

Lo strano destino dell'analisi della struttura latente

di Giovanni Di Franco

Pubblicato in *Sociologia e Ricerca Sociale*, n. 58/59, 1999, pp. 317-364.

1. L'analisi della struttura latente e il linguaggio delle variabili

Scrivere dell'analisi della struttura latente (d'ora in poi Asl) dopo quasi mezzo secolo dalla originaria proposta di Lazarsfeld (1950a; 1950b) significa essenzialmente interrogarsi sulle ragioni del declino del cosiddetto linguaggio delle variabili nella ricerca sociale. Nell'ottica di Lazarsfeld e della sua scuola, infatti, l'Asl costituiva il necessario completamento della grammatica del linguaggio delle variabili.

Nel convegno dell'*American Sociological Association* del 1946 a Cleveland, Lazarsfeld avviò il suo programma di esplicitazione e di formalizzazione dell'analisi empirica nelle scienze sociali tenendo la relazione inaugurale dal titolo: *Interpretation of statistical relations as a research operation* (poi pubblicato in Lazarsfeld 1955b; tr. it., 1967). L'autore, sviluppando i contributi d'inizio secolo di Yule sulle cosiddette associazioni illusorie, mostrava che l'introduzione di una variabile di controllo, capace di rendere spuria la relazione tra altre due variabili, poteva costituire il fondamento di un modello generale di analisi multivariata in grado di "spiegare, interpretare e specificare" una originaria relazione bivariata oggetto di analisi.

Nagel valutò il contributo di Lazarsfeld nei seguenti termini: "codifica in maniera chiarificatrice i principali tipi di interpretazione che gli studiosi di scienze sociali frequentemente presentano quando spiegano relazioni empiricamente stabilite di dipendenza statistica" (Nagel, 1961; tr. it., 1968, p. 524).

Inoltre, in questo contributo, Lazarsfeld propose un criterio per individuare rapporti causali tra le variabili:

"Con questa analisi si può chiarire, almeno entri certi limiti, un ultimo punto. Si può suggerire una chiara definizione della *relazione causale* tra due attributi. Se si ha una relazione tra x e y ; e se le relazioni parziali tra x e y non scompaiono per nessun antecedente, allora si dovrebbe definire causale la relazione originaria. Qui è indifferente che le operazioni necessarie siano effettivamente svolte o solo rese plausibili con dei ragionamenti. Questi ragionamenti teorici, per inciso, consistono sempre in una delle quattro operazioni discusse qui, ad eccezione di quei casi in cui il ragionamento mira al chiarimento di sequenze temporali in cui si trovano variabili malamente scelte" (Lazarsfeld, 1955b; tr. it., 1967, p. 411, corsivi nel testo).

L'intero programma metodologico di Lazarsfeld e della sua scuola, teso alla formalizzazione di un linguaggio delle variabili, è chiaramente espresso nell'introduzione dello stesso autore al manuale di Hyman (1955):

"Il successo di ogni sforzo scientifico dipende da tre elementi: una chiara identificazione degli oggetti da investigare, una teoria immaginativa relativa al modo in cui essi sono collegati, e acute intuizioni sui problemi specifici dell'evidenza e dei dati che si dimostrano più adeguati per la materia di cui si tratta. [...] In altre parole noi ci siamo assunti il compito di chiarire a noi e agli altri i seguenti tipi di problemi: a) Come riconosciamo e classifichiamo degli oggetti sociali complessi? [...] b) Queste variabili come sono connesse fra di loro a vari livelli di complessità? [...] c) Così noi possiamo descrivere il mondo del sociale per mezzo di un complesso di variabili e studiare poi le interrelazioni fra di esse. Tuttavia il mondo del sociale non è immobile, né del resto vogliamo che lo sia: per conseguenza il problema dei cambiamenti nel tempo diviene una terza area cruciale di investigazione e chiarificazione. [...] d) La formazione di variabili, le loro

interrelazioni e i mutamenti connessi lungo un certo periodo di tempo, formano il nucleo centrale di quello che potremmo chiamare il “linguaggio della ricerca empirica”, tuttavia non vogliamo qui sostenere che questo linguaggio possa esprimere tutti gli interessi degli scienziati sociali, ma è vero invece che sono necessarie delle specifiche investigazioni per decidere, anche in linea di principio, se possiamo esprimere certe idee mediante una totalità, una *gestalt*, oppure se allo svolgimento di certe funzioni possano essere studiate con profitto usando un tale linguaggio di variabili. Il modo per scoprirlo è quello di prendere gli scritti degli scienziati sociali che hanno usato un approccio meno atomistico e più qualitativo e quindi di chiederci: ci guadagniamo in chiarezza se traduciamo le loro affermazioni in un linguaggio di variabili oppure facendo questo sforzo perdiamo alcuni aspetti essenziali del loro lavoro? [...] Gran parte delle discussioni nelle scienze sociali, anche quelle che si svolgono a un livello astratto o puramente qualitativo, possono essere chiarite se si trovano delle risposte alle seguenti domande. Quante variabili sono incluse in ogni particolare sequenza concettuale? Qual è la natura specifica di queste variabili? In che modo si pensa che siano interrelate? Queste tre domande chiarificano il *significato* delle proposizioni; esse non hanno necessariamente alcuna connessione con il problema dell’evidenza quantitativa, ma sono collegate alla ricerca empirica soltanto in modo indiretto. Nell’effettuare una particolare ricerca empirica concreta si è necessariamente obbligati a rispondere a queste domande; perciò si deve acquisire una certa capacità nel trattare con esse. E’ a questo punto che si rivela l’utilità dell’analisi sociologica quantitativa [...] l’investigatore decide quale informazione desidera raccogliere e da quale popolazione; una volta che ha raccolto le informazioni le trasforma in qualche sorta di codice, che generalmente viene poi trasferito su una scheda meccanografica perforata; a questo punto egli è libero di interrelare i vari elementi del suo codice a qualsiasi livello di complessità egli desideri. [...] Dal nostro punto di vista l’essenza della “*survey*” è di essere limitata al linguaggio delle variabili, ma allo stesso tempo essa ne può esaurire tutte le possibilità. [...] Il metodologo è il primo ad affermare che questo è uno sforzo processivo che non viene mai veramente completato e in cui il lavoro di uno studioso costruisce sugli sforzi dei suoi predecessori e attende i successi dei suoi successori” (Lazarsfeld, 1955d, pp. 13-21; corsivi nel testo).

In questo primo segmento del linguaggio delle variabili tutte le operazioni di ricerca coinvolgono variabili manifeste, ovvero presenti nella matrice dei dati; ma quando, nelle ipotesi del ricercatore, la variabile di controllo non è stata rilevata direttamente (caso ovviamente molto frequente nella ricerca sociologica) vengono meno i presupposti per effettuare le operazioni indicate dall’autore. Per superare questo limite, Lazarsfeld (1950a) propose l’AsI ispirandosi a tecniche psicometriche come l’analisi fattoriale e le varie forme di *scaling*. Dalla prima tradizione riprende il concetto di variabile latente come risultato di un procedimento di inferenza da un insieme di variabili manifeste e dalla seconda l’idea di costruire un procedimento di classificazione degli individui e/o delle variabili lungo una dimensione ad essi soggiacente. Detto questo, bisogna riconoscere a Lazarsfeld il merito di aver riformulato questi strumenti in una logica depurata da implicazioni scientiste, come ad esempio la pretesa di “misurare oggettivamente” (Spearman, 1904) la variabile latente o l’assumere una visione deterministica del legame tra variabili latenti e variabili manifeste, e soprattutto di essere stato in grado di predisporre strumenti d’analisi compatibili con la natura categoriale delle variabili prevalentemente in uso nella ricerca sociale. Infatti, a proposito della distinzione tra manifesto e latente l’autore afferma:

“Questa terminologia non implica altro che la distinzione tra informazioni ottenute dall’osservazione diretta e l’informazione ottenuta usando degli assunti ulteriori e/o traendo inferenze dai dati originari” (Lazarsfeld, 1950a, p. 364)

e , per quanto riguarda l’esito di tale analisi:

“Dal momento che questa “formalizzazione” del processo di costruzione del test è un postulato non è soggetto a verifica diretta, esso può essere giudicato solamente in termini di ragionevolezza e di utilità” (Lazarsfeld, 1950a, p. 379).

Infine, sul legame tra l'Asl e l'analisi fattoriale e sul problema della natura delle variabili l'autore afferma che:

“La presente teoria ha numerose similarità e differenze con l'analisi fattoriale. La più grande familiarità riguarda la logica. [...] E' un dato di fatto, il presente sistema è stato sviluppato dal desiderio di adattare l'analisi fattoriale a dati qualitativi. Precedenti autori hanno applicato l'analisi fattoriale a dicotomie producendo tabelle tetracoriche tra tutte le coppie di *items* e calcolando i coefficienti di correlazione di queste tabelle. Poi questi coefficienti sono immessi nella matrice di correlazioni e viene effettuata l'analisi fattoriale. Questo procedimento sembra essere ingiustificato perché ci sono altri coefficienti che possono essere calcolati sulle tabelle tetracoriche rispetto al coefficiente di correlazione di Pearson. L'analisi della struttura latente supera questa difficoltà non richiedendo il calcolo di coefficienti di correlazione. L'unico concetto che è necessario matematicamente è la nozione di indipendenza delle variabili, che ha un significato univoco per le dicotomie. [...] L'analisi della struttura latente non è altro che l'analisi fattoriale dove la matrice dei prodotti incrociati gioca il ruolo dei coefficienti di correlazione. [...] Nell'analisi della struttura latente però non si assume sempre una relazione lineare tra gli *items* e i fattori latenti in quanto si usano correlazioni superiori all'ordine zero” (Lazarsfeld, 1950b, pp. 469-72).

Dal punto di vista tecnico, quindi, l'Asl costituisce un tentativo di applicare i principi dell'analisi fattoriale alle variabili categoriali senza violare la loro natura. In altri termini, questo significa che non è necessario che le relazioni tra le variabili manifeste abbiano distribuzione multinormale. Un'altra somiglianza tra il modello dell'analisi fattoriale (in particolar modo nella sua versione confermativa) e l'Asl consiste nel fissare a priori la struttura latente e poi controllarne la capacità di riprodurre i dati manifesti; mentre, a differenza della fattoriale, l'Asl non presenta niente di simile alla rotazione dei fattori. Nel 1955, in un saggio dedicato ai recenti sviluppi nella modellistica dell'Asl, l'autore ribadisce la necessità di trattare le variabili categoriali in modo appropriato:

“Circa vent'anni fa il testo statistico più usato era quello originariamente scritto da Yule e curato da Kendall nelle edizioni successive. In un certo senso, era un libro curioso. I primi capitoli trattavano quanto gli autori definivano la statistica degli attributi. Il seguito del testo trattava le variabili quantitative e la loro distribuzione. Da allora in poi sono comparsi innumerevoli testi di statistica, ma sono tutti imperniati sulla statistica delle distribuzioni quantitative e sui molti sviluppi che questa ha avuto dacché apparve il volume di Yule e Kendall. La statistica degli attributi non viene affatto trattata, oppure è presentata nella stessa forma che aveva decenni fa. Probabilmente l'aspetto più curioso di questa situazione è quanto è accaduto nell'analisi dei fattori. Spesso il materiale disponibile consisteva in domande dicotomiche a due termini come: sì no, vero falso. Fra dicotomie del genere si può formare una tavola tetracorica. Allo scopo di trattare questo materiale in termini quantitativi furono escogitati stratagemmi di ogni sorta, come correlazioni tetracoriche e coefficienti simili. Molti autori avevano notato che questa traduzione di dati qualitativi in una formula quantitativa dava pessimi risultati. Eppure questa pratica era eseguita meccanicamente. Fu Louis Guttman il primo a far rilevare che i dati qualitativi richiedevano un proprio trattamento, e a proporre soluzioni originali e importanti per certi specifici problemi di scala. Ma, purtroppo, gli autori più giovani furono attratti dall'idea della misura senza rendersi conto di un'implicazione molto più generale: la maggior parte dei dati delle ricerche sociologiche sono di natura qualitativa e pertanto l'intero problema della statistica degli attributi deve avere preferenza assoluta. Specialmente se si vuole applicare la psicometria in campi come le ricerche sugli atteggiamenti o la sociometria, i testi correnti di statistica sugli atteggiamenti non sono quasi di alcuna utilità. L'analisi della struttura latente si inserisce nella tendenza moderna a superare le pseudo-quantificazioni e ad elaborare procedimenti analitici adeguati al tipo di dati qualitativi delle ricerche sociologiche” (Lazarsfeld, 1955a; tr. it., 1967, pp. 536-7).

In diverse occasioni Lazarsfeld, nella sua copiosa produzione – in un'antologia curata da Boudon (Lazarsfeld, 1993) è presente un'esauriente bibliografia – ha cercato di divulgare, in termini non specialistici, i fondamenti logici e le notevoli aspettative riposte nei modelli di Asl. Ad esempio, nel saggio del 1951, scritto con Barton:

“Recentemente è stata elaborata la tecnica della “analisi della struttura latente”, che estrae un sistema di punteggi da un modello matematico “adattato” ai dati empirici. Qui possiamo soltanto accennare a questo complicato argomento. [...] Non possiamo neppure discutere qui dei più complessi approcci formali-matematici questi problemi che sono stati ricordati prima, come l’analisi della struttura latente. L’analisi della struttura latente è stata sviluppata per potere fare per i dati qualitativi ciò che l’analisi dei fattori fa per i dati quantitativi. Quantunque sia soltanto ai suoi primi passi, l’analisi della struttura latente porta in sé la speranza di sistematizzare un certo numero dei problemi qui discussi in questo saggio: il problema di combinare elementi osservabili per formare concetti e categorie; il problema della relazione di un indicatore di un concetto complesso con altri; il problema di combinare gli indicatori in indici quantitativi; e altri ancora” (Lazarsfeld e Barton, 1951; tr. it., 1967, pp. 268-306).

Nel brano citato sono chiaramente espressi tutti i nodi problematici dell’Asl, ma anche le speranze che si riponevano in questo strumento per risolvere i numerosi problemi ancora sul tappeto.

Come ha notato Capecchi (1964, pp. 291-2), prima dell’introduzione dell’Asl i due procedimenti usati nelle ricerche sociali per limitare la numerosità delle classi (aggregare le classi tra di loro simili, ad esempio quando tra due sequenze cambia una sola risposta, oppure assegnare un punteggio, ad esempio il valore uno, per ogni risposta positiva ad ogni domanda) erano deterministici e piuttosto arbitrari:

“Ora è appunto per ovviare a queste arbitrarietà e a queste incertezze nella determinazione di un limitato numero di classi omogenee che è stata proposta da Lazarsfeld e dalla sua scuola l’analisi della struttura latente. [...] E l’elemento base costitutivo che si contrappone agli approcci precedentemente ricordati è che l’analisi della struttura latente comprende modelli tutti di tipo probabilistico mentre prima si trattava solo di approcci deterministici. Con l’analisi della struttura latente non si parte quindi da una decisione deterministica e arbitraria sulla composizione delle classi ma ci si propone di raggiungere questi obiettivi: a) numero minimo di classi omogenee in modo da verificare il modello e numerosità relativa dei soggetti; b) relazione probabilistica classi - domande: si cerca di individuare quali sono le probabilità di rispondere alle varie domande da parte dei soggetti appartenenti ad ognuna delle diverse classi omogenee; c) relazioni probabilistica soggetti - classi: si cerca di individuare, data una sequenza di risposte individuale, qual è la sua diversa probabilità di appartenere alle classi omogenee. Questi elementi incogniti sono stati da Lazarsfeld chiamati “latenti” in contrapposizione agli elementi noti (le frequenze di risposta) che sono definiti come “dati manifesti”. Si parlerà perciò di classi latenti, parametri latenti e struttura latente indicando con questo aggettivo “latente” gli obiettivi “non manifesti” che il ricercatore si propone di raggiungere” (Capecchi, 1964, p. 293).

Tornando a Lazarsfeld, nel saggio del 1959 *Problems in Methodology*, l’autore fornisce una definizione di variabile, di concetto e di tratto:

“[per variabile] io intendo ogni strumento tassonomico o ordinale, per cui si possano fare distinzioni tra persone o collettivi: la dimensione di una città, la situazione finanziaria di una società (dare - avere), il Q.I. di un individuo. Ciascuna di queste è una variabile. Il termine “concetto” verrà usato in un senso piuttosto ristretto. Concentrerò la mia attenzione su ciò che potremmo chiamare i concetti classificatori, come, ad esempio, la “coesione” dei gruppi, l’“aggressività” della gente, la “burocratizzazione” di una istituzione. In questo modo escluderò i concetti che vengono definiti verbalmente senza essere usati direttamente per scopi classificatori, come, ad esempio, le nozioni di “ruolo” e di “schema di riferimento”. Parlerò occasionalmente di questi concetti classificatori come di “tratti”. Nel caso di individui questo è del tutto convenzionale: siamo abituati a considerare i tratti come disposizioni più durevoli in contrasto con episodi evidentemente particolari di comportamento. Sul piano collettivo esiste la stessa relazione. [...] Per quanto si cerchi di dare una chiara definizione di un tratto, le nostre parole rimangono necessariamente imprecise in quanto non sono in grado di trasmettere con pienezza le nostre intenzioni concettuali né forniscono il lettore o l’ascoltatore di uno strumento inequivoco che gli consenta di decidere quando una certa persona o collettività possiede questo tratto, nonché il grado

in cui lo possiede. La traduzione di un tratto definito verbalmente in una “variabile” - espediente per classificare oggetti concreti secondo questo tratto - è sempre più o meno indeterminata” (Lazarsfeld, 1959; tr. it., 1967, pp. 188-9).

E poco dopo, Lazarsfeld chiarisce il rapporto tra le variabili latenti e gli indicatori:

“Dietro ogni tentativo di classificazione di questo tipo sta ciò che chiameremo un’*osservazione stimolante*: esistono variazioni e differenze che debbono essere spiegate. La spiegazione consiste in una proprietà latente o vagamente concepita, riguardo alla quale le persone o le collettività differiscono. Possiamo generalmente distinguere quattro passaggi nella traduzione di questa “immagine” in strumenti di ricerca empirica: 1) L’immagine originaria, la classificazione proposta viene espressa in parole e comunicata per mezzo di esempi; si fanno sforzi per una definizione; 2) Nel corso di questa verbalizzazione, spesso chiamata analisi concettuale, vengono nominati vari “indicatori” e questo aiuta a decidere dove collocare un determinato oggetto concreto (persona o gruppo o organizzazione) riguardo al nuovo concetto classificatore. Con l’allargarsi della discussione sul concetto aumenta il numero degli indicatori appropriati. Chiamerò l’insieme di questi “l’universo degli indicatori” (il termine è stato suggerito da Louis Guttman); 3) Generalmente questo universo è molto vasto e per fini pratici dobbiamo scegliere un *sottoschema* di indicatori che diventa la base per un lavoro empirico; 4) Alla fine dobbiamo raggruppare gli indicatori in un qualche tipo di indice. Quest’ultimo punto è stato trattato estesamente nell’attuale letteratura sulle misurazioni. Possiamo pensare a procedimenti molto semplici, come la somma delle risposte esatte per misurare il grado di conoscenza della geografia; o a procedimenti che richiedono modelli matematici complessi” (Lazarsfeld, 1959; tr. it., 1967, pp. 190-1, corsivi nel testo).

Nella sua trattazione, Lazarsfeld si sofferma sui punti 1-3, lodando gli autori dell’analisi sulla personalità autoritaria per la chiarezza con cui hanno documentato “insolitamente bene” il modo con cui sono arrivati alla scelta degli indicatori. Poi prosegue affermando:

“Quasi tutti i concetti classificatori hanno origine nel modo seguente: si osservano alcune variazioni empiriche; esse debbono essere spiegate da una nozione più generale, un “tratto sottostante”. Gli indicatori di questo tratto indicano la nuova unità da costruire, ma la loro scelta è dettata anche dall’osservazione iniziale” (Lazarsfeld, 1959; tr. it., 1967, p. 197).

L’autore definisce questo procedimento “piuttosto ingegnoso”; a proposito del problema dell’intercambiabilità degli indicatori afferma:

“Le discussioni riguardanti una qualsiasi variabile assumono spesso tocchi pirandelliani. Alcuni suggeriscono degli indicatori ed altri obiettano che essi non riescono ad afferrare “l’intero significato” della classificazione proposta. Quando si prendono in considerazione più indicatori viene elevata l’obiezione che quello non è un “concetto unitario” e che deve essere diviso in tre o quattro unità più “reali”. E’ sorprendente che in questa atmosfera qualcuno abbia il coraggio di avanzare e di effettuare una ricerca. Ma questo è stato fatto. E come risultato è stata lentamente sviluppata la teoria dell’intercambiabilità degli indici. L’esperienza ha dimostrato che, dato un largo universo di elementi, non fa molta differenza quale gruppo di essi venga scelto per formare lo strumento classificatore” (Lazarsfeld, 1959; tr. it., 1967, pp. 205-6).

Confrontando poi i due indicatori di produttività accademica definisce la natura probabilistica del rapporto di indicazione:

“A prima vista questo sembra un risultato piuttosto scoraggiante; l’eccellenza misurata da un indice è, in più di un terzo dei casi, diversa da quella misurata da un altro. Questo risultato è, tuttavia, inevitabile e di importanza limitata. E’ inevitabile perché gli indicatori possono avere, al massimo, una relazione probabilistica col fattore sottostante che si ricerca. [...] Questo non ha niente a che vedere col fatto che alcune delle cose che vogliamo classificare sono inafferrabili sul piano sociale o psicologico. Che due uomini siano amici, questo spesso si può constatare mediante

un'osservazione esterna, ma l'amicizia in se stessa non è un oggetto concreto che si possa scorgere direttamente. Per dedurre la sua esistenza si richiedono indicatori. Per dare un nome all'intero processo si potrebbe usare il termine "processo diagnostico". La certezza della inferenza da dati manifesti a caratteristiche latenti dipende da molti fattori: uno di questi è il grado con cui una domanda, usata come indicatore, è soggetta a varie interpretazioni. Raramente è possibile formulare una domanda cui le risposte attribuiscano un unico significato. Vi possono influire le esperienze degli interrogati immediatamente precedenti all'intervista. [...] In breve tutti gli indicatori si riferiscono ad una classificazione sottostante solo in modo probabile, tenendo presente che ogni classificazione di questo tipo nella ricerca sociale risulta essere necessariamente "impura" (Lazarsfeld, 1959; tr. it., 1967, pp. 207-8).

E finalmente Lazarsfeld chiude il cerchio, riaffermando la possibilità, attraverso l'uso di modelli matematici, di risolvere i problemi di indicazione e di classificazione (che nel suo linguaggio corrisponde alla misurazione). Si noti nel testo l'uso delle virgolette quando si riferisce alla purezza delle classificazioni e/o delle misure:

"Se questa regola della "intercambiabilità" degli indici è una delle basi della ricerca, se ne riceve beneficio ad un prezzo inevitabile e considerevole. Poiché non possiamo mai giungere a classificazioni "pure", un certo numero di casi viene necessariamente malclassificato e così i risultati empirici sono meno chiari di quanto sarebbero se potessimo in qualche modo avere misure precise per le variabili a cui uno studio si riferisce. La nozione che abbiamo ora sviluppato illumina in parte un altro problema, quello dello "operazionismo". Se le cose stanno come si è detto, in relazione ad un tipico problema dell'operazionismo come quello se l'intelligenza sia ciò che viene misurato da un test di intelligenza, allora vi sono tanti tipi di intelligenza quanti sono i tests. Altrimenti qual è la relazione tra i vari tests e il concetto "sottostante"? Si può dare una risposta precisa soltanto se la esprimiamo mediante modelli matematici che mostrano come una classificazione prestabilita consista di parametri che mediante delle equazioni vengono messi in relazione ai dati empirici e da questi possono essere calcolati. Ma anche senza tale precisione si può comprendere facilmente il triplice nesso che emerge da una analisi accurata dell'effettiva *pratica di ricerca*. Le cosiddette definizioni nominali sono essenzialmente dichiarazioni di scopi: l'investigatore comunica come meglio può la classificazione cui egli tende quando parla di intelligenza, coesione sociale, o posizione nella società. Egli poi procede ad una *specificazione del significato* sviluppando il suo universo degli indicatori, processo che, per principio, è illimitato. Ai fini di ricerca noi dobbiamo dunque lavorare con sottogruppi di questo universo - campioni di indicatori, se la parola campione è usata con una certa ampiezza. I vari tests di intelligenza, per esempio, sono campioni diversi di questo tipo. Sebbene essi possano essere soltanto moderatamente correlati l'uno con l'altro, essi sono intercambiabili nella loro funzione predittiva, ad esempio, circa il successo nella vita universitaria. [...] Non è la relazione degli indici l'uno con l'altro il problema cruciale, ma la loro relazione con le variabili esterne. [...] Noi non abbiamo in alcun modo esaurito il problema della formazione degli indici. Alcuni degli anelli mancanti sono stati discussi altrove. Così in molti studi accurati, prima della lista degli indicatori c'è una specificazione delle "dimensioni" secondo cui si dovrebbero cercare gli indici. [...] Abbiamo analizzato procedimenti oggi largamente usati nella ricerca sociale; probabilmente questi condurranno ad una sovrapproduzione di concetti e variabili. Non possiamo dire quali di queste sopravviverà in una fase più sistematica della sociologia empirica. Nel prossimo futuro ci possiamo attendere una più ampia azione reciproca tra gli studi empirici e una meditazione più astratta. Le esigenze di problemi specifici spingeranno verso la formazione di variabili *ad hoc*. Esse verranno classificate attraverso inventari e sforzi teorici, e saranno scelte le più basilari; a volte un'idea concettuale suggerirà una nuova variabile che può prendere il posto di numerosi tipi di classificazioni precedenti. [...] La combinazione di parecchie variabili — la chiameremo analisi a più variabili — ha una flessibilità sorprendente" (Lazarsfeld, 1959; tr. it., 1967, pp. 211-4, corsivi nel testo).

L'ultima parte del brano riportato mostra come per Lazarsfeld l'uso di procedimenti multivariati consenta un miglioramento delle pratiche della ricerca sociale. A questo punto gli elementi essenziali del linguaggio delle variabili sono individuati.

2. Il linguaggio delle variabili e il linguaggio della matematica

Abbiamo visto in che cosa consiste per Lazarsfeld il linguaggio delle variabili e come si collochi in questo contesto l'Asl. Rimane da chiarire quale fosse per l'autore il ruolo da affidare alla formalizzazione matematica nella ricerca sociale. Una prima chiara risposta si trova nell'introduzione al volume del 1954 *Mathematical Thinking in the Social Sciences*:

“Il ruolo del pensiero matematico nelle scienze sociali è divenuto l'oggetto di molte discussioni, controversie e tentativi pieni di speranza. L'origine di questo aumentato interesse è almeno duplice. Il successo della matematica nelle scienze naturali ed il suo prestigio e fascino costituiscono una tentazione per molti dei suoi interessati. In aggiunta, i sociologi e gli psicologi sociali hanno avvertito sempre di più la necessità per un linguaggio più rigido e preciso. Questo è vero per entrambe le tradizioni che sono state spesso indicate come l'approccio “macroscopico” per la spiegazione del comportamento umano. Chi procede con esperimenti o fa concrete osservazioni sui fenomeni della interazione sociale si trova confrontato con numero così grande di fattori che non può tenere testa ad essi adeguatamente se utilizza solo il linguaggio discorsivo o l'intuizione. Anche colui che costruisce grandi sistemi teorici, che non sono basati su innumerevoli dettagli, adesso lavora in un diverso clima intellettuale. Circa cento anni fa il compito sembrava quello di fare previsioni sullo sviluppo futuro della società. Cinquanta anni fa l'interesse era focalizzato sui concetti di base idonei a classificare i fenomeni sociali più rilevanti. Oggi il trend è verso il far emergere le variabili di base dalle quali possono essere derivati tutti gli specifici concetti e le interrelazioni. Anche quelli che non credono in nessun modo all'uso della matematica cercano di utilizzare qualche rudimento di formalizzazione in modo da chiarire le loro ipotesi sottostanti e derivare certi risultati specifici da modelli più generali. Nessuno può prevedere gli sviluppi del futuro intellettuale. Anche il più ardente ottimista non asserirebbe che la matematica ha già portato ad importanti scoperte nel campo delle scienze sociali. Il suo migliore argomento sarebbe che la matematica contribuisca alla chiarezza del pensiero e, permettendo una migliore organizzazione della conoscenza, faciliti le decisioni su ciò che occorre nel successivo lavoro. Per contro si può facilmente dire che il pessimista, che afferma una intrinseca incompatibilità tra i problemi degli scienziati sociali e la struttura del lavoro matematico, non è stato capace in nessun caso di portare argomenti validi; non vi è nessuna idea o proposizione in questo settore che non possa essere espressa in linguaggio matematico anche se l'utilità di fare così può essere messa in dubbio. [...] Come evolverà il ruolo della matematica nelle scienze sociali e più difficile da prevedere perché né la matematica né le scienze sociali sono qualche cosa di immutabile. Vi sono molti esempi storici nelle scienze naturali dove l'oggetto di analisi ha forzato nuovi sviluppi nel campo matematico. E per contro si può dire che gli effetti delle applicazioni matematiche influenzano il modo con cui le scienze sociali formulano il loro problema. Tuttavia non c'è nessuna ragione per sedersi e attendere passivamente per vedere ciò che accadrà. Lo sviluppo di una scienza del comportamento umano e delle relazioni sociali che sia utile e creativa è un compito criticamente importante per i nostri giorni. Nessuno può prevedere la rapidità di questo sviluppo. Ma l'applicazione del pensiero matematico è uno degli strumenti che possono aiutare; perciò esso deve essere utilizzato nel modo più razionale e vigoroso possibile. [...] Sono necessarie persone che siano ugualmente addestrate sia nella matematica che in qualche settore delle scienze sociali” (Lazarsfeld, 1954a, citato in Capecchi, 1967, pp. clxxxi-clxxxii).

Una seconda risposta si trova nel quinto paragrafo del saggio del 1958 *Evidence and Inference in Social Research*. L'autore individua due funzioni nell'uso di modelli matematici nelle scienze sociali: una sintattica ed una linguistica (o semantica):

“Come si usino i modelli, non ci sono dubbi che una importante funzione è l'aiuto a predire il comportamento. Ma c'è un'altra funzione dei modelli nelle scienze sociali che potrebbe essere detta funzione linguistica. [...] La funzione linguistica dei modelli matematici può essere divisa in tre aspetti: 1) una funzione organizzativa; 2) una funzione analitica e 3) una funzione di mediazione. La funzione organizzativa dei modelli matematici permette di ordinare una grande mole di variabili e di relazioni tra queste. [...] La funzione analitica permette nel caso contrario, ossia quando non si dispone di tutti i dati, di mettere a fuoco quali dati manchino nel modello. Questo uso analitico dei modelli matematici spesso conduce all'idea di quali nuovi studi empirici siano necessari. [...] Infine la funzione mediativa dei modelli consente la comunicazione

transdisciplinaria. [...] Solamente dopo che i problemi sono stati formalizzati è realmente possibile lavorare con un approccio interdisciplinare e scambiare mutui contributi da una disciplina a un'altra. Così la funzione linguistica di un modello matematico aiuta ad organizzare un'abbondanza di materiali, aiuta a cogliere deficienze di dati e aiuta a mediare tra procedure che sono formalmente simili ma terminologicamente differenti - una funzione dei modelli matematici che è stata spesso sottovalutata" (Lazarsfeld, 1958, pp. 124-8).

In sostanza, quindi, l'applicazione del linguaggio matematico alla ricerca sociale da un lato costituirebbe un potente strumento per ordinare ed organizzare insiemi di variabili e le loro relazioni in modo ordinato e coerente; dall'altro, fornirebbe una specie di esperanto utile a scienziati di discipline diverse per comunicare tra loro. In un altro passo del già citato saggio del 1959, Lazarsfeld torna sull'argomento della funzione linguistica della matematica, ma individua anche i rischi connessi a un eccesso di formalizzazione nell'attività di ricerca empirica:

"Le piuttosto vaghe relazioni tra le proposizioni che sono un aspetto inevitabile delle teorie proposizionali, anche se vengono ostentatamente presentate *more geometrico*, possono venire migliorate dalla matematica o anche da semplici formalizzazioni. [...] L'uso della formalizzazione nei dati sociologici per ora difficilmente può portare a nuove scoperte, ma può mettere in evidenza elementi non notati o chiarire relazioni tra proposizioni. Quest'ultimo punto risulta particolarmente chiaro nei pochi casi in cui si abbia una successiva matematicizzazione di un inventario sistematico, come ha fatto Herbert Simon per gli scritti di Festinger sulla comunicazione. [...] Simon dimostra che l'intero modello teorico diventa più complesso se tali incisi vi vengono incorporati, ma che nello stesso tempo alcune delle proposizioni risultano essere logici derivati di un più piccolo sottoinsieme proposizionale. [...] E' possibile che la tradizionale ricchezza di pensiero sociale possa essere tradotta in un linguaggio che considera gli oggetti della ricerca sociale come combinazioni di proprietà isolate che chiameremo "variabili", e le idee generali come relazioni tra queste "variabili"? La scheda I.B.M. è forse sul punto di privare le scienze sociali di ogni significato?" (Lazarsfeld, 1959; tr. it., 1967, pp. 185-7, corsivi nel testo).

Come è noto, all'approccio del linguaggio delle variabili, in generale, e all'Asl in particolare, sono state mosse aspre critiche soprattutto da parte di Mills e di Sorokin. Capecchi (1967) individua in queste critiche due componenti: a) la prima di tipo contingente, secondo la quale Lazarsfeld e la sua scuola sarebbero i principali esponenti della ricerca empirica "quantofrenica", incapace di fornire contributi sostanziali alla teoria e ai problemi politico-sociali; b) l'altra, di tipo non contingente, individua gli effettivi limiti della metodologia ed i problemi che si presentano quando si vogliono "misurare" i fenomeni sociali. In questa sede, interessano soprattutto le critiche di Sorokin all'Asl e al tema della misura nella ricerca empirica. Per prima cosa Sorokin critica l'atteggiamento quantofrenico:

"Al momento attuale, lo studio quantitativo dei fenomeni psicosociali è uno dei metodi essenziali di ricerca. Ora, fin tanto che il metodo si mantiene strettamente matematico ed è applicato a quei fatti psicosociali che si prestano all'analisi quantitativa si dimostra altamente fruttuoso e merita di essere coltivato in modo sempre più esteso. Ma quando il vero metodo quantitativo è rimpiazzato da imitazioni pseudomatematiche, quando ne viene fatto un uso sbagliato o se ne abusa in tutti i modi, quando viene applicato a fenomeni che non si prestano ad alcuna quantificazione, quando viene ad essere una vera e propria manipolazione nel vuoto dei simboli matematici o una semplice trascrizione di formule matematiche sulla carta, senza rapporti reali con i fatti psicosociali, allora la sua applicazione fallisce nel modo più clamoroso. In queste condizioni, l'uso del metodo matematico diventa una pura preoccupazione quantofrenica che non ha nulla in comune con la vera matematica e, soprattutto, non ci aiuta a penetrare nel mondo psicosociale. Negli ultimi decenni, a tutto danno delle scienze psicosociali, questa preoccupazione e mania quantofrenica è rapidamente dilagata e minaccia ora di coinvolgere moltissime ricerche non quantitative e di infrangere anche un buon numero di ricerche veramente quantitative. La corrente in questo senso è così forte che lo stadio attuale delle scienze psicosociali può ben essere chiamato *l'era della quantofrenia e della numerologia*. Tale malattia si manifesta sotto molte e

diverse forme e tocca tutti i settori della sociologia, della psicologia, della psichiatria e dell'antropologia" (Sorokin, 1956; tr. it., 1965, p. 110, corsivi nel testo).

In questo passo Sorokin dimostra di essere più realista del re, nel senso che esprime una concezione della "vera matematica" o delle "vere ricerche quantitative", assumendo una corrispondenza biunivoca tra matematica e la misurazione quantitativa che è falsa. Non esiste "La Matematica", esistono le matematiche. Inoltre, Sorokin non ha presente che la matematica "quando è certa non dice nulla del mondo reale, e quando dice qualcosa a proposito della nostra esperienza è incerta" (Einstein, 1966; tr. it., 1997, p. 125).

Ma veniamo alle critiche di Sorokin, dirette in primo luogo ai procedimenti di misura applicati alla ricerca sociale:

"Quanto al successo di questi sforzi si può prevedere in anticipo che se le qualità quantificate hanno delle unità, allora possono essere misurate o rapportate ad una scala cifrata e le loro misure possono essere espresse in numeri. Se le qualità quantificate non presentano unità allora non possono essere misurate in modo efficace. Se, a dispetto di questo, vengono tuttavia rese metriche qualità non misurabili, i risultati saranno per forza di cose fittizi ed arbitrariamente imposti ai fenomeni anziché presentarne un quadro obiettivo. La ragione di questo stato di cose è bene espressa dall'eminente fisico P. Appel: "Nelle formule matematiche le lettere indicano i numeri; queste formule si possono applicare solo a quantità misurabili che possano venir espresse con numeri. In geometria analitica x , y , z indicano numeri. Nelle equazioni di meccanica razionale i parametri x , y , z sono numeri". Dove non vi sono né unità né numeri, tutte le formule e le equazioni o sono prive di significato, o rappresentano un punteggio ed una classificazione soggettiva frutto del fanatismo dei devoti di una quantificazione fuori posto" (Sorokin, 1956; tr. it., 1965, p. 128).

E quindi a Lazarsfeld e al modello dell'AsI:

"Queste critiche si possono applicare, a maggior ragione, all'ipotesi delle "strutture continue latenti" che costituisce la premessa principale della convinzione di Lazarsfeld nella possibilità di classificare scalarmente i dati e i fenomeni qualitativi. Quest'ipotesi consiste nell'enunciazione del postulato secondo cui "esiste una serie di classi latenti tali che il rapporto tra due o più elementi (items) di un test è da attribuirsi alla esistenza di queste classi fondamentali e a queste sole [...] Ogni atteggiamento ha così due aspetti: uno associato alle categorie latenti e un altro che è specifico di ogni elemento (item)". In contrasto con Guttman, per il quale un atteggiamento è una reazione osservabile empiricamente, secondo Lazarsfeld l'atteggiamento è una inferenza compiuta sulle classi latenti, inferite, a loro volta, dai dati manifesti. "Il continuo latente è [in tal modo] una costruzione ipotetica". Abbiamo qui un esempio lampante di pura metafisica introdotta nelle scienze psicosociali moderne. "Pura metafisica", perché Lazarsfeld non ha alcuna base matematica, logica o empirica su cui fondare il suo postulato, secondo cui tutti o moltissimi elementi manifestamente non scalari rappresentano in realtà un *continuo* misurabile, per cui, quando si considerino tutte le classi latenti di questo *continuo*, gli elementi, apparentemente discontinui o non scalari, diventano continui e misurabili. [...] Di conseguenza, da un punto di vista matematico, Lazarsfeld non ha alcun fondamento per ammettere che i dati qualitativi sono sempre continui nelle loro categorie manifeste e latenti. Come tutte le altre teorie di fisica sociale la teoria di Lazarsfeld è, in larga misura, fondata su antiche tesi fisiche e matematiche, ed ignora completamente la teoria dei quanta e la microfisica moderna. Infatti la vera essenza della teoria dei quanta è il principio della discontinuità, dei "salti quantici" nel passaggio di piccole costellazioni di atomi da un livello di energia ad un altro. [...] Se il principio di discontinuità è il principio chiave in questo campo non c'è alcuna base per credere che i fenomeni psicosociali non possano essere anch'essi discontinui e imprevedibili. Dato che il postulato di Lazarsfeld ignora "le forze che solidificano il modello molecolare" enunciate da Plank, Delbrück, Heitler, London e Schrödinger e, poiché come abbiamo detto, ignora totalmente la teoria dei quanta della fisica moderna, possiamo dire a ragione che manca di una seria base fisica e matematica" (Sorokin, 1956; tr. it., 1965, pp. 132-3, corsivi nel testo).

Addirittura si invoca la teoria dei quanta e la fisica atomica. Lazarsfeld ha solo detto “ragioniamo come se ...”, non voleva mandare un missile sulla luna. Sorokin è certamente più scienziato di Lazarsfeld e conosce la matematica meno bene del metodologo austriaco, tra l’altro laureato in matematica. Un conto è la condivisibile condanna degli abusi e, soprattutto, di un uso inconsapevole di strumenti matematici nella ricerca sociale; un altro conto è negare qualsiasi possibilità all’adozione di metodi formalizzati nella ricerca sociale: sarebbe come buttare il bambino insieme all’acqua sporca.

3. La lunga latenza dell’analisi della struttura latente

Finora abbiamo inserito l’Asl nel contesto più generale del linguaggio delle variabili, parlandone in modo univoco, come se si trattasse di un unico procedimento. In realtà sono stati proposti da Lazarsfeld e dalla sua scuola numerosi modelli di Asl, tanto che quest’ultima si può definire come un concetto di genere che include diverse specie. Se si costruisce una tipologia, usando la dicotomia categoriale/cardinale sia per le variabili manifeste sia per le variabili latenti, otteniamo la tabella 1.

Tab. 1 I principali modelli di analisi della struttura latente

Variabili Manifeste	Variabili Latenti	
	Cardinali	Categoriali
Cardinali	Analisi Fattoriale	Analisi del Profilo latente
Categoriali	Analisi dei Trattati latenti	Analisi delle Classi latenti

Se si accetta questa impostazione, l’analisi fattoriale può essere considerata una specie del genere Asl. In questa sede, per ragioni di spazio ci occuperemo prevalentemente dell’analisi delle classi latenti che è idonea al trattamento di variabili categoriali manifeste (in prevalenza dicotomiche) e inferisce variabili latenti categoriali. Per gli altri modelli dell’Asl si rinvia a Lazarsfeld (1950a; 1950b; 1954b e 1955a), Lazarsfeld ed Henry (1968) e Capecci (1964; 1965).

Nell’analisi delle classi latenti (d’ora in poi Acl), il punto di partenza è costituito da un insieme di variabili manifeste categoriali che si assume siano interconnesse per costruzione. In altri termini, ciò significa che esse sono state scelte per formare gli *items* di una scala unidimensionale per la rilevazione di un dato atteggiamento o che sono ritenuti indicatori alternativi di uno stesso concetto. Ma, dato che il rapporto di indicazione non è bi-univoco, non tutta l’associazione rilevata tra le variabili manifeste sarà dovuta ad al rapporto con un unico tratto latente. Questi assunti sono analoghi a quelli dell’analisi fattoriale, dove le variabili manifeste sono riprodotte in parte dalla loro relazioni con i fattori comuni e in parte da una componente che si denomina l’unicità di ogni variabile.

Ora, dato che la struttura delle interconnessioni è dovuta soprattutto ai legami semantici di queste variabili con una comune proprietà (ad esempio un atteggiamento), si ipotizza che, se ci fosse un modo per far risaltare questa proprietà latente riscontreremmo, in ognuna delle categorie di questa variabile latente, l’annullamento delle associazioni tra ogni coppia di variabili manifeste.

Questo è il cosiddetto principio dell’indipendenza locale, che costituisce il fondamento dell’Acl e in generale di tutti i modelli di Asl. Questo principio rappresenta una radicalizzazione del principio della relazione spuria, che a sua volta è alla base di altri procedimenti multivariati (come ad esempio, l’analisi fattoriale classica). Ricolfi sostiene che: "Il [loro] principio chiave è il concetto di relazione indiretta o *correlazione spuria*. L’idea di base è che, in un mondo rappresentato mediante sistemi di equazioni di tipo lineare, la

correlazione tra due variabili possa essere dovuta non già all'effetto di una di esse sull'altra bensì dall'intervento di una o più variabili ulteriori, causalmente o logicamente antecedenti alle prime due. In termini statistici questa ipotesi equivale ad affermare che il coefficiente di correlazione fra due variabili è diverso da zero, ma si annulla quando si introducono una o più variabili di controllo" (Ricolfi, 1992a, p. 79, corsivi nel testo)¹.

Secondo Lazarsfeld, l'applicazione del principio dell'indipendenza locale nell'Acl comporta che :

“Un sistema dicotomico di m items è detto riducibile in λ classi se le seguenti condizioni sono soddisfatte: a) esistono λ sistemi omogenei di m items; b) un'unica corrispondenza è stabilita tra ogni frequenza nel sistema originale e una frequenza in ognuno dei sistemi omogenei; c) addizionando i corrispondenti items nei sistemi omogenei noi otteniamo il corrispondente item nel sistema originale” (Lazarsfeld, 1950a, p. 382).

Detto in termini diversi, le associazioni tra gli m items, registrate sul totale dei casi, devono annullarsi all'interno di ogni classe latente. Per far questo è necessario individuare a priori il numero di classi latenti all'interno delle quali si riscontra l'indipendenza locale tra gli items. Si devono, dunque, stimare i parametri incogniti di un sistema a 2^m equazioni (pari al numero delle sequenze di risposte manifeste), con un numero di incognite dato dalla formula $\lambda(m+1)$; dove λ è uguale al numero di classi latenti prefissate e m è uguale al numero di variabili dicotomiche manifeste. Se il numero dei dati manifesti (2^m), che corrisponde al numero di equazioni, è uguale al numero delle incognite [$\lambda(m+1)$], il sistema si dice identificato, ossia è risolvibile. Nei frequenti casi in cui il numero dei dati manifesti è maggiore del numero delle incognite il sistema si dice sovraidentificato, e si rendono necessarie delle condizioni per la sua riducibilità.

La ricerca delle condizioni di riducibilità dei sistemi d'equazioni divenne così “il problema” matematico che assillò per molti anni gli studiosi dell'Acl. Queste difficoltà di ordine matematico hanno reso inaccessibile l'Acl alla stragrande maggioranza dei potenziali utenti. Lo stesso Lazarsfeld, e con lui molti suoi allievi, si impegnò per molti anni nel tentativo di arrivare ad una soluzione soddisfacente. Non a caso tra la prima presentazione del modello nel 1950 e l'opera sistematica scritta con Henry passano ben diciotto anni.

Non deve quindi sorprendere il fatto che questo strumento di analisi dei dati, che era presentato come un settore importante del linguaggio delle variabili, non venne impiegato in modo diffuso nella ricerca sociale. Se si considera la letteratura, a parte le poche applicazioni

¹ Per illustrare la differenza tra il principio della correlazione spuria e quello dell'indipendenza locale dobbiamo richiamare il teorema della scomposizione della covarianza che, date due variabili cardinali x e y ed una variabile dicotomica z che divide il campione in due gruppi, si esprime con la seguente equazione:

$$\text{covarianza totale} = \text{covarianza tra i gruppi} + \text{covarianza all'interno dei gruppi}.$$

Per covarianza tra i gruppi si intende la quota di covarianza che viene riprodotta dall'individuazione dei due gruppi lungo le due categorie della variabile z , mentre la covarianza interna ai gruppi indica la covarianza che rimane dopo aver introdotto la variabile di controllo z . "La differenza fra il principio della correlazione spuria e il principio di indipendenza locale è che il primo richiede soltanto l'annullamento della media delle covarianze locali (i gruppi devono essere mediamente privi di struttura), il secondo richiede anche l'annullamento delle singole covarianze locali (ogni gruppo deve essere privo di struttura):

$$\begin{array}{ll} \text{Correlazione spuria:} & p_1 \text{ cov}(xy | 1) + p_2 \text{ cov}(xy | 2) = 0; \\ \text{Indipendenza locale:} & \text{cov}(xy | 1) = \text{cov}(xy | 2) = 0. \end{array}$$

Si vede bene, dal confronto delle due formule, come il secondo principio sia una radicalizzazione del primo: si può avere correlazione spuria senza indipendenza locale – ad esempio quando due relazioni locali di segno opposto si elidono – ma non viceversa" (Ricolfi, 1992a, p. 81).

di Lazarsfeld e della sua scuola, non ne sono reperibili molte altre. Già nel 1964, ovvero in un periodo di pieno sviluppo dell'applicazione di modelli matematici nelle scienze sociali, a proposito dell'applicazione dell'Asl in casi concreti di ricerche empiriche, Capecchi faceva notare che: "tenendo presente l'interesse di questo modello non si può perciò dire che le applicazioni siano state copiose" (Capecchi, 1964, p. 315). Per Clogg ciò era dovuto soprattutto ai problemi di ordine matematico: "queste difficoltà comportarono che i metodi della struttura latente furono largamente inaccessibili per la maggioranza dei ricercatori sociali. La completa assenza di analisi convincenti dal punto di vista sostantivo di dati sociali mediante l'analisi della struttura latente nelle riviste scientifiche sociologiche testimonia questa situazione. [...] Il grosso potenziale della tecnica della struttura latente come un linguaggio generale per l'espressione di teoria sociale è stato virtualmente disatteso" (1981, pp. 216-7).

E' molto probabile che proprio l'indisponibilità di programmi automatici di calcolo abbia frenato per molti anni la diffusione delle applicazioni dell'Asl nelle ricerche sociologiche, date le difficoltà di ordine matematico e computazionale. Però, nella seconda metà degli anni settanta, grazie alla diffusione dei personal computer e, soprattutto, ai lavori di Goodman (1974a; 1974b), queste difficoltà potevano essere superate. Infatti, a partire dalla fine degli anni settanta cominciarono ad essere disponibili i primi programmi per l'Asl per i ricercatori sociali che non richiedevano da parte degli utenti delle competenze matematiche specialistiche (Clogg, 1977). I più recenti contributi di Clogg (1981) e di McCutcheon (1987), molto più accessibili degli originari lavori di Lazarsfeld (1950a; 1950b) e soprattutto dell'opera sistematica di Lazarsfeld ed Henry (1968), non hanno modificato sostanzialmente lo stato di latenza dell'analisi della struttura latente. Ciò è confermato dal fatto che l'Asl è oggi assente nel più diffuso programma commerciale di analisi statistica dei dati tra i ricercatori delle scienze sociali.

La spiegazione di questo stato di cose è diretta conseguenza di due fattori: uno temporale e uno tecnologico. Per quanto riguarda il primo punto, Lazarsfeld (e la sua scuola), pur avendo iniziato a lavorare al programma del linguaggio delle variabili a metà degli anni quaranta, produce dei risultati consistenti solo negli anni '60: in particolare *The Algebra of Dichotomous System* (1961b; tr. it. parziale, 1967), e il già citato lavoro con Henry del 1968. Nel primo contributo, in particolare, si mostra la necessità di passare a considerare frequenze di ordine superiore sulla cui base formulare indici di dipendenza condizionati e non solo indici di dipendenza parziali (Capecchi, 1967, p. lxxviii).

Inoltre questi risultati non forniscono programmi di calcolo da utilizzare nella pratica della ricerca sociale.

Su questo punto è bene fare una precisazione ulteriore. Tutti i procedimenti di analisi dei dati che possiamo definire modelli (Di Franco, 1997) dedicati all'elaborazione di variabili categoriali (quindi anche i modelli log-lineari, ad esempio) sono, rispetto ai più semplici e più conosciuti modelli per variabili cardinali, più difficili da gestire da parte di utenti non sufficientemente competenti dal punto di vista sia teorico (ossia delle conoscenze sostantive nel campo d'indagine) sia statistico, anche quando questi siano implementati su supporti informatici di analisi dei dati. Questo perché con le variabili categoriali la struttura delle relazioni possibili porta alla necessità di considerare frequenze di ordine successivo al secondo, ovvero i cosiddetti effetti di interazione di ordine superiore². Inoltre, per analizzare tutte le combinazioni tra anche solo una decina di variabili categoriali si rendono necessari campioni di ampiezza molto superiore ai normali standard della ricerca sociale. Non a caso,

² Ad esempio, con appena tre variabili dicotomiche abbiamo una frequenza di terzo ordine, tre di secondo e tre di primo. I tipi di relazioni possibili non sono solo lineari (come in quasi tutti i modelli per variabili cardinali), ma si devono considerare altre forme di relazione.

nelle poche volte che in letteratura si reperiscono applicazioni di modelli per variabili categoriali le variabili coinvolte non sono quasi mai più di quattro o cinque.

E' comprensibile, anche se non è giustificabile, il perché ancora oggi, malgrado siano a disposizione strumenti informatici che rendono disponibili modelli per variabili categoriali, si continuano ad applicare i modelli per variabili cardinali anche su variabili categoriali, violando più o meno consapevolmente le condizioni di applicabilità di questi modelli.

Ma, al di là di queste perduranti difficoltà tecniche nell'uso dei modelli per variabili categoriali c'è, a mio avviso, una ragione ancora più importante di ordine ideologico. In questo caso dobbiamo chiamare in causa lo scientismo di alcuni ricercatori sociali, che forse nasce da un senso di inferiorità rispetto ad altre scienze ritenute più sviluppate, che ha prodotto una sorta di deriva del linguaggio delle variabili verso i lidi di un linguaggio fortemente matematizzato. Voglio precisare che non ritengo responsabile Lazarsfeld di questa deriva, ma i suoi successori che, concependo il linguaggio matematico più come fine che come mezzo, hanno attinto a piene mani da riviste come *Psychometrika*, *Econometrika* e *Biometrika* e hanno trovato piena soddisfazione nella fusione della tradizione psicometrica (nelle due varianti dell'analisi fattoriale e dei modelli di misurazione) con quella econometrica (consistente nei modelli d'equazioni strutturali) compiuta nel programma LISREL (si vedano i riferimenti in bibliografia ad autori come Jöreskog, Sörbom, Lawley, van Thillo, Saris e Stronkhorst).

Così, a partire dagli anni '60, si assiste ad un cambio di rotta del linguaggio delle variabili, soprattutto per merito di autori come Simon (1954), Blalock (1961) e Boudon (1965), i quali elaborano i procedimenti che prendono il nome di modelli causali in sociologia, e poco dopo con Duncan (1966), che importa nella ricerca sociale il modello della *path analysis*. In questo contesto la maggior parte degli sforzi dei ricercatori sono rivolti alla soluzione di problemi tecnico-matematici: come rendere i sistemi d'equazioni identificabili e come valutare la bontà dell'adattamento ai dati empirici (*goodness of fit*)³. Si trascura, inoltre, la natura artificiale e riduttiva di questi modelli lineari e, soprattutto, i loro assunti (linearità delle distribuzioni, additività degli effetti, errori di rilevazione trascurabili, casualità e indipendenza reciproca dei residui). Con questo non si vuole generalizzare, sostenendo che questa modellistica sia totalmente inutile; tuttavia, senza una buona informazione sulla natura delle proprietà organizzate in quelle variabili, la loro organizzazione in un modello matematico può essere un vuoto esercizio che solleva più problemi di interpretazione di quanti non ne risolva (Perrone, 1977).

Possiamo dire che qualsiasi processo di analisi dei dati richiede molte iterazioni e specificazioni successive, con una continua tensione dialettica tra premesse teoriche e stime empiriche delle conseguenze di tali premesse. Per queste ragioni possiamo considerare tutta la modellistica in uso nella ricerca sociale come un insieme di strumenti per controllare, raffinare e migliorare le articolazioni interne di una teoria. Il loro uso è quindi vincolato alla conoscenza dell'oggetto di indagine; conoscenza necessaria all'individuazione delle variabili rilevanti e dei legami esistenti tra loro. D'altra parte è bene ricordare che la migliore tradizione di sociologia empirica si è sviluppata senza aderire appieno né alla posizione positivista né a quella ermeneutica. Inoltre, nei lavori di Lazarsfeld e della sua scuola si riconosce spesso la necessità che il ricercatore si sforzi nell'interpretare il senso ed i significati intersoggettivi dell'agire sociale sia nel momento della concettualizzazione che in quello della operativizzazione e dell'interpretazione dei dati. Mostrando così di aver appreso

³ In sostanza, il compito di un modello è quello di riprodurre i dati empirici. Se si lavora con variabili cardinali, il punto di partenza è, di solito, una matrice delle correlazioni. Il ricercatore definisce, in base alle sue ipotesi sulla struttura delle relazioni tra le variabili, il ruolo da assegnare alle variabili inserite nel modello e poi controlla se il suo modello è in grado di riprodurre la matrice delle correlazioni empiriche. Se l'esito di questo controllo è soddisfacente, si dice che il modello si adatta ai dati.

l'insegnamento di Weber quando afferma che: “[E viceversa] i dati statistici [...], ovunque concernano il corso oppure le conseguenze di un atteggiamento che racchiude in sé qualcosa di interpretabile in maniera comprensibile, risultano per noi “spiegati” solo se vengono anche realmente interpretati in modo dotato di senso nel caso concreto” (1922, 2° ed. 1951; tr. it. 1958, p. 254).

La necessità di interpretare o comprendere i risultati delle analisi, oltre a quella di spiegarli, non è solo una prerogativa del momento esplorativo di una ricerca, ma deve essere considerata una guida costante anche nel corso di tutte le operazioni tecniche di controllo empirico. Questa conclusione dovrebbe essere condivisa, a mio parere, da chi ha veramente esperienza diretta di ricerca sociale empirica: non c'è operazione di ricerca, dalla fase di progettazione di un'indagine a quella dell'analisi e interpretazione dei dati, che non richieda costanti sforzi di comprensione ed interpretazione tesi all'acquisizione di una consapevolezza nelle scelte da compiere (Marradi, 1996).

La radicalizzazione del linguaggio delle variabili in senso matematico, fortunatamente, non è stata l'unica strada intrapresa dai ricercatori delle scienze sociali. C'è, infatti, tutta un'altra tradizione di ricerca che, a partire intorno agli anni '70, sostituisce il linguaggio delle variabili con quello che potremmo definire il linguaggio dei casi. In quest'altra tradizione l'obiettivo di fondo è di tipo prevalentemente esplorativo: si accantonano in un primo momento le istanze inferenziali e si usano le tecniche di analisi multivariata (come l'analisi delle corrispondenze multiple, l'analisi in componenti principali e la *cluster analysis*) per sintetizzare in modo parsimonioso l'informazione disponibile (Di Franco, 1997).

Negli ultimi anni si assiste ad una forte integrazione tra gli strumenti di analisi multivariata dei dati provenienti da diverse tradizioni (integrazione che mira ad una ibridazione degli stessi)⁴, mentre continua a ritmo sostenuto la produzione di nuovi strumenti come le reti neurali (Di Franco, 1998), i metodi di classificazione *fuzzy*, le reti neuro-*fuzzy*, etc.

Nel commentare l'attuale situazione nel panorama dell'analisi dei dati nella ricerca sociale Ricolfi sostiene, con un certo pessimismo: “Al vecchio progetto, forse un po' riduzionista, di *riconduurre* al linguaggio delle variabili *tutto* il lavoro sui dati, sembra sostituirsi una tendenza di segno opposto, quella a *dilatare* il concetto di analisi multivariata fino ad includervi *qualsiasi* tecnica di analisi dei dati. Al primato della *semantica*, implicito nell'idea lazarsfeldiana del linguaggio delle variabili, succede oggi un insidioso primato della *sintassi*, una tendenza all'uso di tecniche sempre più sofisticate al di fuori di uno schema di riferimento concettuale adeguato. In questa situazione, in cui lo scarto fra complessità *tecnica* degli strumenti d'analisi e l'irrilevanza *sostantiva* di tante applicazioni sembra destinato a crescere indefinitamente, diventa difficile non guardare con qualche nostalgia ai tempi di Lazarsfeld, a un'epoca in cui non era infrequente che “operazioni di ricerca” relativamente semplici conducessero a risultati di grande portata teorica” (Ricolfi, 1993, p. 23, corsivi nel testo).

A conclusione della prima parte del presente lavoro, concordando con l'auspicio formulato da Ricolfi, ossia di fare alcuni passi indietro rispetto al piano sintattico per cercare di farne alcuni in avanti su quello sostantivo, ci proponiamo, nei seguenti paragrafi, di valutare se e come sia possibile riscoprire l'AcI nelle concrete operazioni di ricerca sociale.

⁴ Ad esempio, da alcuni anni si assiste ad un avvicinamento ed un'integrazione tra la tradizione descrittivo-esplorativa della scuola di analisi dei dati francese e la tradizione inferenziale della scuola anglosassone.

4. Il modello dell'analisi delle classi latenti in Lazarsfeld

Cominciamo la presentazione del modello delle classi latenti prendendo in considerazione un saggio del 1954 dal titolo *A Conceptual Introduction to Latent Structure Analysis* (tr. it., 1967, pp. 447-524), nel quale Lazarsfeld cerca di esprimere i concetti relativi all'Acl in modo non eccessivamente formalizzato. Il punto di partenza dell'Acl consiste nell'applicazione di un modello matematico a un problema di misura (che, come visto, corrisponde per Lazarsfeld ad un problema di classificazione). Si tratta di un problema tipico dell'approccio psicometrico e consiste nel predisporre un certo numero di test (per la rilevazione di un atteggiamento o di una data capacità) in una batteria di *items*. Per semplicità, assumeremo che questi siano costituiti solo da dicotomie (ma il fatto che possano essere politomie non comporta sostanziali conseguenze, se non nella prolissità delle formalizzazioni matematiche di cui ci varremo). Queste batterie vengono sottoposte ad un campione di soggetti e poi si attribuisce a ciascun individuo, sulla base delle risposte fornite, un punteggio che rappresenta lo stato dei soggetti sull'atteggiamento o sulla proprietà considerata. Nella maggior parte dei casi, soprattutto nella ricerca sociale, gli *items* sono composti da frasi che in qualche modo sono semanticamente legate alla proprietà di cui sono ritenuti essere dei validi indicatori. Tralasciando i problemi legati alla scelta degli indicatori (e quindi alla costruzione delle scale), in questa sede ci soffermiamo sulle procedure di analisi dei dati. Secondo Lazarsfeld, il primo problema che l'Acl vuole risolvere è:

“[...] rendere misurabili tali serie di rilevazioni qualitative. I tipi di trattamento ai quali si possono sottoporre tali tests dettagliati sono limitati. L'analisi della struttura latente mira a fornire modelli matematici che permettano di mettere in relazione le risposte date ai vari tests. Lo scopo principale del modello è *mettere in evidenza i presupposti impliciti in questo tipo di "misurazione"*. Non si richiede che gli esecutori della misurazione siano consapevoli di questi presupposti, né si richiede che un altro modello non riproduca altrettanto bene le varie operazioni che furono eseguite o che si potrebbero ideare. Ma affermiamo che l'analisi della struttura latente dà logica forma di assiomi alle pratiche e ai dibattiti nel campo della misurazione, e che i suoi assiomi consentono operazioni algebriche che inducono a relazioni non ancora osservate e precisano il significato della nozione di misura nelle scienze sociali.” (Lazarsfeld, 1954b; tr. it., 1967, pp. 447-8, corsivi nel testo).

Con l'Acl è possibile produrre, a partire da un insieme di semplici frasi, delle inferenze sui concetti disposizionali. Questi sono definiti come concetti che non si riferiscono ad una caratteristica direttamente rilevabile, bensì alla tendenza a mostrare particolari “disposizioni” in determinate circostanze (Lazarsfeld, 1954b; tr. it., 1967, p. 449). Lazarsfeld esemplifica questo punto prendendo in considerazione il lavoro di un medico o di un investigatore. Nel lavoro d'indagine dei medici e dei poliziotti, così come nella ricerca sociale, si stabiliscono dei nessi “tra una serie di dati disponibili e una classificazione più basilare su cui converge il loro interesse” (Lazarsfeld, 1954b; tr. it., 1967, p. 451). Questo “paradigma indiziario”, secondo Lazarsfeld, non si applica solo ai concetti di atteggiamento, ma a tutti i concetti disposizionali. Per cui, una volta che si è in grado di costituire un insieme coordinato di indicatori (giustificandone e argomentandone la scelta sulla base di considerazioni teoriche, di intuizioni o di precedenti esperienze empiriche) che permettano le operazioni di rilevazione dei dati, si pone il problema di inferire dai dati manifesti la loro struttura latente. Su come effettuare questo passaggio Lazarsfeld afferma:

“E' su questo punto che la maggior parte degli autori sono molto vaghi. Se si deve sviluppare un coerente sistema d'idee, includente operazioni concrete, allora si deve stabilire *con quanta precisione si debbano trarre inferenze dalle osservazioni concrete ai concetti "sottostanti"*. In questo scritto indagheremo una simile possibilità; non si avvanzerà la pretesa che sia l'unica, e neppure la migliore. Ma per quanto ne sappiamo, è il primo tentativo di attuare

praticamente il programma implicito nel tipo di ragionamento che Weber, James e Henderson consideravano un utile approccio alle scienze sociali” (Lazarsfeld, 1954b; tr. it., 1967, pp. 452-4, corsivi nel testo).

A questo punto si introduce il concetto di probabilità, essenziale nel modello dell’Acl, come “tendenza ad agire”:

“Questo è appunto un significato del termine probabilità nel presente testo. [...] Ne accettiamo a fondamento la cosiddetta interpretazione in termini di frequenza. [...] Inizieremo con una *classe referenziale* R di n elementi di cui s godono della proprietà A; la nostra attenzione si dirige poi verso la proporzione $p = s/n$. Ora, supponiamo d’aver motivo di ritenere che il valore di p rimanga pressoché invariato se aumentiamo la dimensione n della classe referenziale o se ne consideriamo solo una parte scelta arbitrariamente. [...] Allora questa ideale proporzione (limite di p) di elementi che godono della qualità A nella classe referenziale R si chiama la probabilità di A in R. [...] Tale probabilità sarebbe utile per misurare se Brown è pessimista sul futuro politico. Qui la classe referenziale è l’insieme delle interviste ripetute in condizioni di “lavaggio del cervello”. Questa naturalmente è un’idealizzazione cui, in situazioni concrete, ci si può solo approssimare. Ma anche la classe di tutti i trentenni, come viene impiegata nel calcolo delle probabilità di uso comune, risulta, a ben pensarci, un’idealizzazione che, nelle ricerche pratiche, esige le sue approssimazioni [...] quando parliamo di probabilità in questa sede, le classi referenziali sono sempre interviste o osservazioni ipotetiche ripetute, *fatte sullo stesso soggetto*, col presupposto che tutte le risposte precedenti siano dimenticate dal soggetto appena date. [...] Nello stesso modo tendenze e probabilità possono variare nel tempo ed essere tuttavia definite in termini di una classe referenziale di indagini ripetute che si presumono condotte in condizioni di stabilità” (Lazarsfeld, 1954b; tr. it., 1967, pp. 454-6, corsivi nel testo).

Quindi, ricorrendo ad esempi un po’ troppo irrealistici prosegue:

“Un meccanismo probabilistico è quindi una struttura, di qualunque tipo, la quale, in seguito a ripetuti esperimenti, dà risultati che si approssimano sempre di più a una percentuale di risposte determinata in precedenza. *Perciò la probabilità è una proprietà della struttura.* [...] Chiameremo equivalenti due meccanismi probabilistici quando, in base alla nostra conoscenza della loro struttura fisica o ai risultati di precedenti esperimenti, prevediamo che ciascuno produca la stessa proporzione di risposte positive in periodo di tempo abbastanza lungo. [...] D’ora innanzi quindi avrà senso dire che in un certo momento un individuo ha una probabilità p di rispondere positivamente a una data domanda. Ciò significherebbe che, con una superiore conoscenza psicofisiologica, noi saremmo disposti a sostenere quanto segue: se la struttura impegnata nel rispondere a tale domanda rimanesse costante, allora la percentuale di risposte positive date dall’individuo su molte interviste ripetute, sarebbe p per cento. Neppure allora saremmo in grado di prevedere quale risposta egli potrebbe dare a un’intervista successiva” (Lazarsfeld, 1954b; tr. it., 1967, pp. 458-9, corsivi nel testo).

Più avanti Lazarsfeld chiarisce quali sono gli assunti impliciti del modello:

“Vogliamo chiarire i presupposti impliciti in un determinato tipo di misurazione. Questa misurazione è costituita dall’uso di un test composto da più domande per classificare le persone secondo una caratteristica “sottostante”, x. Operiamo con domande perché ci aspettiamo che, in qualche modo ancora indeterminato, queste siano *indicative* di quanto vogliamo effettivamente stabilire. In altre parole, v’è sempre il presupposto che la tendenza di un individuo a rispondere positivamente ad una data domanda sia in qualche modo in relazione con la sua caratteristica sottostante. [...] Così il valore espressivo di una risposta o di un’osservazione singola sarà sempre alquanto dubbio. Il buon senso ci dice che una maggiore quantità di domande migliorerà le nostre inferenze. Di conseguenza usiamo un test costituito da una batteria di domande o elementi. Il modello d’analisi della struttura latente ci consente d’essere più precisi sul modo di produrre queste inferenze; esso evidenzia le ipotesi implicite in vari procedimenti di misura. Sviluppando il modello, troveremo utile incorporare la nozione di meccanismo probabilistico in uno schema più complesso. Per riuscirvi, ci costruiremo un mondo simile a quello dei racconti di fantascienza. [...] Come vengono scelte le diverse domande? Per rispondere a questo interrogativo, dobbiamo tornare

alla nostra discussione iniziale. Si scelgono determinate domande perché, per ipotesi, esprimono qualche atteggiamento o proprietà sottostante che può designarsi con x . Cioè, un test non è formato da una raccolta casuale di domande; esso comprende invece soltanto quelle che il ricercatore ritiene riveleranno quanto sta tentando di misurare. [...] Se le domande scelte esprimono effettivamente un tratto sottostante, *la probabilità* di rispondere positivamente ad ognuna (*non* la risposta effettiva in una determinata situazione) sarà predeterminata dal tratto sottostante. [...] Un'ultima ipotesi sulle diverse classi latenti è che si può allinearle od ordinarle secondo il valore tipico di ognuna. Questa è un'ipotesi che non sempre può essere formulata. Vi sono alcuni casi in cui è necessario parlare di classi non ordinate, per esempio quando si tratta di continui latenti che hanno più di una dimensione” (Lazarsfeld, 1954b; tr. it., 1967, pp. 460-3, corsivi nel testo).

Agli assunti esplicitati nel brano sopra riportato si deve aggiungere anche il principio dell'indipendenza locale, in quanto si deve postulare che l'intervistato si collochi costantemente in una classe latente.

Per rappresentare la struttura latente si devono quindi inferire sia le frequenze relative di ogni classe latente, sia le probabilità latenti associate ad ogni variabile manifesta rispetto alle classi latenti. Se queste sono in numero finito, come nel caso dell'Acl, si ha la cosiddetta *strutturazione*; se vi è un numero infinito di classi, si parla di *traccia* latente (Lazarsfeld, 1954b; tr. it., 1967, p. 465). Né le frequenze relative delle classi latenti (v^x) né le varie probabilità latenti (p^x) sono note. Si vogliono stimare queste incognite partendo dai dati assumendo: a) che le classi del sistema “sottostante” – e quindi le persone che contengono – siano caratterizzate da differenti probabilità di rispondere positivamente a ciascun *items*; b) che una determinata risposta di un individuo ad un particolare elemento sia una “manifestazione” di questa probabilità. Per risolvere il problema della stima dei parametri incogniti, si considerano le sequenze di risposte alle variabili dicotomiche manifeste. Il numero di possibili sequenze di risposte varia al variare del numero e delle categorie delle variabili manifeste usate. Se questo insieme è costituito da k domande dicotomiche, le possibili sequenze sono 2^k . Per questo insieme di sequenze si calcolano le frequenze delle sequenze (ovvero il numero di persone che hanno dato la stessa sequenza di risposta). Pur essendoci una dipendenza tra le frequenze di sequenze e le sequenze di risposte, le prime differiscono dalle seconde in quanto esse rappresentano tutte le persone che hanno fornito la stessa sequenza di risposte. Inoltre, mentre la sequenza di risposte è un'informazione qualitativa, la frequenza di sequenze è espressa come una proporzione variante fra zero ed uno e pertanto può essere calcolata.

Ora, dato che le variabili manifeste sono state scelte perché si suppone che ciascuna esprima in qualche modo una stessa caratteristica sottostante, queste non saranno statisticamente indipendenti ovvero esibiranno tra di loro delle associazioni più o meno forti. Questo significa che le persone che danno una risposta positiva alla domanda i hanno una maggiore probabilità di dare una risposta positiva anche alla domanda j (ammesso che le due domande abbiano la stessa polarità). Si può esprimere formalmente questa associazione con la seguente disequaglianza:

$$p_{ij} - p_i p_j > 0$$

che, in genere, vale per ogni coppia di variabili manifeste:

“La disequaglianza implica che le due domande “vanno insieme” e questa è l'espressione dell'idea che le nostre domande furono scelte insieme come indicatori della caratteristica che vogliamo “realmente” misurare. Nell'algebra dell'analisi della struttura latente l'eccedenza di p_{ij} su $p_i p_j$ è simbolizzata con

$$|ij| = p_{ij} - p_i p_j,$$

forma nota anche come il prodotto in croce di una tabella tetracorica. Tali prodotti in croce si possono formare a vari livelli. Ad esempio, si possono selezionare le persone con reazione positiva

alla domanda k e poi calcolare che relazione vi sia tra questi soggetti e altre due domande. Ciò sarebbe il simbolo

$$l_{ij;kl} = p_{ijk} - p_{ik} p_{jk}$$

Un importante teorema dell'analisi della struttura latente mostra che, muovendo dalle "marginali di ordine zero" p_i e da un certo numero di questi vari prodotti in croce, siamo in condizioni di calcolare tutte le frequenze di sequenze. Questi prodotti in croce e le loro generalizzazioni sono proprio il mezzo principale per calcolare i parametri latenti dai dati noti" (Lazarsfeld, 1954b; tr. it., 1967, p. 471).

A questo punto si possono introdurre le equazioni di base del modello. Una volta prodotte le stime delle frequenze relative delle classi latenti (v^x), per ogni soggetto collocato in una data classe latente x , si suppone che: $p_{ij}^x = p_i^x p_j^x$ (per il principio dell'indipendenza locale). In questo modo è possibile calcolare i valori attesi, applicando le equazioni del modello, ad esempio:

$$p_{ijklm} = \sum_x v^x p_i^x p_j^x p_k^x q_l^x q_m^x$$

Questa equazione in sostanza ci dice che la proporzione di casi con una certa sequenza di risposte (risposta positiva alle variabili i, j, k e negativa alle variabili l ed m) può essere espressa come la somma, su tutte le classi latenti, di un prodotto ottenuto in ciascuna classe latente separatamente. In ogni classe latente si moltiplica la frequenza relativa della classe per tutte le probabilità latenti che appartengono a ciascuna domanda in quella classe latente. Ad esempio, se ci sono cinque domande, possiamo stimare quale valore avrà $p_{234/15}^x$ (i pedici indicano una sequenza dove si sono registrate risposte positive alle domande 2, 3 e 4 e negative alle domande 1 e 5). La stima, per il principio di indipendenza locale, sarà: $p_{234/15} = p_2^x p_3^x p_4^x q_1^x q_5^x$. (il simbolo q_i è usato per indicare la proporzione di risposte negative alla domanda i).

Riassumendo, nei termini di Lazarsfeld:

"L'equazione di sopra esprime le ipotesi generali della teoria della probabilità [...] Essa riassume le due ipotesi fondamentali implicite nel modello. Questo sono, innanzitutto, che, entro una classe latente omogenea x , la frequenza di risposte positive a ciascuna domanda i tende alla probabilità latente della classe p_i^x , e, in secondo luogo, che entro una simile classe omogenea, la risposta a una domanda è indipendente dalla risposta a qualsiasi altra domanda. E' importante rendersi conto che questa indipendenza è proprietà di una classe latente e non significa che esista indipendenza nei dati manifesti dell'intero gruppo. In altre parole, anche se le risposte a due domande possono essere senza relazione in singole classi latenti, sono quasi certamente in relazione nell'intero campione. [...] E' il risultato del fatto che la probabilità di rispondere positivamente a ciascuna domanda varia sostanzialmente da classe a classe; sono le variazioni di classe di queste probabilità rispetto a entrambe le domande che producono la relazione osservata nell'intero campione. [...] La caratteristica latente spiega le interdipendenze fra risposte ed osservazioni. O, con maggiore precisione, queste interdipendenze fra i dati manifesti *definiscono* la caratteristica latente. Si tratta di una classificazione ipotetica che, se venisse elaborata, spiegherebbe, nel senso suddetto, dette relazioni manifeste. Tuttavia a differenza di altre condizioni di ricerca, in cui è possibile chiedere alle persone quanto s'interessino di questioni politiche, questa caratteristica latente non ha, in alcun senso, esistenza tangibile. *E' un costrutto la cui esistenza è inferita dai dati manifesti*" (Lazarsfeld, 1954b; tr. it., 1967, pp. 471-7, corsivi nel testo).

Le equazioni di base del modello hanno due applicazioni: servono a stimare i parametri latenti; una volta stimati i parametri latenti, questi si usano per calcolare i dati attesi dal modello.

I valori attesi dal modello consentono di: 1) valutare la bontà dell'adattamento ai dati; 2) di *assegnare un punteggio* ai tests, risalendo dai parametri latenti ai dati manifesti. (Lazarsfeld, 1954b; tr. it., 1967, pp. 477-9).

Giunti a questo punto Lazarsfeld presenta un esempio di applicazione dell'analisi della struttura latente tratto dalla ricerca *The American Soldier* (Stouffer, a c. di, 1950).

Lazarsfeld prende in considerazione le seguenti quattro domande, assunte come indicatori dello stato d'animo dei soldati al fronte:

- 1) Di solito, come ti pare di sentirti, di buon umore o depresso? (risposta positiva: di solito sono di buon umore);
- 2) Se toccasse a te scegliere: credi che serviresti meglio la Patria come soldato o come addetto a un lavoro sedentario di guerra? (risposta positiva: come soldato);
- 3) Credi che in complesso l'esercito ti dia l'occasione di farti valere? (risposta positiva: "un'ottima occasione" e "una buona occasione");
- 4) In generale, come ti pare che l'esercito sia diretto? (risposta positiva: "E' diretto molto bene" e "E' diretto abbastanza bene") (Lazarsfeld, 1954b; tr. it., 1967, p. 479).

Le risposte positive a queste quattro domande indicano uno stato di morale alto. Lazarsfeld nota che le prime due domande hanno un taglio differente rispetto alle ultime due. Nelle ultime due domande la risposta positiva comprende un'affermazione intermedia oltre ad una estrema, mentre nelle prime due domande solo la posizione estrema è considerata risposta positiva. Senza aggiungere altro Lazarsfeld passa a presentare i risultati dell'Acl, stimando un modello a tre classi latenti (vedi tabella 2).

Tab. 2 La struttura delle classi latenti di quattro domande sul morale

Domande	Risposte positive	Probabilità Latenti		
		Classe latente 1	Classe latente 2	Classe latente 3
domanda 1	Buon umore	0,66	0,14	0,14
domanda 2	Meglio come soldato	0,63	0,18	0,18
domanda 3	Esercito dà occasione	0,86	0,86	0,30
domanda 4	Esercito è ben diretto	0,90	0,90	0,52
Numero soggetti appartenenti alle classi latenti		1155,4	388,1	1116,5

Fonte: Lazarsfeld, 1954b; tr. it., 1967, p. 480.

La tabella 2 presenta le stime delle frequenze delle tre classi latenti (1.155,4 nella prima classe latente: morale alto; 388,1 nella seconda: morale medio e 1.116,5 nella terza: morale basso) e le probabilità latenti dei soggetti appartenenti alle tre classi di fornire una risposta positiva alle varie domande. Possiamo usare queste probabilità latenti alla stessa stregua dei *factor loadings* nell'analisi fattoriale. Queste probabilità consentono, quindi, di stimare quante persone della prima classe latente risponderanno positivamente alla prima domanda (66%) e, leggendo per colonna lungo le tre classi latenti, di interpretare semanticamente le classi latenti. Lazarsfeld interpreta la prima classe latente in termini di morale alto, la seconda in termini di morale medio e la terza in termini di morale basso.

Continuando la lettura della tabella 2, si nota qualcosa di decisamente strano. Infatti, leggendo le probabilità latenti per riga, riscontriamo che, per ogni domanda, due delle tre classi latenti presentano la stessa probabilità. Secondo Lazarsfeld questa è "una conseguenza del calcolo, derivante dal basso numero di domande disponibili per questa analisi" (Lazarsfeld, 1954b; tr. it., 1967, p. 480). In altri termini, le probabilità latenti delle domande 1 e 2, così come quelle delle domande 3 e 4, sono pressoché sovrapponibili. Trattandole come tracce, cioè supponendo che il numero di classi, anziché finito, sia infinito, si ottiene la figura 1 (si noti che, per ragioni grafiche, l'ordine delle classi è rovesciato: la prima classe è quella del morale basso, la seconda è quella del morale medio e la terza è quella del morale alto), Lazarsfeld commenta così questa rappresentazione grafica:

“Vi sono due cose da notare riguardo a queste tracce. In primo luogo, le domande 1 e 2 sono più discriminanti delle domande 3 e 4, nel senso che soltanto coloro col morale alto rispondono positivamente. [...] I soldati che abbiano appena un po' di morale alto ammettono che l'esercito dà occasioni di farsi valere ed è diretto abbastanza bene. Ma ci vuole un morale molto più elevato perché un soldato dica che di solito è su di morale e che serve meglio come soldato che come civile. In secondo luogo, le tracce delle domande 3 e 4 fanno pensare che le rispettive risposte riflettano qualcosa di più che la valutazione del morale. Si noterà che in ogni classe latente vi sono più risposte positive a queste domande che alle prime due. Infatti c'è una probabilità superiore al 50 per cento che i membri della classe col morale più basso, la classe 3, rispondano positivamente alla quarta domanda. Una lettura attenta della terza e della quarta domanda fornisce qualche indicazione su ciò che potrebbe essere questo elemento aggiunto. Mentre i primi due elementi toccano esclusivamente questioni personali del soldato, le ultime due domande richiedono anche valutazioni dell'esercito come istituzione. In altri termini i giudizi su di una certa “realtà” si sovrappongono a manifestazioni soggettive dello stato di animo dei soldati” (Lazarsfeld, 1954b; tr. it., 1967, pp. 481-2).

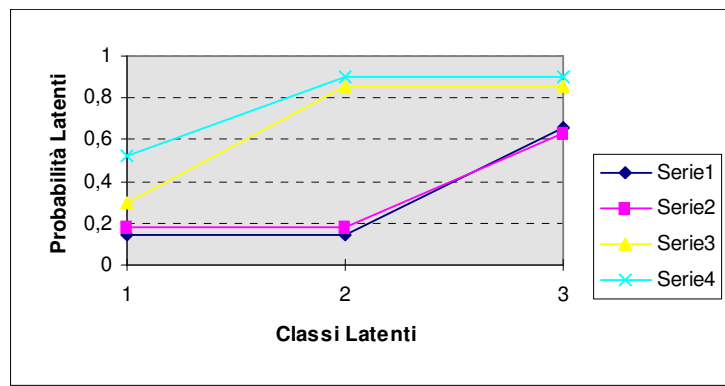


Figura 1 La strutturazione delle probabilità latenti delle tre classi rispetto alle quattro domande nell'esempio di Lazarsfeld

Lazarsfeld prosegue l'esemplificazione del modello presentando le modalità di controllo dell'adattamento del modello ai dati. A questo scopo si calcolano le frequenze delle sequenze attese sulla base dei parametri stimati e poi queste ultime si confrontano con le frequenze delle sequenze empiriche. Se le differenze non sono significative (in senso statistico, sulla base di un test di significatività come il chi quadro, ad esempio) il modello si adatta ai dati; se le differenze sono statisticamente significative allora il modello deve essere modificato. Per il calcolo delle frequenze attese si segue questa procedura:

“Nel fare i calcoli si usa l'equazione $p_{ij}^x = p_i^x p_j^x$. Sulla base dei parametri latenti, vediamo quanti nella classe 1 fornirebbero la sequenza + — + — (positiva sulle domande 1 e 3, negativa sulle domande 2 e 4). Elaboreremo la nostra risposta passo a passo. Secondo i nostri calcoli vi sono 1.155 soldati in questa classe latente. Essi hanno una probabilità 0,66 di rispondere positivamente alla domanda 1. Il risultato è $v^1 p_1^1 = 762$ soldati. E' probabile che un certo numero di questi risponda negativamente alla seconda domanda; si sa inoltre che la probabilità di tale reazione negativa è data da $q_2^1 = 1 - p_2^1 = 0,37$. Ora, ricorrendo all'ipotesi che entro ogni classe latente le risposte a differenti domande sono date indipendentemente, è possibile stabilire in via di previsione quanti darebbero risposta positiva alla prima domanda e negativa alla seconda, rivelando, in altri termini, una sequenza del tipo (+ — ...). Il risultato è $(v^1 p_1^1) q_2^1 = 281$. Al fine di trovare quanti di questi uomini rispondano positivamente anche alla terza domanda, dobbiamo moltiplicare 281 per p_3^1 ; così facendo, arriviamo ad un totale di 247 soldati. Infine, moltiplicando quest'ultimo numero per $q_4^1 = 0,10$ troviamo che nella prima classe 25 soldati presentano la sequenza (+ — + —). In altri termini, al fine di determinare quanti uomini risposero con questa particolare sequenza, abbiamo eseguito il prodotto $v^1 p_1^1 q_2^1 p_3^1 q_4^1$ percorrendo dall'alto in basso la

prima colonna della tabella 1. Si perviene così ad una regola elementare, derivata dalla suddetta equazione: per calcolare la frequenza di una determinata sequenza di risposte in una data classe latente, si moltiplicano le corrispondenti probabilità p e q e la frequenza di classe, disponendo i fattori in una colonna simile a quelle della tavola 1. Se si vuole trovare quante persone dell'intero campione è prevedibile che abbiano data una particolare sequenza si addizionano queste sequenze per tutte le classi latenti, come indica l'equazione $p_{234/15} = \sum_x v^x p_2^x p_3^x p_4^x q_1^x q_5^x$ (Lazarsfeld, 1954b; tr. it., 1967, pp. 482-3, corsivi nel testo).

Nella tabella 3 sono riportate tutte le 16 frequenze attese delle sequenze per le tre classi latenti. La prima colonna indica le 16 sequenze di risposte; le successive tre colonne riportano il numero stimato di soldati di una determinata classe latente che ha fornito una data sequenza di risposte; la quarta colonna fornisce le frequenze teoriche di ogni sequenza (data dalla somma delle colonne 2, 3 e 4); l'ultima colonna, infine, riporta le frequenze delle sequenze manifeste (ossia quelle rilevate nel campione).

La lettura della tabella 3 permette di valutare quanto il modello adottato corrisponda alle rilevazioni manifeste. Lazarsfeld, a questo punto, commenta:

“Un analogo raffronto degli altri numeri delle due colonne rivela che in generale esiste una stretta corrispondenza fra frequenze teoriche e rilevate. Vi sono alcuni scostamenti, ma in generale sono irrilevanti. Quindi abbiamo un alto grado di sicurezza che il particolare modello scelto per queste quattro domande sul morale, corrisponde effettivamente alla situazione manifesta. Le conclusioni che si traggono da questo modello non presentano ampie variazioni rispetto a quanto fu effettivamente rilevato” (Lazarsfeld, 1954b; tr. it., 1967, p. 484).

Tab. 3 La struttura latente del test sul morale dei soldati

Sequenze di risposte	Frequenze Teoriche			Totale delle frequenze teoriche	Frequenze rilevate
	Classe latente 1	Classe latente 2	Classe latente 3		
++++	376,9	7,8	3,6	388,3	385
+---++	217,2	35,3	16,5	269,0	267
---+++	193,4	46,6	21,9	261,9	252
+++--	40,0	0,8	3,3	44,1	42
++--+	60,9	1,2	11,3	73,4	71
---++	111,5	212,5	99,7	423,7	439
+----	3,7	0,6	47,1	51,4	54
+----	3,4	0,8	62,2	66,4	59
---+-	11,8	22,6	91,0	125,4	123
----+	18,0	34,4	310,5	362,9	353
-----	2,0	3,6	283,6	289,2	286
++---	6,5	0,1	10,4	17,0	25
+--+--	23,1	3,7	15,2	42,0	36
+---+	35,1	5,7	51,6	92,4	98
-++--	20,6	4,9	20,0	45,5	56
-+--+	31,3	7,5	68,6	107,4	114
Totale	1155,4 43,4%	388,1 14,6%	1116,5 42%	2660,0 100,0%	2660

Fonte: Lazarsfeld, 1954b; tr. it., 1967, p. 485

Da questo punto in poi Lazarsfeld si dedica alle procedure di classificazione e di attribuzione di punteggi alle sequenze di risposte. Dato che il modello postula che ogni individuo del campione appartenga a una particolare classe latente, si può stimare a quale classe l'individuo appartiene. Si introduce così il concetto di reclutamento, che permette di stimare quanti individui, fra tutti quelli con una particolare sequenza di risposte, provengono da ciascuna delle diverse classi latenti. Ciò si può stabilire scorrendo le righe della tabella 3. Stando alle frequenze teoriche, ad esempio, si scopre che 269 intervistati rispondono alle

quattro domande sul morale con la sequenza (+ — + +). Di questi, l'81% proviene dalla prima classe, il 13% dalla seconda classe e il 6% dalla terza classe. Calcolando in modo analogo le percentuali di reclutamento per tutte le altre sequenze, è possibile trarre inferenze sulla classe latente da cui proviene un dato individuo. Dai dati manifesti conosciamo la sequenza presentata da quell'individuo; dai parametri latenti veniamo a conoscere la probabilità che un individuo con una data sequenza di risposte ha di appartenere ad una data classe latente. Così abbiamo un mezzo per assegnarlo ad una classe latente o ad un'altra, collocandolo nella classe individuata dalla *maggioranza* degli intervistati che hanno risposto con quella sequenza. Cioè, lo si assegna alla classe *modale*.

E' ovvio che quest'attribuzione di classe degli intervistati ha soltanto un certo grado di probabilità ed è, pertanto, un metodo piuttosto approssimativo per ordinare le sequenze di risposte. E' inevitabile commettere una quantità d'errori più o meno gravi.

Un problema anche più serio sorge a proposito delle sequenze che non sono unimodali. In questo caso si presentano delle difficoltà nell'assegnare le sequenze di risposta ad una classe e assegnandole comunque ad una sola classe si commetterebbero degli errori. Inoltre, la classificazione nelle classi latenti basata sul criterio della moda presenta l'inconveniente di dividere le sequenze in un numero limitato di gruppi.

Per ovviare a questi inconvenienti, Lazarsfeld propone un altro metodo di classificazione che permette classificazioni più sensibili (Lazarsfeld, 1954b; tr. it., 1967, p. 488). Questo metodo consiste nell'attribuire dei valori medi alle classi latenti. Ad esempio, si assegna il valore "più uno" alla prima classe, il valore "zero" alla seconda classe e il valore "meno uno" alla classe 3. Poi, per ciascuna sequenza, si assegna un punteggio, calcolandone il valore medio. Si può interpretare un punteggio come la posizione media in termini di classe latente degli intervistati che presentano una data sequenza. Usando la posizione latente media, possiamo quindi fare distinzioni più sottili e, inoltre, calcolando una misura di dispersione, si possono anche distinguere due sequenze che presentino valori medi simili. (Lazarsfeld, 1954b; tr. it., 1967, p. 489).

Conclusa l'esposizione dell'esempio proposto da Lazarsfeld, chiariamo perché questo esempio non ci convince. Intanto Lazarsfeld non dice che con quattro variabili dicotomiche e con tre classi latenti il sistema di equazioni non è risolvibile perché ci sono 16 equazioni e 15 incognite. Per cui è necessario imporre delle condizioni di riducibilità affinché il sistema di equazioni possa essere risolto. Questo aspetto, come detto, ha costituito per molti anni un serio problema a cui sono state date diverse soluzioni parziali⁵. Ora è molto probabile che, a seconda della tecnica di riduzione usata, si ottengano risultati diversi.

Un altro problema non chiarito da Lazarsfeld in questo esempio è quello del perché ha scelto un modello con tre classi latenti. Sappiamo, infatti, che il numero delle classi può essere considerato un dato manifesto, in quanto deve essere stabilito a priori. Inoltre, considerando i risultati ottenuti nella tabella 2, è molto strano che le probabilità di rispondere positivamente alle prime due domande siano le stesse tanto per i soggetti della seconda quanto per quelli della terza classe, così come è strano che siano identiche le probabilità di rispondere positivamente alle domande tre e quattro per i soggetti che sono nella prima e nella seconda classe latente. Infine, non condivido l'interpretazione della seconda classe latente in termini di "morale medio", dato che questa sembra piuttosto costituire una classe residuale. Ricordando che le variabili manifeste sono costituite da dicotomie, è piuttosto strano pensare ad un *continuum latente* con categorie ordinate. Ritengo più plausibile ipotizzarlo come costituito da una classe di soldati ottimisti (la prima), una classe di soldati pessimisti (la terza) e una classe residuale o tutt'al più neutra rispetto alle altre due.

⁵ Si vedano, senza considerare i tanti articoli apparsi nel decennio '50-'60 sulla rivista *Psychometrika*, tra gli altri, Lazarsfeld, 1950a, 1950b, 1954a, 1955b; Capecchi, 1964, 1965; Lazarsfeld ed Henry, 1968.

Per tentare di chiarire questi dubbi ho replicato l'analisi⁶ con il programma MLLSA (*Maximum likelihood latent structure analysis*, Clogg, 1977). In primo luogo ho valutato la possibilità di adattare un modello a due classi latenti, come sembra logico fare a partire dalle probabilità latenti presenti nella tabella 2. Nella tabella 4 si riportano i risultati dell'adattamento dei modelli d'indipendenza a due e a tre classi latenti.

Tab. 4 Analisi esplorativa dei dati dell'esempio di Lazarsfeld

Modello	L^2	χ^2	Gradi di libertà	Significatività 0,05
Indipendenza	611,430	757,099	11	Respinto
due classi	42,352	43,366	6	Respinto
tre classi	4,509	4,603	1	Accettato

Fonte: nostra elaborazione sui dati di Lazarsfeld, 1954b; tr. it., 1967, p. 480.

Tab. 5 Struttura delle classi latenti di quattro domande sul morale

Domande	Risposte positive	Probabilità latenti		
		Classe latente 1	Classe latente 2	Classe latente 3
domanda 1	Buon umore	0,63	0,36	0,17
domanda 2	Meglio come soldato	0,77	0,02	0,19
domanda 3	Esercito dà occasione	0,86	0,90	0,30
domanda 4	Esercito è ben diretto	0,90	1,00	0,54
Numero soggetti appartenenti alle classi latenti		986,7	414,4	1258,9

Fonte: nostra elaborazione sui dati di Lazarsfeld, 1954b; tr. it., 1967, p. 480.

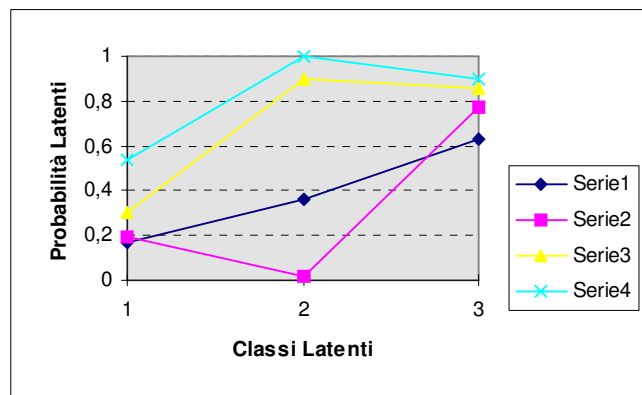


Figura 2 La strutturazione delle probabilità latenti delle tre classi rispetto alle quattro domande nella nostra elaborazione dell'esempio di Lazarsfeld

Il primo dubbio, quindi, si risolve pensando che il modello a due classi fu provato, ma si trovò che non adattava i dati. Usando il programma MLLSA, che produce stime di massima verosimiglianza per i parametri latenti, si ottengono le probabilità latenti presentate nella tabella 5. Come si può notare, ora i parametri latenti sono molto diversi rispetto a quelli ottenuti da Lazarsfeld. Questo si può vedere chiaramente nella figura 2 dove le stime sono trattate come tracce (si noti che per ragioni grafiche, l'ordine delle classi è rovesciato: la

⁶ Questo è possibile sfruttando una positiva caratteristica del trattamento multivariato delle variabili categoriali che permette di replicare le analisi senza disporre della matrice dei dati, ma solo avendo a disposizione le frequenze delle sequenze.

prima classe è quella del morale basso, la seconda è quella del morale neutro e la terza è quella del morale alto). In particolare si nota una maggiore coerenza delle probabilità latenti distribuite nelle tre classi latenti.

Infine nella tabella 6 sono riportati i dati analoghi a quelli della tabella 3 nell'esempio di Lazarsfeld.

Come si può riscontrare confrontando le relative tabelle che presentano i risultati, ci sono significative differenze nei risultati della nostra elaborazione rispetto a quella originaria di Lazarsfeld.

Affronteremo nel paragrafo successivo un altro punto importante trascurato nell'esempio appena riportato: per il momento quello che ci preme sottolineare è che l'uso di una tecnica matematica piuttosto che un'altra non è semplicemente una questione di gusto, ma comporta degli effetti sulle stime prodotte dal modello. Di conseguenza, non si dovrebbe mai dimenticare che: a) un modello matematico è un costrutto artificiale ottenuto sulla base di un insieme di postulati e di vincoli e b) i parametri latenti sono il risultato di operazioni di stima.

Tab. 6 La struttura latente del test sul morale dei soldati

Sequenze di risposte	Frequenze Teoriche			Totale delle frequenze teoriche	Frequenze rilevate
	Classe latente 1	Classe latente 2	Classe latente 3		
++++	371,7	2,2	6,3	380,2	385
+---	109,3	130,7	27,5	267,5	267
----	220,5	4,1	31,2	255,8	252
+++--	40,7	0,0	5,4	46,1	42
++---	59,8	0,2	14,8	74,8	71
---++	64,8	237,3	136,4	438,5	439
+----	1,9	0,0	55,9	57,8	54
-+---	3,9	0,0	63,4	67,3	59
---+-	7,1	0,1	117,9	125,1	123
----+	10,4	25,4	320,8	356,6	353
-----	1,1	0,0	277,3	278,4	286
++---	6,6	0,0	12,8	19,4	25
+--+-	12,0	0,0	23,8	35,8	36
+---+	17,6	14,0	64,7	96,3	98
---+-	24,1	0,0	27,0	51,1	56
-+-+	35,4	0,5	73,4	109,3	114
Totale	986,7	414,4	1258,9	2660,0	2660
	37,1%	15,6%	47,3%	100,0%	

Fonte: nostra elaborazione sui dati di Lazarsfeld, 1954b; tr. it., 1967, p. 485

5. L'analisi delle classi latenti dopo Lazarsfeld

Dopo i contributi di Lazarsfeld, vogliamo esaminare i contributi presenti nella letteratura metodologica e statistica più recente. In questo paragrafo ci varremo dei lavori di Clogg (1981), di McCutcheon (1987) e di Coppi (1998).

Clogg, dopo il dovuto riconoscimento alla primogenitura di Lazarsfeld, afferma:

“L'obiettivo dell'analisi della struttura latente è quello di caratterizzare la variabile latente che spiega l'associazione osservata d'interesse, e questo è ottenuto: 1) stimando la distribuzione delle frequenze relative della variabile latente, 2) stimando le frequenze relative delle variabili osservate per ogni categoria della variabile latente e 3) inferendo da (1) e (2) il significato sostantivo della variabile latente per il problema della ricerca” (Clogg, 1981, p. 216).

Passa poi a considerare gli aspetti di tipo matematico computazionale:

“Ad esempio, il metodo del determinante usato nella stima dei parametri del modello (Lazarsfeld, Henry, 1968) spesso produce stime di probabilità che non sono ammissibili (al di fuori dei valori tra zero ed uno). Il programma LASYS, basato sul metodo del determinante, può per un dato insieme di variabili dicotomiche produrre diversi insiemi di stime di parametri per gli stessi dati, e spesso regole convenzionali basate sulla pratica piuttosto che procedure statistiche rigorose guidano la scelta fra le stime rivali. Metodi efficienti di massima verosimiglianza furono sviluppati per alcuni casi speciali, come furono sviluppati metodi certi per determinare l'identificabilità dei parametri del modello. Ma questi metodi furono di limitata utilità pratica e a nostra conoscenza poco usati” (Clogg, 1981, p. 216).

Infine, presenta gli sviluppi negli algoritmi matematici che hanno reso disponibili i metodi dell'Asl per i ricercatori sociali, attraverso il programma MLLSA (*Maximum likelihood latent structure analysis*) (Clogg, 1977). Questi contributi matematici sono dovuti soprattutto ai lavori di Goodman (1974a, 1974b). L'algoritmo proposto per le stime dei parametri è basato su un procedimento iterativo di riduzione proporzionale dei parametri stimati (Goodman, 1974a) e, quando correttamente applicato, evita le difficoltà poste da altre tecniche e da altri algoritmi.

Inoltre il programma MLLSA fornisce una procedura per esaminare l'identificabilità dei parametri del modello. Questo è il punto che avevamo lasciato in sospeso alla fine del paragrafo precedente. Per identificabilità dei parametri del modello si intende che per un sistema d'equazioni esiste un solo insieme di soluzioni; quando il modello non è identificabile vuol dire, quindi, che ci sono diverse (anche moltissime) soluzioni possibili. Questa condizione non è sempre soddisfatta. Lazarsfeld ed Henry (1968) mostrano (usando tecniche diverse da quella di MLLSA) che i modelli con tre classi latenti per quattro variabili dicotomiche non sono identificabili in generale: questo spiega perché nella nostra replica dell'esempio di Lazarsfeld, riportato nel paragrafo precedente, abbiamo ottenuto risultati diversi rispetto all'applicazione originale nella stima dei parametri latenti.

Goodman ha proposto due metodi per la determinazione dell'identificabilità dei parametri. Il primo consiste nel calcolare una nuova matrice contenente le derivate prime del vettore dei parametri stimati. La trasformazione potrebbe essere non singolare se, e solo se, la matrice delle derivate prime è di rango uguale rispetto al numero di componenti non ridondanti dei parametri latenti. Il secondo metodo consiste nel provare l'adattamento del modello usando valori di partenza che sono vicini, ma leggermente diversi, dalle prime stime ottenute.

Goodman (1974a) ha dimostrato come modelli non identificabili possono essere sostituiti da modelli identificabili (Clogg, 1981).

L'algoritmo implementato sul programma MLLSA, basato su un procedimento iterativo di riduzione proporzionale dei parametri stimati, supera completamente il problema dell'inversione delle matrici ed è facilmente programmabile per ottenere una varietà di modelli ristretti (cioè che presentano altri vincoli oltre a quelli indispensabili per definire un modello) o non ristretti (ovvero senza vincoli ulteriori). Il programma produce soluzioni di massima verosimiglianza e garantisce sempre l'ammissibilità dei parametri stimati (Clogg, 1981, p. 221). Per stimare le probabilità attese, è necessario fornire un insieme iniziale di valori di prova (o sperimentali) che devono essere scelti in modo da soddisfare i vincoli del modello (poiché i parametri della struttura latente sono probabilità, essi devono essere non negativi e soddisfare il vincolo secondo il quale la loro somma deve essere uguale a 1). Poi, sostituendo questi valori nelle formule per le stime di massima verosimiglianza, si ottengono i primi valori attesi dei parametri i quali andranno riscaldati proporzionalmente, in più iterazioni, fino ad ottenere le stime di massima verosimiglianza che approssimano in modo ottimale i dati empirici. Una descrizione completa di quest'algoritmo è data da Goodman (1974a).

Passando al contributo di McCutcheon (1987), dobbiamo innanzitutto dire che non ne condividiamo l'impostazione epistemologica eccessivamente causalistica. Questo autore può essere un buon esempio della cosiddetta deriva del linguaggio delle variabili di cui abbiamo parlato in precedenza. Si evince dal seguente brano:

“Poiché noi crediamo che ciascuno degli indicatori osservati è causato da una variabile di interesse non osservata, o latente, noi ci aspettiamo che fra le misure osservate vi sia covariazione, e studiamo le strutture delle interrelazioni fra gli indicatori osservati per comprendere e caratterizzare la variabile latente sottostante. La premessa base dello studio delle variabili latenti è che la covariazione attualmente osservata fra le variabili osservate è dovuta alla relazione di ogni variabile manifesta con la variabile latente — che la variabile latente “spiega” le relazioni tra le variabili osservate. Se una variabile come questa esiste, e può essere caratterizzata, allora controllando per questa variabile latente risulterà in diminuzione la covariazione tra tutte le variabili osservate per il livello di mutazione della covariazione. Conseguentemente, la variabile latente è detta essere la “vera” fonte delle covariazioni originariamente osservate” (McCutcheon, 1987, p. 5-6).

Mettendo da parte le convinzioni dell'autore circa i rapporti causali tra variabili latenti e variabili manifeste, ne riprendiamo alcune considerazioni rispetto al rapporto tra l'Acl e l'analisi fattoriale. In primo luogo McCutcheon sottolinea che l'Acl, essendo una procedura di analisi per dati categoriali, non implica gli assunti spesso violati della multinormalità delle distribuzioni e della natura continua delle proprietà. Poi prosegue indicando la possibilità di usare l'Acl nella doppia funzione esplorativa-confermativa così come avviene per l'analisi fattoriale. Con la prima funzione i ricercatori possono esplorare un insieme di variabili categoriali manifeste per individuare delle strutture latenti; con la seconda, i ricercatori possono controllare empiricamente se delle ipotesi circa una data struttura latente sono compatibili con i dati.

Sostanzialmente, McCutcheon propone l'Acl come strumento utile per l'analisi delle tipologie quando si lavora con variabili categoriali. Dato il significativo ruolo giocato dalle tipologie nella teoria e nella ricerca delle scienze sociali nonché le difficoltà incontrate con altre tecniche di analisi dei dati impiegate per identificare e per controllare tipologie latenti, l'Acl costituisce una tecnica estremamente importante per le scienze sociali (McCutcheon, 1987, p. 7-8). E' interessante sfruttare questo strumento in funzione esplorativa soprattutto per confrontare una tipologia sia in ricerche *cross sectional* (ossia su campioni diversi nello stesso tempo) sia in ricerche longitudinali (un campione in diversi momenti nel tempo).

Anche questo autore si sofferma sulla presentazione del programma MLLSA e dell'algoritmo che produce le stime di massima verosimiglianza, affermando che gli stimatori di Goodman forniscono un decisivo passo avanti rispetto ai precedenti approcci per la stima

dei parametri latenti. Nonostante ciò, McCutcheon segnala alcuni inconvenienti di questa tecnica di stima dei parametri:

- a) ci può essere più di una soluzione per le equazioni, ossia potrebbe esistere più di un insieme di parametri latenti per ogni dato numero t di classi latenti. In altre parole, le stime di massima verosimiglianza possono rappresentare un massimo locale anziché globale⁷. Per ovviare a questo inconveniente il ricercatore deve provare più di un singolo insieme di valori iniziali delle stime. Come ha dimostrato Goodman, con la tecnica delle stime di massima verosimiglianza, da diversi insiemi di valori iniziali spesso si ottengono le stesse stime finali;
- b) il secondo punto è che il numero dei parametri stimabili è limitato dai gradi di libertà disponibili nella tabella multidimensionale delle variabili manifeste. Così, solamente quando c'è un numero positivo di gradi di libertà il modello può essere contemporaneamente stimato e controllato;
- c) infine, quando una molteplicità di valori di stime dei parametri può essere associato con una data soluzione (problema dell'identificazione), il modello non è identificato (Goodman, 1974a). Una condizione necessaria e sufficiente per determinare l'identificabilità locale del modello delle classi latenti è fornita da Goodman (1974a). Modelli non identificati possono diventare identificabili imponendo delle restrizioni ad uno o più parametri (McCutcheon, 1987, p. 25-7).

Infine, lo statistico Coppi (1998) presenta l'Acl, senza fare il minimo riferimento ai contributi di Lazarsfeld, come un caso del modello lineare generalizzato. Nella sua trattazione, molto formalizzata, considera il modello dell'Acl come un modello statistico parametrico con $kp+k-1$ parametri (con k = classi latenti e p = variabili manifeste) articolato nelle seguenti fasi:

- 1) stima dei parametri del modello;
- 2) valutazione dell'attendibilità di tali stime;
- 3) verifica della bontà dell'adattamento del modello;
- 4) assegnazione degli individui alle classi latenti;
- 5) interpretazione delle classi latenti.

Anche Coppi si sofferma sulle stime di massima verosimiglianza, ottenibili con procedimento a più iterazioni, e sul test del rapporto delle massime verosimiglianze per la valutazione della bontà dell'adattamento del modello ai dati. In particolare l'autore dimostra che il procedimento di stima converge agli effettivi valori teorici dei parametri oppure a dei massimi relativi della funzione di massima verosimiglianza. A causa di quest'ultima possibilità è opportuno provare più volte il procedimento a partire da diversi insiemi di valori iniziali (Coppi, 1998, pp. 279-301).

Possiamo quindi concludere che nei più recenti contributi in tema dell'Acl vi sia un consenso nel ritenere le stime di massima verosimiglianza come la tecnica più affidabile per la soluzione del sistema d'equazioni per la stima dei parametri latenti del modello.

⁷ Nell'analisi della struttura latente, nei modelli di equazioni strutturali e anche nelle reti neurali, ovvero in tutti quei modelli e/o tecniche che fanno uso di quella che potremmo definire la matematica della complessità, quello che realmente conta non è la stima dei singoli parametri ma l'insieme (o la struttura) dei parametri stimati. Ciò significa che possono esistere più insiemi di parametri che riproducono in modo soddisfacente i dati empirici (ovvero l'adattamento è sempre relativo e mai assoluto). Ecco che interviene il problema dei minimi locali (per le reti neurali, cfr. Di Franco, 1998) o dei massimi locali (per l'analisi della struttura latente).

6. Un'applicazione dell'analisi delle classi latenti ai dati di una recente ricerca

Dopo aver passato in rassegna i più recenti contributi in tema di Acl, ci proponiamo di effettuare delle semplici applicazioni concrete di questo strumento per valutarne l'efficacia. In questa sede più che su problemi di assegnazione degli individui alle classi latenti, ci soffermeremo sull'analisi delle relazioni tra la variabile latente e altre variabili sociologiche, seguendo l'impostazione di McCutcheon (1987). Presenteremo comunque i risultati dell'analisi per lasciare al lettore la possibilità di valutarli.

Applichiamo l'Acl ad alcuni dati di una ricerca ancora inedita sugli stili di vita e sui valori di un campione di italiani. Sono state effettuate 1.608 interviste ad un campione di italiani stratificato per sesso, età, titolo di studio e residenza geografica. Le interviste sono state effettuate da un gruppo di intervistatori con un questionario semistandardizzato il quale, tra le altre, conteneva una batteria di quattordici domande che offrivano agli intervistati la possibilità di una scelta secca tra due frasi. Di seguito si elencano le quattro domande che abbiamo scelto per l'Acl e le percentuali di risposta per ciascuna frase.

D1)

- a) La vera famiglia è quella sancita dal matrimonio e dalla nascita di figli (53,2%);
- b) L'unione tra due persone deve fondarsi sulla libera scelta di entrambi i partner (46,8%).

D2)

- a) Quando in un matrimonio l'amore finisce, o ci sono altri gravi problemi, è giusto che i coniugi divorzino, anche se ci sono i figli (52,9%);
- b) Quando ci sono dei figli, l'unione matrimoniale deve resistere e superare qualsiasi problema (47,1%).

D3)

- a) Per la donna la famiglia è più importante di qualsiasi lavoro (86,8%);
- b) Per la donna il lavoro viene prima della famiglia (13,2%)

D4)

- a) E' giusto che l'uomo, nel mondo del lavoro, occupi delle posizioni di maggiore responsabilità rispetto alla donna (32%);
- b) Le donne hanno dimostrato in molti campi di essere migliori degli uomini; esse dovrebbero avere più responsabilità nel mondo del lavoro ed anche in politica (68%).

Le quattro coppie di frasi esprimono opinioni diverse sulla famiglia, sul divorzio, su cosa sia più importante per le donne e sul ruolo della donna nel mondo del lavoro. Ripetiamo che gli intervistati dovevano, necessariamente, effettuare una scelta secca per ogni coppia di frasi. Prese complessivamente, queste quattro coppie di frasi potrebbero essere considerate indicatori di una visione del mondo tradizionale (famiglia sancita dal matrimonio e dalla nascita dei figli, divorzio non ammesso, donne che devono preferire la famiglia al lavoro e comunque sottomesse agli uomini nel mondo del lavoro) opposta ad una visione del mondo moderna (famiglia fondata su una libera scelta dei partner, approvazione del divorzio, donne che preferiscono il lavoro alla famiglia e che dovrebbero assumere maggiori responsabilità nel mondo del lavoro nonché in politica).

Nella tabella 7 si riporta l'incrocio tra le quattro variabili prese in considerazione

Tab. 7 L'incrocio tra le quattro variabili selezionate per l'AcI

D1 La vera famiglia è:	D2 Divorzio	D3 Cosa è importante per la donna	D4 Prevalenza uomo donna nel lavoro	
			Uomo	Donna
Matrimonio + figli	Sì	Famiglia	93	171
Matrimonio + figli	Sì	Lavoro	4	20
Matrimonio + figli	No	Famiglia	274	257
Matrimonio + figli	No	Lavoro	13	22
Unione libera	Sì	Famiglia	73	371
Unione libera	Sì	Lavoro	17	104
Unione libera	No	Famiglia	32	126
Unione libera	No	Lavoro	6	25

Muovendoci in un'ottica prettamente esplorativa, cominciamo a valutare il modello di indipendenza (che equivale ad ipotizzare una sola classe latente), il modello a due classi latenti e il modello a tre classi latenti (vedi tabella 8).

Tab. 8 Analisi esplorativa dei dati sul pubblico della televisione

Modello	L^2	χ^2	Gradi di libertà	Significatività 0,05
Indipendenza	530,328	616,640	11	respinto
due classi	7,200	6,849	6	accettato
tre classi	0,923	0,916	1	accettato

Il modello d'indipendenza, che muove dall'assunto secondo il quale fra le quattro variabili dicotomiche vi è indipendenza a livello manifesto, deve essere respinto, come poteva facilmente evincersi dalla semplice lettura della tabella 7. Gli altri due modelli, a due e tre classi, risultano entrambi corroborati dal test del chi quadro del rapporto di massima verosimiglianza. Nelle tabelle 9 e 10 si riporta la stima dei parametri latenti per i due modelli.

Tab. 9 Le probabilità latenti stimate per il modello a due classi

		Classe Latente 1	Classe latente 2
Famiglia	Matrimonio figli	.9518	.1864
	Unione libera	.0482	.8136
Divorzio	Sì	.2346	.7730
	No	.7654	.2270
Prev. nel lavoro	Uomo	.9558	.7975
	Donna	.0442	.2025
Importante per donna	Famiglia	.5244	.1496
	Lavoro	.4756	.8504
Stima delle probabilità delle classi latenti		.4504	.5496

Tab. 10 Le probabilità latenti stimate per il modello a tre classi

		Classe Latente 1	Classe Latente 2	Classe Latente 3
Famiglia	Matr. Figli	.7001	.0027	.9440
	Unione libera	.2999	.9973	.0560
Divorzio	Sì	.5084	.8325	.2356
	No	.4916	.1675	.7644
Prev. nel lavoro	Uomo	.8821	.7703	.9600
	Donna	.1179	.2297	.0400
Importante per donna	Famiglia	.0762	.1640	.6692
	Lavoro	.9238	.8360	.3308
Stima delle probabilità delle classi latenti		.2796	.3662	.3542

Nel primo caso, nel modello a due classi (vedi tabella 9), si individua una prima classe, costituita dal 45% dei casi, di persone tradizionaliste. Si può notare che le domande più discriminanti in questo caso sono quella sulla famiglia e quella sulla prevalenza nel mondo del lavoro da parte dell'uomo (in entrambi i casi, per un soggetto collocato nella prima classe latente, la probabilità latente di rispondere positivamente a queste domande è del 95%). Al contrario, sempre per questa prima classe, la domanda meno discriminante è quella sul considerare o meno la famiglia come la cosa più importante per la donna (probabilità latente del 52%). La seconda classe latente, quella dei modernisti, raccoglie il 55% del campione. In questo caso la domanda più discriminante è quella che valuta il lavoro come la cosa più importante per la donna (probabilità latente dell'85%).

Il modello a tre classi latenti, vedi tabella 10, individua una prima classe (28% dei casi) che potremmo definire familisti-materialisti (importanza del lavoro per la donna, prevalenza nel lavoro dei maschi, famiglia tradizionale e indifferenza verso il divorzio); una seconda (37% dei casi) che rappresenta i moderni e una terza (35%) che raccoglie i tradizionalisti in maniera più nitida rispetto al precedente modello.

Tab. 11 La struttura latente del modello a due classi

Sequenze di risposte	Frequenze teoriche		Totale Frequenze Teoriche	Freq. rilevate
	Classe latente 1	Classe latente 2		
++++	81,05	15,18	96,23	93
—+++	4,10	66,30	70,40	73
+—++	264,47	4,46	268,93	274
— — ++	13,40	19,48	32,88	32
++—+	3,75	3,85	7,60	4
—+—+	0,19	16,84	17,03	17
+— — +	12,23	1,13	13,36	13
— — — +	0,62	4,94	5,56	6
+++—	73,50	86,33	159,83	171
—+ + —	3,72	376,95	380,67	371
+—+—	239,84	25,35	265,19	257
— — + —	12,16	110,70	122,86	126
++ — —	3,40	21,93	25,33	20
—+ — —	0,17	95,74	95,91	104
+ — — —	11,09	6,44	17,53	22
— — — —	0,56	28,12	28,68	25
Totale	724,25	883,74	1608	1608
	45%	55%	100,0%	

Nelle tabelle 11 e 12 sono riportate le strutture latenti dei due modelli.

Tab. 12 La struttura latente del modello a tre classi

Seq. di risposte	Frequenze teoriche			Totale Frequenze Teoriche	Frequenze rilevate
	Classe latente 1	Classe latente 2	Classe latente 3		
++++	10,75	0,17	81,38	92,33	93
—+++	4,61	61,77	4,83	71,19	73
+—++	10,40	0,03	264,04	274,49	274
— — ++	4,45	12,43	15,66	32,53	32
++—+	1,44	0,05	3,39	4,88	4
—+—+	0,62	18,42	0,20	19,23	17
+— — +	1,39	0,01	11,00	12,39	13
— — — +	0,60	3,71	0,65	4,95	6
+++—	130,38	0,85	40,23	171,46	171
—+ +—	55,85	314,86	2,39	373,12	371
+—+—	126,07	0,17	130,52	256,76	257
— — +—	54,00	63,35	7,74	125,11	126
+ + — —	17,43	0,25	1,68	19,35	20
—+ — —	7,46	93,89	0,10	101,44	104
+ — — —	16,85	0,05	5,44	22,33	22
— — — —	7,22	18,89	0,32	26,43	25
Totale	449,52 28%	588,90 37%	569,58 35%	1608,00 100,0%	1608

Considerando gli scopi illustrativi di questo esempio, non commentiamo le precedenti due tabelle. Si noti che sono state predisposte nella stessa forma della tabella 3 del secondo paragrafo, per cui valgono le stesse note per la loro lettura.

Infine, sulla base delle probabilità latenti modali, si possono classificare i casi che hanno fornito una data sequenza di risposte. Anche in questo caso non ci interessiamo particolarmente dei risultati di questa procedura. Valgono comunque le avvertenze indicate da Lazarsfeld (vedi paragrafo 4) su questo punto. Nella tabella 13 si presenta la classificazione dei casi solo per il modello a due classi latenti.

Tab. 13 Classificazione dei casi alle classe latenti per il modello a due classi con il criterio della probabilità modale

Sequenze di risposte	Frequenze Teoriche	Frequenze rilevate	Classe assegnata	Probabilità modale
++++	96,23	93	1	.8422
—+++	70,40	73	2	.9416
+—++	268,93	274	1	.9834
— — ++	32,88	32	2	.5922
++—+	7,60	4	2	.5072
—+—+	17,03	17	2	.9888
+— — +	13,36	13	1	.9152
— — — +	5,56	6	2	.8886
+++—	159,83	171	2	.5402
—+ +—	380,67	371	2	.9902
+—+—	265,19	257	1	.9044
— — +—	122,86	126	2	.9010
+ + — —	25,33	20	2	.8658
—+ — —	95,91	104	2	.9982
+ — — —	17,53	22	1	.6326
— — — —	28,68	25	2	.9804
Totale	1608	1608		

E' più interessante mostrare una diversa applicazione dell'Acl. Assumendo come valido il modello a due classi latenti (quello che individua la prima classe di tradizionalisti e la seconda di modernisti), ci chiediamo ora se queste due strutture latenti varino rispetto ad altre variabili sociologiche. Per questo secondo esempio prendiamo in considerazione il genere, il titolo di studio (dicotomizzato in "fino alla licenza media" e "diplomati + laureati") e l'età (dicotomizzata in "18-49 anni" e "50 anni e oltre") degli intervistati. Nelle tabb. 14, 15 e 16 sono presentati i risultati delle stime dei parametri latenti rispetto alle suddette variabili stratificatrici.

Tab. 14 Le probabilità latenti stimate per il modello a due classi stratificato per il genere dei rispondenti

		Classe 1 Maschi	Classe 2 Maschi	Classe 1 Femm.	Classe 2 Femm.
Famiglia	Matr. Figli	1.0000	.1671	.9344	.1883
	Unione libera	.0000	.8329	.0656	.8117
Divorzio	Sì	.2372	.6819	.2447	.8448
	No	.7628	.3181	.7553	.1552
Prev. Nel lavoro	Uomo	.9572	.8099	.9556	.7860
	Donna	.0428	.1901	.0444	.2140
Imp. per donna	Famiglia	.5691	.2491	.4716	.0675
	Lavoro	.4309	.7509	.5284	.9325
Genere	Maschio	1.0000	1.0000	.0000	.0000
	Femmina	.0000	.0000	1.0000	1.0000
Stima prob. delle classi latenti		.2242	.2688	.2325	.2830

Tab. 15 Le probabilità latenti stimate per il modello a due classi stratificato per il titolo di studio dei rispondenti

		Classe 1 basso	Classe 2 basso	Classe 1 alto	Classe 2 alto
Famiglia	Matr. Figli	.9619	.2432	.8661	.0059
	Unione libera	.0381	.7568	.1339	.9941
Divorzio	Sì	.2050	.7633	.3971	.8034
	No	.7950	.2367	.6029	.1966
Prev. nel lavoro	Uomo	.9636	.8289	.8860	.7294
	Donna	.0364	.1711	.1140	.2706
Imp. per donna	Famiglia	.5774	.1834	.2507	.0865
	Lavoro	.4226	.8166	.7493	.9135
Tit. studio	Elem. media	1.0000	1.0000	.0000	.0000
	Dipl. laurea	.0000	.0000	1.0000	1.0000
Stima prob. delle classi latenti		.3613	.3918	.1010	.1459

Tab. 16 Le probabilità latenti stimate per il modello a due classi stratificato per l'età dei rispondenti

		Classe 1 18-49	Classe 2 18-49	Classe 1 >50	Classe 2 >50
Famiglia	Matr. figli	.8650	.1309	.9891	.2914
	Unione libera	.1350	.8691	.0109	.7086
Divorzio	Sì	.3348	.8184	.1833	.6801
	No	.6652	.1816	.8167	.3199
Prev. nel lavoro	Uomo	.9408	.7716	.9620	.8458
	Donna	.0592	.2284	.0380	.1542
Imp. per donna	Famiglia	.4356	.1454	.5798	.1547
	Lavoro	.5644	.8546	.4202	.8453
Età	50 anni e oltre	.0000	.0000	1.0000	1.0000
	18-49 anni	1.0000	1.0000	.0000	.0000
Stima prob. delle classi latenti		.1988	.3540	.2616	.1856

Considerando il campione nel suo complesso, abbiamo stimato che il 45% dei soggetti sono tradizionalisti a fronte del 55% di modernisti.

Rispetto al genere non si riscontrano differenze nella struttura latente (tra i maschi il 44,5% è stimato tradizionalista e il 55,5% modernista; tra le femmine rispettivamente il 45,1% e il 54,9%). L'unica differenza riguarda la domanda sulla maggiore importanza tra la famiglia e il lavoro per la donna (vedi tabella 14).

Le altre due variabili, il titolo di studio e l'età, mostrano invece un impatto più rilevante sulla struttura latente (vedi tabelle 15 e 16). Al crescere del titolo di studio aumenta la frazione di modernisti. Infatti tra i soggetti con titolo di studio basso (fino alla licenza media) le quote di tradizionalisti e di modernisti sono rispettivamente del 48% e del 52% a fronte del 41% e del 59% tra i soggetti forniti di diploma o di laurea.

Tab. 17 La struttura latente del modello a due classi stratificato per il genere dei rispondenti

Seq. di risposte	Classi Latenti				Freq. Teo	Freq. ril	FreqTeo	Freq rile
	1 Maschi	2 Maschi	1 Femmine	2 Femmine				
++++	44,80	9,94	38,52	3,84	54,74	51	42,35	42
—+++	0	49,53	2,70	16,56	49,52	50	19,26	23
+—++	144,09	4,64	118,89	0,71	148,73	153	119,60	121
— — ++	0	23,10	8,35	3,04	23,10	21	11,39	11
++ — +	2,00	2,33	1,79	1,05	4,34	3	2,83	1
— + — +	0	11,63	0,13	4,51	11,62	14	4,63	3
+ — — +	6,44	1,09	5,52	0,19	7,53	7	5,71	6
— — — +	0	5,42	0,39	0,83	5,42	6	1,22	0
+++ —	33,92	29,95	43,16	53,07	63,88	72	96,24	99
— + + —	0	149,30	3,03	228,78	149,31	145	231,80	226
+ — + —	109,10	13,97	133,21	9,75	123,07	117	142,97	140
— — + —	0	69,65	9,35	42,03	69,65	73	51,39	53
++ — —	1,52	7,03	2,01	14,45	8,55	5	16,46	15
— + — —	0	35,04	0,14	62,29	35,04	37	62,43	67
+ — — —	4,88	3,28	6,14	2,65	8,16	11	8,84	11
— — — —	0	16,35	0,43	11,44	16,35	14	11,88	11
Totale	346,75 22%	432,25 27%	373,80 23%	455,20 28%	779,00 49%	779	829,00 51%	829

Considerando l'età, che è indubbiamente connessa con il titolo di studio, registriamo un rapporto inverso. Tra i soggetti con cinquanta anni e oltre le quote sono del 58,5% di

tradizionalisti e di 41,5% di modernisti; tra i soggetti da 18 fino a 49 anni i tradizionalisti scendono al 36%, di converso i modernisti salgono al 64%.

Si potrebbe continuare l'esplorazione delle relazioni tra queste variabili combinando, ad esempio, il titolo di studio e l'età in una tipologia e poi stimarne la struttura latente.

Con questo semplice esempio, comunque, riteniamo di aver dimostrato l'utilità dell'applicazione dell'AcI nelle operazioni di analisi dei dati categoriali nella ricerca sociale.

Infine, nelle tabelle 17, 18 e 19 si riportano le strutture latenti dei tre modelli stratificati.

Tab. 18 La struttura latente del modello a due classi stratificato per il titolo di studio dei rispondenti

Seq. Di risposte	Classi Latenti				Freq. Teo	Fre rile	Fre Teo	Fre rile
	1 Basso	2 Basso	1 Alto	2 Alto				
++++	63,74	17,78	12,40	0,07	81.51	80	12.47	13
—+++	2,52	55,33	1,92	11,82	57.85	60	13.74	13
+—++	247,17	5,51	18,83	0,02	252.67	255	18.85	19
— — ++	9,79	17,16	2,91	2,89	26.96	26	5.80	6
++—+	2,41	3,67	1,60	0,03	6.08	3	1.62	1
—+—+	0,10	11,42	0,25	4,39	11.52	13	4.63	4
+— — +	9,34	1,14	2,42	0,01	10.49	11	2.43	2
— — — +	0,37	3,54	0,37	1,07	3.91	3	1.45	3
+++—	46,65	79,14	37,07	0,74	125.80	133	37.81	38
—+ — —	1,85	246,36	5,73	124,86	248.19	238	130.61	133
+ — + —	180,90	24,55	56,28	0,18	205.42	201	56.46	56
— — + —	7,17	76,40	8,70	30,56	83.58	89	39.25	37
+ + — —	1,76	16,34	4,77	0,27	18.11	15	5.05	5
— + — —	0,07	50,85	0,74	46,32	50.93	58	47.06	46
+ — — —	6,85	5,07	7,24	0,07	11.91	14	7.31	8
— — — —	0,27	15,77	1,12	11,34	16.05	12	12.45	13
Totale	580,93 36%	630,07 39%	162,36 10%	234,64 15%	1211,00 75%	1211	397,00 25%	397

Tab. 19 La struttura latente del modello a due classi stratificato per l'età dei rispondenti

Seq. di risposte	Classi Latenti				Freq. Teo	Fre ril	Fre Teo	Fre ril
	1 ">50"	2 ">50"	1 "18-49"	2 "18-49"				
++++	42,53	7,74	37,93	6,84	50.26	49	44.77	44
—+++	0,47	18,82	5,92	45,44	19.28	19	51.35	54
+—++	189,51	3,64	75,36	1,52	193.15	195	76.88	79
— — ++	2,09	8,85	11,76	10,08	10.93	11	21.84	21
++—+	1,68	1,41	2,39	2,03	3.09	2	4.41	2
—+—+	0,02	3,43	0,37	13,45	3.45	5	13.82	12
+— — +	7,49	0,66	4,74	0,45	8.16	8	5.20	5
— — — +	0,08	1,61	0,74	2,98	1.70	1	3.73	5
+++—	30,83	42,28	49,15	40,22	73.11	78	89.37	93
—+ + —	0,34	102,81	7,67	267,06	103.14	100	274.72	271
+ — + —	137,35	19,89	97,65	8,93	157.26	154	106.57	103
— — + —	1,51	48,36	15,24	59,26	49.88	51	74.49	75
+ + — —	1,22	7,71	3,09	11,91	8.93	6	15.00	14
— + — —	0,01	18,74	0,48	79,05	18.75	21	79.54	83
+ — — —	5,43	3,63	6,14	2,64	9.06	11	8.79	11
— — — —	0,06	8,82	0,96	17,54	8.88	8	18.50	17
Totale	420,61 26%	298,39 19%	319,60 20%	569,40 35%	719,00 45%	719	889,00 55%	889

7. Conclusioni

Con l'esempio appena proposto, riteniamo di aver mostrato alcune delle possibili applicazioni dell'Acl ai dati sociologici.

Una ulteriore applicazione dell'Acl potrebbe seguire la proposta di Alberto Marradi (1981) nell'uso dell'analisi fattoriale come strumento utile per la definizione e il raffinamento dei concetti oggetto di studio. In modo analogo alla proposta di Marradi, l'Acl potrebbe essere utilizzata per esplorare il dominio semantico, che un insieme di variabili categoriali permette di definire, in fasi successive di progressivo affinamento, fino a giungere alla costruzione di indici sintetici per ciascuna classe latente che attribuiscono dei punteggi a ciascun individuo su ognuna di queste.

In definitiva, pensiamo che si possa inserire l'Acl nella "cassetta degli attrezzi" dei ricercatori sociali. Sarà cura di questi ultimi applicare l'Acl consapevolmente a problemi sostantivi di ricerca, evitando di esibire questo strumento come vessillo di "scientificità".

In generale, dobbiamo a Lazarsfeld l'aver introdotto nella ricerca sociale alcuni strumenti di analisi dei dati (sia a livello mono- che bi- e multi-variato) ed in particolare la logica dell'analisi multivariata, l'Asl e le tecniche nello studio di *panels*, appositamente costruiti sulle esigenze di quest'ultima e che rispettano la natura della variabili usate.

Altri meriti di Lazarsfeld sono l'aver lasciato un patrimonio di ricerche empiriche che sono ancor oggi fonte preziosa di insegnamenti per tutti ricercatori sociali, soprattutto per l'insegnamento fondamentale di puntare con determinazione all'obiettivo usando tutto il buon senso di cui si dispone. Infine, tra i meriti maggiori di Lazarsfeld c'è l'aver posto in luce la ricchezza di possibilità che si annidano dietro le strutture e i procedimenti più semplici (ad esempio nelle operazioni di tipologizzazione, nell'elaborazione di tre variabili, etc.), cercando di codificare e formalizzare i procedimenti della ricerca empirica in un modo sempre finalizzato alla funzione didattica e di trasmissione delle competenze (ad esempio con il *Planning Project for Advanced Training in Social Research* della Columbia University; cfr. Lazarsfeld, 1955d, p. 13).

Se poi, per alcuni aspetti o in alcuni passaggi della sua produzione, ha peccato di scarso rigore terminologico o di non sufficiente chiarezza concettuale, o se, in altri momenti, si è lasciato trasportare troppo dall'entusiasmo nel convincimento di riuscire ad elevare la sociologia empirica al rango di scienza matura, questi peccati non possono modificare il peso e l'importanza della figura di Lazarsfeld nella metodologia della ricerca sociale.

Riferimenti bibliografici

- H. M., Blalock jr., 1961, *Causal inferences in nonexperimental research*, Chapel Hill, The University of North Carolina Press; tr. it. *L'analisi causale in sociologia*, Padova, Marsilio, 1967.
- R., Boudon, 1965, "A Method of Linear Causal Analysis", in *American Sociological Review*, vol. 30, pp. 365-374.
- V., Capecchi, 1964, "I modelli di classificazione e l'analisi della struttura latente", in *Quaderni di Sociologia*, vol. XIII, pp. 289-340.
- V., Capecchi, 1965, "Analisi della struttura latente e analisi dei fattori", in *Quaderni di Sociologia*, vol. XIV, pp. 33-68.
- V., Capecchi, 1967, "Metodologia e ricerca nell'opera di Paul F. Lazarsfeld", in Lazarsfeld, 1967, pp. VII-CXCVI.
- M., Cardano e R., Miceli, (a c. di), 1991, *Il linguaggio delle variabili. Strumenti per la ricerca sociale*, Torino, Rosenberg & Sellier.
- C. C., Clogg, 1977, *Unrestricted and restricted maximum likelihood latent structure analysis: a manual for users*, University of Park, PA, Working Paper 1977-09.
- C. C., Clogg, 1981, "New developments in latent structure analysis", in Jackson e Borgatta, (a c. di), 1981, pp. 215-246.
- R., Coppi, 1998, *Lezioni di analisi statistica multivariata*, Roma, Dipartimento di Statistica, Probabilità e Statistiche applicate, Università degli studi di Roma "La Sapienza".
- G., Di Franco, 1997, *Tecniche e modelli di analisi multivariata dei dati. Introduzione all'applicazione per la ricerca sociale*, Roma, Seam.
- G., Di Franco, 1998, "Reti neurali artificiali e analisi dei dati per la ricerca sociale: un nuovo paradigma?", in *Sociologia e Ricerca Sociale*, n. 56, pp. 35-75.
- G.J., Di Renzo, (a c. di), 1966, *Concepts, Theory and Explanation in the Behavioral Sciences*, New York, Random House.
- O. D., Duncan, 1966, "Path Analysis: sociological examples", in *American Journal of Sociology*, vol. 72, pp. 1-16.
- O. D., Duncan, 1975, *Introduction to structural equation models*, New York, Academic Press.
- A., Einstein, 1966, *The Quotable Einstein*, Gerusalemme, The Hebrew University of Jerusalem and Princeton University Press; tr. it., *Pensieri di un uomo curioso*, Milano, Mondadori, 1997.
- L. A., Goodman, 1974a, "Exploratory latent structure analysis using both identifiable and unidentifiable models", in *Biometrika*, n.61, pp. 215-231.
- L. A., Goodman, 1974b, "The analysis of systems of qualitative variables when some of the variables are unobservable. Part I-A: Modified latent structure approach", in *American Journal of Sociology*, n. 79, pp. 1179-1259.
- H. H., Hyman, 1955, *Survey design and analysis. Principles, cases and procedures*, Glencoe, The Free Press; tr. it., *Disegno della ricerca e analisi sociologica*, Padova, Marsilio, 1967.
- D. J., Jackson, E. F., Borgatta (a c. di), 1981, *Factor Analysis and Measurement in Sociological Research. A Multi-dimensional Perspective*, Beverly Hills, Sage.
- K. G., Jöreskog, 1963, *Statistical estimation in factor analysis: a new technique and its foundation*, Stoccolma, Almqvist and Wiksell.
- K. G., Jöreskog, 1966, "Testing a simple structure hypothesis in factor analysis", in *Psychometrika*, n. 31, pp. 165-178.
- K. G., Jöreskog, K. G., 1967, "Some contributions to maximum likelihood factor analysis", in *Psychometrika*, n. 32, pp. 443-482.

- K. G., Jöreskog, D. N., Lawley, 1968, "New methods in maximum likelihood factor analysis", in *British Journal of Mathematical and Statistical Psychology*, n. 21, pp. 85-96.
- K. G., Jöreskog, 1969, "A general approach to confirmatory maximum likelihood factor analysis", in *Psychometrika*, n. 34, pp. 183-202.
- K. G., Jöreskog, 1970, "A general method for analysis of covariance structure", in *Biometrika*, n. 57, pp. 239-251.
- K. G., Jöreskog, M., van Thillo, 1973, *Lisrel: a general computer program for estimating a linear structural equation system involving multiple indicators of unmeasured variables*, Research Report 73-5, Uppsala, Uppsala University, Dept. of Statistics.
- K. G., Jöreskog, 1976, *Analyzing psychological data by structural analysis of covariance matrices*, Research Report 76-9, Uppsala, University of Uppsala.
- K. G., Jöreskog, D., Sörbom, 1979, *Advances in Factor Analysis and Structural Equations Models*, Cambridge, MA. Abt Books.
- P. F., Lazarsfeld, 1950a, "The logical and mathematical foundation of latent structure analysis", in Stouffer et al. 1950, pp. 362-412.
- P. F., Lazarsfeld, 1950b, "The interpretation and computation of some latent structures", in Stouffer et al. 1950, pp. 413-472.
- P. F., Lazarsfeld, A. H., Barton, 1951, "Qualitative measurement in the social sciences: classification, typologies and indices", in Lerner e Lasswell (a c. di), 1951, pp.155-192; tr. it. Lazarsfeld, 1967, "Classificazioni, tipologie e indici", pp. 231-306.
- P. F., Lazarsfeld (a c. di), 1954a, *Mathematical Thinking in the Social Science*, Glencoe, Ill., The Free Press.
- P. F., Lazarsfeld, 1954b, "A Conceptual Introduction to Latent Structure Analysis", in Lazarsfeld (a c. di), 1954a, pp. 349-387; tr., it., Lazarsfeld, 1967, "Introduzione al concetto di analisi della struttura latente", pp. 447-524.
- P. F., Lazarsfeld, 1955a, "Recent Developments in Latent Structure Analysis", in *Sociometry*, vol. XVIII, n. 4, pp. 391-403; tr. it. Lazarsfeld, 1967, "Recenti sviluppi nell'analisi della struttura latente", pp. 525-540.
- P. F., Lazarsfeld, 1955b, "Interpretation of statistical relations as a research operation", in Lazarsfeld e Rosenberg (a c. di), 1955c, pp. 115-125; tr. it., Lazarsfeld, 1967, "L'interpretazione delle relazioni statistiche come operazione di ricerca", pp. 393-411.
- P. F., Lazarsfeld, M., Rosenberg (a c. di), 1955c, *The language of social research: a reader in the methodology of social research*, Glencoe, Ill., The Free Press.
- P. F., Lazarsfeld, 1955d, "Introduction", in Hyman, 1955; tr. it., 1967, pp. 13-26.
- P. F., Lazarsfeld, 1958, "Evidence and Inference in Social Research", in *Daedalus*, vol. 87, n. 4, pp. 99-130; tr. it. parziale in Cardano, Miceli (a c. di), 1991, pp. 121-132.
- P. F., Lazarsfeld, 1959, "Problems in methodology", in Merton, Broom e Cottrel (a c. di), 1959, pp. 39-78; tr. it., Lazarsfeld, 1967, "Problemi di metodologia", pp. 179-229.
- P. F., Lazarsfeld, 1961a, "Notes on the history of quantification in sociology: trends, sources and problems", in *Isis*, LII, n. 168, pp. 277-333; tr. it. in Lazarsfeld, 1967, "La quantificazione in sociologia: origini, tendenze e problemi", pp. 3-108.
- P. F., Lazarsfeld, 1961b, "The algebra of dichotomous systems", in Solomon (a c. di) 1961, pp. 111-157; tr. it. parziale, Lazarsfeld, 1967, "L'algebra dei sistemi dicotomici", pp. 413-444.
- P. F., Lazarsfeld, 1966, "Concept formation and measurement in the behavioral sciences: some historical notes", in Di Renzo, 1966.
- P. F., Lazarsfeld, 1967, *Metodologia e ricerca sociologica*, Bologna, Il Mulino.
- P. F., Lazarsfeld e N. W., Henry, 1968, *Latent Structure Analysis*, Boston, Houghton Mifflin.
- P. F., Lazarsfeld, 1993, *On social research and its language*, Chicago, The University of Chicago Press.

- D., Lerner, H.D., Lasswell, (a c. di), 1951, *The policy sciences*, Stanford, Stanford University Press.
- A., Marradi, 1981, "Factor analysis as an aid in the formation and refinement of empirically useful concepts", in Jackson e Borgatta (a c. di), 1981, pp. 11-49.
- A., Marradi, 1996, "Metodo come arte", in *Quaderni di Sociologia*, n. 40, pp. 71-92.
- A. L., McCutcheon, 1987, *Latent Class Analysis*, Beverly Hills, Sage.
- R. K., Merton, L., Broom, L.S., Cottrel jr., (a c. di), 1959, *Sociology today: problems and prospects*, New York, Basic Books Inc.
- E., Nagel, 1961, *The Structure of Science: Problems in the Logic of Scientific Explanation*, New York, Harcourt, Brace & World; tr. it., *La struttura della scienza: problemi di logica della spiegazione scientifica*, Milano, Feltrinelli, 1968.
- L., Perrone, 1977, *Metodi quantitativi della ricerca sociale*, Milano, Feltrinelli.
- L., Ricolfi, 1992a, *Helga. Nuovi principi di analisi dei gruppi*, Torino, Franco Angeli.
- L., Ricolfi, 1992b, "Sul rapporto di indicazione: l'interpretazione semantica e l'interpretazione sintattica", in *Sociologia e Ricerca Sociale*, n. 39, pp. 57-79.
- L., Ricolfi, 1993, *Tre variabili. Un'introduzione all'analisi multivariata*, Milano, Franco Angeli.
- M., Rosenberg, 1968, *The Logic of Survey Analysis*, New York, Basic Books.
- W. E., Saris, L. H., Stronkhorst, 1984, *Causal modelling in nonexperimental research. An introduction to the LISREL approach*, Amsterdam, Sociometric Research Foundation.
- H.A., Simon, 1954, "Spurious Correlation: a causal interpretation", in *Journal of the American Statistical Association*, vol. 49, pp. 467-479.
- H., Solomon (a c. di), 1961, *Studies in item analysis and prediction*, Stanford, Stanford University Press.
- D., Sörbon, K. G., Jöreskog, 1976, *Cofamm: confirmatory factor analysis with model modification user's guide*, Chicago, National Educational Resources Inc.
- P. A., Sorokin, 1956, *Fads and Fobles in Modern Sociology and Related Sciences*, New York, Regnery; tr. it., *Mode ed utopie nella sociologia moderna e scienze collegate*, Firenze, Editrice Universitaria G. Barbera, 1965.
- C., Spearman, 1904, "General intelligence objectively determined and measured", in *American Journal of Psychology*, vol. 15, pp. 201-293.
- S. A., Stouffer et al., 1950, *Studies in Social Psychology in World War II. Vol. IV Measurement and Prediction*, Princeton, NJ Princeton University Press.
- M. Weber, 1922, 2° edizione 1951, *Gesammelte Aufsätze zur Wissenschaftslehre*, Tübingen, Mohr; tr. it., *Il metodo delle scienze storico sociali*, Torino, Einaudi, 1958.