L'ESPERIENZA DELLA CAMPANIA.

MIGLIORARE LE STATISTICHE SUL TURISMO ATTRAVERSO L'USO DEI BIG DATA.

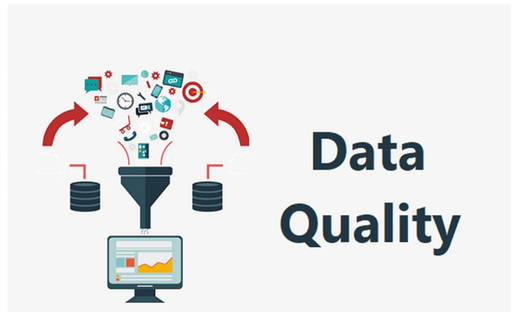
MILeS 2022

MILANO IMPRESA, LAVORO E SOCIETÀ

ANTONELLA BIANCHINO - ISTAT bianchin@istat.it | **DANIELA FUSCO -** ISTAT dafusco@istat.it

OBIETTIVO: MIGLIORARE LA QUALITÀ DEI DATI

- Gli Istituti nazionali di statistica utilizzano ordinariamente i Dati Amministrativi (AD) a fini statistici. Per AD si intendono le informazioni prodotte dalla Pubblica Amministrazione a fini amministrativi che, elaborate da istituti di ricerca, diventano utilizzabili a fini statistici. Il rapido sviluppo dell'Information Technology (IT) delle Pubbliche Amministrazioni (PA) ha reso possibile questo processo di integrazione. Tuttavia, gli errori non campionari caratterizzano gli AD, riducendo il loro potere informativo.
- L'uso delle statistiche multi-fonte e l'introduzione dei Big Data possono migliorare l'accuratezza delle statistiche prodotte.
- Lo studio analizza lo stato di attività delle strutture ricettive della Regione Campania valutato combinando: due fonti AD regionali (Rilevatore Turistico Regionale e Turismo Web), la fonte Big Data TripAdvisor e il Registro statistico delle Imprese attive (Asia).



DATI AMMINISTRATIVI: LIMITI E OPPORTUNITÀ

Opportunità:

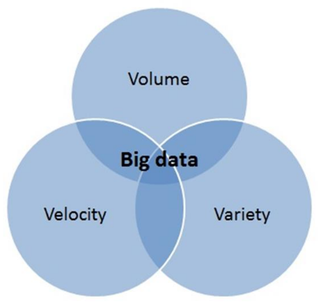
- Ampliamento dei contenuti informativi della produzione statistica.
- Riduzione del disturbo statistico.
- Disponibilità di dati su intere popolazioni.
- Riduzione dei costi.

Limiti:

- Differenza di definizione tra la popolazione target e la popolazione coperta dalla fonte amministrativa.
- Identificazione errata delle unità statistiche.
- Ritardi e/o registrazioni amministrative mancanti.

I tre fattori causano errori di copertura (sopra o sotto).

BIG DATA: LIMITI E OPPORTUNITÀ



I Big Data sono potenzialmente una fonte di dati molto interessante per le statistiche ufficiali, soprattutto in combinazione con fonti di dati più tradizionali come indagini campionarie e registri amministrativi. Tuttavia presentano dei limiti molto affini ai campioni non probabilistici (potrebbero non essere rappresentativi dell'intera popolazione, ma di una parte di essa con specifiche caratteristiche).

FASE 1: VALUTAZIONE DELLA QUALITÀ DEI AD

INDICATORS	*SOURCE*	
	Turismo web	*Rilevatore T. Regionale*
Contact	+	+
Purpose	+/o	+
Respons Burden	+	+
Feedback	+	+
Security	+	+
INDICATORS	*METADATA*	
	Turismo web	*Rilevatore T. Regionale*
Clarity	+	+
Comparability	+/o	+
Unique key	+	+
Data treatment	-	-
INDICATORS	*DATA*	
	Turismo web	*Rilevatore T. Regionale*
Technical check	+	+
Over coverage	+	+
Under coverage	?	?
Data treatment	-	-
Not responding units	+	+
Missing partial response	+/o	o/-
Measurment	+/o	o/-
Sensibility	+	o

La valutazione si basa sul modello sviluppato da Statistical Netherlands, utilizzando specifici indicatori per le tre Hyperdimensions, Fonte, Metadata, Data.

Nella valutazione Hyperdimension Data possiamo notare valutazioni più basse per il Rilevatore Turistico Regionale soprattutto a causa delle variabili di linkage (indirizzo, codice fiscale, partita IVA) che presentano errori e dati mancanti.

Legenda tabella: + Buono, o Ragionevole, - Insufficiente, ? Non chiaro, / Intermedio tra i valori indicati

FASE 2: VALUTAZIONE DEI BIG DATA

NUMERO DI STRUTTURE PER HUB

Province	*Source*					
	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
Total	**7.185**	**7.059**	**13.506**	**5.457**	**14.815**	**12.196**

Per la valutazione degli Hub delle strutture ricettive, è stata considerata solo la copertura. I totali sono stati comparati con i dati ufficiali dell'Istat, utilizzati come benckmark. Come mostra la tabella, TripAdvisor presenta un numero di strutture più simile ai dati Istat.

Nota: dati Istat 2018, dati HUBs 2019

INTEGRAZIONE E STATO DI ATTIVITÀ

Sono stati integrati i due AD regionali, successivamente Asia e infine TripAdvisor.

Pre-processing:

- Codifica di Provincia e Comune.
- Rimozione delle stop word.
- Rimozione delle parole ricorrenti.
- Correzione dei numeri di telefono.
- Eliminazione del rumore.
- Conversione dei DUG.
- Assegnazione di Provincia e Comune mancanti.

Scelta delle variabili di Match:

- Partita Iva o Codice Fiscale.
- Comune.
- Denominazione struttura.
- Titolare.
- Nome strada.
- Indirizzo.
- Numero di telefono.
- E-mail.

Metodi di riduzione dello spazio di ricerca:

- Blocking, raggruppa i dati in base alle modalità della variabile scelta.
- SimHash, Locality-Sensitive Hashing (LSH), metodo per ridurre lo spazio vettoriale di un insieme di dati.

Metodi di integrazione applicati:

- Deterministico con regole.
- Probabilistico.

Alla presenza in ogni fonte è stato 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 di attività** 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".

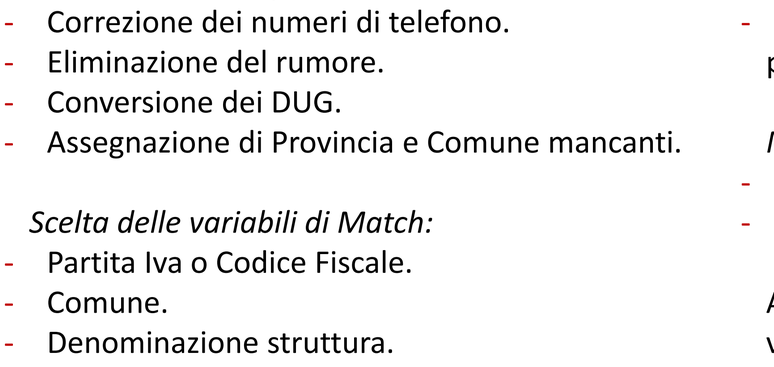
Le restanti strutture sono state classificate "Escludibili".

RISULTATI E CONCLUSIONI

PROBABILITÀ DI ELEGGIBILITÀ PER PROVINCIA

Provincie	Probabilità											Totale
	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**

ELEGGIBILITÀ DELLE STRUTTURE, VALORI IN PERCENTUALE, LIVELLO PROVINCIALE



ELEGGIBILITÀ STRUTTURE E CONFRONTO CON DATI ISTAT, LIVELLO PROVINCIALE

Province	Strutture totali	Escludibili	Eleggibili	Attive	Eleggibili + Attive	Dati Istat
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**

INDICATORI DI QUALITÀ*, LIVELLO PROVINCIALE

Province	Tasso di sovra-copertura	Tasso di eleggibilità	Tasso di attività
Caserta	**15,50%**	35,20%	49,30%
Benevento	**19,80%**	49,00%	31,30%
Napoli	**37,10%**	21,60%	41,30%
Avellino	**2,30%**	45,20%	52,50%
Salerno	**20,20%**	46,20%	33,70%
Campania	**25,90%**	36,40%	37,80%

La distribuzione delle probabilità 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%. Considerando solo le strutture eleggibili, quindi con probabilità maggiore di 0,45, il dataset integrato si riduce a 9.838 strutture

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

Il tasso di copertura regionale è pari a +25.9% (over coverage), rispetto al +75.2% ottenuto utilizzando solo le fonti amministrative.

* Tasso di copertura=$(Unità\ non\ eleggibili+(1-\alpha)\cdot Unità\ non\ risolte)\ /(Unità\ risolte+Unità\ non\ risolte+\ \alpha\cdot Unità\ non\ risolte)\cdot 100$

Istat | Istituto Nazionale di Statistica