguali, che rappresentano oltre il 77% del complessivo abbinato e che, se si è disposti ad ammettere uno scarto tra i ricavi degli Sds e quelli dell'indagine non superiore al 5%, salgono ad oltre il 90% del totale e dovrebbero costituire un campione ragionevolmente rappresentativo. Difatti, si è osservato che il totale dei dati abbinati dà luogo a differenze economicamente irrilevanti e tanto più questo sarà vero se escludiamo quel 10% di casi con differenze superiori al 5%. Infine, si osserva che la parte dei dati dei ricavi degli Sds che non è stato logicamente possibile abbinare con Pmi, pari a circa 2.970.000 osservazioni, ha i principali indicatori di locazione, di scala e di forma molto simili da un punto di vista economico a quelli della parte abbinata, per cui si può applicare il campo di variazione, rilevato sulla parte abbinata, magari calcolandolo con esclusione di quel 10% di casi che ha differenze di un certo rilievo, per selezionare i dati degli Sds sulla parte non abbinata.

Se si ha invece l'obiettivo di trattare informazioni a livello aggregato, allora sulla base delle precedenti evidenze, il gruppo dei dati diversi pur avendo livelli maggiori dei dati eguali, mostra che tra Sds e Pmi le differenze sono da un punto di vista economico assolutamente irrilevanti, sia in valori di sintesi (medie e deviazione standard) sia in molti valori specifici.

3.3 Applicazione della procedura 2 alla variabile spese del personale dipendente degli Studi di settore

Come esempio dell'applicazione della procedura 2 si presenta il caso della variabile costo del personale. Poiché per costo del personale nella rilevazione Pmi si intende il costo del personale dipendente, la variabile degli Sds da prendere in considerazione per ottenere un aggregato confrontabile in definizione si ottiene detraendo le prestazioni diverse dal lavoro dipendente dalla variabile definita nel modo seguente:

- *spese per lavoro dipendente e per altre prestazioni diverse da lavoro dipendente afferenti l'attività dell'impresa* (a);

dove le altre prestazioni diverse sono le componenti seguenti:

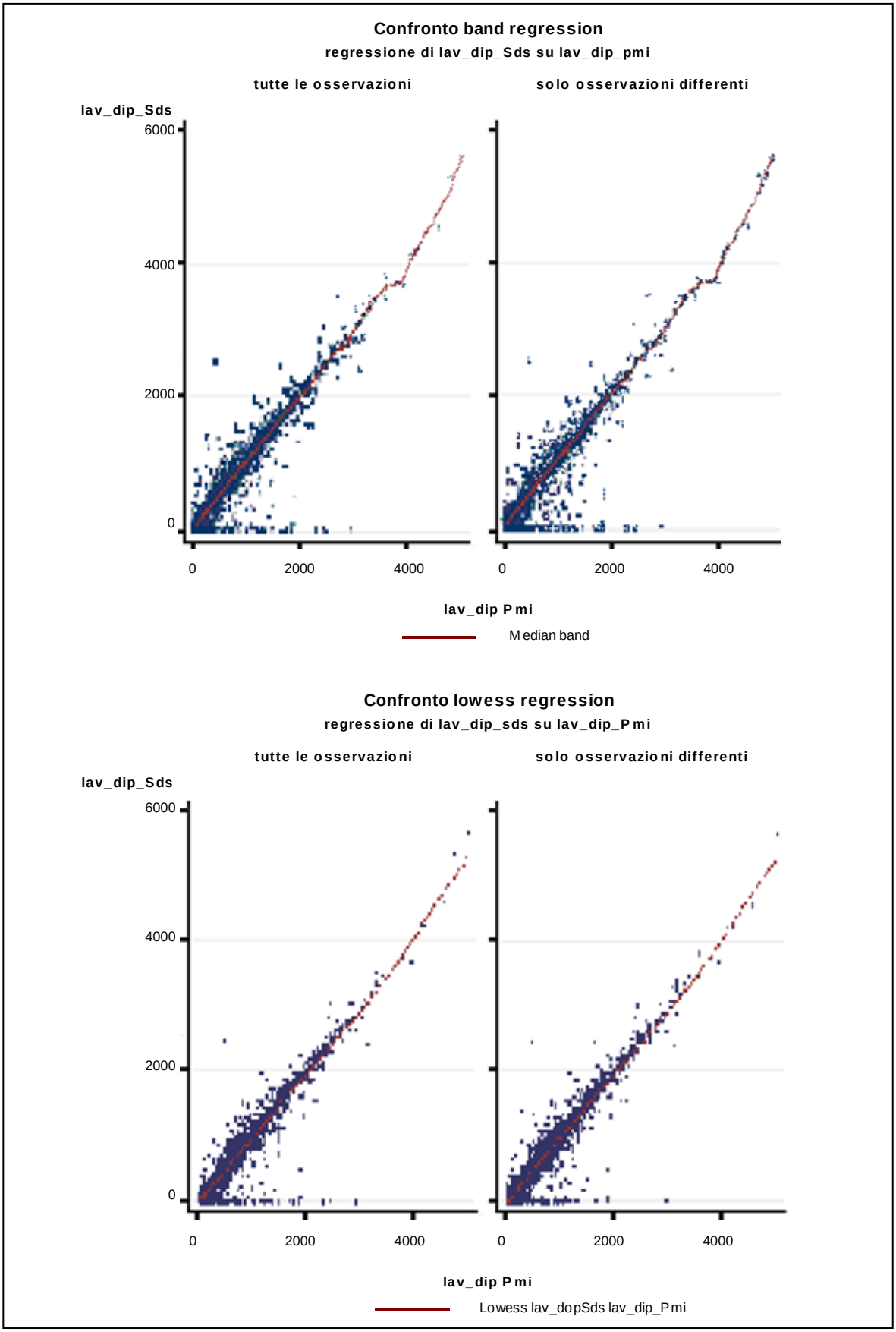
- di cui per *prestazioni rese da professionisti* (b)
- di cui per *personale di terzi distaccato presso l'impresa o con contratto di lavoro interinale o di somministrazione di lavoro* (d)
- di cui *per lavoratori coordinati e continuativi* (e).

Nell'indagine Pmi il costo del personale è rappresentato dalla variabile *costo per il personale* (c).

È da escludersi pertanto di trovarsi nel caso A (schema 1) mentre è possibile trovarsi di fronte al caso B oppure C. Si analizzerà allora la somma algebrica $x = a + (b + d + e)$, variabile di analisi, mediante l'uso della variabile di controllo diretto *c*.

Dopo un esame preliminare sui valori arrotondati alle migliaia di euro e su un totale di casi con variabili entrambe non nulle pari a 18.225, gli indicatori di sintesi delle variabili *a, b, d, e* mostrano che la precedente espressione algebrica relativa al costo del personale degli Sds calcolata sui valori medi delle variabili interessate produce un valore sensibilmente vicino a quello medio del *costo del personale* di Pmi dando così una prima giustificazione al confronto tra le variabili con le definizioni modificate. Tuttavia si nota che la modifica applicata, pur avendo sensibilmente avvicinato le due variabili, ha eguagliato di fatto il 54% dei casi, circa la metà, facendo supporre che ci si trovi nel caso *c* (definizioni non riconciliabili) anziché B (definizioni differenti ma riconciliabili). Nel caso dei ricavi (paragrafo 5.2) la modifica apportata aveva eguagliato una quota ben superiore pari al 77% dei casi, qui essa, invece, eguaglia una parte sensibilmente inferiore pari appunto al 54% delle imprese. Ciò premesso si è pensato di trattare l'applicazione come il caso *c* di variabili per le quali è possibile definire un'ipotesi di relazione funzionale e seguendo lo schema 3, si attiva la *procedura statistica 2* iniziando quindi con le regressioni esplorative. Siamo quindi nella fase 1 - alternativa 2 dello schema 3.

Figura 2 - Grafici di regressione m-band e lowess su tutte le osservazioni e solo sulle osservazioni differenti



Dalla *regressione m-band* eseguita su tutte le osservazioni e solo sulle osservazioni differenti (primi due grafici a sinistra) si osserva che le due variabili presentano fondamentalmente una relazione lungo una retta a 45° attorno alla quale gravita la gran parte dei punti. Ci sono però alcune curvature dopo i 3/4 milioni di euro e la presenza di un disturbo dato da alcuni punti estremi aventi in ascissa la variabile Pmi positiva e in ordinata la variabile Sds nulla che comportano forse eteroschedasticità. Un esame complementare con la *regressione lowess* conferma le precedenti osservazioni (ultimi due grafici a destra).

Dopo che si è valutato la forma lineare della relazione si procede sui soli dati differenti a stimare tale relazione impiegando tecniche di tipo robusto che si fanno precedere per motivi di controllo da una *regressione OLS* con compensazione dell'eteroschedasticità stimata con il metodo di White-Huber. Dalla *regressione lineare OLS* si ottiene che la relazione lineare ha un coefficiente angolare significativo molto vicino ad 1 (0.978885) e uno spostamento verso l'alto pari a circa 12 mila euro (11.99779) dato dall'intercetta che, evidentemente, assume il compito di correggere la variabile *x* del costo del lavoro dipendente Sds avvicinandola alla variabile di controllo *c* di Pmi. La varianza spiegata dal modello è molto alta ($R^2 = 0.94$).

Se i dati fossero omoschedastici cioè con errori standard dei coefficienti correttamente stimati si potrebbe individuare già in questa fase quei dati degli Sds tra quelli non abbinati a Pmi compresi nella fascia di confidenza costruita attorno alla retta di regressione che risultano compatibili e quindi validati. Poiché esiste conferma di eteroschedasticità,[16] il precedente quadro può essere ritenuto solo un riferimento per le successive analisi imperniate su alcune tecniche di *regressione robusta*, quali la *regressione quantile* e la *regressione M*.

La regressione quantile rispetto al modello precedente fornisce un'intercetta molto minore (circa 1.870 euro di maggiorazione) e un coefficiente angolare più vicino a 1, con un alto pseudo R^2. Anche la regressione robusta di tipo M[17] che è idonea a fronteggiare *outlier* di tipo verticale[18] nonché situazioni di eteroschedasticità, conferma le precedenti osservazioni. Per i casi in esame si considerano pertanto i valori massimo e minimo della variabile *costo per il personale* di Pmi e si costruisce un intervallo di variazione per la variabile dipendente *costo del personale* degli Sds da utilizzare come intervallo di valori ammissibili anche per i casi non abbinati.[19]

3.4 Applicazione della procedura 3 alle variabili dei costi degli Studi di settore

I costi sostenuti dai professionisti per la loro attività, presenti nel quadro G degli Sds dell'anno 2007, sono le variabili per le quali si suppone di trovarsi nella fattispecie D collegata alla *procedura statistica 3*, quella caratterizzata dall'assenza di variabili di controllo su cui basare un procedimento di validazione. La variabile *costi totali* dei professionisti degli Sds risulta dalla somma delle componenti seguenti:

- *spese per lavoro dipendente* (indicata con sps_dip);
- *spese per lavoro in somministrazione* (indicata con sps_lav_int);
- *spese per co.co.co* (sps_coco);
- *spese per compensi a lavoro di terzi* (comp_terzi);
- *spese per consumi* (consumi);
- *altre spese* (altro_sps).

Un primo esame delle variabili è fornito dalle seguenti statistiche che forniscono una panoramica sulla configurazione media dei costi di uno studio professionale, anche se va detto che sarebbe stato più opportuno disaggregare i precedenti dati in base ad alcune attività economiche. Le varia-

[16] Test di Breusch-Pagan/Cook-Weisberg applicato al modello OLS risulta avere $p<0.001$ ed è in accordo con la regressione robusta White Huber prima stimata che mostra errori standard dei coefficienti sensibilmente diversi.

[17] "Esposta nel paragrafo 2.2.1.2.

[18] Si tratta di dati in cui il punto è eccessivo (in più o in meno) come ordinata ma ha ascissa regolare.

[19] Una regressione ancora più selettiva, di tipo MM e cioè idonea a stimare la relazione anche in presenza di casi estremi sia sulla x sia sulla stessa c, ha permesso di selezionare oltre il 73% dei casi.

bili dei costi riferite a 687.287 imprese hanno la composizione riportata nella prima colonna da cui si rileva la predominanza delle altre spese (spese per attrezzature professionali e spese varie), mentre nelle forme di lavoro impiegate appare prevalere il lavoro dipendente sia pure strettamente affiancato dalle spese per compensi a lavoro di terzi.

Tavola 2 - Peso percentuale e principali indici di posizione

COMPONENTI DEI COSTI	Peso % sul totale costi	Media	Mediana	Sqm	Minimo	Massimo
spese per lavoro dipendente	27,0	5.413,1	0	23.792,7	0	1.113.772
spese per lavoro in somministrazione	0,2	40,5	0	2.012,4	0	600.148
spese per co.co.co.	1,2	241,0	0	32,68,2	0	762.899
spese per compensi a lavoro di terzi	23,6	4.770,7	0	25.578,7	0	2.906.008
spese per consumi	7,9	1.574,3	969	2.577,0	0	512.229
altre spese	40,2	8.081,7	2.780	24.886,7	0	3.803.453
Totale	**100,0**					

Fonte: nostre elaborazioni su dati Istat

La presenza di alti valori massimi rispetto alla media, con conseguente dilatazione dello scarto quadratico medio, induce il sospetto della presenza di dati estremi. Una verifica preliminare sull'esistenza dei dati estremi è data dal confrontare, in termini assoluti e relativi, alcuni indici di dispersione standard con indici robusti agli *outlier*. Si paragona pertanto lo scarto quadratico medio (sqm) con la mediana degli scarti delle differenze in valore assoluto dalla mediana (mad)[20] e con l'intervallo interquartile (iqr).

Tavola 3 - Principali indici di posizione e variabilità

COMPONENTI DEI COSTI	Sqm	Mad	Irq	Sqm/media (cv)	Mad/mediana	Iqr/mediana
spese per lavoro dipendente	23.792,7	0	0	4,4	n.d.	n.d.
spese per lavoro in somministrazione	2.012,4	0	0	49,7	n.d.	n.d.
spese per co.co.co.	3.268,2	0	0	13,6	n.d.	n.d.
spese per compensi a lavoro di terzi	25.578,7	0	1.000	5	n.d.	n.d.
spese per consumi	2.577,0	969	2.018	1,6	1,0	2,1
altre spese	24.886,7	2.307	6.099	3,1	0,8	2,2

Fonte: nostre elaborazioni su dati Istat

Si osserva che per tutte le variabili lo scarto quadratico medio è sensibilmente diverso dal mad, e che ciò si ripete anche quando si confronta il coefficiente di variazione standard (cv) con l'omologo (mad/mediana) misurato in termini robusti. Si rileva inoltre che leggendo lungo le righe della tavola, il quoziente mad/mediana, quando calcolabile, non appare divergere molto dal rapporto iqr/mediana. Entrambe queste circostanze danno conferma all'ipotesi che esistono effettivamente alcuni dati estremi.

La rilevazione puntuale dei medesimi avviene normalmente con opportune tecniche di analisi che però, come noto, sono di difficile lettura quando esse devono trattare una grande mole di dati come quella qui in esame. Per affrontare tale problema si è allora fatto ricorso a un doppio passaggio: si è prima applicata una tecnica di rilevazione multivariata degli *outlier*, le componenti principali residue sintetizzate nelle due statistiche Hisq e Disq (Khattre e Naik, 2000), e poi le loro risultanze sono state lette con l'ausilio di grafici box-plot.

Scendendo nei dettagli dei risultati della procedura, si osserva che le sei variabili di costo esaminate vengono riassunte nella misura dell'85% della loro variabilità complessiva dalle prime quattro

[20] L'indice è basato su un doppio calcolo della mediana: primo, si computano in valore assoluto gli scarti tra le singole osservazioni della variabile x e la loro mediana, cioè xi-mediana(x); secondo, si calcola la mediana di tali scarti, da cui l'acronimo mad, che sta per median of absolute deviation.

componenti principali, per cui le ultime due possono essere ritenute residuali. Questo significa che i punti che giacciono in uno spazio R^4 se sono non estremi sono rappresentati anche sulla quinta e sesta componente con coordinate piccole, mentre tali quinta e sesta componente saranno non piccole nel caso contrario, cioè con gli *outlier*. Ne consegue che la somma delle distanze euclidee dei punti dal sottospazio ottimale a quattro dimensioni (la statistica Disq), che equivale a misurare la somma dei quadrati delle componenti residue, eventualmente standardizzate dal rispettivo autovalore (la statistica Hisq), registrerà valori alti per i punti estremi,[21] Poiché come detto l'elevata numerosità dei punti in esame non permette di trovare una *cutoff* nitido, si è calcolato sulle statistiche Disq e Hisq il grafico box plot che rileva cautelativamente come estremi i punti eccedenti le soglie *far high fence* e *far low fence* (rispettivamente i punti inferiori al primo quartile meno tre volte l'intervallo interquartile o superiori al terzo quartile più tre volte l'intervallo interquartile). Il grafico box plot individua, oltre alle due soglie indicate, anche altre due soglie di livello più contenuto, date dalla *high fence* e *low fence* (pari, rispettivamente, alle precedenti ma con fattore di molteplicità di 1,5 per l'intervallo interquartile). Per motivi cautelativi si sono assunti come dati estremi solo quelli eccedenti le due soglie più estreme. Sono stati così individuati 78.535 punti estremi su un totale di 687.752 (11%).

L'analisi fattoriale, eseguita sui dati regolari con il metodo di estrazione degli assi del fattore principale e con estrazione di quattro assi, ha rilevato un valore complessivo dei *Residui quadratici medi fuori diagonale* = 0,065589707, che supera di poco il limite usuale presente in letteratura pari a 0,05 ma tale risultato va naturalmente valutato ricordando l'elevato numero di osservazioni presenti. Anche la stessa statistica riferita alle singole variabili dell'analisi ha dato i risultati ammissibili. Le unicità - cioè la parte della variabilità di ogni variabile non spiegata dai fattori - resta in generale al di sotto della soglia dello 0,40 che è il limite superiore normalmente suggerito in letteratura.

Tavola 4 - Correlazioni residue con unicità della diagonale

	sps_dip	sps_lav_int	sps_coco	comp_terz	consumi	altro_sps
sps_dip	**0.24362**	-0.03495	0.01973	0.01186	-0.01948	-0.06518
sps_lav_int	-0.03195	**0.00542**	-0.00344	0.00020	0.03650	0.00077
sps_coco	0.01973	-0.00344	**0.00250**	-0.00186	-0.02847	0.00746
comp_terz	0.01186	0.00020	-0.00186	**0.00942**	0.03029	-0.04307
consumi	-0.19478	0.03650	-0.02847	0.03029	**0.33461**	-0.12733
altro_sps	-0.06518	0.00077	0.00746	-0.04307	-0.12733	**0.01974**

Fonte: nostre elaborazioni su dati Istat

Una rotazione degli assi ha portato ai risultati presentati in tavola 5 da cui si osserva che il caricamento delle variabili sui fattori è nitido, non essendovi variabile che carica su due o più fattori congiuntamente, che il primo asse indica un collegamento, segno di correlazione, tra le tre variabili di costo evidenziate in grassetto, mentre gli altri tre assi non fanno altro che rappresentare, ciascuno, una variabile di costo non correlata alle altre.

Tavola 5 - Schema fattoriale ruotato

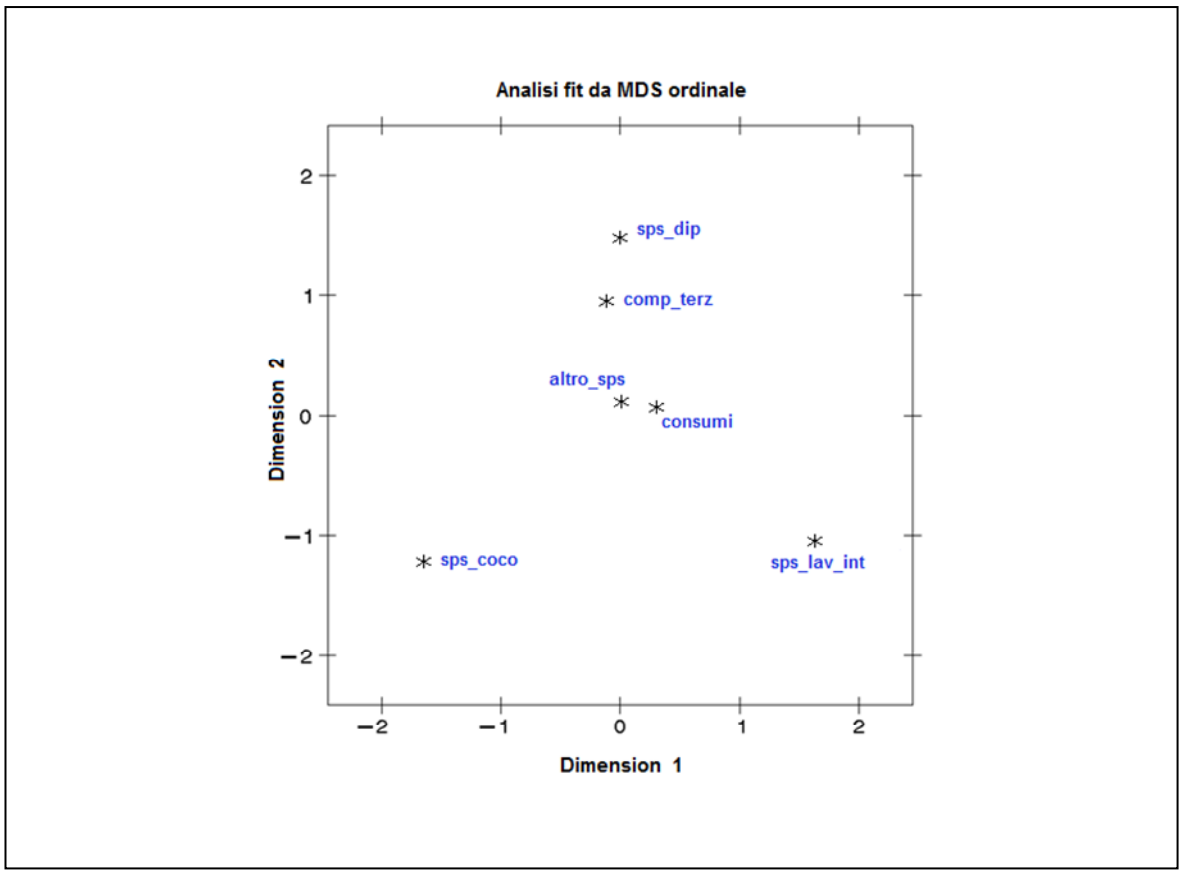
	Factor1	Factor2	Factor3	Factor4
sps_dip	0.85902	0.08263	0.10776	0.00564
sps_lav_int	0.03469	0.00551	0.99665	0.00624
sps_coco	0.07760	0.04637	0.00639	0.99467
comp_terz	0.22550	0.96812	0.00609	0.04945
consumi	0.80554	0.07137	-0.04442	0.09709
altro_sps	0.84924	0.28272	0.00411	0.03764

Fonte: nostre elaborazioni su dati Istat

[21] In tal caso si può eseguire un test sui limiti oltre i quali le distanze rilevate indicano dati anomali, nell'ipotesi restrittiva che i dati provengano da una normale *p*-variata (Khattre e Naik, 2000 op. cit,. pp. 65-66).

Una verifica con la tecnica del *multidimensional scaling* di tipo moderno ha confermato i risultati precedenti, così come riportato nella figura 3 aggiungendo un'ulteriore informazione e cioè che anche la variabile *compensi per lavoro di terzi* (comp_terzi) potrebbe appartenere al gruppo delle 3 variabili individuate dall'analisi fattoriale. Andando a riprendere la matrice delle correlazioni tra variabili e fattori tavola 5 si osserva in effetti che la variabile *compensi per lavoro di terzi* carica sul primo fattore con un coefficiente pari a 0,22 che, considerando che si tratta di microdati con altissima numerosità, è da ritenersi di una certa rilevanza seppure inferiore a quella delle altre 3 variabili prima evidenziate.

Figura 3 - Analisi multidimentional scaling per le variabili dei costi dei professionisti (quadro G) - Anno 2007



In conclusione, la *procedura statistica 3* qui applicata ha prodotto i seguenti risultati: esistono dati estremi stimabili in circa il 10% del totale e qualora si esegua un'analisi sulla parte dei dati non estremi si osserva una partizione delle variabili in due blocchi: un primo, con le spese per lavoro dipendente, per consumi, per altre spese e forse anche per compensi di terzi che risultano correlate positivamente (spese complementari), e un secondo dato dalle altre voci di spesa (forme di lavoro non dipendente) che risultano incorrelate tra loro e dalle variabili del primo blocco. Il primo gruppo di variabili potrebbe pertanto essere interpretato come quello costituito dalla parte strutturale degli studi professionali, mentre il secondo sembra rappresentare quella parte degli studi professionali che si attiva – e si disattiva – in modo indipendente, con aspetti di possibile succedaneità rispetto al primo blocco.

Se queste osservazioni risultano conformi alla comune conoscenza ed esperienza sul settore in esame, si può allora sostenere che la procedura 3 qui esposta, avendo confermato la buona qualità delle variabili in esame, ha pure convalidato le medesime.

Conclusioni

Questo documento riporta alcune esperienze maturate presso la *DCAR, Servizio DAM*, nell'ambito dello studio di un processo di validazione delle variabili economico-contabili contenute nei quadri F e G degli Sds. La peculiarità specifica dell'archivio degli Sds risiede nella straordinaria ricchezza di informazioni presenti sia in termini di unità (4 milioni di imprese rilevate) sia di variabili (circa 200 tipologie di questionario ciascuno con decine di variabili rilevate), ed è questo un ulteriore motivo per cui lo si è scelto come archivio pilota per lo studio qui descritto.

L'attenzione del processo di validazione è stata posta sulle variabili di tipo economico-contabile poiché sono di rilevante importanza sia nei confronti della rilevazione sulle piccole e medie imprese Pmi, sia per le stime degli aggregati di Contabilità Nazionale, essendo in linea con la IV direttiva comunitaria.

Un problema da tenere sotto controllo è l'inevitabile discrepanza che da sempre esiste tra le finalità fiscali sottostanti l'archivio amministrativo degli Sds e gli scopi statistici perseguiti in Istituto. A esso si aggiunge il fatto che le norme legislative sottostanti la rilevazione di dati amministrativi e fiscali possono cambiare nel tempo, e ciò può avere una ricaduta sulle definizioni e quindi sui valori dei dati rilevati, a tal punto da non renderli più confrontabili con i dati delle indagini statistiche. Un modo per garantire la continuità dei flussi informativi e la loro stabilità nel tempo è quello di mantenere un rapporto stabile e costruttivo con l'ente che gestisce i dati amministrativi. A tal proposito, nei confronti degli Sds, l'Istat ha avuto un ruolo attivo nella gestione dei rapporti con l'Agenzia delle Entrate soprattutto nel ridisegnare il questionario della rilevazione.

Alle delineate criticità si è risposto anche attraverso un complesso di procedure statistiche atte a valutare la qualità dei dati amministrativi in esame, procedure che possono essere estese a tanti altri archivi amministrativi e costituire pertanto un approccio più generale per la trasformazione di un archivio amministrativo in un registro statistico.

Riferimenti bibliografici

Agresti A. 2002. *Categorical Data Analysis*. Wiley.

Allison P.A. 1999. *Logistic Regression Using Sas.* Cary, NC: Sas Institute.

Bakker B.F.M. 2009. *Methodenreeks, ondedeel Micro-integratie.* The Hague, Statistics Netherlands.

Basilevsky A. 1993. *Statistical Factor Analysis and Related Methods*. Wiley, p. 508 e subs.

Bernardi A. 2011. Rappresentazione sintetica di indicatori di qualità per i dati amministrativi. *Istat Working Papers* n. 6/2011.

Bernardi A., F. Cerroni, V. De Giorgi. 2008. A Methodological Process for Assessing Variables coming from Administrative Sources: an Application to the Tax Authority Source (Sector Studies). Lavoro presentato alla European Conference on Quality in Official Statistics Q2008. Roma, 9-11 luglio.

Bernardi A., F. Cerroni, V. De Giorgi. 2010. Analysis on Economic Fiscal Data for Statistical Uses. Lavoro presentato al Seminario Essnet Meets "Using Administrative Data in the Production of Business Statistics". Roma, 18-19 Marzo.

Cameron A.C., P.K.Trivedi. 2009. *Microeconomics Using Stata*. Stata Press.

Casciano M.C., V. De Giorgi, F. Oropallo, G. Siesto. 2010. Experimental Analysis in the Estimation of SBS Variables for Small Firms by Using Administrative Data. Lavoro presentato al Seminario Essnet "Using Administrative Data in the Production of Business Statistics". Roma, 18-19 marzo.

Casciano M.C., V. De Giorgi, O. Luzi, F. Oropallo, G. Siesto. 2011a. Integration of administrative data to estimate structural business statistics (SBS) variables in the sample survey on Italian small and medium enterprises in *SIS2011 Statistical Conference. Alma Mater Studiorum-Università di Bologna. June 8, 2011- June 10, 2011. Statistics in the 150 years from Italian Unification. Book of Short Paper*. Quaderni di Dipartimento. Serie Ricerche 2011, n. 10: ISSN 1973-9346.

Casciano M.C., A. Cirianni, V. De Giorgi, T. Di Francescantonio, A. Mazzilli, O. Luzi, F. Oropallo, M. Rinaldi, E. Santi, G. Seri, G. Siesto. 2011b. Utilizzo delle fonti amministrative nella rilevazione sulle piccole e medie imprese e sull'esercizio di arti e professioni. *Istat Working Papers* n.7/2011.

Cerroni F, V. De Giorgi. 2008a. Nota su rilevanza e utilizzo di nuove fonti amministrative per l'analisi delle imprese. Il quadro economico degli Studi di Settor*e*. Nota tecnica Servizio DAM/B. Roma, Giugno.

Cerroni F, V. De Giorgi. 2008b. The Tax Authority Source as an Example of the Use of an Administrative Source as a Statistical One. Lavoro presentato alla conferenza IAOS2008. Shanghai: 14-16 ottobre.

Cerroni F, V. De Giorgi. 2010. Uso di Dati Amministrativi a Fini Statistici nel Progetto di Gemellaggio Internazionale con l'Istituto di Statistica della Tunisia. Lavoro contenuto nella relazione finale del porgetto di gemellaggio Istat-INS Tunisia n.TU/07/AA/OT/02. Roma, aprile.

Cerroni F, V. De Giorgi, M. Mantuano. 2011. L'impatto della crisi sui risultati economici delle imprese: analisi microeconomica del settore tessile e del settore IT. Lavoro presentato al convegno "L'analisi dei dati di impresa per la conoscenza del sistema produttivo italiano: il ruolo della statistica ufficiale". Roma, 21-22 novembre.

Daas P.J.H., S.J.L. Ossen, J. Arends-Tóth. 2009. Framework of Quality Assurance for Administrative Data. Lavoro presentato alla 57a sessione dell'ISI. Durban, South Africa: 16-22.

ESC. 2007. Pros and Cons for Using Administrative Records in Statistical Bureaus. Lavoro presentato alla seminario su "Increasing the efficiency and productivity of statistical offices". Economic and Social Council conference of European statisticians. Ginevra, Svizzera.

Everitt B., S. Landau, M. Leese. 2001. *Cluster Analysis.* A Hodder Arnold Publication: 4th edition.

Hamilton L.C. 2009. *Statistics with Stata.* Brooks/Cole, Cengage Learning.

Istat., *Conti economici delle imprese. Anno 2003*.

Istat, *Rapporto Annuale. La situazione del paese nel 2008*, Roma, Maggio 2009.

Istat, *Rapporto Annuale. La situazione del paese nel 2009*, Roma, Maggio 2010.

Istat, *Struttura e Competitività del Sistema delle Imprese Industriali e dei Servizi. Anno 2008.*

Khattre R., D.N. Naik. 2000. *Multivariate Data Reduction using Sas*. Crystal Dreams Publishing, p. 300.

SAS. 2009. On-line Documentation. Sas/Stat Module, example 28.2.

Piccolo D. 2000. *Statistica*. Il Mulino, p.66.

Van del Laan P. 2000. Integrating administrative registers and household surveys. Netherlands Official Statistics, pp. 7-15.

Wallgren A., B. Wallgren. 2007. *Register-based Statistics: Administrative Data for Statistical Purposes*. Wiley, marzo.

Zuliani A., C. Scala. 1964. *Complementi di Statistica Metodologica*. Ed. Kappa, p. 264.

Informazioni per gli autori

La collana è aperta ad autori dell'Istat e del Sistema statistico nazionale, e ad altri studiosi che abbiano partecipato ad attività promosse dal Sistan (convegni, seminari, gruppi di lavoro, ecc.). Da gennaio 2011 essa sostituirà Documenti Istat e Contributi Istat.

Coloro che desiderano pubblicare sulla nuova collana dovranno sottoporre il proprio contributo alla redazione degli Istat Working Papers inviandolo per posta elettronica all'indirizzo iwp@istat.it. Il saggio deve essere redatto seguendo gli standard editoriali previsti, corredato di un sommario in italiano e in inglese; deve, altresì, essere accompagnato da una dichiarazione di paternità dell'opera. Per la stesura del testo occorre seguire le indicazioni presenti nel foglio di stile, con le citazioni e i riferimenti bibliografici redatti secondo il protocollo internazionale 'Autore-Data' del *Chicago Manual of Style*.

Per gli autori Istat, la sottomissione dei lavori deve essere accompagnata da una mail del proprio dirigente di Servizio/Struttura, che ne assicura la presa visione. Per gli autori degli altri enti del Sistan la trasmissione avviene attraverso il responsabile dell'ufficio di statistica, che ne prende visione. Per tutti gli altri autori, esterni all'Istat e al Sistan, non è necessaria alcuna presa visione. Tutti i lavori saranno sottoposti al Comitato di redazione, che valuterà la significatività del lavoro per il progresso dell'attività statistica istituzionale. La pubblicazione sarà disponibile su formato digitale e sarà consultabile on line.

Gli articoli pubblicati impegnano esclusivamente gli autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat. Si autorizza la riproduzione a fini non commerciali e con citazione della fonte.