

FRANCESCO MARIA SANNA

METODI PER L'ANALISI MULTIVARIATA DI INDICATORI TERRITORIALI

1. *Introduzione.* – La conoscenza, a vari livelli di dettaglio (non solo territoriale), del quadro di riferimento demografico, economico, sociale di un paese è andata assumendo via via maggiore rilevanza sia per gli studiosi delle diverse discipline, sia per i decisori pubblici.

Per rispondere a queste esigenze conoscitive si è assistito ad una progressiva diffusione di rilevazioni per lo studio di problemi specifici, all'ampliamento delle rilevazioni già esistenti, alla proliferazione di metodologie e studi volti a sempre meglio dominare i fenomeni d'interesse.

Questo processo è stato, tra l'altro, favorito dal sempre maggiore grado di disaggregazione al quale le informazioni sono rese disponibili; ciò consente analisi sempre più approfondite e puntuali, rese possibili dallo sviluppo e dalla diffusione di mezzi informatici sempre più potenti. La ricchezza informativa delle elaborazioni possibili può peraltro compromettere la comprensione di fenomeni articolati.

Se da un lato è vero che è ormai usuale ricorrere a classificazioni delle unità amministrative ai vari livelli (province, comuni, unità sanitarie locali, ecc.) in base ad una molteplicità di indicatori demografici, economici, sociali (evidentemente, l'impiego di indicatori diversi può portare, ed in genere porta, a differenti raggruppamenti tipologici), dall'altro occorre rilevare che sempre più spesso ci si scontra con la difficoltà di reperire in tempi brevi validi indicatori relativi ai fenomeni analizzati e di operarne una sintesi efficace e non contraddittoria (gli indicatori dovrebbero essere facilmente interrelabili, aggregabili e manipolabili secondo le specifiche esigenze dell'utente).

«Data la grande disponibilità del *software* per gli elaboratori elettronici, si assiste, nelle applicazioni, ad analisi condotte su tabelle dei dati senza che sia stato sufficientemente approfondito il significato dei dati rilevati e la legittimità della particolare applicazione. Oggi è molto facile ottenere dall'elaboratore un *output* relativo a una matrice di dati, quale essa sia. Si tratta di

vedere che significato ha» (Rizzi, 1985, p.20).

Si pongono perciò problemi di due ordini:

– quelli connessi alla scelta degli indicatori (individuazione degli indicatori disponibili, identificazione delle componenti degli indicatori di sintesi, sia qualitativi sia quantitativi, eventuale definizione di standard e/o valori-soglia, ecc.);

– quelli riguardanti i metodi preferibili, nei diversi contesti, per la sintesi degli indicatori stessi. E' su questi ultimi che viene concentrata l'attenzione in questo lavoro.

2. *L'analisi dei dati.* – Le metodologie statistiche che vanno sotto il nome di «analisi dei dati» consentono di giungere a sintesi informative che e-nucleano da masse sempre più considerevoli di dati (non dominabili con gli strumenti elementari di analisi tipici della statistica descrittiva) le loro caratteristiche fondamentali, quelle che, in definitiva, costituiscono l'essenza dei fenomeni di interesse. Ciò sia a fini confermativi (si vogliono verificare ipotesi già formulate) sia a fini esplorativi, nel qual caso occorre inizialmente considerare un elevato numero di caratteri.

Ogni studio che voglia tener conto simultaneamente di molteplici indicatori al fine di giungere a classificazioni tipologiche territoriali non può prescindere da metodi di analisi statistica multivariata che, con la minima perdita di informazione, consentono di analizzare fenomeni complessi, sia mediante rappresentazioni grafiche di facile interpretazione e notevole espressività (analisi in componenti principali), sia mediante aggregazioni automatiche per l'individuazione di tipologie (*cluster analysis*). Degli indicatori prescelti, solo alcuni, quelli più significativi, concorrono alla formazione delle tipologie (le cosiddette «variabili attive»), mentre gli altri vengono utilizzati *a posteriori* a scopo esclusivamente descrittivo, per meglio delineare le caratteristiche dei gruppi che si formano.

Un problema che sempre più spesso si pone è quello della concomitanza, nell'analisi, di variabili qualitative e quantitative. E' noto che per analizzare un carattere nominale si possono utilizzare variabili *dummies*, che segnalano la presenza/assenza di una determinata modalità in corrispondenza di ciascuna unità statistica. Il carattere viene perciò analizzato considerando le possibili combinazioni lineari delle sue variabili indicatrici. Ma ciò equivale ad una opportuna quantificazione del carattere stesso, in cui i coefficienti della combinazione lineare sono le quantificazioni delle modalità corrispondenti. Anche nel caso di caratteri ordinali si può procedere analogamente, imponendo però dei vincoli alle quantificazioni, in modo da renderle coerenti con le informazioni d'ordine note *a priori*. Infine, in talune situazioni potrebbe essere utile riquantificare anche caratteri quantitativi. Si tratta, in definitiva, di una parametrizzazione dei dati.

Ulteriori problemi possono sorgere, al momento della scelta degli indicatori da includere nell'analisi, allorché si ritenga che gli indicatori correntemente disponibili costituiscano delle *proxies* delle «vere» variabili d'inter-

se; può essere allora conveniente utilizzare preliminarmente gli indicatori disponibili per individuare, mediante tecniche di analisi esplorativa ed opportune modellizzazioni, la struttura latente del fenomeno. In tal modo è possibile esprimere i fenomeni osservati (per i quali si dispone di un gran numero di indicatori, non necessariamente quantitativi) in termini di un numero molto minore di «variabili latenti».

Inoltre, l'opportunità, sempre più avvertita, di seguire l'evoluzione temporale di fenomeni territorialmente diffusi comporta la necessità di analizzare matrici di dati a più di due indici (variabili $\times$ individui $\times$ occasioni) (1).

Prima però di applicare una qualsiasi tecnica di analisi multivariata, occorre dare risposta ad almeno due interrogativi, che riguardano:

a) il peso con il quale ciascun individuo entra nell'analisi, che non necessariamente deve essere il medesimo per ogni individuo. «Un esempio saliente di ponderazione si riscontra quando i dati multidimensionali sono riferiti a zone (solitamente quelle amministrative) di un territorio. E' allora molto naturale ponderare ciascuna zona con un indice del suo "grado di importanza"» (Zani, 1984, p. 9). In qualche caso può essere utile introdurre anche una matrice dei pesi delle variabili; l'individuazione dei pesi da attribuire sconta però, in genere, un maggior grado di soggettività rispetto alla scelta dei pesi da attribuire agli individui;

b) l'opportunità di introdurre trasformazioni (non neutrali rispetto alla successiva analisi) negli indicatori. In particolare, spesso (dati espressi in unità di misura diverse o di diversi ordini di grandezza) si pone il problema di una eventuale standardizzazione; talune analisi statistiche, basandosi su metodi che non risultano invarianti rispetto a trasformazioni lineari dei dati, forniscono risultati diversi secondo che si operi sui dati originari oppure sui corrispondenti valori standardizzati.

Inoltre, nelle analisi socio-economiche molto spesso le variabili in gioco risultano altamente correlate; al fine di eliminare tale correlazione è opportuno effettuare una trasformazione lineare della matrice dei dati (che indichiamo con $X_{n,p}$), di matrice di correlazione $R_{p,p}$ in una matrice $Y_{n,p}$ avente matrice di correlazione I , le cui componenti $Y_1, Y_2, \dots, Y_p$ risultino cioè a due a due non correlate.

Nel seguito vengono passate in rassegna le principali metodologie impiegate per la soluzione di questi problemi, per ciascuna delineando tratti salienti e potenzialità (2).

3. L'analisi in componenti principali. – Questo tipo di analisi fornisce una rappresentazione sintetica di una tavola di dati (di dimensione $n \times p$, in

(1) Nel caso che l'analisi non debba essere ripetuta nel tempo ma si voglia tener conto dell'evoluzione pregressa dei fenomeni, si possono più semplicemente considerare nel modello sia indicatori strutturali che indicatori dinamici.

(2) Per gli sviluppi formali dei vari metodi si rimanda alle opere citate in bibliografia.

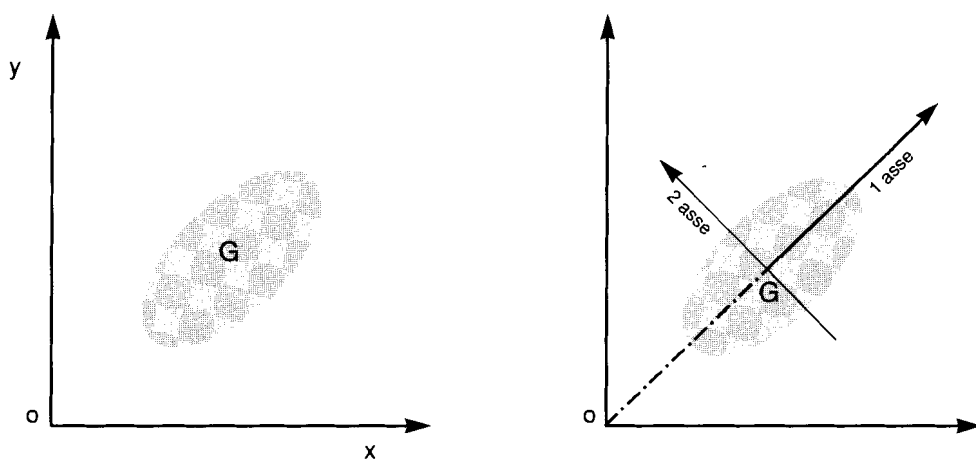


FIG. 1 – OUTPUT DI UN'ANALISI IN COMPONENTI PRINCIPALI.

cui n rappresenta gli individui e p il numero di indicatori considerati), di cui evidenzia le caratteristiche più importanti.

Nella tavola originaria, ciascun individuo è caratterizzato da un vettore di p informazioni (gli indicatori); geometricamente ogni individuo è rappresentato da un punto in uno spazio S_p a p dimensioni. Sintetizzare una tavola di questo tipo significa proiettare gli n punti-individuo in S_q , un sottospazio affine di dimensione $q < p$, individuato in maniera tale che la nuvola degli n punti di S_p subisca la minore deformazione possibile (cioè, con la minore perdita possibile di informazioni). Per $q = 2$ si ottiene una rappresentazione grafica su un piano.

La tecnica consente di analizzare, una volta individuato il centro di gravità della nuvola di punti (cioè il «profilo medio» degli individui), in quale modo e in quale misura i singoli individui si discostano da esso. Il nuovo sistema di riferimento (la cui origine è rappresentata dal baricentro G della distribuzione, fig. 1) viene ad essere costituito dagli «assi fattoriali», combinazioni lineari degli indicatori originari, tali da cogliere il massimo di variabilità insita nella nuvola di punti. Le componenti principali così individuate sono ortogonali, per cui i coefficienti di correlazione lineare fra di esse sono sempre nulli.

In genere si riesce a cogliere una quota significativa di informazioni (misurata dalla percentuale cumulata degli autovalori) considerando i primi 2 (o 3) assi fattoriali. Ciò consente di limitare a dimensioni dominabili (anche visivamente) la rappresentazione degli n punti-individuo, di descrivere il profilo degli individui mediante un numero molto ridotto di «fattori» e di in-

dividuare le principali relazioni tra individui e indicatori (3).

L'analisi in componenti principali può essere usata come primo passo per successive analisi multivariate, sia al fine di ridurre le dimensioni della matrice dei dati originari (a prezzo, comunque, di una perdita di informazione), sia come criterio «naturale» di ponderazione delle variabili (si tratta, in questo caso, di svolgere le successive analisi classificatorie sulle componenti principali, che risulta ragionevole pesare ciascuna con il rispettivo autovalore, che ne misura il contributo alla spiegazione della variabilità totale).

3.1 Lo «scaling» multidimensionale. – Alla base dei metodi di *scaling* multidimensionale, sviluppatisi originariamente in campo psicometrico, vi è la convinzione che alle similarità (o dissimilarità) riscontrate in un gruppo di oggetti possano essere associate delle distanze in uno spazio metrico. Questi metodi possono dare un contributo non indifferente per il superamento di alcune difficoltà che insorgono nell'analisi multivariata di dati territoriali.

I dati territoriali, contenuti in una matrice $X_{n,p}$ e sottoposti ad analisi esplorative (ad esempio in componenti principali, onde ridurre essenzialmente le dimensioni delle informazioni), presentano spesso problemi che rischiano di compromettere la validità dei risultati delle analisi effettuate. I coefficienti di correlazione tra le variabili, infatti, risentono dell'articolazione spaziale adottata; inoltre, in presenza di forti disomogeneità territoriali (come nel caso italiano), essi non andrebbero calcolati sul complesso delle unità considerate, ma distintamente su differenti sottoinsiemi di esse (senza comunque dimenticare che tali sottoinsiemi non possono essere considerati indipendenti). Le soluzioni fornite dai metodi di analisi fattoriale basati sulle matrici di correlazione risentono della impossibilità di separare l'effettiva interrelazione tra le variabili dagli effetti derivanti dalla scala territoriale adottata.

Utilizzando metodi di *scaling* multidimensionale si passa dalla considerazione dei valori dei coefficienti di correlazione a quella della loro posizione in graduatoria; ciò «riflette un declassamento di tali misure di concordanza, che si giustifica proprio con la componente di imprecisione che esse contengono» (Bracalente, 1984, p. 173). A fronte di questo declassamento sta una maggiore robustezza di tali metodi e quindi una tendenziale mag-

(3) Alcune principali regole di lettura dei risultati dell'analisi in componenti principali:

- la prossimità tra due punti-individuo segnala la similarità dei loro profili in termini degli indicatori considerati;
- la prossimità al centro di gravità di un punto-individuo segnala un profilo non dissimile da quello medio;
- i punti-indicatore che risultano avere un peso più rilevante sui risultati dell'analisi sono quelli che più si discostano dal centro di gravità;
- la vicinanza tra un punto-individuo e un punto-indicatore consente di affermare che, per quell'individuo, quella variabile assume un valore alto, ricordando però che la posizione dell'individuo non dipende solo da quell'indicatore, ma dal complesso degli indicatori considerati.

giore stabilità dei risultati; tali metodi inoltre superano il vincolo della linearità delle relazioni tra variabili e fattori e consentono, perciò, in genere, una maggiore riduzione di dimensioni.

Metodi di *scaling* multidimensionale permettono altresì di differenziare l'importanza che, in diverse ripartizioni territoriali, assumono alcuni fattori la cui influenza è peraltro comune a tutto il territorio nazionale (4).

4. *La cluster analysis.* – La *cluster analysis* o analisi dei gruppi ha per scopo quello di produrre raggruppamenti omogenei di individui, in termini di caratteristiche di maggior rilievo. Ciascun gruppo presenta caratteristiche sue peculiari, che lo distinguono dagli altri. I gruppi che si individuano possono corrispondere a tipologie effettivamente esistenti, ma possono delineare altresì raggruppamenti che sono tali solo rispetto agli indicatori prescelti e al metodo impiegato. La distinzione è sostanziale sul piano concettuale, in quanto:

a) nel primo caso i gruppi esistono di per sé e si conformano a ben precisi modelli; si tratta di individuarli;

b) nel secondo caso si dovrebbe più propriamente parlare di «analisi tipologica», ovvero di ricerca di gruppi che presentino omogeneità rilevanti di comportamento, senza che essi siano necessariamente riconducibili a un qualche modello teorico.

Dal punto di vista operativo, invece, la distinzione non appare rilevante.

In relazione al tipo di algoritmo di classificazione prescelto, questa tecnica consente di trattare simultaneamente un numero molto elevato di variabili (anche se gli appesantimenti in termini elaborativi possono costituire un pesante vincolo) osservate su collettivi spesso molto numerosi (ad esempio, nel caso di un'analisi a livello comunale estesa a tutto il territorio italiano si tratterebbe di oltre 8000 individui), e porta all'individuazione di un numero assai ristretto (in genere inferiore a 10) di tipologie (5).

Occorre tenere presente che, poiché ciascun gruppo viene a formarsi in base all'analisi simultanea di più variabili (le cosiddette «variabili attive», che costituiscono un sottoinsieme delle variabili osservate), non è di norma possibile imputare ad un solo indicatore il principale motivo della presenza o assenza di un determinato individuo in un certo gruppo.

Inoltre, quando gli individui sono rappresentati da unità amministrative, la scelta (talora obbligata) di operare a determinati livelli di aggregazione (ad esempio, la provincia) può «oscurare» (in tutto o in parte) le disomogeneità eventualmente presenti all'interno di ciascuna unità elementare considerata.

(4) La scelta di muoversi sulle linee qui esposte costituisce, di fatto, un possibile modo (non certo l'unico) di passaggio da un approccio esplorativo di una realtà multidimensionale, in assenza di informazioni *a priori*, ad un approccio in cui l'informazione *a priori* esiste e deriva dai dati disponibili.

(5) Vari tentativi, nessuno dei quali conclusivo, sono stati compiuti per giungere a regole che forniscano il numero ottimale di *clusters*.

I metodi di *clustering*, pur nella loro varietà, presentano tutti due momenti «cruciali» comuni.

Il primo è rappresentato dalla scelta di quale distanza, o di quale indice di similarità, tra coppie di elementi, sia preferibile utilizzare. Tale scelta può avere influenza notevole sui risultati dell'analisi: non è, infatti, detto, ad esempio, che passando dalla distanza euclidea a quella della città a blocchi l'ordinamento delle distanze tra coppie di unità statistiche resti il medesimo e, di conseguenza, che i risultati rimangano gli stessi cambiando la metrica (6).

La seconda importante scelta da effettuare concerne l'algoritmo di classificazione (7): esistono infatti più famiglie di algoritmi di classificazione:

- algoritmi scissori, che procedono per dicotomie successive dell'insieme degli n individui e che forniscono una gerarchia di partizioni;
- algoritmi aggregativi, che procedono invece alla costruzione di classi per aggregazioni successive di individui, a due a due, e che forniscono pure una gerarchia di partizioni;
- algoritmi che conducono direttamente a delle partizioni (ad esempio, il metodo di aggregazione intorno a centri mobili).

Gli algoritmi del primo tipo, pur essendo i più ricchi di proprietà matematiche, non sono di particolare rilievo in questo contesto, in quanto trovano applicazione in situazioni nelle quali il numero di individui è limitato. Gli algoritmi aggregativi sono utilizzabili su masse anche cospicue di dati e, rispetto ai primi, presentano il vantaggio di riuscire a individuare anche gruppi che non siano di forma sferica. Quando però le masse di dati da trattare sono molto vaste, per motivi legati ai tempi (e ai costi) di elaborazione, si ricorre in genere a metodi come quello dell'aggregazione intorno a centri mobili, che generano direttamente delle partizioni.

In ogni caso, i metodi di uso corrente assegnano ogni elemento a uno e un solo gruppo e, per gli elementi assegnati al medesimo gruppo, non distinguono tra quelli che si situano in prossimità del centroide del gruppo (che hanno cioè un profilo molto vicino a quello che caratterizza il gruppo) e quelli che si collocano invece in posizione più distante.

E' bene tener presente che, nella pratica, sono piuttosto rare situazioni in cui non vi siano individui per i quali la classificazione in un gruppo o in un altro non sia relativamente arbitraria. In alcune situazioni questo rappresenta il prezzo da pagare per garantire l'univocità dell'assegnazione di ogni individuo a uno e un solo gruppo; in altre, si può pensare di attribuire gli individui che presentano somiglianze con due o più gruppi «in parte» a cia-

(6) In statistica, lo spazio di riferimento è a n dimensioni; i casi particolari $n = 2$ e $n = 3$, a differenza di quanto avviene in campo fisico, non hanno alcun significato particolare. Non vi è motivo di privilegiare la distanza euclidea rispetto ad altre; anzi, in campo socio-economico sono numerose le situazioni nelle quali la metrica della città a blocchi risulta preferibile.

(7) Non sempre, infatti, algoritmi diversi portano sostanzialmente ai medesimi risultati: ciò avviene spesso se i gruppi hanno una forma più o meno sferica e allorquando i dati risultano avere effettivamente una struttura sottostante (ma ricordiamo che le tecniche di *clustering* si applicano anche in situazioni nelle quali una vera e propria struttura sottostante non esiste).

scuno di essi. A tal fine sono stati proposti metodi di classificazione «sfocata», basata sulla teoria dei *fuzzy sets*, cui si farà in seguito un breve cenno.

Infine, la *cluster analysis* può essere utilizzata come stadio preliminare per una successiva analisi discriminante, mirante ad attribuire a gruppi già individuati eventuali nuove unità statistiche (8).

4.1 Un possibile criterio per la selezioni delle variabili. – «Nel campo delle scienze economiche e sociali si pone l'accento sul principio della "parsimonia", in base al quale si ritiene opportuno limitare il numero di caratteri a quelli con il maggior contenuto d'informazione e più strettamente legati all'indagine in esame» (9).

In concreto, si dispone di un insieme di indicatori, tutti potenzialmente utilizzabili, e si mira a ridurre, senza perdita rilevante d'informazione, lo spazio delle variabili. Siano:

$X = [X_1, X_2]$ la matrice a blocchi i cui elementi sono la matrice delle k variabili da considerare (X_1) e la matrice delle $(p - k)$ variabili da eliminare (X_2) e

$$R = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$$

la matrice di correlazione, le cui componenti sono le matrici di correlazione fra sottoinsiemi di variabili.

Si tratta di scegliere la partizione in due gruppi delle p variabili che ottimizza una particolare funzione obiettivo (se, come spesso accade, si opera su variabili standardizzate, si considerano le corrispondenti matrici Z , Z_1 , Z_2). Il problema si presenta più semplice per k prefissato (allorquando si sia cioè prestabilito il numero di variabili da considerare), ma può essere affrontato e risolto anche per k ignoto (si tratta di analizzare un numero assai più elevato di partizioni). Metodi basati sulla sensibilità dell'insieme delle variabili all'eliminazione di una di esse rientrano come caso particolare in questo schema (la matrice X_2 è in tal caso un vettore $n \times 1$).

Un valido criterio di selezione appare quello basato sull'indice di ridondanza proposto da Stewart e Love:

$$RI^2 (Z_2/Z_1) = \text{tr}(R_{21}R_{11}^{-1}R_{12}) / (p - k)$$

(8) Ad esempio, l'istituzione di una nuova provincia può determinare la necessità di classificare le nuove unità che si vengono a costituire (la nuova provincia stessa e le province preesistenti i cui indicatori risultino modificati) in un quadro di riferimento costruito sulle unità (province) preesistenti e che non hanno subito modificazioni.

L'attribuzione si effettua sempre introducendo delle ipotesi sulla struttura dei dati; la correttezza dell'attribuzione dipende perciò dalla sussistenza delle ipotesi stesse.

(9) Milioli, 1988, p. 289. L'articolo contiene un'interessante rassegna dei criteri di ottimizzazione proposti da vari autori.

che assume valori nell'intervallo $[0,1]$; in particolare, risulta massimo allorché dalle variabili in X_2 non si può trarre alcuna informazione aggiuntiva rispetto a quella contenuta in X_1 . Tra tutte le partizioni analizzate, si sceglierà quella che massimizza tale indice, quella cioè che isola le $(p-k)$ variabili che danno la minor informazione addizionale rispetto alle k variabili considerate (10).

4.2. La scelta dell'algoritmo: algoritmi aggregativi.— Gli algoritmi aggregativi (o di classificazione ascendente gerarchica) sono del tipo *bottom-up*, operano cioè dal basso verso l'alto, aggregando le unità statistiche in modo da ricostruire un «albero». Si tratta di metodi validi quale che sia la matrice di distanze (o similarità) adottata.

Una volta calcolate tutte le distanze, il metodo aggrega i due individui che risultano «più vicini» in termini della distanza prescelta; ricalcola quindi le distanze tra il nuovo elemento (rappresentato dai 2 individui aggregati) e i restanti $(n-2)$ individui; aggrega i due individui «più vicini»; ecc. Il procedimento viene iterato $(n-1)$ volte, fino ad ottenere un solo gruppo, e può essere rappresentato graficamente mediante un dendrogramma (esemplificato in fig. 2), nel quale in ordinata sono rappresentate le distanze. Tagliando il dendrogramma con una retta parallela all'asse delle ascisse a una determinata altezza, si ottiene la partizione in un numero di classi che sarà tanto più elevato quanto più la retta di taglio è prossima agli elementi terminali dell'albero (gli individui).

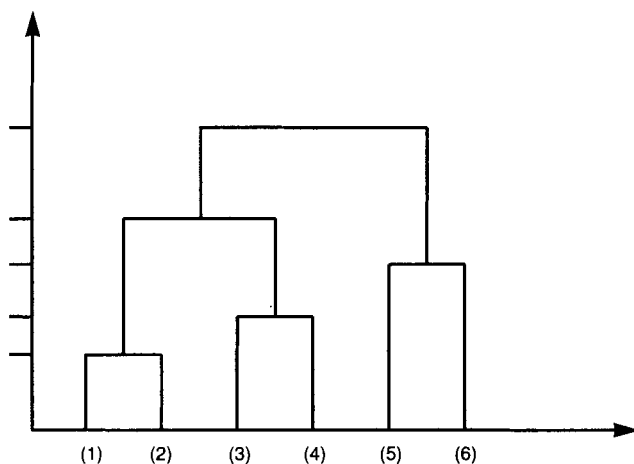


FIG. 2 – ESEMPIO DI DENDROGRAMMA.

(10) L'indice può essere espresso come media aritmetica semplice dei coefficienti di determinazione multipla fra ciascuna variabile di Z_2 e quelle di Z_1 .

4.3 La scelta dell'algoritmo: algoritmi che conducono direttamente a partizioni. – Si tratta di algoritmi che, richiedendo una mole molto più contenuta di calcoli rispetto a quella richiesta dagli algoritmi aggregativi (e il vantaggio è ancora maggiore rispetto agli algoritmi scissori), sono largamente utilizzati su grandi masse di dati, pur se le loro caratteristiche formali sono senz'altro inferiori rispetto a quelle possedute da altri algoritmi di classificazione automatica.

In particolare ci soffermiamo sul procedimento di aggrégazione intorno a centri mobili: l'algoritmo di classificazione procede ad una scelta casuale di k individui, intorno ai quali, sulla base di prefissati criteri di similarità, vengono aggregati tutti i rimanenti; si individuano i baricentri dei k gruppi e si riaggregano gli individui intorno ai nuovi centri; l'operazione viene reiterata fino ad ottenere una classificazione stabile.

Il procedimento può essere replicato h volte, ciascuna a partire da k individui scelti a caso ogni volta; i gruppi che ogni volta si ottengono non sono, evidentemente, gerarchizzabili (un individuo classificato nel gruppo A la prima volta non necessariamente apparterrà ancora al medesimo gruppo nelle successive). Si possono considerare come stabili quei raggruppamenti di individui che risultano appartenere sempre alla medesima classe in ciascuna partizione.

E' possibile utilizzare questa tecnica per ridurre il numero di individui sui quali applicare poi un algoritmo di classificazione gerarchica ascendente.

4.4. Le classificazioni «sfocate». – Dati n individui, sui quali sono state osservate p variabili, da classificare in k gruppi, si definisce k -partizione sfocata degli n individui la matrice $U_{n,k}$ i cui elementi u_{ij} ($i = 1, 2, 3, \dots, n; j = 1, 2, \dots, k$) forniscono il valore della funzione di appartenenza dell'individuo i -mo al gruppo j -mo, sotto i seguenti vincoli

$$0 \leq u_{ij} \leq 1 \text{ e } \sum_{j=1}^k u_{ij} = 1$$

Le procedure di classificazione sfocata differiscono dalle usuali (11) per il calcolo, a ciascun passo, dei centroidi sfocati di ogni gruppo, definiti come vettori dei valori medi delle variabili, ponderati con le potenze s -me dei valori della funzione di appartenenza (più elevato è s , minore è il peso degli individui che presentano valori più bassi della funzione di appartenenza). La validità dei risultati può essere giudicata mediante vari indici, tra i quali il più noto è il coefficiente di nitidezza F dato da

$$F = \sum_{i=1}^n \sum_{j=1}^k (u_{ij})^2 / n$$

(11) Se tutte le u_{ij} assumono soltanto i valori 0 o 1, si ritorna al caso classico di una partizione propria degli n individui.

5. *Le matrici a più indici.* – Ci limitiamo a considerare le cosiddette «matrici cubiche» (variabili $\times$ individui $\times$ occasioni), che interessa analizzare nell'ottica dei modelli di classificazione, con attenzione particolare all'evoluzione dei fenomeni nel tempo.

Nel caso di matrici a 3 indici, nelle diverse occasioni si possono presentare le seguenti situazioni:

a) stesse unità e stessi caratteri; b) stesse unità e caratteri differenti; c) differenti unità e stessi caratteri.

In presenza di un riferimento amministrativo costante – le stesse unità territoriali – il primo caso è quello di maggior rilievo.

Rispetto al trattamento classificatorio dei dati, ci si possono porre obiettivi diversi (compatibilmente con la situazione che si ha di fronte), in quanto si potrebbe richiedere di avere la stessa classificazione per le diverse occasioni, oppure classificazioni distinte per le varie occasioni. In particolare, se in ciascuna occasione sono stati osservati gli stessi caratteri (ma non necessariamente le stesse unità), si potrebbe individuare lo stesso numero di classificazioni per ogni occasione, lasciando aperta la possibilità che i *clusters* corrispondenti nelle varie occasioni contengano unità diverse; se invece le unità sono le stesse nelle varie occasioni (ma non necessariamente i caratteri osservati), allora si potrebbe individuare un'unica classificazione per le varie occasioni, al cui interno i caratteri possono assumere un peso diverso; infine, se tanto unità che caratteri sono gli stessi, è possibile individuare una sola classificazione che colga gli elementi di costanza nei dati osservati nelle diverse occasioni (trattandosi dell'obiettivo più vincolante, non è detto però che debba necessariamente essere posto).

Recentemente sono stati sviluppati alcuni approcci che tengono conto in modo più rigido dell'informazione *a priori* disponibile; in particolare, interessano gli approfondimenti sui metodi di classificazione con vincoli di contiguità. Questo tipo di informazione è in genere rilevante (ma spesso trascurato nelle analisi multivariate classiche) nell'analisi di dati spaziali.

6. *Qualche considerazione conclusiva.* – Gli strumenti passati in rassegna consentono tutti, sotto angolazioni diverse, di giungere a sintesi interpretative di fenomeni complessi diffusi sul territorio. La necessità di giungere a raggruppamenti tipologici dai confini ben definiti, l'opportunità di non ignorare l'obiettivo presenza di «zone d'ombra», l'utilità derivante dal disporre di una graduatoria degli individui per definire priorità di intervento orienteranno volta a volta il ricercatore nella scelta. Così la tradizionale *cluster analysis* sarà privilegiata allorché risulti necessario (o comunque preferibile) riferirsi ad ambiti dai confini amministrativi ben individuati: quando invece si intenda privilegiare un'indagine mirata alla comprensione di realtà che costituiscono un *continuum* (e come tale si ritenga di doverle trattare) il ricorso a metodi di classificazione «sfocata» appare preferibile, pur con le inevitabili complicazioni (con riguardo, ad esempio, ad una decisione di ammissibilità di un individuo a fruire di uno specifico beneficio) che di-

scendono dalla mancata attribuzione di alcuni individui (al limite tutti) ad un solo raggruppamento tipologico; l'analisi in componenti principali trova un'indicazione preferenziale in presenza della necessità di giungere ad una semplificazione drastica di fenomeni complessi (sempreché gli indicatori prescelti siano idonei e poche componenti riescano ad assorbire una quota rilevante di variabilità); i risultati di un'analisi in componenti principali possono fungere a loro volta da *input* per successivi approfondimenti, ecc.

Allorché l'approccio, almeno in fase iniziale, è di tipo esplorativo appare opportuno applicare l'analisi in componenti principali sugli indicatori prescelti, per passare poi alla classificazione delle unità in gruppi omogenei, utilizzando le coordinate delle unità stesse sulle componenti più rilevanti. In tal modo la classificazione verrà effettuata su un numero ridotto di indicatori incorrelati (le componenti principali). La scelta dell'algoritmo classificatorio risulta vincolata dal numero delle unità sulle quali operare (province, comuni); quanto più è elevato tale numero, tanto più netta sarà l'indicazione a favore di un'analisi non gerarchica, quale quella basata sull'aggregazione intorno a centri mobili. In caso di impiego di tale metodo, una sola applicazione potrebbe fornire risultati caratterizzati da una certa instabilità; è perciò opportuno calcolare un certo numero di partizioni diverse, sulle quali individuare i «raggruppamenti stabili» cioè quei gruppi di unità che in tutte le partizioni risultano appartenere al medesimo *cluster*. Lo studio critico degli alberi di classificazione consente poi di eliminare i gruppi stabili di scarsa numerosità e/o di più difficile collocazione per il loro comportamento anomalo (in assoluto o relativamente all'area di appartenenza).

Quale che sia il metodo scelto, per la migliore comprensione dei fenomeni sarebbe opportuno disporre di dati al massimo livello di disaggregazione. Il condizionale è peraltro d'obbligo: basti pensare alle macroscopiche differenze, in termini sia di quantità di dati disponibili sia di periodicità di raccolta e diffusione dell'informazione stessa, tra dati a livello regionale, provinciale, comunale.

Talune affermazioni a scala regionale (ad esempio sul livello e le caratteristiche dello sviluppo economico di un'area) si rivelano improprie generalizzazioni se il dettaglio viene spinto al livello provinciale; analogamente, le province presentano al loro interno profonde disomogeneità, per cui il loro profilo risulta, in vari casi, determinato da uno o pochi comuni egemoni.

Ma è la stessa tradizionale articolazione amministrativa – regioni, province, comuni – che andrebbe rimessa in discussione, prendendo atto che spesso nessuna di queste suddivisioni ha un'effettiva valenza interpretativa e che l'omogeneità all'interno delle varie aree è andata progressivamente decrescendo nel tempo: «Analisi che abbiano adeguato livello di approfondimento consentono non di rado di prospettare forti dubbi sulla "individualità" regionale. Emblematico, per tutti, il caso della Basilicata, vera e propria "non regione" [...] L'unità provinciale, a sua volta, sopravvive ad una profonda crisi, che negli anni '70 aveva maturato un dibattito culturale favorevole alla sua abolizione [...] Sotto l'etichetta comunale sono comprese, infi-

ne, situazioni che presentano insostenibile variabilità, già per la sola dimensione demografica» (Marbach, 1987, p. 219). Per giunta, nel nostro paese non si riscontra la tendenza, emersa invece in molti altri paesi europei nel dopoguerra, di ridurre, talora drasticamente, il numero dei comuni.

Appare perciò di fondamentale rilievo, quale che sia il metodo prescelto per l'analisi, l'annunciata disponibilità dell'ISTAT a fornire, in occasione dei censimenti 1991, dati su una griglia di circa 350.000 subaree, che il ricercatore potrà variamente aggregare, in relazione alle proprie esigenze conoscitive, prescindendo (entro limiti) dalla corrispondenza di tali aggregazioni con una tra le usuali suddivisioni amministrative; né va trascurato il fatto che molte indagini campionarie correnti potrebbero utilizzare in futuro il medesimo impianto-base. In tal modo si riuscirebbe a costruire un cospicuo patrimonio informativo, suscettibile di sistematici aggiornamenti, su maglie elementari del territorio, variamente aggregabili in relazione alle diverse esigenze (senza però dimenticare che l'eventuale mancato rispetto degli usuali confini amministrativi creerà problemi, non sempre sormontabili, di raccordo e di aggiornamento dei dati).

RIFERIMENTI BIBLIOGRAFICI

A - Problemi metodologici

- BAHRENBERG G. e altri (a cura di), *Recent Developments in Spatial Data analysis*, Brookfield, Gower, 1984.
- BARTHOLOMEW D.J., *The Foundations of Factor Analysis*, in «Biometrika», Cambridge, 1984, pp. 221-232.
- BELLACICCO A. e LABELLA A., *Le strutture matematiche dei dati*, Milano, Feltrinelli, 1979.
- BENZECRI J.-P., *L'analyse des données*, Parigi, Dunod, 1973, 2 voll.
- BRACALENTE B., *I metodi di scaling multidimensionale nell'analisi esplorativa di dati economici territoriali*, in «Atti XXXII Riun. Scient. della Soc. Ital. di Statistica», Napoli, Liguori, 1984, vol. II, pp. 171-181.
- CAILLEZ F. e PAGES J.-P., *Introduction à l'analyse des données*, Parigi, Smash, 1976.
- CLARK D., DAVIES W.K.D. e JOHNSTON R.J., *The Application of Factor Analysis in Human Geography*, in «The Statistician», Edinburgo, 1974, pp. 259-281.
- COPPI R., *Prospettive metodologiche nell'analisi di matrici di dati a più indici*, in «Atti delle Giornate di studio del Gruppo Italiano aderente all'IFCS», Palermo, ILA Palma, 1989, pp. 13-31.
- COPPI R. e BOLASCO S. (a cura di), *Multiway Data Analysis*, Amsterdam, North Holland, 1989.
- DAVISON M.L., *Multidimensional Scaling*, New York, J. Wiley & Sons, 1983.
- DI CIACCIO A., *Metodi di classificazione per l'analisi simultanea di caratteri qualitativi e quantitativi*, in «Atti XXXV Riun. Scient. della Soc. Ital. di Statistica», Pa-

- dova, CEDAM, 1990, vol. II, pp. 215-221.
- DI CIACCIO A. e DE LEEUW J., *Cluster analysis per matrici di dati a due o più indici e con livello di misurazione misto*, in «Atti delle Giornate di studio del Gruppo Italiano aderente all'IFCS», Palermo, ILA Palma, 1989, pp. 83-96.
- DIDAY E. e altri (a cura di), *Data Analysis and Informatics*, Amsterdam, North Holland, 1980.
- ESCOUFIER Y., *À propos du choix des variables en analyse des données*, in «Metron», Roma, 1986, pp. 31-47.
- EVERITT B.S., *Cluster Analysis*, Londra, Heinemann, 1974.
- FABBRIS L., *Analisi esplorativa di dati multidimensionali*, Padova, CLEUP, 1983.
- GRIFFITH D.A., *Advanced Spatial Statistics: Special Topics in the Explanation of Quantitative Spatial Data Series*, Dordrecht, Kluwer Academic, 1988.
- HARTIGAN J.A., *Clustering Algorithms*, New York, J. Wiley & Sons, 1975.
- HAWKINS D.M. (a cura di), *Topics in Applied Multivariate Analysis*, Cambridge, Univ. Press, 1982.
- JAMBU M. e LEBAUX M.O., *Cluster Analysis and Data Analysis*, Amsterdam, North Holland, 1983.
- LAURO N. e SICILIANO R., *Recenti sviluppi e nuove tendenze nell'analisi multidimensionale dei dati*, in *Metodi statistici per la tecnologia e l'analisi dei dati multidimensionali*, Padova, Univ. di Padova, 1988, pp. 163-199.
- LEBART L., MORINEAU A. e FÉNELON J.-P., *Traitement des données statistiques: méthodes et programmes*, Parigi, Dunod, 1979.
- LEONI R. (a cura di), *Alcuni lavori di analisi statistica multivariata*, Firenze, SIS-Commissione «Analisi statistiche multiple», 1982.
- MILIOLI M.A., *La scelta delle variabili nell'analisi multidimensionale dei dati: i metodi basati sulla matrice di covarianza*, in *Metodi statistici per la tecnologia e l'analisi dei dati multidimensionali*, Padova, Univ. di Padova, 1988, pp. 289-296.
- RIZZI A., *L'analisi dei dati*, Roma, La Nuova Italia Scientifica, 1985.
- SOCIETÀ ITALIANA DI STATISTICA, *Atti delle Giornate di Studio su «Classificazione e analisi dei dati: metodi, software, applicazioni – Analisi dei dati spaziali: metodi tecniche, applicazioni»*, Pescara, Univ. «G. D'Annunzio», 1991.
- STEWART D. e LOVE W., *A generical Canonical Correlation Index*, in «Psychological Bull.», Washington, 1968, pp. 160-163.
- UNIVERSITÀ DEGLI STUDI DI PADOVA, FACOLTÀ DI SCIENZE STATISTICHE, DEMOGRAFICHE ED ATTUARIALI, *Due temi di analisi statistica multivariata*, Padova, CLEUP, 1979.
- WRIGLEY N., *Categorical Data Analysis for Geographers and Environmental Scientists*, Londra, Longman, 1985.
- ZANI S., *Analisi multidimensionale dei dati: recenti orientamenti e prospettive*, in «Atti XXXII Riun. Scient. della Soc. Ital. di Statistica», Napoli, Liguori, 1984, vol. I, pp. 3-26.
- ZANI S., *Un metodo di classificazione «sfocata»*, in *Metodi statistici per la tecnologia e l'analisi di dati multidimensionali*, Padova, Univ. di Padova, 1988, pp. 281-288.

B - Problemi applicativi

- BRACALENTE B., *Il censimento dell'industria e le indagini economiche correnti in Italia*, in «Atti XXXV Riun. Scient. della Soc. Ital. di Statistica», Padova, CEDAM, 1990, vol. I, pp.179-189.
- CISLAGHI C., *Le informazioni censuarie e le indagini correnti per le analisi dei fenomeni demografici sociali e sanitari*, in «Atti XXXV Riun. Scient. della Soc. Ital. di Statistica», Padova, CEDAM, 1990, vol. I, pp.167-177.
- GOLINI A. e altri, *La demografia dei Comuni italiani*, Roma CNR-IRP, 1988.
- MARBACH G., *Indicatori di reddito e di consumo in Italia: problemi delle disaggregazioni territoriali*, in «SIS-Convegno 1987: Informazione ed analisi statistica per aree regionali e subregionali», Perugia, Galeno, 1987, pp.217-230.
- MARBACH G., *Il reddito nei Comuni italiani 1987*, Roma, Quaderni del Banco di Santo Spirito, 1989.
- SIESTO V., *I problemi di misurazione dell'economia sommersa*, in «Atti XXXIV Riun. Scient. della Soc. Ital. di Statistica», Siena, Nuova Immagine, 1988, vol. I, pp. 87-110.
- SOCIETA' ITALIANA DI STATISTICA, *Convegno 1987: Informazione ed analisi statistica per aree regionali e subregionali*, Perugia, Galeno, 1987.

METHODS FOR THE MULTIVARIATED ANALYSIS OF TERRITORIAL INDICATORS. – The work examines the main methods of multivariate analysis of territorial data. It dwells upon peculiarities of each one of these and indicates, even though summarily, the fields of elective employment of the different methods. The attention is chiefly turned to the problems the application of each one of these methods can create in the different contexts, either with reference to the complexity of calculation (which could condition the choice of algorithm) or to the scale of territorial disaggregation of data (administrative references not always and not necessarily must turn out to be preferable and favourite).

Università di Roma «La Sapienza», Dipartimento di Studi geoeconomici, statistici, storici per l'analisi regionale.