

ARCHIVI TERRITORIALI: IL VALORE DEI BIG DATA COME NUOVA FONTE DI VALIDAZIONE.

Armando d’Aniello¹, Daniela Fusco², Luigi Praitano³

SOMMARIO

Gli archivi amministrativi locali costituiscono un importante patrimonio informativo, sia per conoscere le realtà territoriali sia a supporto dei decisori politici.

Tuttavia tali archivi contengono errori caratteristici di questa tipologia di fonte che non la rendono utilizzabile senza accurati accorgimenti. Gli errori più comuni sono il lag temporale tra data di presentazione dell’atto e periodo di riferimento dello stesso, la perdita di informazione per ritardo nell’aggiornamento della fonte e la presenza di unità catalogate con classificazioni obsolete.

In questo studio si dimostra come i Big Data possano intervenire nella risoluzione e nel superamento di tali limiti. A tale scopo sono state integrate due fonti amministrative della Regione Campania contenenti l’elenco delle strutture turistiche ricettive (Rilevatore Statistico Regionale e Turismo web) e confrontate con l’elenco delle strutture presenti su Tripadvisor e il Registro statistico delle imprese attive (Asia). Attraverso tecniche di integrazione è stato ottenuto un archivio unico. Lo studio mostra quale sia l’effetto dell’introduzione della fonte Big Data nella determinazione dello stato di attività delle strutture ricettive presenti nei registri di impresa di derivazione locale.

¹ Università degli Studi di Napoli PARTHENOS, Napoli, e-mail: armando.daniello@studenti.uniparthenos.it.

² Istat, Dipartimento per lo sviluppo di metodi e tecnologie per la produzione statistica, Napoli, e-mail: dafusco@istat.it (corresponding author).

³ Regione Campania, Ufficio di statistica, Napoli, e-mail: luigi.praitano@regione.campania.it.

2. Introduzione

I dati di fonte amministrativa sono entrati da tempo a far parte della statistica ufficiale. Con essi si fa riferimento alle informazioni prodotte dalla PA per fini di natura amministrativa che, trattati dagli istituti di ricerca, divengono utilizzabili a fini statistici.

Tale processo di integrazione è stato reso possibile dal rapido sviluppo informatico della PA che ha reso disponibili informazioni strutturate e facilmente utilizzabili su imprese, istituzioni e individui.

Tuttavia tali dati scontano le conseguenze di errori non campionari caratteristici di questa tipologia di fonte. In particolare: il lag temporale tra data di presentazione dell'atto e periodo di riferimento dello stesso, la perdita di informazione per ritardo nell'aggiornamento della fonte e la presenza di unità catalogate con classificazioni obsolete. Tali aspetti possono ridurre il potere informativo degli archivi amministrativi.

Questo contrasta con la necessità della PA di avere dati corretti per i fini di governance preposti alla raccolta delle informazioni

In questo studio si dimostra come i Big Data possano intervenire a migliorare la copertura degli archivi della PA. In particolare verrà analizzato il database della Regione Campania contenente l'elenco delle strutture turistiche ricettive al fine di valutare se tali strutture possano essere considerate attive. A tale scopo sono state integrate le due fonti amministrative regionali (Rilevatore Statistico Regionale e Turismo web), l'elenco delle strutture presenti Tripadvisor e il Registro statistico delle imprese attive (Asia). Lo studio mostra l'effetto dell'introduzione della fonte Big Data nella determinazione dello stato di attività delle strutture ricettive presenti nei registri regionali.

3. Gli archivi amministrativi territoriali

La statistica ufficiale per fronteggiare la crescente richiesta di nuove informazioni, più tempestive ed a maggior dettaglio territoriale, dovuta ad una necessità di analizzare e governare i rapidi cambiamenti sociali ed economici del Paese, a cui si è aggiunta una minore disponibilità di risorse stanziare per la Pubblica Amministrazione, è stata indotta a sviluppare le condizioni ed a promuovere sempre più l'uso sistematico delle fonti amministrative a fini statistici.

Questo processo è stato favorito anche da diversi fattori:

- Lo sviluppo della digitalizzazione nella Pubblica Amministrazione, che ha reso disponibile una grande mole di informazioni strutturate su imprese, istituzioni ed individui.
- Il crescente costo delle indagini campionarie connesso alla problematica della diminuzione dei tassi di risposta, che rende necessario il contenimento del fastidio statistico sui rispondenti.
- Il progresso tecnologico che ha aumentato la capacità di archiviare, processare ed analizzare una quantità sempre maggiore di dati.

La tendenza ad un progressivo utilizzo delle fonti amministrative, che caratterizza anche il contesto europeo, ha comportato una revisione del tradizionale processo di produzione statistica e la costruzione di nuovi strumenti metodologici e di valutazione della qualità, dei processi e dei dati, definiti da Eurostat.

Nello specifico i dati di fonte amministrativa possono essere impiegati in tre tipologie di processi di produzione statistica:

- Produzione diretta di dati da una fonte.
- Produzione diretta di dati da più fonti.
- Produzione indiretta dei dati, attraverso l'utilizzo dei dati delle fonti a supporto delle indagini statistiche.

4. Le fonti analizzate

Le esigenze che hanno indotto la statistica ufficiale a sviluppare condizioni finalizzate all'uso sistematico delle fonti amministrative sono le stesse che hanno anche promosso lo sviluppo e la ricerca di nuove fonti di dati, da integrare anch'esse con quelle già presenti (statistiche ed amministrative): i Big Data. L'obiettivo infatti è estrarre il più possibile valore dai dati, citando quanto descritto dall'allora presidente dell'Istat G. Alleva, "la capacità di estrarre valore dai dati è legata alla capacità di integrare dati che provengono da fonti diverse" (Alleva G., 2017).

Volendo precisare, con il termine Big Data intendiamo una raccolta di dati così estesa in termini di volume, velocità e varietà da richiedere tecnologie e metodi analitici specifici per l'estrazione di valore (De Mauro et al., 2016). Questi dati cui si fa riferimento sono originati dall'uso degli strumenti digitali e registrano eventi o spesso "comportamenti" (spontanei) degli utilizzatori.

I principali vantaggi apportati dall'utilizzo di fonti esterne a fini statistici consistono nel:

- Aumento di accuratezza, coerenza e completezza delle informazioni statistiche prodotte, mediante l'integrazione di più fonti.
- Ampliamento dei contenuti informativi della produzione statistica, attraverso la diffusione di nuove informazioni sui fenomeni e sulle popolazioni già oggetto di indagine o di dati relativi a fenomeni e realtà non ancora analizzati dal punto di vista statistico.
- Riduzione del disturbo statistico, eliminando o ridimensionando indagini correnti, sostituibili, interamente o parzialmente, con dati di fonte amministrativa, diminuendo i rischi di non risposta.
- Possibilità di ottenere una copertura totale delle popolazioni di riferimento delle statistiche, consentendo di aumentare il dettaglio territoriale di riferimento delle informazioni prodotte, anche al di sotto del livello comunale.
- Riduzione dei costi per la Pubblica Amministrazione.

Nonostante uno dei vantaggi appena descritti sia la riduzione dei costi, l'impiego di queste fonti non è un'operazione a costo zero, infatti occorre verificare che i dati contenuti negli archivi amministrativi, rilevati per fini diversi da quelli statistici, siano comparabili con quelli della statistica ufficiale. Per ottimizzare i processi risulta quindi indispensabile che concetti, definizioni e classificazioni siano quanto più standardizzati ed omogenei con quelli adottati della statistica ufficiale e ciò spesso si traduce in una problematica non di poco conto, che richiede lavori impegnativi finalizzati a trattare i dati amministrativi e renderli statisticamente utilizzabili.

L'armonizzazione a monte delle fonti amministrative è dunque un punto cruciale sia per un'attività statistica ottimizzata e sostenibile sia per la realizzazione di una sinergia tra i sistemi informativi della Pubblica Amministrazione. L'adozione di concetti, definizioni e classificazioni della statistica ufficiale da parte degli enti della Pubblica Amministrazione infatti, può risultare un utile strumento per l'interscambio informativo tra gli enti stessi e per il miglioramento dell'efficacia della loro azione amministrativa. Quanto appena descritto si può considerare certo poiché concetti, definizioni e classificazioni di cui sopra sono condivisi e consolidati all'interno non solo del sistema statistico nazionale ma anche di quello internazionale.

Analogamente per quanto concerne l'impiego delle fonti amministrative, anche quello dei Big Data comporta delle problematiche:

- | | |
|--|--|
| <ul style="list-style-type: none">• La popolazione statistica obiettivo è diversa da quella fonte.• Definizioni e classificazioni spesso non coincidono con quelli della statistica ufficiale.• L'accesso ai dati può essere problematico. | <ul style="list-style-type: none">• Occorre valutare la disponibilità e la qualità dei dati.• Si possono presentare problemi tecnologici dovuti al trattamento di ingenti moli di dati.• Può risultare difficoltoso estrarre informazioni statisticamente rilevanti. |
|--|--|

Sono dunque necessarie onerose attività per estrarre informazioni di valore dai Big Data. Se al momento i metodi utilizzati non sono considerati sufficienti, si prevede che nell'imminente futuro i Big Data saranno utili, così come le fonti amministrative oggi, per aumentare la tempestività delle informazioni, ampliare le opportunità di analisi e migliorare la qualità delle stime.

4.1. Il Rilevatore Statistico Regionale e Turismo web

Il Rilevatore Statistico Regionale della Campania è un applicativo della Regione Campania, ha lo scopo di fornire un sistema di comunicazione tra le strutture ricettive e gli uffici turistici competenti per la trasmissione in via telematica dei dati di movimentazione turistica. Permette di facilitare gli adempimenti obbligatori per legge agli operatori del settore e di disporre in tempo reale dei dati delle presenze turistiche a vantaggio di una migliore programmazione turistica o territoriale.

Il popolamento del Rilevatore Statistico Regionale avviene per diretto contatto del titolare della struttura con l'Ufficio regionale competente. Quindi in via teorica, in attesa che il titolare riceva la SCIA dagli Uffici comunali competenti si ha una preregistrazione sul portale in maniera tale che già si possa avere un invio dei dati di movimento da parte della struttura.

La piattaforma Turismo Web è il portale della Regione Campania dedicato alle strutture ricettive ubicate in regione per la comunicazione on-line dei prezzi ed ai Comuni della regione per il censimento anagrafico delle strutture ricettive. Per poter ottenere le credenziali della piattaforma (quindi registrarsi), la struttura deve essere obbligatoriamente in possesso di SCIA.

La piattaforma Turismo Web consta di due sezioni:

1° Sezione (Comunicazione prezzi da parte delle strutture)

In essa tutte le strutture ricettive ad eccezione degli Agriturismi hanno l'obbligo di registrarsi e comunicare i prezzi.

2° Sezione (Comunicazione da parte del Comune)

Tramite tale sezione i comuni assolvono ad un obbligo di legge, due volte all'anno, nell'inviare l'elenco delle strutture sul territorio competente. Gli Uffici regionali competenti supervisionano tale flusso informativo. Le strutture non possono accedere a questa sezione.

Le eventuali modifiche della singola struttura vengono prese in carico con gli aggiornamenti semestrali dei comuni.

L'anagrafica di Turismo Web deriva dalla 2° Sezione.

Entrambe le fonti fanno riferimento all'anno 2018.

4.2. ASIA

Il Registro statistico delle imprese attive (Asia) è costituito dalle unità economiche che esercitano arti e professioni nelle attività industriali, commerciali e dei servizi alle imprese e alle famiglie e fornisce informazioni identificative (denominazione e indirizzo) e di struttura (attività economica, addetti dipendenti e indipendenti, forma giuridica, data di inizio e fine attività, fatturato) di tali unità.

Asia nasce nel 1996, in base al Regolamento del Consiglio Europeo n. 2816/93 relativo al coordinamento comunitario dello sviluppo dei registri d'impresa utilizzati a fini statistici, poi abrogato e sostituito dal Regolamento CE n. 177/2008.

Le unità di analisi sono le imprese, le variabili comprese nel registro sono classificate secondo tre tipologie: variabili identificative (ragione sociale, indirizzo e altri caratteri per l'esatta individuazione dell'unità sul territorio); variabili di stratificazione (attività economica dell'impresa classificata secondo la classificazione Ateco, forma giuridica, dimensione dell'impresa, in termini di addetti indipendenti e dipendenti medi annui e di fatturato); variabili demografiche (data di nascita e cessazione dell'impresa, data di eventi quali scopri, fusioni o procedure concorsuali, fallimenti, liquidazioni, eccetera).

Ai fini della produzione dell'informazione statistica, le imprese sono classificate per attività economica, definita in base ad un livello specifico della nomenclatura Ateco.

Sono escluse dal campo di osservazione le imprese appartenenti ad alcuni settori quali: Agricoltura, silvicoltura e pesca (sezione A della classificazione Nace Rev.2); amministrazione pubblica e difesa; assicurazione sociale obbligatoria (sezione O); attività di organizzazioni associative (divisione 94); attività di famiglie e convivenze come datori di lavoro per personale domestico; produzione di beni e servizi indifferenziati per uso proprio da parte di famiglie e convivenze (sezione T); organizzazioni ed organismi extraterritoriali (sezione U); le unità classificate come istituzioni pubbliche e istituzioni private non profit. Il Registro è aggiornato annualmente attraverso un processo di integrazione di informazioni provenienti sia da fonti amministrative, gestite da enti pubblici o da società private sia da fonti statistiche oltre che attraverso una rilevazione campionaria di controllo della copertura di ASIA, di aggiornamento delle unità locali (IULGI) e di completamento dei registri satellite.

Nel 2017 (ultimo dato definitivo disponibile) Asia contiene 4 milioni e 398 mila per 4 milioni e 747 mila unità locali e un totale di 17 milioni e 59 mila addetti (Istat, 2018). Il maggior numero di imprese e unità locali (oltre il 79 per cento) è impiegato nei servizi, cui corrisponde circa il 69 per cento di addetti (oltre il 35 per cento nel commercio, trasporto e magazzinaggio, alloggio e ristorazione). Nell'industria in senso stretto sono presenti il 9,2 per cento di imprese a cui corrisponde il 23,4 per cento degli addetti complessivi. Lombardia e Lazio sono le regioni con più imprese (rispettivamente 18,5 e 10,0 per cento) e addetti (23,6 e 11,1 per cento) e le uniche (ad eccezione della provincia autonoma Bolzano) in cui gli addetti delle unità locali sono inferiori (e anche di molto) a quelli delle imprese.

4.3. Tripadvisor

In questo studio, al fine di individuare le strutture presenti sul web, si è deciso di utilizzare le informazioni contenute sugli HUB (siti web che contengono le informazioni relative alle strutture ricettive). Sono stati confrontati i dati contenuti nelle principali piattaforme. Si è deciso quindi di utilizzare Tripadvisor poiché la numerosità delle strutture presenti è congrua con quanto pubblicato dall'Istat⁴.

Tripadvisor® è tra le piattaforme di viaggi più grande del mondo, utilizzata da circa 463 milioni di viaggiatori ogni mese. Viene utilizzata per consultare oltre 859 milioni di recensioni e opinioni relative a 8.6 milioni di alloggi, ristoranti, esperienze, compagnie aeree e crociere.

In questo lavoro sono stati estratti i dati delle strutture ricettive registrate in Campania mediante web scraping, i dati estratti sono relativi a marzo 2020. I dati sono stati ricavati attraverso scraping su due livelli in cui si sono estrapolate informazioni riguardo la denominazione, l'indirizzo, il comune e la tipologia delle strutture.

5. La qualità delle fonti

Per quanto concerne i due dataset regionali sono state effettuate le valutazioni attraverso il metodo proposto da Statistics Netherlands: per verificare la possibilità d'impiego a fini statistici sono valutate le tre hyperdimension: Source, Metadata e Data (Daas P., 2016). Queste sono esaminate mediante delle checklist singolarmente per ogni fonte di dati e nell'ordine in cui sono state citate.

Per ogni hyperdymension sono state valutate le dimensioni di cui essa è composta mediante specifici indicatori di qualità, calcolati con metodi di misura qualitativi o quantitativi. In questo caso, la prima hyperdimension, Source, non è stata valutata perché le due fonti sono in collaborazione con l'Istituto. Per quanto riguarda la hyperdimension Data, sebbene sia presente una tecnica di valutazione, l'autore delle checklist specifica che la sua valutazione è contingente sia ai dati sia all'utilizzo cui sono destinati,

⁴ La rilevazione annuale della capacità delle strutture ricettive rileva le principali informazioni di carattere strutturale degli esercizi ricettivi. La rilevazione è svolta in conformità al Regolamento (UE) n. 692/2011 del Parlamento Europeo e del Consiglio del 6 luglio 2011 che regola le Statistiche Europee sul Turismo

per questo motivo in questo lavoro sono stati costruiti solo gli indicatori di qualità effettivamente calcolabili tra quelli proposti dall'autore, più altri da noi introdotti. Sulla base delle checklist sono stati dunque costruiti i relativi indicatori di qualità riguardo le hyperdimension Metadata e Data. Le valutazioni complessive sulle fonti inerenti alle dimensioni delle due hyperdimension valutate sono riportate nelle seguenti tabelle:

Tabella 1 - Hyperdimension Metadata, indicatori.

METADATA		
DIMENSIONI	FONTI	
	Turismo WEB	Rilevatore Turistico Regionale
1. Chiarezza	+	+
2. Comparabilità	+/ o	+
3. Chiave unica	+	+
4. Trattamento dei dati	-	-

Fonte: nostra elaborazione su dati Regione Campania

Tabella 2 - Hyperdimension Data, indicatori.

DATA		
DIMENSIONI	FONTI	
	Turismo WEB	Rilevatore Turistico Regionale
1. Controlli tecnici	+	+
2. Sovra copertura	+	+
3. Sotto copertura	?	?
4. Unità non rispondenti	+	+
5. Voci mancanti	+/ o	o /-
6. Misurazione	+/ o	o /-
7. Sensibilità	+	o

Fonte: nostra elaborazione su dati Regione Campania

+	Buono	?	Non chiaro
o	Ragionevole	/	Punteggio intermedio
-	Scarso		

La valutazione della hyperdimension Data costituisce la fase dell'analisi più approfondita ed è qui che si incontrano le problematiche sui dati: per il dataset proveniente da Turismo Web non si riscontrano particolari difetti, ma il dataset ricevuto dal Rilevatore Statistico Regionale presenta carenze in merito alle dimensioni "voci mancanti" e "misurazione", in dettaglio vi è un cospicuo numero di campi vuoti nei dati per la prima, in particolar modo per la variabile chiave contenente le partite iva o i codici fiscali dei titolari delle strutture ricettive (il 90%), e la seconda dimensione presenta un elevato numero di campi compilati in modo errato, specie per la variabile chiave inerente all'indirizzo della struttura (il 75%).

Per quanto riguarda il dataset ricavato dal HUB Tripadvisor la qualità è stata valutata semplicemente considerando il numero di strutture in esso presenti attraverso un confronto tra più fonti sia ufficiali sia fonti web. Nella tabella seguente si mostra la diversa numerosità delle strutture nelle fonti.

Tabella 3 - Numerosità delle strutture per fonte e provincia.

Provincia	Fonte					
	Istat	Tripadvisor	Google	Booking	Kayak	Hotels.com
Caserta	439	353	671	160	5.871	2.515
Benevento	635	244	415	63	141	343
Napoli	3.453	3.714	7.408	4.148	7.103	4.805
Avellino	403	231	399	406	108	1.660
Salerno	2.255	2.517	4.613	680	1.592	2.873
Totale Campania	7.185	7.059	13.506	5.457	14.815	12.196

Fonte: nostra elaborazione su dati Istat e internet

L'analisi ha evidenziato come il numero di strutture registrate in Tripadvisor, rispetto ad altri HUB, si possa ritenere più attendibile se comparato con i dati Istat, seppur relativi all'anno 2018.

6. Il record linkage per l'integrazione dei dati

Prima di procedere alla trattazione del lavoro, per chiarire la terminologia utilizzata e comprendere le operazioni svolte si riassume la teoria di base del record linkage.

Il record linkage è definito, nella forma più generale, come una tecnica algoritmica il cui scopo è identificare quali coppie di record di due basi di dati, intese in questo contesto come matrici unità-variabili, corrispondono ad una stessa unità (Belin, Rubin, 1995).

Nella pratica se esiste un codice identificativo o un insieme di variabili "prive di errori" che può svolgere il compito di un codice identificativo, il problema del ricongiungimento dei record di due dataset è banale e si svolge mediante merge.

I metodi per il record linkage si occupano del caso in cui un unico codice identificativo non è presente e tra le restanti variabili ce ne sono alcune (o tutte) in grado di identificare le unità ma sono riportate con errore. Questi errori comportano il mancato riconoscimento di alcuni veri match oppure l'interpretazione di falsi match come veri. L'obiettivo è dunque minimizzare gli errori di abbinamento.

Le prime operazioni da effettuare prima ancora di procedere alle vere e proprie fasi del record linkage consistono nella:

- Raccolta di informazioni relative ai dati da utilizzare, in particolar modo per quanto concerne le definizioni di unità e popolazione.
- Richiesta e controllo dei dati ottenuti dalla fonte.
- Preparazione dei dati per il record linkage.

Svolti questi passaggi preliminari si può procedere alle fasi proprie del record linkage, che, così come definite da T. B. Jabine e F. J. Scheuren, sono tre (Jabine T.B., Scheuren F.J., 1986):

1. *Preprocessing* (Pre-processamento).
2. *Matching* (Applicazione dell'algoritmo di record linkage).
3. *Analysis* (Analisi dell'output).

Da un punto di vista analitico l'applicazione del record linkage richiede innanzitutto la presenza di:

- Due basi di dati che abbiano in comune un insieme non vuoto di unità. Definiamo gli insiemi di unità presenti nelle due basi di dati rispettivamente A e B di numerosità v_A e v_B .
- La presenza di k ($k \geq 1$) variabili chiave presenti nelle due basi di dati e definiamo la variabile che osserva congiuntamente le k variabili chiave con (X_1, \dots, X_k) , che quindi è in grado di identificare univocamente le unità.

Ipotizzando che le variabili chiave siano prive di errori, che non abbiano subito modifiche nell'intervallo di tempo intercorso tra le due rilevazioni e che non siano presenti mancate risposte parziali, basterebbe confrontare l'esatta corrispondenza dei valori di (X_1, \dots, X_k) tra le unità delle due rilevazioni per identificare quelle che sono state osservate in entrambe. Poiché queste condizioni nella realtà non si verificano, per controllare la corrispondenza delle unità provenienti dalle due rilevazioni, si rende necessaria una procedura di decisione basata sul confronto tra le modalità assunte dalle variabili chiave.

Dati i due insiemi A e B definiti in precedenza, le procedure di record linkage confrontano ogni unità (record) di A con ogni unità (record) di B sulla base delle variabili chiave scelte. Tutte le coppie da confrontare rappresentano gli elementi che costituiscono l'insieme dato dal prodotto cartesiano tra i due insiemi di partenza.

Sulla base di quanto definito, l'obiettivo del record linkage consiste nella bipartizione dell'insieme $A \times B$ in due sottoinsiemi disgiunti ed esaustivi: M ed U , tali che:

$$M \cap U = \emptyset, \quad M \cup U = A \times B = \Omega,$$

$$\text{dove} \quad M = \{ (a, b) \in A \times B : a = b \}, \quad U = \{ (a, b) \in A \times B : a \neq b \}.$$

Considerando l'insieme di coppie che costituiscono $A \times B$ come una matrice C , detta di configurazione, in cui ogni riga fa riferimento ad ogni unità di A ed ogni colonna ad ogni unità di B , i cui elementi $c_{a,b}$ assumono valore:

$$1 \text{ se } (a, b) \in M, \quad 0 \text{ se } (a, b) \in U,$$

l'obiettivo del record linkage può essere nuovamente definito nel determinare gli elementi della matrice C .

Sulla matrice di configurazione possono essere stabiliti vincoli di abbinamento, nel nostro caso è stata applicata una procedura di record linkage *one to one* per cui il vincolo sulla matrice consiste nello stabilire che per ogni riga e colonna la somma degli elementi può essere al più uguale ad uno:

$$\sum_{a=1}^{v_A} c_{a,b} \leq 1, \forall b \in B, \quad \sum_{b=1}^{v_B} c_{a,b} \leq 1, \forall a \in A.$$

Definendo $\mathbf{X}_A = \{x_{a,j}^A\}$ la matrice di v_A righe e k colonne, dove il termine $x_{a,j}^A$ rappresenta la modalità della variabile X_j sull'unità a in A , con $j = 1, \dots, k$, $a = 1, \dots, v_A$, ed ugualmente $\mathbf{X}_B = \{x_{b,j}^B\}$ la matrice di v_B righe e k colonne, dove il termine $x_{b,j}^B$ rappresenta la modalità della variabile X_j sull'unità b in B , con $j = 1, \dots, k$, $b = 1, \dots, v_B$, la funzione di confronto è rappresentata formalmente in questo modo, per semplicità supponiamo che \mathbf{y} sia un vettore ottenuto confrontando ad una ad una le singole variabili chiave:

$$\mathbf{y} = (y_1, y_2, \dots, y_k), \quad \text{con} \quad y_j = f(x_{a,j}^A, x_{b,j}^B), \quad j = 1, \dots, k$$

Essa serve a determinare il livello di disuguaglianza che insiste tra valori delle variabili di *matching*. In linea di principio, ci si aspetta livelli bassi di diversità per le coppie che costituiscono i match e livelli più elevati per gli *unmatch*.

Si definiscono le funzioni di confronto utilizzate in questo lavoro:

- Funzione di uguaglianza: la funzione valuta solamente se i valori delle variabili di matching hanno valore uguale o diverso.

- Funzione di confronto di *n-grams*: il *n-grams* di una stringa è dato da tutte le sottostringhe di ampiezza *n* che si possono formare con i caratteri consecutivi al suo interno. La funzione derivante dal confronto di due stringhe mediante *n-grams* assume valori in (0,1) ed è data dal rapporto tra il numero di *n-grams* in comune ed il numero medio di *n-grams* delle stringhe.

Si definisce \mathbf{Y} “variabile vettore dei confronti”, costituita da tutti i valori di \mathbf{y} , e D lo spazio dei possibili valori assumibili da \mathbf{Y} , un pilastro fondamentale del record linkage è la distribuzione di frequenza di \mathbf{Y} osservata sulle coppie di record che compongono Ω :

$$P(\mathbf{Y} = \mathbf{y} \mid (a, b)), \quad \mathbf{y} \in D.$$

\mathbf{Y} è costruita sulla base di due distribuzioni di probabilità condizionate, $m(\mathbf{y})$ ed $u(\mathbf{y})$, a seconda che la coppia associata al valore di \mathbf{y} appartenga rispettivamente ad M od a U :

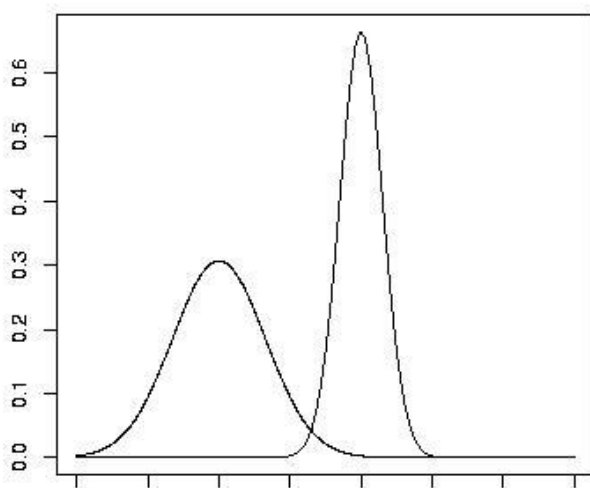
$$m(\mathbf{y}) = P(\mathbf{Y} = \mathbf{y} \mid (a, b) \in M), \mathbf{y} \in D$$

$$u(\mathbf{y}) = P(\mathbf{Y} = \mathbf{y} \mid (a, b) \in U), \mathbf{y} \in D$$

Le due distribuzioni permettono quindi di visualizzare quali valori della variabile sono più frequenti tra le coppie in M e quali tra le coppie in U . In generale ci si aspetta che la funzione di densità di \mathbf{Y} abbia un andamento regolare, infatti la funzione $m(\mathbf{y})$ confronta le differenze per coppie di record che costituiscono un match, le quali dovrebbero essere caratterizzate, a meno di errori, dalle stesse modalità delle variabili chiave, per cui la densità sarà concentrata nei valori che rappresentano bassi livelli di diversità. Mentre la funzione $u(\mathbf{y})$ confronta le differenze per coppie di record costituenti unmatch, per cui viceversa, la densità sarà concentrata nei valori che rappresentano livelli di diversità elevati. Un esempio del tipico andamento della funzione \mathbf{Y} è mostrato nella figura 1.

Osservando il grafico si sottolinea che più aumenta la “distanza” tra le due distribuzioni più è semplice il compito di distinguere la categoria d’appartenenza della generica coppia (a, b) . In genere la distribuzione dei match è più concentrata rispetto a quella dei non-match. Va aggiunto che poiché le distribuzioni dipendono dal tipo di funzione di confronto scelto, l’ideale sarebbe quello di utilizzare funzioni il più discriminanti possibile ma questo comporta elevati costi computazionali e più parametri da stimare, per cui nella pratica si utilizzano le funzioni poco complesse e meno discriminanti.

Figura 1 - Andamento della funzione \mathbf{Y} , costituita dalle funzioni di densità $m(\mathbf{y})$ (destra) ed $u(\mathbf{y})$ (sinistra).



Fonte: Scanu M. (2003).

Per quanto concerne il nostro caso di studio prima di procedere in dettaglio alle operazioni svolte è necessario definire che è stata svolta una dapprima integrazione tra i dataset regionali. Il nuovo dataset ottenuto è stato poi integrato a sua volta con il registro Asia attraverso un merge basato sui codici fiscali dei titolari,

codici delle partite iva, province e comuni delle strutture. In seguito, è stato ricavato mediante web scraping il dataset dei dati contenuti nel HUB Tripadvisor, il quale dopo essere stato preparato per il record linkage, è stato integrato, eliminandone i residui, con quello ricavato precedentemente dalle fonti regionali e dal registro Asia.

6.1. Pre-processamento

Questa prima fase ha l'obiettivo di rendere compatibili ed omogenei i dati all'interno dei dataset da integrare. Di solito questa è la fase che nel complesso richiede maggior tempo. L'insieme di azioni che compongono la fase di pre-processamento sono di seguito elencate:

- Scelta delle variabili chiave: nella pratica viene scelto il numero minimo di variabili che, congiuntamente, identificano univocamente le unità. Le variabili chiave sono scelte tra quelle che sono: universali, variabili a cui tutte le unità rispondono, permanenti, variabili che restano immutate nel tempo, accurate, non sensibili, in merito alla riservatezza dei dati delle unità. Va detto che difficilmente è possibile trovare ed utilizzare variabili chiave che rispettino contemporaneamente tutti i requisiti esposti.
- Miglioramento della qualità dei dati: in genere è un'operazione che andrebbe svolta in fase di pianificazione di un'indagine allo scopo di ottenere dati quanto più accurati e completi, nel caso in cui i dati provenissero da rilevazioni statistiche.
- Standardizzazione delle variabili: consiste nella trasformazione delle modalità delle variabili di modo da aumentarne il potere discriminante e rendere più evidenti le differenze tra i record per l'elaboratore. Un esempio è l'eliminazione dei titoli dai nomi degli individui, delle imprese e degli indirizzi.
- *Blocking and Sorting*: è un metodo che consiste nell'ordinare e dividere in gruppi i record nelle due basi di dati secondo le modalità di una o più variabili. È un'operazione effettuata quasi sempre perché permette di ridurre lo spazio di ricerca in cui vengono svolti i confronti e quindi complessità computazionale del problema.

Nel nostro caso per quanto riguarda il pre-processamento dei due dataset regionali, una volta valutata la qualità si è proceduto alla fase di trattamento e standardizzazione dei dati, il processo si è articolato in diverse operazioni svolte in modo dettagliato ed in un preciso ordine:

1. Conversione di tutti i caratteri alfabetici in maiuscolo.
2. Conversione dei dati nei record riguardanti la provincia ed il comune delle strutture nei relativi codici Istat.
3. Sono stati resi omogenei gli indirizzi delle strutture attraverso la conversione delle più svariate abbreviazioni presenti nei dati. Sono state rimosse le diverse annotazioni riguardanti ad esempio gli indirizzi senza civico, le specifiche della scala, piano, CAP, etc..
4. Sono state rimosse dai dati delle variabili riguardanti le denominazioni delle strutture le parole ricorrenti che causano l'abbinamento di falsi match durante l'applicazione degli algoritmi di linkage (i.e.: "HOTEL", "B&B", etc.).
5. Correzione dei numeri di telefono delle strutture, in particolare è stato fatto in modo che per ogni record fosse presente un solo numero di telefono, ove presente, scritto correttamente: privo di caratteri speciali, alfabetici e spazi, aggiungendo lo zero iniziale per i numeri fissi ove mancante.
6. Rimozione di spazi e caratteri speciali presenti all'inizio di ogni valore dei dati.
7. Rimozione di tutti i caratteri speciali dai dati ad eccezione di "@" e "." da quelli riguardanti indirizzi e-mail, pec e siti web.
8. Correzione degli indirizzi mail, pec e siti web ove presenti, in particolare è stato fatto in modo che per ogni record fosse presente un solo indirizzo ed un solo sito scritti correttamente.
9. Sono state aggiunte due variabili ad entrambi i dataset derivanti dalla scomposizione degli indirizzi: una variabile per il titolo (DUG) ed una per i nomi degli indirizzi (NOMESTR) – i.e.: Via Giuseppe Verdi -> DUG=Via / NOMESTR=Giuseppe Verdi.

Quest'operazione è stata fatta allo scopo di eliminare il "rumore" creato dai titoli degli indirizzi nel calcolo delle funzioni di confronto similmente a quanto fatto per le denominazioni.

10. Aggiunta di una variabile contatore (COUNT) finalizzata ad avere un riferimento fisso associato ad ogni record nelle fasi successive.

Per quanto riguarda invece il dataset ricavato mediante web scraping dal Hub Tripadvisor si è proceduto in modo simile, si elencano le principali operazioni:

1. Conversione di tutti i caratteri alfabetici in maiuscolo.
2. Rimozione della parte di testo riguardante le recensioni dai dati relativi alla tipologia della struttura.
3. Correzione della variabile relativa al comune della struttura, inserendo il valore della località ove mancante.
4. Correzione della variabile tipologia, in particolare inserendo il valore della seconda variabile relativa alla tipologia ove il valore della prima risultasse vuoto e viceversa
5. Aggiunta di una variabile contenente la classificazione Istat delle strutture (alberghiere ed extra alberghiere) ottenuta dalla tipologia.
6. Correzione della variabile relativa ai numeri di telefono.
7. Correzione della variabile relativa ai civici degli indirizzi.
8. Conversione dei DUG degli indirizzi.
9. Conversione dei Comuni nei relativi codici territoriali Istat.
10. Assegnazione manuale dei Comuni mancanti e successiva conversione.
11. Rimozione parole ricorrenti dalle denominazioni delle strutture.

6.2. Applicazione dell'algoritmo

L'applicazione dell'algoritmo utilizzato per l'abbinamento dei record delle due basi di dati richiede innanzitutto la scelta di un approccio: deterministico o probabilistico.

- L'approccio deterministico è basato sulla decisione a priori di regole che se rispettate definiscono i match. Secondo tale approccio, due record sono considerati match se coincidono nei valori delle variabili chiave scelte oppure soddisfano un sistema di regole che attribuisce un punteggio ad ogni coppia di record, e se tale punteggio supera una determinata soglia la coppia è considerata match. La definizione delle regole dipende dai dati e dalla conoscenza della popolazione di riferimento. Nell'approccio deterministico l'incertezza nella corrispondenza, stabiliti i criteri di matching, è ridotta al minimo, ma il tasso di abbinamento può essere basso. Ad ogni modo è efficace nel caso in cui sono presenti degli identificatori univoci per le unità o variabili chiave, riportati senza errore. Inoltre, in presenza di errori o valori mancanti, il metodo deterministico è poco robusto.

- L'approccio probabilistico basa la scelta dei match sulla probabilità di abbinamento dei record attraverso metodi statistici che includono procedure di stima e test sui dati. Secondo tale approccio, viene definito un modello che genera i dati osservati, si stabilisce una regola di decisione che ha l'obiettivo di essere "ottima", si stimano gli elementi utili all'applicazione della regola di decisione, vengono definite delle probabilità di errore.

È opportuno sottolineare che nella pratica possono anche essere utilizzati in modo combinato applicando sulle variabili di alta qualità il metodo deterministico e sui residui, intesi come le restanti unità non abbinate delle due basi di dati, il metodo probabilistico.

Per quanto riguarda l'approccio probabilistico, precedentemente sono state definite le distribuzioni m ed u , ed il loro meccanismo nel contesto del record linkage. Sulla base di queste distribuzioni Fellegi e Sunter nel 1969 (Fellegi I.P., Sunter A.B., 1969) hanno sviluppato una procedura di abbinamento dei record definita "ottimale".

Innanzitutto, ad ogni coppia di record (a, b) bisogna associare una decisione tra tre possibili alternative:

- A_m : stabilire che la coppia è un match.

- A_u : stabilire che la coppia è un non-match.
- A_\emptyset : stabilire che non si hanno sufficienti informazioni per decidere se la coppia costituisce o meno un match.

Le scelte si basano totalmente sulle informazioni che si hanno a disposizione, i confronti $\mathbf{y}_{a,b}$.

Le decisioni A_m ed A_u vengono definite “decisioni positive”. A_\emptyset , definito anche “match incerto”, è una decisione che si prende nel caso in cui si rende necessaria una verifica manuale per stabilire lo status della coppia e bisogna considerare che questo tipo di scelta comporta costi per evolverla in una decisione positiva, motivo per cui bisogna cercare di prendere la decisione A_\emptyset nel minor numero di casi.

L’obiettivo è dunque ridurre i match incerti, controllando i possibili errori quando vengono prese decisioni positive.

Supponendo di conoscere la distribuzione di frequenza dei confronti \mathbf{y} sulle coppie in M ed in U , la procedura decisionale è affidata a un meccanismo probabilistico:

- scelgo A_m con probabilità $P(A_m|\mathbf{y})$,
- scelgo A_u con probabilità $P(A_u|\mathbf{y})$,
- scelgo A_\emptyset con probabilità $P(A_\emptyset|\mathbf{y})$,

con $P(A_m|\mathbf{y}) + P(A_u|\mathbf{y}) + P(A_\emptyset|\mathbf{y}) = 1$.

Tale meccanismo si articola in due passaggi:

1. Si trasforma il vettore $\mathbf{y}_{a,b}$, di k dimensioni, in un numero reale attraverso una funzione dei vettori di confronto, definita in questo contesto $t(\mathbf{y}_{a,b})$ ed indicata col nome di “peso”. Ad esempio, considerando come funzione di confronto quella di uguaglianza, una funzione dei vettori di confronto potrebbe essere la somma dei valori dei valori contenuti nei vettori, come nel caso considerato precedentemente per mostrare l’andamento delle distribuzioni $m(\mathbf{y})$ ed $u(\mathbf{y})$:

$$t(\mathbf{y}_{a,b}) = \sum_{h=1}^k y_{a,b}^h.$$

In realtà la funzione ideale è quella che permette di includere tutte le informazioni necessarie, utili a discernere le coppie appartenenti ad M da quelle appartenenti ad U .

Si determinano due intervalli disgiunti dei valori di $t(\mathbf{y}_{a,b})$ per identificare i pesi delle coppie incluse in M e quelli delle coppie incluse in U . Rifacendoci all’esempio precedente bisogna individuare due valori di $t(\mathbf{y}_{a,b})$, τ_λ e τ_μ , con $0 < \tau_\lambda < \tau_\mu < k$, tali che nell’intervallo $[0, \tau_\lambda]$ siano incluse le coppie appartenenti ad U ed in $[\tau_\mu, k]$, le coppie appartenenti ad M . Questo perché le coppie con valori più alti di $t(\mathbf{y}_{a,b})$, per costruzione, corrispondono a coppie di record che probabilmente costituiscono un match e viceversa. L’intervallo di valori $[\tau_\lambda, \tau_\mu]$ include tutte le coppie di record per le quali non si hanno informazioni sufficienti a stabilirne l’insieme di appartenenza.

Gli intervalli sono determinati sulla base del trade-off tra costi ed errori tollerati, analogamente al meccanismo dei test statistici, dove il peso $t(\mathbf{y}_{a,b})$ corrisponderebbe al test rapporto di verosimiglianze.

Passando adesso al nostro caso, per quanto riguarda la prima integrazione tra i dataset regionali, dopo aver effettuato il pre-processamento dei dati, ciascun dataset è stato diviso in due in modo da separare i record relativi alle strutture della Provincia di Salerno da quelli delle strutture delle restanti province della Campania. Quest’operazione è stata effettuata per via dell’elevato numero di strutture risultanti nella provincia di Salerno.

L’integrazione dei record delle strutture è avvenuta eseguendo molteplici prove ed in più passaggi, verificando manualmente l’output di volta in volta, sfruttando opportunamente ed alternativamente le diverse combinazioni di variabili chiave contenute nei dataset. La scelta delle variabili chiave da utilizzare è stata fatta considerandone la completezza, la qualità ed il potere discriminante.

Per quanto riguarda il tipo di approccio scelto, inizialmente è stato adottato un approccio deterministico e

successivamente probabilistico ma data la qualità delle variabili si è optato successivamente per un approccio più combinato il quale si è mostrato più efficiente.

La seconda integrazione tra il dataset regionale integrato ed il registro Asia è stata svolta attraverso un merge mediante le variabili relative ai codici fiscali dei titolari, partite iva, province e comuni delle strutture.

Per la terza integrazione, eseguita tra il dataset ottenuto dalle precedenti integrazioni e quello ricavato da Tripadvisor, è stato adottato esclusivamente un approccio deterministico, diminuendo le soglie di inclusione e controllando l'output ad ogni passaggio.

6.3. Analisi dell'output (fasi)

Nella prima integrazione, tra i due dataset regionali, sono state svolte sei fasi di record linkage per i dati relativi alla Campania meno la provincia di Salerno e sette per i dati di quest'ultima. In entrambi i casi sono stati impiegati in modo combinato il metodo deterministico e quello probabilistico.

Si descrivono e si elencano le definizioni dei metodi di riduzione dello spazio di ricerca, delle funzioni di confronto utilizzate e dei nomi delle variabili impiegate.

Metodi di riduzione dello spazio di ricerca:

- Blocking riduce lo spazio di ricerca raggruppando i dati in base alle modalità della variabile scelta.

- SimHash è una tecnica di Locality-Sensitive Hashing (LSH), un metodo per ridurre lo spazio vettoriale di un insieme di dati.

Funzioni di confronto:

- Equality è una funzione che controlla l'uguaglianza dei valori confrontati.

- Inclusion3Grams è una funzione che confronta gli n-grams (descritta nel paragrafo 2.6), in questo caso tre.

Variabili impiegate per il match:

- Partita Iva o Codice Fiscale.
- Comune.
- Denominazione struttura.
- Titolare.
- E-mail.

- Nome strada.
- Indirizzo.
- Numero di telefono.

Con la procedura applicata la percentuale di match raggiunta è pari all'85,5%.

I risultati prodotti dall'integrazione dei due dataset amministrativi mettono in luce la presenza di una discrepanza tra le strutture ivi contenute. Di fatto, pur tentando l'impiego di ulteriori passaggi mediante diverse tipologie di funzioni associate anche ad altre variabili, non si riescono a perseguire ulteriori risultati. Il motivo principale risiede nelle carenze circa la qualità dei dati relativamente alle variabili chiave, quali indirizzo, titolare, partita iva o codice fiscale ed indirizzo.

Nella terza integrazione, tra il dataset regionale già integrato con Asia e il dataset ricavato da Tripadvisor, sono state svolte sei fasi di record linkage. In questo caso è stato prediletto unicamente il metodo deterministico. In ogni fase del record linkage è stato utilizzato come metodo di riduzione dello spazio di ricerca il Blocking, inizialmente sulla variabile relativa al comune e successivamente su quella relativa alla provincia. Di seguito si elencano le variabili impiegate nelle fasi di record linkage:

- Comune
- Indirizzo

- Nome strada
- Denominazione

Con la procedura applicata la percentuale di match raggiunta è pari al 47,9%.

Anche in questo caso i risultati mostrano una discrepanza tra le strutture presenti nei due dataset e pur provando altri tentativi non si ottengono concreti risultati. Il motivo principale risulta essere l'assenza di

identificatori univoci tra i dataset ed anche la qualità dei dati gioca il suo ruolo come anche nell'integrazione precedente.

Terminate le procedure di record linkage è stato quindi costruito il dataset integrato tra tutte le fonti, escludendo le strutture presenti univocamente in Tripadvisor. Nel dataset sono state aggiunte variabili dummy per segnare la presenza della struttura nelle fonti.

7. La determinazione dello stato di attività

Sulla base del dataset integrato contenente le informazioni riguardo la presenza nelle diverse fonti impiegate delle strutture ricettive è stato costruito un modello finalizzato alla determinazione del relativo stato di attività. Sono state innanzitutto classificate le fonti in base alla loro attendibilità, il registro Asia è stata considerata la fonte più attendibile proprio per il modo in cui è costruito ed aggiornato. Tra le due fonti regionali è stata considerata maggiormente attendibile Turismo Web per via alla valutazione della qualità. Infine, la fonte meno attendibile considerata è Tripadvisor in quanto più soggetta ad errore.

Alla presenza in ogni fonte è stato quindi associato un valore di probabilità relativo allo stato di attività della struttura. In particolare, alle strutture registrate in Asia 0.45, in Turismo Web 0.25, nel Rilevatore Statistico Regionale 0.20, in Tripadvisor 0.10. Sono state poi definite delle soglie per classificare lo stato delle strutture:

- Valori uguali o superiori a 0.55 -> Eleggibili attive
- Valori compresi tra 0.55 e 0.45 -> Eleggibili
- Valori minori di 0.45 -> Escludibili

La classificazione è stata determinata mediante i seguenti criteri:

- Le strutture presenti in Asia e in una qualsiasi delle altre fonti sono state classificate "Eleggibili attive".
- Le strutture presenti univocamente in Asia o solo nelle due basi di dati regionali sono state classificate "Eleggibili".
- Le restanti strutture sono state classificate "Escludibili".

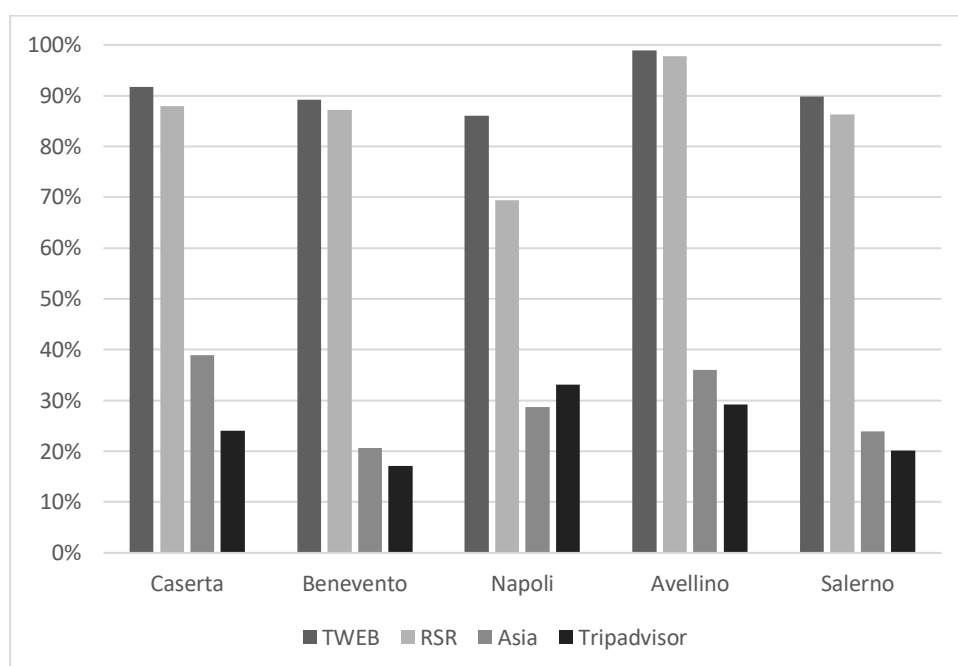
8. Risultati

Le 13.274 strutture contenute nell'archivio della Regione Campania, sono per 67% presenti in entrambe le fonti informative regionali. Quasi tutte quelle della provincia di Avellino (97%) hanno questa caratteristica, per la provincia di Caserta sono l'80%, Benevento e Salerno il 76% e nella provincia di Napoli circa la metà (55%). Il 20% sono solo nella fonte Turismo web e l'11% solo in RSR.

Relativamente alla fonte web, sono state individuate circa la metà delle strutture presenti su TripAdvisor (7.059). Analizzando la distribuzione provinciale, le meno presenti sul web sono quelle della provincia di Benevento (32%) seguite dalla provincia di Salerno (39%). Ben il 73% di quelle della provincia di Napoli sono state registrate sulla piattaforma on-line (Grafico 1).

In merito all'ultima fonte considerata, meno del 30% sono presenti nell'archivio Asia e la gran parte sono allocate nella provincia di Caserta (39%) e in quella di Avellino (36%).

Grafico 1 - Distribuzione delle unità per fonti, livello provinciale.



Fonte: nostra elaborazione su DB integrato

La distribuzione delle probabilità (Tabella 4) vede l'attribuzione alla grande parte delle strutture di una probabilità pari a 0,45 (36%), mentre la probabilità pari a 1 è stata attribuita al 21%. In valore relativo, tale probabilità è stata ottenuta maggiormente nelle provincie di Caserta (34%) e Avellino (35%).

Tabella 4. – Probabilità di eleggibilità per provincia

Province	Probabilità											Totale complessivo
	0,2	0,25	0,3	0,35	0,45	0,55	0,65	0,7	0,75	0,8	1	
Caserta	24	48	14	2	199	59	5	16	4	2	193	566
Benevento	71	63	9	8	374	81	2	25		2	129	764
Napoli	572	906	132	288	1.104	643	10	278	4	94	1.080	5.111
Avellino	3	3	1	3	198	72	1	4			153	438
Salerno	548	570	90	81	2.952	628	7	201	10	43	1.265	6.395
Campania	1.218	1.590	246	382	4.827	1.483	25	524	18	141	2.820	13.274

Fonte: nostra elaborazione su DB integrato

Considerando solo le strutture eleggibili, quindi con probabilità maggiore di 0,45, il dataset integrato si riduce a 9.838 strutture (Tabella 5).

La distribuzione per provincia non si discosta molto da quella iniziale: la maggior parte delle strutture (il 52%) si trova nella provincia di Salerno, seguita dalla provincia di Napoli (37%). Le restanti strutture si dividono tra il 6% della provincia di Benevento e il 4% di Caserta e Avellino.

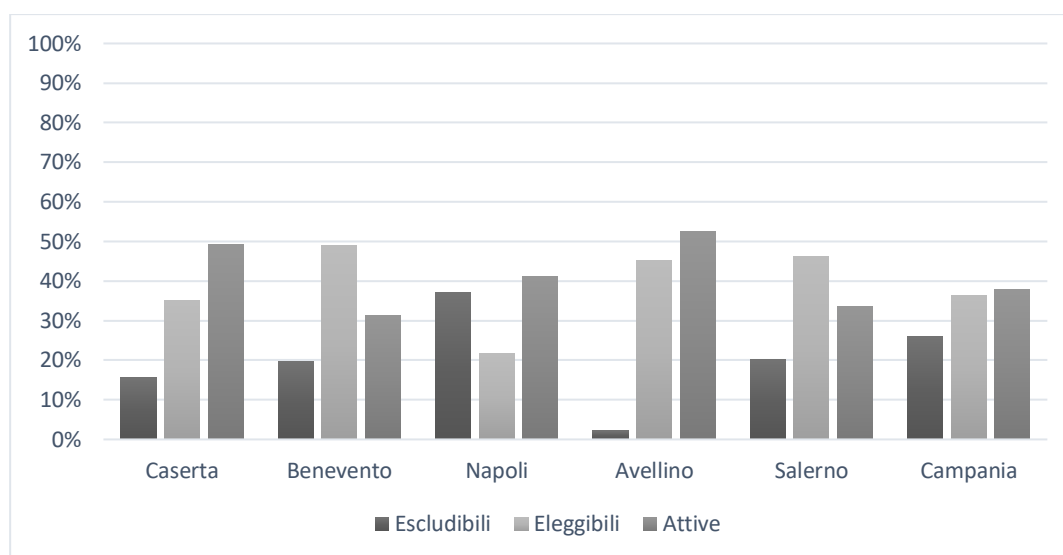
La maggiore quota di strutture eleggibili attive, quindi con una maggiore probabilità di esistenza, rispetto al totale provinciale si trovano a Napoli (66%). La quota minore a Benevento (39%).

Il grafico 2 mostra invece la distribuzione delle categorie per provincia a livello percentuale. Da come si evince, la provincia di Napoli ha il maggior numero di strutture escludibili, la provincia di Avellino il maggior numero di strutture eleggibili attive e quella di Benevento la maggiore incidenza di strutture eleggibili.

La classificazione realizzata mostra come la numerosità delle strutture definite eleggibili od attive si discosti poco dai dati Istat (Tabella 5) ad eccezione della provincia di Salerno dove si evidenzia una forte differenza.

Di seguito (Grafico 2) si mostra il risultato della classificazione delle strutture a livello provinciale.

Grafico 2 - Eleggibilità delle strutture, valori in percentuale, livello provinciale.



Fonte: nostra elaborazione su DB integrato

Tabella 5 - Eleggibilità strutture e confronto con dati Istat, livello provinciale.

Province	Strutture totali	Escludibili	Eleggibili	Attive	Eleggibili + Attive	Dati Istat (2018)
Caserta	566	88	199	279	478	439
Benevento	764	151	374	239	613	635
Napoli	5.111	1.898	1.104	2.109	3.213	3.453
Avellino	438	10	198	230	428	403
Salerno	6.395	1.289	2.952	2.154	5.106	2.255
Campania	13.274	3.436	4.827	5.011	9.838	7.185

Fonte: nostra elaborazione su DB integrato e dati Istat

Considerando come benchmark il dato Istat, lo scostamento percentuale della numerosità di strutture attive è del 36% in più nell'archivio integrato. In particolare la differenza è determinata dalla quantità delle strutture della provincia di Salerno che, seppur ridotta di più di 7.000 unità rispetto al dato iniziale, resta sempre di un valore doppio rispetto a quanto rilevato dall'indagine sulla capacità delle strutture ricettive. Ciò potrebbe dipendere dalla presenza di strutture gestite non in forma imprenditoriale che non fanno parte del campo di osservazione dell'indagine Istat, ma potrebbero essere presenti nel dataset integrato con probabilità pari a 0.45.

9. Conclusioni

Gli archivi della Pubblica Amministrazione costituiscono un patrimonio informativo di ampie dimensioni e possono rappresentare un'importante risorsa per acquisire e fornire le informazioni necessarie per conoscere le realtà territoriali. Tale patrimonio informativo può servire da base condivisa per tutti coloro che partecipano al processo di formazione dei programmi regionali oltre che da sistema omogeneo per la diffusione delle informazioni socio-economiche ufficiali.

Un aspetto fondamentale per il loro utilizzo è la valutazione della qualità degli archivi, poiché c'è da tenere conto di alcuni aspetti quali il lag temporale tra la presentazione degli atti amministrativi (anagrafici o di

impresa) e il loro perfezionamento; la perdita di informazioni che può derivare dal ritardo nell'aggiornamento degli archivi; la presenza di unità non più esistenti o con informazioni obsolete.

Il lavoro svolto mostra in che modo i Big Data possano intervenire nel processo produttivo come fonte secondaria a parziale risoluzione di alcune criticità dei dati amministrativi.

In particolare possono contribuire alla riduzione del lag temporale consentendo l'eliminazione delle unità non più esistenti. Nello specifico, essendo stata utilizzata come variabile secondaria, ha contribuito alla definizione dello stato di attività delle strutture ricettive. Tuttavia, un avanzamento del lavoro potrebbe essere quello di utilizzare i Big Data come fonte principale, a integrazione della fonte amministrativa. Questo consentirebbe di individuare le strutture non ancora registrate a causa dei ritardi nella registrazione degli atti amministrativi da parte della PA.

La tabella 6 mostra alcuni indicatori di qualità calcolati sul database finale, calcolati allo scopo di valutare il risultato ottenuto.

In particolare si notino i valori del tasso di sovra-copertura, che rappresenta la discrepanza tra la lista e la popolazione obiettivo. Infatti, per «sovra-copertura» si intende la presenza nella lista di unità che non appartengono alla popolazione obiettivo. Esso è stato calcolato seguendo le indicazioni Eurostat come rapporto tra:

$$(Unità\ non\ eleggibili + (1-\alpha) \cdot Unità\ non\ risolte) / (Unità\ risolte + Unità\ non\ risolte + \alpha \cdot Unità\ non\ risolte) \cdot 100$$

dove:

Unità risolte: un'unità è risolta se in corso di rilevazione è stato possibile accertare se era eleggibile o meno

Unità non risolte: un'unità è non risolta se in corso di rilevazione non è stato possibile accertare nemmeno se era eleggibile o meno (spesso chiamate anche unità con eleggibilità sconosciuta)

Unità non eleggibili: l'unità non appartiene alla popolazione oggetto di indagine pur essendo presente nell'archivio o lista di estrazione delle unità di rilevazione

α (alpha) = frazione delle unità non risolte che si stimano essere eleggibili. Eurostat raccomanda di porre $\alpha=1$ (Eurostat, 2010).

Considerato che nella costruzione di un registro è considerato accettabile e desiderato l'errore di sovra-copertura, il valore del totale Campania pari a 25,9% rappresenta un valore accettabile. Si consideri che su questo dato influisce il limite del lavoro sperimentale, superabile con la messa a regime, dato dalla diversa data di riferimento delle fonti prese in considerazione.

Come mostrano il tasso di eleggibilità, dato dal numero di strutture eleggibili sul totale, e il tasso di attività, dato dal numero di strutture eleggibili attive sul totale, il lavoro di integrazione dei due archivi amministrativi con Asia e i Big Data, ha consentito di "scartare" il 26% delle strutture iniziali perché considerate con una bassa probabilità di esistenza. Per completare il controllo di qualità sarebbe necessaria un'analisi micro delle strutture non considerate eleggibili, attraverso dei controlli campionari con l'ausilio degli enti provinciali del turismo che nello specifico hanno il ruolo di collettori dei dati.

Tabella 6 – Indicatori di qualità, livello provinciale.

Province	Tasso di sovra-copertura	Tasso di eleggibilità	Tasso di attività
Caserta	15,5%	35,2%	49,3%
Benevento	19,8%	49,0%	31,3%
Napoli	37,1%	21,6%	41,3%

Avellino	2,3%	45,2%	52,5%
Salerno	20,2%	46,2%	33,7%
Campania	25,9%	36,4%	37,8%

Fonte: nostra elaborazione su DB integrato

Bibliografia

Istat (2020), *Rapporto annuale 2020 – la situazione del paese*. Istituto nazionale di statistica Via Cesare Balbo, 16 - Roma

Alleva G. (2017). *Il valore dei dati nell'era dei Big Data*. Intervento del Presidente dell'Istituto nazionale di statistica, Università di Napoli Federico II Dipartimento di Scienze Politiche.

Andrea De Mauro, Marco Greco e Michele Grimaldi, (2016). *A Formal definition of Big Data based on its essential features*, in *Library Review*, vol. 65, n. 3.

Daas P. (2016). *On the quality of registers*. Joint UNECE/Eurostat Expert Group Meeting on Register-Based Censuses. The Hague, The Netherlands, 10-11 May 2010.

Scanu M. (2003). *Metodi statistici per il record linkage*. Sistema statistico nazionale, Istituto nazionale di statistica. Metodi e norme. N.S.; 2003/16.

Belin, Rubin (1995). *A method for calibrating false-match rates in record linkage*. JASA, 694-707.

Jabine T.B., Scheuren F.J. (1986). *Record linkages for statistical purposes: methodological issues*. Journal of Official Statistics, 2, 255-277.

Ivan P. Fellegi, Alan B. Sunter (Dec., 1969). *A theory of record linkage*, Journal of the American Statistical Association: Vol. 64, No. 328, pp. 1183-1210.

ABSTRACT

The local administrative archives constitute an important informative heritage, both to know the territorial realities and to support the political decision-makers.

However, these archives contain errors characteristic of this type of source. So, without precautions, the information contained are not usable. The most common type of errors are the time lag between the date of presentation of the procedures and their reference period, the information loss due to delay in updating the source, the presence of units classify with obsolete classifications.

This study shows how Big Data can intervene in solving and overcoming these limits. For this purpose, two administrative sources of the Campania Region about tourism have been integrated. They contain the list of tourist accommodation structures (Regional Statistical Detector and Web Tourism) and have been compared with the list of structures existing on Tripadvisor web site and with the Business Statistical Register (Asia) developed by Istat. A unique archive has been obtained through statistical integration techniques. The study shows what is the effect of the introduction of the Big Data source in determining the state of activity of the enterprises available in local business registers.