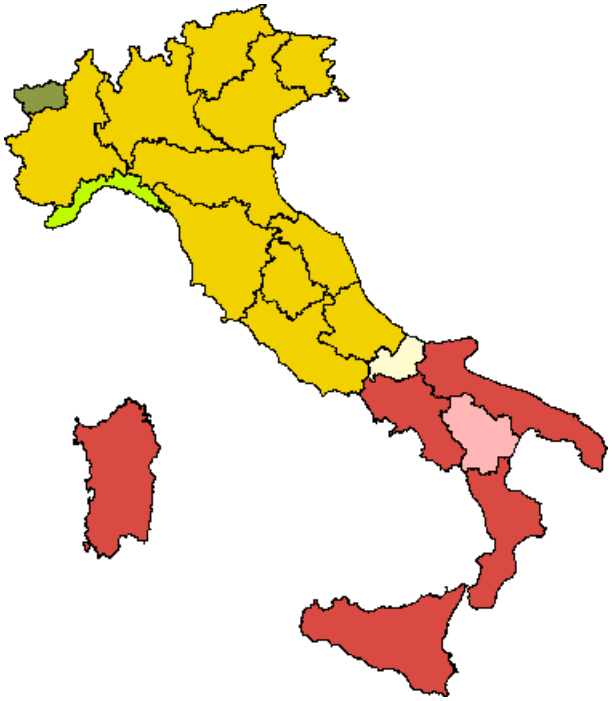
Università degli Studi della Calabria
Arcavacata di Rende (Cosenza)
Dipartimento di Economia e Statistica

Ricerca

Classificazione di serie storiche brevi: il caso delle regioni italiane



Coordinatore **Autore**

Settembre 2006

Note

1 La ricerca è stata realizzata usando le strutture del dipartimento di Economia e Statistica dell'Università della Calabria.

2 Il finanziamento della ricerca è dovuto alla Regione Calabria con Legge Regionale n.10 del 22 settembre 1998 (legge finanziaria).

Indice

Note **ii**

1 Classificazione di serie storiche brevi: il caso delle regioni italiane **1**

1.1 Introduzione . . . 1
1.1.1 Natura delle serie storiche . . . 3
1.1.2 Variazioni relative . . . 5
1.1.3 Variazioni relative medie . . . 8
1.1.4 Scomposizione di una serie storica . . . 9
1.1.5 Distanza tra serie storiche . . . 12
1.2 Metodologia Distatis . . . 20
1.2.1 Trasformazione delle matrici di distanze . . . 22
1.2.2 Matrice sintesi delle distanze parziali . . . 23
1.2.3 Rappresentazione di unità e indicatori . . . 27
1.2.4 Sintesi delle macrodeterminanti . . . 28
1.3 Analisi dei gruppi . . . 31
1.3.1 Concetto di cluster . . . 31
1.3.2 Formazione dei gruppi . . . 35
1.3.3 Scelta del legame . . . 36
1.4 Applicazione a dati simulati . . . 40
1.5 Scelta degli indicatori . . . 44
1.5.1 Quadro comunitario di sostegno . . . 44
1.5.2 Conti economici regionali . . . 54
1.6 Risultati dell'analisi . . . 58
1.6.1 Analisi in blocco . . . 58

1.6.2 Analisi per blocchi 63
1.6.3 Conclusioni 70

Bibliografia **72**

Sitografia **75**

Elenco delle figure

1.1 Serie storiche da considerare simili 9
1.2 Scomposizione dei movimenti di una serie storica 11
1.3 Confronto di due serie storiche 14
1.4 scambio di due segmenti 15
1.5 Confronto tra una serie storica e sue ritardate 19
1.6 Descrizione della distanza tra unità in base a serie storiche 21
1.7 Schema generale di scelta degli indicatori. 28
1.8 Approccio per determinanti separate. 29
1.9 Gruppi naturali 32
1.10 definizione grafica di cluster 32
1.11 Diversità di strutture 34
1.12 Curve di Andrews dei dati simulati 41
1.13 Posizionamento degli indicatori e delle unità nel dataset simulato 42
1.14 Analisi dei gruppi nel dataset simulato 43
1.15 Esito della classificazione. Analisi in blocco. Dati Q.C.S. 60
1.16 Esito della classificazione. Analisi in blocco. Dati C.E.R. 61
1.17 Esito della classificazione. Analisi in blocco. Dati Q.C.S.+C.E.R. 62
1.18 raggruppamenti delle regioni regioni in base a 170 serie storiche macroeconomiche 63
1.19 Pesi fattoriali nella componente trasversale. Dati Q.C.S. 64
1.20 Esito della classificazione. Analisi per blocchi. Dati Q.C.S. 65
1.21 Pesi fattoriali nella componente trasversale. Dati C.E.R. 66
1.22 Esito della classificazione. Analisi per blocchi. Dati C.E.R. 67
1.23 Pesi fattoriali nella componente trasversale. Dati Q.C.S.+C.E.R. . 68
1.24 Esito della classificazione. Analisi per blocchi. Dati Q.C.S.+C.E.R. 69

Elenco delle tabelle

1.1 Avvisi di gara in *Gazzetta Ufficiale* nel periodo 1/1-30/6/1993 . . 5
1.2 Sistema di ponderazione degli scarti tra variazioni relative 18
1.3 Valori delle serie storiche del dataset simulato 41
1.4 Coefficiente di rappresentatività del fattore ponderale. Dati Q.C.S. 64
1.5 Coefficiente di rappresentatività del fattore ponderale. Dati C.E.R. 66
1.6 Coefficiente di rappresentatività del fattore ponderale. Basi Unificate 68

CAPITOLO 1

Classificazione di serie storiche brevi: il caso delle regioni italiane

Parole chiave: analisi gerarchica dei gruppi, distanza tra serie storiche, metodologia Distatis, Quadro comunitario di sostegno, Conti economici regionali.

1.1 Introduzione

L'obiettivo del presente studio è di proporre l'applicazione della metodologia Distatis già rivelatisi molto utile per la sintesi delle strutture di dati a tre vie, per l'individuazione di macroregioni omogenee, non necessariamente contigue, a partire da serie storiche brevi rilevate sulle singole regioni italiane. Più specificatamente, tenteremo di individuare, grazie ad una nuova tecnica che prevede l'uso combinato delle serie storiche e dell'analisi dei gruppi, regioni con una simile evoluzione temporale al fine di orientare le politiche di sviluppo. Pur trattandosi di una metodologia generalizzabile a tutte le regioni essa viene in particolare proposta per individuare, qualora fossero presenti, i divari territoriali interni maturati nelle regioni Obiettivo 1 e non Obiettivo 1 per graduare e meglio finalizzare le politiche regionali di sostegno alla crescita. Si ricorda che le regioni dell'obiettivo 1 sono quelle in ritardo di sviluppo, cioè con un prodotto interno lordo pro-capite

inferiore al 75% della media comunitaria. Per quanto riguarda l'Italia, rientrano nella categoria: Basilicata, Calabria, Campania, Puglia, Sardegna, Sicilia e, inoltre, il Molise, unica regione in sostegno transitorio.

Il progetto di realizzare un nuovo sistema di classificazione delle regioni italiane con la Calabria in primo piano è stato ispirato dalla necessità di rispondere con strumenti tecnici adeguati alla crescente domanda di analisi di impatto sulle politiche pubbliche, di scenari sugli effetti di strategie di sviluppo e di previsioni a medio e lungo termine sulle tendenze di sviluppo delle economie regionali. Le esigenze conoscitive di questo tipo si stanno diffondendo anche a livello regionale, sulla base della diffusione della cultura della valutazione, che ha nella politica regionale europea un forte stimolo. Anche l'ampliamento delle competenze delle Regioni a statuto ordinario, peraltro in corso di definizione legislativa, e la maggiore autonomia finanziaria regionale hanno ormai creato l'esigenza di inserire nei documenti di programmazione finanziaria scenari sulle tendenze delle economie locali e sulla possibile evoluzione delle entrate di bilancio.

La classificazione di serie storiche in gruppi caratterizzati da un comune andamento tendenziale è un importante problema statistico che interessa molti settori di ricerca

- Economia. La presenza di strutture evolutive simili fra industrie diverse può rivelare risposte comuni ai moti congiunturali.
- Finanza. Costruzione di *leading indicators* per l'andamento di un fondo titoli o di un portafoglio a scopo di previsione e controllo.
- Marketing. Tipologia delle sequenze di contatti all'interno di un sito per la costruzione di profili degli utenti.
- Medicina. Confronto di un elettrocardiogramma con altri analoghi che hanno preceduto problemi cardiaci a scopo di prevenzione e cura.
- Biologia. Esperimenti in forma di serie storica trovano ampio spazio nello studio di sistemi biologici quali la risposta alle variazioni di temperatura o altre condizioni di stress.

Si veda comunque Liao [23] per una buona panoramica sulle applicazioni delle tecniche di raggruppamento alle serie storiche. Il problema specifico affrontato in questo lavoro riguarda la possibilità di classificare le regioni italiane in gruppi considerando l'evoluzione temporale, per un arco di tempo breve, di indicatori opportunamente scelti.

Lo schema del presente lavoro è il seguente. Dopo alcune note introduttive sulla natura delle serie storiche e sulla misurazione delle varazioni relative, lo studio dà spazio all'argomento fondamentale dell'analisi dei gruppi: la definizione di un criterio di distanza fra serie storiche. In particolare verrà discussa la problematica della definizione di una metrica fra serie storiche brevi basata sul confronto

delle matrici dei tassi loro tassi di crescita. Gran parte delle tecniche statistiche non sono appropriate per i piccoli dataset ed è perciò necessario seguire nuove strade per identificare strutture con le poche osservazioni disponibili. Nel primo paragrafo verrà definita una nuova misura di distanza/dissimilarità tra due serie storiche. Questa produce una matrice di distanza per ciascun indicatore e la metologia Distatis, discussa nel paragrafo 1.2, servirà a condensare in una unica matrice le informazioni sull'andamento congiunturale fornite da un vasto insieme di indicatori. Nel paragrafo 1.3 verrà illustrato l'uso della clustering gerarchica aggregativa a partire dalla matrice di distanze sintetica degli indicatori. La nuova tecnica basata sulla triade Distanza tra serie storiche brevi-metododologia Distatis-Clustering gerarchica verrà applicata ai dati ricavati dal progetto "Informazione statistica territoriale e settoriale per le politiche strutturali 2001-2008i" condotto dall'Istat per sostenere l'attività di monitoraggio e valutazione del Quadro Comunitario di Sostegno 2000-2006. Tali dati saranno integrati con i dati di contabilità rivelatisi una buona base conoscitiva per la formulazione di politiche regionali nell'ambito di piani di programmazione. I conti economici regionali nascono dall'esigenza di avere un supporto statistico a livello sub-nazionale come base conoscitiva per comprendere le singole realtà regionali italiane e per sostenere ed agevolare l'attuazione di politiche di intervento pubblico volte a migliorare l'economia locale. Le stime degli aggregati macroeconomici utilizzati sono calcolate dall'ISTAT sulla base di rilevazioni proprie di dati censuari, di indagini svolte presso le famiglie e le imprese e di rilevazioni di altra fonte esterna sia privata che pubblica.

In questo contesto, la possibilità di impostare politiche regionali, di valutarne l'impatto a livello locale sui principali indicatori socio-economici, di monitorarne l'efficacia nel tempo, è fortemente condizionata dalla disponibilità di un adeguato sistema informativo a livello territoriale, che abbia caratteristiche di sistematicità, completezza, rilevanza, pertinenza e coesione sia interna sia con il quadro nazionale. L'analisi dell'aspetto territoriale dello sviluppo economico e del suo ruolo nelle dinamiche di crescita divergenti permetterà, nel quinto paragrafo, di trarre conclusioni relative al posizionamento della Calabria nelle dinamiche economiche dell'ultimo decennio. nel panorama delle regioni italiane.

1.1.1 Natura delle serie storiche

Non sono possibili costruzioni teoriche laddove non c'è ripetizione. I fenomeni socio-economici effettivamente non si ripetono se considerati solo in modo superficiale e descrittivo. Ad un esame superficiale nessun evento si ripete. Ogni fatto è unico ed è impossibile, nella realtà terrena, la sua replicazione esatta e completa: il sole che vediamo ogni giorno non è mai lo stesso e diversa è in ogni momento la luna; eppure esistono accurate tabelle delle eclissi di sole e di luna che consentono di prevedere esattamente quando questi eventi avverranno. Qui non si intende parlare della mera replicazione di un fatto ché ciò sarebbe inattuabile

ed anche inutile: ammesso che fosse possibile la riproduzione esatta di tutte le condizioni -note, sospettate e sconosciute- anche i loro effetti sarebbero costanti e verrebbe meno il motivo dell'analisi statistica. Piuttosto si intende il ripetersi di una versione scarna ed essenziale dell'avvenimento le cui caratteristiche rilevanti si mantengono intatte nella turbolenza delle varie manifestazioni. L'elemento specifico che appare nelle serie storiche è che i fenomeni sono posti in relazione con i propri valori passati. La serie storica o serie temporale è un insieme di osservazioni disposte cronologicamente per formare una traiettoria nello spazio dei possibili movimenti del fenomeno. Il significato del dato si arricchisce del fatto di essere osservato prima o dopo di un altro ed il tempo di rilevazione stabilisce anche la sequenza di acquisizione. L'ordinamento dei dati non può quindi essere alterato ed ogni raffronto tra serie storiche deve tenere conto di questo vincolo.

L'analisi delle serie storiche è rivolta alla comprensione di fenomeni che si evolvono nel tempo in maniera non deterministica. Gli obiettivi dell'analisi delle serie storiche dipende dal tipo di dati che sono disponibili. E' tuttavia possibile inquadrarli in tre scopi distinti

1. Previsione. Il modello deve fornire previsioni sull'andamento futuro del fenomeno.

2. Interpretazione. Il modello deve mettere in evidenza i legami tra i fenomeni coinvolti in forme e modi riconducibili a precise teorizzazioni.

3. Descrizione. al modello si chiede solo di rappresentare bene la realtà osservata onde distinguerla da realtà dissimili.

Nell'ambito delle serie storiche vale la pena di distinguere tra fenomeni di flusso e di stato (ovvero di *stock*). Nei primi c'è un livello iniziale e uno livello finale; ciò che si rileva in momenti prespecificati è la risultante delle forze di aumento/diminuzione, immissione/emissione, estinzione/rinnovazione che agiscono sul fenomeno da quando finisce l'unità temporale che precede a quando termine l'unità attuale. Quindi, il dato di flusso include il dato rilevato nell'unità precedente (e, indirettamente, di tutte le unità rilevate in passato) e farà parte di tutte le unità che saranno rilevate in futuro. Infatti, gli indicatori incentrati relativi ai flussi nascono come definizioni operative delle variazioni che intervengono in una fissata unità di tempo nella consistenza di un qualche fenomeno di stato. Tali fenomeni riguardano la consistenza nel determinato istante in cui si esegue la rilevazione della grandezza di interesse. La rilevabilità istantanea è infatti l'aspetto peculiare di dei fenomeni di stato o consistenza laddove i fenomeni di flusso diventano visibili solo con il decorso dell'intervallo temporale. In entrambi i casi, alla elaborazione possono arrivare come medie o totali di rilevazioni periodiche: dati annuali come aggregati di dati mensili o dati giornalieri come insieme di dati più frequenti: chiusura giornaliera di un indice di borsa come media dei *fixing* realizzati una volta al minuto nel corso della giornata.

I fenomeni di stato, per le loro caratteristiche e l'insita complessità, sono rilevati saltuariamente (annualmente, ogni cinque o dieci anni) sia per gli immaginabili costi (e rischi di errore) connessi ad una rilevazione più frequente, ma anche perché hanno una dinamica lenta, con cambiamenti contenuti che non invalidano le conclusioni anche se il loro dato non è aggiornato. Non sempre è così: i prezzi, i salari, le tariffe di alcuni beni (quali la benzina e le telefonate) variano fortemente anche per intervalli ristretti di tempo e necessitano di un monitoraggio accurato e costante. La rilevazione dei flussi avviene per unità di tempo regolari: il dinamismo di questi fenomeni è già perturbato per conto proprio senza che ci sia bisogno di aggiungere anche una durata variabile. Se il flusso è stabile, l'arco di tempo ha poca influenza. Se invece il flusso è soggetto a forti oscillazioni, allora la rilevazione per unità di tempo piccole comporta notevoli cambiamenti nella successione dei valori che invece si smussa se si estende per periodi più lunghi.

1.1.2 Variazioni relative

Il correlato empirico di diversi costrutti che più o meno direttamente sono riferiti al mutamento nel tempo è ottenuto dalla seguente impostazione: le unità sono esaminate rispetto ad un indicatore in un dato periodo; la stessa operazione è ripetuta in un periodo diverso e ci si chiede quale sia la variabile che possa esprimere i cambiamenti intervenuti tra le due circostanze. Consideriamo l'esempio riportatao nella tabella 1.1.

Tabella 1.1: Avvisi di gara in *Gazzetta Ufficiale* nel periodo 1/1-30/6/1993

Regione	1991	1992	$\{2\} - \{1\}$	$(\{2\}/\{1\}) * 100$	$(\{1\}/\{2\}) * 100$
Calabria	261.5	567.4	305.9	216.98	46.09
Campania	860.7	804.3	-56.4	93.45	107.01
Lazio	1024	498.9	-525.1	48.72	205.25
Emilia Romagna	341.2	765.8	424.6	224.44	44.55
Veneto	539.2	466.4	-72.8	86.5	115.61
Lombardia	782.6	818.1	35.5	104.54	95.66
Piemonte	481.6	1856	1374.4	385.38	25.95

Nella terza colonna è riportato lo scarto nei due periodi. A parte il segno, questi importi non producono confronti significativi: se nella prima occasione era 100 e nella seconda passa a 200 si verifica un aumento del 100%; se invece era già a $1'000$ e passa a $1'100$, l'aumento sarà solo del 10%. Se siamo interessati alla efficacia informativa dell'indicatore nel corso del tempo questa è di certo resa più evidente dalla variazione relativa.

D'altra parte, un confronto tra tra differenze assolute di serie storiche espresse in unità di misura eterogenee (pollici/centimetri, gradi Celsius/Fahrenheit) oppu-

re con campi di variazione eterogenei (variazioni del tasso di interesse e variazioni del PIL) risulterebbe scarsamente indicativo. I rapporti (detti di coesistenza) calcolati nella quarta e quinta colonna aggirano il problema della diversità di scala, ma danno solo un'indicazione indiretta: ad esempio, per il Piemonte il primo rapporto dice che nel 1992 gli importi sono pari il 385% di quelli del 1991 ovvero che questi sono il 26% di quelli del 1992. Per asseverare il cambiamento tra i due periodi occorre mentalmente ragguagliare il rapporto all'unità: se superiore ad uno indica un aumento ed indica una diminuzione se inferiore ad uno. In definitiva, i rapporti di coesistenza non rappresentano bene l'incremento o il decremento tra le due occasioni come ci riescono le variazioni assolute grazie alla presenza del segno ed alla invarianza rispetto alle traslazioni; le seconde però sono legate alla scala di misurazione della serie da cui invece prescindono i rapporti. Per arrivare ad una misura significativa della variazione è necessario coinvolgere sia le differenze assolute che i rapporti perché tanto la sensibilità alle trasformazioni di centrallità che di variabilità possono incidere sulla percezione corretta del cambiamento intervenuto nella serie storica, specialmente tra l'inizio e la fine di un trend.

Se indichiamo con $H(X_r, X_s)$ il rapporto tra le due variabili componenti osservate nel periodo r ed s, si dirà che $H(.)$ misura la variazione relativa tra i due periodi se gode delle seguenti proprietà:

$$\begin{gathered} H(X_r, X_s) = 0 \text{ se } X_r = X_s \\ H(X_r, X_s) = \text{è continua rispetto a } X_r \text{ e } X_s \\ H(X_r, X_s) < 0 \text{ se } X_r < X_s \\ H(aX_r, aX_s) = H(X_r, X_s) \text{ se } a \neq 0 \end{gathered}$$

Esistono varie scelte per la funzione $H(.)$. Eccone alcuni esempi.

$$\begin{aligned} H_1(X_r, X_s) &= \left(\frac{X_r}{X_s}\right)^{\alpha} - 1 \text{ per } X_s > 0, \alpha > 0 \\ H_2(X_r, X_s) &= \left(\frac{X_s}{X_r}\right)^{\alpha} - 1 \text{ per } X_r > 0, \alpha > 0 \\ H_3(X_r, X_s) &= \frac{X_s - X_r}{\left(0.5\left|X_s\right|^{\alpha} + 0.5\left|X_r\right|^{\alpha}\right)^{1/\alpha}} \\ H_4(X_r, X_s) &= \alpha \ln\left(\frac{X_s}{X_r}\right) \quad X_r > 0, \alpha > 0 \\ H_5(X_r, X_s) &= \frac{X_s - X_r}{\hat{\sigma}(X)}; \\ H_6(X_r, X_s) &= \frac{X_s - X_r}{\max(X) - \min(X)} \end{aligned}$$

Le più note sono la H_1 che è esatta se il tasso di crescita locale del fenomeno è lineare e la H_1 che risulta più affidabile per un tasso di crescita locale in forma

logaritmica; quest'ultima però può essere applicata solo a dati positivi e quindi potrebbe rendersi necessaria una traslazione dei valori del tipo $X' = X + \min X + 1$ che porta a zero il minimo della trasformata logaritmica.

La formula $H(.)$ è detta emisimmetrica se $H(X_r, X_s) = -H(X_s, X_r)$ cioè se scambiando di posto le osservazioni cambia solo il segno, ma non il valore della variazione relativa. Come si può controllare facilmente, le formule 3-8 sono emisimmetriche e non lo sono la 1 e la 2. L'assenza di tale caratteristica può portare a delle incongruenze. Prima di una svalutazione, il cambio Euro/Dollaro era 900 : 1000 cioè erano necessarie 900 euro per acquistare 1000 dollari. Dopo la svalutazione il cambio è 1125 : 1000. L'euro si è svalutato del 20%, ma il dollaro si è rivalutato del 25%. Come è possibile?

$$H_1 = \left(\frac{900}{1125}\right)^{\alpha} - 1 = -20\% \quad H_1 = \left(\frac{1125}{900}\right)^{\alpha} - 1 = +25\%$$

L'incongruenza sparisce se si usa una formula emisimmetrica quale la 3 con $\alpha = 1$:

$$H_3 = \frac{1125 - 900}{\left(\frac{1125 + 900}{2}\right)} = +22\% \quad H_3 = \frac{900 - 1125}{\left(\frac{1125 + 900}{2}\right)} = -22\%$$

Le formule di tipo 3 hanno un altro difetto: raffrontano lo scarto ad un valore di riferimento, ma fittizio il cui ruolo non è di comprensione immediata. La positività del valore di riferimento è necessaria per mantenere il segno negativo quando il valore più recente è inferiore a quello meno recente (tendenza decrescente) ed il segno positivo quando la tendenza è crescente ($X_r < X_s$).

Se $H(X_r + c, Y_s + c) \neq H(X_r, Y_s)$ per $c \neq 0$ la formula risulta sensibile alle traslazioni. Per superare tale problema è necessario inserire al denominatore una differenza come al numeratore, ma che sia sempre positiva. Sia la standardizzazione (formula 5) che la unitarizzazione (formula 6) producono rapporti invarianti rispetto ai cambiamenti di scala ed alle traslazioni, ma i cui valori sono compresi tra zero ed uno nel caso della formula 6 ed hanno media nulla e deviazione standard unitaria nel caso della formula 5. Non si deve peraltro trascurare il fatto che sia la deviazione standard che il campo di variazione sono stime campionarie ottenute in base ai valori già osservati e non rispetto ai valori del processo stocastico di cui è figlia la serie storica. A questo consegue una distorsione (meno accentuata per la standardizzazione che per la unitarizzazione) dei cui effetti occorre diffidare.

Non è superfluo segnalare che le variazioni di cui si è parlato in questo paragrafo hanno un carattere puntuale più che locale dato che considerano solo due termini della serie storica. Le variazioni di tendenza possono essere anche sviluppati a più di due termini. Ad esempio, un carattere meno puntuale avrebbe la valutazione dello scarto basato su tre osservazioni, anche non consecutive, della serie storica (cfr. [22]).

1.1.3 Variazioni relative medie

La rilevazione di una serie storica avviene spesso per unità che sono idealmente frazionabili in sottounità: il quinquennio in anni, l'anno in semestri o trimestri o mesi, il mese in settimane e così via; può perciò insorgere l'esigenza di misurare la variazione relativa attribuibile a ciascun sottoperiodo ovvero il tasso medio di crescita. In generale, se $y(t)$ indica il livello raggiunto dal fenomeno nel tempo t, la variazione relativa media al tempo $t + \Delta t$ potrebbe essere espressa come:

$$\frac{\frac{X(t+\Delta t) - X(t)}{\Delta t}}{X(t)} = \frac{X(t+\Delta t) - X(t)}{X(t)\,\Delta t} \tag{1.1}$$

La $X(t)$ è incognita ed occorre desumere le sue variazioni dai dati osservati. In pratica ci si limita a dividere la variazione relativa tra i due periodi per il numero di sottoperiodi.

$$\Delta_{s,r} = \frac{X_s - X_r}{\delta(s-r)} \qquad \text{con} \quad s > r \tag{1.2}$$

dove δ è un fattore positivo che, opportunamente scelto, può rendere la frazione [1.2] emisimmetrica ed invariante rispetto alle trasformazioni lineari. Il rapporto esprime il tasso di crescita locale presupponendo che ogni subperiodo incluso tra quelli a confronto abbia dato il medesimo contributo al tasso di crescita del periodo intero e che tra un subperiodo e l'altro non si siano verificate altre accumulazioni. Da un altro punto di vista, la definizione [1.2] può essere considerata una stima del coefficiente angolare della retta che passa più vicino a tutti i punti della serie storica.

Un ragionamento analogo può essere applicato alle inclinazioni [1.2] per due periodi

$$\Delta^*_{t,u} = \frac{\Delta_{t,s} - \Delta_{r,u}}{(t-u)} = \frac{\frac{X_t - X_s}{\delta(t-s)} - \frac{X_r - X_u}{\delta(r-u)}}{\delta(t-u)} = \left[\frac{X_t - X_s}{\delta(t-s)(t-u)}\right] - \left[\frac{X_r - X_u}{\delta(r-u)(t-u)}\right] \tag{1.3}$$

con $t \geq s \geq r \geq u$. Se $t - s = r - u$ allora la [1.3] approssima la derivata seconda tra curva sottostante i dati (standardizzati). L'aspetto più peculiare della curvatura di una funzione in un punto è il segno della sua derivata seconda in quel punto. Se è positiva (negativa) indica convessità (concavità) cioè è aperta verso l'alto (basso); se è nullo allora è nulla anche la curvatura in quel punto. Le funzioni lineari hanno tutte curvatura nulla in ogni loro punto. Al momento, tuttavia, non riteniamo utile coinvolgere le differenze seconde nella valutazione della distanza tra due traiettorie. D'altra parte se due serie storiche hanno coincidenti

tutte le differenze prime [1.2] avranno coincidenti anche le [1.3]; il converso non è necessariamente vero.

Il nostro studio ha considerato un caso particolare della formula [1.2]

$$H\left(X_r, X_s\right) = \begin{cases} \dfrac{X_s - X_r}{(s-r)\,\hat{\sigma}\left(X\right)} & \text{se } \hat{\sigma}\left(X\right) > 0 \\ 0 & \text{se } \hat{\sigma}\left(X\right) = 0 \end{cases} \tag{1.4}$$

La presenza del fattore correttivo $\hat{\sigma}\left(X\right)$ rende simili le variazioni di serie storiche che hanno la stessa dinamica pur in presenza di un livello medio e di un fattore di scala molto diversi. La figura che segue illustra alcuni casi di serie storiche considerate identiche per il criterio [1.4].

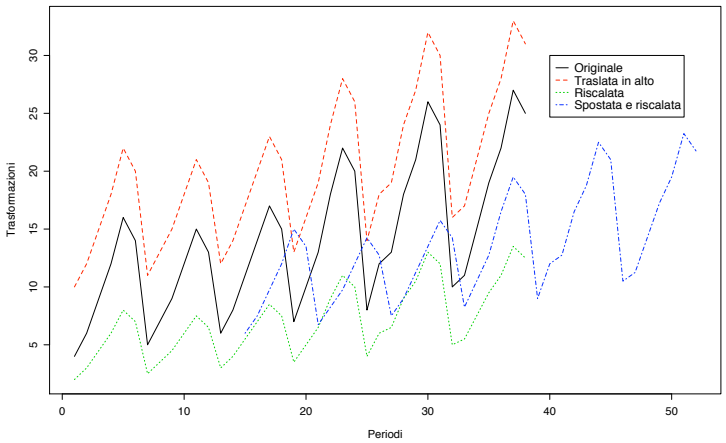


Figura 1.1: Serie storiche da considerare simili

1.1.4 Scomposizione di una serie storica

L'idea tradizionale di una serie storica è che essa sia esprimibile mediante una funzione del tempo: $f\left(t\right)$ ed una componente aleatoria $u\left(t\right)$

$$X_t = g\left[f\left(t\right), u\left(t\right)\right] \tag{1.5}$$

dove $f\left(t\right)$ è la componente sistematica o deterministica considerata effetto della interazione di vari fattori, non direttamente osservabili, ma ben definiti sul piano concettuale. La notazione $u\left(t\right)$ è la componente erratica che riassume tutto quello che non rientra in $f\left(t\right)$ o che non può essere spiegato altrimenti. Si ritiene che il comportamento di $u\left(t\right)$ sia assimilabile a degli errori accidentali non osservabili e non controllabili. La funzione $g\left(t\right)$ è scelta tra le più semplici:

$$\textit{Additivo}]\ X_t = f\left(t\right) + u\left(t\right) \qquad \textit{Moltiplicativo}]\ X_t = \left[f\left(t\right)\right]\left[u\left(t\right)\right] \tag{1.6}$$

Un modello additivo è appropriato quando l'intervento della componente erratica non influenza il livello medio della serie storica, ma agisce sommandosi ad esso. Se invece la fluttuazione erratica impatta proporzionalmente con il livello medio della serie (aumentandolo se è maggiore di uno e diminuendolo se inferiore ad uno) è appropriato il modello moltiplicativo. In comune i due schemi hanno il duplice scopo di offrire una sintesi delle componenti fondamentali della serie storica e di fornire la base per la previsione dei suoi valori futuri.

Alla base della teoria delle serie storiche vi è la convinzione che i fattori determinanti dell'andamento nel passato continuino a esercitare effetti analoghi anche nel futuro. Il principio dell'induzione, tuttavia, non può essere assunto come fondamento generale perché si rivela talvolta falso. L'aver osservato che certi eventi si sono manifestati finora in una data maniera non autorizza a dedurne che essi continueranno a presentarsi nel medesimo modo. Nulla però vieta di pensare che, fra tutte le possibili manifestazioni, la più plausibile sia proprio la loro ripetizione nei modi e nelle forme consuete. Anzi, tale metodo della previsione annunciata, è il criterio tipico adottato nelle scienze empiriche: in presenza di certe condizioni si prevede il verificarsi di un evento. Se questo succede, si ha una conferma, ma sono richiesti ulteriori accertamenti. Più accertamenti arrivano, più sicuri siamo della previsione (senza mai poter raggiungere la certezza).

E' fuori discussione che i fatti economici cambiano. Per certi elementi la variabilità è caotica e oscillatoria. Per altri è orientata in una direzione precisa ed è cumulativa. Alcuni fenomeni si modificano rapidamente, certi altri lentamente o molto lentamente. A volte la loro evoluzione è così lenta da poter essere considerati fermi (almeno per un certo periodo). Lo studio delle serie storiche propone perciò di classificare i fattori secondo la velocità di cambiamento o dei tempi di razione ad essi da parte degli operatori economici e di studiarli prima separatamente e poi nella loro interazione. Non bisogna però dimenticare che le componenti non sono che comode approssimazioni che aiutano a comprendere la dinamica del fenomeno, ma non sono il fenomeno.

E' impossibile elencare tutti i fattori. Nell'ambito del cosiddetto approccio classico, i fattori sono classificati secondo l'arco di tempo in cui esercitano i loro effetti: trend, ciclo, stagionalità, erraticità. La serie storica è la risultante di queste forze che si ritiene stiano agendo all'interno del fenomeno. Ogni qualvolta si procede ad analizzare una serie storica con l'approccio classico se ne effettua la scissione in parti astratte che non esistono se non nella realtà virtuale del modello. Si ipotizza che ogni componente possa agire in modo autonomo rispetto alle altre: tante solitudini che viaggiano su strade diverse, ma che arrivano insieme. Le singole componenti non verranno stimate tutte insieme in simultanea, ma sono ottenute una dopo l'altra, attraverso il metodo della residuazione. Questa è la naturale conseguenza dell'ipotesi che ogni componente eserciti la sua influenza in modo indipendente dalle altre.

Ai fini della nostra indagine è utile scomporre la serie storica in tutti i suoi possibili movimenti da un periodo ad un altro dell'arco di osservazione. In effetti, se i valori riscontrati nei periodi intermedi fossero mancanti oppure risultassero contaminati e quindi da escludere, rimarrebbe solo il confronto tra i dati agli estremi. Inoltre, è ben noto che per molti fenomeni l'approssimazione lineare del tasso di crescita nei confronti troppo ravvicinati è volatile; invece, la [1.4] è più stabile se applicata a ritardi maggiori. Infatti, in questo modo si tiene conto dell'influenza di una occasione non solo sulla successiva, ma su tutte le altre a seguire. L'insieme dei movimenti è condensato nella matrice

$$\mathbf{V}_X = \begin{pmatrix} v_{2,1} & 0 & 0 & \cdots & 0 \\ v_{3,1} & r_{3,2} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ v_{n,1} & v_{n,2} & v_{n,3} & \cdots & v_{n,n-1} \end{pmatrix} \tag{1.7}$$

L'elemento in posizione r, s è dato dalla [1.4]. La serie storica è così disarticolata in una matrice $(n-1, n-1), n > 1$ di tassi di variazione nella speranza di cogliere e descrivere i tratti salienti del suo comportamento e di comprenderne il meccanismo con cui si evolve nel tempo. Si può notare, *en passant*, che la mediana dei valori nella triangolare bassa di [1.7] è il coefficiente angolare della retta robusta proposta da Theil [30].

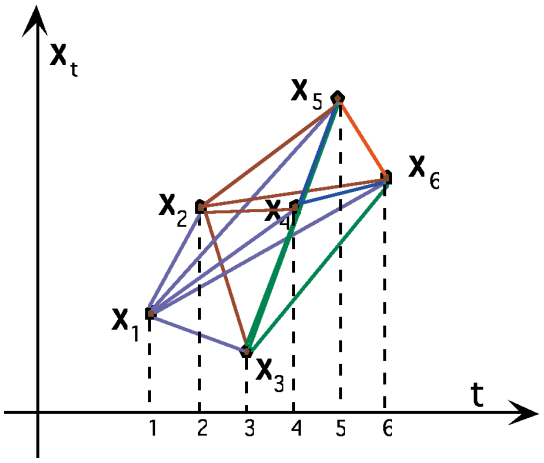


Figura 1.2: Scomposizione dei movimenti di una serie storica

Nulla vieterebbe di considerare anche la parte superiore della matrice [1.7] nonché lo scarto virtuale (nullo, ovviamente) tra un valore al periodo t e se stesso.

$$\mathbf{V}_X^* = \begin{pmatrix} 0 & v_{2,1} & v_{3,1} & \cdots & v_{n,1} \\ v_{2,1} & 0 & v_{3,2} & \cdots & v_{n,2} \\ v_{3,1} & r_{3,2} & 0 & \cdots & v_{n,3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ v_{n,1} & v_{n,2} & v_{n,3} & \cdots & 0 \end{pmatrix} \tag{1.8}$$

in cui l'ordine degli indici nel tasso non è importate dato che $v_{i,j} = v_{j,i}$. La definizione della matrice dei tassi di crescita sarebbe più completa da un punto di vista teorico, ma non per la distanza effettiva visto che si tradurrebbe in un semplice raddoppio dei termini con il solo esito del rallentamento dei calcoli.

Il processo che è sottostante la serie storica è importante e la sua modellazione un compito utile e produttivo che non può certo essere ben realizzato con la scomposzione in matrice di tassi di crescita. Tuttavia, lo scopo della nostra ricerca è l'individuazione di raggruppamenti tra le dinamiche economiche delle regioni e questo può essere deciso sulla base di elementi descrittivi, con mezzi semplici, *model-free* anche perché con pochi dati è difficilissimo distinguere un meccanismo strutturato, magari non lineare da un meccanismo del tutto erratico. Quale che sia il processo che ha generato le serie storiche (e siccome sono brevi, i potenziali processi teorici sono tantissimi) i dati osservati continuano a mostrare o a non mostrare la tendenza al raggruppamento ed è questa che ci interessa individuare e descrivere. Inoltre, riteniamo che sia piuttosto difficile trovare una classe di modelli che possa rappresentare serie storiche molto eterogenee sia come tendenza che come struttura di variabilità. Ad esempio, l'impatto della eteroschedasticità sulla matrice dei tassi di crescita non introduce difficoltà aggiuntive dato che la stima dei rapporti è già pronta a trattare questo tipo di distorsione. D'altra parte se lo scopo è di arrivare ad una matrice di compromesso che riassuma la prossimità tra serie storiche non sembra indispensabile la mediazione dei modelli.

1.1.5 Distanza tra serie storiche

La definizione della prossimità tra oggetti è un problema affrontato in molti contesti dell'analisi dei dati e nell'ambito delle decisioni in condizioni di incertezza. Se poi gli "oggetti sono manifestazioni dinamiche di un fenomeno si aggiunge la necessità di valutare la prossimità tra le loro evoluzioni (trajettorie o serie storiche). La misura della distanza tra due serie storiche può essere affrontata secondo varie angolature. In linea di massima si configurano i seguenti approcci:

1. Conversione delle trajettorie osservate in modelli statistici (ARIMA, trasformate di Fourier, etc.), stima dei loro parametri e calcolo della distanza

tra i modelli in base alla loro formulazione parametrica. (Cfr. Caiado *et al.* [5] per una panoramica) ovvero si realizza una comparazione tra i loro residui.

2. Molti indici di distanza si basano sulle autocorrelazioni incrociate

$$\rho(\tau) = \left(\frac{1}{n\hat{\sigma}_i\hat{\sigma}_j}\right) \sum_{t=1}^{n-|\tau|} (X_{i,t+\tau} - \mu_i)(X_{J,t+\tau} - \mu_j) \tag{1.9}$$

dove $\hat{\sigma}_i, \hat{\sigma}_j$ sono le stime campionarie delle deviazioni standard delle serie storiche i e j, rispettivamente. Le serie storiche spariscono per lasciare il posto alle loro autocorrelazioni dal cui confronto scaturisce la misura della loro prossimità. Da notare che le autocorrelazioni tra le serie storiche in scala originale o in logaritmica possono essere diverse da quelle applicate ai tassi di variazione relative discussi nel paragrfo 1.1.3.

3. La serie storica è sottoposta ad un processo di perequazione con delle funzioni *spline* adoperando opportuni parametri di smussamento. La prossimità tra i fenomeni è poi misurata valutando la conformazione dei profili cono degli indici di correlazione tra funzioni. (Cfr. Heckman e Zamar [15]).

4. La serie storica di n valori è considerata un punto nello spazio euclideo a n dimensioni cosicché la distanza tra due serie storiche diventa assimilabile alla distanza tra due punti. Per fare fronte alla conseguenze della dimensionalità elevate si applicano metodi per ridurre le dimensioni pur conservando le principali tendenze della serie storica.

5. Segmentazione della serie storica in subserie di uguale ampiezza e interpolazione fra i punti di ognuna di un polinomio di grado fissato con i minimi quadrati. La successiva classificazione è applicata ai residui delle subserie.

6. Scomposizione della serie storica in fattori e calcolo della dissomiglianza come sintesi delle distanze tra le componenti. In questo approccio rientra anche la frammentazione della serie storica in sottoserie ottenuta in base all'ottimizzazione di un qualche criterio.

Il primo approccio presuppone una serie di osservazioni piuttosto prolungate laddove, nell'ambito del nostro studio, sono state riscontrate serie storiche che si protraggono per un numero limitato di periodi (da 2 a 26) . Per queste numerosità è difficile definire una metrica seguendo i canoni della modellistica poiché si potrebbero stimare ben pochi parametri, ammesso che l'adattamento sia uniformemente soddisfacente per tutte le serie storiche e che la comparabilità tra le rilevazioni si mantenga inalterata per tutto l'arco delle rilevazioni. Lo stesso tipo di riserva si applica al secondo approccio. Infatti, la stima campionaria del

coefficiente di correlazione è inefficiente se i dati disponibili per il suo calcolo sono pochi come nel caso di serie storiche brevi. Il terzo approccio non può essere realisticamente applicato a serie storiche brevi. D'Urso e Vichi [8], (D'Urso [7]), in analogia con la scomposizione delle serie storiche valutano la sua articolazione in trend, velocità (differenze prime) e accelerazione (differenze seconde) e propongono una metrica che coinvolge, additivamente, le varie componenti con pesi differenziati. La nostra proposta si inquadra in questa impostazione, ma cerca di impegnare maggiormente i valori delle serie storiche brevi di cui si dispone.

Partiamo dalla rilevazione di due indicatori in forma di serie storiche di pari lunghezza n i cui valori siano presentati ordinatamente in un diagramma cartesiano in corrispondenza di ascisse equispaziate. Ogni coppia (X_t, Y_t) produce uno scarto in corrispondenza del periodo t

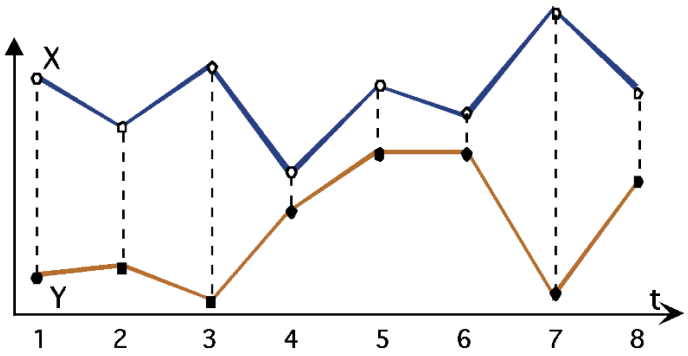


Figura 1.3: Confronto di due serie storiche

Potremmo misurare la distanza in base allo scarto complessivo tra le due serie in cui una, diciamo la Y, fornisce il termine con cui l'altra, la X, si confronta. Occorre perciò trovare un indice che quantifichi i cambiamenti da apportare alla poligonale delle Y perché si possa sovrapporre esattamente -e nella sequenza fissata- alla poligonale delle X. La sintesi dei cambiamenti si potrebbe realizzare, ad esempio, con la metrica di Minkowsky:

$$\left(\sum_{t=1}^{n} |Y_t - X_t| \right)^{1/\alpha} \tag{1.10}$$

La [1.10] esprima la distanza complessiva percorsa qualora ogni Y_t si spostasse, punto-per-punto, sulla corrispondente X_t. In diverse applicazioni la formula [1.10] ha asociato un sistema di pesi non geativi ed a somma unitaria che possa bilanciare il contributo degli scarti in cui un termine sia troppo grande o troppo piccolo rispetto all'altro.

La [1.10] ha un difetto tecnico: fornisce lo stesso valore per ogni permutazione delle coppie (X_t, Y_t) cioè qualunque sia la conformazione che una serie assume

rispetto all'altra la [1.10] darà sempre lo stesso valore se la sua morfologia è costituita dagli stessi segmenti in posizione corrispondente. Nella figura 1.4 sono riportate le due serie storiche già in figura 1.4 in cui però il punto 5 ed il punto 7 hanno scambiato posizione. La forma delle traiettorie è cambiata, ma non è cambiata la somma complessiva delle lunghezze dei segmenti.

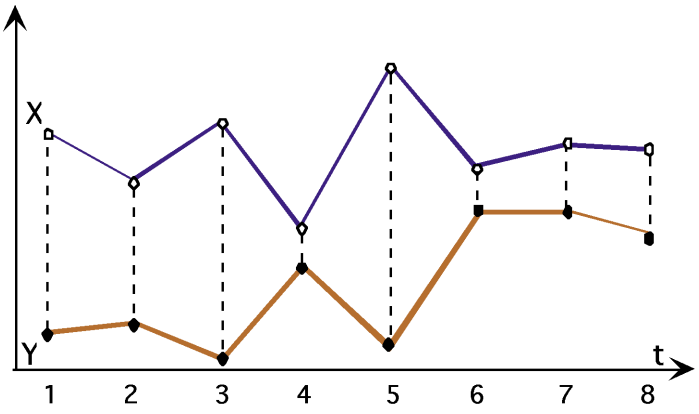


Figura 1.4: scambio di due segmenti

Idealmente, la definizione di distanza tra serie storiche dovrebbe portare alla scelta dell'indice che poi dovrebbe determinare la collocazione univoca delle serie su di una scala graduata dalla minima alla massima distanza. Diciamo subito che questo non è possibile. Ogni indice sintetico non può che essere un riassunto degli scarti e per questa sua natura si potrà verificare il caso di indici uguali in situazioni diverse come sempre succede quando si cerca di condensare in una quantità scalare, aspetti complessi e variegati. Ci sono però requisiti che è comunque opportuno siano soddisfatti e che nel corso del tempo si sono mostrati utili per una misura corretta del fenomeno. Ad esempio, in una fase preliminare della nostra ricerca abbiamo considerato la quantità $d(X,Y) = [1 - r_m(X,Y)]/2$ dove

$$r_m = \frac{\sum_{r=1,s=r+1}^{n-1} \sum_{s=r+1}^{n} [v_{r,s}(X) - \bar{v}_X][v_{r,s}(Y) - \bar{v}_Y]}{\sqrt{\sum_{r=1}^{n-1} \sum_{s=r+1}^{n} [v_{r,s}(X) - \bar{v}_X]^2 \sum_{r=1}^{n-1} \sum_{s=r+1}^{n} [v_{r,s}(Y) - \bar{v}_Y]^2}} \tag{1.11}$$

$$\bar{v}_X = \frac{\sum_{r=1}^{n-1} \sum_{s=r+1}^{n} v_{r,s}(X)}{n(n-1)/2}; \qquad \bar{v}_Y = \frac{\sum_{r=1}^{n-1} \sum_{s=r+1}^{n} v_{r,s}(Y)}{n(n-1)/2}$$

Questo indice è richiamato spesso nella valutazione della distanza tra serie storiche, ma non è privo di inconvenienti. Il coefficiente [1.11] non cambia se i valori di una serie storica o di entrambe subiscono una trasformazione lineare e questo è un vantaggio come si è detto nel paragrafo 1.1.3; tuttavia, può cogliere solo legami di dipendenza lineare o che si approssimino a questa tendenza nell'arco dei valori osservati. Infine, la [1.11] non necessariamente verifica la disuguaglianza triangolare ed è quindi escluso dalla misura della distanza.

La [1.10] ha anche un difetto logico nella sua genericità: perché l'andamento di una serie dovrebbe essere un termine di paragone per l'altra? Qual'è il significato di trasformare l'una nell'altra? Perché trascurare la possibilità di spostarsi da un punto verso tutti quelli ad esso successivi?

La nostra idea, proposta per la prima volta in letteratura, è che la dinamica di una serie storica breve possa essere descritta dalla sua matrice dei tassi di variazione [1.7]. Se le due serie storiche sono simili avranno una matrice dei tassi di crescita simile e se le matrici dei tassi di crescita di due serie storiche sono simili, le due serie storiche potranno essere considerate equivalenti ai fini delle indagini statistiche. In aggiunta, le tecniche di riduzione delle serie storiche, molto diffuse nella archiviazione e ricerca di trajettorie, può incentrarsi sulla matrice [1.7] o sue derivate usando l'analisi delle componenti principali.

Consideriamo due serie storiche X e Y, osservate nei periodi da 1 ad n le cui variazioni relative siano raccolte nelle matrici $\mathbf{V}_X$ e $\mathbf{V}_Y$. A questo punto la misura della distanza tra le due serie storiche diventa un problema di misurare la distanza tra due matrici. Esistono ovviamente moltissimi modi per misurare la distanza tra due matrici. Noi useremo la norma di Frobenius:

$$d\left(\mathbf{V}_X, \mathbf{V}_Y\right) = \sqrt{\sum_{r=1}^{n-1} \sum_{s=r+1}^{n} \left[v_{r,s}\left(X\right) - v_{r,s}\left(Y\right)\right]^2} \tag{1.12}$$

che considera lo scarto al quadrato tra tutti gli $n(n-1)/2$ tassi di variazione stabilendo la distanza tra le due serie dal punto di vista di tutti i valori osservati e che rispetta i requisiti di una norma tra matrici. In particolare

$$\begin{gathered} d\left(\mathbf{V}_X, \mathbf{V}_Y\right) = \mathbf{0} \text{ se e solo se } \mathbf{V}_X = \mathbf{V}_Y \\ d\left(\mathbf{V}_X, \mathbf{V}_Y\right) \geq \mathbf{0}; \\ d\left(a\mathbf{V}_X, a\mathbf{V}_Y\right) = |a|\, d\left(\mathbf{V}_X, \mathbf{V}_Y\right) \ \forall\ a \\ d\left(\mathbf{V}_X + \mathbf{V}_Y\right) \leq d\left(\mathbf{V}_X\right) + d\left(\mathbf{V}_Y\right) \\ d\left(\mathbf{V}_X \mathbf{V}_Y\right) \leq d\left(\mathbf{V}_X\right) d\left(\mathbf{V}_Y\right) \end{gathered}$$

Poiché la metrica $d\left(X, Y\right)$ è ottenuta confrontando di tutti i possibili tassi di variazione ci si aspetta che possa quantificare in modo attendibile la vicinanza/lontananza tra le due serie storiche. Lin *et al.* [] considerano una metrica estesa a tutte le possibili coppie degli n valori, ma compattano il confronto in soli quattro possibili valori della legati alla concordanza/discordanza dello scarto tra il periodo r ed il periodo s delle sue serie. Möller-Levet *et al.* [23] adoperano un caso speciale della [1.12] limitate alle coppie con un solo periodo di ritardo.

E' da osservare che la [1.4] assegna peso $(s-r)^{-1}$ allo scarto tra i corrispondenti tassi di variazione; ne risulta una ponderazione che da risalto maggiore agli scarti ravvicinati rispetto agli scarti osservati per periodi distanziati e questo senza badare alla entità dello scarto. Se appare legittimo frazionare l'evoluzione

della singola serie storica nei sottoperiodi che intercorrono tra quelli a confronto, sussistono però dubbi sul fatto che il peso che da ciò deriva sia coerente con l'ammontare di aumento o diminuzione che si è realizzato. Nel comparare due serie storiche sorge naturale la ricerca di co-presenze di picchi e valli e quindi di dare più rilevanza alle impennate ed alle cadute simultanee rispetto ai periodi di stasi. Questa pratica può essere inserita nella misura della distanza tra serie storiche. A questo risultato si può pervenire dando un peso diverso ad ognuno dei tassi inseriti nella [1.12]. In via preliminare definiamo una quantità che sia legata all'entità dello scarto e che assuma valori in un intervallo definito.

$$w'_{r,s}(X) = \left[\frac{v_{r,s}(X)}{\max|v_{r,s}(X)|}\right]; \quad w'_{r,s}(Y) = \left[\frac{v_{r,s}(Y)}{\max|v_{r,s}(Y)|}\right] \tag{1.13}$$

Le espressioni in [1.13] variano nell'intervallo $[-1, 1]$. Sono nulle se i due termini della serie storica sono uguali; raggiungono l'estremo inferiore in caso del massimo decremento nel periodo e quello superiore in corrispondenza del massimo incremento. Il sistema di pesi si può avvalere di una trasformazione che riduce l'effetto dello scarto tra le variazioni della X e della Y se queste sono concordanti e lo aumenta per le discordanti e che inoltre l'aumento e la diminuzione siano correlate positivamente con l'entità dello scarto. Non riteniamo utile, almeno in questa ricerca, distinguere tra co-movimenti di crescita e co-movimenti di declino per cui i pesi devono essere simmetrici. Lo schema presentato nella tabella 1.2 risponde allo scopo.

Tabella 1.2: Sistema di ponderazione degli scarti tra variazioni relative

V_x/V_y	-1.0 -0.8	-0.8 -0.6	-0.6 -0.4	-0.4 -0.2	-0.2 0.0	0.0 0.2	0.2 0.4	0.4 0.6	0.6 0.8	0.8 1.0
-1.0 -0.8	1	8	9	10	11	12	13	14	15	16
-0.8 -0.6	8	2	8	9	10	11	12	13	14	15
-0.6 -0.4	9	8	3	8	9	10	11	12	13	14
-0.4 -0.2	10	9	8	4	8	9	10	11	12	13
-0.2 0.0	11	10	9	8	5	8	9	10	11	12
0.0 0.2	12	11	10	9	8	5	8	9	10	11
0.2 0.4	13	12	11	10	9	8	4	8	9	10
0.4 0.6	14	13	12	11	10	9	8	3	8	9
0.6 0.8	15	14	13	12	11	10	9	8	2	8
0.8 1.0	16	15	14	13	12	11	10	9	8	1

I pesi attributi alle varie partizioni dell'intervallo $[-1, 1]$ sono tutti positivi e compresi nell'intervallo $(0, 16)$. Dovranno essere poi resi a somma unitaria in base alla effettiva utilizzazione.

La misura di distanza tra serie storiche che proponiamo è la norma ponderata di Frobenius applicata allo scarto tra matrici dei tassi di variazione.

$$d(X, Y) = \sqrt{\sum_{r=1}^{n-1}\sum_{s=r+1}^{n} w_{r,s}(X, Y)\left[v_{r,s}(X) - v_{r,s}(Y)\right]^2} \tag{1.14}$$

La distanza [1.14] coinvolge tutte le possibili inclinazioni dei segmenti delle due serie storiche attribuendo un apposito peso ad ogni confronto. Sarà quindi valido aiuto per distinguere le somiglianze effettive da quelle apparenti anche per serie storiche brevi. Inoltre, la 1.14] è molto realistica dal punto di vista economico perché valuta i tassi medi di crescita ponderati per periodi non lunghi ragguagliando scarti fra fenomeni che nel frattempo non hanno, in genere, subito importanti cambiamenti.

E' evidente che la presenza di un valore anomalo in una delle serie storiche ne produca l'allontanamento dalle altre, ma con serie storiche brevi non è possibile stabilire con obiettività quando un valore, legittimo dal punto di vista della rilevazione, sia da considerarsi anomalo rispetto al contesto, data la fragilità del contesto. La [1.14] ha infine la caratteristica, in genere giudicata positiva, di ritenere distanti una serie storica e la sua ritardata (a parità di lunghezza).

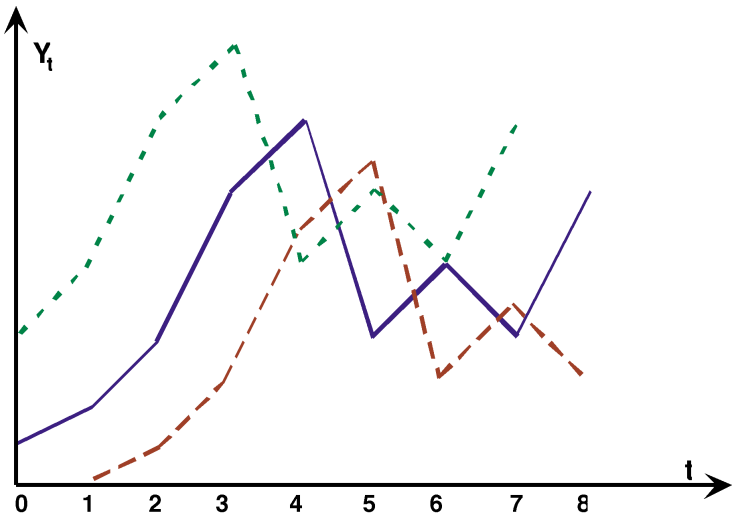


Figura 1.5: Confronto tra una serie storica e sue ritardate

La serie storica con linea continua costituisce la tendenza di riferimento. La curva tratteggiata ha la suastessa conformazione, ma inizia una periodo dopo; la curva punteggiata deriva pure da quella centrale, ma ha un valore anticipato. Poiché la misura di distanza [1.14] confronta i tassi di variazione periodo per periodo non è in grado di accorgersi che la loro struttura è simile sebbene traslata di uno o più periodi.

1.2 Metodologia Distatis

I divari regionali in Italia si presentano particolarmente articolati e differenziati segnalando percorsi di crescita non sequenziali, ma alternativi (almeno in parte) dando luogo a situazioni di benessere sostanzialmente equivalenti, ma a stadi diversi dei sentieri tradizionali dello sviluppo economico. Lo scopo di questo paragrafo è di impostare una analisi dinamica dei differenziali regionali sulla base di indicatori espressi in forma di serie storica e condensati in diverse matrici di distanze tra regioni. La metodologia utilizzata per sintetizzare le diverse matrici di distanza riprende lo studio di Abdi *et al.* [1] predisposto per la sintesi di alcuni studi sulla valutazione della prossimità tra volti umani.

La tecnica Distatis è una generalizzazione della procedura classica di *scaling* multidimensionale il cui scopo è di analizzare una sola matrice di distanze. La base operativa della procedura è attribuita a Escouffier [9] (si veda anche Lavit *et al.*, [21], e Korth e Tucker, [19]).

L'obiettivo della tecnica Distatis invece è l'analisi di un insieme di matrici di distanze che vengono fatte confluire in una matrice di sintesi alla quale potranno essere applicate le usuali tecniche di scaling e/o clustering. La regione j dell'insieme $\{1, 2, \cdots, p\}$ è descritta dalle serie storiche brevi

$$X_{t,i}(j), t = 1, 2, \cdots, n_i, \quad i = 1, 2, \cdots, k$$

dove k denota il numero di indicatori rilervati sulle unità ed n_i è la durata della serie i-esima, potenzialmente diverse per ogni serie.

La prima intuizione originale del nostro lavoro è stata di trasformare le k serie storiche in altrettante matrici più agevoli da gestire in ambito multivariato. Per l'i-esimo indicatore (cioè per l'i-esima serie storica) calcoliamo la distanza ponderata di Frobenius fra le due generiche regioni j_1, j_2

$$d\left[X_i(j_1), X_i(j_2)\right] = \sqrt{\sum_{r=1}^{n_i-1} \sum_{s=r+1}^{n_i} w_{r,s}\left[X_i(j_1), X_i(j_2)\right]\left[v_{r,s}(i, j_1) - v_{r,s}(i, j_2)\right]^2} \quad (1.15)$$

Ripetendo il calcolo della [1.15] per tutte le coppie di regioni otteniamo la matrice delle distanze relativo all'-esima serie storica.

$$\mathbf{D} = \begin{bmatrix} 0 & d_{2,1} & d_{3,1} & \cdots & d_{p,1} \\ d_{2,1} & 0 & d_{3,2} & \cdots & d_{p,2} \\ d_{3,1} & d_{3,2} & 0 & \cdots & d_{p,3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{p,1} & d_{p,2} & d_{p,3} & \cdots & 0 \end{bmatrix} \quad (1.16)$$

La $\mathbf{D}_i$, per costruzione, è simmetrica e definita positiva. Inoltre, poiché ogni suo elemento è frutto del calcolo di una metrica, la $\mathbf{D}_i$ è una matrice metrica nel senso di Gower e Legendre [13]. Ogni indicatore produce una matrice di distanze di ordine (pxp) che fornisce il suo contributo alla classificazione delle unità. Il dataset può essere visualizzato come uno strato di matrici.

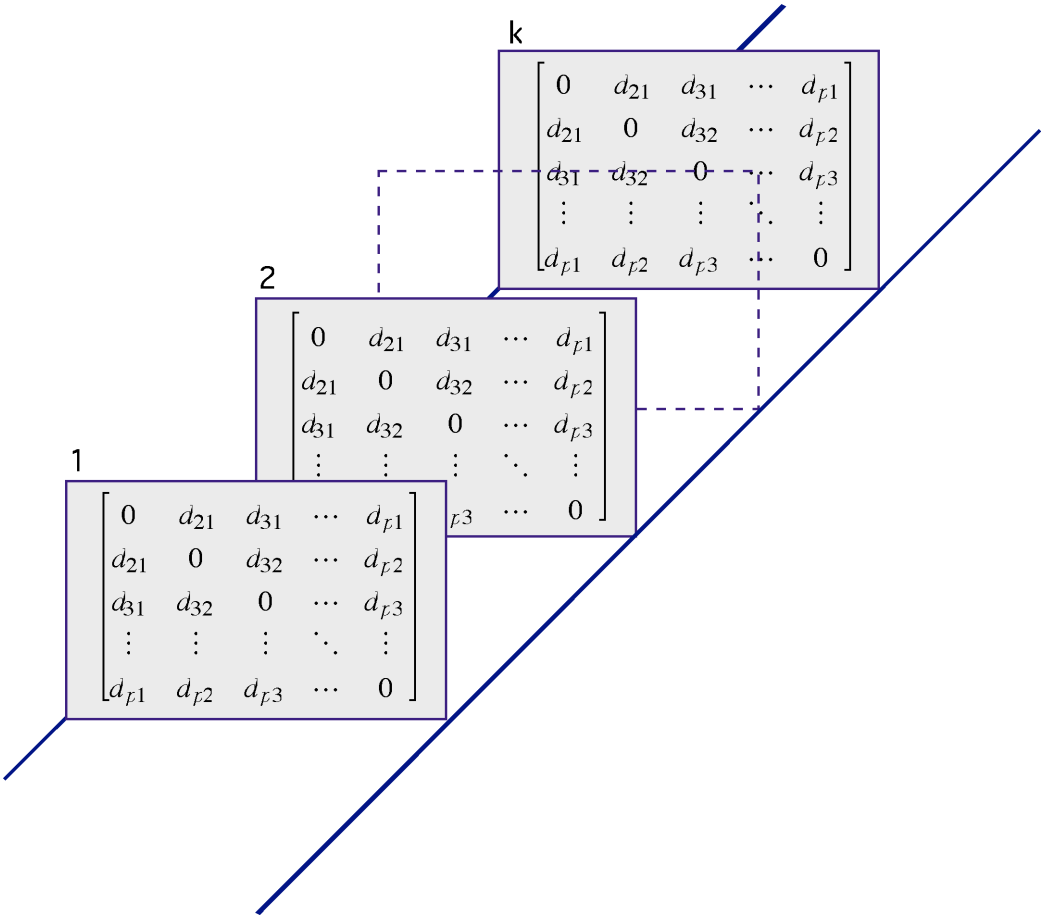


Figura 1.6: Descrizione della distanza tra unità in base a serie storiche

Ogni strato è relativo ad un indicatore ed ogni cella all'interno dello strato è la distanza tra le due serie storiche dell'indicatore rilevato sulle regioni che si incrociano nella cella. Ne consegue che per ogni coppia di regioni si formano k distanze, una per ogni indicatore, derivata dal confronto punto-punto delle corrispondenti serie storiche. In pratica diventa così possibile considerare le interazioni tra i diversi indicatori considerati e tutti i loro effetti ai fini della segmentazione delle regioni.

1.2.1 Trasformazione delle matrici di distanze

Lo scopo del nostro studio è di inglobare in un confronto unico tutte le k matrici $\mathbf{D}_i$ che diciamo parziali perché relative ad uno specifico indicatore e non al complesso degli indicatori. Lo strumento più agevole per realizzare questo obiettivo è il calcolo degli autovalori ed autovettori delle matrici parziali ovvero la loro scomposizione in valori singolari per poi cercarne una matrice sintesi o di compromesso. Questa analisi però non verrà condotta direttamente sulle matrici di tipo [1.16] perché appaiono meno suscettibili di sintesi a causa della presenza di valori nulli sulla diagonale. L'idea generale del metodo Distatis messo a punto da Abdi *et al.* [1] è di trasformare le matrici di distanze in matrici di varianze-covarianze per ciascuno degli indicatori e successivamente tentarne una sintesi. A questo fine si può sfruttare una relazione ben nota nell'ambito dello *scaling* multidimensionale (cfr. Young e Householder [31]) e che risulta molto efficace per la scomposizione delle $\mathbf{D}_i$.

Sia $\boldsymbol{\beta}$ un vettore di pesi a somma unitaria

$$\beta_i \geq 0, \qquad \sum_{i=1}^{p} \beta_i = 1 \tag{1.17}$$

Ogni β_i esprime l'importanza da attribuire all'unità i-esima all'interno nello schema delle distanze tra le unità. In prima approssimazione diamo a tutte le regioni lo stesso peso: $\beta_i = 1/p, \qquad i = 1, 2, \cdots, p$. La scelta è soggettiva. In verità, la presenza della regione Val d'Aosta, per le sue peculiarità, crea spesso difficoltà nelle elaborazioni su base regionale. Dalle analisi territoriali sono spesso escluse le due regioni più piccole, Val d'Aosta e Molise, per l'eccessiva instabilità dei dati disaggregati per cui non sarebbe forse sbagliato attribuire loro un peso diverso rispetto alle altre; ma con quale sistema? Se gli indicatori fossero tutti correlati con la popolazione si potrebbe utilizzare come peso la quota parte di popolazione residente al 31 dicembre trascorso più vicino all'epoca degli indicatori. Altre scelte sono possibili. La diversa ponderazione delle regioni è un argomento che presenta differenti ottiche interpretative con contorni sfumati e quadri descrittivi controversi. Riteniamo quindi opportuno rinviare ad ulteriori ricerche la definizione dei pesi delle unità.

Il vettore [1.17] serve a definire una matrice di centramento

$$\mathbf{C} = \mathbf{I} - \mathbf{u}\boldsymbol{\beta}^t = \mathbf{I} - \frac{1}{p}\mathbf{u}\mathbf{u}^t \tag{1.18}$$

dove $\mathbf{u}$ è un vettore (p x 1) formato da soli uno. La matrice [1.18] ha ordine (pxp), è simmetrica e tale che $\mathbf{u}^t\mathbf{C} = \mathbf{0}; \qquad \mathbf{C}^t\mathbf{C} = \mathbf{C}\mathbf{C}^t = \mathbf{C}$. Inoltre, l'operazione $\mathbf{X}^* = \mathbf{C}\mathbf{X}$, determina delle nuove variabili $\mathbf{X}^*$ la cui somma (ovvero la la media aritmetica di colonna) è pari a zero:

$$\mathbf{u}^t\mathbf{X}^* = \mathbf{u}^t\mathbf{C}\mathbf{X} = \mathbf{0}.$$

La proprietà della matrice **C** più interessante ai fini del metodo Distatis è il suo coinvolgimento nella operazione seguente:

$$\mathbf{S}_i = -\frac{1}{2}\mathbf{C}\mathbf{D}_i\mathbf{C}^t \tag{1.19}$$

La [1.19] è una matrice quadrata di ordine (pxp), simmetrica, definita (o semi-definita) positiva e con elementi sulla diagonale postivi o nulli (è anche detta del tipo di varianze-covarianze perché è riconducibile al prodotto di una matrice per la sua trasposta).

La $\mathbf{S}_i$ contiene le stesse informazioni della matrice di distanze $\mathbf{D}_i$. Infatti, quest'ultima può essere facilmente ricostruita proprio a partire dalla $\mathbf{S}_i$.

$$\mathbf{D}_i = \mathbf{s}_i\mathbf{u}^t + \mathbf{u}\mathbf{s}_i^t - 2\mathbf{S}_i \tag{1.20}$$

dove $\mathbf{s}_i$ è un vettore (p x 1) contenente gli elementi della diagonale di $\mathbf{S}_i$. L'analisi degli autovalori ed autovettori di $\mathbf{S}_i$ è equivalente a quella effettuabile sulla matrice $\mathbf{D}_i$. La ricostruzione [1.20] è unica rispetto alle trasformazioni additive delle coordinate incluse nella matrice $\mathbf{S}_i$ grazie all'effetto di centramento, ma non è unica rispetto a loro rotazioni ortogonali anche se questo più che un problema, costituisce un'opportunità che sarà adeguatamente sfruttata.

La [1.19] deve essere applicata a tutte le k matrici di distanza parziali; tuttavia, la possibilità di una struttura di variabilità disomogenea tra i vari indicatori suggerisce di normalizzare le varie matrici $\mathbf{S}_i$ dividendone gli elementi per l'autovalore massimo $\lambda_{1,i}$ (che è positivo, come vedremo).

$$\mathbf{T}_i = \lambda_{1,i}^{-1}\mathbf{S}_i \tag{1.21}$$

Grazie a questa operazione, l'autovalore massimo di $\mathbf{T}_i$ è pari ad uno per ogni $i = 1, 2, \cdots, k$. Il problema è ora di ottenere una matrice sintesi delle k matrici parziali. E' ovvio che se $k = 1$ allora l'analisi dei gruppi potrà direttamente essere condotta sulla D_1 senza dover fare ricorso alla metodologia Distatis.

1.2.2 Matrice sintesi delle distanze parziali

Le informazioni contenute negli indicatori consentono di condurre una analisi sui vari aspetti per affrontare gli interrogativi e misurare i risultati dell'intervento pubblico sul territorio. A priori, non c'è ragione di ritenere che si tratti di dati equivalenti, a meno di fluttuazioni casuali. Se così fosse, la matrice sintesi delle matrici di distanze parziali potrebbe essere ottenuta generando ogni elemento come media aritmetica semplice dei corrispondenti elementi delle varie matrici. E' chiaro come questa procedura sia in genere inappropriata se gli indicatori non possono essere considerato sullo stesso piano. Se però debbono essere ponderati con pesi diversificati, qual'è il sistema più adatto? A tale domanda possiamo dare una risposta più organica rispetto alla ponderazione delle unità.

Per semplificare l'analisi congiunta delle matrici di distanza parziali tra le unità conviene effettuarne la vettorizzazione cioè convertire ogni $\mathbf{T}_i$ un vettore $\boldsymbol{\tau}_i$ di dimensione (p^2 x 1) formato impilando l'una sull'altra le colonne di $\mathbf{T}_i$ (ovvero le righe, dato che si tratta di matrici simmetriche) ovvero si pone $\boldsymbol{\tau}_i = vec\,(\mathbf{T}_i)$. Se poi si accostano le k diverse vettorizzazioni si ottiene una nuova matrice di dimensione (p^2 x k)

$$\mathbf{Z} = [\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \cdots, \boldsymbol{\tau}_k] \tag{1.22}$$

La matrice $\mathbf{Z}$ racchiude tutte le informazioni sulle distanze parziali derivate dai k indicatori. Con uno sforzo di immaginazione, possiamo considerare ciascuna colonna di $\mathbf{Z}$ una "variabile" (con media nulla) ed ogni sua riga un punto nello spazio euclideo k-dimensionale. Ogni $\boldsymbol{\tau}_i$ induce un proprio schema di classificazione delle regioni ed occorre trovare una matrice delle somiglianze intermedia, di sintesi o di compromesso tra quelle disponibili che ne raccolga gli elementi comuni. Sembra naturale pensare ad una media ponderata delle varie matrici parziali in cui vi sia un sistema di pesi ottimale (cioè tale da favorire le concordanze e di minimizzare le discordanze tra ciò che emerge dai diversi indicatori) che riesca a cementare in un fatto unico le informazioni derivate dai vari indicatori.

In breve si tratta di risolvere il problema di minimo

$$\min_{\boldsymbol{\lambda},\mathbf{T}} \quad \sum_{i=1}^{k} \parallel \mathbf{T}_i - \lambda_i \mathbf{T} \parallel^2 \tag{1.23}$$

Come è ben noto (cfr. Greenacre [14], appendice) la soluzione è ottenuta in modo semplice ed elegante realizzando la scomposizione in valori singolari della matrice $\mathbf{Z}$ cercando un sottospazio che si adatti quanto più possibile (nel senso dei minimi quadrati) alla matrice originale. Come già avvenuto per la distanza tra le matrici dei tassi di variazione, la relazione [1.23] adotta la norma (non ponderata, ma standardizzata) di Frobenius per misurare la divergenza tra le matrici parziali.

$$\mathbf{A} = \left(\mathbf{N}^{-0.5}\mathbf{Z}\right)^t \left(\mathbf{Z}\mathbf{N}^{-0.5}\right) \tag{1.24}$$

dove $\mathbf{N}$ è una matrice diagonale i cui elementi sono tratti dalla diagonale della matrice di varianze-covarianze ($\mathbf{Z^t Z}$). La [1.24] è la matrice (k x k) che esprime i prodotti scalari tra le matrici di distanze parziali (vettorizzate). Si può inoltre notare che la matrice $\mathbf{A}$ risulta invariante rispetto alle traslazioni ed alle modifiche globali di scala.

Concentriamoci ora sull'elemento generico di tale matrice:

$$a_{ij} = \frac{\boldsymbol{\tau}_i^t \boldsymbol{\tau}_j}{\parallel \boldsymbol{\tau}_i \parallel \parallel \boldsymbol{\tau}_j \parallel} = cos\,(\boldsymbol{\tau}_i, \boldsymbol{\tau}_j) \tag{1.25}$$

che varia nell'intervallo $[0, 1]$. Questo non è subito evidente. A questo fine riscriviamo in numeratore [1.25] nei termini originali delle matrici vettorizzate

$$\boldsymbol{\tau}_i^t \boldsymbol{\tau}_j = Tr\left(\mathbf{T}_i^t \mathbf{T}_j\right) = \frac{Tr\left[\left(\mathbf{C}\mathbf{D}_i\mathbf{C}^t\right)\left(\mathbf{C}\mathbf{D}_j\mathbf{C}^t\right)\right]}{4\lambda_{1,i}\lambda_{1,j}} = \frac{Tr\left[\left(\mathbf{C}\mathbf{D}_j\right)\left(\mathbf{C}\mathbf{D}_i\right)\right]}{4\lambda_{1,i}\lambda_{1,j}} \tag{1.26}$$

Come si vede, la [1.25] è la traccia del prodotto tra due matrici simmetriche e non negative (le matrici di distanza parziali $\mathbf{D}_i$ e $\mathbf{D}_j$ centrate in modo che $\mathbf{CT}_i$ e $\mathbf{CT}_j$ abbiano media nulla. Ne consegue che la [1.25] può essere scritta come

$$a_{ij} = \frac{Tr\left[\left(\mathbf{C}^t\mathbf{D}_j\right)\left(\mathbf{C}^t\mathbf{D}_i\right)\right]}{\sqrt{Tr\left(\mathbf{C}^t\mathbf{D}_j\mathbf{C}\right)Tr\left(\mathbf{C}^t\mathbf{D}_i\mathbf{C}\right)}} = R_v\left(\mathbf{D}_i, \mathbf{D}_j\right) \tag{1.27}$$

La espressione in [1.27] è il coefficiente di correlazione vettoriale tra $\mathbf{D}_i$ e $\mathbf{D}_j$. Tale indice, proposto da Robert e Escoufier [27] riporta al numeratore la somma dei quadrati delle covarianze tra le colonne componenti le due matrici a confronto per cui, per costruzione, varia tra zero ed uno. Il primo estremo indica che la sequenza delle distanze riscontrata tra le unità in base agli indicatori i e j è l'una ortogonale all'altra cioè non hanno niente in comune. L'altro estremo si raggiunge man mano che una delle matrici può essere espressa come trasformazione affine dell'altra ovvero c'è esatta sovrapposizione tra le distanze espresse dai due indicatori rappresentati da$\mathbf{D}_i$ e $\mathbf{D}_j$. Vedremo nel prossimo paragrafo perché è così importante assicurarsi che la matrice $\mathbf{A}$ sia positiva.

La scomposizione in valori singolari della matrice $\mathbf{A}$ produce: $\mathbf{A} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^t$ con $\mathbf{P}^t\mathbf{P}$. Supponiamo che $\mathbf{A}$ sia una matrice positiva (cioè costituita da elementi maggiori di zero). Per il teorema di Perron (cfr. Gantmacher, [12] p. 53), l'autovalore massimo λ_1 di $\mathbf{A}$ è reale, semplice, compreso tra la somma di riga minima e massima ed ha associato un autovettore $\mathbf{p}_1$ con elementi tutti dello stesso segno (che possono essere resi positivi moltiplicandoli per -1 senza perdere alcuna generalità). La possibilità di un primo autovettore formato da elementi dello stesso segno è possibile anche con un matrice che non sia positiva. Se $\mathbf{A}$ non è positiva e il suo primo autovettore non ha pesi tutti positivi (o tutti negativi) si può ricorrere alla programmazione quadratica che imponga un primo autovettore con elementi di uguale segno.

Nella prima componente principale si riflettono, chi più chi meno, tutte le matrici vettorizzate di distanza cioè tutti gli indicatori ed è perciò da considerare un fattore ad essi "trasversale" che misura pertanto la dimensione complessiva della distanza attraverso la somma ponderata di tutte le matrici parziali derivate dai singoli indicatori. Ogni elemento del primo autovettore è la proiezione di ciascuna matrice $\mathbf{T}_i, i = 1, 2, \cdots, k$ parziale su tale autovettore. Il primo autovettore, per la sua trasversalità, esprime ciò che vi è di comune nei sistemi di distanze generate dagli indicatori nel corso del tempo. Più una matrice di distanza parziale è prossima alle altre (ovvero ha elementi che più si avvicinano alla media aritmetica di tutti gli elementi delle altre matrici in posizione corrrispondente) maggiore sarà il suo peso. Se gli elementi di $\mathbf{p}_1$ fossero tutti uguali allora le matrici di distanza parziali darebbero informazioni non sovrapponibili e sarebbero da considerare tutte egualmente utili a formare la classificazione delle unità. Questo però avviene anche se la matrice $\mathbf{A}$ ha righe con somma unitaria oltre che costituite da valori positivi. Infatti, in questo caso la $\mathbf{A}$ si configurerebbe come una matrice stocastica con autovalore massimo unitario e il cui autovettore

associato ha elementi uguali in valore oltre che in segno a garantire la presenza bilanciata di tutti gli indicatori anche se nulla di tutto questo è presente nei dati. Se fossero invece le matrici di distanze parziali ad essere uguali (a meno di una trasformazione lineare), allora ne basterebbe una sola per descrivere la struttura delle distanze e solo un elemento di $\mathbf{p}_1$ sarà diverso da zero.

La seconda componente è un fattore "bipolare" che, grazie alla ortogonalità con il primo, esprime le contrapposizioni che rimangono tra due gruppi principali di matrici parziali una volta che si sia eliminata l'influenza del fattore dimensionale. Si tratta di una presenza importante, ma che non ha ancora trovato una collocazione nella procedura Distatis. Non deve però sfuggire la premessa alla procedura stessa e che cioè fra le matrici di distanze parziali associate ai vari indicatori sussistano relazioni di tipo lineare. Se questa ipotesi è eccessiva si può ricorrere ad un coefficiente di correlazione vettoriale non parametrico basato sui ranghi delle distanze (all'interno di ciascuna delle matrici di distanze parziali) quali gli indici di Spearman, Kendall, Blest, Mangano (si veda Tarsitano [29].

A questo punto è naturale adottare il primo autovettore come base per lo schema di ponderazione delle matrici di distanza parziali. In genere gli autovettori sono calcolati in modo che la somma dei quadrati dei loro elementi sia pari ad uno. Per rendere unitaria la loro somma semplice, dobbiamo riscalare gli elementi di $\mathbf{p}_1$ per assicurare un corretto sistema di pesi.

$$\alpha = \left(\frac{1}{\mathbf{u}^t \mathbf{p}_1} \right) \mathbf{p}_1 \tag{1.28}$$

La matrice sintesi delle distanze parziali può quindi essere definita come

$$\mathbf{T} = \sum_{i=1}^{k} \alpha_i \mathbf{T}_i \tag{1.29}$$

Questa è la matrice compromesso ottenuta dall'aggregazione delle matrici parziali in cui il collante è fornito dai pesi [1.28]. Ogni elemento di $\mathbf{T}$ è una media ponderata dei corrispondenti elementi delle matrici $\mathbf{T}_i$ ed è compreso tra il valore minimo e massimo che si riscontra in queste ultime. Ricordiamo che questo schema di ponderazione serve a perequare il contributo delle matrici parziali ovvero degli indicatori, ma deve essere tenuto distinto dallo schema di pesi delle unità $\boldsymbol{\beta}$ di cui si è accennato all'inizio del paragrafo e che ha uno scopo diverso. La [1.29] può essere considerata la matrice più correlata con le matrici parziali $\mathbf{T}_i, i = 1, 2, \cdots, k$.

E' possibile quantificare la rappresentatività della matrice compromesso rapportando il primo autovalore al totale degli autovalori (in percentuale)

$$q = \left(\frac{100\lambda_1}{\sum_{i=1}^{k} \lambda_i} \right) = \frac{\lambda_1}{tr\,(\Lambda)} = \frac{\lambda_1}{k} \tag{1.30}$$

Maggiore è q migliore sarà la rappresentatività della matrice sintetica. In genere, la rappresentatività può essere considerata plausibile se $q > 100k^{-1}$ cioè il coefficiente [1.30] è almeno superiore alla percentuale media spiegata da un indicatore qualsiasi quando tutti gli indicatori si trovano su di un piano di parità.

La matrice delle distanze sintetica si ottiene ricostruendo la matrice di distanze a partire dalla matrice di varianze-covarianze [1.29]

$$\mathbf{D} = \boldsymbol{\theta}\mathbf{u}^t + \mathbf{u}\boldsymbol{\theta}^t - 2\mathbf{T} \tag{1.31}$$

con $\boldsymbol{\theta} = diag\,(\mathbf{T})$. E' su questa matrice che verrà condotta l'analisi gerarchica dei gruppi nel paragrafo 1.3.

1.2.3 Rappresentazione di unità e indicatori

La matrice di sintesi contiene i tratti comuni della distanza tra le unità. Contiene inoltre informazioni sul ruolo degli indicatori che l'hanno costituita, Infatti, la [1.29], in quanto espressa come matrice di varianze-covarianze può essere sottoposta a scomposizione in valori singolari

$$\mathbf{T} = \mathbf{H}\boldsymbol{\Psi}\mathbf{H}^t \qquad \text{con} \quad \mathbf{H}^t\mathbf{H} = \mathbf{I} \tag{1.32}$$

Da questa relazione otteniamo gli pseudo punteggi fattoriali corrispondenti alle unità

$$\mathbf{F} = \mathbf{T}\mathbf{H}\boldsymbol{\Psi}^{0.5} \tag{1.33}$$

Ogni riga della matrice $\mathbf{F}$ rappresenta una unità (regione). Lo scatterplot delle prime due componenti esprime l'approssimazione (a due dimensioni) dei dati relativi alle unità e può dare un'idea dei loro raggruppamenti.

Per ottenere una rappresentazione analoga con gli indicatori possiamo usare le stesse matrici viste per le regioni. In particolare

$$\mathbf{G} = \mathbf{T}\left[\mathbf{H}\boldsymbol{\Psi}^{-0.5}\right] \tag{1.34}$$

La matrice $\left[\mathbf{H}\boldsymbol{\Psi}^{-0.5}\right]$ è un operatore che proietta la matrice di varianze-covarianze di sintesi nello spazio dei fattori comuni degli indicatori. Lo scatterplot delle prime due componenti può dare indicazioni chiarificatrici sulle relazioni tra gli indicatori. La stessa operazione può essere realizzata cono ciascuna delle matrici varianze-covarianze ottenuta per ogni indicatore.

$$\mathbf{G}_i = \mathbf{T}_i\left[\mathbf{H}\boldsymbol{\Psi}^{-0.5}\right] \tag{1.35}$$

Ogni unità può essere descritta dalle coordinate delle prime due componenti della $\mathbf{G}_i$ e questo per tutte le k matrici di distanze parziali. La posizione delle unità nella matrice di sintesi è interpretabile come il baricentro delle posizione che l'unità occupa in ciascuna delle matrici di distanze. Se gli indicatori sono pochi e poche le unità si può ottenere un grafico molto suggestivo (cfr. Abdi *et al.* [1]).

1.2.4 Sintesi delle macrodeterminanti

Le analisi territoriali coinvolgono un numero elevato di variabili e per varie ragioni: perché si vuole limitare il rischio di omettere aspetti importanti (nulla deve essere trascurato o disperso), perché talune caratteristiche si articolano naturalmente su indicatori diversi, perché si voglio estrarre tutti i dati utili da una osservazione che potrebbe non essere ripetibile ed infine per la natura prettamente multidisciplinare di questo tipo di investigazione che deve valutare nel modo più completo le peculiarità delle unità oggetto di studio. D'altra parte un eccesso di misurazioni implica non solo difficoltà di calcolo, ma complica le procedure di stima e di interpretazione dei risultati senza contare i deleteri effetti dovuti alla maledizione della maledizione della dimensionalità secondo la quale anche la duplicazione esatta di un entità diventa distinguibili se si considera un numero di variabili abbastanza grande. E' evidente quindi la necessità di procedere alla scelta di quante e quali variabili utilizzare in effetti.

Anania e Tarsitano, [2] tracciano una precisa linea di demarcazione tra due procedure alternative. Una linea investigativa, detta della "analisi in blocco", persegue l'idea di raccogliere un elevato numero di variabili per poi procedere alla loro riduzione (a mezzo della analisi fattoriale o delle componenti principali) in modo da eliminare inevitabili duplicazioni e ridondanze che si riscontra in molti indicatori di contesto. Una volta ottenuti i "metaindicatori" si procede al raggruppamento delle unità (le regioni, nel nostro caso) in "metaregioni ovvero delle macroregioni o zone omogenee caratterizzate da una combinazione distintiva dei metaindicatori che sembrano individuare specifiche configurazioni di regioni.

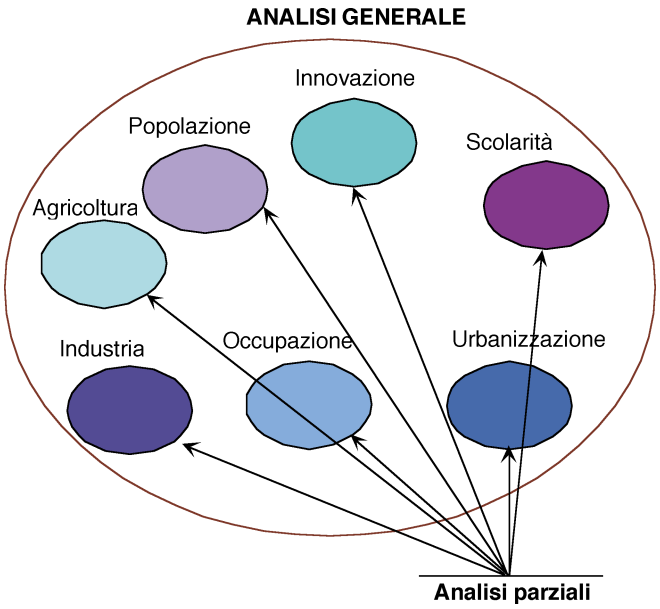


Figura 1.7: Schema generale di scelta degli indicatori.

Non sempre si ritiene valida questa procedura. Innanzitutto, l'esecuzione delle procedure di analisi multivariata non è insensibili alla dimensionalità del problema. Anche la semplice moltiplicazione di due matrici può incorrere in errori di approssimazione che, anche se piccoli, non possono essere trascurati a causa dei loro effetti cumulativi. I problemi di stabilità numerica sono molto più acuti se si analizzano tutti gli indicatori contemporaneamente. C'è poi da considerare che nell'affrontare un problema, si hanno comunque delle idee scaturite dalla sua conoscenza generale che portano ad assegnare delle priorità e quindi a rilevare certi dati piuttosto che altri, ad utilizzarne di già rilevati ed a non rilevarne di nuovi. Un'ipotesi alternativa all'analisi in blocco è che sia possibile scomporre lo sviluppo socio-economico di una regione in pochi aspetti caratteristici (determinanti) analizzabili separatamente. Il presupposto è che si possano individuare variabili tipiche di ciascun aspetto e talmente intrinseche che non sia possibile assegnare una data variabile a più di una determinante, che cioè fra queste non vi siano sovrapposizioni.

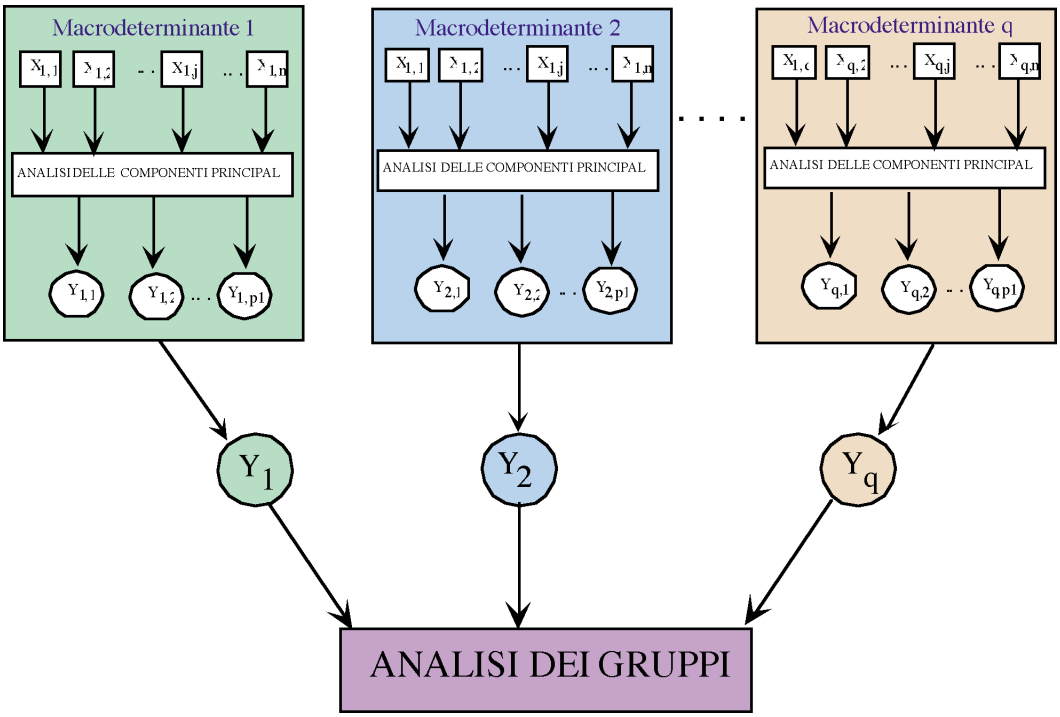


Figura 1.8: Approccio per determinanti separate.

Problemi in questo senso non mancano, perché le scelte devono essere tali da limitare al minimo l'area di sovrapposizione fra blocchi di indicatori. Dopo aver ripartito le variabili originarie tra le determinanti si applicano, separatamente, le

usuali tecniche statistiche di analisi multivariata per la ricerca dei metaindicatori rappresentativi di ciascuna determinante delle regioni ovvero si effettua una "analisi per blocchi". Questa scelta non è dovuta tanto a motivi di semplicità, che pure hanno importanza, quanto per due ragioni: da un lato si vuole restringere lo studio a quegli indicatori che effettivamente si prestano ad essere adoperate come cartine al tornasole di caratteristiche discriminanti e la cui variabilità, da una regione all'altra, sia più l'affioramento di una differenza strutturale invece che un caso accidentale. In questo modo può diventare più semplice identificare i metaindicatori ottenibili con l'analisi fattoriale. Dall'altro lato, si cerca di evitare l'inclusione di variabili tra loro fortemente correlate, che spesso determinano confusioni nel riconoscimento dei fattori.

Supponiamo che i k indicatori disponibili a livello regionale siano suddivisi in h macrodeterminanti $M_1, M_2, \cdots, M_h$ ognuna formata a sua volta da $k_1, k_2, \cdots, k_h$ indicatori con l'evidente vincolo che $\sum_{i=1}^{h} k_i = k$. Nel paragrafo precedente si è visto come realizzare l'analisi in blocco dei k indicatori. Nulla di troppo diverso comporta l'analisi per blocchi se non che invece di una sola matrice delle distanze sintetiche ne verranno determinate h che dovranno a loro volta essere sintetizzate.

Indichiamo con $\mathbf{T}^{(i)}$ la matrice varianze-covarianze di sintesi ottenuta alla fine della procedura Distatis per l'i-esima macrodeterminante (formula 1.29) e sia $\boldsymbol{\tau}_i^{(h)}$ la sua versione vettorizzata. Se ${}_2\mathbf{Z}$ denota la matrice avente per colonne le vettorizzazioni delle matrici sintetiche delle macrodeterminanti possiamo applicare lo stesso schema del paragrafo 1.2.2 a partire dalla relazione [1.24]. Costruiamo la matrice di varianze-covarianze

$$ {}_2\mathbf{A} = \left[{}_2\mathbf{N}^{-0.5}\left({}_2\mathbf{Z}\right)\right]^t \left[{}_2\mathbf{N}^{-0.5}\left({}_2\mathbf{Z}\right)\right] \tag{1.36} $$

dove ${}_2\mathbf{N}$ è la matrice diagonale formata con gli elementi della diagonale della matrice di varianze-covarianze $({}_2\mathbf{Z}{}_2^{\mathbf{t}}\mathbf{Z})$. Nell'ipotesi che la [1.36] abbia solo elementi positivi, possiamo costruire una matrice sintesi di secondo livello

$$ {}_2\mathbf{T} = \sum_{j=1}^{k} \gamma_j \mathbf{T}^{(i)} \tag{1.37} $$

dove

$$ \gamma = \left[\frac{1}{\mathbf{u}^t\left({}_2\mathbf{p}_1\right)}\right]\left({}_2\mathbf{p}_1\right) \tag{1.38} $$

Si determina quindi un terzo sistema di ponderazione $\boldsymbol{\gamma}$ che serve ora a perequare le h determinanti. La classificazione delle unità può quindi essere effettuata sulla matrice di distanze generata dalla trasformazione [1.31] applicata alla [1.37]

$$ {}_2\mathbf{D} = \left({}_2\boldsymbol{\theta}\right)\mathbf{u}^t + \mathbf{u}\left({}_2\boldsymbol{\theta}\right)^t - 2\left({}_2\mathbf{T}\right) \tag{1.39} $$

con ${}_2\boldsymbol{\theta}$ vettore formato dagli elementi della diag $({}_2\mathbf{T})$.

Le due procedure dovrebbero portare a risultati diversi a meno che tra le regioni non sia vigente una struttura di gruppo molto marcata.

1.3 Analisi dei gruppi

E' una metodologia nota anche come *cluster analysis* che ha lo scopo di raggruppare oggetti, individui, variabili, sequenze temporali o spaziali, in classi omogenee rispetto al complesso delle caratteristiche possedute da ciascun elemento; più specificamente, per analisi dei gruppi si intende una procedura multivariata che inizia con un dataset contenente informazioni su un certo numero p unità e tenta di riorganizzarle in gruppi omogenei al loro interno rispetto alle k caratteristiche rilevate su ciascuna di esse. L'obiettivo principale dell'analisi dei gruppi è la riduzione della dimensionalità attraverso la eliminazione delle descrizioni di ogni singola entità per concentrarsi sulla descrizione di poche entità -reali o virtualiche, una per ogni cluster, ne esprimano i tratti più tipici. Se ad esempio, le caratteristiche del cluster fossero fisse, basterebbe una unità qualsiasi per esprimere ogni aspetto del cluster.

Il maggior stimolo allo sviluppo di tale disciplina è da ascrivere a Sneath e Sokal [28]; dopo tale pubblicazione, la letteratura sperimentò una rapida crescita nei campi più differenziati della scienza grazie al concomitante sviluppo degli elaboratori elettronici ed all'aumento crescente delle loro prestazioni. Negli anni '90 la *cluster analysis* conosce un declino soprattutto a causa della natura euristica della sua impostazione. Il desiderio di definire una tipologia stringente ed esaustiva è spesso tanto forte da far dimenticare le debolezze teoriche delle procedeure nonché le precauzioni sempre necessarie nel valutarne i risultati. E' infatti possibile che i dati semplicemente non si prestino ad essere disaggregati, che non ci sia nessuna struttura di gruppo e tentare di imporla può portare a soluzioni poco significative. Potrebbe peraltro capitare che il dataset sia un campione proveniente da una popolazione ben articolata in gruppi, ma che per effetto del caso il campione non rispecchi quella articolazione, ma un'altra, con più o meno gruppi, o con nessun gruppo. Anche in questo caso i risultati hanno scarsa attendibilità perché le procedure inferenziali in grado di fornire indicazioni hanno avuto sviluppi limitati.

1.3.1 Concetto di cluster

E' quasi un luogo comune che non esista una definizione di gruppo o *cluster* su cui concordino tutti gli interessati alla *cluster analisys* e nemmeno la maggior parte. Più o meno si conviene su due caratteristiche almeno in parte confliggenti richieste per costituire un cluster: compattezza all'interno e separazione dall'esterno. Il primo requisito imporrebbe infatti che entità molto prossime o simili siano collocate nello stesso cluster; il secondo che unità molto remote o dissimili siano collocate in cluster diversi.

La prossima figura descrive un dataset che contempla tre cluster naturali.

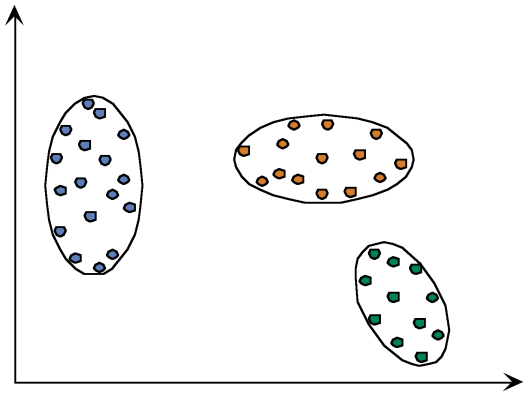


Figura 1.9: Gruppi naturali

L'appartenenza delle unità a tali gruppi è determinata in modo semplice ed univoco da una valutazione anche grezza dei dati. Raggruppamenti di questo tipo sono però rari. Più spesso ci si deve contentare di raggruppamenti in cui le unità formano un cluster solo perché le ragioni per considerarle simili sono più forti di quelle per considerale dissimili. I due requisiti di compattezza e separatezza non sono comunque indipendenti

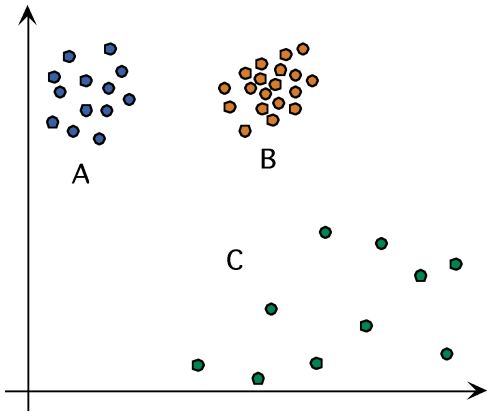


Figura 1.10: definizione grafica di cluster

I due gruppi A e B nella figura, pur essendo molto ravvicinati, sono gruppi legittimi in virtù del loro alto grado di coesione interna. Il cluster C, anche se molto disperso, è pure un legittimo cluster grazie al suo elevato grado di isolamento. Anche la cardinalità dei gruppi ha un ruolo nella definizione: l'omogeneità interna è più richiesta per i cluster meno numerosi che per quelli più numerosi; la separatezza dovrebbe essere più grande per i cluster più numerosi.

I gruppi possono essere diversi, a seconda del metodo utilizzato. Qui ricordiamo una distinzione ta quelle meno contestate: metodi gerarchici e non gerarchici.

Nella clustering gerarchica rientrano quei criteri che prevedono la individuazione di p partizioni (cioè tante quante sono le unità) ciascuna caratterizzata da un diverso numero di gruppi. Le partizioni individuate vanno a costituire una struttura gerarchica (gruppi annidati) detta albero o dendrogramma, molto utile per stabilire una tassonomia delle unità analizzate (le regioni). I criteri di tipo non gerarchico prevedono l'individuazione di una sola partizione (detta mosaico), delle p unità da analizzare, in g gruppi, dove g è un parametro, di solito incognito, che deve pure essere determinato dalla procedura. Così come per i criteri gerarchici, anche per i criteri non gerarchici di raggruppamento la attribuzione delle unità ai vari gruppi è effettuata in modo da ottimizzare una funzione obiettivo predefinita e ben specificata; tali criteri differiscono però da quelli gerarchici sia perché prevedono l'individuazione di una sola partizione delle unità sia perché consentono la rimozione e la riattribuzione delle unità stesse, rendendo così possibile la correzione di partizioni insoddisfacenti; ciò avviene attraverso l'applicazione di procedure di tipo iterativo. L'individuazione della partizione ottima si dovrebbe ottenere considerando tutte le possibili partizioni dell'insieme originario, in realtà però una procedura esaustiva è di impossibile applicazione pratica anche nei casi in cui si dispone di elaboratori ad alta velocità; pertanto la ricerca dell'ottimo è generalmente condotta in un numero ristretto di partizioni.

Mosaico. Sia $C_i = (u_{i1}, u_{i2}, \cdots, u_{ini})$ un cluster cioè un sottoinsieme delle unità comprese nel dataset $B = (u_1, u_2, \cdots, u_p)$. Un mosaico $M = (C_1, C_2, \cdots, C_g)$ di B è una collezione di sottoinsiemi tali che:

$$C_i \neq \emptyset; \quad \bigcup_{i=1}^{g} C_i = B; \quad C_i \cup C_j = \emptyset \text{ per } i \neq j \tag{1.40}$$

dove g è un parametro da determinare. Le cardinalità dei gruppi n_i verificano le relazioni

$$\sum_{i=1}^{n} n_i = n; \quad 1 \leq n_i \leq n - g + 1$$

Ogni cluster contiene almeno una entità e tutte le entità debbono essere classificate in uno ed uno solo dei cluster in funzione di uno specifico criterio di ottimizzazione.

Albero. Un albero $L = (C_1, C_2, \cdots, C_g)$ per il dataset B è una collezione di gruppi tali che

$$\begin{aligned} &1) \quad B \in L;\ \emptyset \in L \\ &2) \quad \text{Se } C_1, C_2 \in L \Rightarrow C_1 \cap C_2 \in \{\emptyset, C_1, C_2\} \end{aligned}$$

cioè o uno include l'altro oppure non hanno elementi in comune. L'albero è un livellamento gerarchico delle unità nel dataset in cui un gruppo è sempre incluso

in un gruppo più grande ottimizzando un criterio a validità loicale, almeno fino al livello in cui si decide che il processo di annidamento possa arrestarsi.

Per ottenere un mosaico si ha come input i valori osservati sugli indicatori (eventualmente trasformati) e misurati su scale intervallari o proporzionali oppure come presenza/assenza, ma deve sempre essere possibile rappresentarli come punti-vettore in uno spazio cartesiano. L'input per ottenere un albero è una matrice delle distanze.

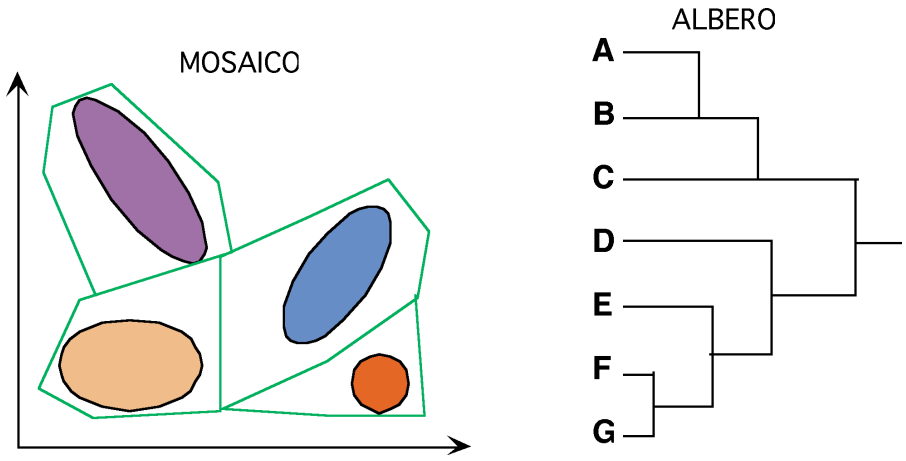


Figura 1.11: Diversità di strutture

Una delle poche certezze acquisite in molti decenni di applicazioni dell'analisi dei gruppi è proprio la diversità tra le due strutture fondamentali, mosaico ed albero, prodotte dal processo di classificazione. Il primo costituisce una suddivisione del dataset in gruppi che si dispongono tutti sullo stesso piano ed ognuno rappresenta, almeno potenzialmente, una categoria di entità omogenee al suo interno, distinta e distante dalle altre. Il numero dei gruppi deve essere stabilito a priori. L'analisi può essere ripetura con un diverso numero di gruppi, ma ogni volta il loro numero è fissato e la procedura non lo può cambiare (tranne in particolari algoritmi come la tecnica ISODATA). Il secondo forma una struttura di tipi e sottotipi di unità che facilita la organizzazione e l'interpretazione senza imporre un particolare numero di cluster; numero che può essere scelto osservando con attenzione il dendrogramma. Occorre però dire che l'albero può essere trasformato in un mosaico tagliando ad un qualche livello la gerarchia. In questo lavoro useremo solo un metodo gerarchico perché più adatto a gestire la matrice di sintesi delle distanze ottenuta alla fine della procedura Distatis discussa nel precedente paragrafo. Inoltre, la grande suggestività del dendrogramma esalterà le distinzioni delle regioni grazie al ridotto numero di unità.

1.3.2 Formazione dei gruppi

La base della analisi dei gruppi gerarchica è il calcolo delle $p(p-1)/2$ misure di distanza. E' questa un aspetto molto attrattivo: i valori originali possono essere ignorati una volta che siano confluiti nella matrice delle distanze; eventuali problemi di difformità tra le scale di misura degli indicatori e/o il loro diverso campo di variazione si ritengono risolti in fase di costituzione della matrice delle distanze. Il punto di partenza di ogni metodo aggregativo è la combinazione delle unità che presentano la minore distanza per formare un nuovo gruppo di due unità. Negli stadi successivi occorre stabilire la misura di distanza tra la singole unità ed un cluster e tra due cluster di unità. E' questa la scelta che caratterizza la particolare strategia aggregativa (ne discuteremo nel prossimo sottoparagrafo).

Si parte da un raggruppamento $L_0 = \{C_i = \{u_i\}, i = 1, 2, \cdots, p\}$ costituita da p cluster contenenti ognuno una unità e si perviene, per fusioni successive della coppia di clusters più prossima, al raggruppamento: $L_{p-1} = \{u_1, u_2, \cdots, u_p\}$ che le contiene tutte. E' quindi necessario definire una metrica o, come si usa dire in questo ambito, di un "legame" $m(C_i, C_j)$ per i cluster del raggruppamento $L_q, q = 0, 1, \cdots, n-1$ che selezioni i due da aggregare. Supponiamo che ciò avvenga per una distanza h_q allo stadio q-esimo; a tale livello il raggruppamento L_q include $p-q$ clusters di cui almeno uno con più di un elemento quando $q > 0$. La sequenza di raggruppamenti (L_q, h_q) per i livelli $q = 1, 2, \cdots, p-1$ contiene tutti i dettagli. Ogni passaggio è quello ottimo per il raggruppamento q-esimo in base al legame $m()$, con riferimento ad una coppia di cluster, ma non vi è garanzia che sia quello più efficiente rispetto a tre, quattro, o più unità né è sicura l'ottimalità del risultato finale. Peraltro, il cluster formatosi al livello q non può più essere diviso, ma solo combinato, nel suo insieme, con altri cluster. Da notare che la formazione dei gruppi è deterministica nel senso che può essere ripetuta esattamente ripartendo dal primo abbinamento delle due unità più vicine.

La matrice delle distanze per il livello $q = 0$ è una matrice del tipo cui si perviene alla del paragrafo precedente. Per quelli successivi la matrice delle distanze si ottiene cancellando la riga e la colonna corrispondenti ad uno dei clusters C_h e C_k interessati da una fusione e sostituendo la riga e la colonna dell'altro con il ricalcolo delle distanze dal nuovo cluster $C_h \cup C_k$ in base alla metrica $m()$; gli altri elementi rimangono invariati e si trasferiscono come tali al raggruppamento successivo fino a che non siano coinvolti in una qualche aggregazione. Le tecniche aggregative più note ricadono nella formula di Lance e Williams ([20]):

$$m(C_i \cup C_j, C_k) = a_i m(C_i, C_k) + a_j m(C_j, C_k) + b m(C_i, C_j) + g\,|m(C_i, C_k) - m(C_j, C_k)| \quad (1.41)$$

con $a_i, a_j, b \in [0, 1]$ e g una costante qualsiasi. Per assicurare che la distanza [1.41] sia sempre non negativa è necessario e sufficiente che $a_i + a_j + b \geq 0$ e $a_i + g \geq 0$ Tale formula consente il calcolo delle dissomiglianze al livello q a partire da quelle già

calcolate al livello precedente $q-1$ cosicché i tempi di esecuzione risultano ridotti. Tuttavia, c'è da considerare che ogni combinazione dei parametri (a_i, a_j, b, g) esprime un diverso algoritmo aggregativo alcuni corrispondenti ad interpretazioni di entità in uno spazio euclideo ed altri che ne prescindono. Diventa perciò importante disporre di criteri che aiutino a discernere tra gli infiniti schemi di aggregazione possibili con la [1.41.

Affinché lo schema di aggregazione sia rappresentabile con un dendrogramma e perciò sia più leggibile è necessario che il minimo delle distanze tra gruppi $C_q = \min m(C_i, C_j)$ sia monotonicamente crescente rispetto a q cioè al livello gerarchico di fusione. Questo implica che tra i parametri della formula generale dei legami valgano le relazioni: $a_i + a_j + b \geq 1$, $a_i + a_j \geq 0$, $g \geq -\min(a_i, a_j)$ ovvero la misura di distanza deve soddifare la disuguaglianza ultrametrica:

$$m(C_i, C_j) \leq \max\{m(C_i, C_k), m(C_j, C_k)\} \tag{1.42}$$

Questa relazione esprime un concetto utile nella clustering gerarchica e riguarda la possibilità di risalire da un insieme di indici di distanza ad una specifica classificazione e viceversa (gerarchia indicizzata. Si veda Gordon [**?**]). La [1.42] vuol dire che, laddove la disuguaglianza triangolare porta a costruire, data una terna di distanze, un triangolo qualsiasi e quindi, almeno potenzialmente, con i tre lati disuguali, la disuguaglianza ultrametrica consente di costruire triangoli equilateri o isoscele, ma non scaleni. La disuguaglianza ultrametrica implica quella triangolare, ma non viceversa.

1.3.3 Scelta del legame

Consideriamo un cluster $C = \left(u_{i1}, u_{i2}, \cdots, u_{i_{n_c}}\right)$ dove $(i_1, i_2, \cdots, i_{n_c})$ è un sottoinsieme di indici in $(1, 2, ..., p)$ ed n_c è la cardinalità di C. Sia, inoltre, e una unità esterna al gruppo C. In che modo può essere definita la distanza o legame tra e e C? Le tecniche sono tante, ma come requisito minimo da chiedere c'è il fatto che se $\mathrm{n}_c = 1$ cioè il cluster si riduce ad una sola entità u_c allora la distanza deve coincidere con la distanza tra tale unità e quella esterna: $m(C, e) = m(u_c, e)$. Devono poi essere rispettate, a nostro parere, le proprietà delle misure di distanza ultrametrica.

Legame singolo $(a_i = 0.5, a_j = 0.5, b = 0, g = -0.5)$

Si forma un insieme con tutte le distanze da ogni entità in C e si considera come distanza tra e e C quella più piccola.

$$m(C, e) = \min\{u \in C \,|m(u, e)\} \tag{1.43}$$

Questa procedura basa la fusione tra due gruppi C_i e C_j di numerosità n_i ed n_j su di un solo legame e gli altri vi entrano solo indirettamente quando formano

la graduatoria per stabilire la distanza minima. Il rischio è che, di minimo in minimo, si vengano a trovare nello stesso gruppo unità molto distanti tra di loro e che ben difficilmente si potrebbero considerare omogenee. Infatti, il legame singolo tenta di formare una catena di entità, ognuna prossima all'altra, ma con una crescente distanza tra quelle inserite subito e quelle inserite più tardi. Il problema riguarda soprattutto la separatezza dei gruppi che può essere debole da impedire il buon funzionamento del legame singolo. Tra più gruppi ben separati possono trovarsi delle entità ibride che determinano il fenomeno del concatenamento che è molto sgradevole, ma spesso in evitabile quando la struttura di gruppo non è molto pronunciata (cioè proprio nei casi più interessanti).

Il vantaggio del metodo del legame singolo è di non essere vincolato a forme particolari dei clusters e di essere l'unico in grado di cogliere clusters che non abbiano forma sferica o ellissoidale. Tale vantaggio è anche più apprezzabile ricordando che, di solito, la forma dei clusters non è nota. Da notare però che, se i cluster hanno proprio forma sferica, la performance del legame singolo ne risulta compromessa.

Legame completo $(a_i = 0.5, a_j = 0.5, b = 0, g = 0.5)$

E' antitetico al legame singolo e misura la distanza tra due gruppi con il valore massimo di tale misura tra ogni possibile combinazione formata con una unità esterna e ed una unità u interna a C

$$m\left(C, e\right) = \max\left\{u \in C \left| m\left(u, e\right)\right.\right\} \tag{1.44}$$

Il meccanismo di formazione dell'albero è lo stesso come per il legame singolo (e come per molte altre procedure agglomerative). Il legame completo tende a formare clusters sferici, compatti e c'è la possibilità che unità molto ravvicinate siano incluse nello stesso cluster solo a livelli gerarchici elevati. Il legame singolo ed il legame completo perseguono finalità diverse rispetto alle caratteristiche dei cluster: l'isolamento esterno e la coesione interna. Il legame singolo trascura la coesione interna per preoccuparsi della separatezza dei clusters; il legame completo cura maggiormente la compattezza di un cluster e considera poco l'isolamento. Da notare che per questi due tipi di legami conta solo l'ordine di grandezza delle distanze e non il valore effettivo. Si ottengono valutazioni analoghe con una qualsiasi trasformazione monotona delle distanze (ad esempio i logaritmi).

Legame delle medie di gruppo $\left(a_i = \frac{n_i}{(n_i+n_j)}, a_j = \frac{n_j}{(n_i+n_j)}, b = 0, g = 0\right)$

Se invece di considerare una delle possibili distanze consideriamo una loro sintesi quale potrebbe essere, ad esempio, la media aritmetica, delle distanze tra l'unità generica e esterna al cluster C e quelle contenute nello stesso C.

$$m\left(C, e\right) = \frac{\sum_{u \in C} m\left(u, e\right)}{n_c} \tag{1.45}$$

Il metodo delle medie di gruppo (*Unweighted pair-group average*) ha un comportamento che è una via di mezzo tra il legame singolo e quello completo. Ha però il limite che se si confrontano un cluster piccolo ed uno grande, quest'ultimo tende ad inghiottire quello minore. Inoltre, poiché si tiene conto della meddia

Metodo di Ward $\left(a_i = \frac{n_i+n_k}{(n_i+n_j)}, a_j = \frac{n_j+n_k}{(n_i+n_j)}, b = \frac{-n_k}{(n_i+n_j+n_k)}, g = -0\right)$

Dato un insieme di valori $\{x_1, x_2, \cdots, x_p\}$ si intende costruire un albero organizzato in modo tale che il passaggio da L_q a L_{q+1} rende minima la somma dei quadrati. Tra i criteri basati sulla minimizzazione della devianza totale interna ai gruppi il più noto è quello suggerito da Ward [**?**]. Il metodo di Ward ad ogni stadio del processo di raggruppamento considera tutte le possibili coppie di aggregazione tra gruppi e procede alla fusione dei due gruppi che minimizzano l'incremento della devianza totale dal centroide del nuovo gruppo formato. All'inizio del processo, quando ogni cluster è formato da un singolo elemento, la devianza interna ai vari gruppi è zero ovunque; quando due entità si fondono in un singolo cluster si introduce una certa variabilità al suo interno che è destinata a crescere in funzione della numerosità del gruppo stesso; il metodo di Ward impone la fusione di quei gruppi che provocano il minor aumento della variabilità interna ai gruppi.

Il metodo di Ward è da considerarsi uno schema piuttosto efficiente, sebbene tenda a favorire il formarsi di gruppi piccoli. Proprio per questo è inappropriato se ci si aspettano gruppi di numerosità molto diversa. A questo fine risulta più indicato il legame delle medie di gruppo. Questo tende a formare gruppi con variabilità simile, laddove il metodo di Ward tende a formare gruppi di simile ampiezza.

Una delle debolezze dell'analisi gerarchica dei gruppi è che metodi diversi applicati sulle stesse unità e basati sulla stessa matrice di distanze possono determinare raggruppamenti piuttosto diversi se non is è in presenza di gruppi naturali ben distanziati. E' quindi essenziale scegliere il tipo di legame più appropriato per i dati. Quale delle quattro proposte risulta essere la misura di distanza preferibile?

In alcuni casi è appropriato il requisito della ammissibilità monotona e cioè che se la matrice originale delle distanze subisce una trasformazione monotona crescente in qualsiasi stadio della gerarchia, il risultante raggruppamento non cambia. Chen e Van Ness [6] dimostrano che la [1.41] gode della ammissibilità monotona se

$$\begin{cases} a_i + a_j &= 1 \\ b &= 0 \\ |g| &= \min\{a_i, a_j\} \end{cases}$$

Le uniche combinazioni monotonicamente ammissibili sono il legame singolo e quello completo (grazie al fatto che ignorano l'effettiva grandezza delle distanze

tra le unità e/o i cluster per considerarne solo l'ordine). Se poi si aggiunge il requisito della ammissibità metrica e cioè le distanze tra gruppi debbono costituire una metrica ad ogni livello di aggregazione, il cerchio si restringe al solo legame completo che quindi sembra la strategia più indicata. Si deve però considerare che, se il legame completo gode delle proprietà indicate non manca di inconvenienti:

1. provoca un effetto di separazione che mantiene lontane entità vicine perché incluse in clusters che contengono -anche solo una- molto distanziate;

2. tende a formare molti gruppi poco numerosi;

3. nasconde gli *outliers* aggregandoli in clusters non manifestamente irragionevoli;

4. può generare parità nei livelli di aggregazione con arbitrarietà nella sequenza gerarchica che potrebbe cambiare secondo l'ordine di considerazione delle unità o di non riproducibilità del risultato nel caso di scelte per sorteggio.

5. E' molto sensibile: basta cambiare una distanza, anche solo leggermente, per dar luogo ad un raggruppamento molto diverso dal precedente.

Jardine e Sibson [18] elencano una serie di criteri ed accertano che solo il legame singolo è in grado di soddisfarli tutti per cui è quello da usare. Sui criteri proposti è mancato però il consenso diffuso necessario per avvalorarne una preferenza nelle applicazioni. In effetti, le sperimentazioni condotte per molti anni portano ad escludere questo metodo in favore del legame di Ward o delle medie di gruppo. Everitt *et al.* ([10], p.62) segnalano che il metodo di Ward risulta sensibile ai valori remoti laddove il metodo del legame medio è, rispetto a questi, relativamente robusto. Inoltre, il metodo di Ward presuppone, almemo concettualmente, che le unità siano dei punti nello spazio euclideo k-dimensionale.

Non esiste un legame che sia sempre superiore agli altri, anche se forse il legame singolo e quello completo trascurano almeno una parte delle informazioni ed hanno carattere estremo che li rende meno richiesti rispetto agli altri due (Milligan [25]). Poiché la classificazione delle regioni potrebbe dar luogo a gruppi di numerosità diversa e poiché la base di riferimento è la matrice sintesi delle distanze parziali utilizzeremo il legame delle medie di gruppo per la nostra indagine sulle regioni italiane.

1.4 Applicazione a dati simulati

Prima di utilizzare la nostra tecnica sui dati reali conviene sperimentarla su di un dataset artificiale di cui si conosce bene la struttura e che sia in grado di mostrare la capacità informativa della nuova tecnica multivariata. Costruiamo una matrice di dati relativa a $p = 20$ unità sulle quali facciamo finta di rilevare $k = 3$ indicatori osservati per periodi diversi, ma brevi. La prima serie storica $X_{t,1}$ è una funzione trigonometrica osservata per $n_1 = 6$ periodi per valori di t corrispondenti all'insieme $\{-0.3, -0.2, -0.1, 0.1, 0.2, 0.3\}$

Cluster 1. Numero di elementi: 6 $\{R_1, \cdots, R_6\}$

$$\begin{cases} \sin(2\pi t) & 1.25\sin(2\pi t) & 1.50\sin(2\pi t) \\ 1.75\sin(2\pi t) & 2.00\sin(2\pi t) & 2.25\sin(2\pi t) \end{cases}$$

Cluster 2. Numero di elementi: 5 $\{R_7, \cdots, R_{11}\}$

$$\begin{cases} \cos(2\pi t) & 1.25\cos(2\pi t) & 1.50\cos(2\pi t) \\ 1.75\cos(2\pi t) & 2.00\cos(2\pi t) \end{cases}$$

Cluster 3. Numero di elementi: 4 $\{R_{12}, \cdots, R_{15}\}$

$$\begin{cases} \sin\left(2\pi t - \frac{\pi}{4}\right) & 1.25\sin\left(2\pi t - \frac{\pi}{4}\right) & 1.50\sin\left(2\pi t - \frac{\pi}{4}\right) \\ 1.75\sin\left(2\pi t - \frac{\pi}{4}\right) \end{cases}$$

Cluster 4. Numero di elementi: 3 $\{R_{16}, \cdots, R_{18}\}$

$$\cos\left(2\pi t - \frac{\pi}{4}\right) \quad 1.25\cos\left(2\pi t - \frac{\pi}{4}\right) \quad 1.50\cos\left(2\pi t - \frac{\pi}{4}\right)$$

Cluster 5. Numero di elementi: 2 $\{R_{19}, R_{20}\}$

$$\sin\left(2\pi t + \frac{\pi}{6}\right) \quad 1.25\sin\left(2\pi t + \frac{\pi}{4}\right)$$

Il secondo indicatore è ricavato dal primo per $n_2 = 5$ periodi con la trasformazione $X_{t,2} = 1 + X_{t,1} - 1.5X_{t,2}^2 \quad \text{per} \quad t = 1, 2, \cdots, 5$. Il terzo indicatore è pure ricavato dal primo per $n_3 = 4$ periodi con la trasformazione $X_{t,3} = X_{t,i}^{-1} \quad \text{per} \quad t = 1, 2, \cdots, 4$. I valori delle pseudo serie storiche sono riportate in tabella

Tabella 1.3: Valori delle serie storiche del dataset simulato

Unità	A_1	A_2	A_3	A_4	A_5	A_6	B_1	B_2	B_3	B_4	B_5	C_1	C_2	C_3	C_4
R_1	-1.0	-1.0	-0.6	0.6	1.0	1.0	-1.3	-1.3	-0.1	1.1	0.6	-1.1	-1.1	-1.7	1.7
R_2	-1.2	-1.2	-0.7	0.7	1.2	1.2	-2.3	-2.3	-0.5	0.9	0.1	-0.8	-0.8	-1.4	1.4
R_3	-1.4	-1.4	-0.9	0.9	1.4	1.4	-3.5	-3.5	-1.0	0.7	-0.6	-0.7	-0.7	-1.1	1.1
R_4	-1.7	-1.7	-1.0	1.0	1.7	1.7	-4.8	-4.8	-1.6	0.4	-1.5	-0.6	-0.6	-1.0	1.0
R_5	-1.9	-1.9	-1.2	1.2	1.9	1.9	-6.3	-6.3	-2.2	0.1	-2.5	-0.5	-0.5	-0.9	0.9
R_6	-2.1	-2.1	-1.3	1.3	2.1	2.1	-8.0	-8.0	-2.9	-0.3	-3.7	-0.5	-0.5	-0.8	0.8
R_7	-0.3	0.3	0.8	0.8	0.3	-0.3	0.5	1.2	0.8	0.8	1.2	-3.2	3.2	1.2	1.2
R_8	-0.4	0.4	1.0	1.0	0.4	-0.4	0.4	1.2	0.5	0.5	1.2	-2.6	2.6	1.0	1.0
R_9	-0.5	0.5	1.2	1.2	0.5	-0.5	0.2	1.1	0.0	0.0	1.1	-2.2	2.2	0.8	0.8
R_{10}	-0.5	0.5	1.4	1.4	0.5	-0.5	0.0	1.1	-0.6	-0.6	1.1	-1.8	1.8	0.7	0.7
R_{11}	-0.6	0.6	1.6	1.6	0.6	-0.6	-0.2	1.0	-1.3	-1.3	1.0	-1.6	1.6	0.6	0.6
R_{12}	-0.5	-0.9	-1.0	-0.2	0.5	0.9	0.2	-1.1	-1.5	0.8	1.1	-2.2	-1.1	-1.0	-6.4
R_{13}	-0.6	-1.1	-1.2	-0.2	0.6	1.1	-0.1	-2.0	-2.5	0.7	1.1	-1.8	-0.9	-0.8	-5.1
R_{14}	-0.7	-1.3	-1.5	-0.2	0.7	1.3	-0.4	-3.0	-3.8	0.7	1.0	-1.5	-0.7	-0.7	-4.3
R_{15}	-0.8	-1.6	-1.7	-0.3	0.8	1.6	-0.7	-4.2	-5.2	0.6	0.8	-1.3	-0.6	-0.6	-3.7
R_{16}	-0.9	-0.5	0.2	1.0	0.9	0.5	-1.1	0.2	1.1	0.5	0.7	-1.1	-2.2	6.4	1.0
R_{17}	-1.1	-0.6	0.2	1.2	1.1	0.6	-2.0	-0.1	1.1	-0.1	0.3	-0.9	-1.8	5.1	0.8
R_{18}	-1.3	-0.7	0.2	1.5	1.3	0.7	-3.0	-0.4	1.2	-0.8	-0.3	-0.7	-1.5	4.3	0.7
R_{19}	-0.4	-0.1	0.2	0.7	0.9	1.0	0.3	0.9	1.1	0.9	0.7	-2.5	-9.6	4.8	1.3
R_{20}	-0.5	-0.1	0.3	0.9	1.1	1.2	0.1	0.8	1.2	0.6	0.2	-2.0	-7.7	3.8	1.1

L'impianto complessivo delle serie storiche simulate può essere percepito rappresentando le unità con le rispettive serie storiche mediante l'applicazione della tecnica delle curve di Andrews (Andrews, [3]).

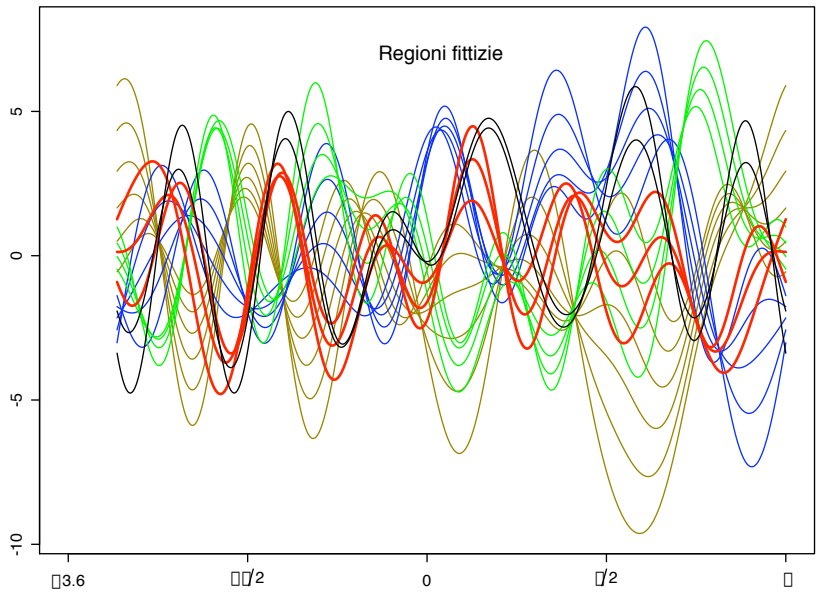


Figura 1.12: Curve di Andrews dei dati simulati

L'idea originale di Andrews è stata quella di inglobare i valori multidimensionali in una entità con infinite dimensioni, ma facili da riportare in grafico: curve periodiche. Infatti, ogni indicatore nel dataset è rappresentato con una componente nella somma della trasformata di Fourier. L'insieme di punti in R^k

diventa un fascio di curve in R^2. I valori osservati costituiscono i coefficienti della combinazione lineare delle armoniche.

$$f_i(t) = \frac{x_{i1}}{\sqrt{2}} + x_{i2}sen(t) + x_{i3}cos(t) + x_{i4}sen(2t) + x_{i5}cos(2t) + \cdots$$

il cui campo di variazione è stabilito in $-\pi < t < \pi$. L'ordine di grandezza degli indicatori incide sulla frequenza, sull'ampiezza della fase e sul periodo delle funzioni armoniche caratterizzando così ogni unità con una specifica curva. In questo senso riteniamo utile applicare il grafico alle variabili standardizzate. Le somiglianze e le differenze tra le unità emergono dal confronto della diversa morfologia delle curve.

La classificazione delle unità è abbastanza evidente dalle curve che delineano una segmentazione in cinque gruppi ben distinti. Una tecnica di analisi multivariata che manchi di individuare questi gruppi non sarebbe da raccomandare.

La matrice sintesi delle distanze è stata ottenuta ponderando le matrici parziali con pesi: 0.335, 0.353, 0.312 determinati dalla procedura Distatis. Il secondo pseudo indicatore ha una proiezione maggiore degli altri due ovvero è più vicino alla media delle distanze.

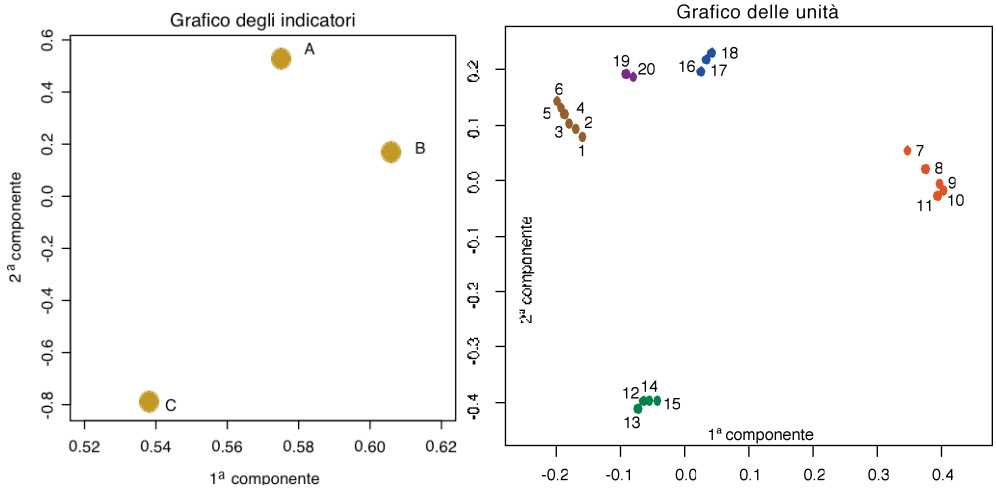


Figura 1.13: Posizionamento degli indicatori e delle unità nel dataset simulato

Gli indicatori A e B foniscono una struttura di distanza più simile o comunque maggiormente differenziata rispetto all'indicatore C che infatti partecipa con peso minore alla matrice sintetica. La struttura dei gruppi appare in tutta evidenza nel grafico a sinistra.

L'applicazione del raggruppamento gerarchico aggregativo (metodo del legame medio) comporta il seguente dendrogramma

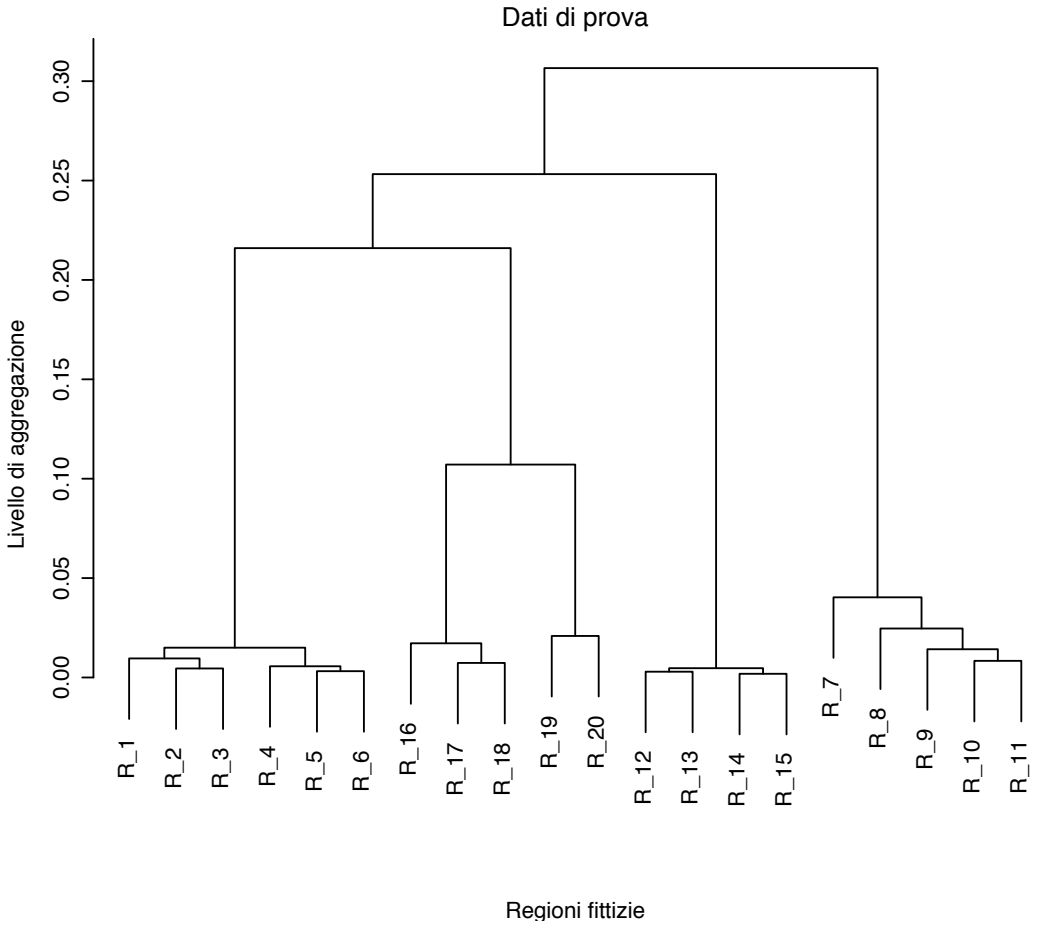


Figura 1.14: Analisi dei gruppi nel dataset simulato

E' facile constatare che la presenza di cluster naturali nell'evoluzione delle unità ha dato luogo a raggruppamenti molto riconoscibili. L'unica incertezza, peraltro legittima se si considera la formula di costruzione, riguarda la collocazione del quinto cluster che potrebbe affiancarsi al quarto.

1.5 Scelta degli indicatori

Oggetto del nostro studio è un tentativo di classificazione o *zoning* delle regioni italiane in aree uniformi secondo alcuni indicatori socio-economici rilevati nelle regioni italiane, con l'obiettivo di individuare e definire le diverse tipologie di regione esistenti in Italia. Una delle questioni fondamentali, piuttosto frequente nelle ricerche territoriali, è la scelta delle variabili regionali da prendere in esame al fine di sfruttare in modo adeguato i giacimenti informativi che in modo crescente si rendono disponibili. Il procedimento di scelta è tanto difficile quanto importante perché una batteria di indicatori ben selezionati e sapientemente organizzati in macrodeterminanti (se si adotta l'analisi per blocchi), dovrebbe fornire valide indicazioni sia per individuare differenti livelli di sviluppo nonché di evidenziare in maniera sintetica gli aspetti più rilevanti della crescita economica attesa per l'area territoriale di riferimento (la Calabria nel nostro caso).

1.5.1 Quadro comunitario di sostegno

Il Quadro comunitario di sostegno (QCS) è il documento approvato dalla Commissione europea, d'intesa con gli Stati membri interessati, sulla base della valutazione del Piano presentato dagli stessi Stati. Il QCS contiene la fotografia della situazione di partenza, la strategia, le priorità d'azione, gli obiettivi specifici, la ripartizione delle risorse finanziarie, le condizioni di attuazione. Il QCS è articolato in assi prioritari e attuato tramite uno o più Programmi operativi.In Italia, il processo formativo di tali documenti è stato avviato nel dicembre 1998 con il coordinamento del Ministero del Tesoro, del Bilancio e della Programmazione Economica (oggi Ministero dell'Economia). I soggetti coinvolti sono stati: regioni, amministrazioni nazionali, enti locali e parti economiche e sociali.

Nell'ambito del progetto: "Informazione statistica territoriale e settoriale per le politiche strutturali 2001-2006" , l'ISTAT [16] si è impegnato a supportare l'attività di valutazione del QCS attraverso la costruzione e l'aggiornamento di un'ampia base dati di indicatori socio-economici regionali. Gli indicatori di "contesto chiave" e le variabili di rottura sono uno degli strumenti di monitoraggio e valutazione adottato in tale ambito. In particolare gli indicatori di contesto costituiscono delle *proxy* per la misurazione del livello di realizzazione di un insieme di obiettivi specifici aggregati per settore o macro aree di intervento. Gli aggiornamenti del dataset sono effettuati tre volte l'anno (indicativamente a marzo, luglio e novembre) per tutto il periodo di validità del QCS, cioè fino al 2008. Ne consegue che per le regioni italiane è ora disponibile una base di indicatori statistici regionali sulla base dei quali è stata condotta l'attività di programmazione e valutazione ex ante degli interventi da attuarsi nelle regioni Obiettivo 1 nell'ambito dei Fondi strutturali 2000-2006.

La banca dati regionale è attualmente composta da 140 indicatori costruiti per le venti regioni italiane, le quattro ripartizioni territoriali (con l'aggiunta

dell'aggregato Centro-nord), le regioni Obiettivo 1, le regioni non Obiettivo 1, e le regioni Obiettivo 1 con l'esclusione del Molise che, fino al 31/12/06, si trova in una condizione di regime di aiuti transitorio. Laddove disponibili, i dati sono forniti in serie storica breve http://www.istat.it/ambiente/contesto/azioneB.html. L'orizzonte temporale di riferimento è costituito, generalmente, dagli anni 1998-2003. Per le nostre elaborazioni abbiamo considerato solo gli aggregati Obiettivo 1 (Molise incluso) e regioni non Obiettivo 1 che sono diventate delle unità virtuali definite con i valori in posizione mediana nella graduatoria degli indicatori rilevati nelle regioni componenti.

Nella discussione del metodo Distatis si è avuto modo di constatare l'importanza di avere indicatori concordanti ovvero che essi si muovano, per quanto possibile, nella stessa direzione del fenomeno studiato: valori alti sistematicamente associati a connotazioni positive del fenomeno (o anche negative, se si vuole, purché il trend sia comune). In questo senso occorre notare che in molte serie economiche la correlazione è positiva o almeno è resa tale da fattori concomitanti che determinano co-movimenti comuni nelle loro manifestazioni.

Gli indicatori predisposti dall'Istat hanno già una articolazione in macrodeterminanti dette "assi prioritari" corrispondenti alle grandi aree di intervento del QCS e che mira a favorire una piena applicazione all'assetto programmatico dei principi di coerenza, concentrazione ed integrazione. In questo senso non è necessario procedere ad una riclassificazione degli indicatori in macrodeterminanti nel caso di analisi per blocchi. Sono state comunque effettuare alcune semplificazioni. In particolare sono stati esclusi gli indicatori per i quali erano assenti, per il periodo considerato, i dati relativi a più di una delle regioni ovvero sono stati accorpati ai dati di altre regioni. Sono state anche escluse le serie storiche la cui durata non consentiva alla corrispondente poligonale di formare almeno un angolo (cioè con meno di tre periodi di rilevazione). Gli indicatori regionali usate nella nostra indagine sono:

- Asse 0: Variabili di rottura (14 indicatori effettivi su 15 disponibili). Questa determinante è finalizzata alla quantificazione dell'impatto complessivo del Programma di Sviluppo del Mezzogiorno in quanto include indici dell'efficacia dell'intervento pubblico nel generare meccanismi endogeni dello sviluppo e capaci di misurare i punti di "rottura" e discontinuità nel contesto socio-economico.

  a) Capacità di esportare. Valore delle esportazioni di merci in perc. del PIL. Periodo 1985-2004.
  b) Grado di apertura dei mercati: importazioni. Valore delle importazioni di merci in percentuale del PIL. Periodo 1985-2004.
  c) Grado di indipendenza economica. Importazioni nette in perc. del PIL. Periodo 1995-2003.
  d) Capacità di attrazione dei consumi turistici. Giornate di presenza (italiani e stranieri) nel complesso degli esercizi ricettivi per abitante. Periodo 1980-2004.

e) Intensità di accumulazione del capitale. Investimenti fissi lordi in perc. del PIL. Periodo 1980-2003.

f) Capacità di attrazione di investimenti esteri. Investimenti diretti lordi dall'estero in Italia su investimenti diretti netti in EU15. Periodo 1997-2004.

g) Partecipazione della popolazione al mercato del lavoro. Forze di lavoro in età 15-64 anni sul totale della popolazione in età 15-64 anni. Periodo 1995-2005.

h) Capacità di offrire lavoro regolare. Unità di lavoro irregolari sul totale delle unità di lavoro. Periodo 1994-2004.

i) Capacità di esportare prodotti a elevata o crescente produttività. Quota delle esportazioni dei prodotti ad elevata crescita della domanda mondiale sul totale delle esportazioni. Periodo 1991-2004.

j) Capacità innovativa. Spesa sostenuta per attività di ricerca e sviluppo intra muros della Pubblica Amministrazione, dell'Università e delle imprese pubbliche e private in percentuale del PIL. Periodo 1982-2003.

k) Capacità di sviluppo dei servizi alle imprese. Unità di lavoro nel settore delle Attività immobiliari e imprenditoriali sul totale delle unità di lavoro dei servizi destinabili alla vendita. Periodo 1995-2003.

l) Capacità di sviluppo dei servizi sociali. Persone di 14 anni e più che hanno partecipato a volontariato sul totale della popolazione di 14 anni e più. Periodo 1995-2004.

m) Capacità di finanziamento. Differenziale dei tassi attivi a breve termine sui finanziamenti per cassa con il Centro-Nord. Periodo 1995-2003.

n) Condizioni di legalità e coesione sociale. Crimini violenti per 10000 abitanti. 1983-2003.

- Asse 1: Risorse naturali (15 indicatori completi su 16 disponibili). L'obiettivo globale dell'Asse è: creare nuove opportunità di crescita e di sviluppo sostenibile; rimuovere le condizioni di emergenza ambientale; assicurare l'uso efficiente e razionale e la fruibilità di risorse naturali; adeguare e razionalizzare reti di servizio efficienti per acqua e rifiuti; garantire il presidio del territorio anche attraverso le attività agricole; preservare le possibilità di sviluppo nel lungo periodo e accrescere la qualità della vita.

  a) Irregolarità nella distribuzione dell'acqua. Famiglie che denunciano irregolarità nell'erogazione dell'acqua. Periodo 1995-2004.

  b) Nuove superfici boscate. Variazione assoluta di nuove superfici boscate (in ettari; base 1995). Periodo 1996-2004. Il dato delle Marche è nullo per tutto il periodo.

  c) Superficie forestata in ettari. Periodo 1995-2004.

  d) Raccolta differenziata dei rifiuti urbani. Rifiuti urbani oggetto di raccolta differenziata sul totale dei rifiuti urbani. Periodo 1996-2004.

  e) Superficie forestale percorsa dal fuoco. Superficie forestale percorsa dal fuoco sul totale della superficie forestale. Periodo 1995-2004.

  f) Spese di ripristino del soprassuolo boscato. Spese di ripristino totale o parziale del soprassuolo boscato (in migliaia di euro). Periodo 1995-2001.

  g) Interruzioni del servizio elettrico. Frequenza delle interruzioni accidentali lunghe (superiori a 3 minuti) del servizio elettrico (numero medio per utente). Periodo 1998-2004.

h) Grado di insoddisfazione dell'utenza per l'erogazione di gas. Grado di insoddisfazione complessiva dell'utenza per i servizi di erogazione di gas. Periodo 1998-2004. Il dato della Sardegna è stato posto pari a zero data l'esiguità degli allacciamenti alla rete del metano.

i) Energia prodotta da fonti rinnovabili. GWh di energia prodotta da fonti rinnovabili su GWh prodotti in totale. Periodo 2000-2004.

j) Potenza efficiente lorda delle fonti rinnovabili. Mw di potenza efficiente lorda delle fonti rinnovabili su Mw di potenza efficiente lorda totale. Periodo 2000-2004.

k) Intensità energetica dell'industria. Intensità energetica dell'industria (migliaia di TEP per milioni di euro di valore aggiunto prodotto dall'industria). Periodo 1995-2003.

l) Popolazione regionale servita da gas metano. Perc. di popolazione regionale che risiede in comuni serviti da gas metano. Periodo 2002-2004.

m) Rifiuti urbani raccolti. Rifiuti urbani raccolti per abitante (in kg). Periodo 1996-2004.

n) Rifiuti urbani smaltiti in discarica per abitante. Rifiuti urbani smaltiti in discarica per abitante (in kg) Periodo 1996-2004.

o) Consumi di energia elettrica coperti da fonti rinnovabili. Produzione lorda di energia elettrica da fonti rinnovabili in percentuale dei consumi interni lordi di energia elettrica. Periodo 2000-204.

- Asse 2: Risorse culturali (4 indicatori effettivi su 9 disponibili). L'obiettivo globale dell'Asse è: stabilire condizioni per nuove opportunità imprenditoriali nel settore della cultura e delle attività culturali; accrescere la qualità della vita dei cittadini, la fiducia e il benessere sociale; valorizzare, tutelare e rendere maggiormente fruibili le risorse culturali del Mezzogiorno.

  a) Grado di partecipazione del pubblico agli spettacoli teatrali e musicali. Spesa media pro-capite del pubblico per attività teatrali e musicali (euro costanti 2003). Periodo 1995-2004.

  b) Grado di diffusione degli spettacoli teatrali e musicali. Biglietti venduti per attività teatrali e musicali per 100 abitanti. Periodo 1995-2004.

  c) Incidenza della spesa per ricreazione e cultura. Percentuale di consumi interni (dei residenti e non) per ricreazione e cultura sul totale dei consumi interni. Periodo 1995-2004.

  d) Volume di lavoro impiegato nel settore ricreazione e cultura. Unità di lavoro (ULA) del settore ricreazione e cultura. Percentuale sul totale. Periodo 1995-2004.

- Area 3: Risorse umane (26 indicatori effettivi su 27 disponibili). L'obiettivo globale dell'Asse è: indurre nuove occasioni di sviluppo espandendo la dotazione, la disponibilità e la qualità delle risorse umane. Far crescere il contenuto scientificotecnologico delle produzioni meridionali; rafforzare la rete di centri di competenza del Mezzogiorno e valorizzare i collegamenti tra sottosistema scientifico ed imprenditoriale. Ridurre i tassi di disoccupazione, accrescere la partecipazione al mercato del lavoro e l'emersione delle attività informali (e quindi la loro produttività), valorizzare le risorse femminili, favorire i processi di recupero della fiducia e benessere sociale e ridurre la marginalità sociale.

a) Tasso di disoccupazione giovanile (nuova serie). Persone in cerca di occupazione in età $15-24$ anni su forze di lavoro della corrispondente classe di età (in percentuale). Periodo 1995-2005.

b) Tasso di disoccupazione (nuova serie). Persone in cerca di occupazione in età 15 anni e oltre sulle forze di lavoro nella corrispondente classe di età (in percentuale). Periodo 1995-2005.

c) Tasso di occupazione (nuova serie). Persone occupate in età 15-64 anni sulla popolazione nella corrispondente classe di età (in percentuale). Periodo 1995-2005.

d) Tasso di occupazione 55-64 anni (vecchia e nuova serie). Persone occupate in età 55-64 anni sulla popolazione nella corrispondente classe di età (in percentuale). Periodo 1995-2005.

e) Incidenza della disoccupazione di lunga durata (nuova serie). Quota di persone in cerca di occupazione da oltre 12 mesi sul totale delle persone in cerca di occupazione (in percentuale). Periodo 1995-2005.

f) Tasso di natalità lorda delle imprese. Nuove imprese sul totale delle imprese registrate nell'anno precedente (in percentuale). Periodo 1995-2005. I dati relativi all'Umbria non sono attendibili. Il valore corrispondente per il 1995 ed il 1996 è stato posto pari al valore riscontrato nel 1997.

g) Tasso di occupazione regolare. Occupati interni regolari sulla popolazione residente media in età 15-64 anni (in percentuale). Periodo 1995-2003.

h) Incidenza della disoccupazione femminile di lunga durata (nuova serie). Quota di persone in cerca di occupazione da oltre 12 mesi sul totale delle persone in cerca di occupazione-Femmine. Periodo 1995-2005.

i) Differenza tra tasso di occupazione maschile e femminile (nuova serie). Differenza assoluta fra tasso di occupazione maschile e tasso di occupazione femminile in età 15-64 anni (in percentuale). Periodo 1995-2005.

j) Differenza tra tasso di attività maschile e femminile (nuova serie). Differenza assoluta fra tasso di attività maschile e tasso di attività femminile in età 15-64 anni (in percentuale). Periodo 1995-2005.

k) Occupati che partecipano ad attività formative e di istruzione (vecchia serie). Adulti occupati nella classe d'età 25-64 anni che partecipano ad attività formative e di istruzione per 100 adulti occupati nella classe di età corrispondente (in percentuale). Periodo 1995-2005.

l) Non occupati che partecipano ad attività formative e di istruzione (vecchia e nuova serie). Adulti inoccupati (disoccupati e non forze di lavoro) nella classe d'età 25-64 anni che partecipano ad attività formative e di istruzione per 100 adulti inoccupati nella classe d'età corrispondente (in percentuale). Periodo 1995-2005.

m) Indice di povertà regionale (popolazione). Popolazione che vive in famiglie al di sotto della soglia di povertà (in percentuale). Periodo 2002-2004.

n) Indice di povertà regionale (famiglie). Famiglie che vivono al di sotto della soglia di povertà (in percentuale). Periodo 2002-2004.

o) Livello di istruzione della popolazione 15-19 anni (vecchia serie). Quota della popolazione di 15-19 anni in possesso almeno della licenza media inferiore (in percentuale). Periodo 1995-2005.

p) Tasso di partecipazione nell'istruzione secondaria superiore. Periodo 1994-2004.

q) Tasso di abbandono al secondo anno delle scuole secondarie superiori. Abbandoni su iscritti al secondo anno del totale delle scuole secondarie superiori. Periodo 1995-2004.

r) Tasso di abbandono al primo anno delle scuole secondarie superiori. Abbandoni su iscritti al primo anno del totale delle scuole secondarie superiori. Periodo 1995-2004.

s) Laureati in scienza e tecnologia. Laureati in disciplina scientifiche e tecnologiche per mille abitanti in età 20-29 anni. Periodo 1998-2004.

t) Addetti alla Ricerca e Sviluppo. Addetti alla ricerca e sviluppo per 1.000 abitanti. Periodo 1995-2003.

u) Incidenza della spesa pubblica in Ricerca e Sviluppo. Spese per ricerca e sviluppo della Pubblica Amministrazione e dell'Università in percentuale del PIL. Periodo 1995-2003. Il dato della Val d'Aosta per il 1998, omesso per motivi di riservatezza dall'ISTAT, è stato ricostruito con la semisomma tra il dato del 1997 e dl 1999.

v) Adulti che partecipano all'apprendimento permanente. Percentuale della popolazione 25-64 anni che frequenta un corso di studio o di formazione professionale. Periodo 2001-2005.

w) Tasso di scolarizzazione superiore (vecchia serie). Percentuale della popolazione in età 20-24 anni che ha conseguito almeno il diploma di scuola secondaria superiore. Periodo 1995-2005.

x) Giovani che abbandonano prematuramente gli studi. Percentuale della popolazione 18-24 anni con titolo di studio inferiore al diploma di scuola secondaria superiore e che non partecipa ad ulteriore istruzione o formazione. Periodo 2000-2005.

y) Livello di istruzione della popolazione adulta (vecchia serie). Percentuale della popolazione in età 25-64 anni che ha conseguito al più un livello di istruzione secondario inferiore. Periodo 1995-2005.

- Asse 4: Sistemi locali di sviluppo (28 indicatori effettivi su 33 disponibili). L'obiettivo globale dell'Asse è: creare le condizioni economiche per lo sviluppo imprenditoriale e la crescita produttiva; aumentare la competitività, la produttività, la coesione e la cooperazione sociale in aree concentrate del territorio, irrobustendo, anche attraverso l'innovazione tecnologica, le filiere produttive (specie in agricoltura e nello sviluppo rurale); promuovere la localizzazione di nuove iniziative imprenditoriali, ivi incluse quelle nel settore turistico, e l'emersione di imprese dall'area del sommerso; assicurare la sostenibilità ambientale dello sviluppo del sistema produttivo, anche utilizzando le migliori tecnologie disponibili e rispettando nel medio e lungo periodo la capacità di carico dell'ambiente.

  a) Indice di infrastrutturazione economica. Indice di infrastrutturazione economica (N.I. Italia=100). Periodo 1995-1997. Il dato per il 1996 è stato stimato ponendolo pari alla semisomma tra il valore del 1995 e quello del 1997.

  b) Incidenza dell'occupazione nei distretti industriali. Quota di occupati interni nei SLL di PMI della regione sul totale degli occupati interni. Periodo 1996-2002.

c) Tasso di crescita dell'agricoltura. Variazione relativa rispetto all'anno precedente del V.A. ai prezzi base dell'agricoltura, della caccia e della silvicoltura. Periodo 1996-2004.

d) Produttività del lavoro in agricoltura. Valore aggiunto dell'agricoltura, della caccia e della silvicoltura per ULA (in migliaia di eurolire 1995). Periodo 1995-2003.

e) Superficie irrigata/irrigabile nelle aziende agricole. Ettari di superficie irrigata/irrigabile sul totale della superficie agricola utilizzata (SAU). Periodo 1995-2000. Il dato del 2003 non è stato utilizzato in quanto mancavano sia quelli del 2001 che quelli del 2002.

f) Produttività dei terreni agricoli. Valore aggiunto dell'agricoltura, della caccia e della silvicoltura ai prezzi base per ettaro di SAU (in migliaia di eurolire 1995). Periodo 1995-2004.

g) Elementi fertilizzanti usati in agricoltura. Elementi fertilizzanti semplici distribuiti per ettaro di SAU (dati in quintali). Periodo 1995-2004.

h) Principi attivi contenuti nei prodotti fitosanitari. Principi attivi contenuti nei prodotti fitosanitari (a) per ettaro di SAU (in kg). Periodo 1997-2004.

i) Peso dell'export del settore agroalimentare. Valore delle esportazioni di prodotti agroalimentari in percentuale rispetto al PIL. Periodo 1995-2004.

j) Occupati delle famiglie agricole in altri settori. Componenti della famiglia agricola (a) che svolgono attività lavorativa extra-aziendale in settori diversi da quello agricolo. Periodo 1995-2000. Il dato del 2003 non è stato utilizzato in quanto mancavano sia quelli del 2001 che quelli del 2002.

k) Produttività del lavoro nell'industria alimentare. Valore aggiunto dell'industria alimentare, delle bevande e del tabacco per ULA dello stesso settore (migliaia di eurolire 1995). Periodo 1995-2003.

l) Variazione della popolazione residente nei comuni rurali. Variazione relativa annua della popolazione residente nei comuni rurali. Periodo 1995-2004.

m) Peso delle società cooperative. Addetti delle società cooperative sul totale degli addetti. Periodo 1996-2003.

n) Produttività del lavoro nell'artigianato. Valore aggiunto per addetto delle PMI iscritte ad albo artigiano (in migliaia di eurolire ed euro correnti). Periodo 1998-2003.

o) Importanza economica del settore della pesca. Valore aggiunto ai prezzi base della pesca, della piscicoltura e servizi connesi sul valore aggiunto totale. Periodo 1999-2003.

p) Produttività del lavoro nell'industria in senso stretto. Valore aggiunto dell'industria in senso stretto per ULA dello stesso settore (migliaia di eurolire 1995). Periodo 1995-2004.

q) Produttività del lavoro nell'industria manifatturiera. Valore aggiunto dell'industria manifatturiera per ULA dello stesso settore (migliaia di eurolire 1995). Periodo 1995-2003.

r) Produttività del lavoro nel commercio. Valore aggiunto del settore del commercio per ULA dello stesso settore (migliaia di eurolire 1995). Periodo 1995-2003.

s) Produttività del lavoro nel turismo. Valore aggiunto del settore del turismo per ULA dello stesso settore (migliaia di eurolire 1995). Periodo 1995-2003.

t) Produttività del lavoro nei servizi di intemediazione monetaria e finanziaria e nelle attività immobiliari e imprenditoriali. Valore aggiunto dei settori "intermediazione monetaria e finanziaria e "attività immobiliari ed imprenditoriali per ULA degli stessi settori (migliaia di eurolire 1995). Periodo 1995-2003.

u) Produttività del lavoro nelle PMI. Valore aggiunto aziendale per addetto nelle piccole e medie imprese (in migliaia di eurolire ed euro correnti). Periodo 1998-2003.

v) Tasso di natalità netta delle imprese. Imprese iscritte meno imprese cessate sul totale delle imprese registrate nell'anno precedente. Periodo 1995-2005.

w) Incidenza della certificazione ambientale. Siti di organizzazioni con certificazione ambientale $ISO/14001$ sul totale dei siti di organizzazioni certificate. Periodo 1998-2005.

x) Investimenti diretti della regione all'estero. Investimenti diretti netti della regione all'estero in percentuale del PIL. Periodo 1997-2004.

y) Andamento dell'occupazione del settore della pesca. Variazione relativa rispetto all'anno precedente delle unità di lavoro del settore della pesca, della piscicoltura e dei servizi connessi. Periodo 1996-2003.

z) Indice di intensità creditizia. Impieghi bancari (consistenza media annua) in percentuale del Pil (a prezzi correnti). Periodo 1998-2004.

A Investimenti in capitale di rischio - *early stage*. Investimenti in capitale di rischio in percentuale del PIL. Periodo 1999-2004.

B Investimenti in capitale di rischio - *expansion e replacement*. Investimenti in capitale di rischio in percentuale del PIL. Periodo 1999-2004.

- Asse 5: Città (12 indicatori effettivi su 12 disponibili). L'obiettivo globale dell'Asse è: migliorare l'articolazione funzionale e la qualità del sistema urbano del Mezzogiorno attraverso la definizione del ruolo delle città nel loro contesto regionale, e in particolare: creare condizioni economiche, amministrative e sociali adatte allo sviluppo imprenditoriale; aumentare la competitività e la produttività strutturale dei sistemi economici territoriali; favorire la localizzazione di nuove iniziative nelle aree urbane e metropolitane specie nei servizi alle persone e alle imprese; combattere la marginalità sociale e favorire i processi di recupero della fiducia sociale; riqualificare il contesto urbano, con particolare attenzione per gli aspetti ambientali.

  a) Diffusione della pratica sportiva. Popolazione di 3 anni e più che esercita pratica sportiva in modo continuativo o saltuario. Periodo 1995-2004. Nel 1996 la domanda relativa all'attività sportiva non era contenuta nell'indagine annuale. Per completare il quadro dell'indicatore il valore relativo all'anno mancante è stato posto pari alla semisomma tra il valore del 1995 e quello del 1997 per tutte le regioni.

  b) Difficoltà delle famiglie nel raggiungere negozi alimentari e/o mercati. Perc. di famiglie che dichiarano molta o abbastanza difficoltà nel raggiungere negozi alimentari e/o mercati. Periodo 1995-2005.

  c) Difficoltà delle famiglie nel raggiungere i supermercati. Perc. di famiglie che dichiarano molta o abbastanza difficoltà nel raggiungere i supermercati. Periodo 1995-2005.

  d) Monitoraggio della qualità dell'aria. Dotazione di stazioni di monitoraggio dell'aria (valori per $100'000$ abitanti). Periodo 2000-2004.

e) Utilizzo di mezzi pubblici di trasporto . Occupati, studenti e scolari, utenti di mezzi pubblici sul totale delle persone che si sono spostate per motivi di lavoro e di studio hanno usato mezzi di trasporto (in percentuale). Periodo 1995-2004.

f) Indice di microcriminalità nelle città 1. Totale delitti legati alla microcriminalità nelle città per $1'000$ abitanti. Periodo 1995-2003.

g) Indice di microcriminalità nelle città 2. Totale delitti legati alla microcriminalità nelle città sul totale dei delitti nei comuni capoluogo. Periodo 1995-2003.

h) Verde pubblico nelle città. Metri quadri di verde urbano a gestione comunale nei comuni capoluogo di provincia per abitante. Periodo 2000-2003.

i) Trasporto pubblico locale nelle città. Linee urbane di trasporto pubblico locale nei comuni capoluogo di provincia per $100Km^2$ di superficie comunale. Periodo 2000-2003.

j) Dotazione di parcheggi di corrispondenza. Numero di stalli di sosta nei parcheggi di corrispondenza dei comuni capoluogo di provincia per $1'000$ autovetture circolanti. Periodo 2000-2003.

k) Indice di attrattività delle università. Rapporto tra saldo migratorio netto degli studenti e il totale degli studenti immatricolati (in percentuale). Periodo 1998/199- 2003/2004. Per i dati della Val d'Aosta relativi agli anni accademici 1998/1999 e 1998/2000 si è utilizzato l'osservazione del 2000/2001, anno di istituzione dell'Università degli studi di Aosta.

l) Indice di attrattività dei servizi ospedalieri. Emigrazione ospedaliera in altra regione per ricoveri ordinari acuti sul totale delle persone ospedalizzate residenti nella regione. Periodo 1999-2003.

- Asse 6: Reti e nodi di servizio (23 indicatori effettivi su 25 disponibili). L'obiettivo globale dell'Asse è: migliorare e creare le condizioni di contesto (nei trasporti, nella società dell'informazione, nella sicurezza pubblica) per lo sviluppo imprenditoriale e la localizzazione di nuove iniziative e per aumentare la competitività e la produttività strutturale dei sistemi economici territoriali, mediante interventi che assicurino la sostenibilità ambientale, promuovano la riduzione degli impatti (riequilibrio modale nei trasporti), rispettino la capacità di carico dell'ambiente e del territorio in generale e favoriscano i processi di recupero della fiducia sociale.

  a) Indice di criminalità diffusa 1. Furti e rapine meno gravi per $1'000$ abitanti. Periodo 1995-2003.

  b) Indice di criminalità diffusa 2. Furti e rapine meno gravi sul totale dei delitti. Periodo 1995-2003.

  c) Indice di criminalità organizzata (anno 1995=100). Periodo 1995-2003.

  d) Indice di criminalità violenta. Crimini violenti per $10'000$ abitanti. Periodo 1995-2003.

  e) Indice di criminalità minorile (escluso il furto). Totale dei minori denunciati per ogni tipologia di delitto escluso il furto sul totale dei minori denunciati. Periodo 1995-2003.

  f) Indice di criminalità minorile. Indice di criminalità minorile (in percentuale). Periodo 1995-2003.

g) Percezione delle famiglie del rischio di criminalità nella zona in cui vivono. Famiglie che avvertono molto o abbastanza disagio al rischio di criminalità nella zona in cui vivono sul totale delle famiglie Periodo 1995-2004.

h) Tonnellate di merci in ingresso ed in uscita per ferrovia sul totale delle modalità. Periodo 1995-2004. Per il 1999 l'indicatore non è stato calcolato poiché non sono disponibili i dati relativi al traporto merci per ferrovia. Per completare la serie si è utilizzata la semisomma tra il 1998 ed il 2000.

i) Indice del traffico merci su ferrovia. Tonnellate di merci in ingresso ed in uscita per ferrovia per 100 abitanti. Periodo 1995-2004. Per il 1999 l'indicatore non è stato calcolato poiché non sono disponibili i dati relativi al traporto merci per ferrovia. Per completare la serie si è utilizzata la semisomma tra il 1998 ed il 2000.

j) Tonnellate di merci in ingresso ed in uscita su strada sul totale delle modalità. Periodo 1995-2004. Per il 1999 l'indicatore non è stato calcolato poiché non sono disponibili i dati relativi al traporto merci per ferrovia. Per completare la serie si è utilizzata la semisomma tra il 1998 ed il 2000.

k) Indice del traffico merci su strada. Tonnellate di merci in ingresso ed in uscita su strada per abitante. Periodo 1995-2004.

l) Indice del traffico aereo. Passeggeri sbarcati ed imbarcati per via aerea per 100 abitanti. Periodo 1995-2003

m) Grado di soddisfazione del servizio di trasporto ferroviario. Grado di soddisfazione dell'utenza per i servizi di trasporto ferroviario. Periodo 1995-2004.

n) Indice di utilizzazione del trasporto ferroviario 1. Frequenza di utilizzazione dei treni in perc. Periodo 1995-2004.

o) Indice di utilizzazione del trasporto ferroviario 2. Lavoratori, scolari e studenti di 3 anni e più che utilizzano il treno abitualmente per recarsi a lavoro, asilo o scuola sul totale. Periodo 1995-2004.

p) Inquinamento causato dai mezzi di trasporto. Emissioni di CO_2 da trasporto stradale (tonnellate per abitante). Periodo 1996-2002.

q) Indice di diffusione dell'informatizzazione nei comuni. Popolazione residente in comuni con anagrafe collegata al sistema INA-SAIA sul totale della popolazione regionale. Periodo 2002-2004.

r) Grado di diffusione di Internet nelle famiglie. Grado di diffusione di Internet: perc. di famiglie che dichiarano di possedere l'accesso a Internet. Periodo 1997-2004.

s) Grado di utilizzo di Internet nelle famiglie. Grado di utilizzo di Internet: perc. di persone di 6 anni e più che dichiarano di aver utilizzato Internet negli ultimi tre mesi. Periodo 2002-2004.

t) Grado di diffusione del personal computer nelle imprese con più di dieci addetti. Percentuale di imprese (con più di dieci addetti) dei settori industria e servizi che dispongono di personal computer. Periodo 2003-2005.

u) Indice di diffusione dei siti web delle imprese. Percentuale di imprese (con più di dieci addetti) dei settori industria e servizi che dispongono di sito web. Periodo 2003-2005.

v) Indice di diffusione della banda larga nelle imprese. Percentuale di imprese (con più di dieci addetti) dei settori industria e servizi che dispongono di collegamento a banda larga. Periodo 2003-2005.

x) Grado di utilizzo di Internet nelle imprese. Percentuale di addetti delle imprese (con più di dieci addetti) dei settori industria e servizi che utilizzano computer connessi a Internet. Periodo 2003-2005.

1.5.2 Conti economici regionali

Il sistema dei conti economici regionali fornisce una descrizione quantitativa dell'attività economica e finanziaria di una regione. Nella contabilità regionale (così come in quella nazionale) il metodo di registrazione dei flussi economico-finanziari è quello della partita doppia mediante la compilazione di una serie di conti dai quali si evincono la formazione, distribuzione e impiego del reddito, l'accumulazione del risparmio, le forme di finanziamento e le interdipendenze tra persone fisiche e istituzioni che operano su un determinato territorio.

L'adozione del Regolamento Comunitario 2223/96 sul Sistema Europeo dei Conti (SEC95), avvenuta nel 2000, ha comportato l'interruzione delle serie dei conti economici regionali costruite secondo il precedente Sistema (SEC79) che, iniziando dal 1980, erano state aggiornate dall'ISTAT fino al 1996. La disponibilità di nuove serie territoriali solo a partire da tale anno non consentiva di analizzare gli andamenti dell'economia regionale per un arco temporale sufficientemente esteso e costituiva, inoltre, un'evidente limitazione alla costruzione di modelli finalizzati alla produzione di stime anticipate o alla previsione degli andamenti economici territoriali. L'ISTAT e la SVIMEZ, avvertendo queste limitazioni, nel 2003 hanno deciso di intraprendere un'iniziativa di ricerca congiunta per la ricostruzione delle serie dei conti economici regionali coerenti con il nuovo schema dei conti SEC95 per gli anni dal 1980 al 1994. Oltre all'intera serie storica ricostruita 1980-1994, la documentazione statistica contenuta nel volume include i dati relativi agli anni 1995-2003 già elaborati dall'ISTAT (Cfr: Istat [17]). La base di dati comprende per il periodo 1995-2003, il conto delle risorse e degli impieghi e, solo fino al 2001, il conto della distribuzione del reddito. L'analisi a 25 branche (derivate dalla classificazione NACE-Rev.1) è effettuata, per tutto il periodo, sul valore aggiunto, sugli occupati interni (dipendenti e indipendenti) e le corrispondenti unità di lavoro, sui redditi da lavoro dipendente. Per questi ultimi, per il 2002 e il 2003, non è fornita l'analisi nelle componenti di retribuzioni lorde e contributi sociali. Per quanto riguarda gli investimenti fissi lordi l'analisi dei dati a 25 branche è effettuata fino al 2001; per il 2002 è disponibile l'analisi a 6 branche e per il 2003 è limitata alle tre macrobranche agricoltura, industria e servizi. La disaggregazione dei consumi finali delle famiglie nei 12 gruppi standard di beni e servizi è disponibile per l'intero periodo, mentre la disaggregazione nelle 10 funzioni di spesa dei consumi finali delle amministrazioni pubbliche è disponibile solo fino al 2002.

Le serie regionali che l'Istat rende correntemente disponibili sono scaturite dal lavoro di adozione del SEC95 concluso nel 2000 e coprono gli anni a partire dal 1995. Nel corso del 2004, l'Istat, in cooperazione con la Svimez, ha concluso e pubblicato la ricostruzione delle serie storiche regionali dal 1980 al 1994 in chiave con le nuove serie SEC95 (Vedi: Istat, Svimez, "Conti economici regionali 1980-1994; serie storiche ricostruite secondo il SEC95" , 16 febbraio 2004.). Queste "serie ricostruite" hanno minori disaggregazioni rispetto alle "serie complete"

che vengono aggiornate ogni anno, ma sono con esse coerenti. Per comodità degli utilizzatori è stato perciò creato un unico file "Serie storiche", nel quale le "serie complete" aggiornate annualmente, vengono riunificate con il segmento 1980-1994 ricostruito, dopo essere ricondotte alle sue stesse analisi.http://www.istat.it/dati/dataset/20051025_00/. In definitiva, possiamo contare su di un periodo di copertura che va dal 1980 e arriva al 2003 (ad eccezione della macrodeterminante: "Investimenti fissi per branca proprietaria" in cui il 2003 risulta mancante).

Conto delle risorse e degli impieghi. (5 indicatori). E' il più importante tra i conti economici e riassume la situazione macroeconomica della regione mettendo in evidenza l'equilibrio tra l'offerta e la domanda finale. L'offerta è rappresentata dalle risorse (somma di prodotto interno lordo ed importazioni dall'estero) mentre la domanda finale è rappresentata dagli impieghi (consumi finali interni, investimenti fissi lordi ed esportazioni verso l'estero).

a) Importazioni nette
b) Investimenti fissi lordi
c) Prodotto interno lordo
d) Variazione delle scorte e oggetti di valore
e) Consumi finali interni

Valore aggiunto ai prezzi base e PIL. (11 indicatori). E' calcolato attraverso indagini sulle imprese e bilanci delle società di capitali. E' considerato esaustivo in quanto contiene una stima del sommerso, proprio perché è calcolato attraverso l'input di lavoro. E' un indicatore della crescita del sistema economico in termini di nuovi beni e servizi prodotti e messi a disposizione della società per impieghi finali.

a) Agricoltura, silvicoltura e pesca
b) Industria in senso stretto
c) Costruzioni, Commercio, riparazioni, alberghi e ristoranti, trasporti e comunicazioni
d) Intermediazione monetaria e finanziaria; attività immobiliari ed imprenditoriali
e) Altre attività di servizi
f) Valore aggiunto ai prezzi base (al lordo SIFIM)
g) Servizi di intermediazione finanziaria indirettamente misurati
h) Valore aggiunto a prezzi base (al netto SIFIM)
i) IVA, imposte indirette nette sui prodotti e imposte sulle importazioni
j) Prodotto interno lordo ai prezzi di mercato

Consumi delle famiglie. (6 indicatori). Spese delle famiglie (a prezzi correnti): rilevate attraverso l'indagine sui consumi delle famiglie. Questa determinante ha lo scopo di rilevare la struttura ed il livello dei consumi secondo le principali caratteristiche sociali, economiche e territoriali delle famiglie residenti, consentendo di conoscere e seguire l'evoluzione, in senso qualitativo

e quantitativo, degli standard di vita e dei comportamenti di consumo delle principali tipologie familiari, in riferimento ai differenti ambiti territoriali. L'indagine rileva l'ammontare complessivo della spesa al momento dell'acquisto del bene, a prescindere dal momento dell'effettivo consumo o utilizzo di questo e dalle modalità di pagamento. In particolare, sono rilevate notizie che riguardano gli individui che compongono la famiglia, caratteristiche dell'abitazione, reddito e risparmio, oltre alle spese per generi alimentari, abitazione, arredamento, abbigliamento e calzature, salute, trasporto e comunicazioni, tempo libero spettacoli e istruzione, altri beni e servizi. E' incluso l'autoconsumo.

a) Abitazione e spese connesse
b) Alimentari, bevande e tabacco
c) Istruzione, cultura ed altri beni e servizi
d) Spese sanitarie
e) Trasporti e comunicazioni
f) Vestiario e calzature

Investimenti fissi per branca proprietaria. (6 indicatori effettivi). Questa determinante raccoglie informazioni sul valore di beni durevoli destinati a fini non militari, acquistati dalle unità produttive residenti per essere utilizzati durante un periodo superiore ad un anno nel processo produttivo, nonché il valore dei servizi incorporati di beni di investimento acquistati.

a) Agricoltura, silvicoltura e pesca
b) Altre attività di servizi
c) Commercio, riparazioni, alberghi e ristoranti, trasporti e comunicazioni
d) Costruzioni
e) Industria in senso stretto
f) Intermediazione monetaria e finanziaria; attività immobiliari ed imprenditoriali

Investimenti fissi per branca produttrice. (2 indicatori effettivi)

a) Costruzioni
b) Macchine, attrezzature, mezzi di trasporto e altri prodotti

Redditi da lavoro dipendente. (6 indicatori). Redditi da lavoro dipendente (a prezzi correnti): calcolati attraverso indagini sulle imprese e dati ricavati dall'archivio INPS dei DM 10 (moduli INPS per il pagamento dei contributi) e localizzati in base alla regione in cui sono guadagnati. Costituiscono il costo sostenuto dai datori di lavoro per la remunerazione dell'attività lavorativa fornita dai propri dipendenti.

a) Agricoltura, silvicoltura e pesca
b) Altre attività di servizi
c) Commercio, riparazioni, alberghi e ristoranti, trasporti e comunicazioni

d) Costruzioni

e) Industria in senso stretto

f) Intermediazione monetaria e finanziaria; attività immobiliari ed imprenditoriali

Unità di lavoro dipendenti. (6 indicatori). Occupati dipendenti, indipendenti e totali (media annua): calcolati attraverso le indagini sulle forze di lavoro, l'archivio Asia e l'indagine sui conti delle imprese. E' il numero di lavoratori che presta la propria attività all'interno della regione.

a) Agricoltura, silvicoltura e pesca

b) Altre attività di servizi

c) Commercio, riparazioni, alberghi e ristoranti, trasporti e comunicazioni

d) Costruzioni

e) Industria in senso stretto

f) Intermediazione monetaria e finanziaria; attività immobiliari ed imprenditoriali

Unità di lavoro indipendenti. (6 indicatori)

a) Agricoltura, silvicoltura e pesca

b) Altre attività di servizi

c) Commercio, riparazioni, alberghi e ristoranti, trasporti e comunicazioni

d) Costruzioni

e) Industria in senso stretto

f) Intermediazione monetaria e finanziaria; attività immobiliari ed imprenditoriali

1.6 Risultati dell'analisi

La necessità di costituire basi di dati a livello locale è ormai una prassi evidente e incontrastata nel nostro Paese e nel corso del tempo non sono mancate risposte adeguate da parte degli organi istituzionalmente preposti ed enti di ricerca. L'articolazione territoriale deriva sia dalle esigenze di studio, che da quelle, più operative, della politica economica; sono molti, infatti, i provvedimenti che, tanto a livello europeo che nazionale e regionale, fanno oramai riferimento a sistemi economici limitati territorialmente. Un aspetto che ancora rimane da inserire a pieno titolo è il tempo cioè dati in forma di serie storica, pur se breve. L'interesse per lo studio delle dinamiche economiche regionali è stato a lungo ostacolato dalla carenza di dati utilizzabili. È in questo quadro che è maturata l'idea di incrementare la quantità di informazioni economiche a livello locale integrando con la loro dimensione congiunturale. Ne è scaturita una metodologia che lungi dal separare struttura e congiuntura (talvolta risultata utile in alcune semplificazioni del funzionamento del sistema economico) fonde in una tecnica unica le esigenze di analisi statica e di analisi dinamica.

La nostra indagine, in particolare, si è concentrata sull'uso congiunto degli indicatori delle due banche dati regionali disponibili in forma di serie storica. In Bonaccorsi e Gallo [4] è riportata una analisi più dettagliata in cui si confrontano contemporaneamente le venti regioni italiane sulla base di un singolo indicatore (nella banca dati Q.C.S.) monitorato durante l'ultimo triennio disponibile.

Nel caso della base di dati Q.C.S. la classificazione si è basata su di un totale di 122 indicatori, la base di dati C.E.R. ne include invece 48. Infine, si è anche deciso di aggregare le due basi di dati ed effettuare una classificazione su tutte le 170 serie storiche disponibili per questa indagine, pur nella consapevolezza della loro diversa natura e finalità.

1.6.1 Analisi in blocco

Le due analisi in blocco producono risultati abbastanza simili. C'è una divisione piuttosto netta tra due gruppi principali: il primo riconducibile alle regioni ricadenti nell'obiettivo 1 ed il secondo costituito dalla maggioranza delle regioni del Centro-Nord. Esistono poi delle unità isolate, in genere le regioni più piccole in termini di superficie territoriale e popolazione residente, che formano degli autogruppi o singoletti: Val d'Aosta, Molise, Basilicata. L'isolamento di una singola area rispetto ai territori contigui impedisce forse di raggiungere quella massa critica che permette di essere visibili nel contesto di una competizione crescente a scala globale e che richiede a ciascuna area la messa in opera di strategie di cooperazione almeno con i bacini territoriali più vicini.

Una disamina più dettagliata dei due dendrogrammi rivela poco di più. Per il dati del quadro comunitario di sostegno (QCS) la Liguria è un singoletto mentre si ritrova accorpata al gruppo virtuale formato dai dati aggregati Italia e aggregati

Non obiettivo 1. Le regioni più rappresentativa dei due gruppi principali sono L'Emilia-Romagna e la Sicilia per i dati Q.C.S. Il centro dei cluster cambia per i dati C.E.R. Per il gruppo Centro-Nord diventa la Liguria e per il Sud-Isole è ora la Puglia. Dal club delle regioni obiettivo 1 manca solo la Basilicata (poco spiegabile se non per le dimensioni ridotte) ed il Molise (di cui è però già nota la natura di regione di transizione). La qualità buona della classificazione è confermata dalla collocazione coerente delle unità virtuali "Obiettivo 1" e "Non obiettivo 1" che nella figura [1.16] appaiono molto distanti e solo la seconda è inserita in uno stesso cluster insieme ai dati dell'Italia intera. Questo non è metodologicamente sbagliato, ma sorprende il fatto che, per i dati C.E.R., non ci sia nessuna delle regioni che stia bene al centro delle regioni "Non obiettivo 1". Questa rappresentatività è invece riscontrata nella figura [1.15] in cui l'Emilia-Romagna è l'unità delegabile a descrivere le regioni fuori dall'obiettivo 1.

Uno degli obiettivi prioritari che si è posta l'Unione Europea è quello di ridurre in maniera effettiva e consistente le disparità regionali, di modo che le diverse aree europee siano in grado di cooperare e dialogare nel campo economico. Ora, nel caso dei ritardi strutturali di una regione, è ovvio che non si tratta di ritardi dovuti a difficoltà occasionali o contingenti, ma ad una penalizzazione profonda del sistema economico che ha origine nella minore dotazione dei fattori dello sviluppo potenziale. Questo obiettivo non sembra certo raggiunto alla luce di ciò che emerge dalle due analisi in blocco che confermano il divario tra le due tipologie di regioni.

L'evidenza di quanto affermato scaturisce in tutta la sua portata nel dendrogramma relativo alle 170 serie storiche delle due banche dati considerate congiuntamente ed il cui esito di classificazione è stato riportato nella cartina 1.18. La struttura dei gruppi diventa più netta e compatta con divari meglio specificati. Le regioni ricadenti nell'insieme "Obiettivo 1" sono state collocate in uno stesso cluster, senza esclusioni anche se Basilicata e Molise sembrano spostarsi verso un altro gruppo. Infatti, le regioni "Non obiettivo 1" sono state segmentate in due gruppi principali. Il primo è formato da Abruzzo, Marche, Umbria, Fiurli-Venezia Giulia, Liguria e Trentino Alto Adige che ha caratteristiche medie rispetto all'intera nazione in quanto è incentrato sulle unità virtuali "Media delle regioni non obiettivo 1" e "Media di tutte le regioni italiane". Il secondo gruppo è formato da regioni che si stanno avvicinando all'Europa date le caratteristiche che mostrano negli indicatori principali nei due dati set (cfr.). La Val d'Aosta forma un singoletto caratterizzato soprattutto dalle sue ridotte dimensioni. Infine c'è la Lombardia che forma un gruppo a parte ed è forse l'unica regione che può entrare a pieno titolo nel clube delle regioni europee più ricche e sviluppate. La Calabria si aggrea subito (è quindi molto vicina) alla Puglia che è poi la regione emblematica del cluster formato dalle regioni "Obiettivo 1".

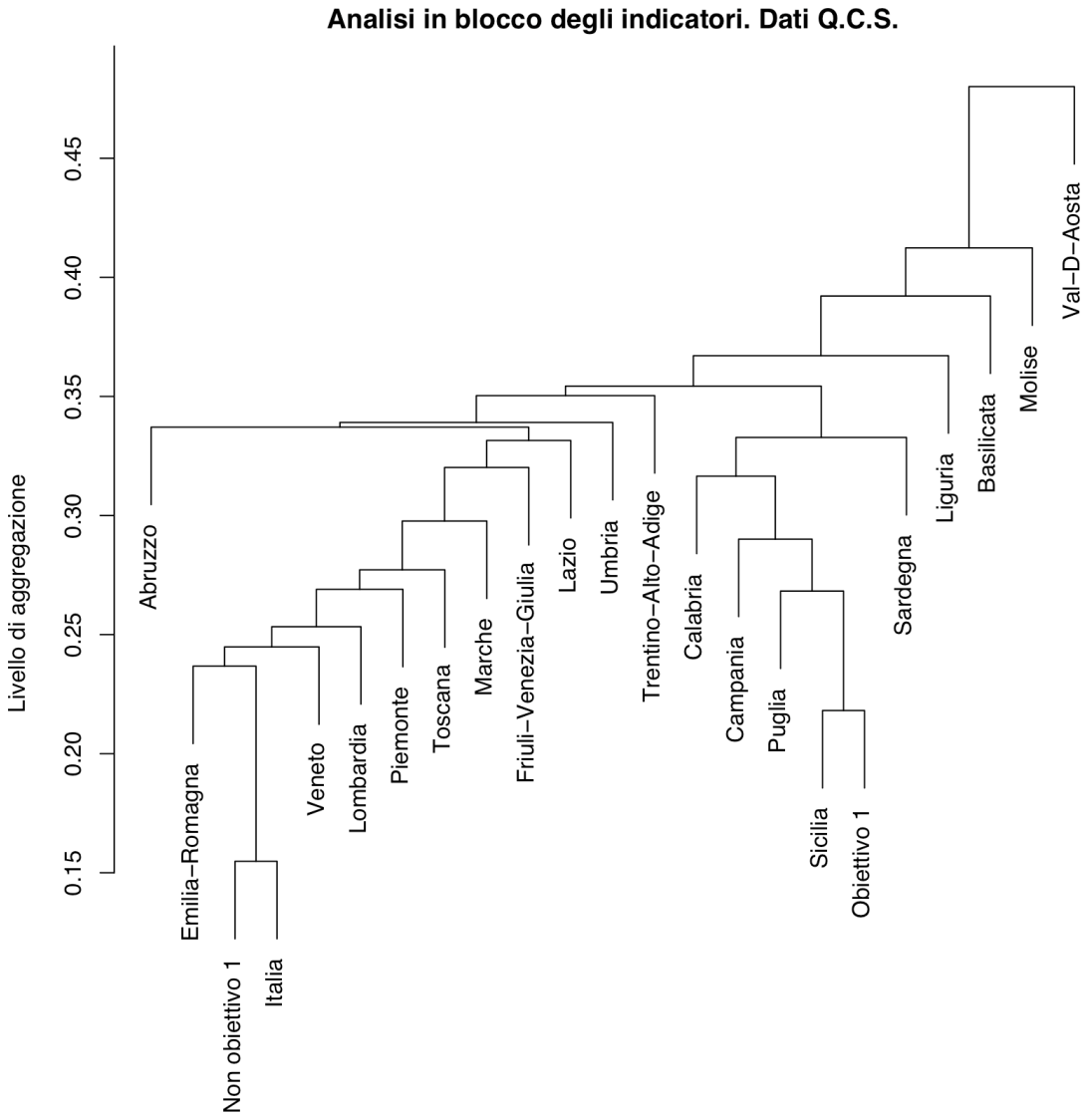


Figura 1.15: Esito della classificazione. Analisi in blocco. Dati Q.C.S.

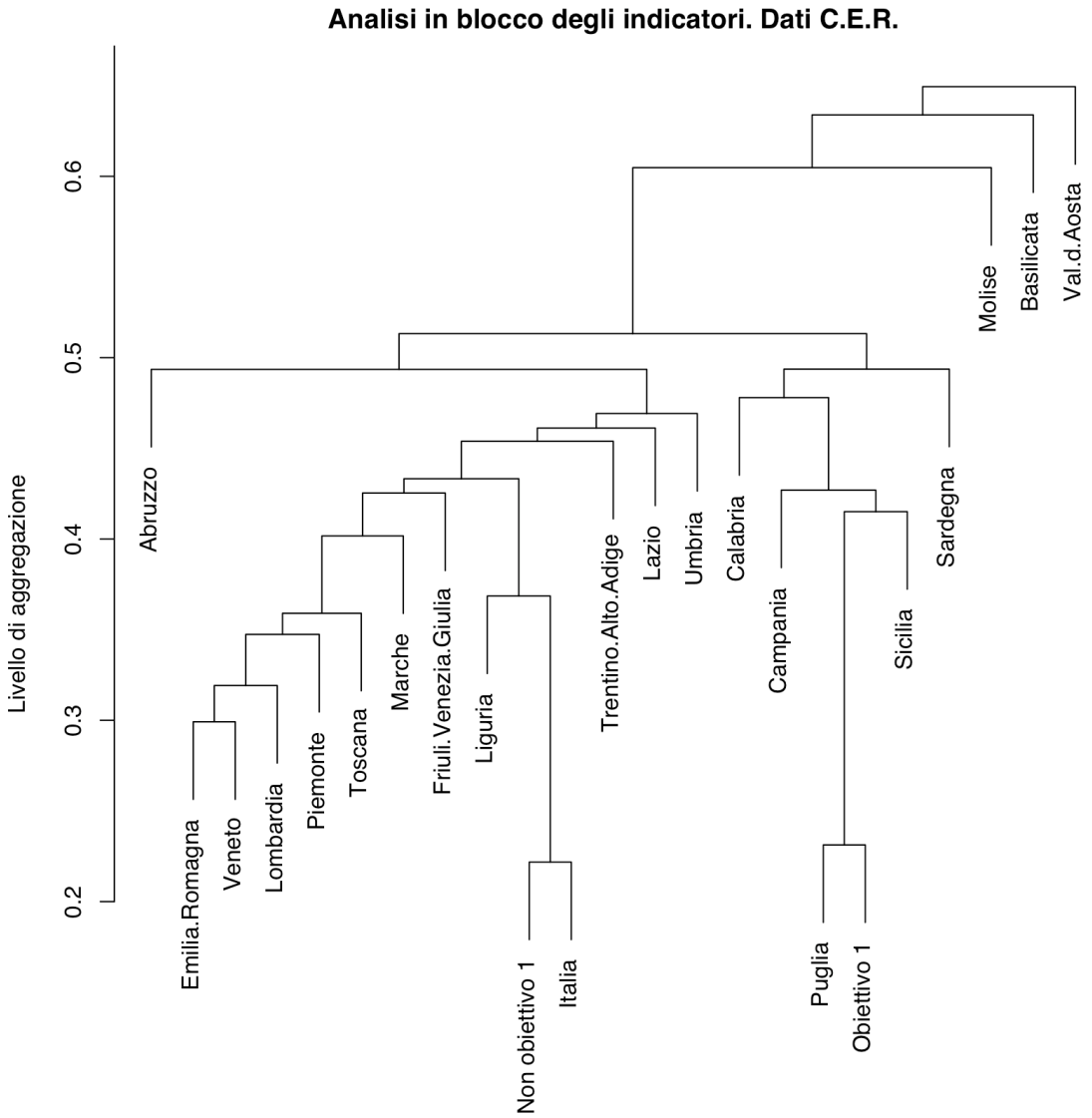


Figura 1.16: Esito della classificazione. Analisi in blocco. Dati C.E.R.

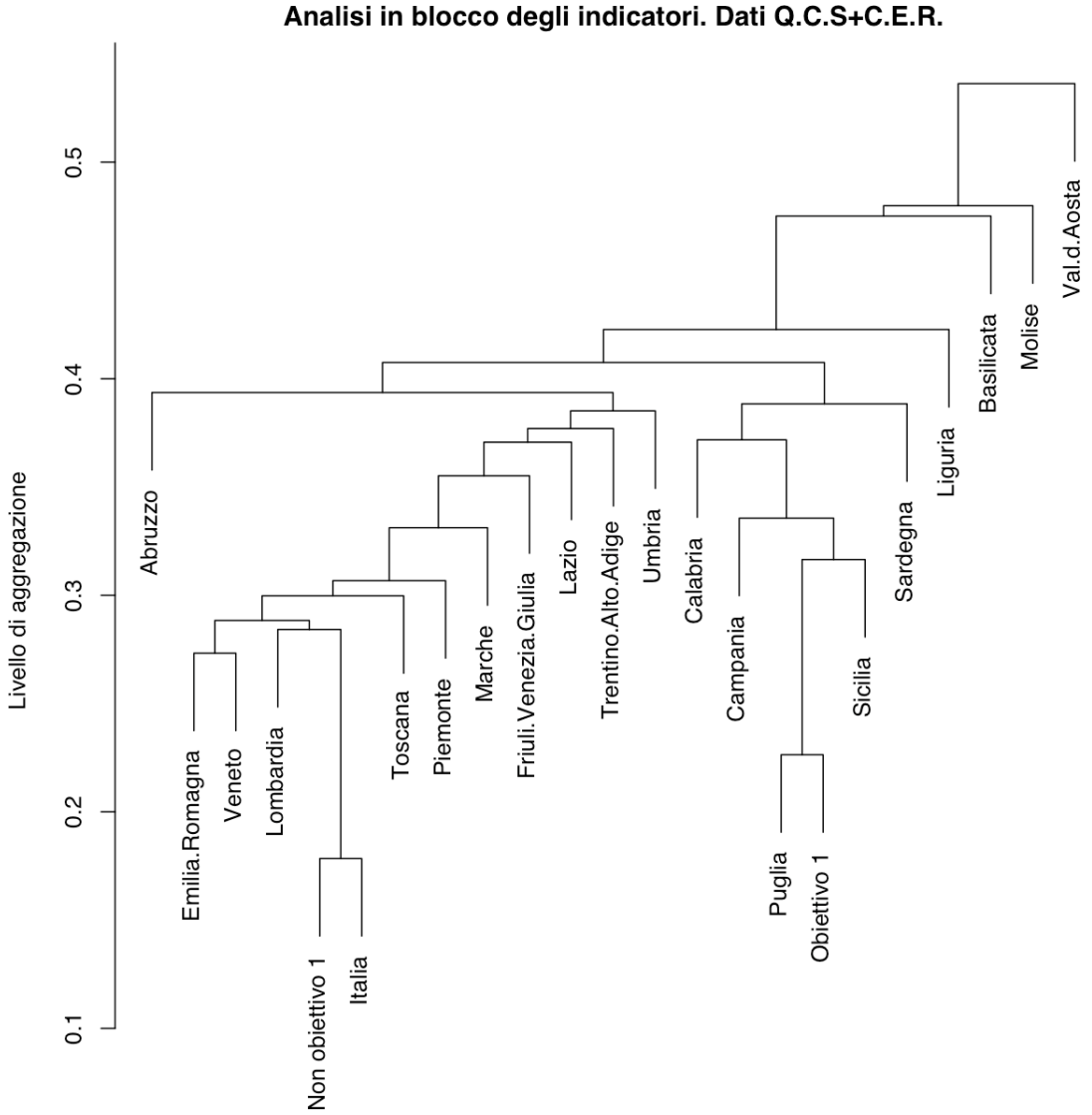


Figura 1.17: Esito della classificazione. Analisi in blocco. Dati Q.C.S.+C.E.R.

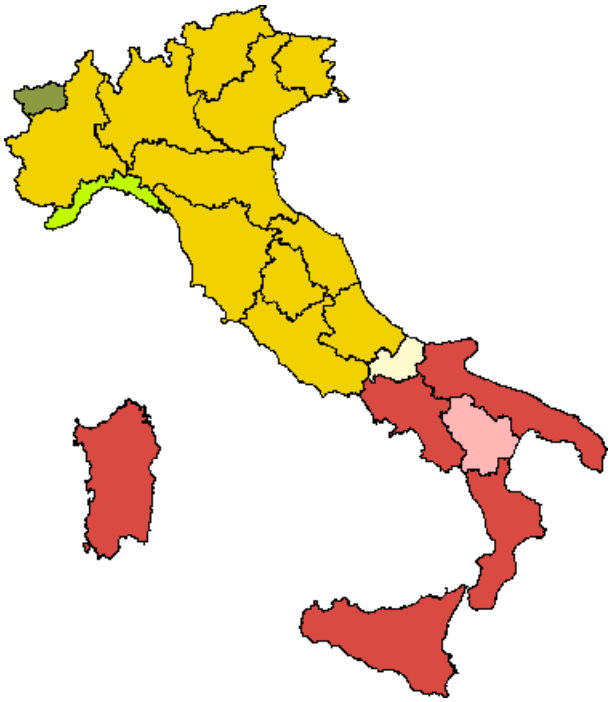


Figura 1.18: raggruppamenti delle regioni regioni in base a 170 serie storiche macroeconomiche

1.6.2 Analisi per blocchi

La disponibilità di statistiche di qualità è un elemento cruciale nei processi di programmazione e di valutazione. Dati e indicatori a livello territoriale disaggregato possono avere un ruolo fondamentale in quanto aiutano a comprendere meglio le peculiarità dei territori e ad accrescere la consapevolezza di chi ha responsabilità politiche ed amministrative. Con l'affinarsi della *governance* i bisogni informativi sono in crescita ed il flusso di dati, soprattutto a livello regionale ha già raggiunto una notevole consistenza.

Nella tabella 1.4 è stato riportato il coefficiente di rappresentatività della prima componente prodotta dal metodo Distatis per i dati Q.C.S. Non si può fare a meno di notare il coefficiente elevato per l'Asse 2: risorse culturali. Tuttavia, occorre correggere l'impressione di prevalenza. Infatti, tale indicazione è molto legata al numero di indicatori introdotti nella macrodeterminante e in alcuni casi, il coefficiente di rappresentatività è inferiore alla percentuale media di rappresentazione ottenibile con uno qualsiasi degli indicatori in essa ricompresi. Il grafico in figura 1.19 conferma lo scentramento deline lo scentramento del blocco di indicatori ricompresi nell'Asse 2: "Risorse culturali" segnalando un forse ancora incompleta maturità di definizione nell'ambito delle variabili riferite al Quadro

Comunitario di Sostegno. Le altre macrodeterminanti sono raggruppate in un unico cluster in cui si può evidenziare una insospettabile prossimità tra il blocco di indicatori "Variabili di rottura" e "Risorse naturali".

Tabella 1.4: Coefficiente di rappresentatività del fattore ponderale. Dati Q.C.S.

Asse	Denominazione	Rappresentatività	$\frac{100}{k_h}$
0	Variabili di rottura	7.52	7.14
1	Risorse naturali	8.05	6.67
2	Risorse culturali	24.35	25.00
3	Risorse umane	3.91	3.85
4	Sistemi locali di sviluppo	2.11	3.57
5	Città	9.42	8.33
6	Reti e nodi di servizio	4.94	4.17

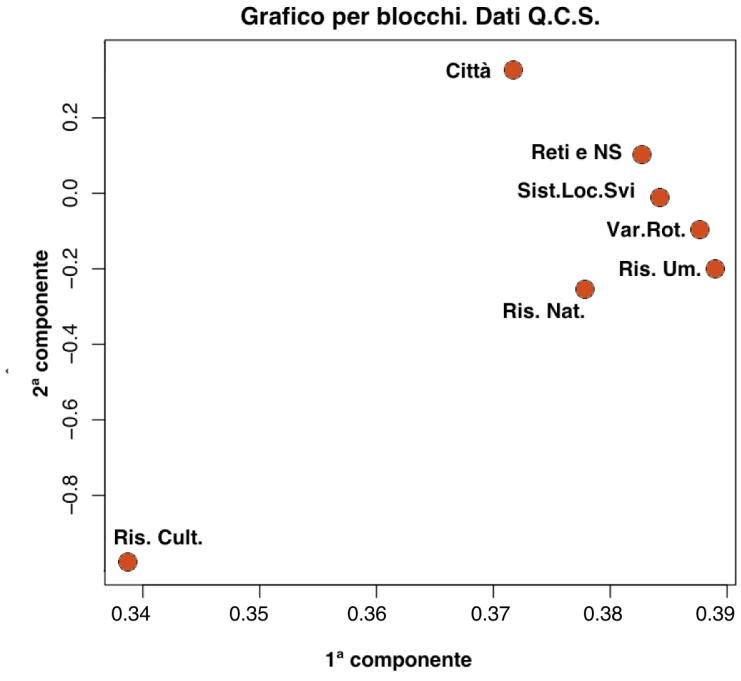


Figura 1.19: Pesi fattoriali nella componente trasversale. Dati Q.C.S.

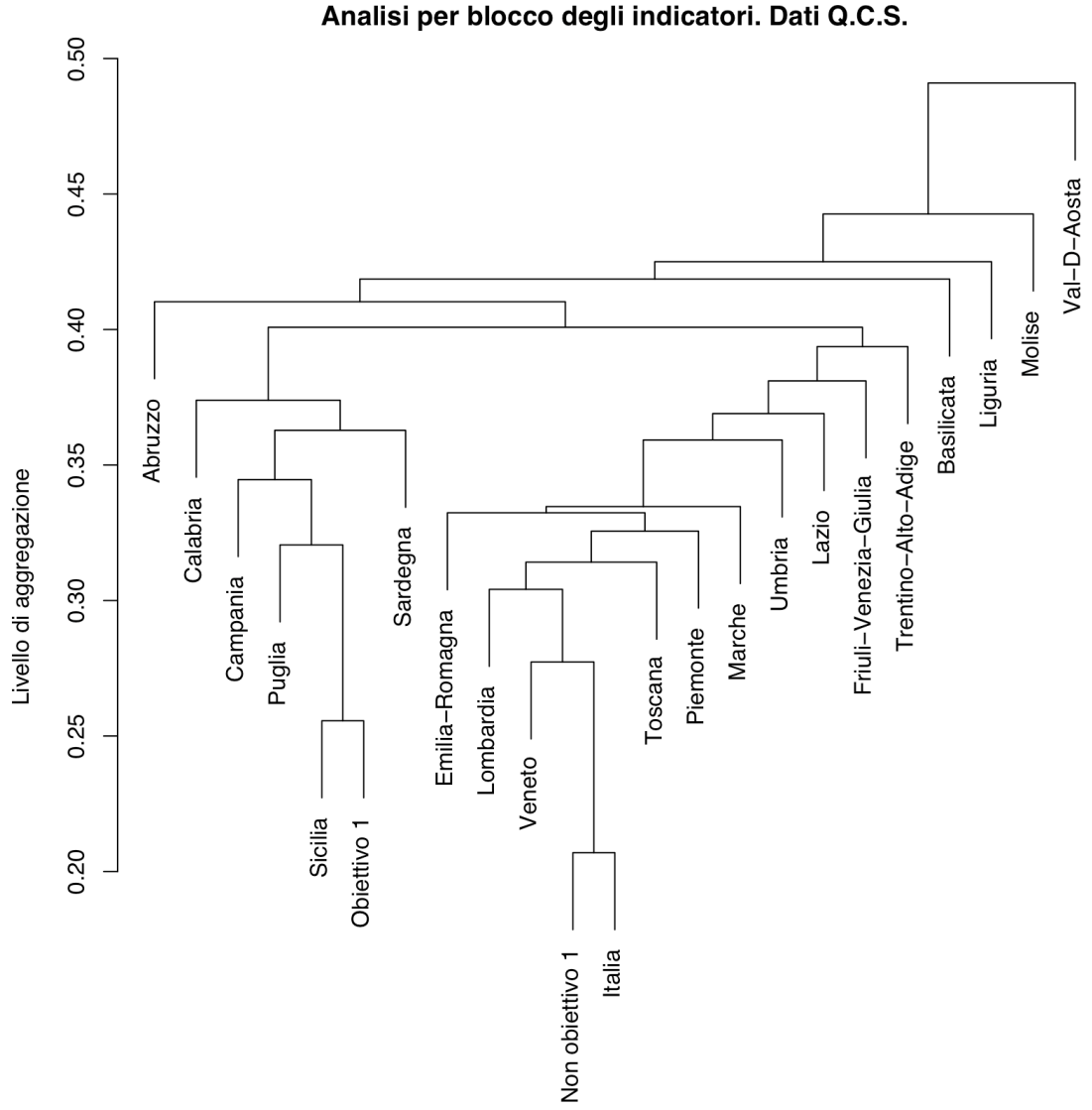


Figura 1.20: Esito della classificazione. Analisi per blocchi. Dati Q.C.S.

Si riscontra una sola novità rispetto alla classificazione in blocco: l'Abruzzo è ora in una posizione più baricentrica tra le regioni dell'Obiettivo 1 e quelle fuori dall'obiettivo.

Le macrodeterminanti della seconda base di dati, la C.E.R., sono più disperse nel piano dei fattori comuni delle distanze parziali. Si nota la prevedibile vicinanza tra il blocco "Redditi da lavoro dipendente" e "Valore aggiunto" nonché la altrettanto prevedibile distanza tra "Investimenti fissi per branca proprietaria" e "Investimenti fissi per branca produttrice". Inaspettata giunge forse la vicinanza tra "Unità di lavoro dipendenti" e "Unità di lavoro indipendenti".

Tabella 1.5: Coefficiente di rappresentatività del fattore ponderale. Dati C.E.R.

Blocco	Denominazione	Rappresentatività	$\frac{100}{k_h}$
1	Risorse ed impieghi	19.16	20.00
2	Valore aggiunto e PIL	8.39	9.01
3	Consumi delle famiglie	17.47	16.67
4	Investimenti fissi per branca proprietaria	17.33	16.67
5	Investimenti fissi per branca produttrice	50.00	50.00
6	Redditi da lavoro dipendente	16.39	16.67
7	Unità di lavoro dipendenti	15.97	16.67
8	Unità di lavoro indipendenti	14.69	16.67

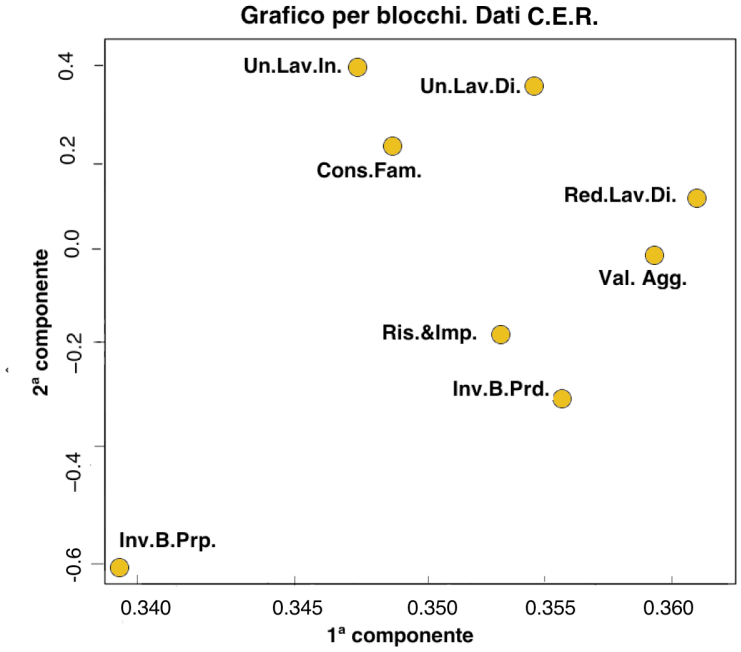


Figura 1.21: Pesi fattoriali nella componente trasversale. Dati C.E.R.

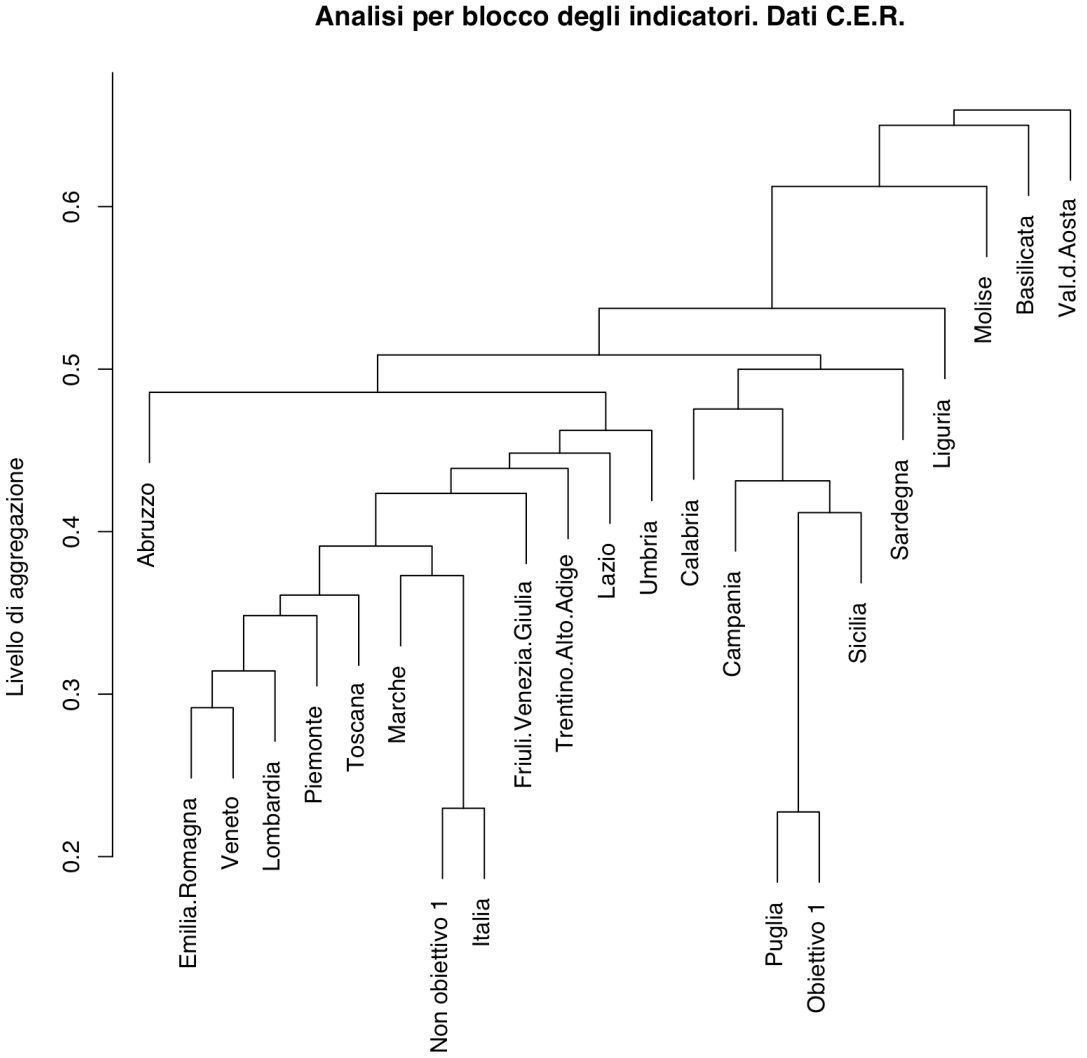


Figura 1.22: Esito della classificazione. Analisi per blocchi. Dati C.E.R.

La sola modifica che interviene nella composizione dei gruppi rispetto alla classificazione in blocco è la posizione della Liguria che era ben inserita nel gruppo delle regioni non obiettivo 1 ed ora si confifura come un singoletto.

Abbiamo ritenuto utile una analisi separata per ciascun blocco anche per le due basi congiuntamente. Dalla base di dati Q.C.S. abbiamo però escluso l'Asse 2: "Risorse culturali in quanto poco coerente con gli altri indicatori. Risultano confermate tutte le relazione emerse a livello di singola analisi

Tabella 1.6: Coefficiente di rappresentatività del fattore ponderale. Basi Unificate

Blocco	Denominazione	Rappresentatività	$\frac{100}{k_h}$
1	Risorse ed impieghi	19.16	20.00
2	Valore aggiunto e PIL	8.39	9.01
3	Consumi delle famiglie	17.47	16.67
4	Investimenti fissi per branca proprietaria	17.33	16.67
5	Investimenti fissi per branca produttrice	50.00	50.00
6	Redditi da lavoro dipendente	16.39	16.67
7	Unità di lavoro dipendenti	15.97	16.67
8	Unità di lavoro indipendenti	14.69	16.67
0	Variabili di rottura	7.52	7.14
1	Risorse naturali	8.05	6.67
3	Risorse umane	3.91	3.85
4	Sistemi locali di sviluppo	2.11	3.57
5	Città	9.42	8.33
6	Reti e nodi di servizio	4.94	4.17

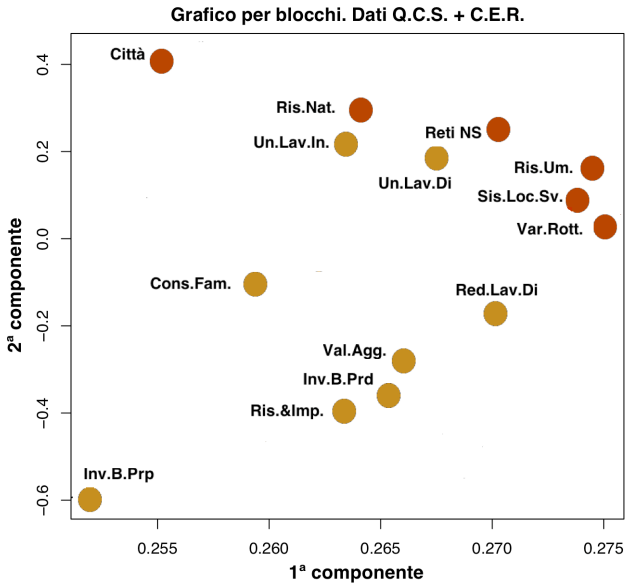


Figura 1.23: Pesi fattoriali nella componente trasversale. Dati Q.C.S.+C.E.R.

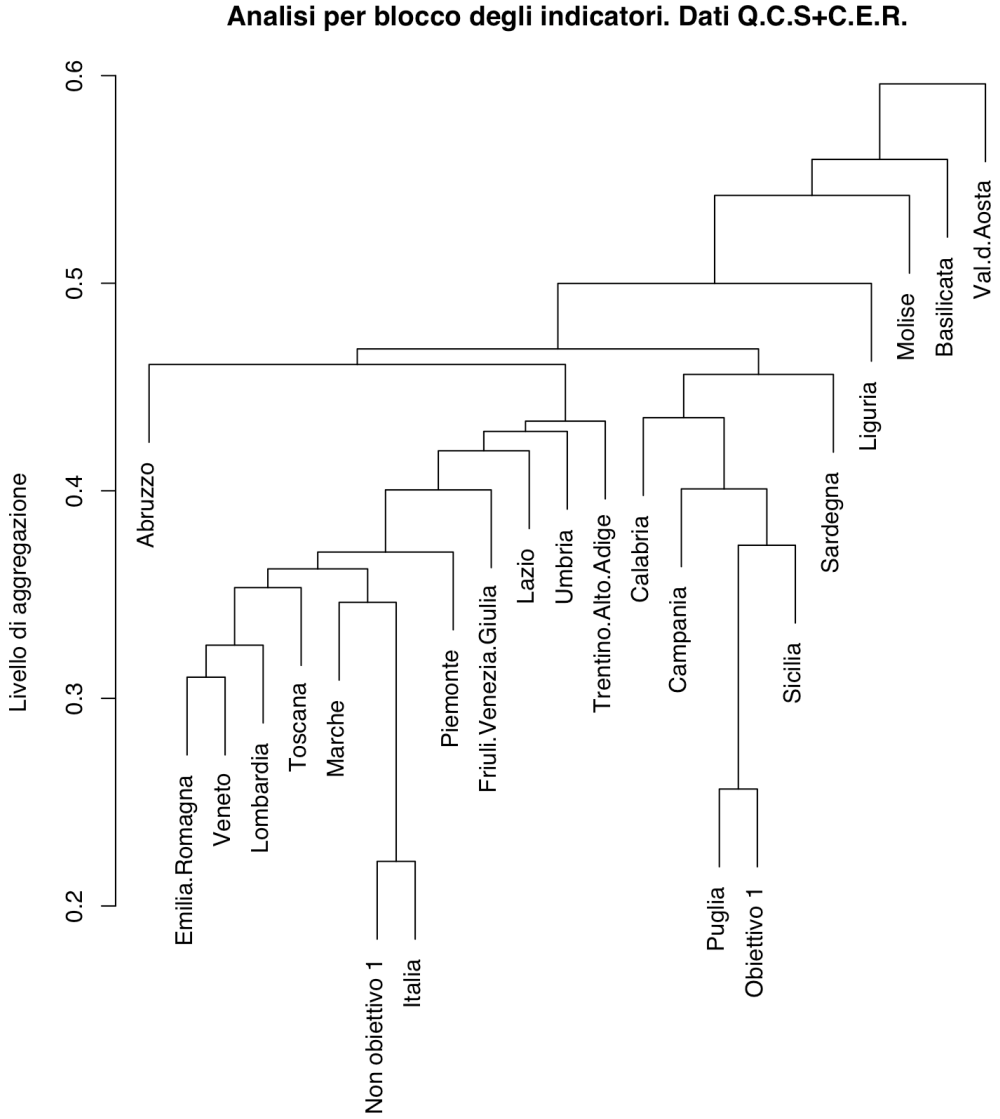


Figura 1.24: Esito della classificazione. Analisi per blocchi. Dati Q.C.S.+C.E.R.

I dendrogrammi dei dati aggregati, in blocco e per blocchi, sono simili rivelando sono alcuni scambi nei gruppi peraltro coinvolgenti unità molto ravvicinate in entrambe le classificazioni.

La mappatura delle regioni italiane che risulta dall'analisi degli indicatori segnala che persiste una netta distinzione tra la dinamica delle regioni obiettivo 1 e le altre. Al netto di alcune regioni quali Molise, Basilicata e Val d'Aosta che si sono configurate come situazioni discoste dalle tendenze in atto, si può dire che le politiche comunitarie di accompagnamento alle dinamiche di crescita localizzata e le concomitanti azioni delle singole regioni, all'epoca prevalente dei dati: 2004, non hanno portato a significative uscite dal gruppo delle regioni arretrate.

1.6.3 Conclusioni

Michelangelo sosteneva che la scultura fosse già all'interno del blocco di marmo e che il compito dell'artista fosse solo quello di farla uscire. Più bravo l'artista, più efficaci i suo attrezzi e meno la statua perde della sua bellezza originale quando appare agli occhi del pubblico. L'analisi statistica dei dati è lo strumento di cui si servono i ricercatori per delineare ed isolare le strutture le tendenze che giacciono nelle rilevazioni empiriche. In questo lavoro sono state utilizzate, separatamente e congiuntamente, due basi di dati orientate alle regioni italiane.

Q.C.S. Indicatori Istat per il quadro comunitario di sostegno

C.E.R. Indicatori istat per i conti economici regionali

La complessità delle tecniche cresce con l'aumentare delle dimensioni dei dataset e della loro contaminazione; sono richieste sensibilità e competenze per ottenere un'opera ben fatta invece di un abbozzo informe.

In questo lavoro abbiamo sviluppato una nuova tecnica per misurare la distanza tra due serie storiche brevi impegnando le loro matrici dei tassi di crescita. Inoltre, abbiamo generalizzato la tecnica Distatis per la sintesi di studi diversi distinguendone l'approccio dell'analisi in blocco di tutti gli indicatori da un approccio in cui gli indicatori sono sintetizzati per blocchi distinti. In questo modo sono state ottenute ben sei diverse classicazioni delle regioni italiane fondate su dati tendenziali relativi a più di vent'anni.

Nella biologia evolutiva le *chance* di sopravvivenza di un individuo tipicamente dipendono dal livello critico di uno specifico tratto fisico. Se le *chance* a favore del mantenimento in vita si avvicinano al massimo, allora quel tratto si evolve verso il suo valore ottimale. Nei dendrogrammi presentati nei paragrafi precedenti si è visto che la regione Basilicata si colloca in un singoletto estraneo sia alle regioni Obiettivo 1 che alle altre regioni. In questa fase della ricerca possiamo solo stabilire che non è più nel primo gruppo; quello che non sappiamo è se si è spostata all'indietro oppure talmente avanti da posizionarsi tra le regioni più avanzate. Se un'analisi di statica comparata confermasse la seconda ipotesi, si potrebbero individuare gli indicatori che hanno favorito questa transizione.

La legge dei minimi in ecologia (*Liebig's law*)[24] afferma: la velocità di un processo di sviluppo che dipende da molti fattori è governato dal fattore più lento. In altri termini, il livello di sviluppo che si raggiunge ha vincoli potenziali in ognuno dei fattori che lo determinano, ma il limite più stringente (*limiting factor*) è costituito dal fattore meno disponibile. La legge dei minimi implica anche un fattore vincolo potenziale che non sempre è operativo o potrebbe esserlo solo in alcune fasi dello sviluppo, ma se diventa effettivo contribusice a determinare il processo di crescita. Ogni fattore requisito può divenire un fattore limitante, ma in ogni dato istante solo un fattore requisito è il fattore limitante attivo: quello presente in misura più ridotta. Nel caso in cui la disponibilità del fattore si azzeri e non possa essere surrogato da altri, il processo si arresta. Se la disponibilità

aumenta ci sarà la tendenza ad utilizzarlo fino ad esaurirlo. La Calabria è collocata nel cluster delle regioni Obiettivo 1 in tutte le sei classificazioni. Se esiste un fattore limitante (o più fattori limitanti) devono essere caratteristiche comuni a queste regioni (e non specifici della Calabria) ed è appunto negli indicatori che discriminano più efficaciamente i due gruppi maggiori di regioni che bisognerà cercare tali fattori.

Bibliografia

[1] Abdi Herve, Valentin D., O'toole A., Edelman B. (2005): DISTATIS: the analysis of multiple distance matrices. *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.*

[2] Anania G., Tarsitano A. (1995): Tecniche di analisi statistica multivariata per l'individuazione dei "sistemi agricoli territoriali in Italia. In *I sistemi territoriali agricoli italiani degli anni '90. Contributi metodologici.*(A cura di G.Cannata). Rubbettino Editore, Soveria Mannelli (Cz). 105-242.

[3] Andrews D.F. (1972): Plots of high-dimensional data. *Biometrics*, **28**, 125-136.

[4] Bonaccorsi A., Gallo G. (a cura di) (2006): *Un benchmarking regionale su scala europea*, Confindustria Toscana, Centro Studi.

[5] A periodogram-based metric for time series classification. *Computational statistics and data analysis*, **50** 2668-2684.

[6] Chen Z. Van Ness J.(1998): Characterization of nearest and furthest neighbor algorithms by clustering admissibility conditions. *Pattern Recognition*, **31**,1573-1578.

[7] D'Urso P. (2000): Dissimilarity measures for time trajectories. *Journal of the Italian Statistical Society*, **9**, 53-83.

[8] D'Urso P., Vichi M. (1998). Dissimilarities between trajectories of a three-way longitudinal data set. In A. Rizzi, M.Vichi, H.H. Bock (eds.) *Advances in Data Sciences and Classification*, 585-592, Springer-Verlag, Berlin.

[9] Escoufier, Y. (1980): L'analyse conjointe de plusieursmatrices de données. In M. Jolivet (Ed.), *Biométrie et Temps*. Paris: Société Française de Biométrie. 59–76.

[10] Everitt B., Landau S., Leese M. (2001): *Cluster analysis. 4th edition.* Arnold, London

[11] Ferlaino F., Molinari P. (2006): Indagine sull'organizzazione amministrativo-territoriale in Italia. Il Piemonte nel quadro nazionale e comunitario. IRES Piemonte.

[12] Gantmacher F.B. (1959). *The theory of matrices, Vol. 2.* Chelsea Publishing Company, New York.

[13] Gower J.C., Legendre P. (1986): Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, **3**, 5-48.

[14] Greenacre M. (1984): Theory and applications of correspondence analysis. Academic Press, London.

[15] Heckman N.E., Zamar R.H. (2000): Comparing the shapes of regression functions. *Biometrika*,**87**, 135-144.

[16] *Informazione statistica territoriale e settoriale per le politiche strutturali 2001-2008.*

[17] *Conti economici regionali 1980-1994: serie storiche ricostruite secondo il SEC95*, 16 febbraio 2004,

[18] Jardine N., Sibson R. (1968): The construction of hierarchic and nonhierarchic classification. *Computer Journal*, **11**, 177-184.

[19] Korth B., Tucker L.R. (1976): Procustes matching by congruence coefficients. *Psychometrika*, **41**, 531–535.

[20] Lance G.N., Williams W.T. (1967): A general theory of classification sorting strategies. II clustering systems. *Computer Journal*,**10**, 271-277.

[21] Lavit C., Escoufier Y., Sabatier R., Traissac, P. (1994): The ACT (STATIS) method. *Computational Statistics & Data Analysis*, **18**, 97–119.

[22] Levenbach H., Reuter B.E. (1976): Forecasting trending time series with relative growth rate models. *Technometrics*, **18**, 261-272.

[23] Liao T.W. (2005): Clustering of time series data. A survey. *Pattern Recognition*, **38**, 1857-1874.

[24] Liebig J. (1840): *Chemistry in its Application to Agricolture and Physiology.* Taylor and Walton, London.

[25] Milligan G.W. (1981): A review of Monte Carlo tests of cluster analysis. *Multivariate Behavioural Research*, **16**, 379-407.

[26] Möller-Levert C.S., Klawonn F., Cho K.H., Wolkenhauer O. (2005): Fuzzy clustering of short time-series and unevenly distributed sampling points.

[27] Robert P., Escoufier Y. (1976): A unifying tool for linear multivariate statistical methods: the R_V-coefficient. *Applied Statistics*, **25**, 257-265.

[28] Sneath P.H.A., Sokal R.R. (1973): *Numerical taxonomy. The principles and practice of numerical classification*. Freeman, San Francisco.

[29] Tarsitano A. (2005): Weighted rank correlation and hierarchical clustering. In Zani S., Cerioli A. (eds.) *Classification and data analysis 2005. Book of short papers*, pp. 517-520.

[30] Theil H. (1950): A rank-invariant method of linear and polynomial regression analysis. Part 3. *Proceedings of Koninalijke Nederlandse Akademie van Weinenschatpen A.*, **53**, 1397-1412.

[31] Young G., Householder A.S. (1938):Discussion of a set of points in therm of their mutual distances. *Psychometrika*,**3**, 19-22.

Sitografia

http://devdata.worldbank.org
http://ec.europa.eu/comm/eurostat/ramon/nuts
http://europa.eu.int/comm/eurostat
http://europa.eu.int/comm/public_opinion/index_en.htm
http://www.amstat.org/publications/jse/jse_data_archive.html
http://www.eea.eu.int
http://www.ires.piemonte.it
http://www.istat.it/ambiente/contesto/infoterr/azioneB.html
http://www.istat.it/dati/dataset/20051025_00/
http://www.istat.it/dati/dataset/20060619_00/
http://www.nationmaster.com
http://www.oecd.org