

XLI CONFERENZA ITALIANA DI SCIENZE REGIONALI

Web Conference, 2-4 settembre 2020

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

Armando D'Aniello - Univ. Pethenope, Daniela Fusco - Istat, Luigi Praitano – Regione Campania

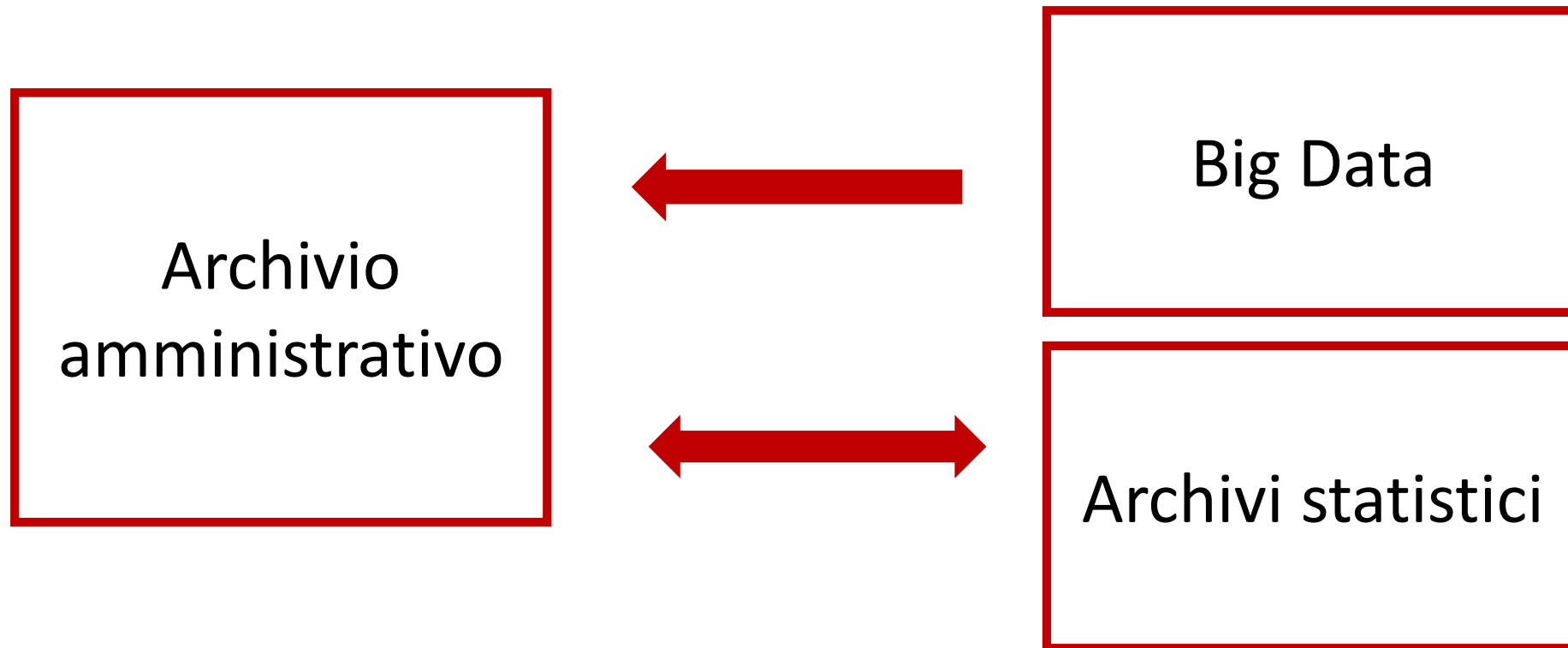
- Gli archivi amministrativi territoriali
- Le fonti utilizzate
- Qualità delle fonti
- Il record linkage per l'integrazione dei dati
- La determinazione dello stato di attività
- Risultati
- Conclusioni

La statistica ufficiale per fronteggiare la crescente richiesta di nuove informazioni, più tempestive ed a maggior dettaglio territoriale, è stata indotta a promuovere sempre più l'uso sistematico delle fonti amministrative a fini statistici.

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, a supporto delle indagini statistiche.

Migliorare la copertura degli archivi della Pubblica Amministrazione



Il **Rilevatore Turistico Regionale** della 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.

Il popolamento del Rilevatore avviene per diretto contatto del titolare della struttura con l'Ufficio regionale.

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.



Volume – Velocità – Varietà

Pregi:

- Aumento di accuratezza, coerenza e completezza delle informazioni statistiche prodotte
- Ampliamento dei contenuti informativi della produzione statistica
- Riduzione del disturbo statistico
- Copertura totale delle popolazioni di riferimento delle statistiche, consentendo di aumentare il dettaglio territoriale di riferimento delle informazioni prodotte
- Riduzione dei costi.

Difetti:

- La popolazione statistica obiettivo può essere diversa
- Definizioni e classificazioni spesso non coincidono con quelli della statistica ufficiale.
- L'accesso ai dati può essere problematico.
- 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.

Tripadvisor® è tra le piattaforme di viaggi più grande del mondo.

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.



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:

- 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 scorpori, fusioni o procedure concorsuali, fallimenti, liquidazioni, eccetera).

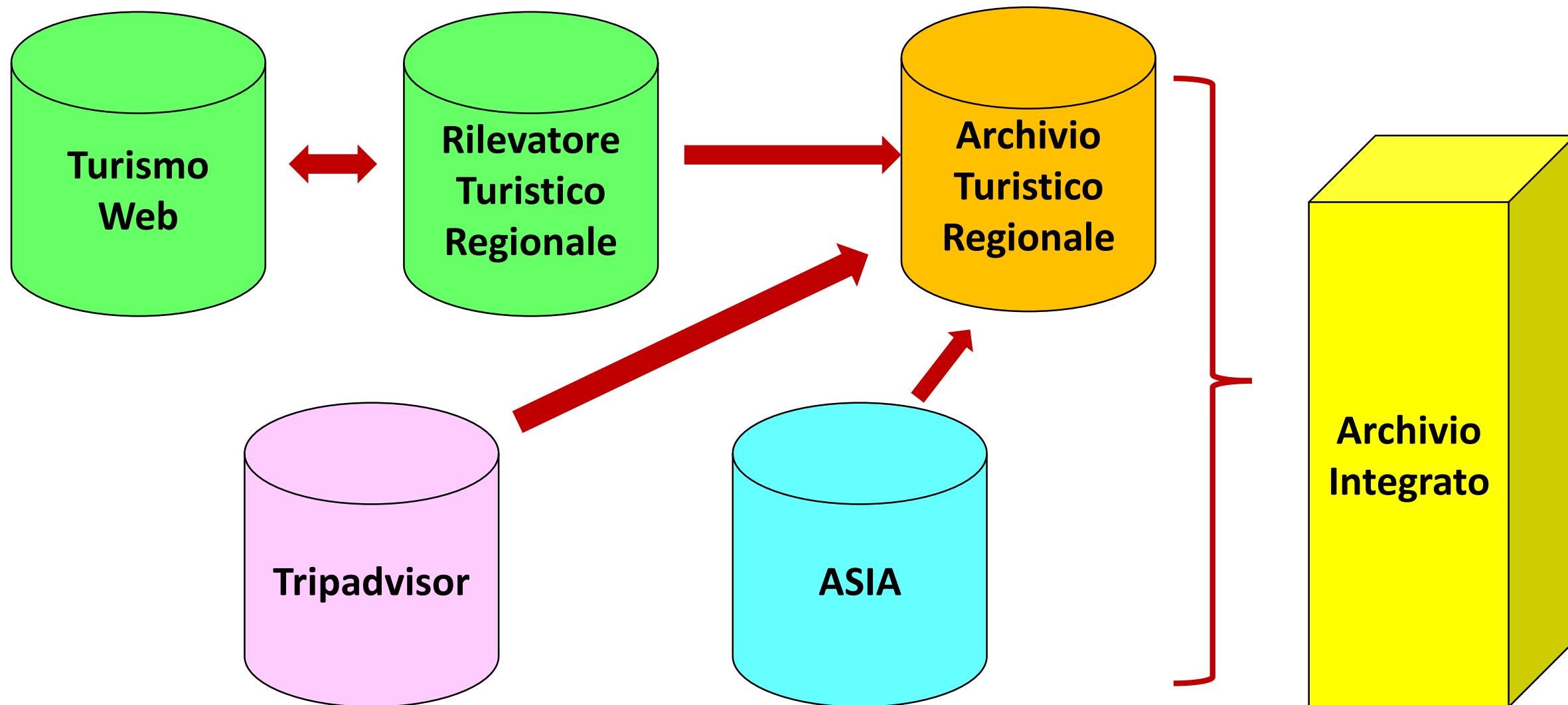
La qualità delle fonti amministrative

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

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

Il record linkage per l'integrazione tra le fonti



Pre-processamento

- Conversione di provincia e comune nei relativi codici Istat.
- Omogeneizzazione degli indirizzi delle strutture attraverso la conversione delle più svariate abbreviazioni presenti nei dati
- Rimozione di parole ridondanti che causano l'abbinamento di falsi match durante l'applicazione degli algoritmi di linkage (i.e.: "HOTEL", "B&B", etc.).
- Correzione dei numeri di telefono delle strutture
- Rimozione di spazi e caratteri speciali
- 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.
- Aggiunta di una variabile contenente la classificazione Istat delle strutture (alberghiere ed extra alberghiere) ottenuta dalla tipologia
- Aggiunta di 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.

13

La scelta dell'algoritmo

- **Fase UNO:** Turismo web + RTR - deterministico con regole e probabilistico;
- **Fase DUE:** Archivio regionale + ASIA – deterministico (merge);
- **Fase TRE:** Tripadvisor – deterministico con soglie alte.

Variabili di match:

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

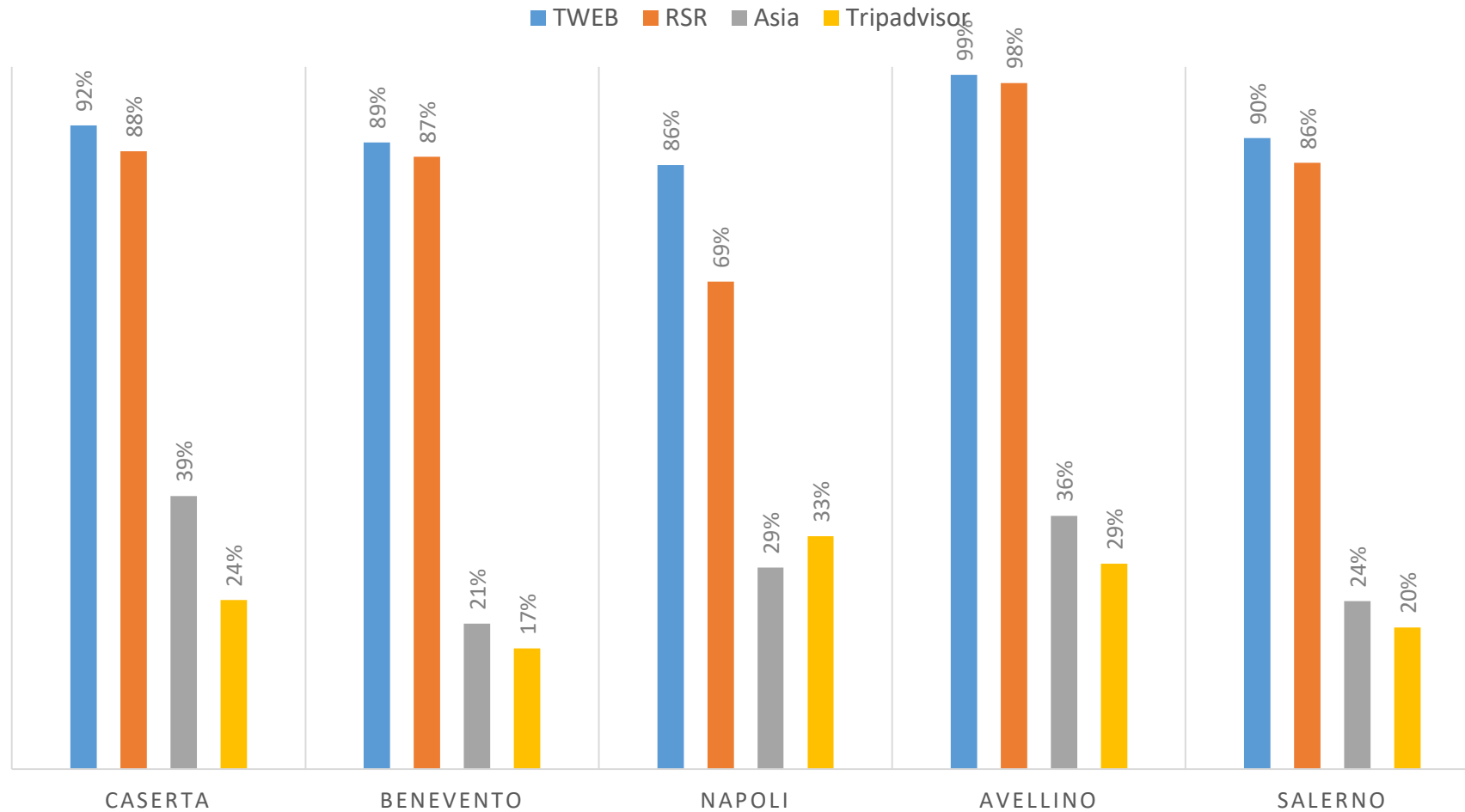
Probabilità attribuite:

- ✓ Asia 0.45
- ✓ Turismo Web 0.25
- ✓ Rilevatore Turistico Regionale 0.20
- ✓ Tripadvisor 0.10

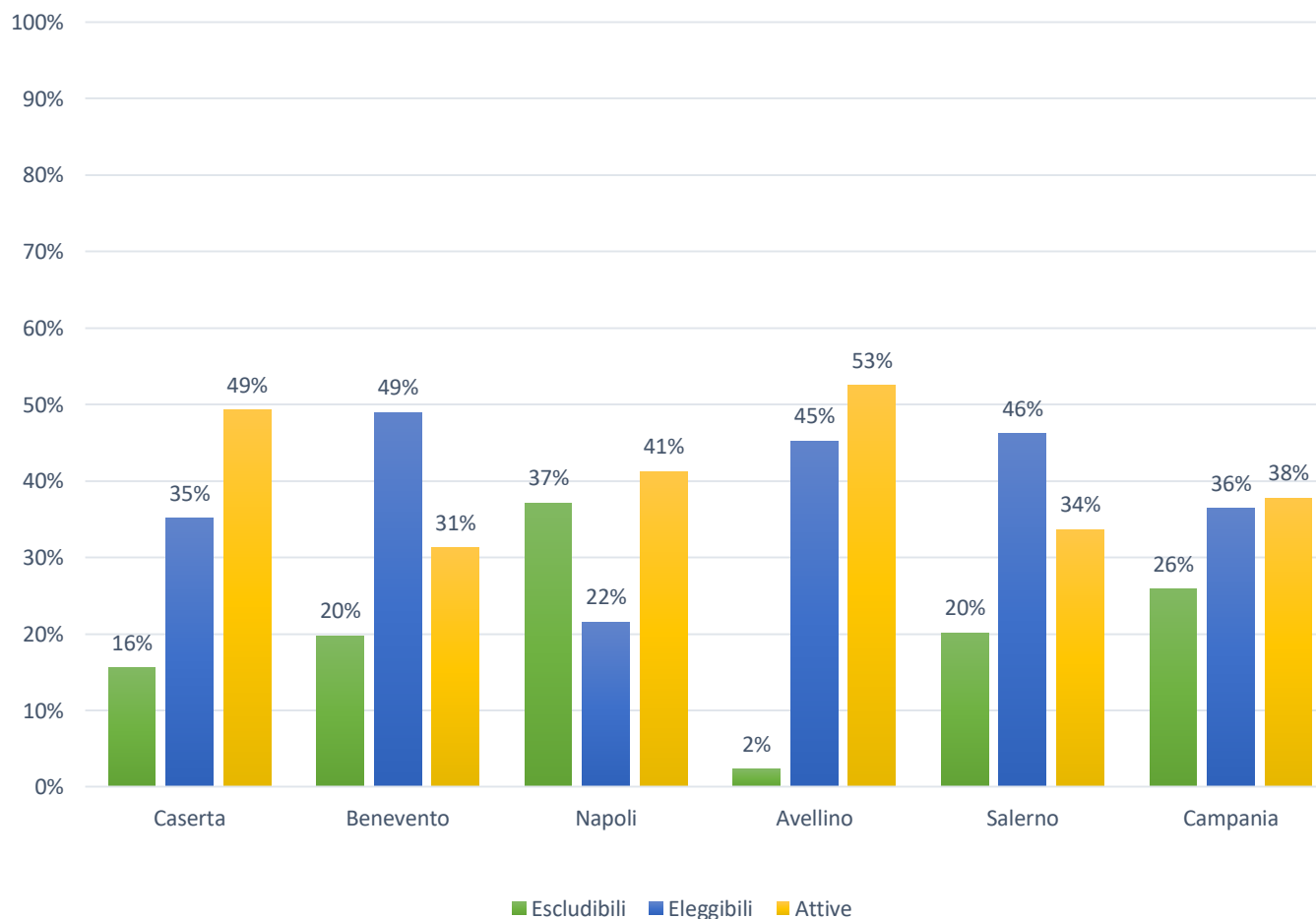
Soglie di inclusione:

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

Distribuzione delle unità per fonti nell'archivio integrato, livello provinciale



Eleggibilità delle strutture nell'archivio integrato, valori in percentuale, livello provinciale



Eleggibilità strutture nell'archivio integrato 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

Indicatori di qualità dell'archivio integrato, 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%

Grazie per l'attenzione...

A hand-drawn orange smiley face with two vertical lines for eyes and a curved line for a mouth, positioned to the right of the text "Grazie per l'attenzione..."