GIS e salute: qual è l'impatto della precisione della georeferenziazione negli studi di epidemiologia ambientale. Due casi studio in Toscana

Daniela Nuvolone, Marco Santini, Pasquale Pepe,
Fabio Voller, Francesco Cipriani

Agenzia regionale di sanità della Toscana, Via Pietro Dazzi 1, 055462431, 0554624330, info@ars.toscana.it

Riassunto

I GIS sono strumenti ampiamente diffusi negli studi di epidemiologia ambientale per la georeferenziazione degli indirizzi di residenza delle persone e per la valutazione dell'esposizione e delle correlazioni con gli esiti sanitari.. Nonostante ciò, le questioni inerenti la qualità dei dati cartografici di base sono scarsamente dibattute nel mondo dell'epidemiologia italiana. Al fine di valutare la precisione della georeferenziazione e l'impatto che questa può avere in studi di epidemiologia ambientale sono stati condotti due casi studio in Regione Toscana, uno nell'area montana di Piancastagnaio (Siena) e l'altro nell'area urbana intorno l'aeroporto di Firenze. Sono stati confrontati i risultati dell'address geocoding effettuato con tre metodi: il grafo prodotto dalla Regione Toscana e i due applicativi commerciali di Google e Bing-Microsoft. Sono stati testati 1549 indirizzi a Piancastagnaio e 3319 indirizzi dell'area di Firenze intorno l'aeroporto. La banca dati regionale mostra performance migliori rispetto ai due applicativi, con differenze più marcate nel comune di Piancastagnaio. Ciò comporta impatti significativi nella valutazione dell'esposizione individuale. Lo studio mostra come la precisione delle operazioni di geocoding abbia un impatto significativo negli studi "ambiente e salute", che impone la conduzione di approfondimenti sulla qualità dei dati cartografici di base.

Abstract

GIS are widely used in environmental epidemiology studies to locate study population by geocoding addresses and to evaluate exposures and relationship with health outcomes. Despite this, quality of geocoding results is poorly discussed by Italian environmental epidemiologists. To evaluate quality of geocoding process and its impacts on evaluation of exposure two case studies have been carried out: one in the mountain area in the municipality of Piancastagnaio and one in the urban area in Florence. Three geocoding systems have been compared: the geographic data base produced by Tuscany Region and two commercial systems (Google and Bing-Microsoft). 1549 addresses in Piancastagnaio and 3319 addresses in Florence have been tested. Tuscany geographical database showed better performance than the two commercial systems, with bigger differences in Piancastagnaio. This implied misclassifications in the evaluation of individual exposures. The study highlighted the impacts of geocoding process in environmental health research and pointed out the need of specifically evaluate the quality of cartographic data.

Introduzione

I sistemi informativi geografici sono da tempo ampiamente utilizzati negli studi di epidemiologia, soprattutto in epidemiologia ambientale, fino ad essere diventati strumenti di fatto irrinunciabili per la conduzione di indagini e ricerche su *environmental health.*

Nel contesto epidemiologico, alla base delle applicazioni GIS vi sono 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. Il caso più comune in epidemiologia ambientale è l'*address geocoding*, ossia l'assegnazione delle coordinate geografiche agli indirizzi di residenza, espressi come nome del toponimo stradale e numero civico. I risultati della georeferenziazione possono essere utilizzati ad esempio per associare l'informazione individuale a livello di indirizzo con le informazioni sanitarie o socio-economiche a livello di micro-area (ad esempio le sezioni di censimento) o per indagare le relazioni tra eventi sanitari e altri fattori che variano nello spazio, come ad esempio le concentrazioni di inquinanti in aria, la distanza da sorgenti di inquinamento o la vicinanza a presidi e servizi sanitari (Rushton et al., 2006, Nuckols et al., 2004).
Le operazioni di geocoding possono, però, introdurre bias ed errori che sono stati ampiamente indagati dalla letteratura internazionale (Zandbergen, 2009, Schootman et al., 2007, Goldberg et al., 2013, Goldberg, Cockburn, 2010, Whitsel et al., 2006, Ward et al., 2005). La qualità della georeferenziazione può essere caratterizzata secondo tre componenti: la completezza, ovvero la percentuale di records di input che vengono effettivamente georeferenziati (anche chiamata *match rate*); l'accuratezza o precisione, ovvero quanto il punto georeferenziato è vicino alla localizzazione "vera" dell'indirizzo (*positional error*); e la ripetibilità, cioè quanto i risultati della georeferenziazione sono sensibili a variazioni del dataset di riferimento, agli algoritmi di *matching*, alle abilità e interpretazione del ricercatore (Lovasi et al., 2007, Zandbergen, 2008). Vari studi internazionali sono stati condotti per indagare e confrontare le performance dei vari sistemi di geocoding che in questi anni si sono resi disponibili grazie ad una crescente produzione e condivisione di informazione geografica (Goldberg et al., 2013, Zhan et al., 2006, Oliver et al., 2005, Mazumdar et al., 2008). Se da una parte i ricercatori internazionali, prevalentemente statunitensi, hanno, quindi, da tempo dedicato risorse alla ricerca e pubblicazione di studi sulla qualità del processo di georeferenziazione e sugli impatti e ricadute che questa ha negli studi di epidemiologia ambientale, dall'altra tale argomento ha invece ricevuto scarsa attenzione da parte dell'epidemiologia italiana.
L'obiettivo del presente studio, oltre a richiamare l'interesse sull'importanza della valutazione della qualità dei risultati dei processi di *address geocoding*, è valutare i possibili impatti che l'uso di sistemi di *geocoding* diversi hanno negli studi di epidemiologia ambientale. Sono presentati due casi studio condotti in Toscana.

Materiali e metodi
Sono state indagate due aree in regione Toscana: il comune di Piancastagnaio, nella zona montana dell'Amiata senese, dove sono presenti centrali Enel per la produzione di energia elettrica da fonte geotermica, e l'area urbana nelle vicinanze dell'aeroporto di Firenze. In entrambe le aree sono in corso studi di epidemiologia ambientale per valutare l'impatto che queste attività possono avere sulla salute. Per la georeferenziazione degli indirizzi di residenza delle coorti di popolazione in studio sono state utilizzate e confrontate tre banche di dati cartografici: il grafo stradale regionale prodotto dal "Settore Pianificazione Integrata della Mobilità e dei Trasporti e Sistema Informativo della Mobilità" della Regione Toscana, e due applicativi commerciali, entrambi ad accesso gratuito, Google e Bing-Microsoft. Il grafo regionale toscano è il frutto di pluriennali investimenti da parte della Regione Toscana finalizzati alla collaborazione con gli enti comunali per la produzione, manutenzione e aggiornamento dell'informazione geografica regionale. Il grafo comprende i dati cartografici relativi a elementi stradali, giunzioni, toponimi, estese amministrative, accessi, civici, cippi chilometrici. Google mette a disposizione un web service dedicato all'*address geocoding*. La Google Geocoding API fornisce gli strumenti per accedere a questo servizio attraverso interrogazioni HTTP. Microsoft permette l'address geocoding attraverso un servizio del tutto analogo a quello di Google, il Bing Maps REST Service. Entrambi gli strumenti permettono l'uso gratuito entro certi limiti e all'interno dei rispettivi termini di servizio. L'output è rappresentato da file XML o JSON che contengono l'indirizzo normalizzato, le informazioni geografiche, il grado di precisione della georeferenziazione e vari codici di errore. A differenza della banca dati regionale la

vocazione di questi servizi è planetaria e, con precisioni ovviamente variabili, permettono la geocodifica di sempre più indirizzi ovunque nel mondo. Tutti e tre i sistemi di georeferenziazione presentano nel file di output alcune variabili relative al tipo di *geocoding* effettuato, che consentono una valutazione della qualità e della precisione secondo scale predefinite. I risultati della georeferenziazione degli indirizzi ottenuti con ciascuno dei tre sistemi sono stati valutati mediante sovrapposizione con ortofoto digitali e carta tecnica regionale. Complessivamente sono stati testati 1549 indirizzi nell'area montana di Piancastagnaio e 3319 indirizzi dell'area intorno l'aeroporto di Firenze.

Risultati

Caso studio comune di Piancastagnaio

Relativamente ai 1549 indirizzi del comune di Piancastagnaio, la banca dati regionale presenta un match rate del 100%: per il 70,1% degli indirizzi in input vi è stato un riconoscimento preciso del numero civico, per l'8% è stata georeferenziata la località, e per il restante 21,8% il punto è stato georeferenziato al civico più vicino. In quest'ultimo caso il sistema regionale fornisce una indicazione della distanza numerica tra indirizzo in input e indirizzo georeferenziato: se ad esempio per "Via Roma 4" non viene trovato il civico 4 e il più vicino risulta "Via Roma 14", il sistema riporta come distanza +10. L'applicativo di Google scarta un solo indirizzo: il 60,5% degli indirizzi in input presenta un matching parziale tra indirizzo testuale in input e la formattazione da parte dell'algoritmo di geocoding. In questa casistica rientrano svariate situazioni, più o meno rilevanti dal punto di vista della qualità della georeferenziazione: si va dal non riconoscimento del numero civico o dell'esponente fino alla modifica sostanziale dell'indirizzo. Dei 1548 indirizzi georeferenziati il 39% riporta come metodo di *geocoding* la dicitura "*rooftop*" ovvero vi è stato il riconoscimento sul numero civico, il 33% riporta "*range interpolated*" che utilizza il metodo dell'interpolazione lungo i tratti stradali, il 22% riporta come metodo di *geocoding* la dicitura "*approximate*" in cui sostanzialmente l'indirizzo non viene riconosciuto e viene localizzato nel centro del comune (è il caso delle località sparse fuori dal centro cittadino), ed il 6% riporta "*geometric center*", che sembrerebbe anch'essa un interpolazione ma meno precisa rispetto a quella effettuta lungo i segmenti stradali. L'applicativo Bing di Microsoft ha un match rate di circa il 98% ma solo per 16 indirizzi (1%) riesce a riconoscere il numero civico; per il resto la georeferenziazione degli indirizzi in input è effettuata sul centro del comune o interpolando su pochi principali assi stradali. Pertanto il *positional error* di Bing è talmente elevato per il caso studio di Piancastagnaio che il confronto è stato effettuato solo tra la banca dati regionale e l'applicativo di Google. Se si sovrappongono i risultati della georeferenziazione con i due metodi e l'ortofoto della Regione Toscana (figura 1), si nota immediatamente la diversa estensione dei territori coperti, sostanzialmente determinata dal fatto che la banca dati regionale riesce a interpolare in maniera più precisa le località sparse fuori dal centro abitato, che invece Google posiziona tutte nello stesso punto, ovvero il centroide del comune. Chiaramente in un contesto montano come quello di Piancastagnaio la presenza di molte case sparse ha un influenza importante sulla qualità della geocodifica.

Anche ad una analisi di maggiore dettaglio, effettuata mediante sovrapposizione con reticolo stradale, carta tecnica regionale, ed ispezione visiva della successione dei numeri civici sullo stesso asse stradale, l'*address geocoding* eseguito con la banca dati regionale mostra performance migliori rispetto all'applicativo di Google. Tale considerazione vale anche quando si confrontano i migliori risultati dell'applicativo di Google, laddove per migliori si intendono quei punti con matching totale tra indirizzo di input e formattazione dell'algoritmo e con metodo di geocoding definito "*rooftop*", con i migliori della banca dati regionale, ovvero quelli riconosciuti esattamente sull'edificio.

In tabella 1 sono riportate le descrittive della differenza in metri tra i risultati della banca dati regionale, considerato il gold standard, con quelli dell'applicativo di Google. La distanza media tra le coppie di punti è 635 m nell'insieme degli indirizzi di input. Si nota come le differenze aumentino nel sottogruppo delle case sparse, mentre diminuiscono nel sottogruppo dei punti

"migliori" del sistema di *geocoding* regionale ed in maniera ancora più significativa nel sottogruppo dei "migliori" dell'applicativo di Google.

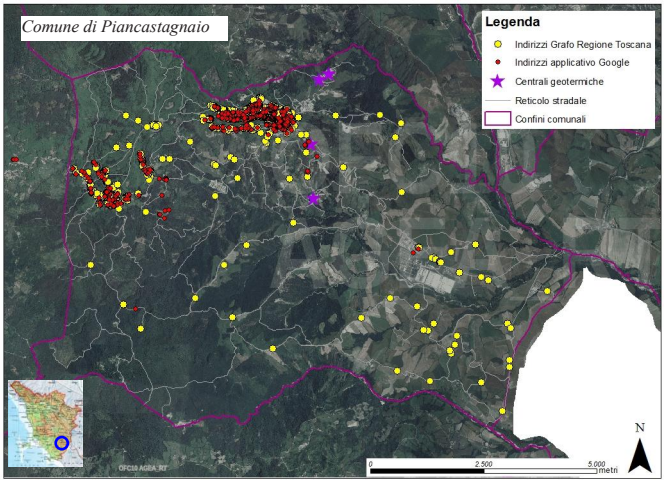

Figura 1 - Geocoding degli indirizzi nel comune di Piancastagnaio: confronto tra banca dati regionale e applicativo di Google.

Tabella 1 - Descrittive della distanza (in metri) tra i due metodi di georeferenziazione.

	min	max	media	sd	25°	50°	75°	95°
Totale indirizzi n=1549	0,4	9034	635	1451	12	32	394	4474
Migliori* sistema Regione Toscana n=1086	0,4	9034	384	1067	10	24	86	2437
Migliori** applicativo Google n=424	0,4	4699	49	238	7	16	34	146
Case sparse n=125	201	8464	3084	2512	697	2633	5437	7634

* Riconoscimento del numero civico e posizionamento sull'edificio ** Matching perfetto tra indirizzo di input e formattazione dell'algoritmo e metodo di geocoding "*rooftop*"

L'impatto che tali differenze tra sistemi di geocodifica diversi potrebbe avere in uno studio di epidemiologia ambientale è stato valutato considerando una ipotetica classificazione della popolazione sulla base di classi di distanza dalle centrali geotermiche presenti nel comune di Piancastagnaio (tabella 2). Sono state considerate varie corone circolari a distanze crescenti, così come riportato nella tabella sottostante. Si notino le diverse numerosità nelle varie classi di distanza tra i due metodi, con le differenze maggiori nella classe 1000-1500m, dove nel sistema di geocoding di Google ricadono molti indirizzi non trovati e georeferenziati nel centro del comune (metodo "*approximate*").

Caso studio comune di Firenze

Per il comune di Firenze l'area in studio comprende il quartiere ovest della città caratterizzato dalla presenza dell'aeroporto "Amerigo Vespucci", del tratto dell'autostrada del Sole A1, del tratto dell'autostrada Firenze Mare A11. Attualmente è in discussione un progetto di riqualificazione ed ampliamento dello scalo aeroportuale, compresa una ipotesi di allungamento e variazione di orientamento della pista. Ed è tuttora in corso anche uno studio di coorte retrospettivo dei residenti

(circa 32.000 persone nel periodo 2000-2013) in questo quartiere finalizzato a valutare gli effetti dell'esposizione al rumore di origine aeroportuale sulla salute dei cittadini.

Tabella 2 - Confronto tra gli indirizzi georeferenziati con i due metodi per fasce di distanza dalle centrali geotermiche.

Fasce di distanza dalle centrali geotermiche	Sistema Regione Toscana	Applicativo Google
Entro 500m	5	26
500-1000m	509	407
1000-1500m	**286**	**569**
1500-1750m	118	119
1750-2000m	83	74
2000-2250m	87	60
2250-2500m	34	12
2000-3000m	8	2
3000-4000m	160	89
4000-5000m	226	172

Dei 3488 indirizzi di input il sistema di geocodifica regionale presenta anche in questo caso un *match rate* del 100%; di questi il 75% presenta una georeferenziazione basata sul riconoscimento del numero civico (in corrispondenza dell'edificio). A differenza del caso studio nel comune di Piancastagnaio, nel contesto urbano di Firenze non si verifica la modalità di georeferenziazione basata sulla interpolazione in corrispondenza di località. Per il restante 25% l'algoritmo interpola l'indirizzo di input al civico più vicino, sempre con indicazione della distanza numerica.
L'applicativo di Google scarta 28 indirizzi in input; il 58% presenta un matching parziale tra indirizzo testuale in input e la formattazione del sistema, il 24% riporta come metodo di geocoding la dicitura "*rooftop*", il 56% riporta "*range interpolated*", il 16% riporta la dicitura "*geometric center*", ed infine il 3,6% è georeferenziato come "*approximate*".
L'applicativo di Microsoft Bing scarta l'1,6% degli indirizzi in input. Il sistema di geocodifica riporta che per il 94,5% la georeferenziazione è avvenuta mediante riconoscimento dell'indirizzo (metodo "*address*"), per il 3,7% mediante riconoscimento del segmento stradale ("*road block*").
Rispetto al caso di Piancastagnaio, la sovrapposizione dei risultati ottenuti con i tre metodi (figura 2), non mostra le divergenze macroscopiche evidenziate nel caso precedente, grazie alla maggiore copertura in ambiente urbano dei dataset dei due sistemi commerciali, Google e Bing. In figura sono riportate anche le isofone dei livelli di rumore prodotto dall'aeroporto, utilizzando l'indicatore Lden, ovvero una misura dell'esposizione della popolazione ai livelli di rumore su tutto l'arco della giornata.
Anche in tale situazione, la geocodifica prodotta dal sistema regionale risulta migliore sia di quella di Google, anche se con divergenze più contenute rispetto al caso di Piancastagnaio, sia rispetto a Bing. Va segnalata la migliore performance dei sistemi regionali anche per la particolarità del comune di Firenze di utilizzare spesso come esponente del numero civico un numero, oltre al tradizionale sistema con lettere dell'alfabeto ("Via Roma 3/A"). Ad esempio "Via Don Lorenzo Milani 133 6", oppure "Via Pistoiese 203 1" sono classiche situazioni in cui i sistemi commerciali non riescono a distinguere tra i due numeri, mentre il sistema regionale, basato su operazioni di geocodifica a livello locale, riesce a gestire tali specificità.
Si riportano due grafici (figura 3) che sintetizzano le differenze tra il sistema regionale e i due sistemi commerciali e consentono di evidenziare eventuali bias direzionali. In questo tipo di grafico, per ogni coppia di punti, la geocodifica del sistema regionale è presa a riferimento e posta al centro. I punti vicino al centro hanno piccole differenze rispetto al riferimento, i punti al di sopra hanno una coordinata più a nord rispetto al riferimento, e così per le 4 direzioni. Dato che i punti presentano

per lo più una distribuzione casuale in tutte le direzioni, si può escludere un bias sistematico tra i diversi sistemi di georeferenziazione.

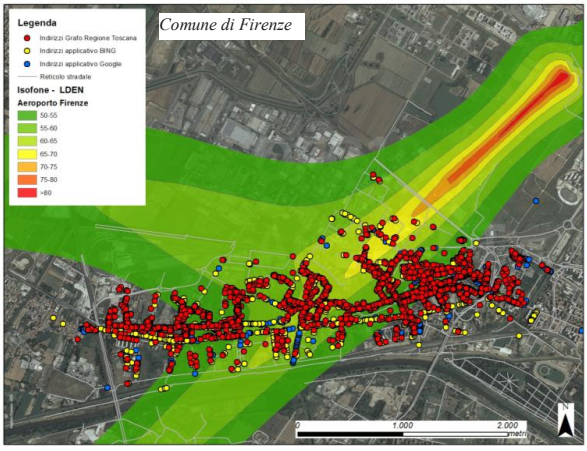

Figura 2 - Geocoding degli indirizzi nel comune di Firenze: confronto tra i tre metodi.

In tabella 3 si riportano le statistiche di confronto tra il sistema di geocodifica della regione Toscana e i due applicativi commerciali. Le differenze sono in generale molto più contenute rispetto al caso di Piancastagnaio e i risultati dell'applicativo Bing-Microsoft sono lievemente migliori rispetto all'applicativo di Google.

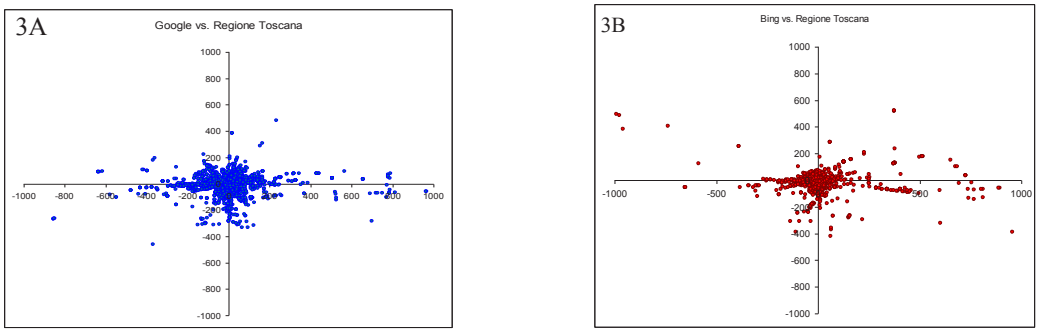

Figura 3 - Distanze e bias direzionali tra il sistema regionale e l'applicativo di Google (3A) e tra il sistema regionale e l'applicativo Bing (3B).

Tabella 3 - Descrittive della distanza (in metri) tra i tre metodi di georeferenziazione.

Confronto con sistema Regione Toscana	min	max	media	sd	25°	50°	75°	95°
Applicativo Google	0,4	1888	122	247	12	50	120	432
Applicativo Bing	0,3	3070	103	337	11	22	51	378

In analogia a quanto esposto per il comune di Piancastagnaio, si riporta per ciascun sistema di georeferenziazione la distribuzione degli indirizzi sulla base di varie classi di decibel, indicative dell'esposizione della popolazione residente al rumore aeroportuale (tabella 4). Si osservano anche in questo caso differenze più contenute rispetto al caso studio di Piancastagnaio.

Tabella 4 - Confronto tra gli indirizzi georeferenziati con i tue metodi per fasce di isofone prodotte dall'aeroporto.

Isofone (dbA)	Sistema Regione Toscana	Applicativo Google	Applicativo Bing
> 80	0	0	0
75-80	0	0	0
70-75	0	0	0
65-70	20	26	35
60-65	440	535	423
55-60	568	580	564
50-55	1053	992	1007
< 50	865	813	917

Discussione

Il presente studio ha analizzato la qualità della georeferenziazione degli indirizzi di residenza ottenuta mediante tre diversi sistemi di *geocoding* e i potenziali impatti che essa può avere in studi di epidemiologia ambientale, nei quali ai residenti vengono generalmente associati, mediante query spaziali, le informazioni sui fattori di rischio. Generalmente negli studi di epidemiologia ambientale svolti in Italia viene data scarsa rilevanza alle questioni inerenti la completezza e la precisione della georeferenziazione; al massimo viene riportato il *match rate* come unica misura dell'accuratezza della procedura di geocodifica, trascurando quasi del tutto il *positional error*.

I due casi studio svolti in regione Toscana affrontano, al contrario, le varie componenti che descrivono la qualità della georeferenziazione, ovvero completezza, precisione e ripetibilità.

Il *match rate* dei tre sistemi confrontati, così come le indicazioni riportate in studi simili condotti soprattutto negli Stati Uniti su altri sistemi di geocodifica (Goldberg et al., 2013), è generalmente molto alto: si raggiungono spesso percentuali di *matching* del 98%. Più complessa è la valutazione della precisione, o *positional error*. Gli autori che si sono occupati di questi argomenti utilizzano generalmente come riferimento, come gold standard, la localizzazione ottenuta mediante acquisizione delle coordinate con GPS. In alternativa il confronto viene effettuato sovrapponendo i risultati della georeferenziazione con le fotografie aeree che rappresentano la realtà al suolo. Nel presente studio, vista la numerosità degli indirizzi in input che rendeva poco praticabile l'utilizzo del GPS, è stato applicato il metodo delle ortofoto, oltre all'uso della carta tecnica regionale e di altri strati informativi utili. La scelta dei due casi studio, oltre che dettata dalla concomitanza nelle due aree di studi di epidemiologia ambientale, è stata motivata anche dall'interesse di valutare la sensibilità dei sistemi di geocodifica a diversi contesti territoriali, rurale/montano quello di Piancastagnaio e prettamente urbano quello intorno l'aeroporto di Firenze.

In entrambe le applicazioni il sistema di geocodifica reso disponibile dalla regione Toscana mostra performance nettamente migliori rispetto ai due applicativi commerciali gratuiti. Nell'area montana di Piancastagnaio l'applicativo Bing di Microsoft risulta del tutto inutilizzabile in studi micro geografici quali quelli di epidemiologia ambientale, in quanto privo di informazione relativa ai numeri civici. L'applicativo di Google, seppur con una precisione inferiore al sistema regionale, riesce a posizionare gli indirizzi nel centro abitato del piccolo comune. Al contrario risulta inefficiente e inaffidabile nel caso di località sparse nel territorio comunale. La differenza media, infatti, tra le coordinate del sistema regionale e quelle restituite dall'applicativo di Google è di circa 600m, ma aumenta fino a oltre 3000m nel caso delle case sparse. Diversa è la situazione per il contesto urbano nell'area intorno l'aeroporto di Firenze. Tutti e tre i sistemi di georeferenziazione mostrano performance nel complesso buone; in quest'area l'applicativo Bing ha una precisione e qualità di poco superiori a quelli dell'applicativo di Google. Anche altri studi hanno mostrato prestazioni migliori dei sistemi di geocodifica in ambiente urbano rispetto ai contesti rurali (Skelly et al., 2002). In questo caso, però, la maggiore qualità del sistema regionale è anche motivata da

elementi a forte connotazione territoriale che i sistemi commerciali, in quanto sviluppati per utenti generalizzati, non possono prevedere e gestire in maniera adeguata. La Regione Toscana, infatti, vanta una esperienza consolidata nella produzione, manutenzione e aggiornamento della banca dati geografica regionale che ha consentito alla regione di dotarsi di una base dati geografica regionale proprietaria che in alcune province ha un livello di copertura di alta qualità.
La valutazione di tutte le componenti che descrivono la qualità del *geocoding* è, pertanto, una operazione di estrema importanza che a causa delle numerose specificità dei territori, va considerata caso per caso e difficilmente può essere generalizzata ad altre aree geografiche. I sistemi di georeferenziazione ad oggi disponibili sono molteplici e sviluppati per le più svariate applicazioni, alcuni ad accesso gratuito, altri a pagamento. La valutazione della qualità dei risultati della georeferenziazione, essendo potenzialmente una operazione onerosa in termini di tempo e risorse, va anche rapportata allo scopo per il quale un sistema viene utilizzato. Nel caso di studi di epidemiologia ambientale a forte dettaglio geografico è necessaria una georeferenziazione di alta qualità per evitare l'introduzione di *bias* geografici che andrebbero ad inficiare i risultati delle analisi di associazione tra esposizione ed eventi sanitari.

Bibliografia

Goldberg D, Cockburn M. (2010), "Improving geocode accuracy with candidate selection criteria", *Trans GIS*, 14(s1): 149–176.
Goldberg DW, Ballard M, Boyd JH, Mullan N, Garfield C, Rosman D, Ferrante AM, Semmens JB. (2013), "An evaluation framework for comparing geocoding systems", *Int J Health Geogr*, 12: 50.
Lovasi GS, Weiss JC, Hoskins R, Whitsel EA, Rice K, Erickson CF, Psaty BM. (2007), "Comparing a single-stage geocoding method to a multi-stage geocoding method: how much and where do they disagree?", *Int J Health Geogr*, 6:12.
Mazumdar S, Rushton G, Smith BJ, Zimmerman DL, Donham KJ. (2008), "Geocoding accuracy and the recovery of relationships between environmental exposures and health", *Int J Health Geogr,* 7: 13–31.
Nuckols JR, Ward MH, Jarup L. (2004), "Using geographic information systems for exposure assessment in environmental epidemiology studies", *Environ Health Perspect*, 112(9): 1007-1015.
Oliver MN, Matthews KA, Siadaty M, Hauck FR, Pickle LW. (2005), "Geographic bias related to geocoding in epidemiologic studies", *Int J Health Geogr*, 4: 29–38.
Rushton G, Armstrong MP, Gittler J, Greene BR, Pavlik CE, West MM, Zimmerman DL. (2006), "Geocoding in cancer research: a review", *Am J Prev Med*, 30(2): S16–S24.
Schootman M, Sterling DA, Struthers J, Yan Y, Laboube T, Emo B, Higgs G. (2007), "Positional accuracy and geographic bias of four methods of geocoding in epidemiologic research", *Ann Epidemiol*, 17(6): 379–387.
Skelly C, Black W, Hearnden M, Eyles R, Weinstein P. (2002), "Disease surveillance in rural communities is compromised by address geocoding uncertainty: a case study of campylobacteriosis", *Aust J Rural Health*, 10(2): 87–93.
Ward MH, Nuckols JR, Giglierano J, Bonner MR, Wolter C, Airola M, Mix W, Colt JS, Hartge P. (2005), "Positional accuracy of two methods of geocoding", *Epidemiology*, 16(4): 542–547.
Whitsel EA, Quibrera PM, Smith RL, Catellier DJ, Liao D, Henley AC, Heiss G. (2006), "Accuracy of commercial geocoding: assessment and implications", *Epidemiol Perspect Innov*, 3: 8–20.
Zandbergen PA. (2008), "A comparison of address point, parcel and street geocoding techniques", *Comput Environ Urban Syst*, 32: 214–232.
Zandbergen PA. (2009), "Geocoding quality and implications for spatial analysis", *Geogr Compass,* 3(2): 647–680.
Zhan FB, Brender JD, De Lima I, Suarez L, Langlois PH. (2006), "Match rate and positional accuracy of two geocoding methods for epidemiologic research", *Ann Epidemiol*, 16(11): 842–849.