

Premessa

Dalla raccolta delle informazioni alla struttura di senso dei dati

di *Luigi Frudà*

I problemi che vengono affrontati e illustrati anche sul piano tecnico-operativo in questo volume ruotano intorno a una questione centrale per ogni ricercatore: il **rapporto analisi-sintesi** che si genera all'interno di una filiera complessa che va dalla formulazione originaria di un problema di rilievo scientifico alla sua traduzione in termini operativi sino alla selezione e progettazione di indicatori empirici, qualitativi e/o quantitativi che siano, i quali producono una base di dati che intercetta, descrive e interpreta il problema sotto analisi. Problema storicamente costante che ha attraversato per intero ogni campo disciplinare e ogni percorso scientifico a noi noto dalla primissima filosofia greca – per limitarsi alla storia scientifica dell'Occidente – ai fondamenti della logica aristotelica, da Cartesio a Kant, da Galileo a Linneo, da Dewey a Popper sino al dibattito epistemologico contemporaneo.

Il nodo con cui ogni ricercatore sa di doversi confrontare è costituito dal fatto che ogni indagine scientifica è costretta ad affrontare una porzione molto piccola di strutture problematiche e di dinamiche molto complesse. In questo processo di riduzione di complessità il ricercatore deve isolare e selezionare, a suo giudizio e secondo ipotesi da porre sotto verifica, parti significative verso le quali orienta tutto il proprio apparato metodologico e tecnico al fine di rilevarne l'ingegneria interna, compositiva e relazionale intra ed extra, nel modo più fedele e analitico possibile. Nelle situazioni più critiche, poco penetrabili e complesse può arrivare a prefigurare modelli nell'ipotesi che la loro formulazione riesca a cogliere per approssimazione parti costitutive, o quanto meno significative, delle dinamiche sotto osservazione.

All'interno di tale quadro la situazione del ricercatore sociale è ulteriormente complicata dal fatto che il suo oggetto di studio è terribilmente dinamico e complesso non soltanto per effetto della pluralità, e spesso enormità, di soggetti e fattori che intervengono in un fenomeno sociale, e quindi anche per gli effetti combinatori che i vari soggetti – qui agenti anche collettivi, istituzionali, strutturali, culturali – vengono a determinare con le

Il rapporto
analisi-sintesi

Complessità sociale
e disegno della
ricerca fra analisi
e sintesi

loro azioni, ma soprattutto per la velocità con cui le dinamiche si innescano, si manifestano ed evolvono.

In questa situazione vi sono due nodi particolarmente critici che vanno posti sotto controllo e soggetti a decisioni operative: il livello di analiticità da mettere in campo cui si collega la massa di informazioni analitiche da produrre e, sul versante opposto, la progettazione del passaggio dai dati analitici prodotti al livello di sintesi ritenuto in ipotesi più efficace.

Certamente fallace l'idea operativa che il ricercatore possa, per prudenza o furbizia (!), tirar dentro il suo progetto la massa più ampia possibile di indicatori analitici. Gli indicatori si danno e si progettano soltanto all'interno di una congruente strutturazione di uno specializzato disegno di ricerca e da qui discende che il ricercatore non ricava alcun vantaggio dall'introdurre procedure caotiche incontrollate e dal produrre ulteriori tassi di rumore analitico interno in una situazione in cui il suo disegno di ricerca è, popperianamente, una congettura che tenta di riprodurre in modo parziale e approssimato una realtà più o meno ampiamente incognita se non del tutto sconosciuta. Radicalizzare quindi analiticamente una progettazione ha scarse possibilità di successo quando questa non è sostenuta da un chiaro quadro di ipotesi, siano esse descrittive o inferenziali.

La formulazione di ipotesi costituisce proprio la via migliore per operare selezioni e scelte appropriate, per includere o escludere taluni indicatori e disegnarne le loro relazioni ipotetiche anche in funzione della loro natura e trattabilità tecnica sul piano della successiva analisi dei dati. Per altro verso il progetto della sintesi non può essere rinviato alle fasi finali di indagine nella presunzione, errata, che comunque si riuscirà a tirar fuori qualcosa avendo una base-dati molto ampia e tale da consentire ampie scorrerie attraverso mappe di ipotesi formulabili *a posteriori*. Ogni sintesi, scientificamente controllata, ha una base analitica altrettanto controllata e mirata; e ciò non esclude per nulla né la revisione, in corso d'opera, del piano di indicatori e ipotesi né intervenienti e possibili esiti di *serendipity*.

Autoconsapevolezza
e ruolo
del ricercatore nella
produzione di sintesi

Nella gestione del delicato rapporto analisi-sintesi il ricercatore deve farsi guidare dalla consapevolezza che ogni eccesso di **analisi**, per le ragioni sopra accennate, rende meno controllabile l'intero impianto di indagine e che ogni **sintesi**, quale ne sia il livello, comporta insieme una perdita e un guadagno di informazioni; una **perdita** perché al procedere del livello di sintesi le informazioni analitiche si sommano fra di loro a scapito del loro originario e singolare contenuto che non è più leggibile sotto forma, per l'appunto, analitica; un **guadagno** perché le singolari informazioni analitiche qualora rimanessero tali sarebbero incapaci di dare un senso e una interpretazione ai dati raccolti che, al contrario, la creazione di sintesi produce a livelli diversificati e controllabili.

Va aggiunto inoltre sul piano non solo scientifico ma anche, e soprattutto, deontologico che il mancato e trasparente controllo del rapporto analisi-

sintesi introduce distorsioni tali da comportare spesso snaturamenti e mistificazioni di dati. Fatto ancora più grave quando tali distorsioni vengono indotte, fuori da ogni consapevolezza, da un uso acritico e automatico di tecniche di trattamento di dati. Basti ad esempio accennare alla situazione in cui l'eccesso di sintesi fa letteralmente sparire la specificità e la singolarità di molti sottoinsiemi di un universo oppure all'abuso o al cattivo uso o all'uso strumentale e mistificante delle medie aritmetiche, senza alcun riferimento congiunto a misure che ne apprezzino gli scarti, per descrivere comportamenti ritenuti, spesso in modo artificioso, di massa. Lo stesso effetto si ha con il ricorso a parallele concettualizzazioni approssimative oppure generalizzanti a tal punto da riuscire del tutto generiche come avviene nell'esercizio abitudinario e scorretto, soprattutto sul piano comunicativo e divulgativo, dello *stretching* concettuale su concetti-universo come "i giovani", "gli elettori", "gli italiani", "il pubblico", "i siciliani", "i lombardi", "i consumatori", "i turisti", "gli stranieri" e così di seguito.

Ma anche al di fuori di questi contesti che configurano prassi, comunicative e non, che alla fine nulla o poco hanno a che fare con il lavoro scientifico si possono dare situazioni in cui anche dall'interno di procedure metodologicamente e tecnicamente rigorose si originano artefatti e anche potenziali mistificazioni connesse a uno scarso controllo delle operazioni di analisi-sintesi e soprattutto alla mancata lettura parallela e comparata sia del dato analitico che del dato sintetico.

Se, ad esempio, si guarda all'andamento sintetico delle variazioni decennali intercensuali della popolazione delle grandi città, si trova che il loro tasso di crescita è, in più di una situazione, ormai prossimo al valore zero oppure risulta negativo anche per due/tre punti percentuali. Su questa base si potrebbe concludere per una attenuazione o derubricazione di alcune politiche come gli interventi sulla tensione abitativa, sui flussi di traffico, sul potenziamento o la creazione di servizi per determinate fasce di popolazione. Al contrario, se scomponiamo i dati analitici per circoscrizione o municipalità, è possibile trovare che in diverse situazioni non vi è decremento, ma incremento di popolazione e mentre in alcune zone si hanno indici di vecchiaia patologici, in altre è registrabile un tasso di giovinezza straordinario a fronte di un dato complessivo sintetico che nasconde la realtà di tali dinamiche. Si comprende bene che da queste diverse letture, ora sintetiche ora analitiche, potrebbero derivarsi azioni diversificate e addirittura opposte di pianificazione sociale e strutturazione urbana. Una corretta lettura analitico-sintetica su base comparata impedirebbe qualsiasi forzatura o mistificazione interpretativa.

Si aggiunga inoltre che questa virtuosa gestione della relazione analisi-sintesi fornisce, in concreto, una formidabile opportunità per sondare la tenuta operativa dei **concetti** messi in campo dal ricercatore proprio nella direzione della loro tenuta analitica e/o sintetica. Nel caso sopra richiamato ap-

pare evidente come non tenga, rispetto alle dinamiche attuali, il concetto anagrafico di “città = comune” e quindi di “residenza”. Il riferimento burocratico-territoriale non è in grado di intercettare per nulla la complessità delle nuove dinamiche urbane che sempre più fanno riferimento non al concetto di “comune” ma a quello di “area metropolitana” sino al concetto limite di “città-regione” e oltre. Se infatti i dati demografici vengono comparativamente letti in analitico e riaggregati per aree territoriali, si scopre spesso che quello che perde la città in senso stretto lo guadagna il sistema limitrofo dei comuni dell’interland sino ad arrivare alla individuazione di vere e proprie neoformazioni urbane satellitari alla città-madre che crescono o sono già cresciute a ritmi sostenutissimi e che proiettano le loro dinamiche in modo totalmente esogeno rispetto alla pura territorialità anagrafica; dal che derivano modalità del tutto diverse nella prefigurazione e progettazione di strutturazioni funzionali e di politiche socioeconomiche. In definitiva il controllo attento del rapporto analisi-sintesi sostiene per intero il processo di indagine e fonda la struttura di senso dei dati sotto osservazione. I vari contributi prodotti in questo volume articolano problematualmente e in modo applicato tale percorso partendo dal presupposto della **multifattorialità** e **multidimensionalità** dei fenomeni studiati dal ricercatore sociale e dalla necessità di un **approccio integrato e strategico** all’analisi dei dati, che traduce da un lato la indispensabilità del governo del rapporto analisi-sintesi e dall’altro la necessità di ridurre la complessità dei fenomeni sotto osservazione senza snaturarli o ipersemplificarli in modo artefatto. Le tecniche di **analisi multivariata** dei dati, che per lo più fanno riferimento a **tecniche fattoriali**, costituiscono il fuoco tematico di questo volume articolato in undici capitoli, dedicati ad altrettanti temi di grande interesse metodologico:

1. l’approccio generale in termini di tecniche esplorative e di modelli di analisi;
2. la logica dell’analisi dei fattori, delle componenti principali e del *multidimensional scaling*;
3. il confronto fra modello fattoriale classico e analisi in componenti principali;
4. l’analisi delle corrispondenze in quanto sintesi di variabili categoriali;
5. il *multidimensional scaling*;
6. l’analisi statistica multidimensionale dei testi;
7. le tecniche discriminanti;
8. le tecniche di formazione e analisi dei gruppi;
9. i modelli statistici di relazione fra variabili;
10. l’analisi dei dati mediante reti neurali artificiali;
11. l’impiego del *data mining* nella ricerca sociale.

Le scelte effettuate, oltre che a esigenze manualistiche, rispondono al criterio di andare incontro a esigenze di puntualità e applicabilità sempre più

avvertite e urgenti nel campo della ricerca sociale. Moltissimi i testi cui ci si potrebbe riferire per ognuno dei passi tematici sopraelencati, pochi i testi realmente comprensibili e aperti alle richieste del ricercatore applicato che non sia anche un profondo conoscitore degli algoritmi interni ai software utilizzati.

Un approccio integrato alla ricerca sociale comporta che il ricercatore metta in atto una strategia di analisi che, oltre al controllo concettuale e operativo del proprio disegno di ricerca, sappia, appunto, integrare fra di loro più tecniche intorno allo stesso oggetto di indagine in modo tale da estrarre il massimo di informazioni possibili dai dati in progetto e dalle matrici che alla fine li ordinano per l'analisi dei dati. Tale strategia implica il rifiuto di rigidità ipotetiche, il ricorso a un piano ben preciso di indicatori analitici, la individuazione puntuale della natura dei dati osservazionali, il ricorso a più combinazioni di differenti tecniche statistiche finalizzate alla individuazione della strategia che meglio si adatta ai dati in analisi e alla messa in evidenza delle caratteristiche descrittive e relazionali dei dati, e, infine, il controllo di congruenza fra concetti sociologici di riferimento e risultati empirici prodotti. Il campo di intervento operativo riguarda l'esame esplorativo delle dimensioni di un fenomeno, le relazioni fra le dimensioni di questo fenomeno e le relazioni fra dimensioni e singole variabili. Le tecniche multivariate rispondono perfettamente a tali esigenze sia per il fatto che possono esplorare descrittivamente e lungo l'asse analisi-sintesi dimensioni e indicatori utilizzati sia perché possono esplorare le relazioni, cioè le interazioni e i rapporti, fra variabili sia sul piano descrittivo che esplicativo aprendo alle esigenze di sintesi che il ricercatore si pone. Lo studio delle relazioni fra variabili (esistenza di una relazione, direzione e intensità di essa, rapporti di dipendenza e indipendenza) ha rilevanza anche per lo sviluppo di analisi predittivo-esplicative. Il ricercatore può fare riferimento a tecniche statistiche come il modello della regressione, l'analisi log-lineare, l'analisi della varianza avendo consapevolezza che i due circuiti, esplorativo ed esplicativo, non sono così conclusi ed esclusivi, per cui esistono, di fatto, forti interscambi operativi, logici e progettuali fra i due ambiti: le tecniche multivariate per molti aspetti precostituiscono una utile propedeuticità per il passaggio dall'ambito esplorativo analitico-sintetico a quello relazionale e potenzialmente esplicativo.

In tale percorso le tecniche fattoriali forniscono, in ogni caso, strumenti formidabili per guadagnare livelli di sintesi potendone controllare parallelamente la base analitica. L'idea di base è quella di ricondurre le variabili a punti in uno spazio geometrico e quindi, come in uno spazio, osservarne le posizioni nel presupposto che, essendo le variabili indicatori di concetti, le loro posizioni nello spazio disegnano lo spazio semantico del concetto: più sono vicini i punti-variabili nello spazio, più è probabile che questi appartengano allo stesso spazio semantico. L'analisi dei fattori, le componenti

Analisi esplorativa
e modelli relazionali

Fattori, componenti
e *multidimensional
scaling*

principali, il *multidimensional scaling* sono alcune delle tecniche che disegnano, con differenti modalità, le variabili in uno spazio a più dimensioni. Attraverso un percorso storico alquanto recente, che va da Charles Spearman ai fondamentali contributi di Louis Leon Thurstone, di Hotelling, di Harman, di Togerson, tali tecniche hanno fissato alcuni basilari modelli di analisi fattoriale dei dati che oggi, grazie a personal computer di media potenza, sono da tempo accessibili a un vasto pubblico di ricercatori producendo anche un derivato effetto culturale nelle modalità di analisi e presentazione dei dati di ricerca empirica.

L'analisi
in componenti
principali

Fra le tecniche sopra richiamate, l'analisi in componenti principali ha avuto una larga diffusione proprio in funzione del fatto che nella ricerca sociale, che adopera una base molto ampia di indicatori, si pone l'esigenza di produrre sintesi intermedie che semplifichino l'insieme di variabili che concorrono a determinare la varianza complessiva del fenomeno sotto analisi. In questo modo, oltre a governare con maggiore precisione la produzione di sintesi descrittive ed esplicative, si va incontro anche a persistenti limitazioni proprie delle tecniche fattoriali, che non possono trattare un numero infinito di variabili. Le procedure, assistite da software specializzati, leggono le relazioni fra variabili a partire dal calcolo dei coefficienti di correlazione e sulla base di particolari algoritmi estraggono dei **fattori** o **componenti principali**, che hanno la caratteristica di avere relazioni con le variabili dalle quali sono estratti ma non coincidono con nessuna delle variabili utilizzate, quindi rispondono perfettamente all'obiettivo di semplificare, in modo diverso, la numerosità delle variabili originarie attraverso, ad esempio, l'eliminazione di ridondanze e sovrapposizioni fra variabili, e attraverso la minore valutazione di tutte quelle variabili che mostrano avere scarsissimo o scarso peso nel produrre sintesi e generare fattori. Altro vantaggio offerto dall'analisi in componenti principali (ACP) è l'attribuzione di un punteggio fattoriale a ogni caso sotto analisi, facendo riferimento a ciascun singolo fattore estratto; il che consente molte altre operazioni aggiuntive di analisi dei dati come l'effettuazione di operazioni di classificazione e raggruppamento. Per il fatto di basarsi sul calcolo di coefficienti di correlazioni è intuitivo che tali tecniche possono trattare soltanto **variabili cardinali**, cioè metriche e capaci di essere espresse con misurazioni a intervalli o a rapporti.

L'analisi
delle corrispondenze
multiple

Nel caso in cui il ricercatore si trova ad avere dati **categoriali** non metrici si può fare ricorso a un altro tipo di tecnica fattoriale come l'analisi delle corrispondenze multiple (ACM), che consente ugualmente, attraverso previe trasformazioni e selezioni delle variabili e delle loro modalità, di estrarre **autovalori**, in pratica equivalenti ai fattori o componenti dell'ACP e utilizzati allo stesso modo non solo per produrre sintesi ma anche per poter procedere a ulteriori operazioni di classificazione e raggruppamento. La misura di variabilità della matrice sotto analisi che viene utilizzata è l'**inerzia**, la

quale è data dalla selezione di un certo numero di variabili dalla matrice-dati originaria e all'interno di ciascuna delle variabili scelte – dette **attive** perché concorrono a determinare gli autovalori – dalla selezione di alcune soltanto delle **modalità** di queste variabili. In questo modo, rilevando la presenza (espressa con il valore 1) o l'assenza (espressa con il valore zero) della x modalità su una particolare matrice, detta matrice di Burt, si trasforma, di fatto, una variabile categoriale in una variabile che può essere trattata come una variabile cardinale metrica. Vantaggio notevole ove si consideri che sono molte le informazioni categoriali e qualitative che entrano in gioco nelle ricerche sociali; si pensi inoltre quale vantaggio viene offerto ai ricercatori che si occupano quasi esclusivamente di informazioni qualitative come quelle che si danno nel campo dell'analisi del contenuto, dell'analisi testuale o della *visual sociology*.

Il *multidimensional scaling*

L'analisi multivariata comprende anche tecniche che sono progettate per trattare, con opportuni adattamenti, sia informazioni di tipo metrico, sia informazioni di tipo categoriale: fra queste ha ampia diffusione nella ricerca sociale il *multidimensional scaling*, che intercetta l'ammontare delle **similarità** e delle **differenze** fra coppie di oggetti o casi o soggetti o unità basandosi sulla rappresentazione parametrica e grafica delle similarità-dissimilarità in uno spazio geometrico. Dal che si fa discendere che la vicinanza o prossimità indica similarità e la distanza dissimilarità. La valutazione di tale prossimità, come intuibile, è deputata a una matrice quadrata, simmetrica o asimmetrica, casi *per* casi detta **matrice di prossimità** e costruita sulla metrica di **misure di distanza** che a loro volta si attivano a partire da indici di associazione o di relazione, da misure della probabilità congiunta del presentarsi di due oggetti espresse in stime delle frequenze degli eventi. Non sfuggerà l'importanza applicata del potersi riferire direttamente anche a dati di natura categoriale come possono essere le risposte a sondaggi di opinione, a scale di atteggiamento, a questionari con ampia base qualitativa così come sempre più spesso avviene anche nelle grandi *surveys* che comparano orientamenti valoriali, comportamenti e mutazioni culturali in macroaree territoriali. In quest'ultima situazione il *multidimensional scaling* (MDS) è particolarmente importante perché può connettere insieme dati strutturali come tasso di occupazione, produzione e consumo di beni e servizi, indici di sviluppo e simili a dati valoriali e di atteggiamento, tipologie qualitative di soggetti e/o collettivi.

L'analisi multidimensionale dei testi

Caso, per certi versi estremo, che rientra con pertinenza all'interno dell'ampia famiglia dell'analisi multivariata è rappresentato dall'applicazione della stessa logica operativa all'analisi di **dati testuali**. La tecnica elettiva più frequentemente utilizzata è l'**analisi delle corrispondenze lessicali** (ACL), che esplora il *corpus* testuale attraverso le **dimensioni tematiche** intercettate dall'analisi del **vocabolario** utilizzato e dalle **parole chiave** individuabili nel testo sotto analisi. Tecnicamente, in assonanza con quanto av-

viene in tutte le altre tecniche fattoriali, si traduce il testo in una matrice costruita a partire dalle **parole** o **forme** in esso contenute; di questa matrice si studiano le interdipendenze. Come nelle altre tecniche fattoriali si perviene alla estrazione di **fattori**, ovvero **autovalori**, interpretabili come dimensioni semantiche, che rappresentati su uno spazio fattoriale permettono la proiezione su di esso delle parole del *corpus* testuale al fine di apprezzarne, attraverso una matrice delle distanze, la vicinanza decodificabile come associazione semantica all'interno del *corpus* analizzato. Ampia la gamma di software specializzati disponibili e altrettanto ampio il gergo sviluppato da questo tipo di tecnica; ma su tutto domina, ed è da ritenere fatto più che mai importante, il ruolo giocato dal ricercatore. Infatti accanto ad alcune poche azioni automatiche delegabili al software, come, ad esempio, la **lessicazione**, altre procedure come la **disambiguazione**, la **lemma-tizzazione** o l'eliminazione delle cosiddette **parole vuote** o **strumentali** (articoli, congiunzioni, preposizioni) rimangono, per ampia parte, nella piena discrezionalità del ricercatore, che deve esercitare il massimo di prudenza e trasparenza per non stravolgere la struttura e il senso originario del testo oggetto di analisi.

Le tecniche
discriminanti

Sulla base di quanto fin qui illustrato si intuisce con evidenza il fatto che l'analisi multivariata dei dati attiva procedure che vanno ben al di là della semplice lettura della matrice-dati originaria e, attraverso operazioni congiuntamente di sintesi e di scomposizione, si dà come obiettivo anche classificazioni più mirate, aggregazioni e comparazioni interne a forte tasso di differenziazione. Tecniche come l'**analisi dei gruppi** e l'**analisi discriminante** rispondono perfettamente a tale esigenza di scomposizione e comparazione dell'universo sotto analisi. L'analisi discriminante, che sotto determinati profili è ritenuta un caso particolare dell'analisi in componenti principali, viene applicata a universi che sono già strutturati in una classificazione ordinata e dei quali si hanno sufficienti informazioni, già esplorate, per leggervi differenziazioni anche marcate. L'ulteriore passo che l'analisi discriminante realizza è quello di porre le classificazioni, o raggruppamenti già noti, in relazione a un insieme di variabili cardinali, in ipotesi discriminanti, che sono presenti e misurabili su tutti i casi analizzati. Tale operazione viene condotta attraverso l'estrazione, dalle variabili cardinali in precedenza selezionate, di **funzioni discriminanti** che, come avveniva per i fattori nell'analisi in componenti principali, consentono di attribuire punteggi, per l'appunto discriminanti, più sintetici a ognuno dei casi in analisi. Sulla base di questi punteggi si effettuano riclassificazioni e comparazioni che, aggiuntivamente alla migliore resa descrittiva e classificatoria dei dati in matrice, presentano anche un guadagno previsionale sulla base di determinati costanti leggibili sulle variabili selezionate: non è un caso che le tecniche discriminanti, per questa loro capacità di differenziazione tipologica

anticipata e fondata su informazioni certe già note, siano adoperate frequentemente dal sistema bancario e nella ricerca economica.

Nel caso dell'analisi discriminante il ricercatore possiede già una chiara impostazione di base delle differenziazioni interne all'universo sotto analisi sino a poter disporre, già dalle fasi iniziali, di raggruppamenti ben delimitati di soggetti o casi; ad esempio, chi vota e chi non vota, chi è iscritto all'università e chi non lo è, chi ha figli e chi no, chi è un cliente solvente e chi insolvente, chi appartiene a una fascia di età e chi a un'altra, chi svolge un lavoro dipendente e chi è un libero professionista, chi è occupato e chi disoccupato o in cerca di prima occupazione. Queste differenziazioni possono essere ulteriormente articolate in forme politomiche più complesse e possono essere scomposte, comparate e ricomposte attraverso più fasi di analisi previa. Nelle situazioni in cui tali differenziazioni non siano derivabili direttamente sia per la complessità, eterogeneità e numerosità dell'universo sotto analisi sia per il numero di variabili coinvolte sia per le incognite proprie di universi poco esplorati, il ricercatore può utilizzare una tecnica come la *cluster analysis*, che in funzione classificatoria e discriminante suddivide un universo in un certo numero di gruppi omogenei al loro interno e differenziati rispetto a ognuno degli altri gruppi. Le matrici che vengono utilizzate per raggiungere tale obiettivo sono delle **matrici di distanza** che impongono l'uso di variabili cardinali metriche. Quando questo criterio di base è soddisfatto, il ricercatore può vantaggiosamente ridurre la complessità interna della matrice e anche le fisiologiche ridondanze presenti ricorrendo a una previa azione di sintesi delle variabili in analisi attraverso la estrazione di fattori con la tecnica delle componenti principali. Su questi fattori ogni soggetto o caso otterrà un proprio punteggio e sulla base di questi punteggi si procede per successive iterazioni alla suddivisione in gruppi omogenei di tutto l'universo. Qualora il ricercatore si trovi a dover analizzare variabili categoriali, per poter attivare le procedure di raggruppamento dovrà previamente intervenire sui suoi dati con altre tecniche multivariate come l'analisi delle corrispondenze multiple (ACM), che consente di trasformare e trattare variabili categoriali alla stregua di variabili cardinali. Le procedure di raggruppamento attivano operazioni di classificazione e sintesi di notevole rilievo rispetto alle indeterminanze che la ricerca sociale affronta riferendosi a fenomeni sempre più complessi e fortemente dinamici dove la semplice classificazione e differenziazione costituisce in ogni caso un obiettivo importante per il ricercatore.

Al di là delle procedure di classificazione, di raggruppamento e di sintesi ogni ricercatore tende, anche proiettivamente, ad andare oltre il livello esplorativo e descrittivo per intercettare filiere di tipo esplicativo-causale. Anche se alla base vi sono relazioni molto strette fra impianto descrittivo ed esplicativo, nessun automatismo procedurale è possibile sul versante delle potenziali prospettive causali dell'analisi dei dati: occorre mettere in

campo uno specializzato disegno di ricerca e fare ricorso a tecniche di ricerca appropriate. La ricerca sociale deve inoltre affrontare due difficoltà aggiuntive: il fatto di avere a base delle proprie analisi un numero altissimo di variabili correlate alla complessità delle dinamiche che studia e la situazione particolarissima di non poter procedere con disegni sperimentali in laboratorio, in quanto il suo oggetto di studio – dinamiche e relazioni sociali – è difficilmente riproducibile in scala laboratoriale. In questa situazione il ricorso a **modelli di relazioni fra variabili**, in quanto rappresentazioni formalizzate e semplificate della struttura di un dato fenomeno, costituisce un ottimo strumento per cogliere nessi esplicativo-causali in situazione di indeterminazione e complessità. Il semplice fatto di poter porre in modo chiaro e sotto verifica relazioni, anche parziali e semplificate, fra variabili consente di sondare la tenuta di ipotesi di **dipendenza o indipendenza** fra di esse anche se sotto la particolare forma della **dipendenza statistica** e della **tendenzialità** stante il quadro problematico, tipico della ricerca sociale, in cui i **nessi potenzialmente causali** difficilmente possono essere letti in forma direzionale causativa, cioè asimmetrica, mentre al contrario molto spesso ci troviamo in presenza di relazioni simmetriche che rinviano a relazioni mutue, cioè a **variazioni concomitanti**. A questo riguardo l'analisi multivariata offre, a seconda delle situazioni di indagine, diversi modelli e tecniche che vanno dalla classica analisi della varianza ai modelli Logit, alla regressione binomiale e multinomiale, alla regressione multipla, alle equazioni strutturali, ai modelli lineari generalizzati (GLM), con il vantaggio che spesso si può operare con semplici trasformazioni sia con variabili cardinali che con variabili categoriali.

Le reti neurali
artificiali

Ulteriori passi possono esser compiuti in questa direzione con l'obiettivo di riprodurre e comprendere meccanismi tipici di fenomeni complessi o parti significative di essi attraverso l'utilizzo di modelli basati sulle **reti neurali artificiali (RNA)**. Si percorre in questo caso un territorio del tutto sperimentale che appare però suscettibile, non solo alla distanza, di notevoli potenzialità. Non è possibile affermare che vi sia corrispondenza o similitudine fra i fenomeni sociali e i processi neurofisiologici che sottostanno al funzionamento dei meccanismi del cervello; ma è certamente possibile, senza sovrapporle o assimilarle, accostare la complessità delle dinamiche sociali o socioeconomiche alla complessità cerebrale. Se si guarda a tali processi in termini di modellistica, si potranno al contrario individuare molti punti di contatto dai quali produrre **simulazioni**, anche se semplificate. Come nei meccanismi cerebrali, si può parlare anche per i sistemi sociali di **unità**, di **connessioni**, di **relazioni**, di **attivazioni** e **inibizioni** di contatti e connessioni, di qualità e peso delle connessioni, di **processi di apprendimento**, di **stati di equilibrio** e altro ancora. Si può tentare di costruire modellizzazioni di tipo neurale che collochino unità ritenute in ipotesi significative su più **strati** e valutare l'effetto sui vari strati dell'attivazione o dell'inibizione di

alcuni particolari processi in una direzione o in un'altra con un certo tipo di intensità energetica in vario modo controllabile. Con questo tipo di approccio si possono simulare, attraverso calcolatori e software specializzati, anche processi di apprendimento connessi a determinati input e avvicinare a situazioni di laboratorio parti di dinamiche sociali complesse impossibili da indagare a tale livello di complessità. Per il momento appare di tutto interesse il confronto sperimentale fra analisi di dati effettuate con metodi statistici tradizionali e analisi condotte con modelli RNA: le esemplificazioni che vengono presentate, pur essendo applicate a fenomeni non particolarmente complessi, fanno concludere per una convergenza di tutto rilievo fra i due approcci con il vantaggio di un più alto livello di identificazione delle variabili coinvolte e di un più alto livello di precisione.

Prospettive altrettanto interessanti possono derivarsi per la ricerca sociale da un altro tipo di approccio alla complessità quale quello del cosiddetto *data mining*, che, letteralmente, va a scavare all'interno di enormi *databases* per esplorarne, attraverso processi di selezione e modellizzazione, la potenziale conoscenza nascosta. Di fatto le società contemporanee producono una quantità infinita di informazioni che vanno a sedimentarsi in vere e proprie miniere digitali alla cui enormità dei valori di massa fisica corrispondono bassissime quote energetiche di esplorazione analitica a fini conoscitivi generali, eccezion fatta per limitate e utilitaristiche analisi settoriali come quelle di tipo aziendale. Da un lato quindi il *data mining* costituisce uno strumento applicato per strutturare analisi descrittive, esplorative e previsionali applicate, di piccolo e medio raggio, soprattutto su tematiche specializzate, per altro verso può essere visto come uno strumento potente e quasi una vera e propria metodologia nell'ambito del processo globale di produzione ed estrazione della conoscenza da grandi masse di dati che molto spesso non sono in relazione fra di loro. In una fase storica di forte interconnessione dei sistemi sociali e culturali un approccio quale quello del *data mining* costituisce una nuova e interessante opportunità per la ricerca sociale soprattutto nella direzione della stima, anche previsionale, della qualità delle mutazioni socioculturali in atto.

Da quanto fin qui sinteticamente illustrato apparirà chiaro al lettore e allo studioso che vi è una base comune nell'articolazione dei contenuti del presente volume, sia in termini di approccio logico che di approccio tecnico e operativo. Al lettore è quindi richiesto molto spesso dagli autori il costante rinvio ai singoli capitoli in un rapporto di reciprocità espositiva che ha guidato, volutamente, l'intero progetto di questo manuale. A fronte di questo piccolo sacrificio si nutre l'ambizione di riuscire a essere comprensibili a un pubblico più ampio del ristretto circuito accademico dei competenti sino a includervi i lettori di ricerche sociali e i divulgatori, e di rendere accessibili, o quanto meno consapevoli, alcuni meccanismi interni agli algoritmi di calcolo e alle logiche di progetto delle singole tecniche di analisi che, molto

Il *data mining*

spesso, rimangono criptici e riservati unicamente al dominio degli specialisti e di quanti posseggono già una buona base di conoscenze matematiche, statistiche e informatiche.

Ai lettori e ai nostri studenti il giudizio finale, a noi tutti la consapevolezza di aver fatto un onesto quanto laborioso tentativo e la speranza che, partendo da qui, altri possano fare meglio di noi.

Ringraziamenti

La redazione di questo volume non sarebbe stata possibile senza la collaborazione di colleghe che hanno variamente contribuito alla rilettura, revisione e uniformazione dei testi e dei riferimenti bibliografici: le dottoresse Brigida Blasi, Francesca Cuppone e Stefania Menchinelli dell'Università di Roma "La Sapienza", la prof. Agnese Vardanega dell'Università di Teramo e la prof. Stefania Vergati dell'Università di Roma "La Sapienza".

A loro va il ringraziamento dei curatori, che restano comunque i soli responsabili di eventuali errori e imprecisioni.