Analisi dei Dati a Tre-vie delle Risposte a Domande Aperte e Indicatori Empirici.

Mary Fraire

Dip.to di Ricerca Sociale e Metodologia Sociologica (RiSMeS)
Facoltà di Sociologia, Università degli Studi di Roma 'La Sapienza'
Corso Italia 38 A, 00198 Roma , Italia
e-mail: fraire@uniroma1.it, fax: 0039-06-84403346

Abstract

A strategy of exploratory multidimensional statistical analysis of open question is proposed with a view to contributing to the assessment of an auxiliary theory to complex social phenomena (Quality of Life, Health etc,) measurement by empirical indicators. After providing a brief exposition of the statistical phases characterizing the measurement process by indicators the A. focuses on two of them. At this aim different and combined textual data analysis (correspondence analysis and three-way data analysis) are applied to the data of a special open question questionnaire employed in an italian Quality of Life survey . After dealing with statistical results and logical observations going out from the strategy of analysis proposed particularly focusing on three-way data analysis, the A. concludes pointing out the original aspects obtained.

Sintesi

L'A. propone una particolare strategia di analisi multidimensionale di dati testuali provenienti dalle risposte libere alle domande aperte di un questionario nell'ambito della problematica della misurazione statistica di fenomeni complessi (es. Qualità della Vita, Salute ecc.) tramite indicatori empirici. Dopo aver brevemente illustrato le principali fasi che caratterizzano il processo di misurazione statistica tramite indicatori, l'A. fissa in particolare l'attenzione su due fasi . A tal scopo diverse e combinate tecniche di analisi dei dati testuali, in particolare l'analisi delle corrispondenze e l'analisi dei dati a 3-vie, focalizzando l'attenzione su quest'ultima, sono applicate come esemplificazione della strategia di analisi proposta, ai dati di un'indagine italiana sulla QdV. I principali risultati statistici e gli aspetti logici emergenti sono quindi esposti nel corso del lavoro mettendo in evidenza quelli più originali.

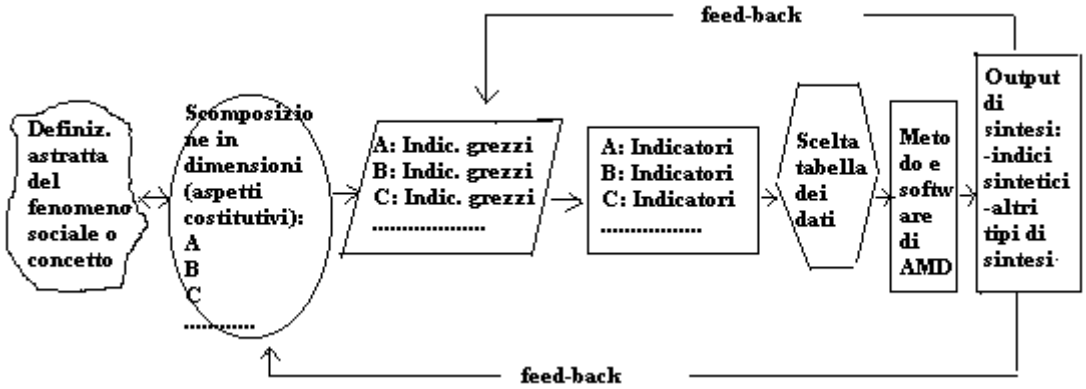
Keywords*:* exploratory multidimensional textual data analysis; three-way data analysis; open questions; complex phenomena; auxiliary theory ; indicators.

1. Introduzione: la problematica considerata

Tra gli impieghi dell'Analisi dei Dati Testuali (ADT) vi è, come noto, l'analisi statistica delle risposte libere date da un campione di *n* individui alle domande aperte di un questionario ossia domande le cui risposte possono essere fornite liberamente dall'intervistato senza alcuna codifica a priori. In generale le domande aperte rispetto a quelle chiuse forniscono un'informazione più ricca ed estesa avente più il carattere di una *espressione* piuttosto che di una *reazione* condizionata dalle modalità proposte.
Quando si tratta di affrontare *collettivi e/o problemi nuovi, poco noti,* le domande a risposta libera sono utili e possono fornire efficaci e nuove informazioni.
Oggetto del presente lavoro è l'impiego dell'analisi statistica multidimensionale dei dati testuali provenienti dalle risposte libere alle domande aperte di un questionario finalizzato alla

costruzione di una teoria ausiliaria *esplicita* alla misurazione statistica di fenomeni sociali o concetti complessi (ad es. Qualità della Vita, Salute ecc.) tramite indicatori empirici.
Non ci si sofferma qui sulla complessa problematica riguardante gli indicatori per la quale si rinvia alla vasta letteratura, ai numerosi studi e ricerche sull'argomento, mentre ci si limiterà ad un aspetto in particolare legato all'osservazione statistica.
Come noto, nell'ambito della problematica della misurazione di fenomeni complessi tramite indicatori empirici , soprattutto nella ricerca di carattere sociale, occorre tener conto del fatto che il linguaggio osservativo non è indipendente dal linguaggio della teoria essendo l'osservazione *intricata* con la teoria. Esiste un *gap* tra concetti (definizioni astratte di un fenomeno) e misure empiriche non colmabile con la logica del certo, regole uniche. Tale *relatività* del sistema di riferimento impiegato-impiegabile per la misurazione empirica, rende particolarmente importante ai fini della ripetibilità delle procedure nelle stesse condizioni , controllabilità, interpretabilità, ecc. delle misure empiriche impiegate, la maggior esplicitazione possibile del processo logico-concettuale di passaggio dal linguaggio teorico al linguaggio osservativo, mediante il quale viene tradotta la definizione concettuale, teorica, del fenomeno complesso (non osservabile statisticamente in modo diretto tramite indici statistici semplici) in indicatori (empiricamente osservabili).
L'approccio metodologico-statistico, tra i tanti esistenti , che qui in particolare si considera è quello di cercare di affrontare la suddetta specificazione delle fasi del processo di misurazione statistica di un fenomeno tramite indicatori (o una loro sintesi in indici sintetici) *in termini formali espliciti* ossia in modo che tale specificazione possa assumere la forma di *una teoria ausiliaria alla misurazione statistica.*
Tale specificazione può assumere ovviamente diversi gradi di definizione, forme di esplicitazione, secondo la natura del fenomeno, le esperienze e conoscenze precedenti, gli scopi della ricerca e così via. Ad esempio tale specificazione può assumere la forma di *modello statistico* di riferimento se è possibile specificare i parametri delle relazioni statistiche (indipendenza, dipendenza, correlazione) ipotizzabili tra tutte le quantità in gioco: a)tra gli aspetti o dimensioni nei quali è stata scomposta la definizione astratta del fenomeno; b) tra queste dimensioni e gli indicatori empirici; c)tra le variabilità residue o specifiche degli indicatori; d) tra dimensioni, indicatori e una loro eventuale sintesi in indici sintetici nonché l'eventuale stima del grado di attendibilità delle misure sintetiche ottenute.
Molto schematicamente per rappresentare le varie *fasi* (statistiche e informatiche) della misurazione statistica di un fenomeno complesso tramite indicatori e una loro eventuale sintesi in indici sintetici si può impiegare l'ipotetigrafia riportata nella Fig.1. (Fraire 1987,1989)

Fig.1 - Le fasi di misurazione statistica di un fenomeno complesso tramite indicatori e indici sintetici.



Per la descrizione delle varie fasi e *feed-back* riportati nella Fig.1 si rinvia a precedenti lavori (Fraire, 1987, 1988,1989,1994) riguardanti in particolare le ultime 4 fasi e la costruzione di una *teoria ausiliaria esplicita* nella forma di *modelli statistici* basati sull'analisi fattoriale.

La strategia di analisi multidimensionale dei dati testuali oggetto del presente lavoro si riferisce invece in particolare al contributo che queste analisi - proprio perché riferite ad aspetti ancora in forma 'lessicale' ma già empirici (statisticamente osservabili) - possono dare nel passaggio dalla seconda alla terza fase ossia dagli aspetti costitutivi o dimensioni nei quali è stato scomposto 'a priori' il fenomeno sociale o concetto agli indicatori empirici corrispondenti.

2. Aspetti metodologici

2.1 La documentazione statistica di partenza: campo d'indagine, questionario, popolazione e le matrici dei dati iniziali.

Per l'esempio di applicazione delle analisi proposte si sono impiegati, a carattere puramente esemplificativo essendo il *corpus* del file di testo non molto grande ma tuttavia significativo sia dal punto di vista del particolare questionario sperimentato che delle risposte ottenute, i dati derivanti da un'indagine sulla Qualità della Vita delle Comunità Montane (C.M.) italiane (INEMO, 1983). Il particolare questionario 'Scheda descrittiva-per problemi' era caratterizzato da una serie di 8 domande aperte miranti ad individuare le '*preoccupazioni sociali rilevanti'* per ciascuna delle seguenti 8 **'***aree di rilevanza sociale':* A: 'Salute' ; B:'Istruzione e Formazione professionale'; C: 'Occupazione e Qualità del lavoro'; D:'Impiego del tempo libero'; E: 'Situazione economica personale'; F:'Ambiente fisico'; G:'Ambiente sociale'; H: 'Sicurezza personale', nelle quali era stata scomposta la definizione astratta di QdV. Per ciascuna area veniva richiesto agli operatori delle 22 C.M. intervistate, appartenenti a tutte le regioni italiane, supposte essere persone capaci di una conoscenza e visione complessiva dei problemi della propria C.M., di esprimere liberamente, per iscritto, secondo il proprio linguaggio, quali fossero i problemi o le preoccupazioni sociali più rilevanti.
A rilevazione avvenuta le risposte alle 8 domande aperte date da ciascuno dei 22 operatori hanno costituito il *file di testo* originario composto quindi da 8 sub-*files* uno per ogni domanda aperta (area di rilevanza sociale) . Il *file numerico* - contenente le variabili sulle caratteristiche delle ***n*** u.s. (qui le 22 C.M.) che potevano essere impiegate come *variabili di raggruppamento (attive)* o *supplementari (passive*) - associato al file di testo delle risposte alle domande aperte (variabili attive) ha riguardato due caratteri statistici:1) RIPGEO = la ripartizione geografica di appartenenza della CM intervistata avente le seguenti modalità: Nord-Ovest; Nord-Est; Centro; Sud; Isole. 2) la REGIONE di appartenenza della C.M..

2.2 Codifiche a posteriori: correzione, disambiguazione e segmentazione del testo e le prime analisi statistiche dei testi.

Si è proceduto quindi alla correzione, disambiguazione e riduzione (segmentazione) del testo mediante l'impiego del *software* SPADT e ad effettuare una prima analisi statistica sulla frequenza, lunghezza ecc. delle forme lessicali sia sul file di testo originale che sul file di testo ' ridotto' (segmentato) . Si è ritenuta tale fase di *codifica a posteriori* molto importante ai fini della scelta della tabella dei dati da sottoporre alle successive analisi statistiche e della significatività e interpretabilità dei risultati stessi delle analisi. E' da osservare che il questionario è stato compilato direttamente dagli intervistati pertanto le operazioni di correzione e disambiguazione (es. alcolismo , alcoolismo, famigliare, familiare ecc.) del file di testo originario sono state essenziali.
Circa la segmentazione del testo, sebbene quella qui impiegata sia stata quella standard proposta da SPADT tuttavia con un ulteriore intervento, ad esempio ampliando il numero delle parole costituenti un segmento, si sarebbero potuti ottenere risultati ancora più significativi, ad esempio, una migliore interpretabilità degli assi fattoriali . Nell'esempio

considerato, come già rilevato, essendo il *corpus* già abbastanza ridotto si è scelto di non intervenire ulteriormente sul testo per non ridurre ulteriormente le occorrenze delle forme lessicali.
Circa la dimensione del *corpus* degli 8 sub-*files* di testo, considerando i files di testo *originario*, ossia non ridotto, si va da un valore dell'indice (cfr. Bolasco, 1999) V/N x100=45% (con N=199 occorrenze e V= 90 parole distinte) per le risposte alla domanda relativa all'area H: 'Sicurezza personale', in cui tuttavia vi erano molte mancate risposte, ad un minimo dell'indice di 17,9% per le risposte relative all'area G: 'Ambiente sociale'.
L'indice migliora notevolmente considerando i sub-*files* di testo *ridotti* (segmentati) in cui nel peggiore dei casi l'indice V/N non supera il 34%. Si abbassa però notevolmente il numero delle occorrenze.
Si sono effettuate quindi diverse analisi sia sui *files* di testo originale che su quelli ridotti, sia per regioni che per ripartizioni geografiche . Nel seguito tuttavia per brevità si farà riferimento solo alle ripartizioni geografiche.

2.3 Le forme lessicali caratteristiche e gli indicatori lessicali empirici.

In queste prime analisi si è cercato di selezionare degli elementi (forme lessicali considerate: parole e segmenti) caratteristici (frasi modali) separatamente per ognuna delle 8 domande (aree di rilevanza sociale) e classificate secondo la mutabile scelta come variabile attiva di raggruppamento sia delle parole che dei segmenti: la ripartizione geografica. Le tabelle dei dati scelte sono state *tabelle di contingenza parole (o segmenti)* ripgeo(5 ripartizioni geografiche),* per ciascuna delle 8 aree di rilevanza sociale.
Una prima analisi dei *files di testo originale* fornisce alcune indicazioni interessanti per l'individuazione di indicatori empirici corrispondenti a ciascuna delle 8 aree attraverso le espressioni verbali, frasi caratteristiche, sintetiche di 'situazioni' o 'preoccupazioni' sociali rilevanti. Non si riportano qui per brevità le risposte caratteristiche, ottenute con il criterio (distanza) del Chi quadrato, date dalle 5 ripartizioni geografiche limitandoci ad osservare che in taluni casi le forme lessicali sono assimilabili a veri e propri indicatori 'lessicali' empirici (es. alto tasso di malati per silicosi; alto tasso di invalidi civili ecc.) con relativa valutazione di entità basata ovviamente sulla conoscenza e percezione personale dell'operatore ossia indicatori soggettivi lessicali empirici.
E' da osservare che con questo tipo di analisi le 8 aree di rilevanza sociale sono analizzate separatamente così come le modalità della variabile di raggruppamento scelta (le 5 ripartizioni geografiche) in altri termini manca la rappresentazione simultanea delle 5 ripartizioni geografiche e delle rispettive preoccupazioni sociali rilevanti (risposte), sia per ogni singola area che per il complesso delle 8 aree di rilevanza sociale. A tal scopo occorreva un approccio multidimensionale per cui si sono impiegate le analisi statistiche descritte nei paragrafi seguenti.

3. Le analisi multidimensionali : analisi delle corrispondenze delle risposte per area di rilevanza sociale.

Con questo tipo di analisi si sono volute analizzare contemporaneamente tutte le modalità della variabile attiva considerata (le 5 ripartizioni geografiche) e le forme lessicali (segmenti) di ognuna , considerando tuttavia le 8 aree di rilevanza sociale separatamente. In particolare, molto schematicamente, si è effettuata l'*analisi delle corrispondenze semplici (ACS)* sulle *tabelle di contingenza segmenti* ripgeo,* per ciascuna delle 8 aree di rilevanza sociale. E' stato possibile in tal modo esaminare sul piano fattoriale delle corrispondenze la relazione tra le 5 ripartizioni geografiche e le forme lessicali più caratterizzanti ciascuna di esse relativamente ad una data area di rilevanza sociale.

Non ci si sofferma qui per brevità sui risultati di tali analisi. Si accenna tuttavia, anche ai fini delle analisi successive, che è risultata significativa la percentuale di inerzia totale spiegata dai primi 3 fattori per ciascuna delle 8 aree: da un minimo dell'80,53% per l'area G: 'Ambiente Sociale' ad un massimo del 92,74% per l'area D: 'Impiego del Tempo Libero'. Rispetto ai piani fattoriali di ciascuna area, in particolare con riferimento al primo piano fattoriale la minima percentuale d'inerzia spiegata dai primi due fattori è risultata per l'area G:'Ambiente Sociale' del 56,42% mentre la più alta del 71,02% per l'area D:'Impiego del tempo libero'. Sempre limitatamente al primo piano fattoriale le ripartizioni geografiche sono state caratterizzate da forme lessicali statisticamente significative (con contributi assoluti generalmente abbastanza alti e netti su uno dei fattori) eccettuata la ripartizione Centro che nella maggior parte delle aree si è collocata vicino al baricentro corrispondente, come noto, se al netto dell'autovalore banale uguale a 1, al profilo medio equivalente all'ipotesi di indipendenza tra i due caratteri considerati con forme lessicali anch'esse statisticamente scarsamente significative. E' da osservare che sebbene non sempre agevole sia l'interpretazione degli assi soprattutto nel caso di dati testuali, mediante l'ACS è stato possibile 'posizionare' e quindi porre in relazione le 5 ripartizioni geografiche rispetto alle forme lessicali (risposte) qui indicanti le preoccupazioni sociali rilevanti dell'area. In tal senso gli assi fattoriali possono essere assimilati a '*variabili lessicali sintetiche*' tanto più significative quanto più i risultati dell'ACS consentono una chiara interpretazione degli assi.
Dall'ACS si sono quindi impiegate le 8 *tabelle di dati quantitativi,* $_{i}X_{n,k}$ (i = 1,..,8 aree di rilevanza sociale; n = 1,..,5 ripartizioni geografiche; k = 1,..,3 coordinate dei primi 3 fattori (punteggi fattoriali) di ciascuna area, per le analisi descritte al punto successivo.

4. L' Analisi dei Dati a Tre-Vie

4.1 La scelta della tabella dei dati a tre-vie

Mediante l'analisi dei dati a tre-vie si sono volute esaminare tutte insieme, contemporaneamente e globalmente, le 8 aree di rilevanza sociale (occasioni), le u.s. (5 ripartizioni geografiche) e le tre variabili (primi 3 fattori di ogni area) ossia le 8 tabelle di contingenza nella forma di tabelle di dati quantitativi suddetta al fine di individuare:
a) attraverso la rappresentazione *globale* di tutte le tabelle quali di esse hanno una struttura simile o diversa e, attraverso la loro distanza dalla tabella 'media' (matrice compromesso), valutare quali di esse contribuiscono maggiormente alla parte di variabilità eccedente quella comune rappresentata dalla matrice compromesso (*analisi dell'interstruttura*);
b) individuare sia per gli individui (medi) che per le variabili (medie) le caratteristiche della *variabilità media* nelle 8 occasioni (*analisi dell'intrastruttura*).
c) individuare e confrontare le *traiettorie* di ciascuna u.s.(qui le 5 ripartizioni geografiche) per l'insieme delle 8 occasioni (aree di rilevanza sociale) considerate ossia i diversi 'percorsi' e quindi il diverso ruolo di ciascuna area di rilevanza sociale nel definire la QdV di ciascuna ripartizione (*analisi delle traiettorie*).
Prima di esaminare molto brevemente i risultati , un cenno meritano le codifiche a posteriori effettuate per la *tabella dei dati a tre-vie* scelta per le analisi.

Come noto, si possono avere diverse codifiche a posteriori di una matrice dei dati a tre-vie (Rizzi, 1989) secondo le situazioni di ricerca in cui : 1) si considerano *diverse le K variabili* rilevate in O occasioni successive ed *uguali le N u.s.;* 2) si considerano *uguali le K variabili* e *diverse le N u.s. rilevate in O occasioni successive;* 3) si considerano *uguali sia le N u.s. che le K variabili* rilevate in O occasioni successive.
Nell'applicazione considerata, la scelta della codifica a posteriori è stata quella di tipo 1) in cui si sono considerate le 8 *tabelle di contingenza* $_{i}X_{n,k}$ [**i** = 1,..,8 occasioni; **n** =1,..,5 u.s.; **k**

= 1,..,3 variabili] aventi *diverse le 3 variabili* rilevate nelle 8 occasioni (aree di rilevanza sociale) e uguali le u.s. (ripartizioni geografiche).
Per le elaborazioni dei dati si è impiegato il *software* ACT- Méthode STATIS. Inoltre si sono considerati i dati centrati e ridotti perché la variabilità delle singole variabili era molto diversa. Infine si è richiesta la normalizzazione delle tabelle per ottenere coefficienti di relazione tra tabelle (l'indice RV di Escoufier) varianti tra 0 e 1.

4.2 Analisi dell'interstruttura

Partendo dalla matrice dei dati a tre-vie ${}_i X_{n,k}$ (i = 1,..,8; n = 1,..,5; k = 1,..,3) suddetta essendo stata richiesta la normalizzazione delle matrici si è ottenuta la matrice RV , quadrata e simmetrica, 8x8, ,contente i coefficienti RV di Escoufier (cfr. ad es. Rizzi, 1985), varianti tra 0 e 1, indicanti la similarità tra le coppie di matrici (per valori di RV prossimi ad 1) e dalla quale è quindi possibile individuare quali sono *globalmente* le matrici che hanno strutture simili, più vicine. Come noto (cfr. ad es. Bolasco 1999) la similitudine tra matrici riguarda la similitudine della nuvola dei punti-unità nelle diverse coppie di occasioni nel senso che gli individui che hanno la stessa struttura hanno le posizioni dei punti omologhi che non sono cambiate (sono stabili) a prescindere dal cambiamento delle variabili nelle diverse occasioni considerate. Nella Tab.1 si riporta la matrice dei coefficienti RV per le 8 tabelle considerate.

Tab.1 - Matrice dei coefficienti RV

	1	2	3	4	5	6	7	8
1	1.000							
2	.668	1.000						
3	.610	.637	1.000					
4	.633	.704	.926	1.000				
5	.802	.658	.779	.660	1.000			
6	.596	.935	.723	.695	.687	1.000		
7	.643	.973	.706	.704	.702	.973	1.000	
8	.767	.620	.750	.838	.670	.624	.577	1.000

Dalla tab.1 si può vedere, ad esempio, che le matrici aventi struttura più simile sono risultate quelle relative alle coppie di occasioni: 2:'Istruz. e Formaz.Prof.' - 6:'Ambiente fisico' (RV=0,935); 2:'Istruz. e Formaz.Prof.' - 7:'Amb.Sociale' (RV=0.973); 3: 'Occupaz. e QdL - 4:'Impieghi del T.L.' (RV=0.926); 6: 'Ambiente Fisico' – 7: 'Ambiente Sociale' (RV=0.973).
Dalla diagonalizzazione della matrice RV, limitatamente qui solo alle prime due componenti principali che spiegano l' 86,46% della variabilità totale con autovalori uguali a $\lambda_1 = 75{,}89\%$ e $\lambda_2 = 10{,}57\%$, è possibile rappresentare sul primo piano fattoriale la relazione tra le prime due componenti e le 8 aree di rilevanza sociale considerate (punti-occasione), che qui per brevità non si riporta.
Si è considerato inoltre il piano delle prime due componenti *centrate* rispetto alla matrice compromesso (WD) che spiega il 72,04% dell'inerzia totale con autovalori rispettivamente uguali a: $\lambda_1 = 43{,}90\%,\ \lambda_2 = 28{,}14\%$.
Come noto la matrice compromesso (WD) è data dalla media ponderata delle matrici di similarità (o distanza) *n×n* tra individui ${}_1 S_{n,n}; {}_2 S_{n,n}; ...; {}_8 S_{n,n}$ (corrispondenti alle matrici originarie ${}_i X_{n,k}$ per i=1,..,8) ponderate con gli autovettori corrispondenti al primo (più grande) autovalore della matrice $C \equiv c_{ij}$ con $c_{ij} = tr({}_i S_j S)$. Poiché si basa solo sul primo autovalore la matrice compromesso WD è 'robusta' (Rizzi, 1987) in quanto poco influenzata dalle piccole variazioni delle matrici di similarità. Circa il significato della matrice compromesso è da osservare che nell'esempio esaminato pur essendo le variabili diverse nelle 8 occasioni considerate esse si possono pensare facenti parte di uno stesso concetto, la QdV, perché rappresentano gli 8 aspetti costitutivi della QdV *definiti a priori*, pertanto anche la

matrice WD ha un senso. La matrice compromesso è in sostanza la sintesi di tutte le matrici considerate attraverso la matrice 'media' più rappresentativa. Significativo è allora il piano fattoriale *centrato* rispetto ad essa attraverso il quale è possibile esaminare la distanza delle diverse matrici dalla matrice compromesso (ossia al 'netto' della parte comune di variabilità) e la loro relazione rispetto ai primi due assi principali. Dal *plot* delle 8 aree di rilevanza sociale sul primo piano fattoriale centrato (72,04% dell'inerzia totale) rispetto alla matrice compromesso , che qui per brevità non si riporta, si possono rilevare le aree di rilevanza sociale (tabelle) maggiormente correlate con il primo asse: 7:'Ambiente sociale'; 6:'Ambiente fisico'; 2:'Istruzione e Formaz.Prof.' ; 8:'Sicurezza sociale'; e le aree di rilevanza sociale maggiormente correlate con il secondo asse: 1:'Salute'; 5: 'Situazione Economica Personale'; 3:'Occupazione e QdL'; 4:'Impieghi del T.L.' .
Il primo asse è quindi caratterizzato da un *cluster* di 4 aree di rilevanza sociale 'sovrastrutturali' rispetto alla QdV (ed è quello che spiega la più alta percentuale di inerzia, 43,90%, dovuta alla variabilità eccedente la variabilità 'media') mentre il secondo asse è più rappresentativo di bisogni 'strutturali' (con percentuale nettamente inferiore d'inerzia spiegata 28,14%) tra questi colpisce l'inserimento degli impieghi del T.L. considerato ormai (l'indagine risale al 1983) tra questi.
Questa classificazione '*a posteriori*' in due clusters delle 8 aree definite '*a priori*' emerge comunque dall'analisi della parte di variabilità delle 8 tabelle eccedente la variabilità 'media'.
Esaminiamo allora più dettagliatamente sia rispetto alle u.s. che alle variabili le caratteristiche della variabilità 'media'.

4.3 Analisi dell'intrastruttura

Dalla diagonalizzazione della matrice compromesso WD si ottengono, limitatamente qui ai primi due fattori, i seguenti autovalori: $\lambda_1 = 32,48\%$ e $\lambda_2 = 28,00\%$ quindi il primo piano fattoriale spiega il 60,48% dell' inerzia totale della matrice compromesso. Si possono allora rappresentare sul primo piano fattoriale sia le variabili (punti variabili-medie) che le u.s. (punti individui-medi) rispetto ai primi due assi compromesso.
Nella Fig.2 si riporta, limitatamente qui alle 5 ripartizioni geografiche (punti individui-medi), la loro rappresentazione rispetto ai primi due assi compromesso.
Sul primo asse si evidenzia la contrapposizione tra NEst (-) e NOvest (+), mentre sul secondo asse la contrapposizione tra Sud (+) e Isole (-). Il Centro come già rilevato nell'Analisi delle corrispondenze è scarsamente correlato (contributi assoluti molto bassi) su entrambi gli assi.

4.4 Analisi delle traiettorie

Per poter esaminare contemporaneamente le ripartizioni geografiche, le diverse strutture delle tabelle e le 8 aree di rilevanza sociale è utile rappresentare le *traiettorie di ciascuna u.s.* rispetto ai piani fattoriali ottenuti dall'analisi dell'intrastruttura (cfr par.4.3) oppure riportare le traiettorie delle u.s. rispetto ai singoli assi fattoriali sviluppati in funzione dell'indicizzazione delle 8 tabelle ed in cui sull'asse verticale vi sono i punteggi fattoriali riferiti all'asse e su quello orizzontale le 8 aree di rilevanza sociale.
Nella Fig.3 si riportano le traiettorie delle 5 ripartizioni geografiche, qui per brevità riferite solo al primo asse fattoriale.Si può notare che la traiettoria della ripartizione NOvest si differenzia nettamente dalle altre nell'articolazione delle 8 aree di rilevanza sociale che costituiscono la definizione di QdV ipotizzata mentre rispetto al secondo asse fattoriale, che qui non si riporta, risulta che è il Sud a differenziarsi nettamente dalle traiettorie delle altre ripartizioni. Inoltre, pur nella diversità dei 'percorsi' delle ripartizioni vi sono delle aree vicine ossia tabelle di una data area simili nella struttura, ad esempio nella Fig.3: l'area 1 :

Fig.2 - Rappresentazione delle 5 ripartizioni geografiche (5 punti unità-medi) rispetto ai primi due assi compromesso (60,48 %)

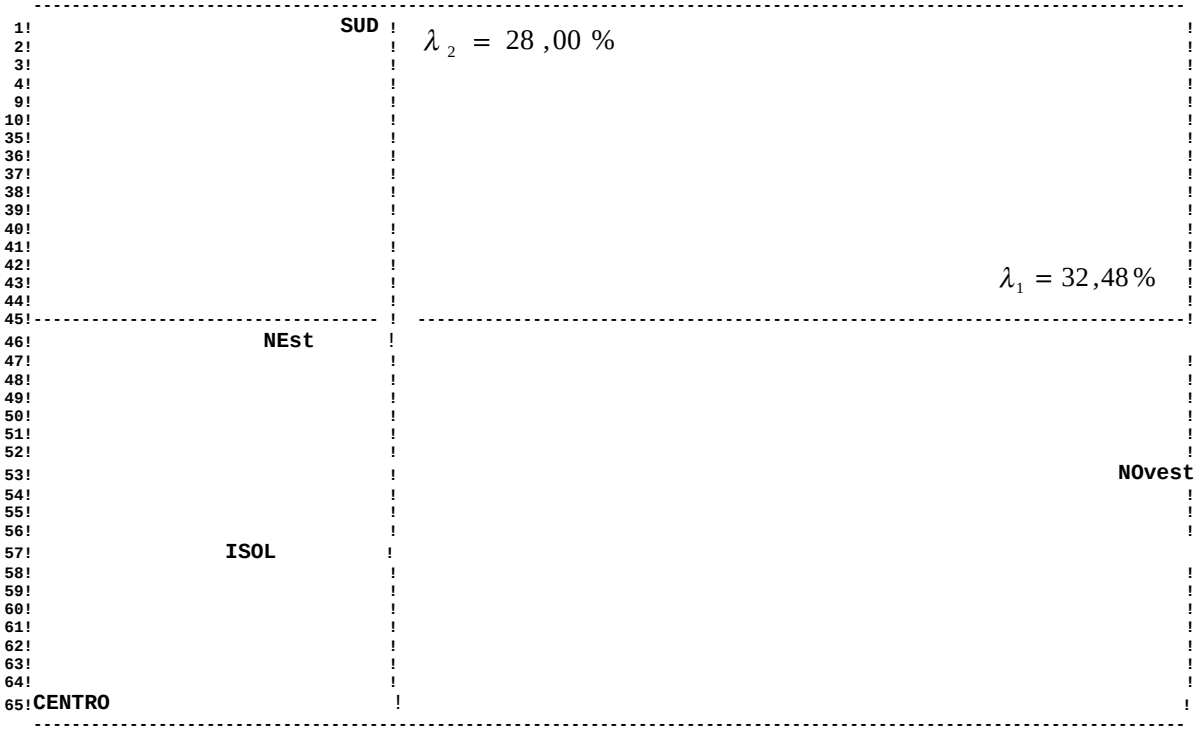
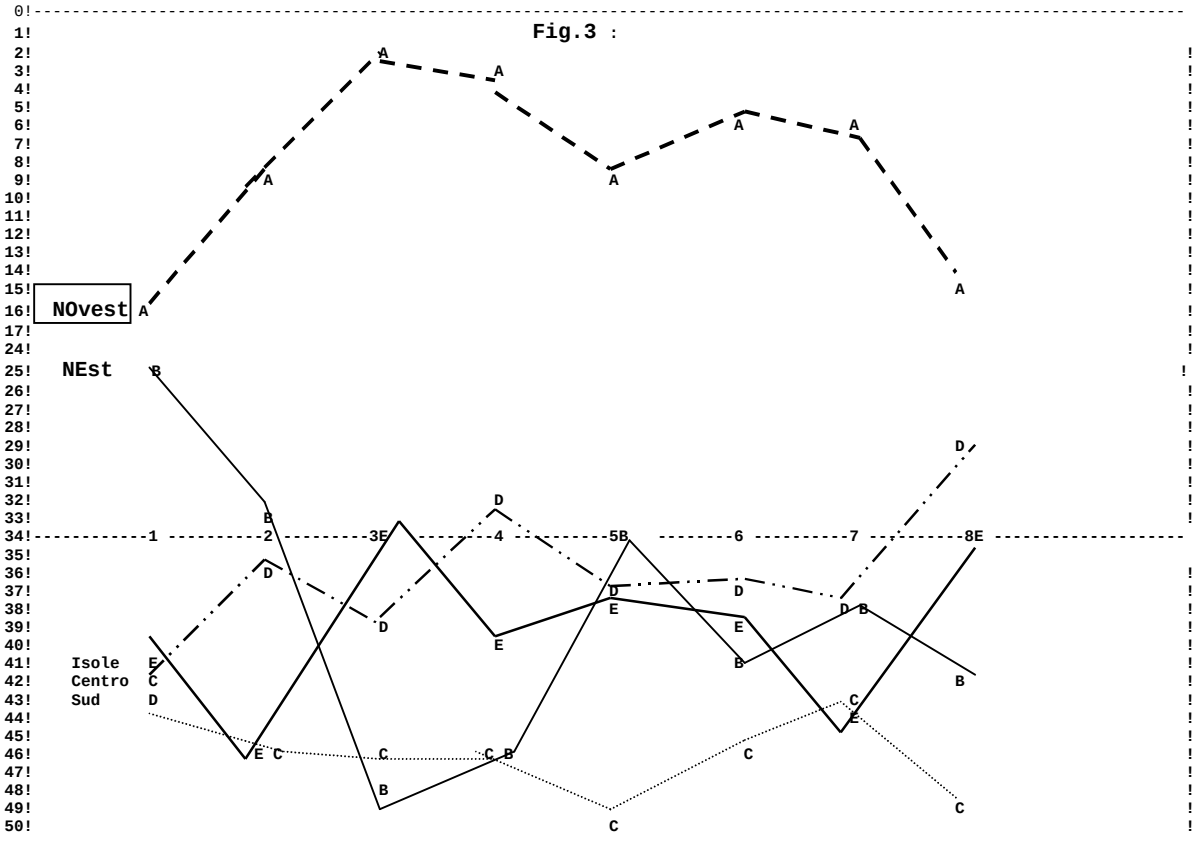


Fig.3 – Traiettorie delle 5 Ripartizioni geografiche rispetto al primo asse fattoriale in funzione delle 8 aree di rilevanza sociale.



*Legenda:Ripartizioni Geografiche.: **A**: NOvest; **B**: NEst; **C**: Centro; **D**:Sud; **E**: Isole. Aree di Rilevanza Sociale:**1**:Salute; **2**:Istruz.eForm.Prof.; **3**:Occup.eQdL; **4**:ImpieghiT.L.; **5**:Sit.Econ.Pers.; **6**: Ambiente Fisico; 7:Ambiente Sociale; 8:Sicurezza Personale.*

'Salute' nelle Isole, Centro e Sud; l'area 7: 'Ambiente sociale' nel NEst e Sud; Centro e Isole; l'area 4: 'Impieghi del T.L.' per il Centro e NEst.

5. Alcune considerazioni conclusive

Attraverso la strategia di analisi proposta è stato possibile *esplicitare empiricamente* il contenuto degli aspetti costitutivi (aree di rilevanza sociale) ipotizzati per la definizione della QdV, ossia le preoccupazioni sociali rilevanti, attraverso indicatori lessicali empirici; individuare gli aspetti (aree) più diversi nella loro struttura globale (analisi dell'interstruttura) e nel contempo quelli contribuenti maggiormente alla differenziazione delle diverse ripartizioni geografiche considerate (analisi dell'intrastruttura). L'emergere inoltre di due *clusters* di aree di rilevanza sociale, assimilabili a bisogni 'strutturali' e 'sovrastrutturali', consente di intravedere i motivi della differenziazione tra le ripartizioni . Infine attraverso l'analisi delle traiettorie delle ripartizioni rispetto agli aspetti costitutivi si può disporre in modo compatto di un'analisi dinamica, ossia dei bisogni nelle diverse aree delle diverse ripartizioni geografiche, in questo caso ponendo l'accento sul carattere 'sincronico' o 'diacronico' di essi. Ad esempio, il NOvest è risultato caratterizzato da una traiettoria diversa legata a preoccupazioni sociali caratteristiche di uno sviluppo socio-economico più avanzato (verificabile anche dagli indicatori lessicali empirici) rispetto a quello delle altre 4 ripartizioni per il primo asse dei bisogni 'sovrastrutturali', mentre è il Sud che presenta una traiettoria diversa rispetto alle altre ripartizioni per il secondo asse dei bisogni 'strutturali' . In alcune ripartizioni c'è invece 'sincronia', pur nella diversità dei percorsi, nelle preoccupazioni sociali di alcune aree. L'esempio qui considerato ha riguardato la QdV ma si potrebbero considerare altri fenomeni come ad es. la Salute in positivo (Well-Health), l'Uso del Tempo giornaliero ecc. , per i quali l'analisi di dati testuali attraverso gli indicatori e le variabili lessicali derivanti dalle risposte a domande aperte, possono fornire *lo scenario globale e articolato, qualitativo e dinamico* caratterizzante *dati* gruppi sociali o popolazioni, di ausilio anche per la misurazione empirica del fenomeno considerato tramite indicatori di tipo quantitativo-oggettivo-descrittivo.

Riferimenti bibliografici

ACT (1989). *Installation e Description de la Méthode STATIS.* CISIA, France

Bolasco S. (1999). *Analisi Multidimensionale dei Dati.* Carocci Ed., Roma.

Fraire M. (1987). *Qualità della Salute:Problemi e Metodi Statistici di Misurazione Tramite Indicatori.* Quaderni del Dip.to SPSA,Università di Roma 'La Sapienza', Serie A-Ricerche n.17/1987

Fraire M. (1989). *Problemi e metodologie statistiche di misurazione di fenomeni complessi tramite indicatori e indici sintetici.* in Statistica, n.2,1989.

Fraire M. 1994 - *Metodi di Analisi Multidimensionale dei Dati*, Ed.CISU, Roma.

INEMO (1983), *Scheda descrittiva per problemi di 22 Comunità montane*, in Inemo-informazioni, n.3/4 luglio/Dicembre 1983.

Rizzi A. (1985), *Analisi dei dati*, Ed. NIS, Roma, 1985.

Rizzi A.(1987), *Sulla matrice media*, Quaderni del Dip.to SPSA,Università di Roma 'La Sapienza', Serie A-Ricerche n.2/1987

Salem A. 1995, *Les unités lexicometriques* - in JADT 1995, Roma Ed.CISU

SPAD.T (1993), *Introduction à SPAD.T intégré.Version 1.5P.C.*, CISIA, Saint-Mandé, France