Spatial enablement e geostatistica applicati ai dati di salute pubblica: problemi, metodi, esempi

Vittorio Casella([a]), Marica Franzini([a]), Marica Teresa Rocca([a]), Daniele Pala([b])

([a]) DICAR, Università di Pavia, via Ferrata 3, Pavia,
(vittorio.casella, marica.franzini, maricateresa.rocca)@unipv.it
([b]) DIII, Università di Pavia, via Ferrata 3, Pavia, daniele.pala02@universitadipavia.it

1. Introduzione

Il presente articolo ha come argomento principale lo *spatial enablement*. Di che cosa si tratta? I dati prodotti dalle società moderne sono in gran parte associati, in modo naturale, alla posizione: la chiamata fatta con un moderno *smartphone* è collocata nello spazio e nel tempo; le agenzie di salute pubblica raccolgono i dati relativi alla salute dei cittadini raggruppandoli per Comuni (implicitamente collocandoli nello spazio) o, nel caso di grandi città, secondo suddivisioni interne alla municipalità; i dati misurati da una stazione per il monitoraggio della qualità dell'aria si riferiscono al punto in cui essa è collocata; i dati anagrafici associano ad ogni residente, un indirizzo. Fare *spatial enablement* con tali dati significa associare loro l'informazione sulla posizione (forse è più appropriato usare il termine *esplicitare*) e sfruttarla ai fini della visualizzazione e dell'analisi. La Figura 1 mostra un esempio relativo a dati di salute pubblica per la Provincia di Pavia per l'anno 2014: nella mappa a destra si possono individuare alcuni andamenti che ben difficilmente si potrebbero cogliere nella tabella mostrata a sinistra. In una presentazione di qualche anno fa, Dan Paull, CEO dell'azienda australiana PSMA, definisce efficacemente lo *spatial enablement* come "Including 'Location' in Your Thinking, Problem Solving and Decision Making" (Paull, 2012).

Lo *spatial enablement* viene applicato a un numero crescente di discipline: si parla di *spatial econometrics*, *spatial epidemiology*, *health geomatics* e molti altri esempi potrebbero essere fatti. Un ambito in cui è applicato è quello *dell'esposomica*, cioè la complessa e articolata disciplina che cerca di correlare la salute delle persone con i fattori ai quali sono esposte, come il patrimonio genetico, l'ambiente, lo stile di vita, le abitudini alimentari.

Il progetto H2020 PULSE (*Participatory Urban Living for Sustainable Environments*) è stato finanziato dalla UE nell'ambito di una *call* denominata, con espressione molto efficace, *Big Data supporting Public Health policies*. Il progetto è molto articolato (PULSE website) ma si può dire, in sintesi estrema, che si occupa del nesso ambiente-salute con un innovativo approccio partecipativo, in cui i cittadini, da una parte ricevono informazioni e consigli dalle agenzie di

salute pubblica e, dall'altra, acquisiscono e donano dati che consentiranno di mettere a punto strategie di governo dell'ambiente e della salute di crescente raffinatezza. Il Progetto è iniziato il 1/11/2016 e ha durata triennale. Uno dei suoi elementi qualificanti è la presenza di cinque città-pilota che forniranno i dati, recluteranno volontari e sperimenteranno le tecnologie sviluppate nel progetto; si tratta di Barcellona, Birmingham, New York, Parigi e Singapore; bisogna inoltre aggiungere che l'assemblea dei partecipanti ha deliberato di accettare come ulteriori città-pilota (sia pure senza budget, ma con uguale dignità in tutti gli altri aspetti) le città di Pavia e Keelung, a Taiwan.

istat	nome	3	10	21	43
18001	ALAGNA	44	7,59	3477	8733
18002	ALBAREDO ARNA	48	6,91	1761	9104
18003	ALBONESE	50	5,72	3385	11174
18004	ALBUZZANO	40	4,08	2670	9538
18005	ARENA PO	47	6,23	2254	14537
18006	BADIA PAVESE	44	5,97	1715	9133
18007	BAGNARIA	50	6,98	2628	12494
18008	BARBIANELLO	46	4,68	2986	14108
18009	BASCAPE'	42	6,05	3457	10672
18010	BASTIDA DE'DOS	46	6,63	2973	13074
18011	BASTIDA PANCA	45	4,82	2204	13555
18012	BATTUDA	41	5,93	3478	10122
18013	BELGIOIOSO	45	5,27	3030	13086
18014	BEREGUARDO	44	4,47	2310	10242
18015	BORGARELLO	38	3,78	2909	10040
18016	BORGO PRIOLO	48	5,8	2430	10921
18017	BORGORATTO M	52	6,43	4117	8739
18018	BORGO SAN SIR	43	5,36	2705	7632
18019	BORNASCO	37	4,22	3115	12320
18020	BOSNASCO	48	5,99	1448	11114

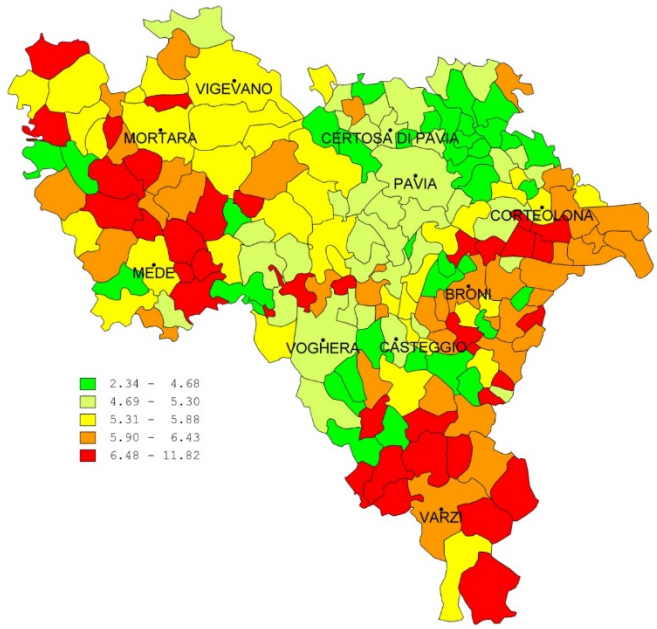


Figura 1 – Esempio di dati di salute pubblica nella usuale forma alfa-numerica (SX) e nella forma, spazialmente abilitata, della mappa (DX)

L'Università di Pavia è uno dei 12 partner ed è presente con tre gruppi di ricerca. Uno di questi è costituito dai docenti, ricercatori e tecnici che operano nel Laboratorio di Geomatica e che si occupano principalmente di effettuare lo *spatial enablement* sui dati delle città-pilota, realizzare un WebGIS avente alcune caratteristiche innovative e specifiche (in collaborazione con l'azienda Genegis GI che si occupa dell'implementazione), eseguire analisi geospaziali dei dati raccolti.

L'articolo si occupa di due aspetti principali. Il primo è il modo in cui le città raccolgono e organizzato i dati, che spesso non è favorevole allo *spatial enablement*. In genere le città dispongono di numerose suddivisioni in poligoni (New York ne ha addirittura otto) e purtroppo alcuni indicatori sono raccolti rispetto a una scomposizione e altri fanno riferimento a un'altra. Si pone dunque il problema di armonizzare i dati. Nella Sezione 2 verranno mostrati alcuni

esempi relativi alla città di New York e si illustrerà la procedura di armonizzazione spaziale che è stata applicata.

Il secondo aspetto riguarda l'analisi geostatistica (tecniche di statistica che tengono conto della posizione) delle mappe relative alla salute e agli aspetti socio-economici. La Sezione 3 mostra alcuni esempi basati sull'uso della GWR (*Geographically Weighted Regression*).

2. Aspetti legati allo *spatial enablement* dei dati delle città

Grazie alla grande disponibilità di open data, la città di New York è quella per cui, nell'ambito di PULSE, è stata svolta finora la più significativa attività di *spatial enablement*. Per tale città (e anche per tutte le altre città-pilota del progetto) i dati di salute e socio-economici non vengono raccolti per l'intera municipalità, ma sono invece aggregati per aree più piccole.

Premettiamo una breve illustrazione di che cosa significhi abilitare spazialmente i dati di salute di New York, in termini pratici. Sui siti di open data si trovano dati esportabili in formato CSV o XLS; essi non sono organizzati in forma rigorosamente tabellare (non sempre, quantomeno), ma vi possono essere ricondotti con manipolazioni o mediante programmi creati ad-hoc (nel nostro caso in ambiente Matlab), se le sorgenti di dati da importare sono numerose.

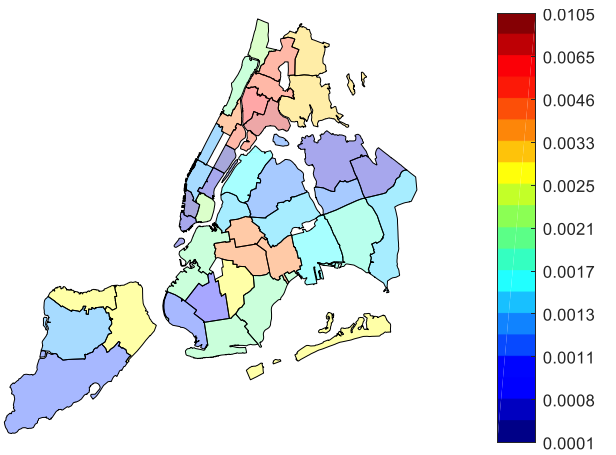
```
,,,,,Adults with Asthma in the Past 12 Months : Summarize
Topic: Health Behavior and Population
Subtopic: Asthma
Indicator Name: Adults with Asthma in the Past 12 Months
Indicator Description: Adults with Asthma in the Past 12 Months
Notes: **Estimate is supressed due to insufficient data.*Estimate is based on small numbers so sh

Year,GeoTypeName,Borough,Geography,Geography_id,IndicatorDescription,Number,Percent,Age-Adjusted
2014,Citywide,New York City, New York City,1,Adults with Asthma in the Past 12 Months,"238,000 ",
2014,Borough,Bronx, Bronx,1,Adults with Asthma in the Past 12 Months,"48,000 ","4.7 (3.5,6.3)","4
2014,Borough,Brooklyn, Brooklyn,2,Adults with Asthma in the Past 12 Months,"59,000 ","3.0 (2.3,4.
2014,Borough,Manhattan, Manhattan,3,Adults with Asthma in the Past 12 Months,"55,000 ","4.2 (3.0,
2014,Borough,Queens, Queens,4,Adults with Asthma in the Past 12 Months,"59,000 ","3.3 (2.4,4.5)",
2014,Borough,Staten Island, Staten Island,5,Adults with Asthma in the Past 12 Months,"17,000 ","4
2014,Neighborhood (UHF 34),Queens, Bayside Little Neck-Fresh Meadows,404406,Adults with Asthma in
2014,Neighborhood (UHF 34),Brooklyn, Bedford Stuyvesant - Crown Heights,203,Adults with Asthma in
2014,Neighborhood (UHF 34),Brooklyn, Bensonhurst - Bay Ridge,209,Adults with Asthma in the Past
2014,Neighborhood (UHF 34),Brooklyn, Borough Park,206,Adults with Asthma in the Past 12 Months,"8
2014,Neighborhood (UHF 34),Brooklyn, Canarsie - Flatlands,208,Adults with Asthma in the Past 12 M
2014,Neighborhood (UHF 34),Manhattan, Central Harlem - Morningside Heights,302,Adults with Asthma
2014,Neighborhood (UHF 34),Manhattan, Chelsea-Village,306308,Adults with Asthma in the Past 12 Mo
2014,Neighborhood (UHF 34),Brooklyn, Coney Island - Sheepshead Bay,210,Adults with Asthma in the
2014,Neighborhood (UHF 34),Brooklyn, Downtown - Heights - Slope,202,Adults with Asthma in the Pas
2014,Neighborhood (UHF 34),Brooklyn, East Flatbush - Flatbush,207,Adults with Asthma in the Past
2014,Neighborhood (UHF 34),Manhattan, East Harlem,303,Adults with Asthma in the Past 12 Months,"7
2014,Neighborhood (UHF 34),Brooklyn, East New York,204,Adults with Asthma in the Past 12 Months,"
```

Figura 2 – Esempio di dati scaricati da siti open data in formato CSV (New York Enviromental & Health Data Portal)

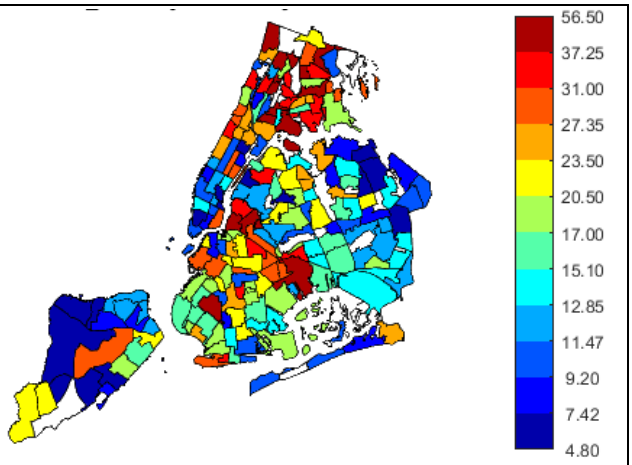
Il frammento mostrato in Figura 2 mostra un piccolo campionario dei problemi che si incontrano: la chiave "1" è usata sia per indicare tutta la città sia per indicare il *borough* del *Bronx*; il separatore delle colonne è la virgola, che è anche usata come separatore delle migliaia; il modo in cui vengono rappresentati

i dati mancanti infine (caso qui non documentato) mal si presta all'elaborazione automatica.

Una volta data una vera organizzazione tabellare ai dati esemplificati, bisogna procurarsi un file grafico (tipicamente in formato *shapefile*) che descrive il perimetro delle aree considerate. Ciò non è banale in quanto i siti che rendono disponibili i dati in forma tabellare non sempre forniscono anche i poligoni. In certi casi, è stato addirittura necessario editare i file dei poligoni, ad esempio per creare la mappa di UHF34 a partire dalla UHF42 (il significato di tali sigle è spiegato nel seguito), una volta capito come la prima è ricavata dalla seconda per aggregazione di poligoni. Spesso è anche necessario armonizzare le chiavi presenti nelle tabelle e nei poligoni, che coincidono in sostanza ma non sempre nella lettera. Superate tali difficoltà, l'unione fra tabelle e poligoni, avente lo scopo di realizzare le mappe mostrate, può essere facilmente svolta in un SW GIS con una operazione di *join* oppure, come fatto da noi, in ambiente Matlab.



A – Ospedalizzazioni per Asma di adulti e giovani adulti, come percentuale sui residenti, nel 2014, riferito alla suddivisione UHF42



B – Tasso di povertà riferito alla suddivisione NTA e all'ultimo censimento (anno 2010)

Figura 3 – Esempi di spatial enablement applicato a dati di salute o socio-economici, con riferimento a suddivisioni poligonali diverse

Per la città di New York vi sono, come la Figura 3 suggerisce, numerose suddivisioni poligonali e in particolare:

- Boroughs: 5 poligoni;
- UHF34 (*United Hospital Fund*): 34 poligoni;
- UHF42 (*United Hospital Fund*): 42 poligoni;
- CD55 (*Community Districts*): 55 poligoni;
- CD59 (*Community Districts*): 59 poligoni;
- CD71 (*Community Districts*): 72 poligoni;
- NTA (*Neighborhood Tabulation Areas*; suddivisioni usate per il censimento): 195 poligoni;
- ZIP (aree relativi ai codici di avviamento postale o *ZIP area*): 200 poligoni;

- ZIP2 (aree relativi ai codici di avviamento postale, valide dal 2014): 262 poligoni.

Purtroppo parametri diversi sono raccolti rispetto a suddivisioni diverse e si pone il problema di una armonizzazione: per fare un esempio, ottenere la mappa dell'indice di povertà riferito ai poligoni ZIP2 a partire dalla stessa mappa riferita a NTA.

Allo scopo è stata definita e codificata in Matlab una procedura di interpolazione. Consideriamo una suddivisione poligonale per cui è nativamente disponibile un certo indicatore (la prevalenza di una patologia o il numero di ricoveri causati da essa). Indichiamo con P_i il generico poligono che fa parte della mappa disponibile; indichiamo con v_i il valore dell'indicatore per il poligono assegnato e con n_i il numero degli abitanti. Consideriamo ora un poligono P_0, avente area A_0 appartenente a una differente suddivisione, per cui non è noto il valore v_0 dell'indicatore considerato: lo si vuole determinare combinando opportunamente i valori relativi ai poligoni P_i che gli si sovrappongono. L'algoritmo determina anzitutto l'area dell'intersezione fra P_i e P_0

$$A_i := \text{area}(P_i \quad P_0) \tag{1}$$

Determina poi il sottoinsieme S degli indici i per cui l'area dell'intersezione è significativa

$$S := \{i \quad \{1,2,\text{K}\ n_p\}; A_i > T_1\} \tag{2}$$

La soglia T_1 dovrebbe valere 0 concettualmente; in pratica, poiché le varie suddivisioni poligonali presentano problemi di congruenza geometrica e si danno valori di A_i maggiori di 0 ma molto piccoli, si è preferito porre $T_1 = 1\,\text{m}^2$. Infine si definisce

$$v_0 := \frac{\bullet_{i\ S} A_i v_i}{\bullet_{i\ S} A_i} \tag{3}$$

in cui si calcola la media dei valori v_i pesati con l'area dell'intersezione A_i. Infine, l'algoritmo effettua una verifica di significatività accertandosi che i vari frammenti derivanti dalle intersezioni ricoprano in modo sostanziale P_0, mediante il test

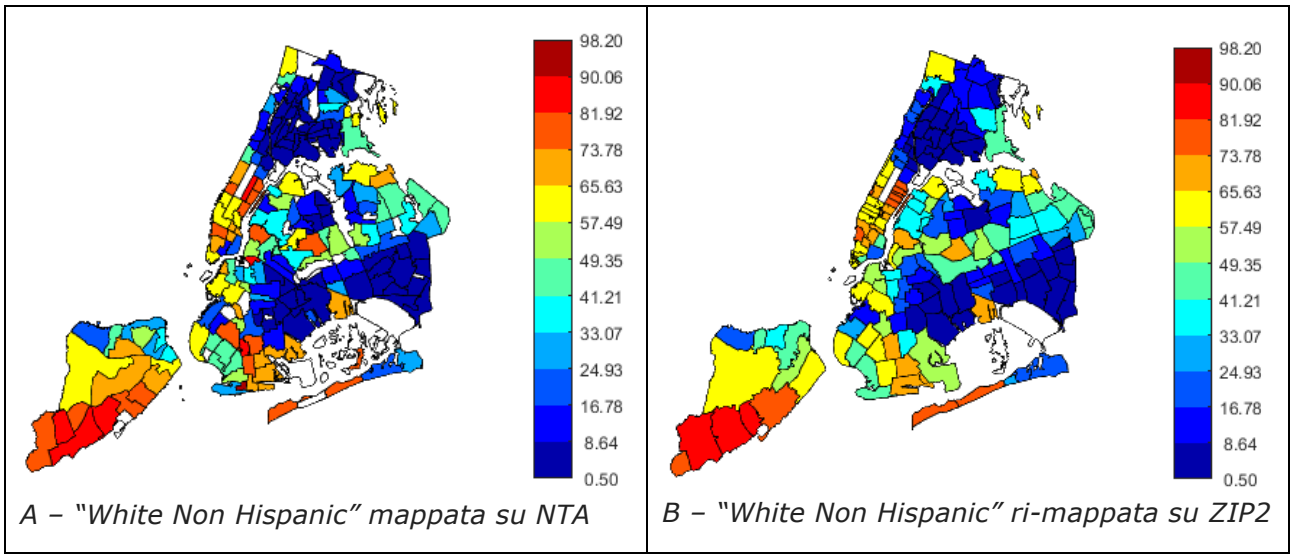
$$\frac{\bullet_{i\ S} A_i}{A_0} > T_2 \tag{4}$$

con $T_2 = 0.7$ negli esempi considerati. Se la condizione di significatività non è soddisfatta, si pone $v_0 = \text{nodata}$. Nella seconda versione dell'algoritmo, la (3) è stata generalizzata nel modo indicato

$$v_0 := \frac{\sum_{i \in S} A_i w_i v_i}{\sum_{i \in S} A_i w_i} \qquad (5)$$

in cui il peso complessivo è uguale al prodotto di A_i con un ulteriore fattore. Se si usa l'algoritmo per interpolare una grandezza estensiva (come ad esempio il numero dei malati per una certa patologia) riteniamo corretto usare la (3) o, equivalentemente, la (5) avendo posto $w_i = 1$. Se si interpola una grandezza intensiva (la prevalenza di una certa patologia ad esempio, cioè il rapporto fra il numero dei malati e il numero dei residenti), per evitare che aree molto e poco popolate pesino allo stesso modo, bisogna introdurre come ulteriore peso il numero degli abitanti, da cui $w_i = n_i$.

La Figura 4 illustra il lavoro di armonizzazione spaziale e mostra, nella sottofigura A, la distribuzione dell'etnia *"White non Hispanic"* riferita alla suddivisione NTA; si tratta di un dataset scaricato direttamente da un database americano. La sottofigura B mostra la stessa grandezza ricampionata sulla suddivisione ZIP2.



A – "White Non Hispanic" mappata su NTA

B – "White Non Hispanic" ri-mappata su ZIP2

Figura 4 – Illustrazione del lavoro di armonizzazione spaziale. A SX, la mappa della quantità "White Non Hispanic" originaria, riferita alla suddivisione NTA. A DX la stessa quantità rimappata sulla suddivisione ZIP2

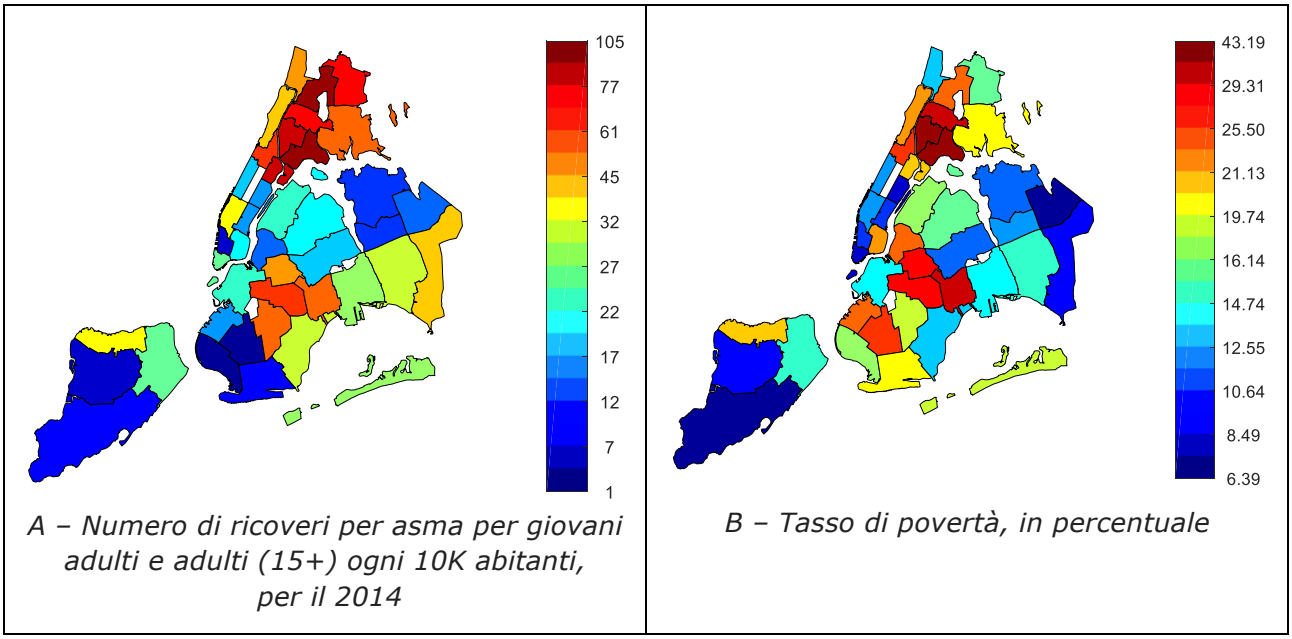
3. Analisi geospaziali con GWR

Nell'ambito del progetto PULSE sono state costruite in modo sistematico le mappe riferite a diversi indicatori di salute e socio-economici, relative ad anni diversi e, talvolta, a fasce di età diverse. Esse possono venire osservate e

comparate visivamente: il WebGIS sviluppato per il progetto H2020 PULSE consente ad esempio di affiancare due mappe ed osservarle in modo sincronizzato nello spazio e nel tempo, facendo *zoom* e *panning*; la semplice osservazione visiva permette di individuare una quantità di elementi di grande interesse.

Tuttavia le mappe possono essere analizzate in termini di geostatistica, una estensione della statistica che tiene conto della dimensione spaziale dei dati. La presente sezione illustra alcuni risultati ottenuti con la GWR (Geographically Weighted Regression, 2003). Preliminarmente, si ritiene utile richiamare alcuni elementi relativi alla regressione ordinaria, che è utile per investigare la correlazione fra una variabile dipendente y e una variabile indipendente x, detta anche regressore o, ancora, variabile esplicativa (*explanatory variable*).

Consideriamo anzitutto le due mappe relative ai ricoveri per asma (Figura 5A) e al tasso di povertà (Figura 5B). Le due mappe sono visivamente correlate in quanto sembra che zone avente il primo indicatore alto presentino anche valori elevati del secondo.



A – Numero di ricoveri per asma per giovani adulti e adulti (15+) ogni 10K abitanti, per il 2014

B – Tasso di povertà, in percentuale

Figura 5 – Due dataset relativi alla suddivisione UHF42

Nel seguito considereremo i ricoveri per asma come variabile dipendente y e il tasso di povertà come variabile esplicativa x. In un approccio classico, a-spaziale, bisognerebbe considerare la tabella dei 42 valori (x_i, y_i), mostrata solo in parte in Tabella 1. Tali valori possono essere rappresentati in un grafico cartesiano, come nella Figura 6 in cui appaiono come croci blu. Calcolare la regressione significa stimare la retta che *passa* per i punti.

x - tasso di povertà	*y* - numero ricoveri
12,78	45,20
15,71	72,10
25,47	87,00
20,39	55,80
38,75	75,80
39,55	81,50
...	...

Tabella 1 – Alcune righe della tabella delle osservazioni

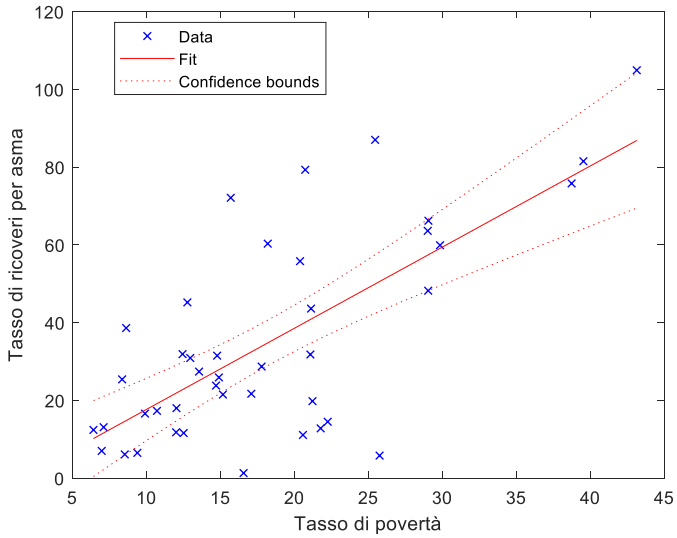


Figura 6 – Dati per la regressione e retta di regressione

Poiché non esiste alcuna retta che passa per i punti mostrati in senso rigoroso, si definisce soluzione del problema quella che minimizza la distanza fra la retta e i punti: si tratta di una definizione corretta in sostanza ma imprecisa, che richiede una formalizzazione. La relazione ideale fra y e x è

$$y = {}_0 + {}_1 x \tag{6}$$

e deve valere per tutti i punti della Figura 6. Si può costruire un formalismo vettoriale ponendo

$$\mathbf{Y} = \begin{matrix} y_1 \\ y_2 \\ \text{M} \\ y_m \end{matrix} \qquad \mathbf{X} = \begin{matrix} 1 & x_1 \\ 1 & x_2 \\ 1 & \text{M} \\ 1 & x_m \end{matrix} \qquad \boldsymbol{\beta} = \begin{matrix} 0 \\ 1 \end{matrix} \tag{7}$$

E in questo modo il complesso delle relazioni del tipo (6) che si possono scrivere per tutti gli m punti equivalgono a

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} \tag{8}$$

La determinazione della retta nel senso della geometria richiederebbe la soluzione del sistema di equazioni

$$\mathbf{Y}_0 \text{ B } \mathbf{X}\boldsymbol{\beta} \tag{9}$$

rispetto a $\boldsymbol{\beta}$; la presenza del pedice in $\mathbf{Y}_0$ sottolinea il fatto che, ora, il vettore contiene delle grandezze misurate. Il sistema indicato non ha una soluzione nel senso classico e rigoroso, come anticipato, dunque si deve definire una soluzione approssimata. Consideriamo il vettore delle osservazioni previste

$$\mathbf{Y}(\boldsymbol{\beta}) := \mathbf{X}\boldsymbol{\beta} \tag{10}$$

si definisce la soluzione $\hat{\boldsymbol{\beta}}$ come quella che minimizza la norma della differenza fra le osservazioni previste (o stimate) e quelle reali

$$\hat{\boldsymbol{\beta}} := \min_{\boldsymbol{\beta}} \left(\left| \mathbf{Y}(\boldsymbol{\beta}) \quad \mathbf{Y}_0 \right|^2 \right) \tag{11}$$

Adottando la norma cartesiana si può scrivere

$$\hat{\boldsymbol{\beta}} := \min_{\boldsymbol{\beta}} \sum_{i=1}^{m} \left(y_{0i} \quad (\mathbf{X}\boldsymbol{\beta}) \right)^2 \tag{12}$$

dove y_{0i} indica l'*i*-esimo elemento di $\mathbf{Y}_0$. La (12) può essere generalizzata in modo da dare ad ogni osservazione un peso diverso

$$\hat{\boldsymbol{\beta}} := \min_{\boldsymbol{\beta}} \sum_{i=1}^{m} w_i \left(y_{0i} \quad (\mathbf{X}\boldsymbol{\beta}) \right)^2 \tag{13}$$

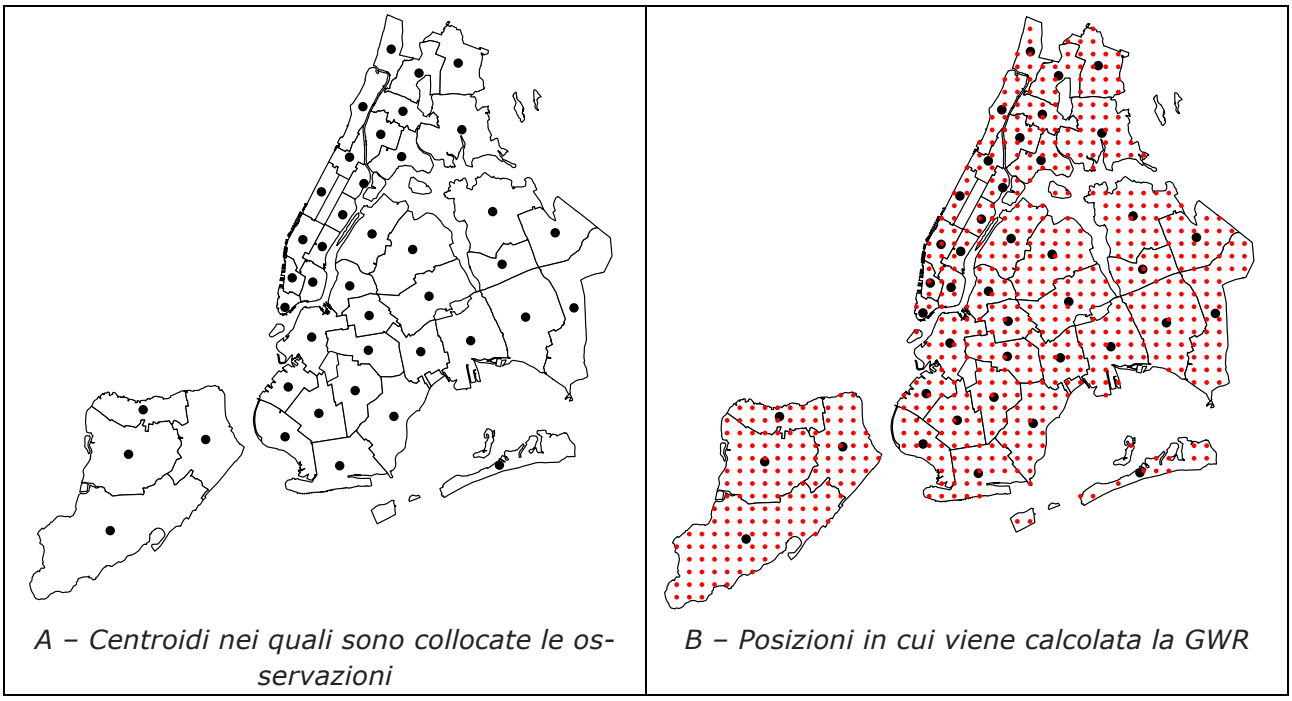
Una volta determinata la retta di regressione rappresentata da $\hat{\boldsymbol{\beta}}$, ci si deve chiedere se questa si adatti bene ai dati originari. Si tratta di un argomento lungo e noi di limiteremo a mostrare (e a definire rigorosamente) il parametro indicato con R^2

$$R^2 = \frac{\sum_{i=1}^{m} \left(\hat{y}_i \quad \bar{Y} \right)^2}{\sum_{i=1}^{m} \left(\hat{y}_{0i} \quad \bar{Y} \right)^2} \tag{14}$$

Dove $\hat{y}_i$ è l'elemento i-esimo di $\hat{\mathbf{Y}}$ e $\bar{Y}$ è il valor medio del vettore $\mathbf{Y}$. Lo R^2 prende valore fra 0 e 1; valori prossimi a 1 indicano che vi è un ottimo accordo fra i punti misurati e la retta, cioè i punti sono quasi allineati; valori prossimi a 0 indicano al contrario che i punti non evidenziano una dipendenza lineare.

Nel caso dei dati mostrati in Figura 6, si ha $R^2 = 0.51$. Tale valore, non altissimo ma comunque non trascurabile, insieme all'osservazione della Figura 6, ci portano a concludere che esiste una correlazione fra le due grandezze considerate e che il tasso di povertà è un determinante dei ricoveri per asma; detto in altri termini, le persone povere tendono a ricorrere maggiormente ai ricoveri per problemi di asma. Vedremo in seguito che il valore di R^2 non particolarmente alto nasconde il fatto che ci sono zone in cui l'accordo fra i dati è molto maggiore e zone in cui avviene il contrario.

Fare la regressione pesata geometricamente (GWR), significa ricordarsi che i dati mostrati da Tabella 1 e in Figura 6 avvengono in un luogo. Nella presente nota, le osservazioni della Tabella 1 sono state collocate nei centroidi dei poligoni (Figura 7A); si tratta di una scelta fatta da molti autori in letteratura, ma non è l'unica possibile.



A – Centroidi nei quali sono collocate le osservazioni

B – Posizioni in cui viene calcolata la GWR

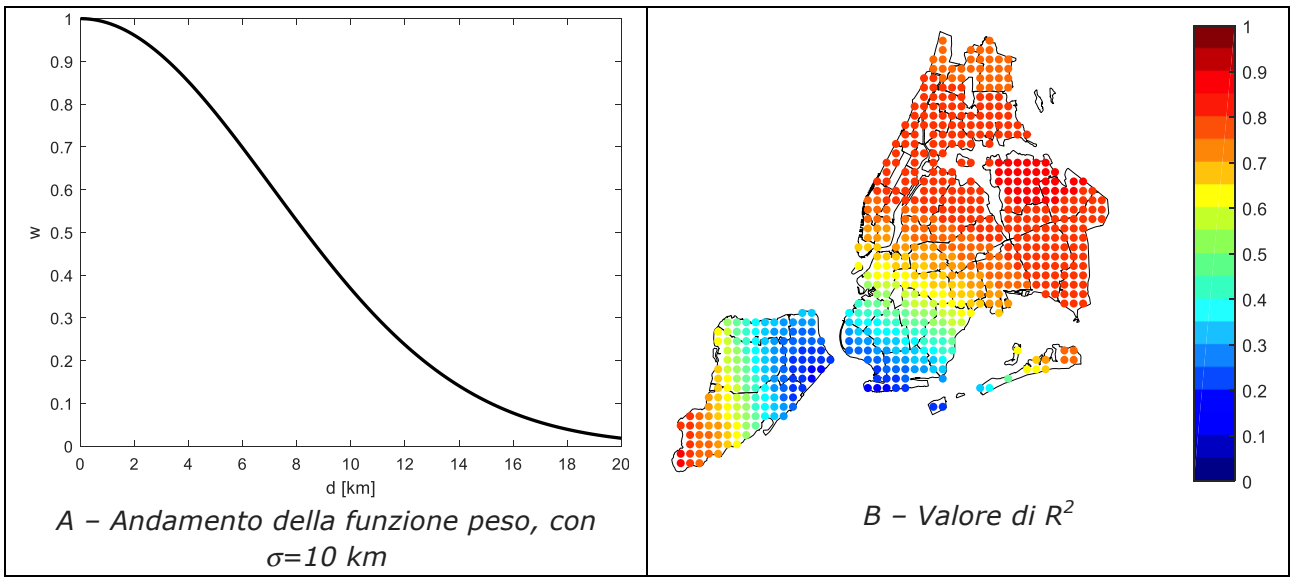
Figura 7 – Aspetti geometrici legati alla GWR

Una griglia regolare, con passo 1 km, è stata sovrapposta al territorio da studiare. I nodi che la costituiscono sono mostrati in rosso nella Figura 7B. Per ognuno dei nodi rossi, è stata calcolata la regressione lineare, come nel caso precedente, usando le osservazioni di tutti i 42 poligoni. La differenza consiste nel fatto che ogni osservazione è pesata (vedi (13)) in funzione della distanza fra il nodo rosso in cui la regressione è calcolata e il nodo nero in cui è collocato la *i*-esima osservazione. Nel nostro caso il peso è calcolato con

$$w_i = e^{\frac{d_{ik}}{}^2} \qquad (15)$$

dove *i* corre sui 42 nodi neri, *k* corre sui nodi neri, d_{ik} è la distanza fra l'*i*-esimo nodo nero e il *k*-esimo nodo rosso, è un parametro che controlla il decadimento del peso al crescere della distanza.

La Figura 8 illustra il calcolo della GWR del tasso di ricoveri per asma come funzione del tasso di povertà. La sottofigura A illustra l'andamento della funzione peso al variare della distanza: essa ha come valore massimo 1 e poi decresce con l'andamento tipico della funzione di Gauss. La figura B illustra i risultati in termini del parametro R^2; per chiarezza, per ognuno dei nodi rossi della Figura 7B è stato fatto un calcolo di regressione, e nella Figura 8B i nodi sono colorati in funzione del valore R^2. Ci sono ampie zone con colori prossimi al rosso, in cui il coefficiente di determinazione è prossimo a 1: in quelle zone, in tasso di povertà è un ottimo predittore del tasso di ricoveri. Vi è tuttavia l'anomalia della zona fra Brooklyn e Staten Island, in cui R^2 è molto basso: evidentemente, il tasso di povertà, da solo, non è un buon predittore per tutta New York.



A – Andamento della funzione peso, con σ=10 km

B – Valore di R^2

Figura 8 – Calcolo della GWR dei ricoveri per asma come funzione del tasso di povertà

La GWR può essere estesa in modo da considerare, invece di uno solo, diversi predittori. Esaminando i dati di Figura 9 emerge che l'etnia e la numerosità della popolazione sono visivamente correlati al fenomeno che vogliamo spiegare, il tasso di ricoveri.

La GWR può essere estesa in modo da tenere conto di diverse variabili indipendenti. La (6) diventa

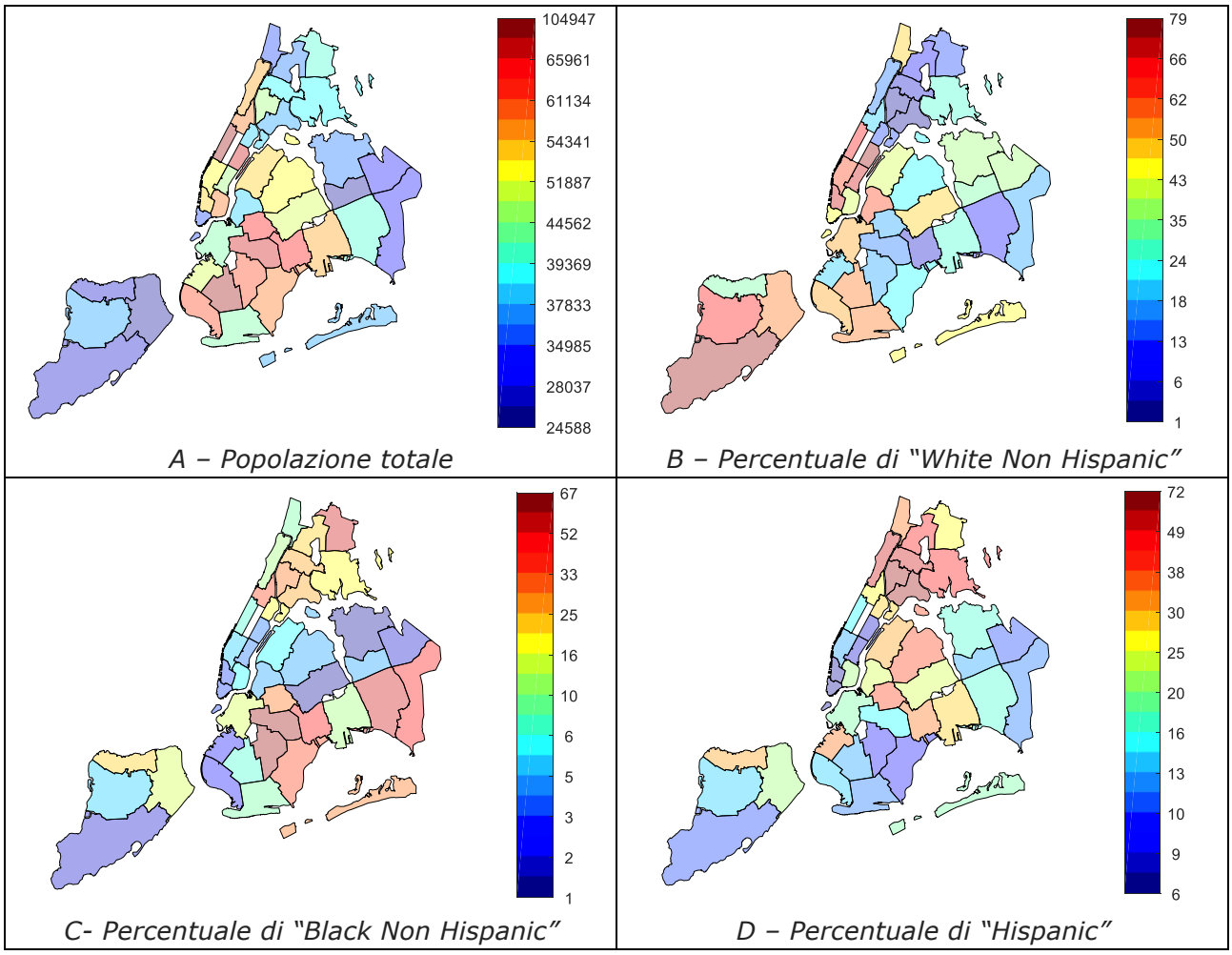
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \mathrm{K}\ \beta_n x_n \tag{16}$$

e le (7) diventano

$$\mathbf{Y} = \begin{matrix} y_1 \\ y_2 \\ \mathrm{M} \\ y_m \end{matrix} \qquad \mathbf{X} = \begin{matrix} 1 & x_{11} & x_{12} & \mathrm{L} & x_{1n} \\ 1 & x_{21} & x_{22} & \mathrm{L} & x_{2n} \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{O} & \\ 1 & x_{m1} & x_{m2} & \mathrm{L} & x_{mn} \end{matrix} \qquad \boldsymbol{\beta} = \begin{matrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \mathrm{M} \\ \beta_n \end{matrix} \tag{17}$$

È stata calcolata la GWR in due altre ipotesi:

- i predittori sono il tasso di povertà e le variabili legate all'etnia;
- i predittori sono quelli elencati sopra e il numero totale di abitanti.



A – Popolazione totale

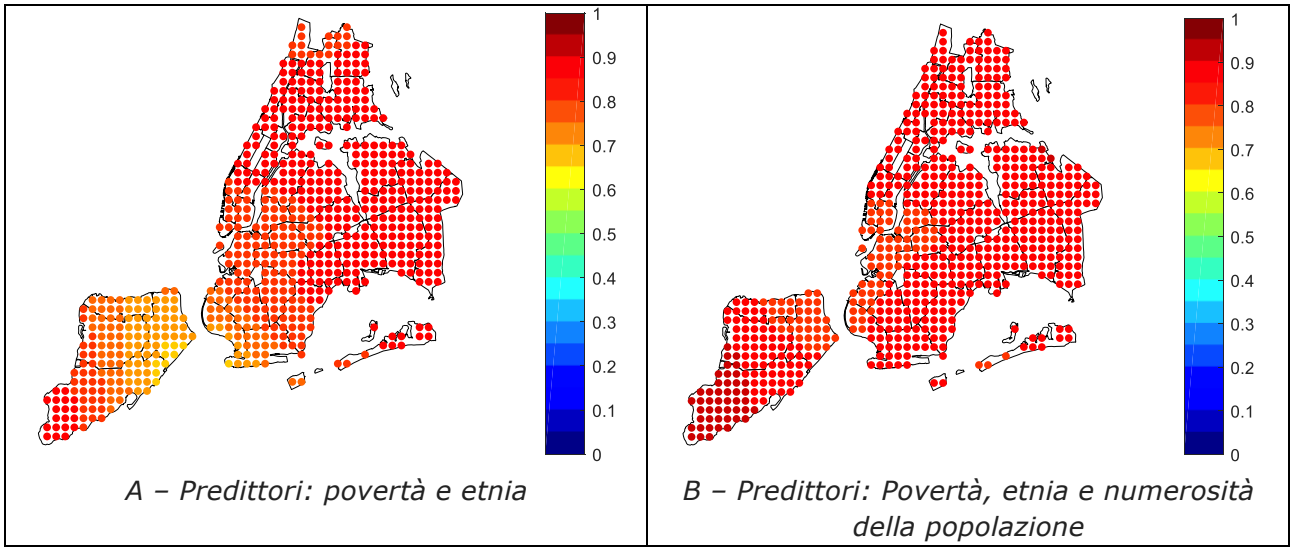
B – Percentuale di "White Non Hispanic"

C- Percentuale di "Black Non Hispanic"

D – Percentuale di "Hispanic"

Figura 9 – Alcuni predittori interessanti

Le Figura 10A e 10B mostrano nuovamente il valore di R^2 ed evidenziano che i due scenari introdotti portano a risultati estremamente migliori.



A – Predittori: povertà e etnia

B – Predittori: Povertà, etnia e numerosità della popolazione

Figura 10 – Risultati della GWR con diversi predittori; la grandezza usata per colorare i nodi è il coefficiente di determinazione R^2

4. Conclusioni

L'articolo fa una rassegna di alcune problematiche emerse nell'esecuzione dello *spatial enablement* di dati di salute e socio-economici relativi alla città di New York. Presenta la soluzione adottate e si sofferma in particolare sul metodo implementato per eseguire l'armonizzazione spaziale.
Successivamente introduce la tecnica della *Geographically Weighted Regression (GWR)* nei suoi elementi fondamentali. Infine presenta alcune analisi geospaziali, basate proprio sulla GWR, in cui si cerca di spiegare il tasso di ricoveri per asma in funzione di diverse variabili esplicative, come il tasso di povertà, la composizione etnica della popolazione e sua numerosità. I risultati sono molto incoraggianti.

L'attività descritta nella nota è stata finanziata dal progetto H2020 PULSE (*Participatory Urban Living for Sustainable Environments*, Grant n. 727816)

5. Bibliografia

Paull D. (2012). Spatial Enablement: including 'local in your thinking, problem solving and decision making'.
https://www.psma.com.au/sites/default/files/2012-may-gsdi-presentation-slides.pdf (ultimo accesso: novembre 2018)

PULSE website. http://www.project-pulse.eu/ (ultimo accesso: novembre 2018)

New York Enviromental & Health Data Portal. http://a816-dohbesp.nyc.gov/IndicatorPublic/PublicTracking.aspx (ultimo accesso: novembre 2018)

Geographically Weighted Regression: The Analysis of Spatially Varying Relationships, by AS Fotheringham, C. Brunsdon, and M. Charlton. Geographical Analysis, 2003.