SAPIENZA
UNIVERSITÀ EDITRICE

ANNALI DEL DIPARTIMENTO DI METODI E MODELLI PER L'ECONOMIA IL TERRITORIO E LA FINANZA

2020

ISSN: 2385-0825 (print)
Registrazione presso il Tribunale di Roma n. 247/2016 del 30/12/2016

ISSN: 2611-6634 (online)
Registrazione presso il Tribunale di Roma n. 105/2019 del 01/08/2019

Sapienza Università Editrice
Piazzale Aldo Moro 5 – 00185 Roma

www.editricesapienza.it
editrice.sapienza@uniroma1.it

Iscrizione Registro Operatori Comunicazione n. 11420

Pubblicato a dicembre 2020

Impaginazione/layout a cura del: Managing Editor

Gianluigi Salvucci*

IL *GEO-BOXPLOT*. UNO STRUMENTO PER LA SINTESI SPAZIALE

Abstract. Il *Boxplot* è un grafico statistico basato sulle caratteristiche fondamentali di una distribuzione. Introdotto da Tukey nel 1969, è largamente utilizzato per la grande forza di esemplificazione. Come noto il grafico evidenzia un rettangolo (Box) con i limiti posti sul primo e terzo quartile; vengono inoltre individuati gli scostamenti dalla mediana e i valori estremi, per visualizzare immediatamente una sintesi della forma della distribuzione. Utilizzando dati geografici, in questa breve nota si discute intorno alla possibilità di effettuare una rappresentazione analoga capace di sfruttare l'elemento geografico. Alla base di questo ragionamento vi è l'introduzione di una funzione cumulata dei valori dei dati geografici basata sulla distanza da un punto prefissato. Una volta calcolata la frequenza cumulata si può considerare il concetto di "media profonda" introdotta dallo stesso Tukey per il *Bagplot*, che consente di dividere il piano della distribuzione bivariata in due parti uguali. Lo strumento sviluppato in ambiente Arcgis, a partire da un punto rappresentativo della distribuzione geografica, il centro mediano, consente di ripartire in due parti uguali in funzio ne della distanza funzionale. Queste rappresentazioni grafiche consentono di confrontare fenomeni diversi avvenuti negli stessi territori, che possono venire sintetizzati attraverso indicatori di concentrazione basati sulle cumulate geografiche proposte.

Keywords: *Boxplot*, Concentrazione, Network Analysis

1. I cinque punti riassuntivi di una distribuzione statistica

Secondo (Tukey, 1970) una distribuzione statistica si caratterizza per 5 punti fondamentali: media, mediana, primo quartile, terzo quartile e scarto interquartilico (Borra, 2015). Questa visione estremamente sintetica ha portato alla creazione di un grafico molto potente per la semplicità con cui restituisce una visione completa delle distribuzioni statistiche, il box plot (Figura 1).

Il grafico è costituito da un rettangolo, *box*, i cui estremi sono il primo e terzo quartile, di conseguenza il lato che li unisce è lo scarto interquartilico. Il segmento che taglia in due è il punto della mediana, mentre i "baffi" sono i segmenti che contengono tutti i valori normali, oltre i quali sono considerati *outliers*. Il grafico consente rapidamente di ottenere informazioni sulla forma della distribuzione che nel caso specifico risulta asimmetrica.

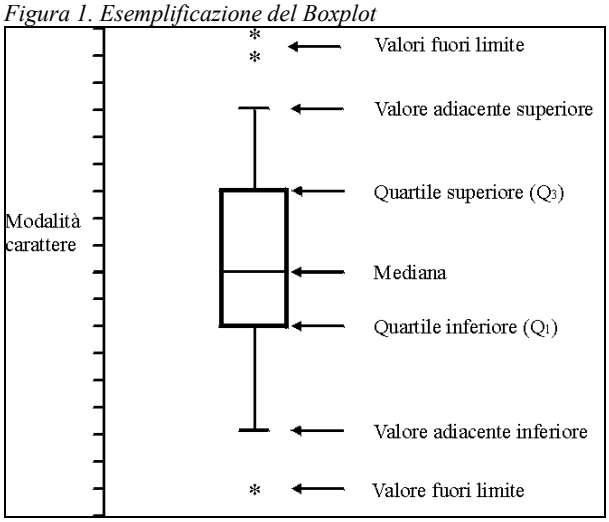
Le distribuzioni presentate nella Figura 2, pur avendo lo stesso dominio hanno una forma diversa. Quella in alto è chiaramente meno variabile rispetto quella inferiore pur avendo lo stesso intervallo di ammissibilità dei valori.

Pur non conoscendo i valori delle distribuzioni si ottengono delle informazioni che potrebbero essere sufficienti per iniziare un ragionamento sui dati a disposizione.

L'analisi di Tukey (1999) è andata oltre spingendosi sull'analisi di distribuzioni bivariate. Si è reso necessario introdurre una definizione diversa di mediana passando dal concetto di punto a quello di area. L'autore introduce il concetto di *Bagplot* sostituendo al *Box* la *Bag*, ovvero un contenitore capace di catturare la metà della distribuzione bivariata, nell'esempio cilindrata e peso di 60 veicoli (Figura 3).

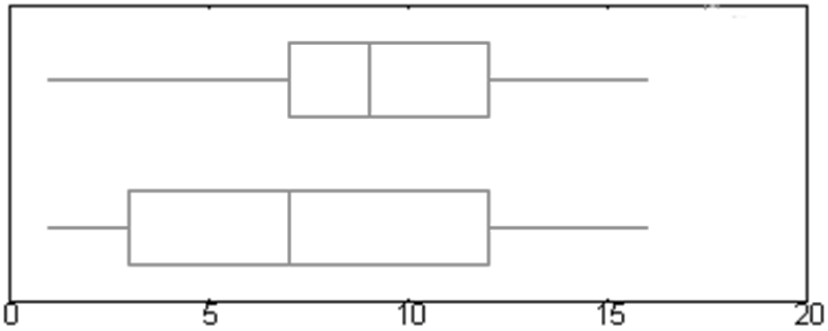
* Istat, Direzione Centrale Statistiche Ambientali, ATA - Servizio Ambiente, Territorio e Registro delle Unità Geografiche e Territoriali – salvucci@istat.it

Figura 1. Esemplificazione del Boxplot



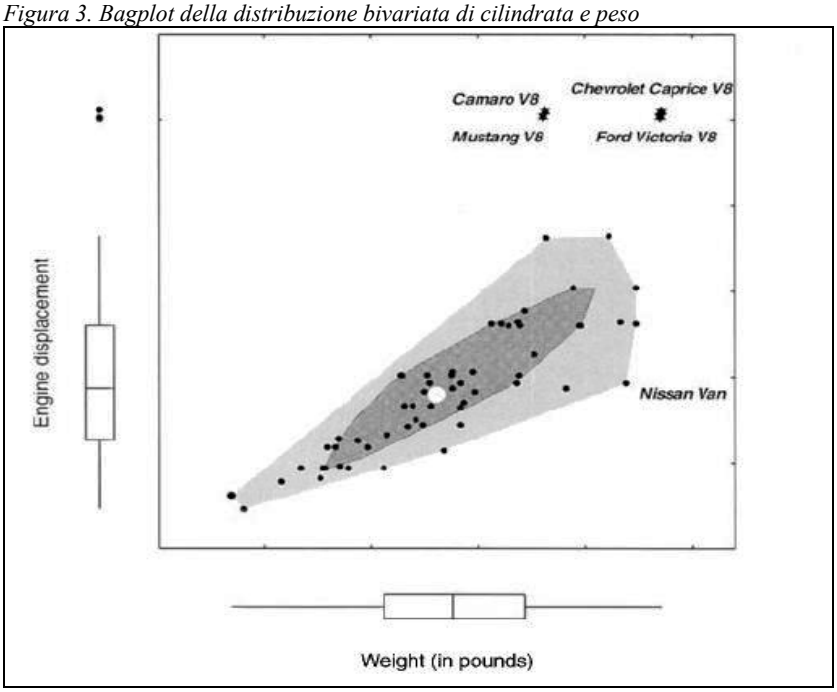
Fonte: http://cirdis.stat.unipg.it/files/Sperimentazione/Box-Plot.html

Figura 2. Boxplot, versione "baffi"



Fonte: Wikipedia

Figura 3. Bagplot della distribuzione bivariata di cilindrata e peso



Fonte: Rosseu Tukey, 1999

Il grafico pone a confronto i due *Boxplot* con la nuvola dei punti includendo nella "busta" più scura il 50% dell'intera distribuzione.

La semplicità con cui questa immagine restituisce un'idea di forma della distribuzione è talmente forte da suscitare l'interesse a trasportare questo insieme di concetti su dati territoriali. In ambito geografico non ci sono rappresentazioni che siano al tempo stesso così semplici ed efficaci. La forma di grafico più simile a questo concetto è la costruzione di una circonferenza basata sulla deviazione standard a partire dal baricentro. Questa configurazione non garantisce di individuare una certa quantità della distribuzione, dal momento che questo dipende dalla forma della distribuzione stessa. Un limite di questo strumento è quello di astrarre dalle differenze territoriali, ovvero di considerare lo spazio isotropico e non il territorio. Per sopperire a questa difficoltà viene proposto lo *Standard Ellipse* di Clark, spesso utilizzata nella configurazione a diverse deviazioni standard (Lee, 2000). In questo caso si costruiscono delle ellissi che sono orientate secondo il primo e secondo fattore della distribuzione spaziale dei punti. Lo strumento incorpora una differenziazione territoriale ma non può tener conto di una distanza funzionale, vale a dire che i punti sono dislocati sul territorio e la loro distanza è solo euclidea. Due punti vicini sulla carta, potrebbero essere estremamente lontani in termini di percorsi effettivi.

Al pari della *Standard Deviation* anche la *Standard Ellipse* di sintesi, pur mostrando il baricentro e le ellissi contenenti quote di distribuzione scelte in base alla deviazione standard, non consente di confrontare tra loro diverse distribuzioni. Questo perché il numero, o l'ammontare degli elementi contenuti nella prima deviazione non potrà essere uguale nelle diverse distribuzioni.

Nell'immenso panorama di strumenti di sintesi dell'analisi ad esempio variogrammi, variogrammi direzionali, co-variogrammi, ecc. mostrano sintesi della autocorrelazione spaziale ma non indagano sulla concentrazione della distribuzione (Ciotoli, 2005) e non sono di facile interpretazione alla stregua di un *Boxplot*.

Se si vuole indagare sul pattern spaziale può essere utile il ricorso al calcolo di indicatori sul livello di clusterizzazione, anche con la visualizzazione di Hot-Spot. Tuttavia questi tipi di analisi propongono il livello di concentrazione e le aree di maggior concentrazione senza poter individuare una porzione predefinita della distribuzione, limitandone la confrontabilità tra popolazioni diverse. Allo stesso modo la restituzione grafica della funzione K di Ripley offre un'immagine di come la distribuzione si concentri a certe distanze, ma non mostra la parte più concentrata dell'intera distribuzione (Janica, 2009).

Lo scopo dell'analisi spaziale, e della geostatistica in generale, è infatti quello di studiare il livello di autocorrelazione spaziale per rafforzare i modelli predittivi nelle diverse forme. Lo strumento proposto ha invece il mero compito di descrivere la distribuzione nel territorio del fenomeno osservato, offrendo un grado di concentrazione.

2. Le tre misure di una distribuzione geografica

Leti (1983) ha insegnato a studiare le distribuzioni aumentando via via il grado di complessità e le capacità di sintesi per i diversi tipi di caratteri: a partire dai dati qualitativi, dove ha senso studiare solo la modalità prevalente e l'omogeneità, fino ai dati quantitativi per analizzare la variabilità intorno ad un centro. Dunque, essendo un dato caratterizzato da un posizionamento geografico, quali informazioni aggiuntive ci può fornire?

La risposta è sicuramente infinita come tutta l'analisi spaziale, ma l'obiettivo è in realtà molto più ambizioso nel voler tradurre geograficamente concetti statistici basilari nella logica di Tukey.

L'obiettivo principale sarebbe semplicemente sostituire al concetto di distanza euclidea, largamente utilizzato per le serie statistiche, quello di distanza geografica funzionale, il percorso necessario a far interagire due luoghi.

Partendo dall'equivalenza tra il concetto di media e quello di baricentro, che sicuramente è condiviso e condivisibile, con dei dati geografici ha senso introdurre il concetto di cumulata?

La risposta sembrerebbe positiva, considerando che vi sono molte applicazioni che, a partire da un punto, cumulano il valore di una popolazione offrendo questa nuova distribuzione (kernel, point density, e altre). Il concetto di cumulata geografica implica un passaggio dal mero ammontare della popolazione alla superficie da essa occupata, ed in questa accezione assume una connotazione geografica: dove e quanto sia denso il fenomeno in quel punto.

L'idea sottostante alla costruzione di un *Geo-Boxplot* è quella di partire da un punto ritenuto significativo dell'intera distribuzione per individuare un'area capace di contenere il 50% della distribuzione stessa.

La cumulata che si introduce vuol essere capace di esaltare gli aspetti geografici e per questo motivo parte dal punto del baricentro percorrendo un grafo stradale. Non interessa che le aree siano contigue, la raffigurazione è di una dimensione parallela in cui la distanza geografica, espressa in minuti di percorrenza, prevale sulla contiguità, definendo lo spazio di interazione di quel sub collettivo.

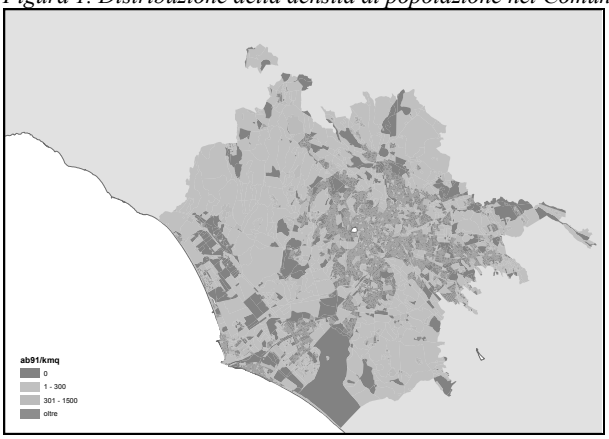
Applicando questa metodologia ai dati geografici si ottengono delle caratteristiche significative, al pari di quanto avviene per le serie statistiche. In questo caso il punto di riferimento da considerare è il baricentro, da cui si parte per la costruzione del *Geo-Boxplot*. Il rettangolo del *Boxplot* viene sostituito da superfici di territorio che potremmo definire geo-quartili, in particolare si propone di considerare l'area del primo quartile e quella del secondo che costituiscono l'area mediana geografica, ovvero quella mediana profonda introdotta da Tukey per il *Bagplot*. Questa procedura consente di creare un parallelismo con lo *Standard Ellipse* di Clark che viene calcolato solitamente per la prima e seconda Deviazione Standard, infatti se la distribuzione dei dati fosse normale, allora con *lo Standard Ellipse* si considererebbero circa il 66% della distribuzione, mentre con questa procedura se ne considera il 50%.

L'area mediana è dunque la porzione di territorio che contiene la metà della popolazione considerata nell'analisi. Si tratta della superficie che a partire dal baricentro e per ordine di tempo riesce a raggiungere il 50% della distribuzione.

3. Un esempio di applicazione: il confronto temporale di popolazione e addetti nel Comune di Roma

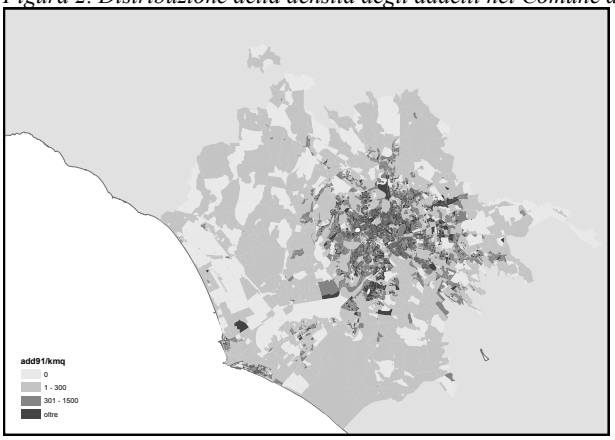
Il confronto tra le due distribuzioni geografiche della popolazione e degli addetti nel Comune di Roma, 1991 e 2011 (Figure 4 e 5) rappresenta un valido esempio di questo ragionamento. Sicuramente una carta per densità con risoluzione a livello di sezione di censimento è esplicativa della diffusione di un fenomeno quale la popolazione, così come per quella degli addetti. In entrambe le raffigurazioni è evidente che esiste una città più compatta, sede del lavoro eppure non si riesce ad individuarla nettamente.

Figura 1. Distribuzione della densità di popolazione nel Comune di Rom



Fonte: elaborazione su dati Istat,1991

Figura 2. Distribuzione della densità degli addetti nel Comune di Roma

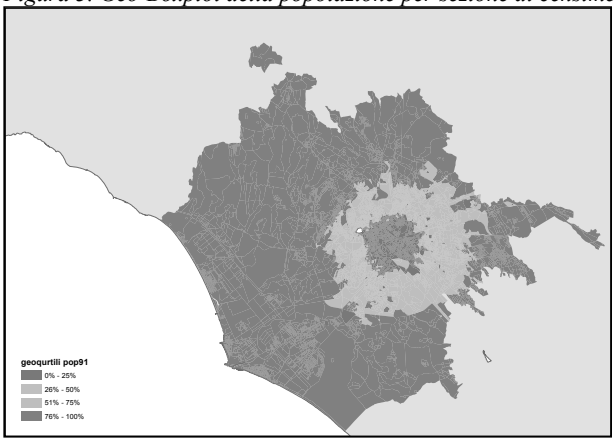


Fonte: elaborazione su dati Istat,1991

4. Il *Geo-Boxplot*

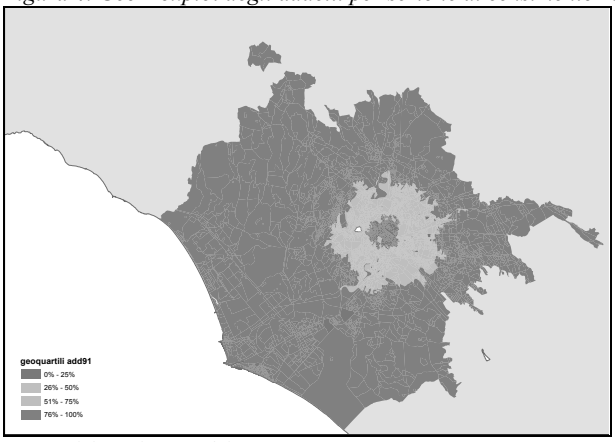
Il passaggio dal *Boxplot* al *Geo-Boxplot* consiste nel sostituire alla scatola una porzione di territorio. Pur avendo le stesse classi di densità le due distribuzioni, di cui al paragrafo precedente, non sono paragonabili richiedendo ulteriori analisi per capire dove e quanto siano evidenti le differenze. La proposta è dunque rappresentata nel confronto delle figure Figura 3 e 4, dove si vede il risultato dell'analisi delle frequenze cumulate sulla base dei percorsi effettivi dal baricentro della distribuzione ad ogni singola sezione.

Figura 3. Geo-Boxplot della popolazione per sezione di censimento 1991



Fonte: elaborazione su dati Istat,1991

Figura 4. Geo-Boxplot degli addetti per sezione di censimento 1991



Fonte: elaborazione su dai Istat,1991

5. Dal *Geo-Boxplot* alla concentrazione spaziale

Il confronto delle distribuzioni della densità degli addetti per sezione nel comune di Roma, in diverse epoche, ci consente di verificare come il fenomeno sia variato nel tempo. Se si confrontano le due carte proposte in Figura 5 e 6, per quanto sia evidente l'espansione nel centro degli addetti, si possono fare diverse considerazioni.

Figura 5 Distribuzione della densità degli addetti nel Comune di Roma (1991)

Fonte:elaborazione su dati Istat,1991

Figura 6 Distribuzione della densità degli addetti nel Comune di Roma (2011)

Fonte:elaborazione su dati Istat, 2011

L'esperimento condotto vuole essere qualcosa in più di una semplice rappresentazione spaziale. Utilizzando i dati della cartografia esposta è possibile costruire un indice di concentrazione alla stregua dell'indice di concentrazione del Gini.

Infatti guardando la distribuzione della cumulata questa si comporta come la curva di Lorenz, con la differenza che si utilizza una distanza temporale.

Rapportando la superficie occorrente a coprire il primo 50% della distribuzione sulla superficie totale si ottiene un rapporto di composizione che è un numero puro e pertanto paragonabile:

Tabella 1. Calcolo dell'Indice di Concentrazione per il 1991

	Addetti	**Popolazione**
Superficie geomediana	61,30	120,38
Totale	1499,62	1499,62
Indice di Concentrazione	4,1%	8,0%

La superficie geomediana degli addetti per il 1991 è pari a 61 km quasi la metà di quella che serve per la popolazione, ne deriva che è possibile affermare che la popolazione romana è dispersa il doppio di quanto accade per gli addetti. Questo semplice ragionamento risulta molto utile per comprendere come si sia evoluto il sistema urbano.

Il fatto che la superficie della geomediana passi da 61 a 94 km significa che c'è stata una diminuzione della concentrazione di quasi la metà del dato iniziale. Ma questo fatto ha riguardato solo il mercato del lavoro?

Tabella 2. Calcolo dell'Indice di Concentrazione per gli addetti

	1991	**2011**
Superficie geomediana	61,30	94,43
Totale	1499,62	1501,28
Indice di Concentrazione	4,1%	6,3%

Considerando i numeri indice con base fissa a quella del 1991 l'incremento di superficie della geomediana che si è registrato per il mercato del lavoro è inferiore di quanto avvenuto per la popolazione. La città si è espansa, tuttavia il mercato del lavoro si è disperso più di quanto abbia fatto quello residenziale rispetto al loro centro naturale.

Tabella 3. indici di concentrazione a confronto

	Addetti		**Popolazione**	
	1991	**2011**	**1991**	**2011**
Superficie geomediana	61,30	94,43	120,38	165,39
Totale	1499,62	1501,27	1499,62	1501,27
Indice di Concentrazione	4,1%	6,3%	8,0%	11,0%
Indice di Concentrazione		1,540		1,373

6. Aspetti critici

La costruzione di questo strumento risente di alcune criticità. Il calcolo di un quantile con un dato suddiviso in classe presuppone un'ipotesi sulla sua distribuzione all'interno della classe e quindi l'individuazione del suo esatto valore. Al momento, tuttavia, non si è in grado di "spezzare" la singola partizione per individuare la porzione di sezione che contiene il 50% esatto degli individui del collettivo.

Supponiamo che 100 individui siano disposti secondo un'unica strada a spirale che parte dal baricentro. Nella Figura 7, con il ragionamento esposto prendendo solo i primi due centri si ottiene una quota di popolazione pari al 30% mentre includendo il terzo si arriva al 70%. In questi casi, se non si può passare ad un dato geografico con una risoluzione migliore che consenta di aggiungere ai primi due centri parte del terzo bisogna effettuare una scelta.

Nel caso teorico si potrebbe sostenere che in entrambe le scelte lo scarto dalla mediana potrebbe essere il medesimo, per cui paradossalmente risulterebbe indifferente. Non si esclude in un prossimo futuro la possibilità di sostituire al terzo poligono una nuvola teorica di punti tali da consentire di individuare esattamente il poligono da includere. Allo scopo si potrebbe pensare di integrare una distribuzione casuale di punti quanti sono gli elementi da collocare piuttosto che distribuirli negli edifici o lungo le strade (Figura 8).

Figura 7. Ipotesi teorica di approssimazione del calcolo

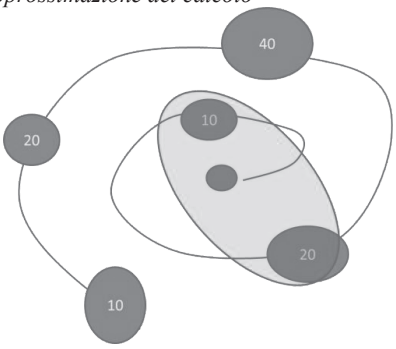
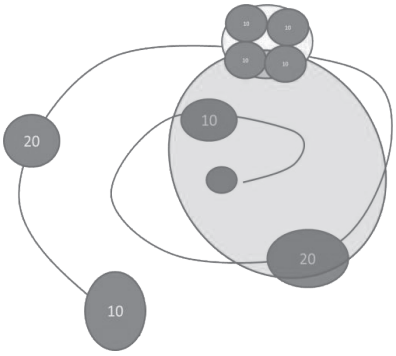


Figura 8. ipotesi migliorativa della geomediana aumentando la risoluzione dei dati con distribuzioni teoriche



7. Conclusioni e sviluppi futuri

L'obiettivo di questa breve nota è stato la proposta di uno strumento che, analogamente al *Boxplot*, consente per i dati geografici di rendere confrontabili fenomeni diversi, almeno dal punto di vista della concentrazione spaziale.

La rappresentazione del *Geo-Boxplot* si basa sulla cumulata geografica, che consiste sostanzialmente nell'individuare una superficie capace di contenere una determinata quota della popolazione. Questa trasposizione del concetto di cumulata dalla statistica consente, in un ragionamento geografico, di spostare l'attenzione dall'ammontare del collettivo, nel caso di una distribuzione statistica, alla superficie necessaria a contenerlo, nel caso di un dato geografico. Dal momento che lo spazio assunto è quello di un piano, a differenza delle modalità di una distribuzione che si muovono su una linea, si rende necessario introdurre un metodo per calcolare la funzione cumulata che viene calcolata ponendo in ordine crescente di lontananza le unità statistiche dal baricentro.

Utilizzando una distanza funzionale, la funzione cumulata incamera quelle differenze territoriali in termini di accessibilità che consentono di ottenere aree contenenti quantili della distribuzione che assumono forme non omogenee.

Restano alcuni punti critici, ovviamente. Uno di questi sta nel fatto che le superfici considerate non discriminano esattamente gli individui in due gruppi perché questo dipende dalla risoluzione dei dati.

Lo strumento proposto può, comunque, avere utilizzi diversi dalla semplice rappresentazione di dati. Ad esempio, introducendo un punto significativo a scelta dell'utente si potrebbe capire il ruolo di quel punto nell'evoluzione territoriale; si pensi ad esempio all'ipotesi che si voglia introdurre una nuova infrastruttura, o un servizio, l'uso di questo strumento potrebbe consentire di verificare quali modifiche nella concentrazione di popolazione o di addetti potrebbero registrarsi.

Riferimenti bibliografici

BORRA S. (2015), *Statistica: metodologie per le scienze economiche e sociali*, McGraw-Hill Education, Milano.

CIOTOLI G. (2005), *Dalla statistica alla geostatistica : introduzione all'analisi dei dati geologici e ambientali*, Aracne, Roma.

JANICA G.M. (2009), *Spazio e misura : introduzione ai metodi geografico-quantitativi applicati allo studio dei fenomeni sociali,* Università degli studi di Siena, Siena.

LEE J. (2000), GIS and statistical analysis with ArcView, John Wiley, New York.

LETI G. (1983), *Statistica descrittiva*, Il Mulino, Bologna.

ROUSSEEUW P.J., RUTS I. and TUKEY J.W. (1999), The Bagplot: A Bivariate Boxplot, *The American Statistician* 53(4), 382–387.

TUKEY J. (1970), *Exploratory data analysis. Preliminary edition,* Addison-Wesley, Reading, Massachusetts.