Regione Siciliana
Assessorato alla Salute
Dipartimento per le Attività Sanitarie
ed Osservatorio Epidemiologico

Soluzioni operative ed aspetti qualitativi per la geo-codifica di popolazioni in ambito sanitario

Sebastiano Pollina A.[3], Paolo Carnà[1], Mauro Ferrante[2], Alfredo Pontillo[1], Alessandro Migliardi[1], Salvatore Scondotto[3]

[1] Servizio Sovrazonale di Epidemiologia, ASL TO3
[2] Università di Palermo
[3] Dipartimento Attività Sanitarie ed Osservatorio Epidemiologico, Regione Siciliana

Schema

- Geocodifica nella ricerca sanitaria
- Principali obiettivi
- Materiali e metodi
- Risultati
- Prospettive future e conclusioni

Le geocodifica nella ricerca sanitaria

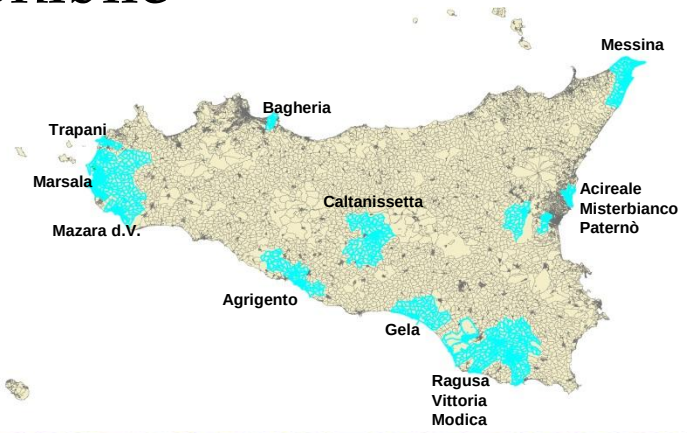
- Il *geocoding* è la trasformazione di un indirizzo testuale in informazioni geografiche
  - consente il posizionamento di un individuo, gruppo e servizio sanitario in un contesto geografico
- Rende possibile effettuare analisi statistiche spaziali circa:
  - Cause di patologia
  - Adeguatezza ed equità nell'accesso ai servizi
  - Sorveglianza Epidemiologica
  - Sviluppo degli obiettivi delle politiche di prevenzione

Obiettivi

- Proporre un indice di qualità delle procedure di geocoding utilizzate:
  - per valutare la qualità dei risultati
  - per tenerne conto nelle fasi di analisi (ad esempio attraverso le stratificazione)

Materiali e Metodi

- Per la Geocodifica sono stati selezionati gli indirizzi residenziali di 13 Comuni Siciliani con popolazione superiore ai 50.000 abitanti
  - I comuni di Palermo, Catania e Siracusa sono stati esclusi dall'analisi
  - Per il comune di Messina la toponomastica ISTAT degli indirizzi non è disponibile



Materiali e Metodi

- La procedura di geocodifica è stata effettuata in tre fasi:
  - Normalizzazione degli indirizzi
  - Linkage con lo stradario ISTAT
  - La localizzazione

Materiali e Metodi

- Sono stati usati due differenti approcci
- Normalizzazione
  - standardizzazione basata su GIS
  - Sviluppo di algoritmi basati su ricorrenza di espressioni regolari
- Linkage
  - Gli indirizzi delle residenza sono stati linkati con la toponomastica ufficiale dello stradario ISTAT
    - La prima procedure è un Matching probabilistico basato sul punteggio di similarità di Jaccard
    - Per la seconda procedura è stato implementato un apposito algoritmo basato su una ricerca Boolean fulltext
- Localizzazione
  - Per la prima procedura è stato usato il software ArcGIS
  - Per la seconda procedura si utilizzato un servizio web server (Google geocoding API)
- È stato implementato un Indice di qualità del Geocoding
- I risultati ottenuti dai due metodi sono stati poi confrontati in termini di distanza tra le coppie di punti trovati con i due sistemi

Qualità dei dati geocodificati

- Match Rate
  - percentuale di record che possono essere geocodificati con un sufficiente livello di qualità
- Match score
  - la confidenza associata ad un particole record descrive il livello di affidabilità su quanto l'indirizzo in input è associato al corretto output in termini di geocoding
- Match type
  - quindi il livello di match geografico
- Spatial accuracy
  - una misura di quanto le coordinate spaziali ottenute corrispondono al luogo reale che si intendeva geocodificare

Geocoding Quality Index (GQI)

$$J(A,B) = \frac{A \cap B}{A \cup B}$$

Jaccard similarity score	*ArcGIS Geocoding*	
	Bad	**Good**
0.00-0.49	Low	Low
0.50-0.81	Low	Medium
0.82-0.99	Low	High
1.00	Low	High

- `google.maps.GeocoderLocationType.ROOFTOP` indicates that the returned result reflects a precise geocode.
- `google.maps.GeocoderLocationType.RANGE_INTERPOLATED` indicates that the returned result reflects an approximation (usually on a road) interpolated between two precise points (such as intersections). Interpolated results are generally returned when rooftop geocodes are unavailable for a street address.
- `google.maps.GeocoderLocationType.GEOMETRIC_CENTER` indicates that the returned result is the geometric center of a result such as a polyline (for example, a street) or polygon (region).
- `google.maps.GeocoderLocationType.APPROXIMATE` indicates that the returned result is approximate.

Boolean full text search score	*Google geocoding API*			
	Rooftop	**Range Interpolated**	**Geometric Center/ Approximate**	**Invalid Address**
0.00-0.32	Low	Low	Low	Low
0.33-0.67	Medium	Medium	Low	Low
0.68-0.99	High	Medium	Low	Low
1.00	High	High	Medium	Low

Risultati

- In totale sono stati selezionati 263.837 indirizzi univoci di residenza associati ad 1.112.014 abitanti (circa 22% dei residenti in Sicilia)

Jaccard's Similarity score	ArcGIS Geocoding		
	Bad	Good	Total
0.00-0.49	73770	0	73770
0.50-0.81	2849	2268	5117
0.82-0.99	39990	44465	84455
1.00	36933	63562	100495
Total	153542	110295	263837

Municipality	N. of unique addresses	%
ACIREALE	16759	6.35%
AGRIGENTO	16644	6.31%
BAGHERIA	21857	8.28%
CALTANISSETTA	13009	4.93%
GELA	29283	11.10%
MARSALA	31164	11.81%
MAZARA DEL VALLO	18048	6.84%
MISTERBIANCO	11223	4.25%
MODICA	17202	6.52%
PATERNÒ	21656	8.21%
RAGUSA	22175	8.40%
TRAPANI	19761	7.49%
VITTORIA	25056	9.50%
Total	**263837**	**100.00%**

Boolean fulltext search score	Google geocoding API				
	Rooftop	Range Interpolated	Geometric Center/ Approximate	Invalid address	Total
0.00-0.32	792	399	555	18132	19878
0.33-0.67	24107	14294	17852	112	56365
0.68-0.99	382	261	1208	13	1864
1.00	110659	47198	27353	520	185730
Total	135940	62152	46968	18777	263837

Risultati

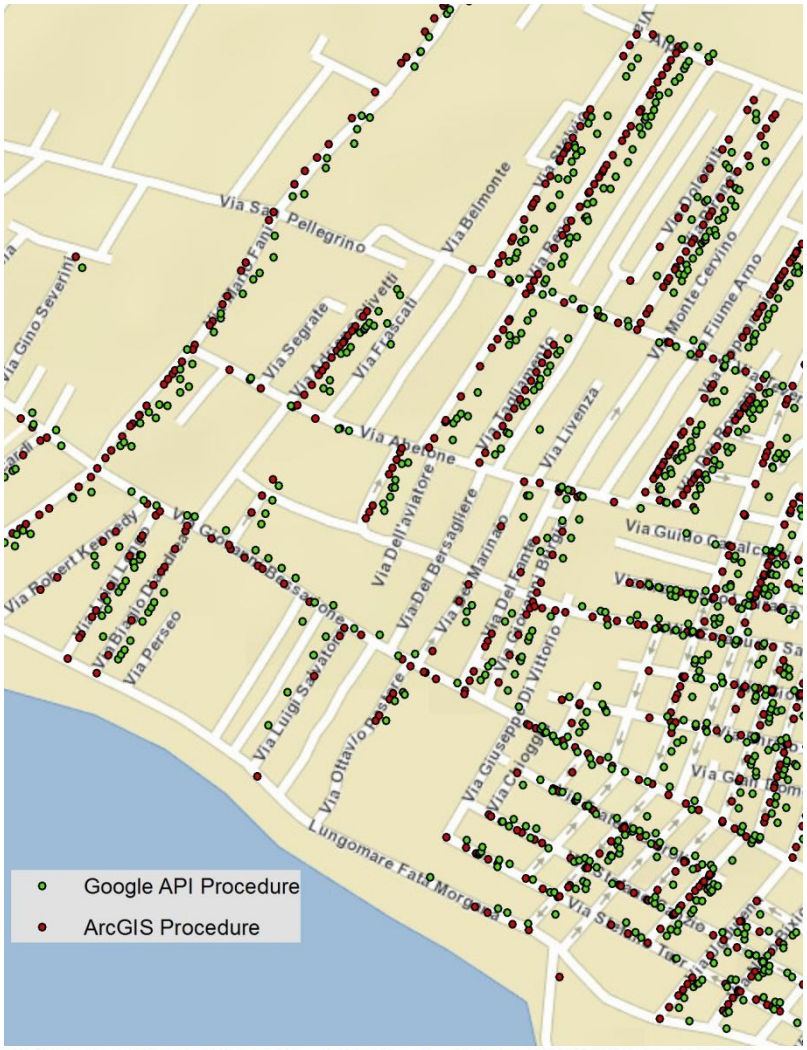
- La% degli indirizzi con qualità medio-alta con almeno una delle due procedure è superiore al 80%
- La tecnica basata su API google è migliore di quella basate su ARC-GIS (75% vs 41%)

	Google procedure GQI			
ArcGIS procedure GQI	Low	Medium	High	Total
	%	%	%	%
Low	51185	54287	48070	153542
%	*19.40*	*20.58*	*18.22*	*58.20*
Medium	949	947	372	2268
%	*0.36*	*0.36*	*0.14*	*0.86*
High	14802	30626	62599	108027
%	*5.61*	*11.61*	*23.73*	*40.94*
Total	66936	85860	111041	263837
%	*25.37*	*32.54*	*42.09*	*100.00*

Risultati

- 90% dei risultati con qualità medio-alta ha una distanza inferiore a 600 m
- 10% degli outliers è stato considerato come bassa qualità (in questa fase)

		Google-based procedure GQI	
ArcGIS-based procedure GQI		Medium	High
Medium	Distance (m)	80.82	106.9
	Std. dev.	*116.83*	*128.97*
High	Distance (m)	95.55	127.92
	Std. dev.	*33.42*	*51.06*



Prospettive future e conclusioni

- Determinare la sezione censuaria per ogni indirizzo geocodificato
- Possibilità di compiere analisi aggiungendo le informazioni dell'unità censuaria inclusi l'indice di deprivazione
- Aumentare la qualità dei risultati, però lavorando contemporaneamente al database dei residenti e alla toponomastica degli indirizzi ISTAT
- Implementare procedure di analisi epidemiologica capaci di tener conto di eventuali errori sistematici derivanti dal geocoding incompleto
- Sviluppare ricerche su base spaziale che considerino la relazione per effetti ambientali e condizioni socio-economiche per i decisori dell'assistenza sanitaria

Gli autori dichiarano di non avere conflitti di interesse

Grazie per l'attenzione

sebastiano.pollinaaddario@regione.sicilia.it