STATISTICA, anno LXII, n. 3, 2002

METODI NON PARAMETRICI MULTIVARIATI: UN'APPLICAZIONE AL CASO DELLA CRESCITA

B. Pacini, G. Pellegrini

1. INTRODUZIONE

La presenza di polarizzazioni e asimmetrie nella distribuzione spaziale del reddito, della ricchezza e dell'occupazione è stata solo recentemente considerata come uno degli elementi che concorrono ai processi di crescita. Da un lato, la "nuova geografia economica", sviluppatasi dai lavori di Krugman (1991), ha mostrato come esternalità pecuniarie e non pecuniarie e spillover spaziali possano portare a una distribuzione delle attività economiche disuguale sul territorio; dall'altro, lo sviluppo di tecniche econometriche indirizzate all'analisi spaziale, legato in particolare ai lavori di Anselin (1988), ha permesso di misurare l'intensità dei legami economici nello spazio.

L'importanza di aggregazioni, polarizzazioni e cluster nello spazio e degli effetti che essi possano avere sullo sviluppo del territorio ha quindi guadagnato base teorica e sostegno empirico nella letteratura internazionale. In Italia, d'altronde, questo non ha destato particolare sorpresa, dato che la presenza di cluster spaziali è sempre risultata evidente almeno rispetto a due dimensioni: una dimensione macro, a livello regionale e circoscrizionale, segnalata dalla presenza di una vasta area territorialmente contigua, riconducibile per larga parte nei confini storici del Mezzogiorno, con livelli di reddito e di occupazione ancora sensibilmente inferiori a quelli del resto d'Italia, e il cui processo di convergenza, arrestatosi all'inizio degli anni settanta (Fabiani e Pellegrini,1999), sembra essere ripreso solo negli ultimi anni; una dimensione micro, a livello locale, individuata dalla presenza di forti difformità di performances nelle imprese di piccola e media dimensione inserite in cluster specializzati, chiamati distretti industriali, rispetto a imprese di pari dimensione e settore ma isolate nel territorio. Quest'ultimo aspetto appare di particolare interesse in quanto suggerisce la presenza di interazioni ed esternalità che si formano tra imprese e tra aree limitrofe a livello locale che possano condizionare le prospettive di sviluppo sia del territorio di localizzazione sia anche di quelli contigui. La ricerca si è quindi indirizzata a studiare la dinamica nel tempo della distribuzione spaziale di grandezze rappresentative dei processi di crescita di

un'area e dei fattori responsabili di addensamenti e polarizzazioni nell'evoluzione delle stesse distribuzioni.

Questo ha richiesto lo sviluppo di tecniche di analisi statistica ed econometrica per molti versi innovative. Quah (1997a,b) ha proposto di utilizzare tecniche non parametriche per individuare la dinamica dell'intera distribuzione di variabili esemplificative dei processi di crescita. In particolare, queste tecniche permettono di verificare: l'esistenza, in un contesto statico, di agglomerazioni e cluster spaziali che si manifestano in una multimodalità della distribuzione[1]; la presenza, in un contesto dinamico, di mobilità all'interno della distribuzione; le variabili che, utilizzate per condizionare la distribuzione, possano modificarne la forma. L'analisi prospettata da Quah si basa sull'utilizzo di due diversi strumenti metodologici: da un lato suggerisce il ricorso ad un operatore matematico, detto *kernel* stocastico, che consente di rappresentare la dinamica delle distribuzioni descrivendo l'evoluzione tra distribuzioni diverse; dall'altro predispone uno schema di condizionamento (*conditioning*) che consente di spiegare la distribuzione osservata in funzione di variabili esplicative ritenute rilevanti. Lo schema è analogo a quello di un modello di regressione, ma concettualmente la variabile da spiegare consiste in questo contesto nell'intera distribuzione di probabilità, e, in particolare, nelle sue caratteristiche di forma e nell'eventuale presenza di comportamenti anomali.

La tecnica di *conditioning* ha riscosso seguito nella recente letteratura[2], tuttavia è stata applicata soltanto in contesti di tipo univariato, ovvero prendendo in considerazione una sola variabile alla volta per studiarne gli effetti sulle caratteristiche della distribuzione. Questo è attribuibile principalmente alla natura del metodo proposto: sebbene l'approccio consenta teoricamente l'utilizzo di più variabili per il condizionamento, la metodologia utilizzata in letteratura non ammette l'estensione al caso multivariato. Lo schema univariato non è tuttavia soddisfacente per molti aspetti: la forma della distribuzione di una variabile di crescita economica è infatti influenzata da più fattori contemporaneamente. Disporre di un approccio di tipo multivariato al problema consente di tenere conto congiuntamente dei diversi fattori in gioco e di poterne valutare l'impatto netto.

Scopo del presente lavoro è proporre una versione multivariata della tecnica originariamente proposta da Quah, che risulti anche di facile implementazione. L'idea di base è quella di sfruttare le analogie tra lo schema di *conditioning* e la problematica della stima non parametrica di una funzione di regressione basata sull'utilizzo di approssimatori locali. Seguendo questa linea di ricerca, prenderemo in considerazione alcuni stimatori locali di regressione che, opportunamente inseriti all'interno della struttura di condizionamento, saranno applicati allo studio degli effetti di spillover spaziali sulla crescita. Come è stato precedentemente sottolineato, l'economia italiana è un eccellente campo di prova di queste applicazioni. La variabile di cui vogliamo studiare la distribuzione spaziale è il tasso di occupazione (non agricolo), che approssima per molti versi il grado di

[1] Il modello esemplificativo più noto è quello detto di *twin peaks*; si veda Quah (1997a).
[2] Oltre ai successivi lavori di Quah, si rinvia ad esempio a Overmann e Puga (1999).

sviluppo di un territorio, inteso come la capacità di un'area di generare posti di lavoro e quindi reddito per chi ci vive. Per meglio misurare la presenza di effetti spaziali, l'analisi viene condotta a un livello territoriale fine: si è scelta la griglia dei sistemi locali del lavoro (SLL), una suddivisione territoriale basata sulle caratteristiche di autocontenimento del mercato del lavoro locale e quindi particolarmente adatta ai nostri scopi. La distribuzione del tasso di occupazione per i 784 SLL italiani (nel 1996) viene quindi messa in relazione con variabili che colgono la struttura geospaziale, lo sviluppo strutturale e la presenza di spillover di prossimità per SLL, con lo scopo di verificarne l'influenza sulla forma della distribuzione e spiegarne l'eventuale multimodalità.

La struttura del lavoro è la seguente: nella sezione 2 si richiama lo schema di condizionamento proposto da Quah e lo si reinterpreta individuando un contesto più generale, che include la possibile esistenza di fenomeni di relazione tra fattori localmente variabili nonché l'eventuale presenza di errori di misura con effetti locali; nella sezione 3, si propone una metodologia alternativa alla luce dello strumento della regressione non parametrica, che viene estesa nella sezione 4 al caso multivariato con riferimento a due possibili diversi metodi di approssimazione locale già presentati nella sezione precedente; nella sezione 5 trovano spazio i risultati empirici sulle caratteristiche della distribuzione del tasso di occupazione per SLL, con riferimento, in primo luogo, al caso univarato e, successivamente, alla valutazione dell'effetto congiunto di più fattori sulla forma della distribuzione. Alcune considerazioni conclusive sono riportate nella sezione 6.

2. LO SCHEMA DI CONDIZIONAMENTO UNIVARIATO

In questa sezione ci proponiamo di reinterpretare la tecnica di *conditioning* al fine di inserirla in un contesto più ampio, all'interno del quale collocare metodi alternativi per la valutazione dell'effetto di variabili rilevanti sulla distribuzione oggetto di studio.

Lo schema proposto originariamente da Quah (1997b) può essere formalizzato nel seguente modo: data una variabile casuale X (il tasso di occupazione, nella nostra applicazione), osservata su un insieme I (i SLL italiani, sempre nel nostro caso),

$$X = \{X(i) : i \in I\} \tag{1}$$

definiamo una nuova variabile, che indichiamo con X^{**} , ottenuta come una media ponderata delle osservazioni originarie su X con pesi $w(i)$ funzione di altre variabili ritenute rilevanti:

$$X^{**} \equiv \sum_{i \in I} w(i) X(i) . \tag{2}$$

Lo schema di condizionamento, che indichiamo con *G*, richiede la definizione di un insieme di pesi *W* e la specificazione di una relazione tra *X* e *X*** come segue:

$$X^{*} = X \mid G \equiv \varphi \left\{ X(i), X^{**}(i) \right\}, \tag{3}$$

dove *X** è la variabile condizionata e φ è l'operatore che descrive la relazione funzionale che sussiste tra le osservazioni originarie e i dati ottenuti come media locale ponderata. La relazione funzionale che lega le due variabili *X* e X^{**} è di tipo deterministico, come quanto proposto nella formulazione di Quah (1997b), e quindi possiamo identificare questo come uno *schema di condizionamento deterministico*.

Lo schema *G* sostanzialmente consiste della specificazione di due diversi elementi: *W* e φ. L'insieme di pesi *W* determina l'influenza di ciascuna osservazione *x(i)* contenuta in *I* sul valore *x**(j)*. Ad esempio, possiamo assumere come variabile condizionante la distanza tra i diversi mercati del lavoro, che quindi determinerà la struttura di ponderazione: i mercati *i*-esimi più vicini alla *j*-esima osservazione (sulla base di una specificata misura di distanza d_{ij}) avranno maggiore rilevanza nella stima di *x**(j)* rispetto alle altre osservazioni:

$$x_j^{**} \equiv \sum_{i \in I} \left(\frac{n \cdot d_{ij}}{\sum_j \sum_i d_{ij} \cdot x_i} \right) \cdot x_i \,. \tag{4}$$

I pesi rappresentano, quindi, l'importanza relativa della distanza di ciascuna osservazione da *x(j)*. Consideriamo, sempre nel nostro esempio, che i pesi siano ottenuti semplicemente come reciproco della distanza da uno specifico mercato locale del lavoro e che la distanza sia calcolata tra i centroidi dei due mercati locali. Quindi, in questo caso, la procedura standard di condizionamento si basa sulla creazione per ciascun SLL di un SLL che potremmo definire "gemello", dove il tasso di occupazione "condizionato" è pari ad una somma ponderata dei tassi di occupazione di tutti i mercati locali, opportunamente normalizzata dividendo per il tasso di occupazione condizionato medio. Per la densità di popolazione, nel caso più semplice, possiamo supporre un peso pari a 1 per lo specifico SLL, e pari a 0 in tutti gli altri casi. Come operatore φ consideriamo, seguendo il suggerimento di Quah(1997b), il rapporto tra dati originari e dati ponderati, specificando la seguente relazione:

$$x_j^{*} = \varphi \left(x_j, x_j^{**} \right) = \frac{x_j}{x_j^{**}} \,. \tag{5}$$

Non essendo richiesto nello schema la specificazione della forma funzionale della distribuzione di *X**, l'approccio utilizzato è di tipo non parametrico. L'analisi

si concentra quindi sulle caratteristiche della distribuzione di probabilità di questo rapporto rispetto a quella iniziale (X), al fine di mettere in evidenza eventuali cambiamenti nella forma delle funzione di densità di probabilità dovuti allo schema di condizionamento adottato. Questi vengono individuati attraverso semplici analisi grafiche o più formalizzate procedure di verifica d'ipotesi sulla simmetria e/o unimodalità della distribuzione.

Si ritiene utile, a questo stadio della trattazione, identificare un quadro di condizionamento più generale all'interno del quale poter specificare diversi modi di procedere nella definizione dei due elementi che caratterizzano lo schema di condizionamento G, e cioè i pesi W e l'operatore φ. Lo schema di condizionamento di tipo deterministico, nel quale la specificazione della relazione tra osservazioni originarie e variabile condizionante non contempla la presenza di un errore casuale che può determinare variazioni locali, può essere considerato come caso particolare di uno schema di condizionamento più ampio, di tipo stocastico, che inserisce nello schema (3) anche un termine di errore casuale nel seguente modo:

$$X^{*}(i) = \varphi\left\{X(i), X^{**}(i), \varepsilon(i)\right\}. \tag{3.a}$$

Tale rappresentazione è, a nostro parere, di portata più generale, in quanto può rendere conto di ulteriori fenomeni: da un lato, ad esempio, della presenza di una relazione tra due fenomeni che sia localmente variabile, rappresentabile nel seguente modo:

$$X^{*}(i) = \varphi\left\{X(i), X^{**}(i) + \varepsilon(i)\right\}, \tag{3.a.1}$$

e, dall'altro, della possibilità che la stessa variabile X sia affetta da errore di misura, situazione che può essere rappresentata all'interno della struttura di condizionamento come:

$$X^{*}(i) = \varphi\left\{X(i) + \varepsilon(i), X^{**}(i)\right\}. \tag{3.a.2}$$

Dello schema generale, lo schema deterministico è quindi un caso particolare, nel quale il termine casuale ha varianza nulla. Un'adeguata considerazione dal punto di vista statistico della presenza della componente stocastica (qui ipotizzata, per semplicità, di tipo additivo) si rende necessaria al fine di contestualizzare e focalizzare ulteriormente l'analisi.

3. UNA METODOLOGIA ALTERNATIVA

Pur mantenendo il carattere non parametrico dell'approccio, esistono modi alternativi di selezionare la struttura dei pesi: il problema del condizionamento può essere riformulato mediante il ricorso alla letteratura sulla stima non parametrica locale di funzioni di regressione (si veda, tra gli altri, Wand e Jones, 1995).

Assumiamo che la media condizionata di X dato $Z=z$ possa essere ottenuta come media ponderata localmente dei valori assunti dalla variabile X con pesi che dipendono dalle realizzazioni della variabile condizionante Z:

$$\hat{m}(z) \equiv \mathrm{E}(X \mid Z = z) \equiv X^{**} \equiv \sum_{i \in I} w(i, Z) X(i) . \tag{6}$$

Questa è l'espressione generale di uno stimatore non parametrico locale di una funzione di regressione. La scelta di approssimatori di tipo locale, oltre ad essere a nostro parere l'estensione più immediata dell'approccio originario sopra presentato, si dimostra particolarmente idonea allo studio in oggetto, in quanto stiamo cercando di evidenziare fenomeni di eterogeneità (vale a dire peculiarità locali) nella distribuzione del tasso di occupazione e non siamo, invece, interessati ad individuare relazioni medie che rimangono costanti all'interno dell'intero campione di SLL. Inoltre in questo modo è possibile tenere conto nelle stime della possibilità di shock stocastici nella relazione tra la variabile condizionante e la variabile condizionata.

In letteratura esistono molti diversi possibili stimatori che soddisfano la (6): la classe di stimatori *kernel* fornisce procedure di stima piuttosto semplici, la cui trattabilità matematica costituisce elemento di forte attrattiva sul piano applicato. Facendo esplicitamente riferimento al tradizionale stimatore di regressione di Nadaraya e Watson, possiamo definire i seguenti pesi:

$$w(i, Z) = \frac{k_h(z - z_i)}{\sum_{i \in I} k_h(z - z_i)} \text{ ,dove } k_h(z - z_i) = \frac{1}{h} k\left(\frac{z - z_i}{h} \right). \tag{7}$$

Lo stimatore di regressione per la j-esima osservazione risulta dalla seguente espressione:

$$\hat{m}(z_j) = \frac{\sum_{i \in I} k_h(z_j - z_i) \cdot x_i}{\sum_{i \in I} k_h(z_j - z_i)} \tag{8}$$

dove h è il parametro, detto *bandwidth*, che controlla il grado di lisciamento (*smoothing*) della funzione, e k è la funzione *kernel*, che determina la forma della funzione di ponderazione (spesso una funzione di densità di probabilità simmetrica e unimodale, come ad esempio la densità normale).

Considerando le espressioni dalla (6) alla (8) risulta evidente l'analogia con lo schema di condizionamento precedentemente presentato: possiamo, infatti sostituire $\hat{m}(z_j)$ a X^{**} nell'equazione (3) e nella (5). Nel fare ciò, il criterio che stiamo osservando nella scelta dei pesi è il seguente: nel determinare il j-esimo dato ponderato si assegna un peso maggiore alle osservazioni per le quali la variabile Z assume valore vicino a z_j. Quando andiamo a calcolare il rapporto indicato nell'espressione (5), dividiamo ciascun tasso di occupazione per una media delle

osservazioni che presentano caratteristiche similari rispetto alla variabile condizionante. Ad esempio, se consideriamo come variabile esplicativa la densità di popolazione nei mercati locali del lavoro, ciascun tasso di occupazione sarà rapportato ad una media di osservazioni relative a SLL che presentano più o meno la stessa densità di popolazione.

Un approccio alternativo, e più generale, al calcolo di valori medi ponderati localmente è dato dagli stimatori di regressione lineari locali, che possono essere interpretati come soluzione del seguente problema di minimi quadrati ponderati localmente:

$$\min_{(\alpha,\beta)} \sum_{i\in I} \left[x_i - \alpha - \beta(z_j - z_i)\right]^2 k_h(z_j - z_i) \tag{9}$$

e danno luogo a stimatori per i quali $\hat{m}(z_j)$ si ottiene come:

$$\frac{1}{n}\sum_{i\in I} \frac{\left[\sum_{i\in I}(z_j - z_i)^2 k_h(z_j - z_i) - \sum_{i\in I}(z_j - z_i)k_h(z_j - z_i)\right]\cdot k_h(z_j - z_i)\cdot x_i}{\sum_{i\in I}(z_j - z_i)^2 k_h(z_j - z_i)\sum_{i\in I} k_h(z_j - z_i) - \left[\sum_{i\in I}(z_j - z_i)k_h(z_j - z_i)\right]^2} \tag{10}$$

espressione riconducibile allo stimatore di Nadaraya e Watson semplicemente considerando il solo termine costante nella soluzione del problema di minimo. Naturalmente ulteriore flessibilità è ottenibile generalizzando a stimatori di regressione polinomiale locale, cioè incrementando il numero di termini polinomiali nella (9).

4. ESTENSIONE AL CASO MULTIVARIATO

La naturale estensione del quadro teorico presentato nella sezione precedente consiste nella considerazione di due o più variabili condizionanti che contribuiscano congiuntamente alla determinazione della variabile media ponderata X^{**}. Questo permette di sviluppare l'analisi anche in un contesto multivariato. Supponiamo, ad esempio, di essere interessati a mostrare se e come la vicinanza e la densità di popolazione possano congiuntamente influenzare la forma della distribuzione del tasso di occupazione. In tal caso, riproponendo la struttura precedente, X^{**} dovrà essere, per ciascuna osservazione, una media ponderata delle diverse realizzazioni del tasso di occupazione con pesi *w(i)* nella (2) ottenuti come funzione di entrambe le variabili, le vicinanza geografica e la densità di popolazione.

I metodi di stima locale non parametrica per funzioni di regressione godono di sufficiente flessibilità da consentirci di estendere facilmente l'insieme di variabili condizionanti, senza necessità di assumere alcun tipo di relazione tra dette variabili. Nel contesto in cui ci muoviamo, infatti, ciò costruisce un elemento necessario, in quanto non si dispone di sufficiente informazione per specificare una relazione di tipo parametrico.

Al tempo stesso, rimanere in ambito non parametrico consente un guadagno in robustezza rispetto a possibili diverse specificazioni di modelli parametrici. Consideriamo ancora una volta lo stimatore di regressione *kernel* di Nadaraya e Watson, che può essere facilmente esteso al caso multivariato considerando un vettore $\mathbf{Z}$ di d variabili esplicative nella espressione della media locale ponderata:

$$\hat{m}(\mathbf{z}) \equiv E(X \mid \mathbf{Z} = \mathbf{z}) \equiv \sum_{i \in I} w(i, \mathbf{Z}) X(i) , \tag{11}$$

e definendo la struttura di ponderazione attraverso una matrice di *bandwidth* d-dimensionale (simmetrica e definita positiva) H e una funzione *kernel* d-dimensionale k , ottenendo il seguente stimatore:

$$\hat{m}(\mathbf{z}) = \frac{\sum_{i \in I} k_H(\mathbf{z} - \mathbf{z}_i) \cdot x_i}{\sum_{i \in I} k_H(\mathbf{z} - \mathbf{z}_i)} \tag{12}$$

dove $k_H(\mathbf{z}_i - \mathbf{z}) = \mid H \mid^{-1/2} k(H^{-1/2}\mathbf{z})$.

In modo analogo possiamo generalizzare lo stimatore di regressione lineare locale, includendo nella funzione da minimizzare un vettore $\mathbf{Z}$ di variabili esplicative. La possibilità di disporre di un modello con un'unica funzione *smooth* di tutti i predittori (senza imporre alcuna restrizione sulla relazione tra predittori stessi) garantisce una notevole flessibilità. Si osservi, tuttavia, che la dimensione del vettore di variabili esplicative è strettamente connessa ad un problema tipico della stima non parametrica, noto in letteratura come *curse of dimensionality*, che determina una notevole minore attendibilità delle stime della media condizionata, nonché una maggiore difficoltà computazionale, all'aumentare del numero di regressori. Per il caso bivariato (al quale peraltro limiteremo la nostra applicazione), lo stimatore si ottiene , ancora una volta, come soluzione del seguente problema di minimi quadrati ponderati localmente:

$$\min_{(\alpha,\beta,\gamma)} \sum_{i \in I} \left[x_i - \alpha - \beta(z_{1j} - z_{1i}) - \gamma(z_{2j} - z_{2i}) \right]^2 k_h(z_{1j} - z_{1i}, z_{2j} - z_{2i}) \tag{13}$$

Come ulteriore semplificazione, anziché ricorrere alla specificazione di una generica funzione *kernel* bivariata possiamo considerare il caso particolare del prodotto di due funzioni *kernel* univariate (si veda, ad esempio, Bowman e Azzalini, 1997). Le proprietà degli stimatori localmente lineari, peraltro, sono particolarmente apprezzate rispetto allo stimatore classico di Nadaraya e Watson, in quanto mostrano una minore distorsione in corrispondenza dei valori estremi dei predittori.

5. RISULTATI EMPIRICI

Nell'analizzare le caratteristiche della distribuzione del tasso di occupazione per i SLL italiani (con riferimento all'anno 1996), le variabili di condizionamento utilizzate (e per le quali riportiamo i risultati ottenuti) sono state le seguenti:

a) *spillover di prossimità*, misurati per ogni SLL come somma ponderata dei tassi di occupazione degli SLL rimanenti con pesi ottenuti come reciproco della distanza dal SLL di riferimento. Si osservi che altre misure di prossimità (utilizzando ad esempio una misura di distanza quadratica o esponenziale) hanno fornito risultati analoghi;

b) *sviluppo strutturale,* misurato dalla quota di occupazione agricola;

c) *struttura geospaziale,* misurata dalla densità di popolazione.

Si è inizialmente proceduto al condizionamento a ciascuna delle variabili singolarmente prese, utilizzando uno stimatore *kernel* di N.W. univariato, con scelta ottimale della *bandwidth* e funzione *kernel* quartica, per la determinazione delle variabili ponderate X^{**}. I valori della funzione di regressione stimati sono, quindi, assunti come variabile X^{**} e l'operatore φ corrisponde al rapporto tra dati originari e valori ponderati.

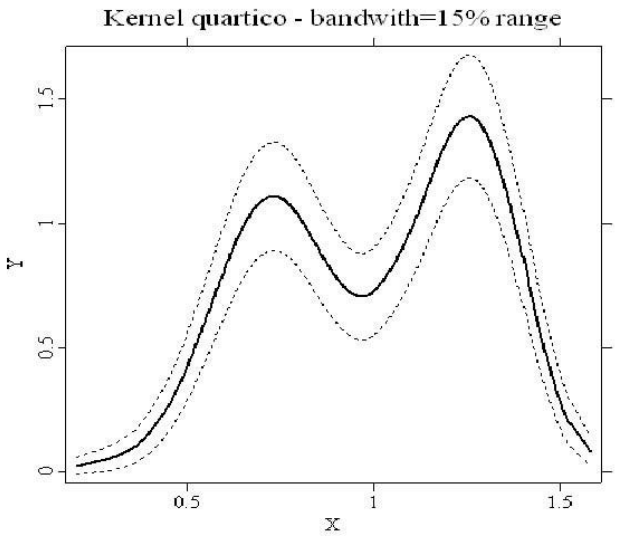

Figura 1 – Distribuzione del tasso di occupazione (rispetto alla media) per SLL (anno 1996).

Al fine di visualizzare le caratteristiche delle distribuzioni ottenute si è poi proceduto ad una stima di densità *kernel* con bande di confidenza al 95% (basate sull'approssimazione asintotica della distribuzione dello stimatore in termini globali). La distribuzione non condizionata è riportata in Figura 1 e mostra una evidente bimodalità, che scompare quando si procede al condizionamento rispetto alla presenza di *spillover di prossimità* o allo *sviluppo strutturale* (rispettivamente nelle Figure 2 e 3). In entrambi i casi, la distribuzione si concentra attorno al valore centrale, pur mantenendo caratteristiche di leggera asimmetria (in partico-

lare nel secondo caso). Un diverso comportamento si osserva condizionando alla *struttura geospaziale*: come si vede in Figura 4 questa variabile non spiega la forma della distribuzione del tasso di occupazione; infatti la bimodalità permane anche dopo il condizionamento e non si osservano variazioni significative nella caratteristiche della distribuzione. In Tavola 1 sono riportate, per completezza rispetto all'analisi grafica, alcune statistiche sulle distribuzioni non condizionata e condizionate, relativamente, in particolare, a curtosi, simmetria e normalità delle diverse distribuzioni.

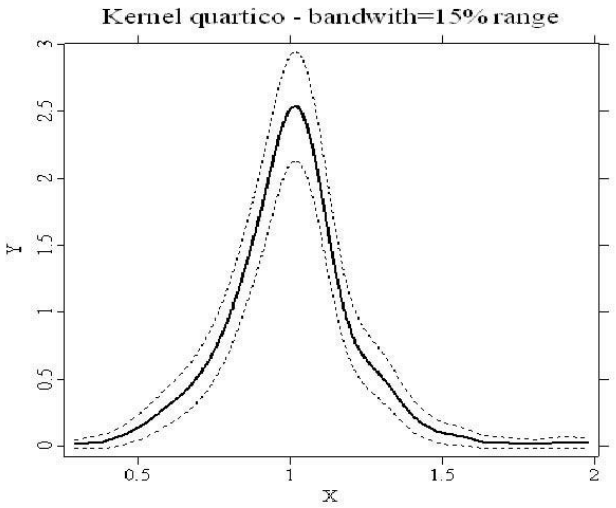

Figura 2 – Distribuzione condizionata alla presenza di *spillover di prossimità*.

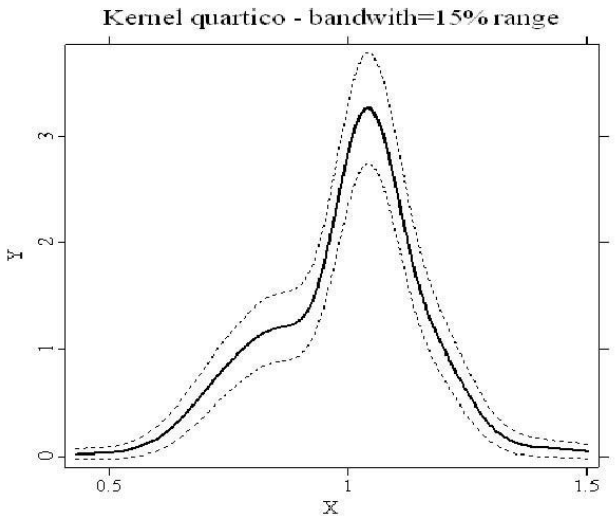

Figura 3 – Distribuzione condizionata allo *sviluppo strutturale*.

Al fine di analizzare l'effetto congiunto dei due fattori *spillover di prossimità* e *sviluppo strutturale*, i quali hanno entrambi effetti sulla bimodalità della distribu-

zione, si è utilizzato uno stimatore di regressione lineare locale, in particolare nella versione LOESS (originariamente proposto da Cleveland, 1979 e, più recentemente, ripreso tra gli altri in Fan e Gijbels, 1996), che prevede un ammontare variabile di *smoothing* per tenere conto del diverso addensamento dei dati all'interno del campo di variazione dei predittori. La distribuzione condizionata risultante dall'applicazione dell'operatore φ è mostrata (ancora una volta mediante lo strumento della stima non parametrica delle funzione di densità con bande di confidenza al 95%) in Figura 5. Il condizionamento alle due variabili congiuntamente considerate determina una un'elevata concentrazione attorno al valore centrale e una sostanziale simmetria della distribuzione (si veda anche Tavola 1).

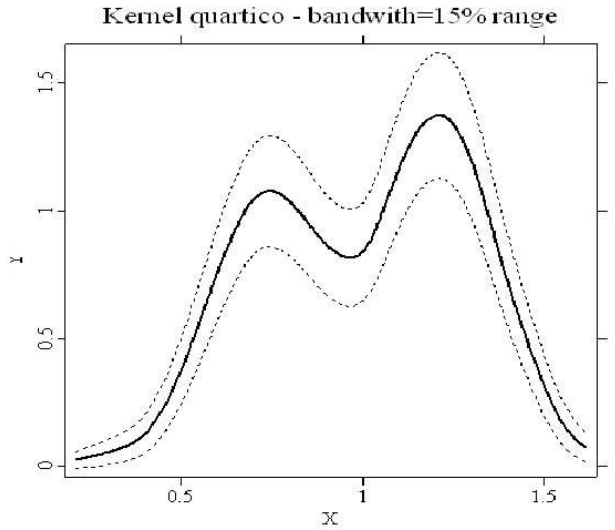

Figura 4 – Distribuzione condizionata alla *struttura geospaziale*.

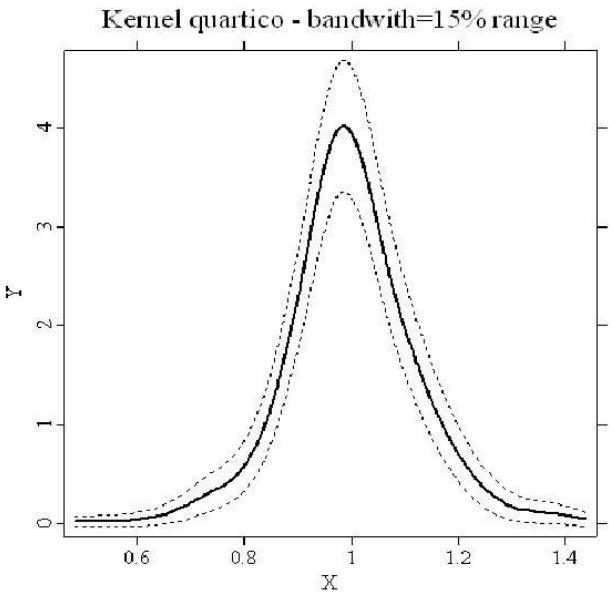

Figura 5 – Distribuzione condizionata agli spillover di prossimità e allo sviluppo strutturale.

TAVOLA 1

Caratteristiche delle distribuzioni non condizionata e condizionate

Variabile	*Distribuzione non condizio-nata*	*Distribuzione condizionata*			
		Spillover spaziali	Sviluppo strutturale	Struttura geospaziale	Spillover e Svil. Strutt.
Media	0.9999	0.9998	0.9985	1.0041	0.9993
Mediana	1.0406	1.0082	1.0268	1.0434	0.9925
Dev. Standard	0.2907	0.2045	0.1590	0.2829	0.1204
Skewness	-0.2395	0.4169	-0.3380	-0.1967	0.0110
Curtosi	1.8899	5.5043	3.4968	2.0706	4.6448
Test D	0.1067	0.0835	0.1029	0.0930	0.0587
Lilliefors	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Test A*	15.495	6.6960	8.2454	9.5885	4.4773
Anderson-Darling	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Indice I di Moran (*)	0.923	0.456	0.471	0.889	0.417
	(71.7)	(35.5)	(36.7)	(69.1)	(32.5)
Indice G di Geary (*)	0.133	0.415	0.510	0.147	0.555
	(35.7)	(24.4)	(20.2)	(35.2)	(18.5)

(*) Tra parentesi si indica il valore dell'indice standardizzato sotto l'ipotesi di distribuzione normale

Il processo di condizionamento agli spillover di prossimità e, in misura minore, ai fattori di sviluppo strutturale modifica la forma della distribuzione in quanto scomputa parte almeno degli effetti dovuti alla prossimità. Una ulteriore verifica di questo risultato proviene dall'applicazione di test di autocorrelazione spaziale. Nel lavoro sono stati utilizzati per questo l'indice I di Moran e l'analogo indice G di Geary.[3] Entrambi gli indicatori segnalano come l'autocorrelazione spaziale della variabile del tasso di occupazione si riduca sensibilmente dopo il condizionamento, pur rimanendo significativa (Tavola 1). La presenza di effetti di autocorrelazione spaziale anche dopo il condizionamento testimonia come esistano altri fattori agglomerativi sul territorio che influenzano la distribuzione (es. la presenza di cluster industriali, di disponibilità di infrastrutture e altre forme di capitale sociale, la specializzazione settoriale). Si noti comunque come l'autocorrelazione spaziale sia minore nel caso del condizionamento multivariato, che evidentemente riesce a scomputare fattori di aggregazione territoriale specifici alle due variabili considerate e tra loro perlomeno parzialmente non correlati.

Le Figure 6, 7, 8 e 9 forniscono una visuale alternativa dei risultati ottenuti: in ciascuna di esse viene riportata una stima di densità bivariata che mostra una sorta di "evoluzione" della distribuzione, da condizionata a non condizionata, con riferimento alle variabili condizionanti già analizzate precedentemente (singolarmente e congiuntamente). La chiave di lettura di queste rappresentazioni è la seguente: nei casi in cui la variabile condizionante non risulta rilevante nella determinazione della forma della distribuzione allora la densità bivariata tende a

[3] La distanza geografica tra i SLL è stata calcolata sulla base della localizzazione spaziale dei comuni centroidi. Nel calcolo degli indici si è assunto come indicatore dello spazio di prossimità la distanza massima, ovvero la distanza tra centroidi tale che tutti i SLL avessero almeno un SLL contiguo. Risultati analoghi si ottengono utilizzando altre misure di distanza (es. distanza media e mediana).

disporsi sulla diagonale del piano definito da valori originari e valori condizionati: ciò significa che la struttura della distribuzione del tasso di occupazione rimane la stessa (è il caso di Figura 8), ovvero a tassi di occupazione non condizionati elevati continuano a corrispondere tassi di occupazione condizionati elevati. Nei casi in cui, invece, si tratti di variabili significative (singolarmente o congiuntamente) nella spiegazione del *pattern* di occupazione, allora la forma campanulare si allontana in vario modo dalla diagonale indicando cambiamenti sostanziali nella caratteristiche della distribuzione. In particolare, gli spillover di prossimità influenzano la distribuzione specie per quanto riguarda le classi più elevate della variabile, con un effetto positivo. Il modello quindi attribuisce parte rilevante delle performances occupazionali di SLL a alta occupazione agli effetti di spillover: infatti, condizionando a questi, le performances sarebbero inferiori (Figura 6).

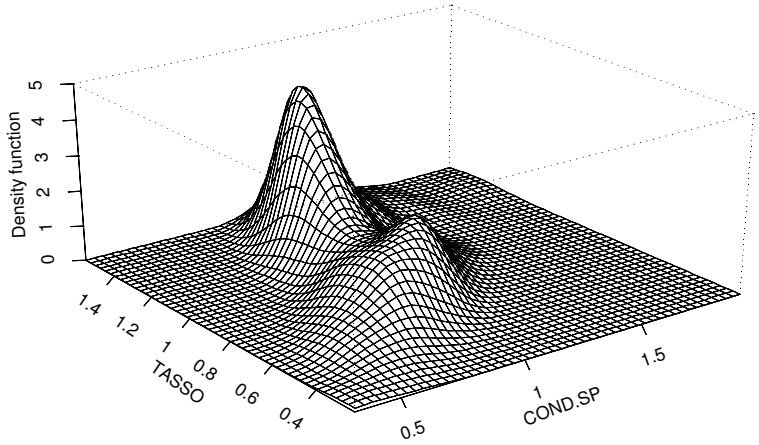

Figura 6 – Evoluzione da non condizionata a condizionata agli *spillover di prossimità*.

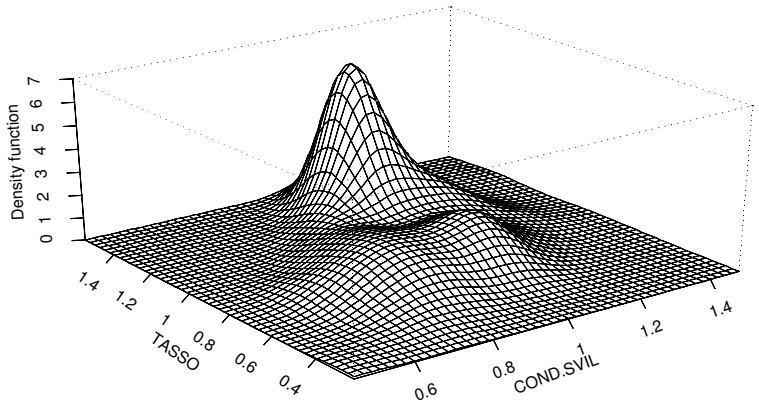

Figura 7 – Evoluzione da non condizionata a condizionata allo *sviluppo strutturale*.

L'effetto dei fattori strutturali risulta rilevante specie per le aree con tasso di occupazione relativamente basso: infatti, se si potesse prescindere dalla struttura settoriale, le performances occupazionali previste dal modello sarebbero migliori. Condizionando a entrambe le variabili la mobilità tra classi (rappresentata

dall'allontanamento dalla bisettrice) è evidente, segnalando ridotta o nulla correlazione tra la distribuzione non condizionata e quella condizionata.

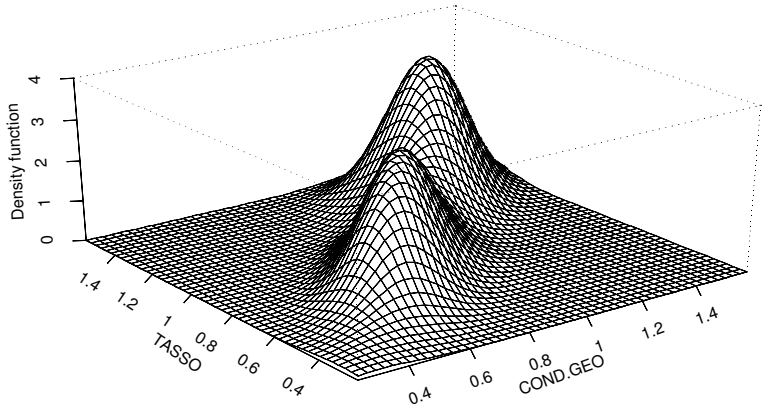

Figura 8 – Evoluzione da non condizionata a condizionata alla *struttura geospaziale*.

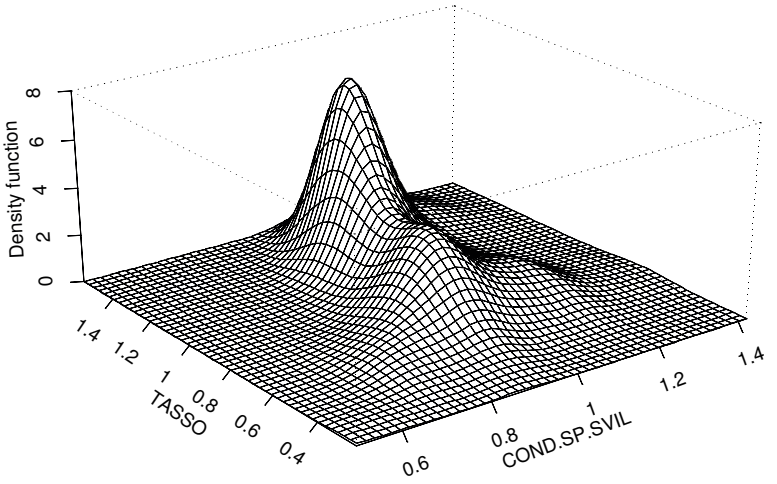

Figura 9 – Evoluzione da non condizionata a condizionata a *spillover* e *sviluppo strutturale*.

6. CONCLUSIONI

Nel presente lavoro la metodologia di analisi nota come *conditioning*, proposta da Quah (1997a,b) per l'identificazione dei fattori da ritenersi responsabili della presenza di meccanismi di sviluppo e di polarizzazione nella forma e nella dinamica della distribuzione di variabili di crescita economica, è stata reinterpretata ed estesa alla luce delle analogie con le tecniche di regressione non parametrica, basate su criteri di approssimazione locale.

In questo nuovo contesto ha trovato spazio anche la formalizzazione di uno schema di condizionamento di portata più ampia, che ammette la presenza di relazioni non solo deterministiche ma anche stocastiche. Questo ha permesso lo sviluppo di tecniche capaci di analizzare l'effetto congiunto di più variabili sulla forma della distribuzione.

Sul piano empirico le potenzialità dell'approccio proposto sono state applicate allo studio della distribuzione del tasso di occupazione nei 784 sistemi locali del lavoro italiani per l'anno 1996. I risultati ottenuti mostrano la rilevanza di alcune variabili (singolarmente e congiuntamente) nello spiegare alcune peculiarità presenti nella distribuzione non condizionata.

In particolare, si è osservato come il condizionamento agli *spillover di prossimità* e allo *sviluppo strutturale* sia in grado di eliminare la caratteristica di bimodalità presente nella distribuzione originaria e accentui la concentrazione della distribuzione stessa, segnalando in tal modo il ruolo di queste variabili nella spiegazione di polarizzazioni di crescita. L'effetto congiunto si mostra ancora più rilevante, eliminando anche le caratteristiche di leggera asimmetria nella distribuzione che permanevano con l'impostazione univariata. Il risultato conferma l'esistenza di effettivi spillover spaziali di crescita: la forma della distribuzione dello sviluppo è fortemente influenzata dal comportamento dei sistemi locali limitrofi. La presenza di un forte disequilibrio nel mercato locale è compensata principalmente da flussi provenienti da mercati del lavoro vicini. Questa compensazione può essere parziale: da qui l'esistenza di rilevanti differenze tra mercati del lavoro tra loro distanti. In Italia, questo processo è rafforzato dalla forma della nazione e dall'orografia del territorio. I SLL sono collocati lungo l'asse Nord-Sud, con le principali aggregazioni e concentrazioni economiche situate nelle regioni settentrionali. La storia dello sviluppo economico in Italia mostra che lo sviluppo è sorto nelle aree del Nord e si è diffuso nelle aree contigue. Questo è in linea con il modello *core-periphery*: lavoratori e posizioni lavorative sono concentrate nel *core* del sistema economico. Inoltre, i risultati mostrano come questa polarità sia rafforzata dal diverso grado di sviluppo strutturale, la cui gradazione è anch'essa collocata prevalentemente lungo l'asse Nord-Sud, ed è per molti versi endogena al processo di crescita.

Alla luce dei risultati ottenuti, si ritiene utile procedere ulteriormente nell'analisi cercando di valutare se gli effetti che abbiamo rilevato mutino o meno nel tempo. A tal fine si rende necessario "mappare" la distribuzione nel tempo mediante l'appropriato uso di stimatori di densità multidimensionali. Nel far ciò, oltre agli aspetti di natura applicata, particolare attenzione dovrà essere posta anche alle questioni computazionali, la cui complessità si accresce notevolmente passando da un'analisi di tipo statico allo studio della dinamica delle distribuzioni.

Dipartimento di Scienze statistiche "Paolo Fortunati"
Università di Bologna

BARBARA PACINI
GUIDO PELLEGRINI

RINGRAZIAMENTI

Si desidera ringraziare un anonimo *referee* e i partecipanti alla presentazione di una versione precedente di questo lavoro nella XLI Riunione scientifica SIS, Università di Milano Bicocca, giugno 2002, per gli utili commenti. La ricerca è stata finanziata dal Murst (Fondo ex 40%) nell'ambito del progetto "Metodi e modelli statistici per l'analisi spaziale" e dal Dipartimento di Scienze Statistiche dell'Università di Bologna (Fondo ex 60%).

RIFERIMENTI BIBLIOGRAFICI

L. ANSELIN, (1988), *Spatial Econometrics: Methods and Models*, Dordrecht, Kluwer Academic.

A.W. BOWMAN, A. AZZALINI, (1997), *Applied Smoothing Techniques for Data Analysis*, Oxford University Press, New York.

W. S. CLEVELAND, (1979), *Robust Locally Weighted Regression and Smoothing Scatterplots*, "Journal of the American Statistical Association", 74, pp. 829-836.

S. FABIANI, G. PELLEGRINI, (1999), *Convergenza e divergenza nella crescita delle province italiane*, in Ricerche quantitative per la politica economica 1997, Banca d'Italia-CIDE.

J. FAN, I. GIJBELS, (1996), *Local Polynomial Modelling and Ist Applications*, Chapman & Hall, London.

P. KRUGMAN, (1991) *Geography and Trade,* MIT Press, Cambridge MA.

H. G. OVERMANN, D. PUGA, (1999), *Unemployment cluster across European regions and countries*, Working paper UT-ECIPA-DPUGA-99-03, University of Toronto, July.

D. T. QUAH, (1997a), *Empirics for Growth and Distribution: Polarization, Stratification and Convergence Clubs*, "Journal of Economic Growth", 2.

D. T. QUAH, (1997b), *Regional cohesion from local isolated actions: II. Conditioning*, CEP Working Paper n. 379, LSE.

S.J. SHEATER, M.C. JONES, (1991), *A reliable data-based bandwidth selection method for kernel density estimation*, "Journal of the Royal Statistical Society", Serie B, 53, pp. 683-690.

M.P. WAND, M.C. JONES, (1995), *Kernel smoothing*, Chapman and Hall, London.

RIASSUNTO

Metodi non parametrici multivariati: un'applicazione al caso della crescita

L'analisi non parametrica della forma e della dinamica della distribuzione di una variabile è stata recentemente utilizzata per identificare meccanismi di sviluppo e di polarizzazione (Quah, 1997a,b). La metodologia utilizzata per lo studio di quale variabile sia responsabile della presenza di tali polarizzazioni è nota come *conditioning*, ed è stata principalmente sviluppata in letteratura con riferimento al solo caso univariato. Nel presente lavoro tale proposta viene analizzata alla luce delle analogie con le tecniche di regressione non parametrica basate su criteri di approssimazione locale. In questo nuovo contesto trova agevole applicazione anche lo studio dell'effetto congiunto di più variabili sulla forma della distribuzione. Un'esemplificazione viene fornita con riferimento allo studio della distribuzione del tasso di occupazione nei 784 sistemi locali del lavoro italiani.

SUMMARY

Nonparametric multivariate methods: an application to the empirics of economic growth

Nonparametric analysis of shape and dynamics of probability distributions has recently been used to identify clustering and polarization phenomena in regional economic development (Quah, 1997 a, b). The technique used to investigate which variables are relevant in determining such polarization phenomena is known as *conditioning.* So far this kind of approach has been applied only in a univariate context. In this paper, we propose a more general framework, referring to the literature on nonparametric local regression techniques. Following this alternative approach we are able to study the joint effect of two or more variables on the shape of the distribution of growth. An empirical illustration is given for the distribution of employment rate in 784 Italian Local Labour Markets.