La zonizzazione statistica in ambito urbano. L'esempio del Comune di Firenze

(Versione Preliminare)

Chiara Bocci
bocci@ds.unifi.it
Alessia Conte
a.conte@comune.fi.it
Andrea Giommi
giommi@ds.unifi.it
Alessandra Petrucci
alex@ds.unifi.it
Emilia Rocco
rocco@ds.unifi.it

Sintesi

Negli ultimi anni, è notevolmente aumentata la necessità di avere informazioni statistiche a livello territoriale dettagliato. Nel caso del Comune di Firenze molte delle analisi statistiche sono riferite, oltre che alla sua superficie totale, ai cinque quartieri che la partizionano ma si avverte la necessità di un riferimento a livelli territoriali di dimensione inferiore.
Il lavoro si propone di presentare alcuni esempi di identificazione di zone "omogenee" effettuati nell'area urbana del Comune di Firenze attraverso l'impiego di metodi di aggregazione spaziale che prendano in considerazione sia fattori oggettivi che specifici requisiti spaziali.

Parole chiave

Zonizzazione, cluster territoriali

1. Introduzione

La parte più rilevante dell'attività dell'Ufficio Comunale di Statistica di Firenze riguarda l'elaborazione e l'analisi di dati provenienti sia da fonti statistiche che amministrative per la produzione di informazioni utili allo sviluppo delle politiche di governo del Comune stesso ma anche utilizzabili da una più ampia utenza sia del settore pubblico sia di quello privato. Sul sito web del Comune sono facilmente individuabili le attività dell'Ufficio e soprattutto le numerose pubblicazioni, la maggior parte delle quali di natura periodica. Buona parte delle statistiche derivano da elaborazioni di dati censuari o di natura amministrativa (archivio anagrafico, registri camerali, ecc.), una

parte cospicua è il frutto dell'elaborazione di dati provenienti da indagini organizzate dall'Ufficio.
L'informazione prodotta fa riferimento al territorio comunale nel suo complesso e in qualche caso a sottoinsiemi dello stesso: i cinque quartieri in cui il Comune si suddivide sono punti di riferimento fondamentali, ma anche altre suddivisioni del territorio possono risultare importanti. I cinque quartieri in effetti hanno dimensioni notevoli ed è avvertita da tempo dai responsabili dell'ufficio l'esigenza di individuare al loro interno aree "omogenee" per caratteristiche demografiche, sociali, economiche, spesso correlate con le caratteristiche dell'edilizia cittadina, che rappresentino un punto di riferimento per l'analisi dei dati, per la progettazione delle indagini, per la diffusione delle informazioni, ad un livello inferiore a quello del quartiere. Nel tempo e anche recentemente, sono state effettuate operazioni di "zonizzazione" del territorio comunale legate ad obiettivi più o meno specifici. Si ricorda una suddivisione del territorio in venti aree rispondenti ad esigenze di equa distribuzione dell'attività di rilevazione nello svolgimento delle indagini ISTAT (Forze di lavoro e Consumi). La suddivisione in venti aree che risale al 1995 (a) non aveva alcun legame anche indiretto con obiettivi di analisi; (b) non rappresentava una vera e propria zonizzazione in quanto non aveva alla base unità areali bensì "aree di circolazione"; in altri termini non si trattava di una partizione del territorio comunale basì di un elenco di venti tragitti mediante i quali raggiungere tutte le abitazioni del territorio comunale. La partizione tuttavia ha costituito la base per la progettazione dell'indagine sulle Forze di lavoro che il comune svolge in parallelo a quella dell'ISTAT a partire dal 1996.
Ancor prima di questa "zonizzazione" esisteva una suddivisione in 69 effettive aree ricavate sulla base dei dati provvisori del censimento della popolazione del 1971. La genesi di tali aree non è oggi del tutto chiara ma certamente furono ricondotte a una partizione dei quartieri cittadini che, fino alla fine degli anni '80, erano quattordici.
Nel 2000, tali aree sono state revisionate con variabili di riferimento di tipo demografico, vincolando i loro confini a quelli dei cinque quartieri prima citati. Il risultato è rappresentato da una zonizzazione in 72 aree del territorio comunale che però non sembrano rispondere adeguatamente alle finalità di analisi, diffusione delle informazioni e base progettuale di indagini che si vorrebbero realizzare.
Si è allora dato avvio ad un nuovo lavoro di zonizzazione che muovendo dagli obiettivi precisati si appoggi sulle moderne tecniche di analisi spaziale attualmente disponibili. Si tratta di un lavoro in progress che non potrà esaurirsi in un lasso breve di tempo ma del quale ci pare interessante presentare fin d'ora le linee di sviluppo, soprattutto relativamente alle metodologie già disponibili e applicabili. Alcune metodologie proposte di recente sono illustrate nel paragrafo 2; nel paragrafo 3 si descrive la metodologia adottata in rapporto ai dati disponibili per la zonizzazione; alcuni risultati preliminari, in forma cartografica, relativi a tre quartieri sono riportati nel paragrafo 4, nel quale si avanzano anche alcune considerazioni sugli sviluppi futuri del lavoro.

2. Metodi per la formazione di *cluster* territoriali

I metodi che consentono di pervenire ad una zonizzazione del territorio, nel nostro caso i quartieri del Comune, sono diversi. E' senz'altro utile descrivere anche se in modo necessariamente sintetico le loro principali caratteristiche dato che queste consentono anche una loro valutazione in rapporto agli obiettivi che si vogliono perseguire.

Per semplicità chiameremo unità spaziale l'entità elementare cui si riferiscono i dati disponibili e di cui è nota la collocazione sul territorio grazie ad un sistema di coordinate riferite ad uno o più punti appartenenti all'unità. Nella maggior parte delle situazioni che ci interessano, il punto rappresenta il centroide di una superficie le cui caratteristiche espresse da indicatori quali medie, proporzioni, rapporti, ecc sono associate allo stesso centroide.
I metodi che consentono di aggregare unità spaziali in modo da partizionare un'area di più ampie dimensioni si rifanno principalmente a procedure di *cluster* analisi cui si aggiungono vincoli che impongono alle unità componenti ogni *cluster* di essere territorialmente contigue.
Le procedure che tendono all'individuazione di *cluster* spaziali hanno spesso l'obiettivo di individuare aree nelle quali un determinato fenomeno assume valori significativamente superiori a quelli che si registrano in altre parti del territorio. Ciò avviene ad esempio in campo epidemiologico quando si delineano aree nelle quali di possono osservare alti tassi di incidenza di una particolare patologia, o in campo ambientale quando la variabile obiettivo dello studio può essere rappresentata da un inquinante diversamente distribuito sul territorio. Queste tipologie di studi hanno portato allo sviluppo di metodiche di *cluster* spaziale ad hoc che tuttavia possono essere adeguatamente utilizzate, eventualmente con opportune modifiche, anche in situazioni in cui la zonizzazione ti un territorio tende soltanto ad evidenziare aree omogenee rispetto ad una batteria più o meno ampia di caratteri demografici e socio economici associabili alle popolazione residente e per i quali può essere interessante la diffusione periodica delle informazioni provenienti da fonti statistiche e/o amministrative.
Un primo metodo che possiamo prendere in esame e che ha una larga diffusione proprio in campo epidemiologico è il metodo di scansione statistica denominato SaTScan (Kulldorff, 1997). In SaTScan una "finestra circolare" (*kernel*), a raggio di lunghezza variabile tra 0 e una misura prefissata, viene spostata su una superficie collocandosi di volta in volta in una posizione equivalente ad un centroide di riferimento. Per ciascuna posizione e per un numero molto elevato di misure di raggio, viene valutata la verosimiglianza di osservare, per una variabile di riferimento, la somma dei valori interni alla circonferenza in rapporto alla somma dei valori dell'area da esaminare esterna alla circonferenza. Le unità incluse nella circonferenza che corrisponde al massimo valore della verosimiglianza formano il cluster che ha la più piccola probabilità di essere osservato per puro effetto di fattori casuali. Il procedimento viene iterato sulla superficie esterna ai cluster individuati ai passi successivi fino ad ottenere una partizione dell'intera area di studio.
SaTScan è anche il nome del software che traduce in termini informatici la procedura descritta ed è disponibile gratuitamente sul Web. Poiché è stato sviluppato per fini epidemiologici ha la caratteristica di lavorare con variabili di riferimento di tipo univariato, anche se il programma informatico contiene indicazioni per l'estensione al caso multivariato.
La scansione di tipo circolare di cui si avvale SaTScan può portare a non identificare correttamente *cluster* la cui forma sia piuttosto irregolare come possono avere zone con caratteristiche di omogeneità negli studi di carattere socio economico.
Un approccio analogo a quello appena visto è alla base del metodo denominato AMOEBA (A Multidirectional Optimal Ecotope Based Algorithm), proposto da J. Aldstadt e A. Getis (2006) nel quale la scansione territoriale non è vincolata dalla forma circolare. AMOEBA è un algoritmo proposto sia per la costruzione di matrici di pesi

spaziali, utilizzabili per la costruzione di modelli SAR, sia per la formazione di *cluster*, in base ad una variabile di studio di tipo univariato.
Tralasciando la procedura di costruzione di una matrice di pesi spaziali, la formazione di cluster avviene come segue. Per una data unità spaziale di partenza viene calcolato il valore di un indice funzione della variabile di riferimento per l'aggregazione (variabile di studio). Un possibile indice è stato proposto da Getis e Ord (1992); tale indice che indichiamo con G_i^* (vedi appendice) assume valore positivo se l'unità nella posizione *i* ha per la variabile di studio un valore superiore alla media di tutte le unità e negativo nel caso opposto. L'indice viene successivamente ricalcolato per gli aggregati di unità (aree) che si formano aggiungendo alla prima unità tutte le possibili combinazioni delle unità ad essa confinanti. In questa prima fase si aggrega alla prima unità quella combinazione di unità confinanti per la quale G_i^* ha lo stesso segno dell'unità di partenza ed è massimo in valore assoluto. Non si aggrega nessuna unità se il valore assoluto di G_i^* calcolato per l'unità di partenza resta superiore a quello relativo alle possibili aggregazioni. Se alcune unità si sono aggregate alla prima fase, si ripete la stessa operazione per tutte le unità confinanti con queste ultime, senza però ritornare a valutare quelle escluse nella fase precedente. Il processo di formazione del primo *cluster* termina quando non ci sono più insiemi di unità confinanti a quelle già aggregate che incrementano il valore assoluto dell'indice G_i^*. Formato il primo *cluster*, che può essere costituito anche dalla solo unità di partenza, si ripete la procedura partendo da una qualsiasi unità esterna ad esso. E' possibile dimostrare che la scelta dell'unità di partenza influisce minimamente sulla formazione dei cluster che possono essere di forma assai irregolare. In altri termini si ottengono aree sostanzialmente equivalenti per contenuto e forma sia che l'unità di partenza sia "centrale" rispetto all'area finale che si viene a formare sia che questa sia posizionata in prossimità del confine con un'area adiacente.
Anche AMOEBA come SaTScan è dunque sviluppato in riferimento a studi di tipo univariato. E' comunque possibile estenderlo al caso di più variabili di riferimento nell'aggregazione individuando un opportuno indice multivariato da massimizzare o minimizzare nel processo di aggregazione. Di AMOEBA non abbiamo trovato disponibile alcun software. Non si è potuto ancora implementarlo ma ciò non dovrebbe richiedere eccessive difficoltà.
Un altro interessante metodo per la formazione di *cluster* territoriali proposto recentemente da Patil e Taillie (2004) è indicato dall'acronimo ULS (*Upper Level Set scan statistic*). Si tratta ancora di uno strumento concepito per individuare aree nelle quali uno o più fenomeni presentino valori anomali o inusuali rispetto alla norma, ma può essere, come i metodi appena visti, utilizzato in generale per l'individuazione di cluster spaziali. Il metodo introdotto da Patil e Taillie tende anche ad eliminare le incongruenze che presentano i procedimenti più comuni per l'individuazione di *cluster* territoriali che procedono secondo gli algoritmi classici imponendo sul risultato finale una serie di vincoli territoriali o gli stessi vincoli durante ogni fase di aggregazione delle unità. I due modi di procedere non portano normalmente allo stesso risultato. Quando i vincoli territoriali vengono calati sul risultato di un'aggregazione effettuata con uno qualsiasi dei criteri classici si ha normalmente uno smembramento di alcune unità dai gruppi in cui sono confluite senza essere legate da vicoli di contiguità con altre unità. Se, per esempio, si volesse ottenere una partizione delle unità in *k* gruppi e si pensasse di adottare il criterio *k-means* il più delle volte si otterrebbe un numero di gruppi

maggiore di *k* poiché dai *k* gruppi derivati dalla procedura tradizionale si dovrebbero poi scorporare alcune unità per il mancato rispetto del vincolo di contiguità. Per tornare a k raggruppamenti si dovrebbe poi procedere ad una nuova aggregazione delle unità scorporate a gruppi per i quali i suddetti vincoli risultino rispettati. Nel procedimento che aggrega le unità, imponendo ad ogni passo il rispetto di vincoli di contiguità, questo non avviene e di conseguenza si mantiene il controllo sulla numerosità dei raggruppamenti della partizione finale che, però, spesso risulta diversa da quella che si avrebbe seguendo il criterio precedentemente descritto.
Nel metodo ULS il risultato non dipende dal momento in cui vengono imposti i vincoli e questo lo rende in qualche senso più coerente e preferibile rispetto ai due precedenti.
Il metodo è discusso nel prossimo paragrafo unitamente ai dati utilizzati nella sua applicazione.

3. Una prima applicazione del metodo ULS

Descriviamo il metodo ULS come un approccio alla formazione di cluster spaziali, partendo dall'ipotesi che le informazioni di riferimento nella formazione dei cluster, siano rappresentate da una variabile scalare *X*.
Supponiamo di effettuare la scansione della variabile *X* su uno spazio bidimensionale *R* partizionato in *A* aree elementari (le sezioni di censimento) che denotiamo con a_i (i = 1, 2, ...,*N*). Per ogni unità elementare indichiamo con G_i il valore di una variabile di risposta, non negativa, funzione di *X*.
Il metodo di scansione ULS identifica cluster di unità contigue che hanno elevati valori di G_i in rapporto a un valore soglia variabile tra estremi predefiniti. Per un dato valore di soglia *g*, l'insieme delle unità contigue con valore maggiore di *g*, formano una "zona" e denotato con Z_j. Utilizzando la stessa notazione di Patil et al. (2006), denotiamo questo insieme di unità con:

$$U_g = \{a_i : G_i \geq g\}$$

Prendiamo ancora dal lavoro citato un grafico che illustra come le zone si formino in corrispondenza di due diverse soglie. Il metodo individua al livello *g* le zone Z_1, Z_2 e Z_3, cioè la porzione della superficie totale che ha un valore di *G* superiore o uguale alla soglia prescelta. Quando la soglia scende al livello *g'* le zone individuate sono Z_4, Z_5 e Z_6 poiché diventa questa la superficie "eccedente" quella soglia.
La stessa figura mette anche in evidenza la casistica osservabile in corrispondenza di livelli decrescenti della soglia *g*:

(a) una nuova zona può formarsi dall'unione di due zone preesistenti;
(b) una zona preesistente ha una superficie maggiore;
(c) "emerge" una nuova zona non presente al precedente valore di soglia.

La capacità di scansione della soglia può essere ben evidenziata anche mediante un grafico ad albero quale quello riportato nella figura 2, tratto anch' esso dal lavoro citato. I nodi dell'albero (rappresentati da cerchietti) sono le unità spaziali posizionate a vari livelli in verticale a seconda del valore della variabile risposta.
Le linee che connettono i nodi indicano la presenza di una relazione di contiguità. Ad una certa soglia il nodo o i nodi più bassi individuano zone comprendenti i nodi più alti legati dai rami dell'albero.

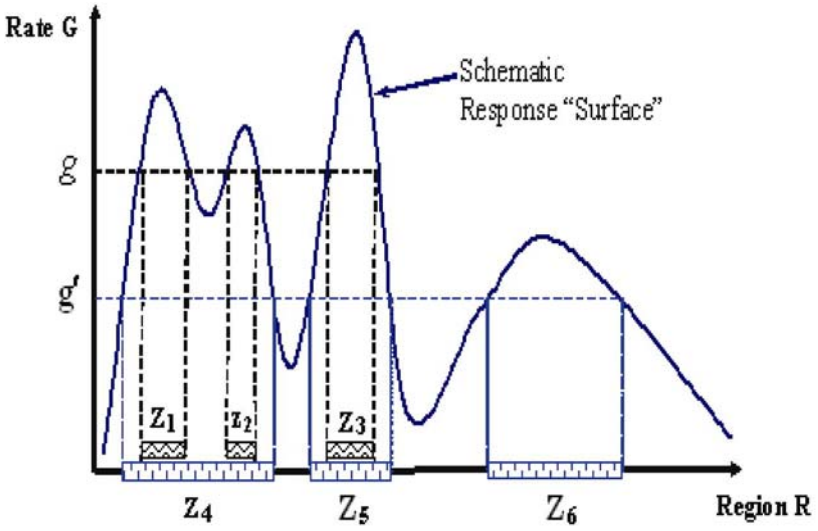


Fig. 1. Schematizzazione della superficie di risposta. Zone che si ottengono in corrispondenza a due livelli *g* e *g'* della soglia di scansione (fonte: Patil et al., 2006).

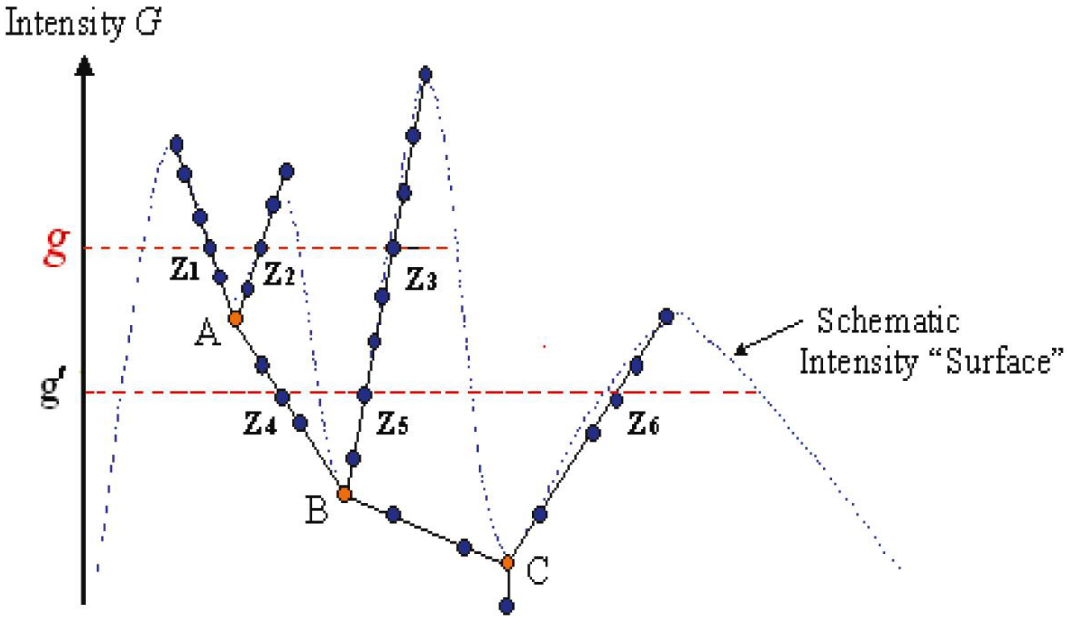


Fig. 2. Schematizzazione dell'albero ULS con cerchietti ad indicare le unità spaziali e le linee (rami dell'albero) ad indicare la contiguità spaziale tra unità ai vari livelli della soglia *g*. La linea tratteggiata schematizza la superficie di risposta (fonte: Patil et al., 2006).

Abbiamo effettuato una prima applicazione del metodo ULS a tre quartieri del Comune di Firenze avendo a disposizione una batteria di variabile e/o indicatori in massima parte provenienti dal censimento della popolazione del 2001 e in parte dall'anagrafe. Nel presente lavoro presentiamo una scelta delle numerose applicazioni per ovvi motivi di spazio. I risultati delle applicazioni sono rappresentati da carte in cui le diverse colorazioni, che comunque seguono confini delle sottostanti sezioni di censimento, rappresentano aree con un livello superiore (colorazione più intensa) o inferiore (colorazione più tenue) rispetto ad una soglia, il più delle volte individuata per semplicità nel valore mediano o in altro quartile della distribuzione della variabile risposta. Da un quartiere all'altro, la soglia delle variabili di risposta può comunque essere fissata in corrispondenza di quartili diversi.
Teniamo a precisare che le carte prodotte non devono essere lette come un primo tentativo di zonizzazione comunale ma piuttosto come una prima esemplificazione dell'applicazione di una metodica che richiede comunque una maggiore messa a punto e uno sviluppo strettamente correlato al particolare obiettivo che stiamo perseguendo. Inoltre, alcuni dei dati del censimento della popolazione del 2001 che abbiamo utilizzato, variabili di natura demografica e sulle abitazioni, hanno, ovviamente, subito evoluzioni nei cinque anni decorrenti dalla data di riferimento e conseguentemente le aree che si dovessero desumere da questi non avrebbero oggi molto valore. E' vero che una zonizzazione debba mantenere una certa validità nel tempo, ma questo appare come un motivo in più perché essa si fondi su dati aggiornati e venga poi modificata o revisionata con cadenza opportuna. Un ulteriore considerazione riguarda il fatto che, almeno allo stadio attuale della metodologia, sembra inevitabile un intervento finale sulla cartografia, non solo per la lettura dei confini delle aree, ma anche per la loro definitiva enucleazione che riteniamo non possa prescindere da quegli elementi legati alla conoscenza soggettiva del territorio che è difficile e il più delle volte impossibile trasmettere in un qualsiasi algoritmo statistico-matematico, poi tradotto in termini informatici.
Nelle successive tavole 1a,b,c, 2a,b,c e 3a,b,c abbiamo riportato i risultati della scansione territoriale ULS sui quartieri 1, 4 e 5 del Comune per le variabili: densità di popolazione (misurata dal numero di dimoranti abituali sulla superficie della sezione di censimento in m^2) e livello di istruzione (misurato come proporzione di dimoranti abituali, di oltre venti anni, in possesso di un titolo superiore a quello di scuola media inferiore sul totale dei dimoranti abituali con più di venti anni). Le carte sono state ricavate sia per ciascuna variabile (tavole 1a,b, 2a,b e 3a,b) sia come intersezione delle due variabili (tavole 1c, 2c e 3c). Queste ultime tavole prefigurano una prima possibilità di analisi di tipo non univariato anche se molto semplice in quanto riferita a due sole variabili.

4. Considerazioni conclusive

Le tavole 1a,b,c, 2a,b,c e 3a,b,c ricavate applicando in successione a due variabili il metodo di scansione statistica del territorio ULS proposto da Patil e Taillie (2004) evidenziano aree territoriali che corrispondo a seconda della colorazione più o meno intensa "zone" in cui due caratteri sono congiuntamente o singolarmente al di sopra o al di sotto di una certa soglia. Le aree non colorate (bianche) sono sezioni di censimento prive di persone abitualmente dimoranti. La soglia, indicata nella legenda della tavola, è

stata scelta anche per motivi di semplicità in corrispondenza di un quartile della distribuzione della variabile risposta e differisce da un quartiere all'altro in rapporto alla distribuzione della stessa variabile.
Le carte potrebbero rappresentare la base di una effettiva zonizzazione anche se per le considerazioni già svolte nel precedente paragrafo vengono proposte solo come esempio applicativo del metodo utilizzato, ad un particolare stadio del suo sviluppo.
Le linee di sviluppo riguardano principalmente la possibilità di includere nell'analisi un numero maggiore di variabili e conseguentemente la definizione della procedura migliore per realizzarla.
Due possibili linee di ricerca sono già evidenti e riguardano, in primo luogo, la valutazione dei risultati che si otterrebbero con un'intersezione di tre o quattro variabili (carte) anziché due come quella del presente lavoro. In secondo luogo, dovranno essere valutate variabili risposta complesse, desumibili ad esempio, dalle prime componenti principali o dal calcolo di misure di similarità in rapporto ad un "centroide" di quartiere.
Parallelamente si procederà all'implementazione di una metodologia alternativa a ULS, ad esempio AMOEBA, e alla valutazione anche per questa delle possibilità di una sua generalizzazione a variabili risposta di tipo multivariato.

5. Bibliografia

Aldstadt, J. and Getis A. (2006). Using AMOEBA to create spatial weights matrix and identify spatial clusters. *Geographical Analysis*, **38,** 327—343.

Anselin, L. (1994). Local indicators of spatial association - LISA. *Geographical Analysis,* **27**, 91-115.

Cressie, N. A. C. (1993). *Statistics for spatial data*. rev. ed. Wiley. New York.

Everitt, B., Landau, S., and Leese, M. (2001). *Cluster Analysis.* Arnold, London.

Getis A. and Ord J. K. (1992). The analysis of spatial association by distance statistics. *Geographical Analysis*, **24**, 189-206.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods*, **26**, 1481-1496.

Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: detection and inference. *Statistics in Medicine,* **14**, 799—810.

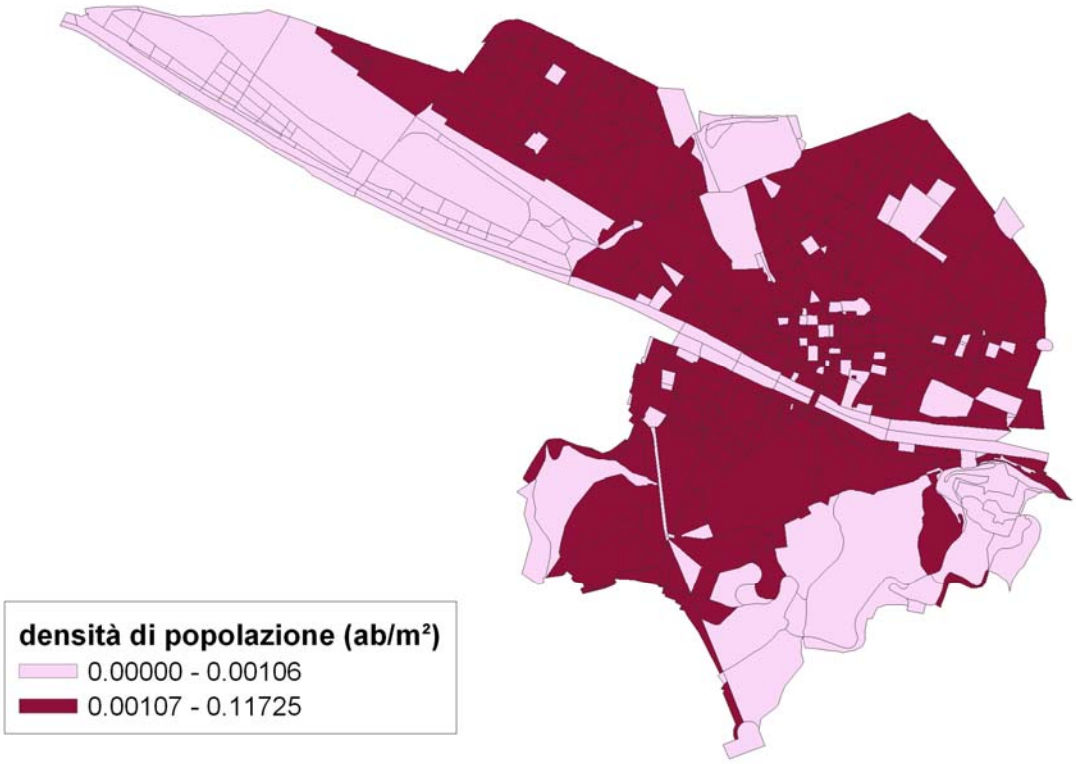
Legendre, P. (1987). Constrained clustering. In *Developments in Numerical Ecology*, P. Legendre and L. Legendre, eds. Springer-Verlag, Berlin. pp 289—307.

Modarres, R. and Patil, G. P. (2006). Hotspot Detection with Bivariate Data. To appear in the *Journal of Statistical Planning and Inference*.

Patil, G. P. and Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics,* **11**, 183—197.

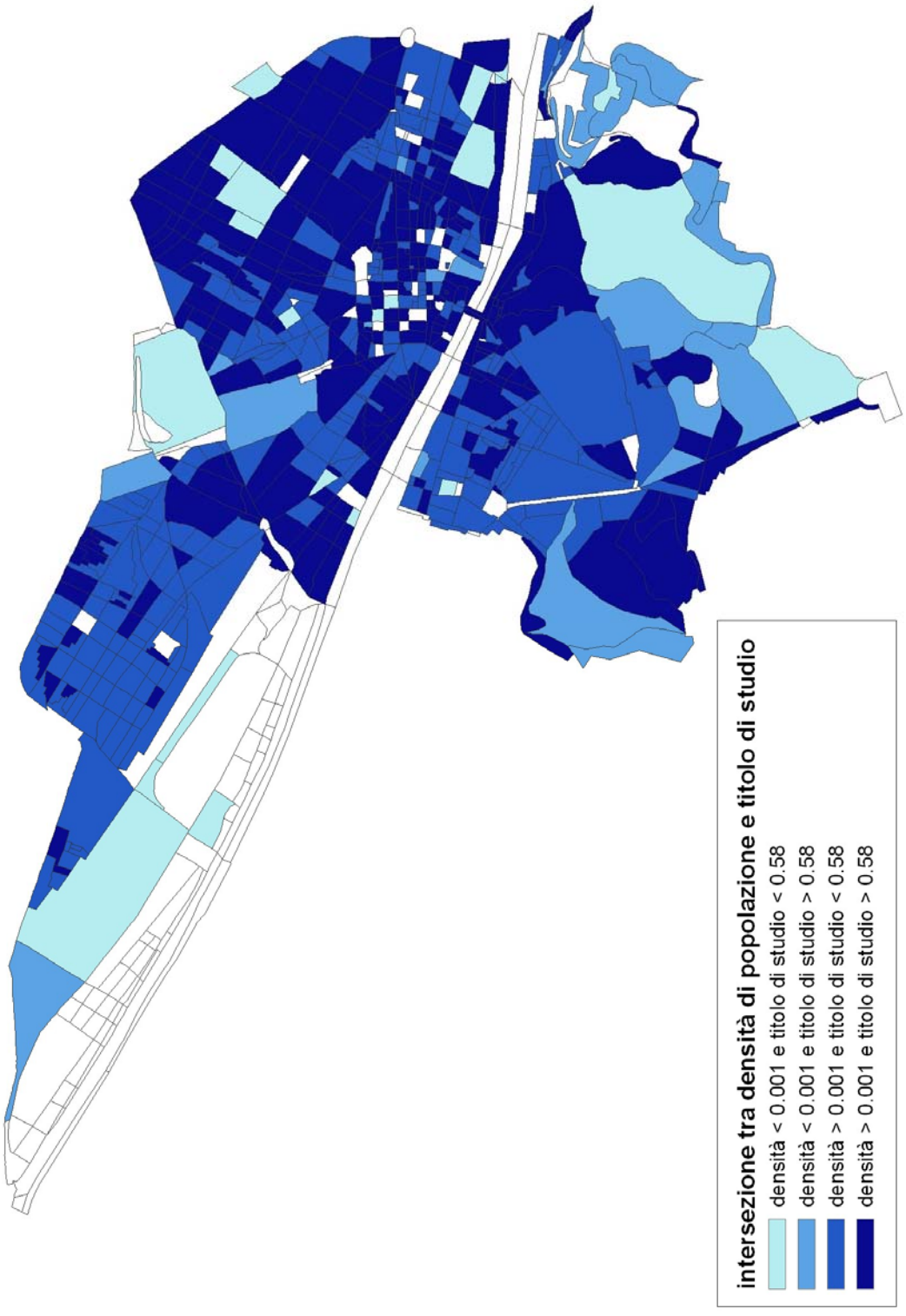
Patil, G. P.,Modarres R., Myers W. L., Patankar P. (2006). Spatially constrained clustering and upper level set scan hotspot detection in surveillance geoinformatics. *Environmental and Ecological Statistics*, **13**, 365-377.

Zani S. (1980), Alcuni contributi della statistica multivariata alla suddivisione del territorio, *Atti della XXX riunione della SIS*, Trento.

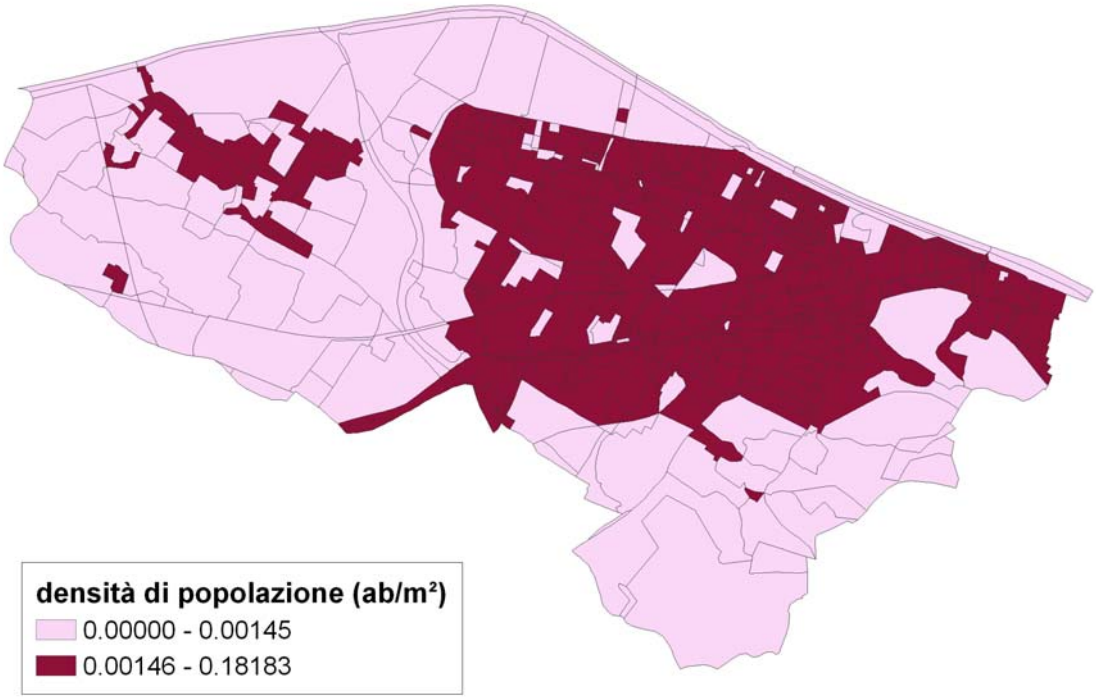


Tav. 1a. Quartiere 1, Densità di popolazione.

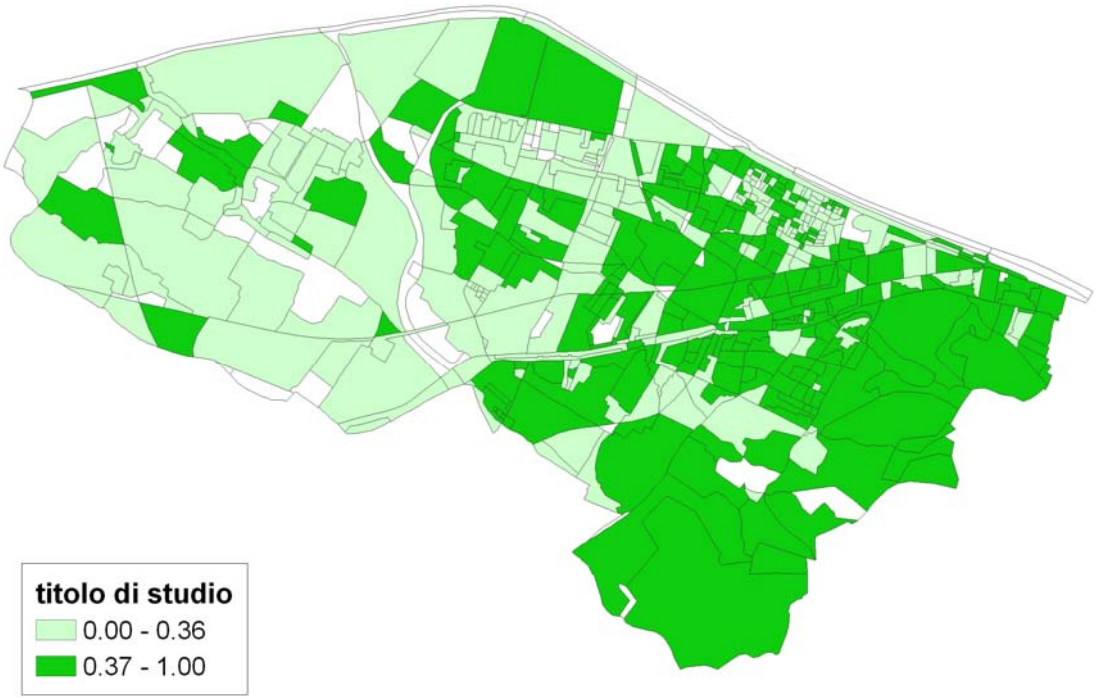
Tav. 1b. Quartiere 1, Titolo di studio.



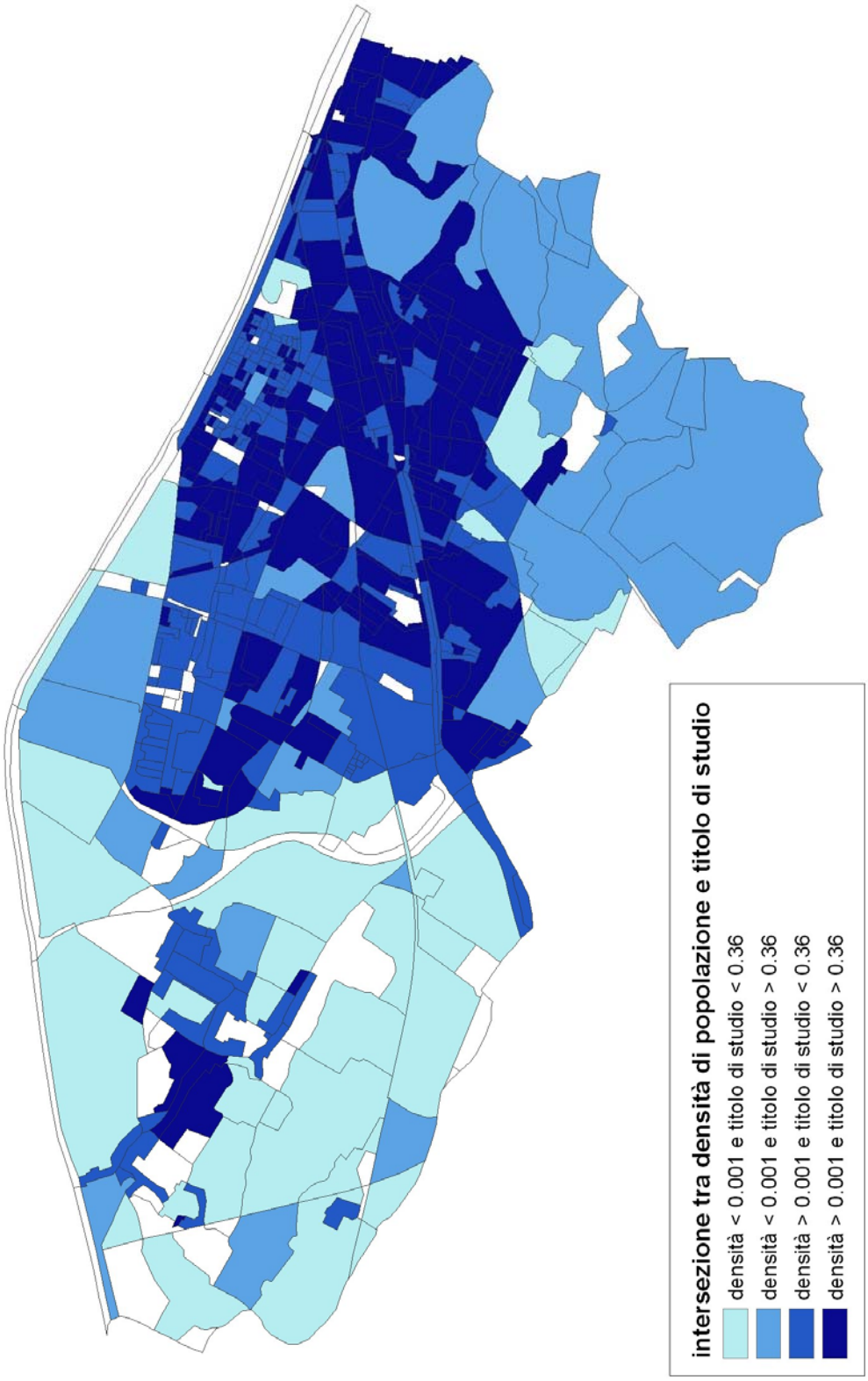
Tav. 1c. Quartiere 1, Intersezione.



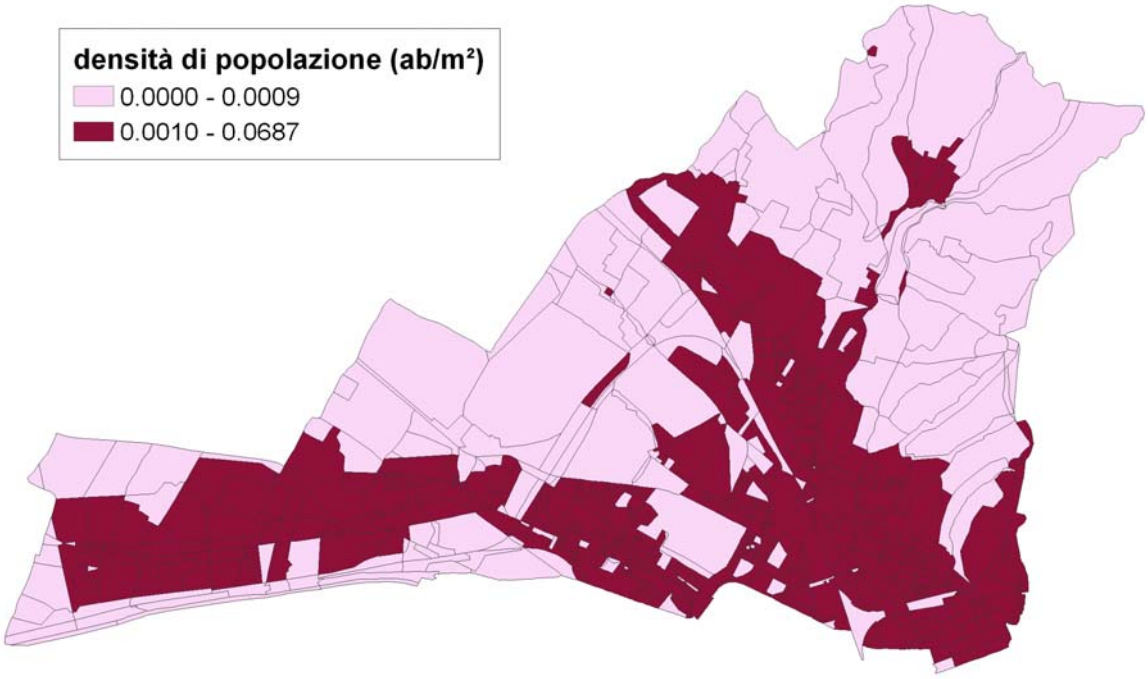
Tav. 2a. Quartiere 4, Densità di popolazione.



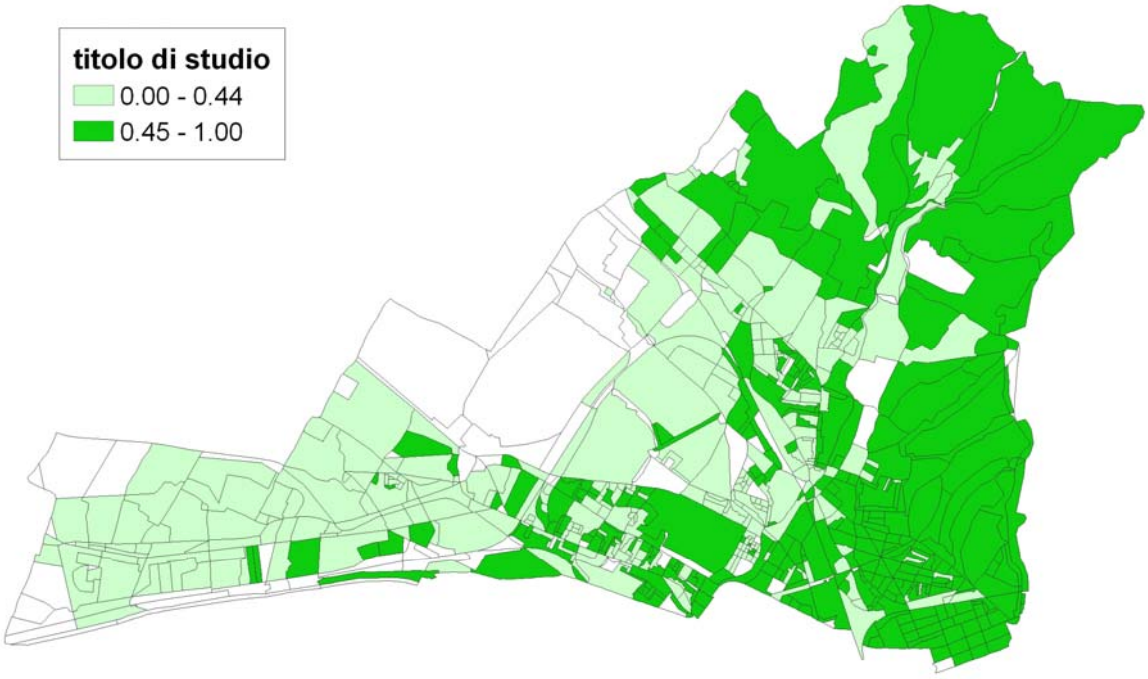
Tav. 2b. Quartiere 4, Titolo di studio.



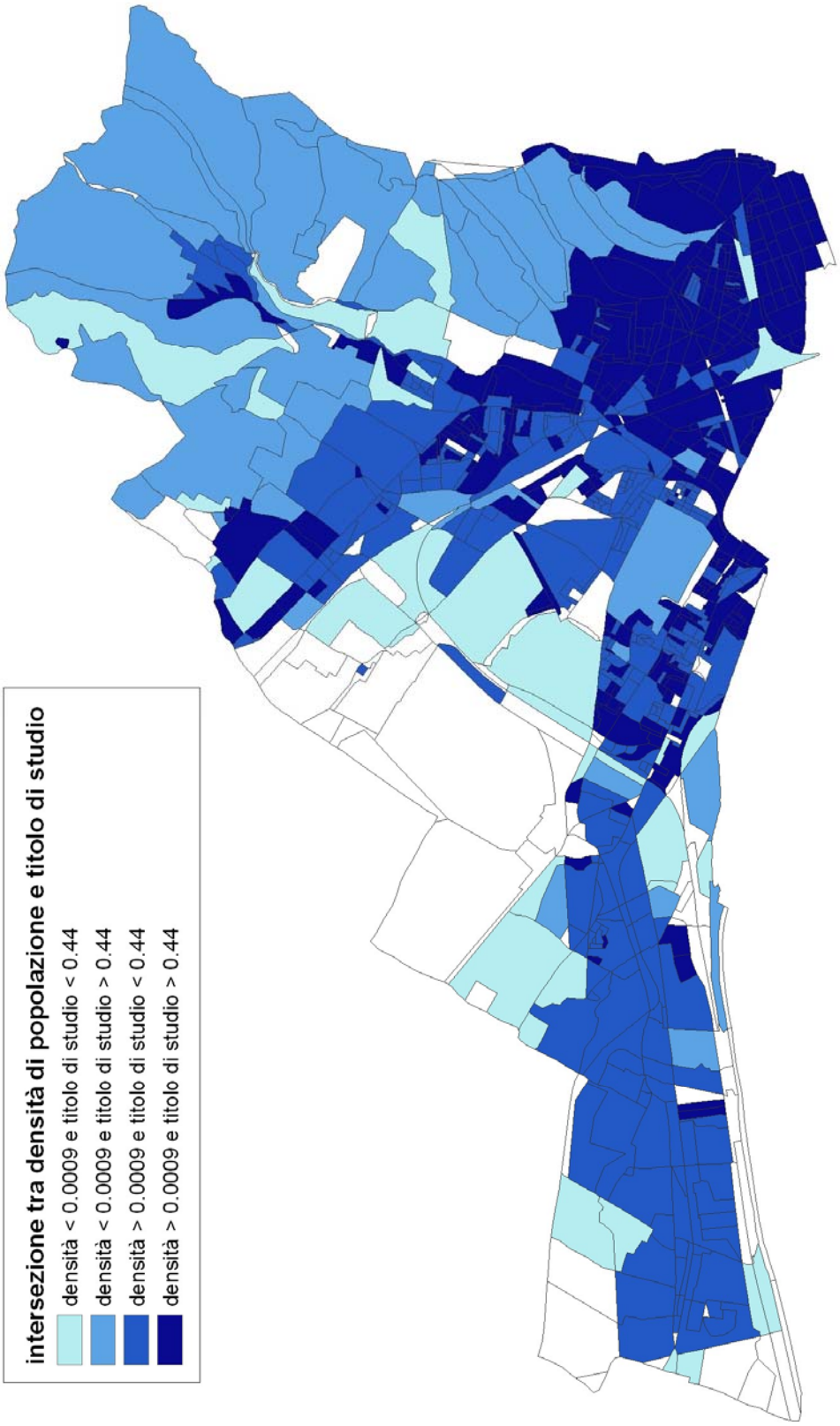
Tav. 2c. Quartiere 4, Intersezione.



Tav. 3a. Quartiere 5, Densità di popolazione.



Tav. 3b. Quartiere 5, Titolo di studio.



Tav. 3c. Quartiere 5, Intersezione.

Appendice

L'indice G_i^* di Getis e Ord è definito dalla seguente espressione:

$$G_i^* = \frac{\sum_{j=1}^{N} w_{ij} x_j - \bar{x} \sum_{j=1}^{N} w_{ij}}{S \sqrt{\frac{\left[N \sum_{j=1}^{N} w_{ij}^2 - \left(\sum_{j=1}^{N} w_{ij} \right)^2 \right]}{N-1}}}$$

Nella quale N è il numero delle unità spaziali x_j è il valore della variabile di riferimento nella posizione territoriale j, $\bar{x}$ è il valore medio della stessa variabile su tutta l'area di riferimento, w_{ij} è una variabile indicatore pari a 1 se l'unità j è nella stessa regione dell'unità i e 0 altrimenti e:

$$S = \sqrt{\frac{\sum_{j=1}^{N} x_j^2}{N} - \bar{x}^2}$$