

Garantire la riservatezza dei dati sanitari geocodificati: valutazione delle strategie di geomasking per i dati a livello individuale

Giorgia Stoppa¹

¹ Unità di Biostatistica, Epidemiologia e Sanità Pubblica. Dipartimento di scienze cardio-toraco vascolari e Sanità pubblica. Università di Padova. Via Loredan 18 35031 Padova (IT),
giorgia.stoppa@phd.unipd.it

I sistemi informativi geografici (GIS), ampiamente utilizzati negli studi di epidemiologia, soprattutto in epidemiologia ambientale, sono ad oggi diventati strumenti essenziali per la conduzione di indagini e ricerche. Nel contesto epidemiologico, quindi, risultano di fondamentale importanza le procedure di georeferenziazione (o geocoding), ovvero il processo attraverso il quale le descrizioni testuali di una localizzazione geografica (indirizzo di residenza, sezione di censimento, CAP) vengono trasformate in dati spaziali digitalizzati. In particolare, negli studi a livello individuale si parla di address geocoding, ossia l'assegnazione delle coordinate geografiche agli indirizzi di residenza. Le coordinate possono poi essere utilizzate, ad esempio, per indagare le relazioni tra eventi sanitari e altri fattori che variano nello spazio, come ad esempio le concentrazioni di inquinanti [1,2].

Tuttavia, le operazioni di geocoding possono, però, introdurre distorsioni ed errori che sono stati ampiamente indagati dalla letteratura internazionale [3-5]. L'errore posizionale può influenzare, ad esempio, le stime dei tassi di malattia, le stime di esposizione e le statistiche per l'identificazione dei cluster [6-8]. Ma ben più importanti sono i problemi di riservatezza dei dati in quanto si tratta di informazioni personali, tali cioè da permettere l'identificazione del soggetto. Per impedire tale identificazione si ricorre al cosiddetto geomasking, cioè al mascheramento della reale posizione geografica del soggetto e che nello stesso tempo mantenga le proprietà geografiche dei dati stessi. Il geomasking si ottiene aggiungendo un rumore stocastico o deterministico alle coordinate spaziali. Sono state sviluppate numerose tecniche di geomasking, ad esempio: Random direction and fixed radius; random perturbation within a circle; Gaussian displacement; donut masking; Bimodal Gaussian Displacement. Tuttavia, tecniche quali donut geomasking e binomial gaussian displacement non vengono molto utilizzate nella pratica in quanto introducono una distorsione eccessiva senza una riduzione significativa del rischio di identificazione [9]. In questo lavoro dopo aver passato in rassegna il problema del mascheramento della posizione geografica sarà presentata un'applicazione in epidemiologia ambientale.

Parte del lavoro è stato supportato dal Master in geostatistica per la salute dell'uomo, degli animali e dell'ambiente dell'Università degli Studi di Padova (<https://www.unipd.it/corsi-master/geostatistica-salute>).

Riferimenti bibliografici

1. Rushton G, Armstrong MP, Gittler J, Greene BR, Pavlik CE, West MM, Zimmerman DL: Geocoding in cancer research: a review, *Am J Prev Med*, 30(2): S16–S24 (2006).
2. Nuckols JR, Ward MH, Jarup L.: Using geographic information systems for exposure assessment in environmental epidemiology studies, *Environ Health Perspect*, 112(9): 1007-1015 (2004).
3. Zandbergen PA.: Geocoding quality and implications for spatial analysis, *Geogr Compass*, 3(2): 647–680 (2009)
4. Schootman M, Sterling DA, Struthers J, Yan Y, Laboube T, Emo B, Higgs G.: Positional accuracy and geographic bias of four methods of geocoding in epidemiologic research, *Ann Epidemiol*, 17(6): 379–387 (2007).
5. Goldberg DW, Ballard M, Boyd JH, Mullan N, Garfield C, Rosman D, Ferrante AM, Semmens JB.: An evaluation framework for comparing geocoding systems, *Int J Health Geogr*, 12: 50 (2013)
6. Goldberg, D. W. and Cockburn, M. G. (2012) The effect of administrative boundaries and geocoding error on cancer rates in california. *Spat. Spatiotemporal Epidemiol.* 3(1), 39–54 (2012).
7. Mazumdar, S., Rushton, G., Smith, B. J., Zimmerman, D. L. and Donham, K. J. : Geocoding accuracy and the recovery of relationships between environmental exposures and health. *Int. J. Health Geogr.* 7, 13 (2008).
8. Jacquez, G. M. and Waller, L. A. The effect of uncertain locations on disease cluster statistics. *Quantifying spatial uncertainty in natural resources: Theory and applications for GIS and remote sensing* pp. 53–64 (2000).
9. Zandbergen, P. A.: Ensuring confidentiality of geocoded health data: Assessing geographic masking strategies for Individual-Level data. *Adv Med*, 2014:567049 (2014)