

UNIVERSITA' DEGLI STUDI DELLA CALABRIA

FACOLTA' DI ECONOMIA

DIPARTIMENTO DI ECONOMIA E STATISTICA

Corso di laurea in Scienze Statistiche ed Attuariali

TESI DI LAUREA

**L'INDIVIDUAZIONE DI GRUPPI DI COMUNI DELLA
PROVINCIA DI CATANZARO ADATTI AD OSPITARE
INSEDIAMENTI INDUSTRIALI**

Relatore:

Chiar.ma prof.ssa
Maria Rosaria Ferrante

Studentessa:

Carmela Monteleone
Matr. 41641

Anno Accademico 2000/2001

INDICE

<i>Introduzione</i>	p. 4
CAPITOLO PRIMO: I DISTRETTI INDUSTRIALI	p. 5
I.1: <i>Introduzione</i>	
I.2: <i>Il territorio</i>	
I.3: <i>La popolazione</i>	
I.4: <i>L'economia</i>	
I.5: <i>Le infrastrutture</i>	
I.6: <i>Come nasce un distretto industriale</i>	
I.7: <i>I vantaggi indotti da un distretto industriale</i>	
CAPITOLO SECONDO: ANALISI IN COMPONENTI PRINCIPALI	p. 15
II.1: <i>Introduzione</i>	
II.2: <i>Determinazione delle Componenti Principali</i>	
II.3: <i>Il caso della matrice di correlazione</i>	
II.4: <i>Proprietà delle Componenti Principali</i>	
II.5: <i>I punteggi fattoriali</i>	
II.6: <i>La scelta dei fattori</i>	
II.7: <i>Interpretazione dei fattori</i>	
II.8: <i>La scelta delle variabili</i>	

CAPITOLO TERZO: CLUSTER ANALYSIS.....p. 47III.1: *Introduzione*III.2: *Le fasi della Cluster Analysis*III.3: *Verifica dell'esistenza dei gruppi*III.4: *Le tecniche di clustering*III.5: *Metodi gerarchici di classificazione*III.6: *Metodi non gerarchici di classificazione*III.7: *Altri metodi di classificazione*III.8: *Scelta del metodo di classificazione*III.9: *Interpretazione dei risultati***CAPITOLO QUARTO : L'IDIVIDUAZIONE DI GRUPPI DI COMUNI
POTENZIALMENTE ADATTI AD INSEDIAMENTI INDUSTRIALI**.....p.85IV.1: *Introduzione*IV.2: *La riduzione della dimensionalità dell'informazione attraverso
l'ACP*IV.3: *L'individuazione dei gruppi di comuni***CAPITOLO QUINTO : LA CONNOTAZIONE DEI GRUPPI**.....p. 113V.1: *Introduzione*V.2: *I fattori utilizzati nell'analisi*V.3: *Descrizione dei gruppi*V.4: *Conclusioni***Appendice A: Misure di distanza**.....p. 141**Appendice B: Descrizione delle variabili**.....p. 146**Appendice C: Descrizione delle unità**.....p. 155**Bibliografia**.....p. 158

Introduzione

Il presente lavoro ha l'obiettivo di individuare, sul territorio della provincia di Catanzaro, gruppi di Comuni caratterizzati da particolari peculiarità socio-economiche favorevoli alla possibilità di insediamenti industriali.

Il primo capitolo ha lo scopo di fornire alcune nozioni basilari su quello che si è pensato possa essere il tipo di insediamento industriale più adatto al contesto socio-economico in questione: il distretto industriale.

Nel secondo e terzo capitolo sono descritti gli strumenti statistici impiegati per raggiungere l'obiettivo sopra delineato. In particolare, nel secondo capitolo viene illustrata l'Analisi in Componenti Principali, utilizzata per rendere più agevole il trattamento del problema in esame in quanto consente una significativa riduzione della mole di informazioni, eliminando quelle ridondanti e mettendo in evidenza le variabili più significative. Nel terzo capitolo vengono proposte alcune tecniche di classificazione, strumenti statistici che consentono di ottenere gruppi di unità omogenei al loro interno ed eterogenei tra loro.

Nel quarto capitolo viene sviluppata l'analisi empirica del fenomeno oggetto di studio.

Il quinto capitolo contiene infine una lettura dei risultati ottenuti e l'individuazione sul territorio di questi ultimi.

CAPITOLO PRIMO: I DISTRETTI INDUSTRIALI

I.1 Introduzione

Negli ultimi anni la teoria dello sviluppo economico ha rivolto il proprio interesse alla dimensione locale, in particolar modo a quelle dimensioni locali le cui caratteristiche, partendo da condizioni di arretratezza, contengono il germe per la nascita di nuove possibilità e prospettive economiche. E' questo il caso dei territori meridionali, i quali stanno sempre più facendo notare le proprie potenzialità di sviluppo ed affermazione in campo economico.

Le condizioni economiche del meridione non sono, in generale, incoraggianti, ma un'analisi più dettagliata di questa realtà territoriale ha permesso agli studiosi di rilevare al loro interno alcune zone caratterizzate da una buona capacità di generare sviluppo.

Le caratteristiche peculiari di tali zone hanno fatto sì che si rivolgesse l'attenzione ad una particolare forma di insediamento industriale: il distretto. Per "distretto industriale", tradotto dal termine anglosassone "Cluster", s'intende un "insieme di imprese e di istituzioni, geograficamente prossime ed

economicamente interconnesse”¹; esso è caratterizzato dall’interdipendenza di una serie di piccole aziende anziché di grandi complessi, e chiama in causa soprattutto comuni limitrofi che si sono trovati riuniti in “distretti” specializzati in qualche produzione. Il riconoscimento giuridico dei distretti industriali è effettuato, per l’Italia, dalla legge 5 ottobre 1991, n. 317.

L’essere caratterizzati dalla relazione di piccole aziende ha portato a pensare ai distretti come ad una plausibile forma di sviluppo della realtà

locale; questo perché, in una realtà come quella del mezzogiorno, caratterizzata da zone in cui lo sviluppo è pressoché assente, sembra logico pensare che la prima possibile tipologia di insediamento industriale sia quella delle piccole aziende, più facili da gestire da chi non è abituato al contatto con realtà economiche più complesse.

La nozione di distretto è direttamente collegata al concetto di “concentrazione geografica”²; la concentrazione geografica si struttura su una serie di elementi fondamentali, cioè:

- il territorio;
- la popolazione,
- l’economia;

¹ Porter, 1998.

² Krugman, *Geography and trade*, 1991

- le infrastrutture.

I.2 Il territorio

Krugman (1991) sostiene che due aree uguali si distinguono per una serie di vantaggi cosiddetti “first-nature”, quali le risorse naturali, il clima, la posizione geografica; ne deriva l’importanza da attribuire, soprattutto in zone dove scarseggiano altre caratteristiche economiche, agli aspetti fisici peculiari ed a tutto ciò che possa contraddistinguere ed avvantaggiare una zona rispetto ad altre.

L’importanza che ricopre il territorio risulta evidente se si pensa ai vantaggi oggettivi che esso può offrire. Un imprenditore che dovrà scegliere una zona dove insediarsi economicamente preferirà, ovviamente, un territorio che non presenti particolari vincoli naturali, come ad esempio la presenza di un’elevata percentuale di zone montuose, al fine di assicurarsi una più agevole possibilità di comunicazione con le zone circostanti.

Per poter effettuare questa scelta è però necessario avere a disposizione un’analisi dettagliata del territorio, che offra un’immagine reale, oltre che delle risorse prettamente economiche, della struttura fisica della zona di interesse e che permetta una scelta cosciente in prospettiva delle esigenze economiche specifiche di ogni soggetto. A tale scopo sarà quindi importante avere a disposizione una

banca dati costruita “ad hoc” che permetta di accedere in modo agevole alle notizie di maggiore interesse.

I.3 La popolazione

La popolazione ricopre un’importanza predominante in quest’analisi perché costituisce al tempo stesso il fattore “bacino d’utenza” ed il fattore “manodopera”. E’ risaputo il peso che ha, come attrazione di una qualsiasi attività economica, la presenza in un determinato territorio della domanda di prodotti da immettere sul mercato finale, domanda che agisce da impulso alla produzione, ma che subisce allo stesso tempo il “fascino” della presenza in loco di output economico, innescando così un “moto circolare” tra domanda ed offerta che è un importante stimolo per la nascita di nuove imprese.

Se la popolazione vista come bacino d’utenza è molto importante, non è da meno il suo essere “manodopera”. A riguardo è bene fare alcune precisazioni. Oltre all’ovvia distinzione tra manodopera specializzata e non, la teoria economica pone in evidenza la distinzione tra manodopera sottoposta a vincoli localizzativi, quali gli addetti nel settore agricolo e minerario, e non. La concentrazione geografica delle attività economiche è favorita dall’aumento del peso delle attività non sottoposte a vincoli localizzativi; ciò è dovuto al fatto che sono gli addetti di questi settori a costituire la manodopera con la maggiore mobilità, mobilità che permette di reperire forza lavoro più agevolmente.

La manodopera specializzata costituisce il punto cardine sia per le nuove imprese che si affacciano sul mercato, sia per quelle già esistenti che vogliono migliorare la propria competitività; sono, infatti, le risorse umane qualificate che permettono non solo la nascita di attività più competitive ed all’avanguardia, ma anche un migliore utilizzo e combinazione dei fattori produttivi. Non è però da sottovalutare la presenza di manodopera non qualificata, la cui abbondanza in certe zone permette di accedere alla forza lavoro a costi più competitivi.

I.4 L’economia

Quando si parla di economia se ne può intendere sia l’aspetto oggettivo, relativo alla presenza di unità economiche, sia quello “teorico” relativo al concetto di “economie di scala”.

La presenza sul territorio di una serie di industrie già esistenti favorisce la nascita di nuovi complessi industriali per i quali le prime hanno la doppia utilità di essere fornitori di beni intermedi o richiedenti di beni finali, permettendo l’innescio di quelle interdipendenze economiche che sono alla base della nascita di un distretto industriale; risulta ovvio che, il concentramento economico si preferisce in quelle zone in cui è maggiore la presenza di imprese già esistenti. Una certa importanza ricopre inoltre la presenza di cosiddetti “poli di sviluppo”,

imprese di particolare dimensione che possono assumere il ruolo di “unità trainanti” dell’intera economia circostante.

Per quel che riguarda le economie di scala, ricordando che quando il livello di attività del sistema si espande sono necessarie minori quantità di fattori per unità di prodotto, ci si limita a sottolineare che, tanto più importanti sono le economie di scala, tanto più forti saranno le tendenze all’agglomerazione; ciò è dovuto al fatto che, sotto tali condizioni, converrà concentrare la produzione in un’unica unità economica dalla quale servire il mercato.

I.5 Le infrastrutture

La necessità di concentrare la produzione in un’unica area dalla quale servire il mercato conduce direttamente al fattore “infrastrutture”, intese sia come strutture commerciali, finanziarie e di servizi, sia come reti di comunicazione. In particolare, sono proprio le reti di comunicazione a giocare un ruolo determinante nella localizzazione economica, essendo esse direttamente collegate ai costi di trasporto, che sono i costi sui quali si basa la teoria di ottimizzazione di un’impresa. Risulta evidente che si sceglierà di localizzare un complesso

economico in una zona se, a parità di altre condizioni, risulterà più agevole economicamente servire un ampio mercato dalla zona stessa.

Un certo rilievo ha la presenza nella zona di istituzioni finanziarie. La centralità assunta dalla funzione finanziaria nei processi decisionali delle imprese fa sì che vi sia interesse per un’analisi delle offerte del sistema finanziario, ed in particolare delle banche. Per le imprese minori si riscontra una certa fragilità sia sotto il profilo finanziario che economico, fragilità che porta a ricorrere a finanziamenti esterni, soprattutto bancari. La banca è, sicuramente, il principale punto di riferimento per il soddisfacimento delle esigenze finanziarie delle imprese.

Da sottolineare è l’importanza delle banche locali, che data la loro capillarità nella presenza sul territorio, si distinguono non solo per il patrimonio informativo sulla clientela non facilmente appropriabile dai concorrenti, ma anche per la capacità di costruire relazioni di clientela di lunga durata ed a carattere esclusivo, che permette loro di offrire combinazioni di servizi disegnate in base alle esigenze proprie dei clienti.

I.6 Come nasce un distretto industriale

Un distretto industriale nasce:

- combinando fattori produttivi presenti nella zona o acquisibili dall'esterno;
- avendo la possibilità di sfruttare un'adeguata tecnologia;
- per l'azione di una o più imprese motrici;
- se è in grado di raggiungere una soglia critica di domanda e quindi di produzione;
- se nella zona vi è una situazione socio-economica che non impedisce lo sviluppo e vi sono istituzioni che possono sostenerlo;
- se diventa competitivo.

In sostanza, la nascita di nuove imprese è il frutto dell'interazione tra caratteristiche sociali e culturali dell'ambiente in cui esse si sviluppano e della peculiarità del bagaglio delle conoscenze tecnologiche, organizzative e di mercato dell'imprenditore nascente. La localizzazione di nuove imprese può essere vista come il risultato di un processo competitivo in cui le diverse imprese mettono in gioco la propria capacità di controllo dei mercati. Fattore fondamentale nel processo competitivo tra le imprese è il superamento di particolari difficoltà oggettive, dette "barriere d'entrata"; in un ambiente competitivo la capacità di superare questi ostacoli iniziali (relativi soprattutto alle prime fasi di avvio di un'azienda) rappresenta il primo passo che ogni entità economica deve compiere per entrare nel mercato. La capacità di cui si parla dipende in via principale dall'adozione di innovazioni, dall'evoluzione della

tecnologia nei processi produttivi, dall'evoluzione della domanda dei prodotti e dei servizi e dall'evoluzione dell'offerta dei fattori produttivi.

Affinché nasca un distretto industriale è necessario che una determinata zona del territorio risponda, se non a tutte, almeno alla maggior parte delle caratteristiche definite precedentemente (condizioni territoriali favorevoli, presenza di manodopera e di un bacino d'utenza, presenza di strutture economiche già esistenti e di infrastrutture adeguate). In genere tali caratteristiche possono essere individuate in zone abbastanza ampie, come ad esempio gruppi di comuni, ma ciò non toglie che se un singolo comune presenta tali condizioni non possa essere preso in considerazione come il punto di riferimento di un insediamento industriale.

I.7 I vantaggi indotti da un distretto industriale

Un distretto industriale si distingue dalle altre forme di insediamento economico per la capacità di far nascere al proprio interno una forza competitiva maggiore rispetto a quella di imprese identiche collocate al di fuori. Tale forza competitiva dipende dalle relazioni che si instaurano tra le imprese facenti parte del distretto, relazioni che permettono un'ottimizzazione delle strutture produttive e decisionali, grazie alla facilità con la quale in tale struttura viaggiano

le informazioni di tipo tecnico-economico. Il distretto ha inoltre il grande vantaggio di essere formato, ad eccezione della presenza di “poli” di elevata dimensione, da una serie di piccole e medie imprese, la cui forza risiede nella propria capacità di adattamento alla domanda e, quindi, di adeguamento della produzione.

La nascita di un distretto è una possibile strada per lo sviluppo di quelle zone che sono considerate “periferiche” dal punto di vista economico; tale sviluppo è agevolato dalla nascita di una nuova attività industriale in quanto genera nuova domanda sia per ciò che riguarda la manodopera sia per la richiesta di prodotti complementari da altre imprese, situazione che getta le basi per la nascita di nuove attività, e che favorisce l’utilizzo di quelle già esistenti.

Si possono così sviluppare imprese collegate, per la fornitura di beni e servizi ma anche per l’utilizzo di nuovi prodotti. Tale situazione non è altro che il “collante” sul quale si basa la nascita e la sopravvivenza di un distretto e rappresenta al tempo stesso il punto di forza di quest’ultimo, che fonda la propria competitività sulla possibilità di appoggiarsi alle imprese circostanti.

CAPITOLO SECONDO: ANALISI IN COMPONENTI PRINCIPALI

II.1 Introduzione

L’Analisi in Componenti Principali (ACP) è una tecnica statistica multivariata, delineata nei suoi principi teorici da K. Pearson nel 1901 e sviluppata successivamente da Hotelling nel 1933, che fa parte dei metodi di analisi fattoriale e può essere impiegata con riferimento a variabili quantitative.

L’obiettivo principale dell’ACP è di ridurre le dimensioni del fenomeno studiato conservando la struttura informativa dei dati iniziali, cioè il grado di variabilità che essi esprimono, in modo tale da poter basare l’analisi su un numero inferiore di variabili (le Componenti Principali) rispetto al numero di variabili di partenza. In tal modo risulta più agevole e facile interpretare il fenomeno oggetto di studio.

Più in dettaglio, le Componenti Principali sono combinazioni lineari delle variabili originarie, incorrelate tra di loro ed elencate in ordine decrescente della loro varianza; l’incorrelazione delle CP è un requisito necessario se si vuole costruire un nuovo sistema di riferimento che non abbia ridondanze e che incorpori soltanto l’informazione utile. Poiché sono ordinate in senso decrescente di varianza, è possibile, facendo riferimento solo ai primi fattori, basare l’analisi

su un numero ridotto di variabili fittizie, le Componenti Principali appunto, senza perdere informazioni essenziali riguardanti il fenomeno.

Sostituendo i dati rilevati (eventualmente standardizzati) nelle combinazioni lineari che rappresentano le CP, si ottengono i cosiddetti “punteggi fattoriali” (scores); tali punteggi hanno il compito di assegnare alle unità una nuova scala di valori, calcolata in base alle nuove variabili di riferimento (le CP).

Da un punto di vista geometrico, le p caratteristiche rilevate su ogni unità sono rappresentate da vettori p -dimensionali e_i , di conseguenza, le n unità definiscono una nuvola di punti; la multidimensionalità di tale nuvola ne rende particolarmente difficile l'interpretazione. L'obiettivo dell'ACP è quello di “adattare” tale nuvola ad un sottospazio di dimensioni più ridotte, quindi più facilmente interpretabile (si pensi ad una retta di regressione che si “adatta” ad una nuvola di punti in \mathbb{R}^2). Come per lo spazio bidimensionale, anche in questo caso si procede alla ricerca di un sottospazio, la cui ottimalità è rappresentata dalla capacità di minimizzare le distanze dei punti dal sottospazio stesso; questo risultato può essere raggiunto massimizzando “l'inerzia” della nuvola di punti, che è pari alla somma dei quadrati delle M -norme³ dei vettori unità:

$$I = \sum_i p_i \|x_i\|_M^2$$

³ Cioè le norme calcolate in base ad una metrica M , dove per norma si intende la lunghezza del vettore, o il suo modulo.

Si supponga di scomporre detta inerzia in due sottospazi, R_1 ed R_2 , ortogonali e componibili in somma diretta⁴; in virtù di tale scomposizione, ogni i -esimo punto, rappresentato dal vettore x_i , avrà due immagini, una in R_1 ed una in R_2 , le cui norme saranno rispettivamente $?(e)$ e $?(e)$. La tecnica si prefigge lo scopo di cercare il sottospazio ottimale, per il quale si abbia la migliore proiezione delle distanze dei punti; ciò si traduce nel ricercare quelle direzioni strutturali (assi fattoriali), fra loro a due a due incorrelate, tale che sia minima in media la distanza dei punti della nuvola da essi.

Dopo aver trovato gli assi ottimali, si ottengono su di essi le coordinate delle unità (gli “scores” precedentemente citati), che rappresentano il valore delle unità secondo la nuova variabile; tali coordinate sono date dal prodotto scalare tra il vettore originario (x_i) ed il versore⁵ del nuovo asse.

Calcolare le coordinate dei punti rispetto agli assi del nuovo sottospazio significa trasportare i dati in un nuovo sistema di riferimento che rappresenta quello originario, ma che ha il pregio di fornire una descrizione più semplice dei dati ed allo stesso tempo di contenere solo le informazioni necessarie⁶.

⁴ La somma diretta di due sottospazi è l'insieme somma di sottospazi disgiunti, cioè tali da avere intersezione nulla. (Rizzi, 1988)

⁵ Un versore è un vettore di norma unitaria che indica il verso del vettore cui si riferisce.

⁶ La descrizione approfondita del calcolo degli assi e delle coordinate è rimandata ai paragrafi successivi.

II.2 Determinazione delle Componenti Principali

Prima di passare alla parte relativa al calcolo vero e proprio delle CP, è bene dare alcune indicazioni geometriche sulle quali si basa la logica di tale calcolo.

L'obiettivo di una riduzione fattoriale è quello di rappresentare la nuvola degli n punti-unità, appartenente allo spazio a p dimensioni. Poiché i dati si considerano centrati rispetto alle relative medie, l'origine dello spazio che deve rappresentare gli individui è posta nel baricentro della nuvola di punti. La distanza di ogni punto da questa origine è data dalla lunghezza, o norma, del vettore unità e misura quindi uno scostamento dalla media.

Il sottospazio vettoriale individuato dalle CP dovrà essere tale da ottimizzare tali distanze.

Partendo dalla matrice dei dati originari, X_{np} , composta da n unità e p variabili, illustriamo nel seguito la tecnica per determinare le Componenti Principali.

La prima CP, Y_1 , è, per definizione, la combinazione lineare delle p variabili iniziali, caratterizzata dal fatto di avere massima varianza; la seconda CP, Y_2 , è la combinazione lineare delle p variabili iniziali avente varianza immediatamente inferiore, con la caratteristica di essere incorrelata (ortogonale) alla componente precedente; e così di seguito per le altre componenti.

Considerando la matrice X degli scostamenti delle variabili dalle loro medie, la prima CP è data da:

$$Y_1 = Xa_1$$

dove a_1 è un vettore colonna p -dimensionale di componenti (a_{11}, \dots, a_{p1}) .

La sua determinazione è subordinata all'impostazione di un problema di massimo vincolato. Si è già detto che Y_1 è una combinazione lineare di X con massima varianza; affinché ciò sia verificato, il vettore a_1 deve essere tale da massimizzare la varianza totale della combinazione lineare Xa_1 , esprimibile in funzione della matrice di varianze-covarianze S :

$$\text{var}(Xa_1) = a_1' S a_1$$

Introducendo il vincolo di normalizzazione per il vettore a_1 :

$$a_1' a_1 = 1$$

il quale implica che la somma dei quadrati delle sue componenti sia pari a 1,

il problema di massimo vincolato assume la forma:

$$\max_{a_1} a_1' S a_1$$

$$\text{s. t. } a_1' a_1 = 1$$

Per risolvere tale problema è necessario impostare la Lagrangiana:

$$a_1' S a_1 + \lambda_1(1 - a_1' a_1)$$

dove λ_1 è il moltiplicatore di Lagrange.

Derivando rispetto ad a_1 l'espressione precedente ed eguagliando a zero si ottiene:

$$(1) \quad (S - \lambda_1 I) a_1 = 0$$

Il sistema (1) è lineare, con p equazioni in p incognite (dove I è la matrice identità), che ammette soluzioni non tutte nulle se il suo determinante è pari a zero:

$$(2) \quad |S - \lambda_1 I| = 0$$

Quest'ultima uguaglianza rappresenta l'equazione caratteristica della matrice S , ed è un polinomio di ordine p , con p soluzioni chiamate "autovalori" (eigenvalues), o "radici caratteristiche". Poiché la matrice S è semidefinita positiva⁷, gli autovalori saranno tutti non negativi.

Per determinare l'autovalore da scegliere, si premoltiplica la (1) per a_1' :

$$a_1' S a_1 - \lambda_1 a_1' a_1 = 0$$

dalla quale deriva che λ_1 coincide con la varianza della combinazione lineare Xa_1 , cioè della prima CP:

$$a_1' S a_1 = \text{var}(Xa_1) = \lambda_1$$

Poiché l'obiettivo è quello di massimizzare tale varianza, si sceglierà come λ_1 il più grande degli autovalori ottenuti dalla (2).

⁷ Una matrice è semidefinita positiva quando ha valori maggiori o uguali a zero.

Dopo aver ottenuto λ_1 , si sostituisce tale valore nella (1), dalla quale, tenendo presente il vincolo di normalizzazione, si ricava il vettore a_1 , che risulta essere l'autovettore corrispondente all'autovalore λ_1 .

Si può quindi enunciare la seguente definizione:

“Si dice prima componente principale di p variabili, espresse in termini di scostamenti dalla loro media, la combinazione lineare:

$$Y_1 = Xa_1$$

in cui a_1 è l'autovettore corrispondente all'autovalore più grande, λ_1 , della matrice di varianze-covarianze S .” (Zani, 1999)

La seconda Componente Principale, Y_2 , è data da:

$$Y_2 = Xa_2$$

che deve essere incorrelata alla precedente, deve cioè soddisfare la condizione:

$$a_1' a_2 = 0$$

Il problema di massimo vincolato, che si risolve in modo analogo al precedente (tenendo conto dell'ulteriore vincolo), risulta:

$$\max_{a_2} a_2' S a_2$$

$$\text{s. t. } a_2' a_2 = 1$$

$$a_2' a_1 = 0$$

Risolvendo il problema si troverà l'autovalore λ_2 e, di conseguenza, l'autovettore a_2 , che permette di definire la seconda CP.

In termini generali, si può introdurre la definizione:

“Si dice v -esima componente principale di p variabili espresse in termini di scostamenti dalla media, la combinazione lineare:

$$Y_v = \sum_{i=1}^p X_i a_{iv} \quad \text{per } v=1, \dots, k \leq p$$

in cui a_{iv} è l'autovettore associato al v -esimo autovalore λ_v , in ordine decrescente, della matrice di varianze-covarianze S .” (Zani, 1999)

Le CP si ottengono dunque dalla risoluzione di un problema di massimo vincolato, che pone in luce l'importanza informativa della varianza; chiedere alle CP di replicare la varianza iniziale significa renderle in grado di rappresentare il fenomeno senza dover perdere informazioni importanti.

Dopo aver delineato le linee teoriche per il calcolo delle componenti principali, è importante enunciare alcune proprietà:

(1) Ogni autovalore è uguale alla varianza della corrispondente CP:

$$\lambda_v = \text{var}(Y_v)$$

La somma di tutti gli autovalori coincide quindi con la varianza totale (che è data dalla traccia della matrice S):

$$\sum_{v=1}^p \lambda_v = \text{tr}(S)$$

pertanto, la quota di varianza totale spiegata dalla v -esima CP è data dal rapporto:

$$\lambda_v / \sum_{v=1}^p \lambda_v$$

nota anche come “tasso di inerzia”.

(2) Il coefficiente di correlazione tra la v -esima CP e la s -esima variabile è dato dal rapporto:

$$r(Y_v, X_s) = (a_{sv} \lambda_v) / (\sqrt{\text{var}(X_s)})$$

dove a_{sv} è l'elemento del vettore a_v corrispondente alla s -esima variabile.

Se si considera un numero di CP pari al numero delle variabili iniziali, si riproduce esattamente la variabilità iniziale; come è ovvio intuire, un tale procedimento non porterebbe alcun miglioramento in termini di dimensioni dello spazio dei dati, e nonostante l'applicazione della tecnica produca un interessante elenco di variabili distinte per capacità informativa (le ultime CP corrispondono a variabili con scarsa percentuale di variabilità riprodotta), non si eliminerebbe il problema relativo alla difficoltà interpretativa.

La proprietà (1) spiega perché le CP abbiano importanza decrescente; esse sono per costruzione associate alla successione decrescente degli autovalori della matrice S , e poiché questi coincidono con la varianza, ne deriva che le CP replicheranno tale tipo di ordinamento. Questo aspetto può essere osservato in pratica calcolando i rapporti tra i singoli autovalori e la variabilità totale, i quali risulteranno sempre più piccoli, indicando che le prime CP spiegano quote maggiori di varianza rispetto alle successive.

Il potere esplicativo dei fattori è rappresentato dalla quantità di varianza totale spiegata da ognuno di essi; detta quantità indica in che misura le norme dei

vettori sono state, in media, ricostruite su ciascun asse fattoriale, indicando la capacità rappresentativa del fattore stesso; si ottiene quindi una graduatoria multidimensionale del potere riproduttivo dei fattori, che cambia in base al numero di componenti che si decide di estrarre.

Il coefficiente di correlazione tra la v -esima CP e la s -esima variabile, definito dalla proprietà (2), ne indica il legame; esso è un buon indicatore del “peso” assunto dalla variabile nella costruzione della CP e viene di solito utilizzato per individuare le variabili più significative.

II.3 Il caso della matrice di correlazione

Nel paragrafo precedente sono state calcolate le CP partendo dalla matrice di varianze-covarianze S e sono state ottenute come combinazioni lineari degli scostamenti delle variabili originarie dalle loro medie; come conseguenza di questo “modus operandi” si ha che il confronto tra le variabili è lecito soltanto nel caso in cui queste ultime siano espresse nella stessa unità di misura, poiché il cambiamento di scala di una sola variabile ha come conseguenza la modifica della varianza totale e quindi influenza in modo significativo i risultati dell'ACP. Si può quindi affermare che si ha una corretta applicazione dell'ACP riferita alla matrice S soltanto nel caso di variabili espresse tutte nella stessa unità di misura.

Nel caso in cui le variabili originarie siano espresse in unità di misura differenti, ed è questa la situazione più frequente nelle analisi reali, si ha la necessità di renderle confrontabili, e ciò può essere fatto considerando le variabili in termini di scostamenti standardizzati, che equivale ad assumere come punto di partenza dell'ACP la matrice di correlazione R .

Va sottolineato che la scelta fatta a priori di utilizzare S oppure R per sviluppare la tecnica ACP condiziona i risultati dell'analisi con riferimento ad alcune proprietà delle CP.

Si parte quindi dalla matrice Z degli scostamenti standardizzati delle variabili di origine. La procedura di calcolo delle CP ricalca le linee già descritte precedentemente, tenendo presente che la varianza totale di p variabili standardizzate (quindi con media nulla e varianza unitaria) è pari a p :

$$\text{var}(Z) = \text{tr}(R) = p$$

La prima CP è data dalla combinazione lineare:

$$Y_1 = Za_1$$

il problema di massimo vincolato ad essa associato assume la forma:

$$\begin{aligned} \max_{a_1} & a_1' R a_1 \\ \text{s. t.} & a_1' a_1 = 1 \end{aligned}$$

La seconda CP è data dalla combinazione lineare:

$$Y_2 = Za_2$$

il cui problema di massimo vincolato risulta essere:

$$\begin{aligned} \max_{a_2} \quad & a_2' R a_2 \\ \text{s. t.} \quad & a_2' a_2 = 1 \\ & a_2' a_1 = 0 \end{aligned}$$

I due problemi di ottimizzazione si risolvono in maniera analoga a quelli dell'ACP basata sulla matrice S.

Nei termini generali, si può quindi definire:

“Si dice v-esima componente principale di p variabili standardizzate la combinazione lineare:

$$Y_v = Z a_v \quad \text{per } v=1, \dots, k \leq p$$

in cui a_v è l'autovettore associato al v-esimo autovalore λ_v , in ordine decrescente, della matrice di correlazione R.” (Zani, 1999)

Le proprietà delle CP calcolate con questo metodo sono simili a quelle delle CP calcolate partendo dalla matrice S:

1) Ogni autovalore è uguale alla varianza della corrispondente CP:

$$\lambda_v = \text{var}(Y_v)$$

la somma degli autovalori è uguale a p:

$$\sum_{v=1, p} \lambda_v = p$$

pertanto la quota di varianza totale spiegata dalla v-esima CP è data dal rapporto:

$$\lambda_v / p$$

2) Il coefficiente di correlazione tra la v-esima CP e la s-esima variabile è dato da:

$$r(Y_v, X_s) = a_{sv} \lambda_v^{-1/2}$$

3) La quota di varianza della s-esima variabile spiegata dalla v-esima CP è uguale al quadrato del corrispondente coefficiente di correlazione:

$$r^2(Y_v, X_s)$$

Poiché le CP sono tra di loro ortogonali, e quindi additive, la quota di varianza della s-esima variabile spiegata dalle prime k CP è data da:

$$\sum_{v=1, k} S_{vs}^2 \quad \text{per } s=1, \dots, p$$

inoltre, la somma dei quadrati dei coefficienti di correlazione riferiti alla v-esima CP è uguale al corrispondente autovalore:

$$\sum_{s=1, p} r_{vs}^2 = \lambda_v$$

Le CP calcolate partendo dalla matrice di correlazione hanno un ruolo importante soprattutto in riferimento alle analisi reali, dove si ha quasi sempre a che fare con matrici di dati in cui le variabili di rilevazione sono espresse in unità di misura differenti.

Osservando le proprietà di questo tipo di CP si nota innanzitutto che esse sono più semplici da interpretare rispetto alle precedenti; questa facilità di interpretazione è attribuibile soprattutto al fatto che la variabilità complessiva della matrice dei dati è pari a p , e ciò porta ad intuire meglio come ogni autovalore rappresenti una “quota” di variabilità totale, cioè come ognuno di essi sia una parte che “compono” la complessità del fenomeno studiato.

Un’importante indicazione viene inoltre fornita dal quadrato del coefficiente di correlazione tra la v -esima CP e la s -esima variabile; nella pratica, infatti, dopo aver scelto il numero dei fattori sul quale basare l’analisi, è bene osservare in quale misura questi riescano a replicare la variabilità di ogni singola variabile, permettendo all’analista di valutare la bontà della riduzione.

II.4 Proprietà delle Componenti Principali

Le proprietà elencate nel paragrafo precedente risultano funzionali per l’interpretazione dei risultati della tecnica ACP. Per comprendere l’ottimalità della trasformazione in termini di CP della matrice dei dati è opportuno enunciare una serie di ulteriori proprietà. Nell’elencare queste proprietà si fa riferimento all’ACP effettuata partendo dalla matrice degli scostamenti dalla media, X .

I PROPRIETA’

Partendo dalla matrice di varianze-covarianze S , si consideri la matrice $A_{n,k}$ dei primi k autovettori di S . Fissato il numero k di tali autovettori, si consideri la

trasformazione lineare:

$$Y_{n,k} = X_{n,p} A_{n,k}$$

Questa trasformazione lineare gode delle seguenti proprietà:

a) sia B una matrice ortonormale⁸; la matrice di varianze-covarianze riferita a Y può essere espressa da:

$$S_Y = B' S B$$

La trasformazione lineare sopra indicata è tale da massimizzare la varianza totale di Y :

$$\text{tr}(S_Y) = \max;$$

b) La varianza generalizzata di Wilks è una misura di variabilità multidimensionale che tiene conto della correlazione tra le variabili ed è data dal determinante della matrice di varianze-covarianze; la trasformazione lineare considerata massimizza la varianza generalizzata di Wilks di Y :

$$|S_Y| = \max;$$

c) $Y_{n,k}$ minimizza la somma dei quadrati dei residui delle variabili; ciò significa che la trasformazione lineare:

⁸ Ciò è composta da colonne ortogonali e normalizzate.

$$Y_{n,k} = X_{n,p} A_{n,k}$$

è il migliore predittore lineare di ordine k della matrice X .

Questa prima proprietà porta ad osservare che la trasformazione indotta dalla tecnica ACP è in grado di conservare un buon livello di informazione iniziale.

II PROPRIETA'

Si considerino le prime k CP; il sottospazio R^k individuato da tali componenti è quello che minimizza nel modo migliore la somma dei quadrati delle distanze perpendicolari dei vettori x_1, \dots, x_n dal sottospazio stesso; in altri termini, le prime k CP individuano il sottospazio migliore per la proiezione della matrice X .

Ricordando l'interpretazione geometrica della tecnica ACP, si ha che tale proprietà risponde ai requisiti richiesti al sottospazio ottimale scelto per adattare la nuvola di punti.

Per l'importanza assunta da questa proprietà, è sembrato opportuno dimostrarla, pur riferendosi al caso semplificato di due sole variabili. Il sottospazio ottimale è in questo caso composto da due CP; si vuole dimostrare che la somma delle distanze al quadrato dei punti dalla prima CP è minima. Le distanze in questione, riferendosi al generico punto i , sono date dalle coordinate dei punti sul secondo asse, y_{i2} . Si vuole dimostrare che:

$$S_{i=1,n} y_{i2}^2 = \min$$

Dalla definizione di seconda componente principale si ottiene la seguente serie di uguaglianze:

$$\begin{aligned} S_{i=1,n} y_{i2}^2 &= S_{i=1,n} (a_2' x_i)^2 = S_{i=1,n} (a_2' x_i x_i' a_2) = \\ &= a_2' (S_{i=1,n} x_i x_i') a_2 = n a_2' S a_2 = n \lambda_2. \end{aligned}$$

Poiché λ_2 è l'autovalore al quale corrisponde la minima varianza (l'ultimo su due), ne segue che l'ultimo termine ha valore minimo:

$$n \lambda_2 = \min S_{i=1,n} y_{i2}^2 = \min.$$

III PROPRIETA'

Si consideri la seguente funzione della matrice X degli scostamenti dei dati dalla media e della matrice S di varianze-covarianze:

$$X' S X = \text{costante}$$

Tale funzione individua una famiglia di ellissoidi p -dimensionali.

Le CP definiscono gli assi principali di tale famiglia di ellissoidi.

L'importanza di questa proprietà si coglie facilmente se si suppone che le n osservazioni siano le realizzazioni di una variabile aleatoria con distribuzione normale multivariata; in questo caso gli ellissoidi p -dimensionali rappresentano i contorni di probabilità costante della variabile aleatoria multipla, e le CP, in quanto assi di questi ellissoidi, giocano il ruolo di "punti di riferimento" per l'interpretazione probabilistica della variabile.

Queste ultime proprietà elencate mettono in risalto l'importanza rivestita dalle CP in quanto "sistema di riferimento privilegiato" attraverso il quale interpretare il fenomeno oggetto di studio, che altrimenti, per l'elevata dimensionalità dello stesso, risulterebbe poco comprensibile; permettono inoltre di affermare che il sottospazio individuato dalle CP è quello che meglio "adatta" la matrice dei dati iniziale.

II.5 I punteggi fattoriali

Il punteggio fattoriale (factor score) di un'unità statistica è il valore (la coordinata) che essa assume su uno specifico asse. Come è noto, la v-esima CP è definita dalla combinazione lineare:

$$Y_v = Xa_v$$

nel caso di variabili intese come scostamenti dalla media, e dalla combinazione lineare:

$$Y_v = Za_v$$

nel caso di variabili standardizzate.

Il punteggio per l'i-esima unità sul generico fattore v è dato da:

$$(1) y_{iv} = a_{v1} x_{i1} + \dots + a_{vp} x_{ip} \quad \text{per } i=1, \dots, n \quad \text{e} \quad v=1, \dots, k$$

nel primo caso, e :

$$(2) y_{iv} = a_{v1} z_{i1} + \dots + a_{vp} z_{ip} \quad \text{per } i=1, \dots, n \quad \text{e} \quad v=1, \dots, k$$

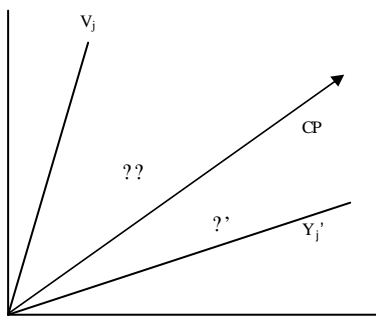
nel secondo.

I punteggi fattoriali non sono altro che il valore assunto dalla nuova variabile (CP) con riferimento all'unità i-esima e risultano importanti in quanto il sistema di misurazione adottato è, per costruzione, ottimale. Per interpretare le caratteristiche delle CP impieghiamo i coefficienti delle relazioni precedenti. Infatti, il segno dei coefficienti a_{vs} indica il tipo di relazione tra la v-esima CP e la s-esima variabile, che può essere di tipo diretto o inverso; il valore numerico del citato coefficiente indica la misura in cui la variabile s-esima concorre alla determinazione dei punteggi della componente v-esima. Queste semplici indicazioni permettono di analizzare più in dettaglio le relazioni esistenti tra le CP e le variabili.

Proviamo ora a dare una spiegazione geometrica dell'importanza assunta dai coefficienti delle combinazioni lineari che rappresentano le CP.

Nello spazio delle unità una CP è costituita dall'insieme delle coordinate degli individui sull'asse fattoriale corrispondente; tali coordinate non sono altro che i punteggi fattoriali espressi dalla (1) o dalla (2). Analizzando nel dettaglio le espressioni (1) e (2) si può notare facilmente che il punteggio rispetto ad un fattore è dato dalla somma dei diversi contributi delle variabili (dove per contributo si intende la quota di variabile che entra a far parte della combinazione lineare); ovviamente, alcune variabili daranno un contributo maggiore di altre.

L'importanza assunta dalle variabili in questo ambito può essere illustrata facilmente da un punto di vista geometrico. Per la dualità esistente tra unità e variabili, le CP dello spazio delle unità corrispondono alle dimensioni fattoriali costruite nello spazio delle variabili; la direzione della generica CP delle unità forma con i vettori delle variabili un angolo tanto più piccolo quanto maggiore è il contributo della variabile nella combinazione lineare che descrive le CP:



Poiché in una combinazione lineare il coefficiente associato ad una variabile ne rappresenta l'angolo, l'importanza di una variabile sarà proporzionale alla grandezza dei coefficienti citati.

Per definizione, le CP, calcolate sia attraverso il metodo basato sulla matrice S sia attraverso quello basato sulla matrice R, sono elencate in ordine decrescente di varianza, e per questo motivo i punteggi riferiti ad una CP risultano avere

maggiore variabilità rispetto ai punteggi riferiti alle CP successive; in conseguenza di ciò, i valori numerici degli "scores" di ciascuna CP ne riflettono l'importanza. Lo svantaggio principale dei punteggi così ottenuti è quello di non essere direttamente comparabili per componenti differenti, per le quali si fa riferimento a diversi autovalori; ciò è dovuto al fatto che, come è noto, ogni autovalore corrisponde alla varianza della rispettiva CP:

$$\text{var}(Y_v) = \lambda_v$$

Per poter rendere confrontabili i punteggi sarà necessario rendere unitaria la loro varianza (dato che la loro media è già pari a zero), e questo risultato può essere raggiunto agevolmente dividendo gli scores per la radice quadrata del rispettivo autovalore. Adottando questo criterio sarà possibile confrontare le unità per mezzo dei loro scores che, come già ricordato, hanno un particolare valore interpretativo.

Notevole importanza ai fini dell'analisi ricopre la rappresentazione grafica dei punteggi fattoriali, che possono essere trattati come nuove variabili da inserire su piani cartesiani individuati da coppie di fattori. I grafici dei punteggi fattoriali così costruiti possono evidenziare particolari andamenti dei punti-unità, quali ad esempio consistenti non linearità, la presenza di dati anomali e, soprattutto, l'esistenza di gruppi di unità nello spazio fattoriale, peculiarità che può essere utilizzata in analisi successive, come ad esempio una "cluster analysis".

Queste analisi grafiche vengono generalmente condotte su coppie di fattori e non attraverso grafici tridimensionali, in modo da ottenere rappresentazioni più facilmente interpretabili, limitandosi spesso a considerare le combinazioni delle prime tre o quattro CP, per i motivi che abbiamo già ampiamente spiegato.

II.6 La scelta dei fattori

Un passo importante della conduzione della tecnica ACP è quello riferito al numero dei fattori che si deve far entrare nell'analisi; teoricamente, questo numero è noto a priori, ma in pratica la sua scelta si basa su diversi criteri.

La scelta del numero dei fattori segue due linee fondamentali, che devono essere soddisfatte contemporaneamente: il numero dei fattori introdotto nell'analisi non deve essere troppo elevato, altrimenti non avrebbe senso una riduzione della dimensione dello spazio delle variabili, ma allo stesso tempo tale numero deve essere "significativo", nel senso che deve essere in grado di rappresentare una buona quota della struttura informativa iniziale.

Si illustrano di seguito i principali criteri per determinare il numero dei fattori.

Nella prassi comune, il numero dei fattori è fissato a priori soltanto in alcuni casi specifici, come ad esempio negli studi di simulazione o negli studi psicologici, ed in poche altre situazioni; in tali casi l'analista imposta il

calcolatore in modo che si fermi una volta estratto il numero di fattori prefissato. La regola generale che deve essere seguita anche nel caso di fattori prefissati è quella riguardante il numero di variabili oggetto del fenomeno studiato: il numero di fattori estratti deve dipendere dal numero delle variabili, nel senso che, per ottenere un buon modello di analisi, tale numero deve aumentare al crescere del numero delle variabili in questione; Harris (1967) sostiene che il rapporto tra il numero di variabili ed il numero di fattori estratti non debba essere inferiore a 2, e più è elevato tale rapporto tanto più l'analisi condotta sarà attendibile.

Nella maggior parte degli studi reali il numero di fattori non viene fissato a priori ma scelto sulla base dei criteri che illustreremo. I criteri principali per la scelta del numero dei fattori sono tre e si basano su:

- 1) la varianza totale spiegata dai fattori;
- 2) il valore numerico degli autovalori;
- 3) la rappresentazione grafica degli autovalori.

Il primo criterio consiste nell'estrarre un numero di fattori tale da spiegare una quota accettabile di varianza totale.

Si ricorda innanzitutto che la quota di varianza totale spiegata dai primi k fattori è data da:

$$S_{v=1,k} / S_{v=1,p}$$

nel caso di analisi condotta sulla matrice di varianze-covarianze, e da:

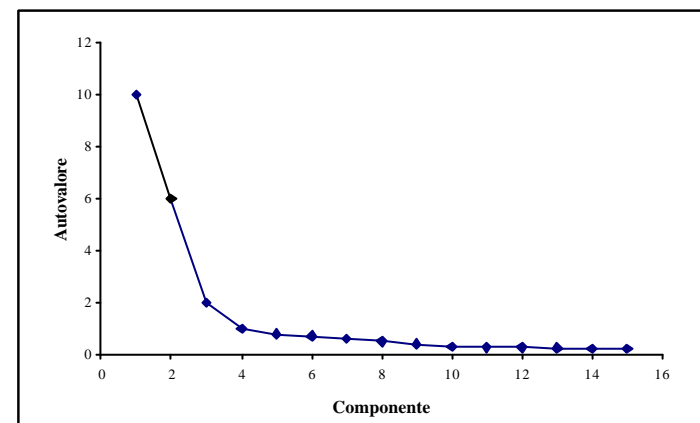
$$S_{v=1,k} = \lambda_v / p$$

nel caso di analisi condotta sulla matrice di correlazione.

In generale, è lecito considerare accettabile un numero di fattori che spiega una quota di varianza totale pari all'80%, ma si accettano valori inferiori a tale percentuale nel caso in cui il numero di variabili osservate sia elevato; la frazione di varianza totale che deve essere spiegata dai fattori è quindi legata al numero di variabili inserite nell'analisi.

Il secondo criterio è particolarmente adatto quando si considera l'ACP condotta sulla matrice di correlazione; Guttman (1954) e Kaiser (1960) sostengono che una semplice regola per determinare il numero dei fattori in questo caso sia quella di considerare quei fattori cui corrispondono autovalori maggiori o uguali ad 1. Hair et al. (1987) giustifica tale criterio se applicato all'ACP, mentre suggerisce un abbassamento della soglia nel caso di altre analisi fattoriali, poiché individua in uno il valor medio degli autovalori di un'analisi in componenti principali. In genere, il numero dei fattori con autovalore superiore ad uno è compreso tra il 16% ed il 33% del totale delle variabili esaminate; questa affermazione implica che adottando tale criterio si otterrà una buona riduzione dei dati.

Il terzo criterio basa la scelta del numero dei fattori sulla rappresentazione grafica degli autovalori (scree test). Su un sistema di assi cartesiani ortogonali vengono riportati sull'asse delle ordinate gli autovalori e su quello delle ascisse il loro ordine di estrazione (si ricorda che gli autovalori sono estratti in ordine decrescente); i punti individuati su tale piano sono poi uniti da segmenti, in modo da formare una spezzata:



Analizzando la spezzata, è possibile individuare la presenza di andamenti particolari; se essa mostra due tendenze, la prima riferita ad una forte pendenza al livello dei primi autovalori e la seconda relativa ad un appiattimento che riguarda i fattori successivi (come nel caso in figura), il criterio adottato porta ad ignorare

questi ultimi considerando significativi solo i primi e, di conseguenza, a scegliere i fattori ai quali tali autovalori si riferiscono; nel caso contrario non sarà possibile effettuare alcuna scelta.

Questo criterio di selezione risulta essere il più severo tra quelli esaminati. Se non ci sono fattori che si distinguono nettamente dagli altri ed allo stesso tempo i primi fattori hanno autovalori che superano di poco l'unità, si può concludere che l'analisi fattoriale non sia il metodo adatto per analizzare i dati in questione.

In alcune occasioni i criteri di scelta del numero di fattori da inserire nell'analisi non portano ad una soluzione univoca, nel senso che più soluzioni soddisfano i requisiti richiesti; in questo caso si decide di dare maggiore importanza alla soluzione che più delle altre soddisfa la logica dell'analisi, escludendo ad esempio quei fattori che pur soddisfacendo i criteri adottati, risultino poco coerenti con l'obiettivo dell'analisi.

II.7 Interpretazione dei fattori

L'interpretazione dei fattori ottenuti applicando la tecnica ACP può essere condotta solo dopo aver accertato l'effettiva esistenza di relazioni tra le variabili, presupposto necessario affinché l'interpretazione stessa sia plausibile.

L'esistenza di relazioni tra le variabili viene verificata principalmente attraverso l'analisi del grafico degli autovalori. Affinché sia accettabile l'esistenza di fattori esplicativi, è necessario che lo scree test (cioè il grafico degli autovalori) presenti un flesso, individuando fattori più significativi di altri; l'assenza di questa struttura, ed in particolare la presenza di un andamento rettilineo, indica l'inefficienza di un'analisi di tipo fattoriale, in quanto i fattori risulterebbero una semplice trasformazione dei dati iniziali, o meglio, non esisterebbero sufficienti relazioni tra le variabili iniziali che giustifichino la costruzione di variabili fittizie atte a riassumere quelle tra di loro correlate.

Questa condizione, che può apparire molto riduttiva, è importante al fine di ottenere un'analisi significativa; ovviamente, sarebbe possibile applicare l'ACP anche nel caso in cui questa non fosse pienamente soddisfatta, ma in tal caso i risultati ottenuti sarebbero poco significativi.

Una prima interpretazione dei fattori si basa sul concetto di "saturazione"; la saturazione, tra un fattore ed una variabile, non è altro che il coefficiente di correlazione tra i due, altrimenti detto "peso fattoriale". L'analisi viene quindi condotta sulla matrice di detti coefficienti di correlazione, ricercando quei coefficienti che presentino il più elevato valore assoluto; quanto più questo valore è alto (sia esso con segno positivo o negativo), tanto più la variabile si considera determinante per quel fattore; in tale caso si dice che la variabile "satura" il fattore.

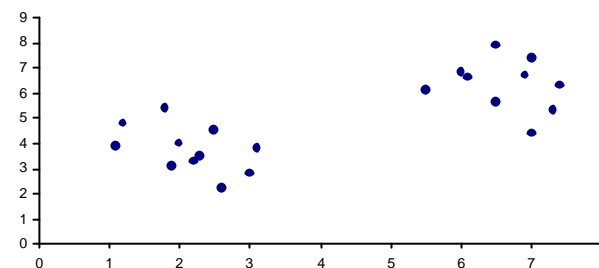
Non si trovano in letteratura delle regole rigide che fissino una soglia oltre la quale si può considerare la saturazione rilevante; Overall e Klett (1972) suggeriscono un valore minimo assoluto, che è pari a 0,35. Effettuando un ragionamento logico e partendo dal fatto che il livello delle correlazioni è legato alla variabilità del fattore, risulta plausibile abbassare tale livello in funzione della radice quadrata degli autovalori.

La conseguenza di questo tipo di ragionamento è che un fattore risulta meglio rappresentato dalle variabili che lo saturano, le quali per questo motivo diventano identificatrici del fattore. Può anche accadere che un fattore sia “bipolare”, cioè abbia alti livelli di saturazione sia con segno positivo sia con segno negativo; in questo caso il fattore risulta denominato dall'accostamento delle variabili che stanno ai due estremi. Nel caso in cui più variabili saturino un solo fattore, si escludono quelle in eccesso, poiché ai fini dell'interpretazione è bene che un fattore sia denominato da una sola variabile; l'esclusione delle variabili in “soprannumero” è giustificata dal fatto che queste saranno con molta probabilità correlate tra di loro, quindi “equivalenti” al fine di identificare il fattore.

Un'altra tecnica di analisi della saturazione può essere condotta riferendosi alla rappresentazione grafica dei pesi fattoriali sui piani definiti da coppie di fattori. Questo tipo di rappresentazione rende l'analista in grado di osservare la presenza di forme o posizioni particolari di classi di entità. Le forme che

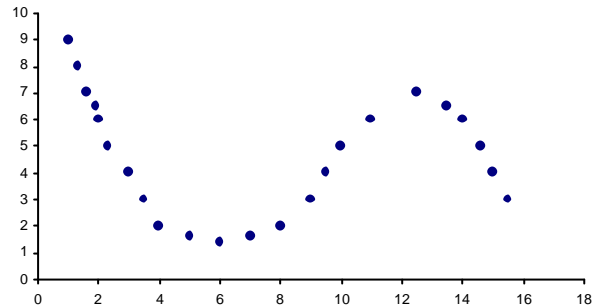
vengono maggiormente rilevate sono la concentrazione di nuvole di punti e le configurazioni curve, denominate “effetto Guttman”.

Il caso della presenza di nuvole di punti in porzioni isolate dello spazio fattoriale sottintende l'appartenenza delle variabili che fanno parte della nuvola ad uno stesso fattore, giustificando l'assunzione della presenza di gruppi:



Il rischio che si corre in questo caso è che uno degli assi passi sopra alla nuvola di punti, rendendo più difficoltosa l'interpretazione grafica; questo inconveniente può essere superato effettuando una rotazione degli assi.

La forma più frequente assunta dai pesi fattoriali negli assi sopra descritti è quella curvilinea, e nella maggior parte dei casi tale forma è quella arcuata, o a “ferro di cavallo”, detta appunto “effetto Guttman”, alla quale può seguire un rientro verso l'interno dei punti terminali:



quest'ultimo effetto è causato in genere dalla mancata standardizzazione delle variabili o dall'attribuzione di un peso errato alle unità.

L'effetto Guttman può però essere soltanto apparente; ciò si spiega ricordando che i pesi fattoriali non sono altro che i coefficienti di correlazione tra le variabili ed i fattori e, su un piano cartesiano descritto da coppie di fattori ortogonali, sono inseriti in un cerchio di raggio unitario e tendono ad assumere forme circolari e, a volte, a mezzaluna.

Interpretare un fattore significa dunque “dargli un nome”, cioè accostarlo alla variabile che meglio lo rappresenta; l'interpretazione dei fattori produce quindi un ordinamento delle variabili che risulta importante nel momento in cui si ha la necessità di individuare quelle più rappresentative nell'insieme iniziale. E' questo l'argomento cardine del problema della scelta delle variabili.

II.8 La scelta delle variabili

Negli studi statistici multivariati si ha spesso a che fare con un elevato numero di variabili; poiché alcune di esse possono fornire informazioni ripetitive o inutili, appesantendo l'indagine e rendendone meno chiara l'interpretazione, può essere utile ai fini dell'analisi individuare un sottogruppo di variabili eliminando quelle ridondanti.

L'utilizzo della tecnica ACP permette di individuare dei criteri di selezione delle variabili, che conducono alla determinazione di un sottogruppo di m variabili ($m < p$), più ristretto rispetto a quello iniziale; per raggiungere tale risultato si parte dal calcolo dell'ACP per tutte le p variabili, se ne eliminano una certa quantità fino ad ottenere la numerosità desiderata (m). I metodi di selezione delle variabili più importanti sono due.

Il primo consiste nell'associare le variabili ad ogni fattore e nell'individuare quelle che presentano il coefficiente di correlazione più alto, in valore assoluto; considerando i primi m fattori, in ordine, le variabili più rappresentative, che saranno quindi scelte, saranno quelle associate a questi ultimi⁹.

Il secondo metodo segue le linee teoriche del primo, ma in questo caso vengono eliminate le variabili con il più elevato coefficiente di correlazione (sempre in valore assoluto) associato agli ultimi ($p - m$) fattori; la logica di

⁹ Si veda a proposito l'interpretazione fattoriale spiegata nel paragrafo precedente.

quest'ultimo criterio sta nel fatto che le ultime CP, con autovalori prossimi a zero, corrispondono a relazioni lineari pressoché costanti per gruppi di variabili, per cui l'eliminazione di una di esse comporta una piccola perdita di informazioni. (Zani, 1999)

Fabbris (1997) suggerisce un metodo alternativo per selezionare le variabili più esplicative; partendo dal presupposto che le variabili sono in genere intercorrelate a gruppi di tre, quattro, o più, per selezionare le variabili che arricchiscono la denominazione dei fattori in modo non ripetitivo, si può effettuare un'analisi di regressione "stepwise", con i punteggi fattoriali come variabili criterio e le variabili analizzate in funzione di variabili esplicative.

Abbiamo qui presentato l'ACP non solo come una tecnica di riduzione dei dati, ma anche come uno strumento per la scelta delle variabili. Tali risultati sono particolarmente funzionali nelle applicazioni reali, le cui difficoltà principali, lo ricordiamo, riguardano la numerosità dei dati e la ridondanza delle informazioni proposte, che ostacolano non solo la fattibilità dell'analisi, ma anche l'interpretazione e la presentazione dei risultati.

CAPITOLO TERZO: CLUSTER ANALYSIS

III.1 Introduzione

L'analisi dei gruppi, o secondo la terminologia anglosassone "cluster analysis" (CA), è un metodo esplorativo che consente di individuare dei gruppi di unità tra loro simili, nell'insieme delle n osservazioni p -dimensionali che costituiscono il data set iniziale; il suo obiettivo principale è quello di isolare gruppi di unità la cui esistenza sia giustificata da un pattern "naturale", nel senso che si richiede che tali gruppi siano intrinseci (anche se non immediatamente visibili) alla struttura dei dati.

Per poter individuare un gruppo, ed in particolare per poter effettuare una distinzione tra i vari tipi di gruppo, è bene enunciare alcune caratteristiche, quali la densità, la dimensione, la forma e la separazione. Dire che un gruppo è "denso" significa che in esso è possibile trovare un agglomerato di punti più fitto rispetto a quello degli altri gruppi; non esiste una misura assoluta della densità, ma intuitivamente essa può essere identificata con la varianza. E' possibile definire la dimensione di un gruppo, in relazione alla misura del suo raggio, soltanto nel caso in cui esso abbia forma regolare, come ad esempio un ipersfera.

La forma è semplicemente la disposizione dei punti-unità nello spazio, che permette di identificare possibili regolarità nella struttura dei dati. La separazione è il modo in cui i gruppi, disposti nello spazio, risultano distinti o, al contrario, si sovrappongono.

Lo stimolo principale per le ricerche sulle metodologie di clustering è stato dato dal lavoro dei biologi Sakal e Sneath, dal titolo “Principles of Numerical Taxonomy”, pubblicato nel 1963; da qui lo sviluppo di un’ampia letteratura a riguardo, che annovera però anche lavori precedenti, quali quelli di Gilmour (1951) e di Cain (1962). Le ragioni principali della rapida crescita di tale metodologia riguardano la possibilità (svilupata soprattutto in tempi recenti) di usufruire di supporti informatici sempre più rapidi, ed il diffondersi del concetto di classificazione come metodologia scientifica. E’ possibile infatti trovare applicazioni delle tecniche di clusterizzazione nella tassonomia, nella psicologia, nella linguistica, nella ricerca di mercato, ecc.

Uno dei risultati più importanti della cluster analysis riguarda la riduzione delle dimensioni dello spazio delle unità, che, per una classificazione che comprende g gruppi, passa da R^n a R^g ; questa riduzione porta grandi benefici per quel che riguarda la facilità di descrizione e di interpretazione del fenomeno.

Da un punto di vista prettamente operativo, la cluster analysis viene utilizzata per:

- effettuare una riduzione dei dati; tale riduzione deve avere il doppio vantaggio di essere semplice e facilmente comprensibile, al fine di evidenziare la struttura informativa dei dati più rilevante in modo accessibile anche a chi non risulta particolarmente esperto, e deve inoltre basarsi su poche dimensioni, in modo da poter essere utilizzata per una agevole presentazione dei risultati di analisi di tipo multivariato;
- generare delle ipotesi di ricerca; una CA consente di individuare delle relazioni tra le unità statistiche che consentono all’analista di impostare un modello che tenga conto di questi legami, evitando che siano applicate tecniche non coerenti con il tipo di data set a disposizione;
- identificare delle tipologie particolari alle quali si possano ricondurre le unità, che verrebbero così distinte tra di loro in base a caratteristiche di gruppo;
- costruire delle tecniche di classificazione automatica, da utilizzare soprattutto per l’inserimento a posteriori di entità all’interno di gruppi o categorie; l’importanza di questo aspetto della CA si coglie facilmente se si pensa alla “tassonomia”, settore proprio delle scienze biologiche che studia la classificazione e la nomenclatura degli esseri viventi e dei fossili;
- stratificare la popolazione; questo aspetto risulta particolarmente rilevante, come è noto, nelle tecniche di campionamento, che vengono effettuate in via preliminare prima di condurre una qualsiasi indagine statistica; stratificando la popolazione è possibile tener conto dei vari “tipi” di unità che ne fanno parte,

evitando la creazione di un campione che, altrimenti, risulterebbe poco rappresentativo;

- attribuire ad entità valori noti soltanto per la classe; capita sovente nelle indagini reali di dover affrontare il problema dei dati mancanti; l'applicazione di una tecnica di clustering dà la possibilità di sostituire i dati mancanti in base alle caratteristiche del gruppo di appartenenza.

La logica del processo di clusterizzazione si rifà alla costruzione di una partizione degli n elementi iniziali in gruppi, caratterizzati dagli obiettivi simultanei di massimizzazione della variabilità esterna ai gruppi e di minimizzazione della variabilità interna ai gruppi, in modo che questi ultimi siano il più possibile rappresentativi.

Per meglio capire l'utilità pratica di questa argomentazione, che in alcune sue parti potrebbe sconfinare nella pura trattazione metodologica, basta pensare a come una CA condotta in via preliminare su un data set di tipo commerciale, costituito ad esempio dai clienti di una determinata zona (o dell'intero paese), possa agevolare una qualsiasi azienda nella pianificazione della propria produzione.

III.2 Le fasi della Cluster Analysis

Un'analisi di raggruppamento pone le proprie basi su una serie di decisioni preliminari, dalle quali deriva l'intero processo di classificazione. Queste decisioni riguardano:

- l'identificazione delle variabili di classificazione e, nel caso in cui tali variabili siano espresse in ordini di misura differenti, la standardizzazione delle stesse;
- la selezione di una misura di distanza tra le unità (nel caso di caratteri quantitativi) o di similarità (nel caso di caratteri qualitativi) e costruzione della relativa matrice (si è già visto nella definizione data di gruppo omogeneo di come una CA si fondi sul concetto di distanza)¹⁰;
- la scelta della tecnica di raggruppamento più adatta alla struttura e all'obiettivo dell'analisi;
- l'identificazione del numero di gruppi entro i quali classificare le unità;
- la scelta (facoltativa) dell'applicazione di altre tecniche multivariate per la lettura dei risultati dell'analisi.

¹⁰ Per una trattazione approfondita delle varie misure di distanza si veda l'Appendice A.

La scelta delle variabili deriva dalle finalità assegnate alla CA, che dipendono in modo quasi esclusivo dalle conoscenze del ricercatore riguardo al fenomeno oggetto di studio. Applicare in modo preliminare un'analisi in componenti principali a tutte le variabili disponibili può fornire un indirizzo privilegiato nella scelta delle variabili da utilizzare nella CA. Se dall'ACP così condotta deriva che k componenti spiegano una quota elevata di variabilità totale, allora la classificazione potrà essere condotta direttamente sugli scores di tali CP che, come noto, racchiudono le informazioni più rilevanti.

Per scegliere la metrica sulla quale incentrare la CA ci si rifà alle caratteristiche proprie delle varie metriche proposte in letteratura (si rimanda anche in questo caso all'Appendice A); questa scelta condiziona i risultati della classificazione, poiché facendo variare il tipo di distanza (nel caso di dati quantitativi), si fa variare anche l'ordinamento tra le unità, con la conseguente creazione di gruppi diversi per quanto riguarda l'omogeneità dei propri elementi.

La scelta della tecnica di raggruppamento parte, come già accennato, dall'obiettivo basilare di creare gruppi di unità con caratteristiche di coesione interna, cioè le unità appartenenti al medesimo gruppo devono essere simili tra loro, e separazione esterna, che equivale a richiedere che i gruppi siano il più possibile distinti tra di loro; si vengono così a formare nel data set iniziale delle classi di unità simili tra di loro, distinguibili in modo più o meno netto a seconda delle tecniche usate.

I metodi di formazione dei gruppi sono numerosi e verranno illustrati nel seguito.

Nello scegliere il numero dei gruppi ci si deve porre come obiettivo principale la creazione di una suddivisione dello spazio dei dati che sia significativa e semplice da analizzare, in modo da rendere tangibile il miglioramento ottenuto passando dalla struttura dei dati originari (non classificati) a quella ottenuta dalla clusterizzazione.

Utilizzare altre tecniche di analisi multivariata dopo aver effettuato una CA non può che produrre un arricchimento del contenuto informativo dell'analisi, e per questo tali tecniche sono di solito quelle grafiche che, come è risaputo, facilitano l'interpretazione e la presentazione dei risultati.

Nella maggior parte dei casi, gli algoritmi di classificazione automatica non producono un risultato certo, in quanto molte scelte sono attribuite all'analista, ma forniscono soluzioni che hanno carattere di ottimi relativi o locali; per questo motivo l'ottimalità della soluzione dipende dalle condizioni iniziali, dalle scelte strategiche effettuate a priori, quali l'accostamento di dati compatibili, la scelta della misura di distanza e dell'algoritmo di classificazione, ma soprattutto dall'esistenza o meno di "gruppi naturali" nell'insieme delle unità che si deve classificare.

III.3 Verifica dell'esistenza dei gruppi

La verifica dell'esistenza "naturale" di gruppi nell'insieme dei dati è un passo importante da compiere prima di svolgere una qualsiasi tecnica di classificazione; risulterebbe infatti inutile applicare un algoritmo di clustering a dei dati che non presentano alcuna struttura agglomerativa. Da un punto di vista pratico, non sarebbe impossibile ottenere dei gruppi da dati senza alcun pattern, ma la soluzione dell'analisi risulterebbe falsata ed artificiale e, soprattutto, non avrebbe alcuna utilità analitica.

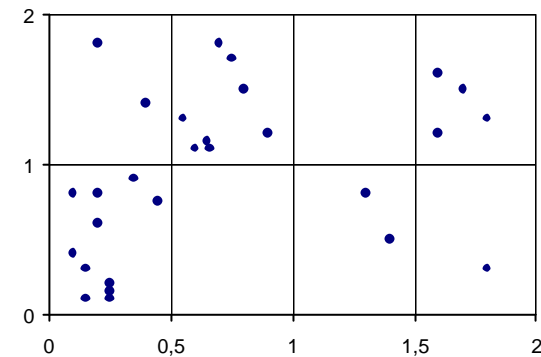
La verifica in questione può essere svolta seguendo vari metodi, alcuni dei quali di natura inferenziale, ma ai fini pratici sembra più opportuno citare un criterio esplorativo semplice e funzionale per lo scopo prefissato. Tale criterio si basa sulla costruzione di un istogramma p-dimensionale, che rappresenta una generalizzazione al caso di p variabili del classico istogramma costruito per una sola variabile. Per meglio illustrare la logica di tale criterio, ci si riferirà al caso banale di due variabili; si costruisce un sistema di assi cartesiani ortogonali dove, una volta fissata un'origine (che può essere arbitrariamente fatta coincidere con i valori minimi), si riportano i punti-variabile, definendo in tal modo uno scatter-plot; quest'ultimo viene quindi suddiviso in rettangoli, che si suppone abbiano

area $h_1 \cdot h_2$, e sia Q il totale dei rettangoli in cui è stato suddiviso lo scatter-plot. Indicando con n_j^{11} , ($j=1, \dots, Q$), il numero di punti che cade nel j-esimo rettangolo; la densità di frequenza:

$$d_j = n_j / (n \cdot h_1 \cdot h_2) \quad \text{per } j=1, \dots, Q$$

esprime il "peso" di ogni rettangolo. La presenza di rettangoli con una forte densità lascia intuire l'esistenza di gruppi nei dati.

A titolo di esempio riportiamo il suddetto grafico per un insieme di 28 unità sulle quali sono state rilevate due variabili:



¹¹ Ovviamente deve risultare:
 $\sum_j n_j = n$

Si ha in questo caso $Q=8$, $h_1=0.5$ e $h_2=1$. Calcolando le densità per i vari rettangoli notiamo in particolare la presenza di un valore elevato per il primo in basso a sinistra, per il quale si ha:

$$d_1=11/(28*0.5*1)=0.78$$

che fa supporre che le unità in esso contenute possano far parte di un gruppo.

Passando dal caso $p=2$ al caso multidimensionale, quando cioè si hanno numerose variabili, va risolto il problema della rappresentazione grafica. Partendo dal presupposto che un grafico di iper-rettangoli p -dimensionali non è fattibile in pratica, il problema può essere risolto in due modi diversi; il primo è quello riguardante la costruzione di diagrammi bidimensionali di coppie di variabili, ma nel caso in cui queste siano in numero elevato non risulterebbe facile analizzare tutti i confronti possibili; il secondo, che di solito è quello che viene preso maggiormente in considerazione, riguarda il ricorso alle CP; se le prime due o tre CP spiegano una quota significativa della variabilità totale, si possono costruire gli scatter-plot bidimensionali riferendosi soltanto a queste ultime e analizzando le densità di ogni rettangolo si possono inoltre ottenere le indicazioni riguardanti il possibile numero di gruppi.

Se si applica una CA senza tener conto di queste informazioni preliminari, si rischia di avere un risultato che è una pura conseguenza dell'imposizione ai dati di una struttura che è a loro estranea, senza poter raggiungere obiettivi concreti; se si applica invece una simile verifica è possibile evitare una distorsione della

struttura del data set, che provocherebbe la perdita di importanza dell'analisi condotta.

III.4 Le tecniche di clustering

Le tecniche di clustering hanno come finalità propria la costruzione di una partizione della matrice dei dati iniziali. Se si considera una matrice di n dati per p variabili, appartenenti ad un insieme dotato di misura di distanza (quindi ad uno "spazio metrico"), la clusterizzazione tende a suddividere gli n elementi in G gruppi, dove il numero dei gruppi, solitamente, non è noto a priori; detti gruppi dovranno essere omogenei al loro interno e separati tra di loro. Indicando con C_g il generico gruppo di unità, la partizione generata può essere espressa da:

$$\{C_1, C_2, \dots, C_g, \dots, C_G\}$$

Sembra opportuno dare, in questa fase preliminare della trattazione, una definizione assiomatica di gruppo omogeneo, in funzione di una matrice di distanze $D(d_j)$:

Si consideri un collettivo di n unità aventi una matrice di distanze D e si fissi una soglia $h > 0$; le possibili distanze tra le unità statistiche di un gruppo sono pari a $n(n-1)/2$. Un gruppo si dice omogeneo quando ognuna delle possibili distanze non è superiore alla soglia h :

$$d_{ij} = h \quad (i \neq j \ 1, 2, \dots, n);$$

inoltre, la media delle distanze di ogni unità da tutte le altre deve essere non superiore ad h :

$$[1/(n-1)] \sum_{j=1, n} d_{ij} = h \quad (i \neq j \ 1, 2, \dots, n)^{12}.$$

La partizione indotta dal processo di clusterizzazione sarà tanto migliore quanto più elevato risulterà il rapporto tra variabilità esterna/variabilità interna, rapporto che cresce all'aumentare del numero dei gruppi. Adottando questo criterio, si vogliono costruire dei gruppi che siano il più possibile omogenei al loro interno ed il più possibile eterogenei tra di loro.

Le varie tecniche di clustering portano a partizioni differenti a seconda del criterio adottato per valutarne l'omogeneità interna. La distinzione principale di queste tecniche si fonda sulle caratteristiche proprie dei gruppi da esse formati, a seconda che questi possano o meno essere inseriti in una struttura gerarchica; da questa peculiarità deriva poi la denominazione di "metodi gerarchici" e "metodi non gerarchici".

I metodi gerarchici portano a suddivisioni in cui ogni classe fa parte di una classe più ampia, la quale è, a sua volta, contenuta da una classe di ampiezza maggiore, fino a giungere alla classe che contiene tutto l'insieme delle unità prese in considerazione. Le tecniche di analisi gerarchica possono essere ulteriormente distinte in agglomerative (di solito le più usate) e scissorie. Le

¹²A. Rizzi "Analisi dei dati" Nuova Italia Scientifica.

tecniche agglomerative partono da gruppi elementari, costituiti da una sola unità, dalla cui fusione vengono generati gruppi sempre più ampi, fino a giungere allo stadio finale in cui si ha un solo gruppo che racchiude tutte le unità. Le tecniche scissorie procedono in senso contrario alle precedenti: esse partono da un gruppo iniziale comprendente tutte le unità, che viene suddiviso in sottoinsiemi sempre più piccoli, fino alla costituzione di una partizione comprendente n gruppi elementari.

I metodi non gerarchici generano dei gruppi che non possono essere inseriti in nessun ordine gerarchico, il cui numero deve essere fissato a priori; questi sono confrontabili soltanto mediante degli indici sintetici della classificazione complessiva, e non gruppo per gruppo. I metodi non gerarchici si distinguono anch'essi in due categorie: quelli che generano partizioni, cioè classi mutuamente esclusive dove una unità può far parte di un gruppo soltanto, e quelli che generano classi sovrapposte, dove si contempla la possibilità di catalogare un'unità in più di una classe; quest'ultimo tipo di metodo è solitamente annoverato tra i metodi "alternativi" di classificazione.

III.5 Metodi gerarchici di classificazione

I metodi gerarchici vengono utilizzati principalmente quando si ha la necessità di investigare la struttura dei dati a differenti livelli. Un metodo di formazione dei gruppi si dirà gerarchico se:

- considera tutti i livelli di distanza;
- i gruppi che si ottengono ad un certo livello di distanza comprendono i gruppi ottenuti ad un livello di distanza inferiore.

Applicando un metodo gerarchico si ottiene una famiglia di partizioni delle n unità, partendo da quella banale in cui ogni gruppo è costituito da una sola unità ed arrivando a quella altrettanto banale in cui tutte le unità sono racchiuse in un unico gruppo (metodo agglomerativo).

Quasi tutte le tecniche di clustering gerarchico partono da una matrice di distanze D calcolata sulle n unità statistiche, e seguono lo schema riportato:

- individuazione nella matrice D delle due unità con la minore distanza; tali unità verranno riunite per formare il primo gruppo. La partizione iniziale, che contava n gruppi, sarà quindi sostituita da una partizione ad $(n-1)$ elementi in cui $(n-2)$ sono costituiti da una sola unità e l'altro sarà formato dalle due unità precedentemente scelte;
- si ricalcola, secondo un criterio prescelto, la distanza tra il gruppo ottenuto e gli altri gruppi, ottenendo una nuova matrice di distanze, che risulterà ovviamente di dimensioni inferiori di uno rispetto alla matrice precedente;

- si individuano nella nuova matrice i due gruppi più prossimi, che verranno riuniti in un solo gruppo, e si ricalcola la distanza tra questo nuovo gruppo ed i restanti, generando un'ulteriore matrice di distanze;

- si itera il processo fino ad un solo gruppo costituito da tutte le unità.

Come è facile intuire, la differenza tra i vari metodi gerarchici risiede proprio nel criterio utilizzato per calcolare la distanza tra i gruppi di unità; un'ulteriore distinzione può essere fatta tra i metodi che si riferiscono alla matrice delle distanze e quelli che si riferiscono alla matrice dei dati.

Si supponga di avere due clusters, C_1 e C_2 , composti rispettivamente da n_1 e n_2 unità; indicando con $d(C_1, C_2)$ la distanza tra i due gruppi e con d_{rs} la distanza tra due unità ($r \in C_1$ e $s \in C_2$); i principali metodi adottati per il calcolo delle distanze tra due distinti gruppi di unità sono:

- **Metodo del legame singolo** (single linkage), anche detto “del vicino più prossimo”, secondo il quale la distanza tra due gruppi è data dal minimo tra le distanze tra coppie di unità appartenenti ai due gruppi:

$$d(C_1, C_2) = \min (d_{rs}) \quad \text{per } r \in C_1 \text{ e } s \in C_2$$

Si aggregano così due gruppi ai quali appartengono le unità con minima distanza. I gruppi che si ottengono con questo metodo hanno una forma allungata, quindi tale tecnica non è idonea nella ricerca di gruppi omogenei, poiché, come suggerito dal nome stesso, i gruppi da esso generati sono costituiti

da unità aventi un “solo legame”. Applicando il metodo del legame singolo è possibile ottenere “l’albero di minima lunghezza” (Minimum Spanning Tree, MST), che permette di verificare la bontà del dendrogramma¹³ ottenuto con questa tecnica; assegnando una lunghezza a tutti gli $n(n - 1)/2$ possibili punti ottenuti in un grafico rappresentante i gruppi, un MST è un albero che non contiene cicli, in cui la somma di tali distanze è minima.

- **Metodo del legame completo** (complete linkage), anche detto “del vicino più lontano”, secondo il quale la distanza tra i due gruppi è data dal massimo tra le distanze tra coppie di unità appartenenti ai due gruppi:

$$d(C_1, C_2) = \max (d_{rs}) \quad \text{per } r \in C_1 \text{ e } s \in C_2$$

Come conseguenza di tale criterio si ha che tutte le distanze tra le unità di un gruppo e quelle di un altro gruppo saranno minori o uguali alla distanza tra i gruppi. Applicando questo metodo si ottengono gruppi di forma circolare, caratterizzati da notevole somiglianza interna. Il dendrogramma rappresentante i gruppi risulta facile da interpretare, ma spesso è difficile trovare una concordanza tra i gruppi generati e il data set iniziale.

- **Metodo del legame medio** (average linkage), anche detto metodo della mediana, secondo il quale la distanza tra due gruppi è data dalla media aritmetica di tutte le $n_1 * n_2$ distanze tra ciascuna unità di un gruppo e ciascuna unità dell’altro gruppo:

¹³ La definizione di dendrogramma sarà data in seguito.

$$d(C_1, C_2) = \frac{1}{n_1 * n_2} \sum_r \sum_s d_{rs} \quad \text{per } r \in C_1 \text{ e } s \in C_2$$

Tale metodo dà maggiore importanza ai gruppi di piccole dimensioni ed è stato proposto per superare gli eccessi nei quali si poteva incorrere adottando i metodi precedenti.

Tra i metodi gerarchici che utilizzano la matrice dei dati i più interessanti sono:

- **Metodo del centroide**, in cui la distanza tra due gruppi è data dalla distanza dei rispettivi centroidi¹⁴:

$$d(C_1, C_2) = d(x_1, x_2)$$

In ogni fase il metodo fonde i gruppi la cui distanza tra i centroidi è minima. Tale metodo presenta delle analogie con il metodo del legame medio; quest’ultimo infatti considera la media di tutte le possibili distanze tra le unità appartenenti a gruppi diversi; il metodo del centroide, seguendo una logica simile, individua preliminarmente il centro di ogni gruppo, quindi ne misura la distanza.

- **Metodo di Ward**, in cui si definisce una funzione obiettivo di minima varianza; dovendo la tecnica di clustering massimizzare la coesione interna tra i gruppi, tale metodo parte dalla Devianza totale delle p variabili:

$$\text{Dev}(T) = \sum_{i \in T_1}^n \sum_{k \in T_2}^p (x_{ik} - \bar{x}_k)^2$$

¹⁴ Il “centroide” è il vettore contenente i valori medi delle p variabili.

dove T è la matrice degli scarti dalle medie. Dev(T) può essere scomposta tra la devianza nei gruppi:

$$\text{Dev}(W) = \sum_{g=1}^G \sum_{k=1}^p \sum_{i \in g} (x_{ikg} - \bar{x}_{kg})^2$$

dove W è la matrice degli scarti dalle medie di gruppo, e la devianza tra i gruppi:

$$\text{Dev}(B) = \sum_{g=1}^G \sum_{k=1}^p n_g (\bar{x}_{kg} - \bar{x}_k)^2$$

dove B è la matrice degli scarti tra le medie di gruppo e le medie generali.

I gruppi scelti per essere aggregati ad ogni passo della procedura dovranno essere quelli che comportano il minor incremento della devianza nei gruppi. La devianza interna ad un gruppo viene calcolata in base agli scostamenti dal centro del gruppo delle unità che ne fanno parte; l'aumento del suo valore indica una perdita di omogeneità interna o, in altre parole, l'introduzione nel gruppo di elementi che gli sono estranei. Essa è pari a zero per la partizione banale in n gruppi ed aumenta a ciascun passo dell'aggregazione fino ad eguagliare la devianza totale quando si giunge alla partizione finale in cui tutte le unità sono riunite nello stesso gruppo.

La distanza tra due gruppi sarà data da:

$$d(C_1, C_2) = (n_1 * n_2 / (n_1 + n_2)) \|x_1, x_2\|^2$$

che non è altro se non il quadrato della distanza euclidea moltiplicato per una quantità proporzionale alla numerosità dei gruppi. Il metodo è stato pensato per applicazioni su distanze euclidee, ma può essere utilizzato per ogni tipo di

distanza. Il dendrogramma risultante offre una chiara partizione delle unità, soprattutto negli stadi finali.

Gli algoritmi gerarchici presi in considerazione possono essere ricondotti ad una formula generale proposta da Lance e Williams (1967). Tale formula esprime la distanza tra un gruppo C_3 ed un gruppo (C_1, C_2) , derivante dall'unione dei due gruppi C_1 e C_2 , tramite le distanze tra i singoli gruppi, cioè tra C_1 e C_2 , C_1 e C_3 , C_2 e C_3 :

$$d [C_3, (C_1, C_2)] = \alpha_1 d_{31} + \alpha_2 d_{32} + \alpha_3 d_{12} + \alpha_4 (d_{31} - d_{32})$$

Facendo variare i parametri α_1 , α_2 , α_3 e α_4 si possono ottenere le espressioni di distanza tra i gruppi enunciati nei principali metodi gerarchici.

Considerando la generica distanza tra un gruppo k ed un gruppo fusione dei gruppi i e j, i valori dei parametri assunti nei vari metodi sono riportati nella seguente tabella:

Metodo	α_1	α_2	α_3	α_4
L. singolo	1/2	1/2	0	-1/2
L. completo	1/2	1/2	0	1/2
L. medio	1/2	1/2	-1/4	0
Centroide	$n_i / (n_i + n_j)$	$n_j / (n_i + n_j)$	$-n_i n_j / (n_i + n_j)^2$	0
Ward	$(n_i + n_k) / (n_i + n_j + n_k)$	$(n_j + n_k) / (n_i + n_j + n_k)$	$-n_k / (n_i + n_j + n_k)$	0

La famiglia di partizioni ottenuta tramite un metodo gerarchico può essere rappresentata da un albero n-dimensionale; in particolare, se si considerano i livelli di distanza tra le partizioni successive, la rappresentazione avviene tramite un particolare strumento grafico, detto “dendrogramma”, del quale si riporta la definizione:

“Si definisce dendrogramma, sull’insieme di n unità A , un’applicazione $D(h):R^+ \rightarrow \mathcal{P}(A)$, dove $\mathcal{P}(A)$ è l’insieme di tutte le partizioni di A e h è il livello di distanza, che soddisfa le condizioni:

1) $D(0)$ è la partizione costituita da n unità distinte, $D(h)$, per $h > 0$ di un opportuno valore soglia, è la partizione formata da un unico gruppo.

2) Se $h' < h$ la partizione $D(h)$ è uguale o più “fine”¹⁵ di $D(h')$; pertanto al crescere del livello di distanza si ottengono partizioni con minor numero di gruppi, costituiti da aggregati di gruppi ricavati ai livelli inferiori.

3) $D(h+\delta) = D(h)$, per $\delta > 0$ sufficientemente piccolo; vi sono cioè degli incrementi del livello di distanza che lasciano invariata la partizione ottenuta”.

(Zani, 1999)

I criteri di taglio del dendrogramma si basano sul calcolo di due indici, l’indice di coesione R^2 , ed il Root-Mean-Square Standard Deviation (RMSSD),

¹⁵ “Una partizione $D(h)$ si dice più fine di un’altra partizione $D(h')$ se ogni gruppo di $D(h')$ è formato da uno o più gruppi di $D(h)$, ma non viceversa.” (Zani, 1999)

che permettono di scegliere la partizione migliore tra quelle individuate dal metodo gerarchico di clustering.

L’indice di coesione è dato dal complemento ad uno del rapporto tra devianza interna ai gruppi e devianza totale:

$$R^2 = 1 - \text{Dev}(W)/\text{Dev}(T)$$

che ovviamente risulta uguale a :

$$R^2 = \text{Dev}(B)/\text{Dev}(T)$$

Esso assume valori nell’intervallo [0,1] ed è perciò confrontabile per partizioni differenti; il valore ottimale di R^2 è quello prossimo all’unità, che implica una classificazione delle unità omogenea all’interno dei gruppi; il suo maggiore inconveniente è che, per la sua struttura, porta a privilegiare le partizioni banali di n gruppi composti da una sola unità. Tale problema può essere superato prendendo in esame l’indice RMSSTD, che considera la componente della devianza Within, la quale a sua volta fa riferimento al gruppo formato al corrispondente passo della procedura di classificazione. Considerando il passo h-esimo ($h=2, \dots, n-1$) della procedura, l’indice RMSSTD è dato da:

$$\text{RMSSTD} = \sqrt{W_h/p(n_h-1)}$$

dove $W_h = \sum_{i=1}^{n_h} \sum_{k=1}^p (x_{ik} - \bar{x}_{kh})^2$ è la devianza delle p variabili nel gruppo definito dal passo h e n_h è la corrispondente numerosità. Come risulta evidente dalla

formula, l'indice RMSSTD perde di significato nel caso banale in cui si considerino n gruppi unitari ($n_h=1$) o un solo gruppo, caso in cui $W_h=W=T$; un forte incremento di RMSSTD tra due passi successivi implica che sono stati riuniti due gruppi fortemente eterogenei tra di loro.

Gli indici R^2 e RMSSTD permettono di valutare il grado di coesione interna ai gruppi; prendendo in considerazione due passi successivi della classificazione, che comprendono rispettivamente $(g+1)$ e (g) gruppi, se risulta che con g gruppi si ha una riduzione modesta di R^2 ed un incremento contenuto di RMSSTD, si passa a considerare la partizione successiva $(g-1)$; se invece si manifesta un salto rilevante nel valore degli indici, si reputa soddisfacente la partizione precedente con $(g+1)$ gruppi.

Un altro indicatore utilizzabile per scegliere tra le diverse partizioni è dato dalla cosiddetta "pseudo-F":

$$F_g = [\text{Dev}(B)/(g-1)] / [\text{Dev}(W)/(n-g)]$$

calcolata per ogni valore di g (numero di gruppi formati). Rappresentando graficamente i valori assunti da F in funzione di g (con F in ordinata e g in ascissa), si possono ottenere diverse situazioni:

- se F cresce al crescere di g , si ha che i dati in esame non risultano adatti ad una suddivisione in gruppi;
- se F diminuisce al crescere di g , tra le unità esiste una struttura gerarchica;

- se F aumenta fino ad un livello massimo e poi decresce, è accettabile pensare ad una suddivisione in un numero di gruppi pari a quello in cui F raggiunge il suo valore massimo.

Per poter effettuare una scelta tra le varie partizioni ottenute dopo aver applicato i metodi gerarchici, è bene considerare alcune proprietà.

La prima proprietà presa in considerazione si riferisce al concetto di "partizione ben strutturata minimale"; la matrice delle distanze ottenuta dalla matrice dei dati iniziali presuppone alcune scelte preliminari, quali il tipo di metrica, eventuali standardizzazioni e ponderazioni delle variabili, ecc.; per questo motivo, affinché i gruppi costituiti partendo da tale matrice abbiano un certo grado di oggettività, è necessario richiedere che la distanza massima all'interno dei gruppi considerati sia minore della distanza minima tra i gruppi. Si arriva quindi alla seguente definizione:

"Una partizione $P = \{C_1, C_2, \dots, C_g, \dots, C_G\}$ di un insieme di n elementi, definita una distanza d , si dice ben strutturata se:

$$\max(d_{ij}) < \min(d_{rs}),$$

per ogni u_i, u_j appartenenti allo stesso gruppo e u_r, u_s appartenenti a gruppi differenti. La partizione ben strutturata contenente il minor numero di gruppi si dice inoltre minimale."¹⁶

¹⁶ Zani, 1999

Si può dimostrare che per ogni matrice di distanze esiste una e una sola partizione ben strutturata minimale.

Un'ulteriore proprietà richiesta ai metodi gerarchici è quella di fornire gli stessi risultati quando si opera una trasformazione crescente delle distanze; in questo caso tali metodi si dicono "invarianti per trasformazione monotona crescente".

Il metodo del legame singolo e quello del legame completo soddisfano quest'ultimo criterio, mentre il metodo del legame medio non lo soddisfa, non godendo l'operatore media aritmetica di tale proprietà.

III.6 Metodi non gerarchici di classificazione

A differenza di quelli gerarchici, i metodi non gerarchici di classificazione forniscono una sola partizione in g gruppi delle n unità considerate, dove g è un numero scelto a priori; questa peculiarità porta il metodo considerato ad impostare i criteri di ottimalità della partizione partendo da questo vincolo iniziale, che se da una parte fa evitare la difficoltà di dover scegliere tra numerose partizioni, dall'altra comporta la necessità di adottare ulteriori criteri che ottimizzino la scelta del numero di gruppi.

I metodi non gerarchici si fondano sulla necessità di collocare le unità all'interno dei gruppi attraverso la specificazione di una funzione obiettivo, che altro non può essere se non quella di minimizzare la devianza interna dei gruppi, secondo un determinato criterio; fissato il numero g dei gruppi che devono essere costituiti, l'algoritmo classifica le unità secondo il criterio prescelto.

I criteri più importanti si rifanno, ovviamente, alla scomposizione della devianza totale:

$$\text{Dev}(T) = \text{Tr}(T) = \sum_{i=1}^n \sum_{k=1}^p (x_{ik} - \bar{x}_k)^2$$

scomponibile in:

$$\text{Dev}(T) = \text{Dev}(W) + \text{Dev}(B)$$

dove, lo ricordiamo, $\text{Dev}(W)$ è la devianza interna ai gruppi:

$$\text{Dev}(W) = \sum_{g=1}^G \sum_{i=1}^{n_g} \sum_{k=1}^p (x_{ikg} - \bar{x}_{kg})^2$$

e $\text{Dev}(B)$ è la devianza tra i gruppi:

$$\text{Dev}(B) = \sum_{g=1}^G \sum_{k=1}^p n_g (\bar{x}_{kg} - \bar{x}_k)^2$$

I vari criteri si prefiggono come scopo o la minimizzazione di W o la massimizzazione di B . La tecnica di gran lunga più usata si riferisce alla minimizzazione della traccia (somma degli elementi diagonali) di W :

$$\min \text{tr}(W)$$

Questo metodo non fa altro che minimizzare la somma dei quadrati delle distanze euclidee all'interno dei gruppi (tale criterio, come si vedrà in seguito, è la base del metodo delle k-medie); esso è invariante per trasformazioni ortogonali, ma porta a diverse soluzioni nel caso si utilizzino dati standardizzati (trasformazioni non lineari); ciò è dovuto al fatto che tale criterio dipende dalla scala di misura delle variabili, in quanto dà un peso uguale alle varianze interne dei gruppi. Non considera le correlazioni tra le variabili, e per questo motivo produce clusters di forma sferica.

Un metodo alternativo è quello che considera il determinante della matrice W:

$$\min |W|$$

Questo metodo è invariante per trasformazioni non lineari, quindi applicabile su dati standardizzati, e risulta molto sensibile alla struttura locale dei dati (Friedman e Rubin, 1967); a differenza del metodo precedente prende in considerazione le correlazioni tra variabili ed identifica clusters di tipo ellissoidale aventi tutti, approssimativamente, la stessa forma.

Le tecniche di classificazione gerarchica sono in genere molto veloci, e ciò porta a ritenerli i più adatti nel caso in cui il numero delle unità sia elevato.

Le fasi della metodologia comprendono:

- scelta della classificazione iniziale delle n unità con un numero di gruppi prefissato a priori;

- calcolo della variazione della funzione obiettivo in base agli spostamenti delle unità tra i vari gruppi, che deve portare ad un miglioramento della coesione interna;

- iterazione del processo finché non viene soddisfatta una regola d'arresto.

L'algoritmo di classificazione gerarchico maggiormente usato nelle analisi reali è quello delle k-medie (MacQueen, 1967), dove k indica il numero dei gruppi che si vuole costruire. La tecnica considerata comprende, una volta fissato k, la scelta di k "poli" iniziali (o "semi"), che sono punti dello spazio p dimensionale (generalmente rappresentati da centroidi), che permettono la costruzione della partizione iniziale, allocando le unità al cluster il cui polo risulta più vicino; dopo avere scelto tali poli, si calcola, per ogni unità, la distanza dai k centroidi e, nel caso una unità risulti più vicina al centroide di un altro gruppo, si rialloca l'unità in quest'ultimo (ciò implica il ricalcolo del centroide sia del vecchio che del nuovo gruppo di appartenenza dell'unità spostata); il processo viene reiterato finché non è più necessario spostare alcuna unità. Per le sue caratteristiche, l'algoritmo considerato è annoverato tra quelli che usano la tecnica "dell'ordinamento rispetto al centroide più vicino", e di solito utilizza una metrica di tipo euclideo, che garantisce la convergenza del processo iterativo. Considerando la t-esima iterazione, la distanza tra l'unità i-esima ($i=1, \dots, n$) ed il centroide del gruppo g-esimo ($g=1, \dots, k$), è data da:

$$d(x_i, x_g^{(0)}) = \sum_{k=1, p}^p (x_{i,k} - x_{k,g}^{(0)})^2$$

Il metodo delle k medie ha come obiettivo quello di generare una partizione che minimizzi la devianza interna dei gruppi, pertanto la bontà della stessa può essere valutata utilizzando l'indice R^2 , mentre non è applicabile in questo caso l'indice RMSSTD, strettamente legato a partizioni gerarchiche.

I maggiori inconvenienti delle tecniche non gerarchiche di clusterizzazione sono legati alla necessità di dover scegliere una partizione iniziale per poter innescare il processo iterativo.

La scelta del numero g di gruppi di cui deve essere costituita la partizione iniziale, può essere effettuata in diversi modi. Se le n unità non sono numerose, si può eseguire l'analisi con diversi valori di g, valutando poi le diverse partizioni in base all'indice R^2 ; questo criterio non offre però alcuna garanzia di trovare un valore ottimo per g, che deve inoltre non essere troppo elevato. Nel caso in cui le n unità siano numerose, il numero g di gruppi può essere fatto derivare da una precedente analisi di tipo gerarchico. Un sistema alternativo consiste nel riferirsi alla rappresentazione grafica dei punti, come quella utilizzata per verificare l'esistenza dei clusters. Beale (1969) propone un algoritmo di tipo probabilistico per individuare un numero di clusters statisticamente significativo; il criterio si riferisce al caso in cui si adotta una matrice di distanze di tipo euclideo.

Si supponga di rilevare k modalità su un collettivo di n unità; considerando m clusters, la variabilità interna complessiva è data da:

$$W(m) = \sum_{g=1, m}^m W_i(m)$$

dove $W_i(m)$ è la somma dei quadrati delle distanze dalle medie aritmetiche, relative all'i-esimo gruppo. Per poter scegliere tra due valori m_1 e m_2 , con $m_1 < m_2$, Beale fa ricorso alla F di Fisher, considerando il rapporto:

$$F(m_2; m_1) = [W(m_1) - W(m_2)]/W(m_2)$$

con $k(m_2 - m_1)$ gradi di libertà al numeratore e $k(n - m_2)$ al denominatore. Tale test fornisce risultati significativi all'aumentare di m; ciò risulta essere un inconveniente nella ricerca di un numero ottimo di gruppi, perciò Beale propone un fattore di correzione per la $F(m_2; m_1)$ che è funzione della diminuzione attesa di $W(m)$ al crescere di m. Poiché si può dimostrare che:

$$E \{ [W(m_1) - W(m_2)]/W(m_2) \} = [(n - m_1)/(n - m_2)](m_2/m_1)^{2k} - 1$$

la $F(m_2; m_1)$ corretta risulta:

$$F(m_2; m_1) = \{ [W(m_1) - W(m_2)]/W(m_2) \} \{ [(n - m_1)/(n - m_2)](m_2/m_1)^{2k} - 1 \}$$

con $k(m_2 - m_1)$ e $k(n - m_2)$ gradi di libertà. Il test è calcolato per ogni coppia m_1 e m_2 , fino ad un massimo prefissato. Se $F(m_2; m_1)$ è significativa ad un prefissato livello di probabilità, si sceglierà un numero di gruppi pari a m_2 .

Per individuare infine i poli che costituiscono i centroidi dei clusters della partizione iniziale, si può banalmente considerare le prime g osservazioni p

dimensionali dell'insieme dei dati, oppure estrarre un campione casuale delle n unità iniziali; la semplicità dei criteri citati è però compensata dalla loro incapacità di garantire un'effettiva rappresentatività dei poli così ottenuti. Una tecnica più affidabile è quella di far coincidere i semi della partizione iniziale con i centroidi dei clusters ottenuti da una partizione gerarchica effettuata in via preliminare.

III.7 Altri metodi di classificazione

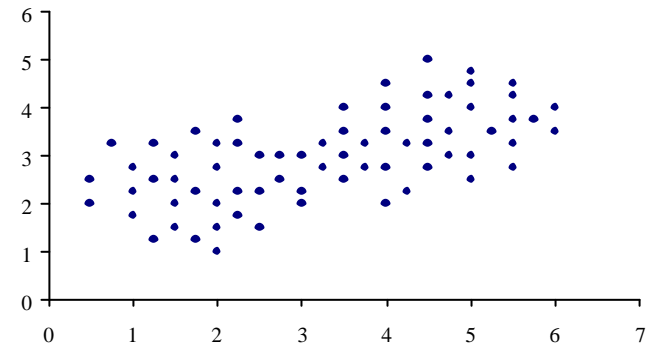
I metodi di classificazione proposti nei paragrafi precedenti hanno come risultato una partizione dell'insieme delle unità iniziali in cui ogni unità fa parte di un solo gruppo del quale, ai fini dell'analisi, assume le caratteristiche. In alcuni casi, però, ci si può imbattere in un certo numero di unità che non presentano caratteristiche "assolute" che permettano di ricondurle ad un unico gruppo. Questo problema ha condotto gli studiosi a ricercare tecniche particolari che giustificano la presenza di una unità in gruppi differenti.

In casi estremi, riscontrabili però nella pratica, è possibile collocare queste particolari unità in gruppi a se stanti; il ricorso a questa particolare

classificazione è contemplato anche nel caso in cui delle unità risultino troppo "distanti" dalle altre (è il caso degli "outliers"); adottando questo criterio si corre però il rischio di incrementare notevolmente il numero dei gruppi, invalidando l'obiettivo di sintesi, proprio di una cluster analysis.

I principali metodi "alternativi" di classificazione sono il cosiddetto "metodo di clumping" ed il "Fuzzy clustering".

I metodi di clumping, studiati in particolare da Gordon nel 1984 ma risalenti ad epoche precedenti, hanno la caratteristica di generare classificazioni non disgiunte, in cui è possibile cioè che una unità sia collocata in gruppi differenti, creando dei clusters sovrapposti:



Tale tecnica risulta valida però soltanto nel caso in cui non ci sia un numero eccessivo di sovrapposizioni; a tale proposito Jardin e Sibson (1971) introducono dei vincoli per limitare il numero delle sovrapposizioni, basati sul cosiddetto metodo B_k ($k=1,2,3,\dots$); secondo questo metodo, il numero massimo di unità che si possono sovrapporre in ogni coppia di gruppi è pari a $(k-1)$; nel caso in cui il numero delle unità che si sovrappongono nei due gruppi superi la soglia consentita, questi verranno fusi a formare un unico cluster.

Zani (1993) sottolinea l'importanza che ricopre la tecnica di clumping nelle classificazioni di zone di un territorio, dove le unità appartenenti a due gruppi rappresentano gli elementi di frontiera tra le due zone individuate; secondo l'autore tale metodologia risulta quindi la più idonea per classificare unità che sono, per loro natura, contigue.

Il secondo metodo citato, introdotto da Zadeh nel 1965, fa riferimento al concetto di "fuzzy set", o insieme sfocato, in cui l'accostamento di una unità ad un gruppo è legata ad una funzione che indica il grado di appartenenza dell'unità al gruppo considerato; la funzione presa in considerazione assume valori compresi nell'intervallo $[0,1]$; in particolare, un valore pari a zero indica l'estraneità dell'unità al gruppo, un valore pari ad uno ne indica l'appartenenza assoluta, mentre i valori intermedi indicano "in che misura" l'unità è legata al gruppo; in tal modo si genera una scala di valori che identificano l'entità del legame tra unità e gruppo.

Secondo il metodo di fuzzy clustering, due unità sono tanto più simili tra loro quanto più prossimo ad uno è il valore della loro funzione di appartenenza al medesimo gruppo (Zadeh, 1977).

Il "grado di sfocatura" di una tecnica di fuzzy clustering è misurato dal cosiddetto "partition coefficient":

$$F = \sum_{i=1,n} \sum_{g=1,G} (u_{ig})^2 / n$$

dove u_{ig} è il valore della funzione di appartenenza dell'unità i -esima al gruppo g -esimo⁶; l'indice assume valori compresi nell'intervallo $[1/G,1]$; in particolare è uguale ad $1/G$ quando ciascuna unità presenta distribuzione uniforme del grado di appartenenza ai diversi gruppi, ed è uguale ad 1 nel caso di gruppi non sfocati, quindi, quanto più piccolo sarà tale indice, tanto più si sarà in presenza di gruppi sfocati.

Per evitare difficoltà di interpretazione legati sia al crescere del numero dei gruppi sia al crescere del loro grado di sfocatura, si introduce un vincolo di contiguità tra le unità assegnate, vincolo che risulterà necessario nel caso già citato di analisi del territorio, in cui le unità di un gruppo devono essere, oltre che simili, anche vicine tra di loro, in modo da poter generare "zone omogenee" di interesse concreto.

⁶ soggetto ai vincoli:
 $0 = u_{i1} = 1$
 $\sum_{g=1,G} u_{ig} = 1$

III.8 Scelta del metodo di classificazione

La scelta del metodo di clustering da adottare nell'analisi di una matrice di dati si basa su una serie di criteri che permettono di valutare la qualità delle varie tecniche; questi criteri riguardano:

- l'oggettività della soluzione; è importante che una tecnica di classificazione abbia la capacità di generare una soluzione che sia riproducibile da chiunque ripeta l'analisi anche in tempi successivi;
- la stabilità della soluzione; nelle analisi reali è possibile imbattersi in dati che presentino errori; la tecnica ottimale deve essere il meno possibile sensibile a piccole variazioni nei dati, in modo tale che l'eliminazione di un'unità non modifichi la struttura dei gruppi;
- l'informatività del risultato; un metodo di classificazione deve essere in grado di produrre dei risultati che incorporino il maggior numero possibile di indicazioni specifiche per l'analisi condotta;
- la semplicità e la rapidità di esecuzione dell'algoritmo di calcolo, in modo da poter condurre agevolmente l'analisi anche su matrici di dati di dimensione elevata.

La scelta principale che deve essere effettuata riguarda le due tipologie di classificazione, ossia quella gerarchica e quella non gerarchica, scelta che deve

tener conto dell'obiettivo finale dell'analisi. A questo proposito si enunciano alcune caratteristiche peculiari delle due metodologie, al fine di rendere l'analista consapevole delle qualità dello strumento che deve adottare.

In genere le tecniche non gerarchiche sono più informative di quelle gerarchiche perché corredate di vari indici che permettono una specifica misura della qualità dei risultati; inoltre i metodi gerarchici risentono maggiormente della presenza di dati anomali e sono particolarmente disturbati da errori di misura. Un importante "pregio" delle tecniche gerarchiche è quello di non richiedere un tempo eccessivo per il calcolo dei risultati, risultati che rischiano però di essere poco rappresentativi nel caso in cui si verificano errori nei primi stadi dell'analisi, in quanto, per le caratteristiche proprie di tali algoritmi, questi errori saranno trascinati fino alla fine.

Nella pratica, al fine di rendere l'analisi il più possibile significativa, si preferisce condurre in successione prima un'analisi di tipo gerarchico e, conseguentemente, un'analisi non gerarchica.

III.9 Interpretazione dei risultati

L'interpretazione dei risultati ottenuti si basa in particolar modo sull'omogeneità interna della classe, considerando questo attributo come il più "desiderabile" ai fini dell'analisi.

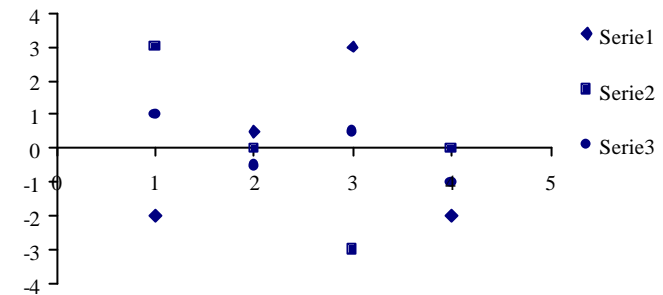
Il rapporto n_{jk}/n_k , dove n_{jk} rappresenta il numero di unità aventi la modalità j tra le n_k unità della classe k -esima, è una misura dell'omogeneità interna al cluster k -esimo e indica la proporzione delle unità del cluster che presentano la modalità considerata. Il rapporto n_{jk}/n_j , dove n_j indica il numero di entità aventi la modalità j tra tutte le n unità, esprime la misura di selettività del cluster rispetto alla modalità j , indicando quante unità che possiedono l'attributo j fanno parte della k -esima classe. Nel caso in cui quest'ultimo rapporto sia inferiore ad uno ($n_j > n_{jk}$), si ha un'elevata probabilità che il gruppo sia omogeneo, nel caso contrario si potrebbero trovare all'interno del cluster unità con caratteristiche differenti, perdendo in omogeneità.

Tra le varie tecniche di interpretazione dei risultati ricoprono un ruolo particolarmente rilevante l'analisi dei profili dei gruppi ed i metodi grafici di analisi di prossimità. Dopo aver individuato la partizione finale, le tecniche gerarchiche e quelle non gerarchiche possono essere trattate allo stesso modo.

Confrontando i valori medi delle singole variabili è possibile stabilire quali variabili siano "discriminanti", quali cioè hanno maggiormente peso nell'identificazione dei gruppi. Dopo aver standardizzato le variabili, si rappresentano i valori medi delle p variabili nei g gruppi finali e li si collega

tramite una spezzata; la spezzata che collega i punti è appunto detta "profilo" del gruppo; quanto più i punti si differenziano su una variabile, tanto più questa variabile risulta discriminante nel formare i gruppi.

Per avere un esempio esplicativo del metodo, consideriamo il caso semplificato di tre gruppi e quattro variabili; i profili dei gruppi sono riportati in figura:



La serie 1 riporta i valori medi delle quattro variabili per il gruppo uno, la serie 2 per il gruppo due e la serie 3 per il gruppo tre; sull'asse delle ordinate troviamo i valori medi delle variabili, mentre su quello delle ascisse i numeri si riferiscono alle quattro variabili. Analizzando il grafico si può notare che le variabili più discriminanti, cioè quelle che maggiormente permettono di distinguere i gruppi, sono la variabile numero uno e la numero tre; ciò porta ad affermare che queste

ultime sono le più importanti per l'analisi condotta, in quanto riescono più di tutte le altre a rappresentare le diverse caratteristiche dei gruppi.

Una volta giunti ad una partizione dell'insieme delle unità, è opportuno effettuare una valutazione del contenuto dei gruppi. Come primo passo è possibile confrontare la media di ogni variabile all'interno dei vari gruppi con la media generale; il confronto tra tali medie deve tener conto della varianza.

Una misura di similarità tra variabili e classi è data dalla statistica test:

$$t_k(X) = \frac{(\bar{X}_k - \bar{X})}{s_k(X)}$$

con:

$$s_k(X) = \sqrt{\left\{ \frac{(n - n_k)}{(n - 1)} * (s^2(X) / n_k) \right\}}$$

dove $s^2(X)$ è la varianza generale della variabile X ed n_k la numerosità del gruppo considerato.

Il valore assunto da t_k esprime la distanza tra la media della classe e la media generale in termini di scarti quadratici medi. E' possibile utilizzare questo valore come una graduatoria tra le variabili caratteristiche di una classe, nel senso che i valori assoluti dei valori-test costituiscono delle semplici misure di similarità fra variabili e classi.

CAPITOLO QUARTO: L'INDIVIDUAZIONE DI GRUPPI DI COMUNI POTENZIALMENTE ADATTI AD INSEDIAMENTI INDUSTRIALI

VI.1 Introduzione

L'analisi in oggetto riguarda gli 80 comuni della provincia di Catanzaro; su tali comuni sono state rilevate 25 variabili, tutte quantitative, delle quali riportiamo un elenco¹⁷:

Var.1: Superficie totale del comune (mq);

Var.2: Altitudine (m);

Var.3: Numero di aree P.I.P.¹⁸ presenti nel comune;

Var.4: Superficie totale di aree P.I.P. presente nel comune (mq);

Var.5: Superficie minima dei lotti delle aree P.I.P. presenti nel comune (mq);

Var.6: Superficie massima dei lotti delle aree P.I.P. presenti nel comune (mq);

Var.7: Totale lotti aree P.I.P. disponibili nel comune;

Var.8: Numero di Aree Industriali presenti nel comune;

Var.9: Superficie totale di Aree Industriali presente nel comune (mq);

¹⁷ Per una descrizione più dettagliata delle variabili rilevate si veda l'appendice B.

¹⁸ Piani per gli Insediamenti Produttivi.

Var.10: Numero medio di infrastrutture presenti nelle eventuali aree del comune;

Var.11: Distanza tra il comune ed il capoluogo (Catanzaro) (km);

Var.12: Distanza tra il comune e l'aeroporto più vicino (km);

Var.13: Distanza tra il comune e la stazione ferroviaria più vicina (km);

Var.14: Distanza tra il comune e l'autostrada A3 (SA-RC) (km);

Var.15: Distanza tra il comune e la strada statale più vicina (km);

Var.16: Numero di laureati;

Var.17: Numero di diplomati;

Var.18: Popolazione Residente;

Var.19: Popolazione Residente compresa tra i 14 ed i 65 anni;

Var.20: Uffici postali;

Var.21: Banche;

Var.22: Numero di Unità Locali;

Var.23: Numero di addetti Unità Locali;

Var.24: Numero di imprese;

Var.25: Numero di addetti imprese.

I dati riguardanti
le variabili numero
1, 4, 16, 17, 18, 19,

22, 23, 24 e 25 sono

stati ricavati da

fonti Istat; i dati

relativi alle altre

variabili sono stati

forniti dalla T.E.A.

s.a.s. (Territorio,

Economia, Finanza

ed Ambiente s.a.s.).

La T.E.A. è una società formata da esperti professionisti con competenza in due aree di interesse: territorio e ambiente, economia e finanza. Le sue attività prevalenti consistono:

- nell'utilizzo di sistemi informativi geografici (G.I.S.) e nell'elaborazione di dati da satelliti per la pianificazione e la gestione del territorio;
- nella predisposizione di studi di fattibilità economico-finanziaria per soggetti pubblici e privati.

I dati forniti sono il frutto di un'indagine commissionata dalla Camera di Commercio, Industria, Artigianato ed Agricoltura alla T.E.A., indagine che si inquadra in un più ampio progetto denominato "Sportello Impresa", il quale ha

come obiettivo la costruzione di un portale di accesso alle informazioni fondamentali per l'attività di governo locale delle attività produttive.

L'indagine è stata condotta con rilevazione censuaria sugli 80 comuni della provincia di Catanzaro, ed ha lo scopo di fornire agli imprenditori informazioni sulle opportunità di localizzazione industriale nella provincia, che risulta essere suddivisa nel seguente modo:

Su un totale di 80 comuni, 43 presentano solo Aree Industriali, 21 solo Aree P.I.P., 7 sia Aree Industriali che Aree P.I.P., 9 né Aree Industriali né Aree P.I.P.

Partendo dal data set fornito dalla T.E.A. ci proponiamo di effettuare un'analisi statistica che consenta l'individuazione sul territorio di gruppi di comuni, distinti per particolari caratteristiche socio-economiche, da utilizzare in studi relativi alla fattibilità di investimenti industriali.

Per ottenere tali risultati abbiamo applicato, in sequenza, un'Analisi in Componenti Principali ed una Cluster Analysis.

IV.2 La riduzione della dimensionalità dell'informazione attraverso l'ACP

Prima di procedere all'esposizione dei risultati ottenuti dall'applicazione della tecnica ACP, è sembrato opportuno proporre una descrizione univariata delle variabili in questione, attraverso degli indicatori semplici ma facilmente comprensibili, e giustificare, attraverso la matrice di correlazione tra le variabili,

l'applicazione della tecnica usata, che trova giusto impiego nel caso di data set caratterizzati dalla presenza di variabili correlate tra di loro. Nella Tabella IV.2.1 riportiamo, per le 25 variabili, i massimi ed i minimi con

i relativi ranges, le medie, le deviazioni standard ed i coefficienti di variazione, in modo da poter effettuare una prima valutazione dell'andamento delle variabili da un punto di vista univariato.

Tabella IV.2.1

	VARIABILE	MINIMO	MASSIMO	RANGE	MEDIA	DEV. ST.	COEFF. V.
1	SUP. TERR.	6.860.000	160.240.000	153.380.000	29.891.875	25.442.117	0,8511
2	ALTITUDINE	6	1.050	1.044	475,6375	195,2189	0,4104
3	N° PIP	0	4	4	0,4500	0,7446	1,6547
4	SUP. TOT. PIP	0	4.840.170	4.840.170	102.738,91	544.709,57	5,3019
5	SUP. MIN. PIP	0	5.150	5.150	460,5000	846,1120	1,8374
6	SUP. MAX. PIP	0	150.000	150.000	3.920,5000	17.430,96	4,4461
7	N° LOTTI PIP	0	114	114	6,7875	18,8409	2,7758
8	N° AI	0	2	2	0,7250	0,6359	0,8772
9	SUP. TOT. AI	0	1.200.000	1.200.000	93.204,83	192.737,47	2,0679
10	MEDIA INFR.	0	6	6	1,9175	2,2554	1,1762
11	DIST. CAPOL.	0	65	65	35,0075	13,9079	0,3973
12	DIST. AEROP.	1,5	89	87,5	47,6813	24,7535	0,5191
13	DIST. FS	1	43,4	42,4	17,1738	9,5755	0,5576
14	DIST. A3	2,6	93	90,4	41,7688	21,1606	0,5066
15	DIST. SS	1	43	42	15,3313	9,9453	0,6487
16	LAUREATI	0	5.924	5.924	147,7125	716,4660	4,8504
17	DIPLOMATI	31	24.290	24.259	755,3875	2.943,57	3,8968
18	POP. RES.	555	96.975	96.420	4.795,50	13.120,76	2,7361
19	POP. ATT.	318	69.040	68.722	3.316,90	9.349,08	2,8186

20	POSTE	1	17	16	1,8000	2,1606	1,2004
21	BANCHE	0	27	27	1,2625	3,8672	3,0632
22	U. L.	13	5.478	5.465	231,4000	713,1652	3,0820
23	ADDETTI UL	14	16.628	16.614	545,7750	2.100,38	3,8484
24	IMPRESE	10	5.077	5.067	216,4500	663,2897	3,0644
25	ADD. IMPR.	12	14.087	14.075	476,6500	1.788,12	3,7514

Ranges elevati sottintendono la presenza di unità molto differenti tra di loro; osserviamo tale fenomeno in modo particolare per le variabili che si riferiscono alle misure di superficie, facendo intuire che nella provincia è possibile trovare comuni che differiscono soprattutto per l'estensione territoriale.

Un indice di variabilità è fornito dal coefficiente di variazione, dato dal rapporto tra la deviazione standard e la media, confrontabile per variabili diverse in quanto depurato dal legame con l'unità di misura; dalla tabella è possibile osservare che i valori assunti da tale coefficiente sono quasi sempre elevati, e ciò implica la presenza di unità molto differenti tra di loro. In particolare, i valori più elevati si hanno con riferimento alle variabili "Superficie totale delle aree P.I.P.", "Superficie massima dei lotti P.I.P.", "Numero di laureati", "Numero di diplomati", "Numero di Unità Locali", "Numero di addetti nelle Unità Locali", "Numero di Imprese", "Numero di addetti nelle Imprese":

Osservando i valori massimi assunti dalle variabili, abbiamo inoltre notato che per la "Superficie territoriale" e le superfici delle aree P.I.P. essi si riferiscono al comune di Lamezia Terme, mentre per le variabili dalla 16 alla 25 (quasi tutte legate alla popolazione, o aventi con essa a che fare) i valori massimi si riferiscono al comune di Catanzaro; in conseguenza di queste osservazioni possiamo già affermare che il

comune di Lamezia Terme risulta essere il più importante per quel che riguarda l'estensione territoriale e la presenza di aree P.I.P., mentre quello di Catanzaro prevale in quanto a popolazione (e ad attività ad essa connesse). In sintesi emerge che i comuni sono abbastanza diversificati, il che giustifica la ricerca di tipologie. La Tabella IV.2.2 rappresenta la matrice di correlazione relativa alle 25

variabili in questione; tale tabella è stata proposta al fine di poter giustificare l'applicazione della tecnica ACP. Come già spiegato ampiamente nella sezione teorica, questa particolare tecnica fattoriale consente una riduzione delle dimensioni del data set, ricavando dalle variabili originarie delle variabili fittizie che hanno il compito di rappresentare le prime. Questa riduzione ha però senso soltanto nel caso in cui esistano delle correlazioni significative tra le variabili; a tale scopo è stata proposta la matrice di correlazione (tabella IV.2.2), osservando la quale è possibile scoprire se esistono le condizioni necessarie per procedere ad una riduzione dei dati di tipo fattoriale¹⁹.

¹⁹ La matrice di correlazione è stata riportata per intero, ma poiché essa è simmetrica sono stati valutati solo i valori al di sotto della diagonale principale.

Tabella IV.2.2

1	2	3	4	5	6	7	8	9	10	11	12	
1	1,00	-0,12	0,33	0,62	0,34	0,58	0,50	0,02	0,16	-0,01	-0,22	-0,19
2	-0,12	1,00	-0,18	-0,19	-0,17	-0,20	-0,17	-0,05	-0,16	-0,04	0,17	-0,02
3	0,33	-0,18	1,00	0,36	0,63	0,40	0,48	-0,19	0,02	0,23	-0,35	-0,29
4	0,62	-0,19	0,36	1,00	0,43	0,96	0,72	-0,13	0,00	0,08	-0,16	-0,23
5	0,34	-0,17	0,63	0,43	1,00	0,43	0,49	-0,35	-0,02	0,08	-0,35	-0,21
6	0,58	-0,20	0,40	0,96	0,43	1,00	0,66	-0,19	-0,05	0,13	-0,14	-0,27
7	0,50	-0,17	0,48	0,72	0,49	0,66	1,00	-0,20	0,11	0,01	-0,29	-0,20
8	0,02	-0,05	-0,19	-0,13	-0,35	-0,19	-0,20	1,00	0,41	-0,13	0,20	0,05
9	0,16	-0,16	0,02	0,00	-0,02	-0,05	0,11	0,41	1,00	-0,05	-0,24	-0,13
10	-0,01	-0,04	0,23	0,08	0,08	0,13	0,01	-0,13	-0,05	1,00	-0,01	-0,18
11	-0,22	0,17	-0,35	-0,16	-0,35	-0,14	-0,29	0,20	-0,24	-0,01	1,00	0,31
12	-0,19	-0,02	-0,29	-0,23	-0,21	-0,27	-0,20	0,05	-0,13	-0,18	0,31	1,00
13	0,01	0,38	-0,25	-0,15	-0,29	-0,10	-0,27	-0,04	-0,22	-0,02	0,31	0,10
14	-0,04	0,11	-0,08	-0,19	-0,01	-0,21	-0,12	-0,15	-0,15	-0,07	-0,04	0,54
15	-0,02	0,59	-0,10	-0,10	-0,14	-0,08	-0,04	-0,19	-0,23	0,05	0,29	0,01
16	0,58	-0,14	0,31	0,49	0,34	0,41	0,60	0,16	0,50	-0,04	-0,30	-0,13
17	0,59	-0,15	0,31	0,50	0,34	0,42	0,61	0,15	0,49	-0,03	-0,30	-0,14
18	0,66	-0,18	0,33	0,66	0,38	0,58	0,69	0,12	0,43	0,00	-0,29	-0,18
19	0,66	-0,19	0,34	0,66	0,38	0,58	0,69	0,12	0,43	0,00	-0,29	-0,18
20	0,64	-0,08	0,35	0,57	0,37	0,50	0,65	0,05	0,38	0,01	-0,25	-0,18
21	0,69	-0,18	0,33	0,68	0,38	0,61	0,70	0,09	0,41	0,01	-0,26	-0,20
22	0,64	-0,18	0,34	0,61	0,37	0,53	0,67	0,13	0,45	-0,01	-0,30	-0,17
23	0,61	-0,16	0,33	0,57	0,36	0,49	0,65	0,14	0,48	-0,02	-0,31	-0,16
24	0,64	-0,18	0,34	0,62	0,37	0,53	0,67	0,13	0,45	-0,01	-0,30	-0,17
25	0,62	-0,17	0,33	0,58	0,36	0,49	0,65	0,14	0,47	-0,03	-0,31	-0,16

(Tab. IV.2.2)

13	14	15	16	17	18	19	20	21	22	23	24	25	
1	0,01	-0,04	-0,02	0,58	0,59	0,66	0,66	0,64	0,69	0,64	0,61	0,64	0,62
2	0,38	0,11	0,59	-0,14	-0,15	-0,18	-0,19	-0,08	-0,18	-0,18	-0,16	-0,18	-0,17
3	-0,25	-0,08	-0,10	0,31	0,31	0,33	0,34	0,35	0,33	0,34	0,33	0,34	0,33
4	-0,15	-0,19	-0,10	0,49	0,50	0,66	0,66	0,57	0,68	0,61	0,57	0,62	0,58
5	-0,29	-0,01	-0,14	0,34	0,34	0,38	0,38	0,37	0,38	0,37	0,36	0,37	0,36
6	-0,10	-0,21	-0,08	0,41	0,42	0,58	0,58	0,50	0,61	0,53	0,49	0,53	0,49
7	-0,27	-0,12	-0,04	0,60	0,61	0,69	0,69	0,65	0,70	0,67	0,65	0,67	0,65
8	-0,04	-0,15	-0,19	0,16	0,15	0,12	0,12	0,05	0,09	0,13	0,14	0,13	0,14
9	-0,22	-0,15	-0,23	0,50	0,49	0,43	0,43	0,38	0,41	0,45	0,48	0,45	0,47
10	-0,02	-0,07	0,05	-0,04	-0,03	0,00	0,00	0,01	0,01	-0,01	-0,02	-0,01	-0,03
11	0,31	-0,04	0,29	-0,30	-0,30	-0,29	-0,29	-0,25	-0,26	-0,30	-0,31	-0,30	-0,31
12	0,10	0,54	0,01	-0,13	-0,14	-0,18	-0,18	-0,18	-0,20	-0,17	-0,16	-0,17	-0,16
13	1,00	-0,01	0,57	-0,19	-0,21	-0,22	-0,22	-0,15	-0,23	-0,23	-0,22	-0,23	-0,22
14	-0,01	1,00	0,13	-0,12	-0,12	-0,15	-0,15	-0,11	-0,16	-0,13	-0,13	-0,13	-0,13
15	0,57	0,13	1,00	-0,15	-0,14	-0,15	-0,15	-0,07	-0,12	-0,14	-0,15	-0,14	-0,15
16	-0,19	-0,12	-0,15	1,00	1,00	0,97	0,97	0,93	0,95	0,98	0,99	0,98	0,99
17	-0,21	-0,12	-0,14	1,00	1,00	0,98	0,98	0,94	0,96	0,99	1,00	0,99	0,99
18	-0,22	-0,15	-0,15	0,97	0,98	1,00	1,00	0,94	0,99	1,00	0,99	1,00	0,99
19	-0,22	-0,15	-0,15	0,97	0,98	1,00	1,00	0,94	0,99	1,00	0,99	1,00	0,99
20	-0,15	-0,11	-0,07	0,93	0,94	0,94	0,94	1,00	0,93	0,94	0,94	0,94	0,94
21	-0,23	-0,16	-0,12	0,95	0,96	0,99	0,99	0,93	1,00	0,99	0,97	0,99	0,98
22	-0,23	-0,13	-0,14	0,98	0,99	1,00	1,00	0,94	0,99	1,00	1,00	1,00	1,00
23	-0,22	-0,13	-0,15	0,99	1,00	0,99	0,99	0,94	0,97	1,00	1,00	1,00	1,00
24	-0,23	-0,13	-0,14	0,98	0,99	1,00	1,00	0,94	0,99	1,00	1,00	1,00	1,00
25	-0,22	-0,13	-0,15	0,99	0,99	0,99	0,99	0,94	0,98	1,00	1,00	1,00	1,00

Escludendo i coefficienti sulla diagonale principale (che sono ovviamente pari ad uno), su un totale di 300 confronti si sono riscontrati 100 coefficienti superiori a 0,45 (valori in neretto ed in rosso), e di questi 100 ben 48 superano

lo 0,70 (valori in rosso). Tale risultato ci consente di affermare che tra le variabili esistono delle correlazioni significative, grazie alle quali è possibile effettuare una riduzione dei dati.

Per l'impiego dell'Analisi in Componenti Principali bisogna tenere conto della natura delle variabili; non essendo queste espresse nella stessa unità di misura è stato necessario ricorrere alla standardizzazione dei dati e quindi procedere alla ricerca degli autovalori partendo dalla matrice di correlazione²⁰.

La prima tabella ottenuta nell'output contiene nelle proprie colonne l'ordine di estrazione degli autovalori, il loro valore numerico, la percentuale di variabilità totale spiegata da ogni fattore e la corrispondente percentuale cumulata:

Tabella IV.2.3

Comp.	λ_i	% di var.	% cum.	Comp.	λ_i	% di var.	% cum.
1	12,335	49,340	49,340	14	0,252	1,006	98,661
2	2,575	10,301	59,641	15	0,208	0,831	99,492
3	2,172	8,688	68,329	16	0,077	0,310	99,801
4	1,555	6,221	74,550	17	0,025	0,100	99,902
5	1,292	5,166	79,717	18	0,017	0,069	99,970
6	0,955	3,820	83,536	19	0,004	0,015	99,985
7	0,708	2,834	86,370	20	0,003	0,010	99,996
8	0,681	2,723	89,094	21	0,001	0,003	99,998
9	0,609	2,435	91,529	22	0,000	0,001	99,999
10	0,467	1,867	93,396	23	0,0002	0,00066	100,000
11	0,397	1,586	94,982	24	2E-05	8,8E-05	100,000
12	0,380	1,521	96,503	25	6E-06	2,4E-05	100
13	0,288	1,152	97,655				

²⁰ I calcoli sono stati effettuati utilizzando il package statistico SPSS, attraverso la procedura di riduzione dei dati di tipo fattoriale.

E' opportuno ricordare che il valore di λ_i è pari alla varianza della corrispondente CP e la percentuale di varianza spiegata è data dal rapporto tra λ_i e la varianza totale, in questo caso pari a 25, quindi la terza colonna si ricava dalla prima (λ_i) dividendo i suoi valori per 25 (varianza totale di 25 variabili standardizzate). La colonna riguardante le percentuali cumulate riveste un ruolo di particolare importanza nella scelta del numero di fattori, come si vedrà in seguito.

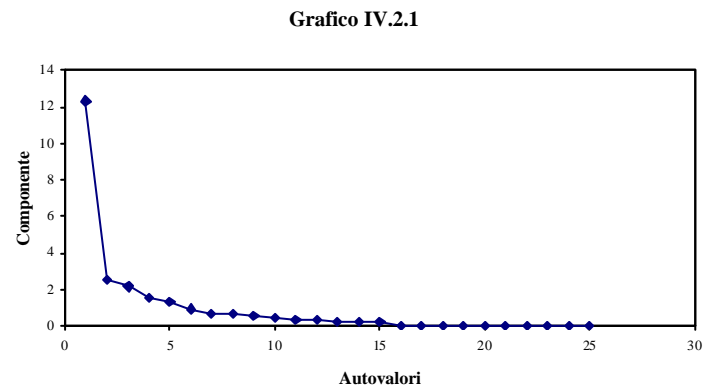
Da una prima analisi della tabella IV.2.3 si nota che i primi autovalori sono più elevati rispetto ai successivi, ed in particolare il primo rispetto agli altri, sottolineando la peculiarità principale della tecnica ACP di fornire i fattori in ordine decrescente rispetto alla varianza; in conseguenza di ciò si ha che nella colonna successiva si riportano percentuali di variabilità spiegata maggiori per i primi fattori, che risultano quindi essere i più significativi.

Per scegliere il numero di fattori significativi si possono seguire diversi metodi. Il primo, e forse il più semplice, consiste nel prendere in considerazione i fattori cui corrispondono autovalori maggiori di uno, metodo adatto nel caso in cui si opera sulla matrice di correlazione; la tabella IV.2.3 rivela che i primi cinque fattori sono associati ad autovalori che superano l'unità.

Un risultato analogo si raggiunge considerando, sempre nella tabella IV.2.3, la percentuale cumulata di varianza spiegata dai fattori; si ha una buona

rappresentazione da parte dei fattori quando tale soglia supera l'80%, e dalla tabella notiamo che i primi cinque fattori sono prossimi a tale quota. Poiché il numero di variabili originarie è elevato, è possibile accettare una quota inferiore di variabilità spiegata; secondo tale logica è lecito quindi considerare anche un numero di fattori pari a tre, cui corrisponde una percentuale cumulata di circa il 68%.

La scelta del numero di fattori può inoltre basarsi sull'andamento degli autovalori, rappresentato nel Grafico IV.2.1, che è stato ottenuto riportando su un sistema di assi cartesiani ortogonali l'ordine di estrazione degli autovalori sull'asse delle ascisse ed il rispettivo valore numerico sull'asse delle ordinate:



Osservando il grafico si nota un brusco cambio di pendenza dopo il secondo autovalore. Tale andamento indica che i fattori più significativi risultano essere i primi, ed in particolare il primo rispetto agli altri, i quali contribuiscono in modo marginale alla spiegazione delle variabili.

Alla luce di
 questo
 ragionamento
 sembra opportuno
 puntare la nostra
 attenzione sui primi
 tre fattori,
 tralasciando i
 successivi il cui
 contributo è del
 tutto trascurabile;
 tale risultato
 soddisfa la regola di
 Harris (1967),
 secondo la quale il

rapporto tra il numero di variabili ed il numero di fattori significativi non deve essere inferiore a due; nel nostro caso si ha infatti che tale rapporto è pari a 8,3. La variabilità spiegata da tali fattori risulta pari al 68% circa.

In relazione ai tre fattori presi in considerazione proponiamo una tabella nella quale vengono riportate le

quote di variabilità di ciascuna variabile spiegate da questi ultimi:

Tabella IV.2.4

VARIABILE	$r_{v=1,2,3}^2$	VARIABILE	$r_{v=1,2,3}^2$
1 SUP. TERR.	0,5457	14 DIST. A3	0,0515
2 ALTITUDINE	0,4956	15 DIST. SS	0,7252
3 N° PIP	0,5256	16 LAUREATI	0,9655
4 SUP. TOT. PIP	0,7205	17 DIPLOMATI	0,9714
5 SUP. MIN. PIP	0,5739	18 POP. RES.	0,9925
6 SUP. MAX. PIP	0,7037	19 POP. ATT.	0,9926
7 N° LOTTI PIP	0,6862	20 POSTE	0,9071
8 N° AI	0,4857	21 BANCHE	0,9779
9 SUP. TOT. AI	0,5662	22 U. L.	0,9909
10 MEDIA INFR.	0,1018	23 ADDETTI UL.	0,9857
11 DIST. CAPOL.	0,3468	24 IMPRESE	0,9912
12 DIST. AEROP.	0,1826	25 ADD. IMPR.	0,9879
13 DIST. FS	0,6086		

Dalla tabella IV.2.4 risulta che per 19 variabili su 25 i primi tre fattori spiegano una quota di variabilità che supera il 50%, e per 10 di loro la percentuale di variabilità spiegata è superiore all'80% (valori in rosso).

Le informazioni riguardanti i legami che intercorrono tra le CP e le variabili sono contenute nella matrice dei coefficienti di correlazione, il cui generico elemento r_{vs} indica il legame esistente tra il fattore v -esimo e la variabile s -esima, tali coefficienti sono anche noti come “pesi fattoriali” e descrivono il cosiddetto “grado di saturazione” di un fattore, cioè la misura in cui una variabile descrive un fattore, e tanto più elevato è tale valore, tanto più il fattore sarà

rappresentato da quella variabile. Overall e Klett (1972) sostengono che le saturazioni più significative sono quelle che superano, in valore assoluto, lo 0,35; in tal caso si dice che il fattore è rappresentato dalla variabile.

Il quadrato del coefficiente di correlazione, r_{vs}^2 , esprime la quota di varianza della s -esima variabile spiegata dal v -esimo fattore, pertanto, più alto sarà tale valore, meglio la variabile sarà spiegata dal fattore.

I coefficienti a_{vs} , ottenibili rapportando i valori r_{vs} alla radice quadrata dei rispettivi autovalori, danno un'indicazione dell'entità e del tipo di legame intercorrente tra le CP e le variabili.

Forniamo di seguito una tabella contenente gli scores dei primi tre fattori:

Tabella IV.2.5

VARIABILE	FATTORE			VARIABILE	FATTORE		
	1	2	3		1	2	3
1 SUP. TERR.	0,6923	-0,137	0,218	14 DIST. A3	-0,17	0,08	0,127
2 ALTITUDINE	-0,227	0,1338	0,6528	15 DIST. SS	-0,187	0,0098	0,8307
3 N° PIP	0,4351	-0,563	-0,141	16 LAUREATI	0,9504	0,2485	0,0235
4 SUP. TOT. PIP	0,7014	-0,445	0,1758	17 DIPLOMATI	0,9565	0,2363	0,0256
5 SUP. MIN. PIP	0,4671	-0,585	-0,114	18 POP. RES.	0,9878	0,115	0,0592
6 SUP. MAX. PIP	0,6296	-0,524	0,1812	19 POP. ATT.	0,9881	0,1143	0,057
7 N° LOTTI PIP	0,7417	-0,359	0,0864	20 POSTE	0,9351	0,1138	0,1405
8 N° AI	0,0681	0,6586	-0,217	21 BANCHE	0,9825	0,0785	0,0805
9 SUP. TOT. AI	0,4251	0,5124	-0,351	22 U. L.	0,9833	0,1498	0,0395
10 MEDIA INFR.	0,0163	-0,319	0,0092	23 ADDETTI UL	0,9748	0,1862	0,0297
11 DIST. CAPOL.	-0,355	0,2559	0,3942	24 IMPRESE	0,984	0,1461	0,0405
12 DIST. AEROP.	-0,242	0,3382	0,0994	25 ADD. IMPR.	0,9769	0,1803	0,0315
13 DIST. ES	-0,273	0,1441	0,7165				

Nella tabella IV.2.5 sono stati messi in evidenza i valori più significativi dei coefficienti di correlazione, che vengono interpretati come degli indicatori del legame che intercorre tra i fattori e le variabili. Notiamo quindi dalla tabella che, per il primo fattore, 18 coefficienti su 25 superano la

soglia di Overall e Klett (valori in blu e rosso), e per cinque variabili tali valori superano lo 0,98 (valori in rosso); come era logico prevedere, per i fattori successivi non abbiamo dei risultati altrettanto ottimali, infatti per il secondo fattore 7 coefficienti su 25 superano lo 0,35 (sempre in valore assoluto) ed i tre migliori sono

compresi tra lo 0,5 e lo 0,6, mentre per il terzo fattore solo cinque fattori su 25 superano la soglia dello 0,35. Per associare un fattore ad una variabile, o meglio, per identificare quelle variabili che sono meglio rappresentate dai fattori, si ricercano i coefficienti più alti (in valore assoluto), che sottintendono un buon livello di

correlazione tra i fattori e le variabili.

Dalla tabella IV.2.5 si evince che il primo fattore risulta legato in modo particolare alle variabili “Popolazione residente”, “Popolazione attiva”, “Banche”, Unità locali” ed “Imprese”, per le quali abbiamo un livello di correlazione che supera lo 0,98.

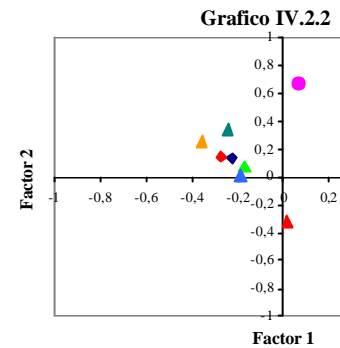
Tale risultato si raggiunge anche interpretando il grafico dei pesi fattoriali, detto anche “cerchio delle correlazioni”, poiché, come è facile intuire, tutti i punti sono compresi in un cerchio di raggio unitario. Prima di proporre tale grafico per il primo ed il secondo fattore, ricordiamo l’interpretazione che bisogna dare ai

punti contenuti
all'interno del
sistema cartesiano:

1) la lunghezza dei vettori che si ottengono congiungendo l'origine degli assi con il punto che identifica la variabile rappresenta la quota di varianza di quella variabile spiegata da fattori considerati, perciò più lungo sarà tale fattore, migliore sarà il livello di rappresentazione del fattore nei confronti della variabile;

2) l'angolo formato dai vettori con ognuno degli assi fattoriali è tanto più piccolo quanto maggiore è il livello di correlazione tra la variabile ed il fattore.

Detto ciò, il
grafico dei pesi
fattoriali per i primi
due fattori assume
la forma:



Seguendo le
indicazioni date è
possibile leggere il
grafico nel seguente
modo:

- Dal punto di vista del primo asse fattoriale, le variabili che presentano la più alta percentuale di variabilità spiegata sono quelle appartenenti al gruppo dei quadratini blu, che risultano infatti avere lunghezza maggiore rispetto alle altre; possiamo inoltre dire che esse hanno con il primo fattore un legame diretto quasi perfetto, in quanto formano con il primo asse un angolo molto piccolo e si avvicinano alla

circonferenza del cerchio (cioè hanno valore prossimo ad uno); come è possibile riscontrare anche nella tabella IV.2.4, tali variabili sono: “Laureati”, “Diplomati”, “Popolazione residente”, “Popolazione attiva”, “Poste”, “Banche”, “Unità locali”, “Addetti UL”, “Imprese”, “Addetti imprese”, che risultano avere un coefficiente di correlazione superiore allo 0,95.

- Dal punto di vista del secondo asse, le variabili che si mettono più in evidenza sono il “N° AI” (puntino fucsia), che formando con tale asse un angolo molto piccolo risulta essere con questo ben correlato in modo diretto, e le variabili “PIP” e “Superficie minima PIP” (puntini blu e lilla), che sono legati al fattore in modo indiretto e raggiungono un buon compromesso tra lunghezza del vettore e scostamento dall’asse.

Come abbiamo
visto le
informazioni che si
colgono dal grafico
IV.2.2 possono
essere lette anche

nella tabella IV.2.4,
ma ci è sembrato
opportuno fornire
un esempio di come
leggere il cerchio
delle correlazioni,
che in genere viene
fornito soltanto per
i primi due fattori.

Dalla tabella
IV.2.4 notiamo
infine che il terzo
fattore presenta
relazioni
significative con le
variabili “Distanza
dalla SS”,
“Distanza dalla FS”
e “Altitudine”.

In conseguenza di questa serie di analisi è possibile “dare un nome” ai fattori, dare cioè un’indicazione degli aspetti che questi sono in grado di cogliere; avremo quindi che il primo fattore risulta essere un buon indicatore della “Capacità di produzione”, che il secondo fattore indica la “Possibilità di insediamenti” (in modo diretto per

quel che riguarda le AI ed in modo inverso per quel che riguarda le aree PIP), e che il terzo fattore può essere visto come un indice sintetico della “Viabilità”.

VI.3 L’individuazione dei gruppi di comuni

Per individuare il numero dei gruppi si è deciso di utilizzare, come suggerito da Punj

G. e Steward D.²¹,
 in sequenza, un
 algoritmo
 gerarchico ed uno
 non gerarchico. Le
 tecniche di
 classificazione
 automatica saranno
 applicate ai
 punteggi fattoriali
 relativi ai primi tre
 fattori, i quali,
 come ampiamente
 spiegato nel
 capitolo riguardante
 l'Analisi in
 Componenti
 Principali, sono i

più rappresentativi
 e consentono di
 analizzare le unità
 senza l'influenza di
 variabili di
 disturbo.

Il risultato più importante al quale si giunge dopo l'applicazione di una classificazione gerarchica è quello relativo alla scelta della partizione ottimale, che sarà il punto di partenza della classificazione non gerarchica, dove, come è noto, è necessario fissare a priori il numero di gruppi in cui suddividere le unità.

L'algoritmo gerarchico scelto per la classificazione delle unità è quello di Ward che, lo ricordiamo, aggrega ad ogni passo i due gruppi che comportano il minor incremento di devianza interna ai gruppi stessi; per calcolare le distanze tra i gruppi viene utilizzato il quadrato della distanza euclidea.

Introducendo queste istruzioni nel package SPSS otteniamo una tabella contenente la successione degli stadi del processo di aggregazione (tabella IV.3.1), che indica per ogni stadio i gruppi che si combinano ed il quadrato della distanza euclidea tra i due; le colonne successive contengono informazioni di riferimento che non riteniamo importanti ai fini dell'analisi (essi riguardano gli stadi futuri in cui è possibile trovare il gruppo appena aggregato).

²¹ "Cluster Analysis in Marketing Research: Review and suggestion for application", Journal of Marketing Research (1983)

Per problemi di spazio, e poiché risultano in effetti i più importanti, riportiamo i valori sopra indicati per gli ultimi dieci passi del processo di aggregazione:

Tabella IV.3.1

Stadio	N [^] gruppi	Cluster accorpati		Distanza Euclidea	Variazione relativa
		Cluster 1	Cluster 2		
70	10	4	15	16,46076	0,1391706
71	9	3	6	18,9432	0,1508098
72	8	2	4	22,06482	0,1647884
73	7	8	40	26,56309	0,2038661
74	6	1	12	35,72273	0,3448261
75	5	2	8	49,92665	0,3976158
76	4	2	3	84,02422	0,6829532
77	3	13	36	118,2363	0,4071695
78	2	1	2	161,0446	0,3620572
79	1	1	13	237	0,4716416

Nelle prime due colonne della tabella IV.3.1 abbiamo gli stadi del processo di aggregazione ed il numero di gruppi di cui è formata la partizione al rispettivo stadio.

La scelta della partizione ottimale si basa principalmente sull'osservazione degli incrementi di distanza tra i gruppi accorpati, o meglio, sugli incrementi relativi; in particolare, si ricerca quello maggiore. Un forte incremento della citata distanza implica l'aggregazione di gruppi che risultano particolarmente distanti rispetto ai gruppi aggregati in precedenza, e ciò porta a privilegiare la partizione precedente a tale salto.

Osservando i dati si nota che il salto più significativo si ha nel passaggio da cinque a quattro gruppi, implicando un valore ottimale (indicativo) del numero di gruppi pari a cinque (riferito allo stadio precedente l'incremento maggiore).

Altri criteri utilizzati per scegliere il numero di gruppi in base al quale formare la partizione ottimale riguardano gli indici R^2 e RMSSTD. Ricordiamo che l'indice R^2 misura la quota di variabilità totale nella matrice dei dati che può essere spiegata dalla partizione considerata, ed assume valori nell'intervallo $[0;1]$, con valori ottimi presso l'unità; l'indice RMSSTD considera invece l'omogeneità del gruppo che si forma nei diversi passi dell'aggregazione ed assume valori ottimali quando tra due aggregazioni successive si evidenziano degli incrementi dell'indice contenuti, poiché un incremento elevato di tale indice implica che al passo cui si riferisce sono stati aggregati due gruppi fortemente eterogenei.

Per avere un altro termine di confronto si è calcolato anche l'indice F_g ; l'indice F_g è funzione del numero g di gruppi di cui è formata la partizione, ed è analizzabile in base all'andamento che assume la curva che lo rappresenta; esso dà indicazioni in merito all'idoneità di una eventuale suddivisione in gruppi dell'insieme dei dati e del numero ottimale di gruppi di cui deve essere formata tale partizione.

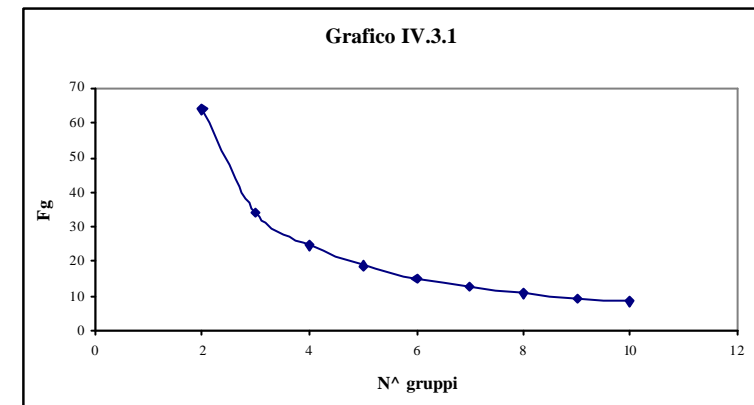
Calcolando questi tre indici per le partizioni ottenute dal passo 70° al 79° (in cui le unità sono raggruppate in un unico cluster), abbiamo ottenuto la tabella IV.3.2:

Tabella IV.3.2

Stadio	N^ gruppi	R ²	Variazione relativa	RMSSTD	Variazione relativa	F _g
70	10	0,52011	-	2619962	-	1,037E+30
71	9	0,51708	-0,00583	2756859	0,05225	1,151E+30
72	8	0,51213	-0,00957	3158094	0,14554	1,298E+30
73	7	0,51117	-0,00187	2353684	-0,25471	1,493E+30
74	6	0,50357	-0,01488	5854853	1,48753	1,769E+30
75	5	0,50071	-0,00567	2857722	-0,51191	2,181E+30
76	4	0,49330	-0,01481	2853403	-0,00151	2,87E+30
77	3	0,46974	-0,04775	6943252	1,43332	4,234E+30
78	2	0,44990	-0,04224	3740943	-0,46121	8,306E+30
79	1	0	-	5089738	0,36055	

Osservando la tabella IV.3.2 si nota che gli incrementi più significativi dell'indice RMSSTD si hanno in due situazioni: passando da sette a sei gruppi e passando da quattro a tre gruppi; in entrambi i casi si registrano variazioni contenute per l'indice R²; ciò significa che ai passi 74 e 77 sono stati aggregati dei gruppi fortemente eterogenei, che portano a privilegiare una partizione che comprende un numero di cluster pari a sette o tre (anche se è più opportuno fare riferimento al primo incremento significativo).

Inserendo in un sistema di assi cartesiani ortogonali i valori assunti dall'indice F_g (sull'asse delle ordinate) ed il numero di gruppi cui questi valori si riferiscono (sull'asse delle ascisse), abbiamo ottenuto il grafico IV.3.1, del quale controlliamo l'andamento:



La curva riportata nel grafico IV.3.1 risulta avere un andamento decrescente, nel senso che i valori di F_g diminuiscono al crescere del numero di gruppi g; ciò significa che i dati presi in considerazione sono adatti ad essere sottoposti ad un processo di clusterizzazione.

Prima di
procedere al passo
successivo

dell'analisi, che
 comporta
 l'applicazione di un
 algoritmo di
 classificazione di
 tipo non gerarchico,
 è bene fare alcune
 osservazioni sui
 risultati ottenuti
 dalla classificazione
 gerarchica.

L'analisi dell'incremento della distanza euclidea dei gruppi che vengono aggregati nelle fasi del processo di agglomerazione (tabella IV.3.1), suggerisce un numero di gruppi pari a 5; l'analisi degli indici R^2 ed RMSSTD (tabella IV.3.2) indicano come più idonea una partizione in sette gruppi, mentre l'andamento della curva F_5 (grafico IV.3.1) conferma la possibilità di una suddivisione in gruppi dei dati. Va inoltre sottolineato che nel processo di aggregazione gerarchica di Ward ci sono due unità che fanno gruppo a sé fino al terzultimo passo, e cioè le unità 13 e 36, riferite ai comuni di Catanzaro e Lamezia Terme, come è possibile osservare nella tabella IV.3.1; è lecito quindi

supporre che questi due comuni si comportino come degli outliers, in quanto caratterizzati da valori anomali nelle variabili rilevate.

Alla luce di queste osservazioni si è deciso di adottare la seguente strategia:

- per la matrice di dati contenente tutte le 80 unità, applicare un algoritmo non gerarchico per un numero di cluster pari a 4, 5, 6 e 7;
- per la matrice di dati contenente 78 unità (cioè le unità originarie meno i comuni di Catanzaro e Lamezia Terme), applicare un algoritmo non gerarchico per un numero di cluster pari a 2, 3, 4 e 5, poiché a questi dovranno essere aggiunti i due formati dagli outliers.

Dopo avere formato le varie partizioni, si procederà al confronto delle stesse per quel che riguarda il grado di omogeneità interna ai gruppi ed il grado di eterogeneità tra i gruppi, individuando quella ottimale.

L'algoritmo non gerarchico scelto è quello di McQueen, noto anche come metodo delle k medie, dove k è il numero di cluster, scelto a priori, in cui si vuole suddividere l'insieme delle unità; tale algoritmo sarà applicato ai punteggi fattoriali dei primi tre fattori.

Abbiamo dunque applicato l'algoritmo di classificazione di McQueen, per $k=4,5,6,7$ alla matrice composta da 80 unità, e per $K=2,3,4,5$ alla matrice composta da 78 unità, riservandoci di considerare gruppi a parte i comuni di Catanzaro e Lamezia Terme che sono stati esclusi dalla matrice iniziale.

Il confronto tra le otto partizioni così ottenute si basa sul grado di omogeneità interna ai gruppi, valutabile in base al valore assunto dalla devianza interna ai gruppi (Dev(W)), ed il grado di eterogeneità tra i gruppi, che si valuta facendo ricorso alla devianza esterna tra i gruppi (Dev(B)), delle quali ricordiamo le formule:

$$Dev(W) = \sum_{g=1}^G \sum_{i=1}^{n_g} \sum_{k=1}^p (x_{ikg} - \bar{x}_{kg})^2$$

$$Dev(B) = \sum_{g=1}^G \sum_{k=1}^p (\bar{x}_{kg} - \bar{x}_k)^2$$

L'ottimalità della partizione dipenderà da questi due valori, cioè maggiore sarà la devianza tra i gruppi, minore sarà la devianza interna ai gruppi, migliore

sarà la partizione ottenuta.

Nella tabella IV.3.3 riportiamo i valori di Dev(B) per le quattro partizioni ottenute applicando l'algoritmo di McQueen alla matrice iniziale composta da ottanta unità:

Tabella IV.3.3 (N=80)

N^ gruppi	Dev(B)	variazione relativa
4	4,55E+16	-
5	2,53E+16	-0,44425406
6	2,54E+16	0,003381308
7	2,59E+16	0,02123179

Tra i valori assunti dalle variazioni relative delle devianze ricerchiamo gli incrementi maggiori per quel che riguarda la devianza tra i gruppi, la quale implica un aumento del grado di separazione degli stessi.

Dalla tabella IV.3.3 notiamo che il maggiore incremento di Dev(B) si ha per K=7.

Nella tabella IV.3.4 riportiamo i valori di Dev(B) per le quattro partizioni ottenute applicando l'algoritmo di McQueen alla matrice di 78 elementi, alle quali sono stati aggiunti i due gruppi composti dai comuni di Catanzaro e Lamezia Terme:

Tabella IV.3.4 (N=78)

N^ gruppi	Dev(B)	variazione relativa
4	2,61E+16	-
5	2,71E+16	0,039774517
6	2,53E+16	-0,06600931
7	3,01E+16	0,189923818

Osservando la tabella IV.3.4 notiamo che anche in questo caso l'incremento maggiore della devianza tra i gruppi si ha per $K=7$; possiamo quindi affermare che la partizione ottimale è quella che è stata ottenuta con tale tecnica, composta da sette gruppi.

Gli ottanta comuni della provincia di Catanzaro vengono così suddivisi in sette gruppi, la cui numerosità è riportata nella tabella IV.3.5:

Tabella IV.3.5

Gruppo	Numerosità
1	20
2	1
3	12
4	31
5	14
6	1
7	1

I gruppi individuati comprendono i comuni:

GRUPPO 1: Albi, Andali, Cardinale, Carlopoli, Cerva, Cicala, Confluenti, Decollatura, Falerna, Gimigliano, Magisano, Martirano Lombardo, Petronà, Santa Caterina dello Ionio, Sellia, Serrastretta, Soveria Mannelli, Taverna, Torre di Ruggiero, Zagarise.

GRUPPO 2: Soverato.

GRUPPO 3: Amato, Botricello, Caraffa di Catanzaro, Cropani, Feroletto Antico, Isca sullo Ionio, Maida, Montauro, San Floro, Simeri Crichi, Sorbo San Basile, Squillace.

GRUPPO 4: Amaroni, Argusto, Belcastro, Cenadi, Centrache, Cortale, Fossato Serralta, Gagliato, Gasperina, Jacurso, Marcedusa, Marcellinara, Martirano, Miglierina, Motta Santa Lucia, Olivadi, Palermiti, Pentone, Petrizzi, Pianopoli, Platania, San Mango d'Aquino, San Pietro apostolo, San Sostene, Sant'Andrea apostolo, San Vito sullo Ionio, Satriano, Settingiano, Soveria Simeri, Staletti, Vallefiorita.

GRUPPO 5: Badolato, Borgia, Chiaravalle Centrale, Curinga, Davoli, Girifalco, Gizzeria, Guardavalle, Montepaone, Nocera Tirinese, San Pietro a Maida, Sellia Marina, Sersale, Tiriolo.

GRUPPO 6: Catanzaro.

GRUPPO 7: Lamezia Terme.

CAPITOLO V: LA CONNOTAZIONE DEI GRUPPI DI COMUNI

V.1 Introduzione

Il risultato ottenuto è una suddivisione in sette gruppi (dei quali tre rappresentati da una sola unità) degli ottanta comuni che compongono la provincia di Catanzaro.

Le tecniche usate hanno consentito di trovare una serie di dimensioni significative che hanno facilitato l'analisi ed hanno creato dei "punti di riferimento" tra le variabili proposte. L'utilizzo di una strategia di raggruppamento ha inoltre permesso di creare nuove "unità" oggetto di studio, i cluster appunto, più idonee ad un'analisi statistico-economica, rispetto alle singole unità di partenza. I cluster individuati risultano caratterizzati da aspetti di ottimalità relativi al grado di coesione interna e separazione esterna, garantiti dalle peculiarità delle tecniche usate. Ci si aspetta dunque che i cluster trovati siano il più possibile distinti tra loro ed abbiano caratteristiche pressoché esclusive.

L'ultimo passo del nostro lavoro consiste quindi nella descrizione dei gruppi, nella ricerca cioè di quelle dimensioni che maggiormente li caratterizzano, in

modo da avere a disposizione utili informazioni per identificare i possibili ruoli che tali zone possono svolgere all'interno di un disegno di sviluppo industriale.

Come è ovvio immaginare, le dimensioni che verranno prese in considerazione dipendono dal tipo di analisi economica proposta, quella cioè relativa ai distretti industriali.

V.2 I fattori utilizzati nell'analisi

Il punto di partenza del nostro lavoro è l'idea che nella zona presa in considerazione possano nascere dei particolari insediamenti

industriali, i distretti appunto; premesso che per un qualsiasi studio che precede un investimento è necessario avere a disposizione, come requisito minimo, una banca dati di tipo socio-economico, riferendoci alla natura del distretto industriale abbiamo pensato fosse particolarmente interessante avere a disposizione una

mappatura socio-economica del territorio, in cui fosse possibile individuare zone (costituite da raggruppamenti di comuni o da singoli comuni di particolari dimensioni) caratterizzati da particolari peculiarità.

Nel primo capitolo, descrivendo i distretti industriali, abbiamo

individuato i fattori che ne favoriscono la nascita, che riguardano il territorio, la popolazione, l'economia e le infrastrutture.

Il territorio e le infrastrutture interessano in modo particolare l'aspetto relativo ai costi di trasporto, elemento fondamentale nei processi di ottimizzazione di un'impresa.

La popolazione, come ampiamente spiegato nel capitolo primo, riveste un ruolo di particolare importanza, in quanto rappresenta innanzi tutto una fonte da cui attingere forza lavoro, ma dalla quale scaturisce anche il bacino d'utenza cui indirizzare la produzione finale.

L'economia deve essere vista

soprattutto come
 capacità del
 territorio di
 generare sviluppo,
 poiché, in zone in
 cui le strutture
 esistenti sono
 carenti, è
 interessante
 monitorarne le
 potenzialità
 economiche.

Il nostro lavoro
 ci consente di
 controllare questi
 aspetti attraverso i
 fattori emersi
 dall'Analisi in
 Componenti

Principali. In
 particolare, i fattori
 denominati
 "Distanze P.I.P.",
 "Distanze A.I." e
 "Altitudine"
 sembrano essere dei
 buoni candidati per
 valutare gli aspetti
 interconnessi al
 problema dei costi
 di trasporto;
 attraverso il fattore
 denominato
 "Popolazione" si
 intende monitorare,
 appunto,
 l'andamento della
 popolazione, ed il

fattore denominato
 “Estensione P.I.P.”
 è stato considerato
 un buon indicatore
 per la valutazione
 di un possibile
 sviluppo economico
 del territorio.

Dopo avere
 attribuito un
 significato a questi
 cinque fattori,
 abbiamo proceduto
 alla descrizione dei
 sette gruppi in
 questione,
 effettuando inoltre
 delle analisi
 comparative tra gli

stessi, finalizzate
 all’individuazione
 di zone del
 territorio preferibili
 rispetto alle altre in
 base ai criteri già
 enunciati.

Grazie al lavoro
 svolto sarà
 possibile
 controllare questi
 aspetti in via
 preliminare
 attraverso i fattori
 emersi dall’Analisi
 in Componenti
 Principali,
 considerando il
 primo fattore come








un indice della capacità di produzione, il secondo della possibilità di insediamenti ed il terzo della viabilità.

Dopo questa descrizione orientativa dei gruppi si passerà ad una descrizione più dettagliata, in cui entreranno in gioco le variabili rilevate ed in particolare quelle che interessano il nostro

tipo di analisi economica.

V.3 Descrizione dei gruppi.

Prima di procedere all'analisi dei gruppi sembra opportuno fornire un supporto grafico necessario a seguire l'analisi degli stessi. Presentiamo quindi di seguito la cartina della provincia di Catanzaro in cui sono stati individuati i sette gruppi di comuni in base a delle colorazioni differenti:

- | | | | |
|---|----------|---|----------|
|  | GRUPPO 1 |  | GRUPPO 5 |
|  | GRUPPO 2 |  | GRUPPO 6 |
|  | GRUPPO 3 |  | GRUPPO 7 |
|  | GRUPPO 4 | | |

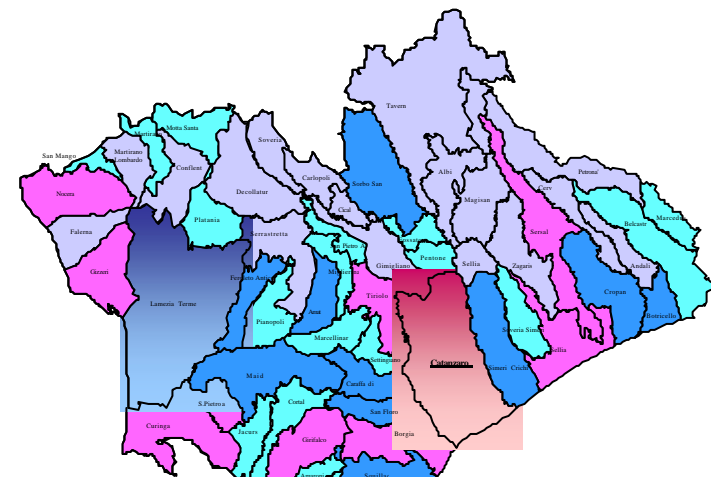
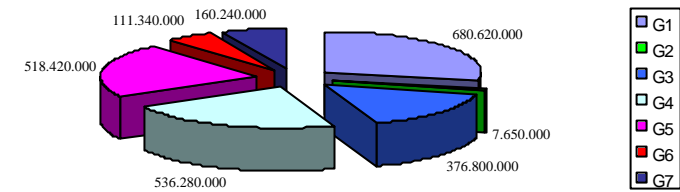


Grafico V.3.1



Forniamo inoltre la porzione di territorio occupata da ogni gruppo, in termini di somma delle superfici di ogni singolo comune facente parte del gruppo:

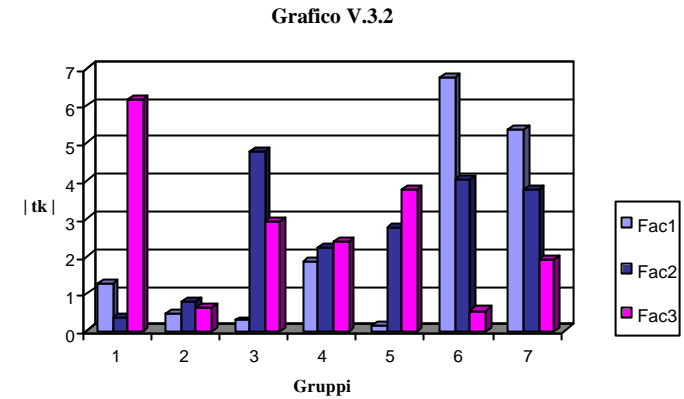
Notiamo innanzitutto che il gruppo che occupa la porzione più grande del territorio è il primo, nonostante non sia il più numeroso (contiene 20 unità contro le 31 del gruppo quattro). In generale è possibile affermare che i gruppi più numerosi occupano porzioni di territorio simili in quanto a superficie, mentre risulta evidente che i comuni che costituiscono i gruppi sei e sette sono caratterizzati da un'estensione territoriale notevole se confrontata con gli altri comuni.

Una prima descrizione dei gruppi proposti viene fatta calcolando l'indice t_k sui punteggi fattoriali dei primi tre fattori. Ricordiamo che tale indice, o meglio il suo valore assoluto, serve ad individuare il fattore, o i fattori, che maggiormente caratterizzano il gruppo.

Abbiamo così costruito il grafico V.3.2 in base alle indicazioni della tabella V.3.1, contenente i valori assoluti dell'indice t_k per i primi tre fattori e per i sette gruppi considerati:

Tabella V.3.1

Gruppo	t1	t2	t3
G1	1,28006	0,34336	6,18172
G2	0,4732	0,7852	0,63203
G3	0,28262	4,76297	2,88634
G4	1,8694	2,24896	2,39585
G5	0,1689	2,7444	3,7525
G6	6,72047	4,02389	0,55009
G7	5,3686	3,75204	1,92611



I valori di t_k più significativi sono quelli che superano, in valore assoluto, il due. In base a tale regola, osservando sia la tabella V.3.1 sia il grafico V.3.2, è possibile notare in particolare che:

- il primo gruppo risulta caratterizzato prevalentemente dal terzo fattore che, lo ricordiamo, rappresenta alcune variabili relative alle distanze ed all'altitudine;
- per il terzo gruppo il fattore più significativo è il secondo, che si riferisce alla presenza di Aree Industriali e di aree P.I.P.;
- il sesto ed il settimo gruppo sono particolarmente legati al primo fattore, anche se non sono da trascurare i valori assunti dal secondo.

Questa prima analisi, che non permette di caratterizzare i gruppi, ha lo scopo di far notare la presenza di gruppi che si distinguono in modo netto dagli altri, in base ad alcune peculiarità che li renderanno o meno preferibili²². L'analisi dettagliata dei gruppi è stata quindi effettuata considerando tutte le variabili.

Proponiamo una serie di tabelle in cui vengono riportati i valori delle medie di gruppo di ogni variabile, che verranno messe a confronto con le medie generali, sia

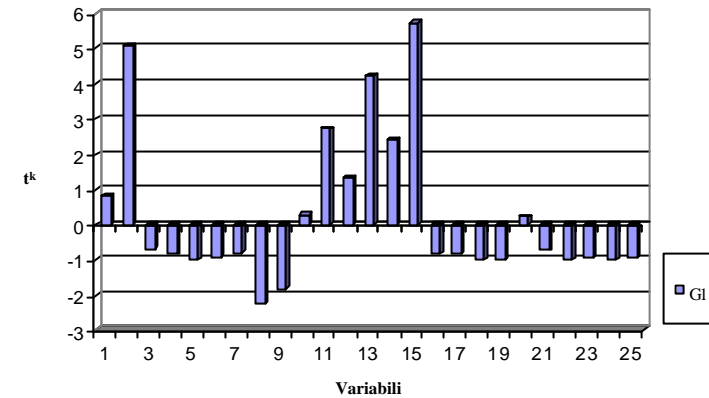
²² Ricordiamo che, nonostante siano stati sottolineati soltanto alcuni fattori, sono significativi tutti i valori che superano il due.

in modo diretto sia in base ai valori assunti da t_k , che indicano quali differenze tra le medie di gruppo e le medie generali siano statisticamente significative.
Le tabelle così costruite saranno accompagnate dai grafici relativi ai valori dei t_k , in modo da avere un ulteriore ausilio grafico di analisi.
Il secondo, il sesto ed il settimo gruppo, poiché sono costituiti da un solo comune, verranno analizzati in seguito.

Tabella V.3.2 (G1, N=20)

VARIABILE	Media gr.	Media gen.	t_k	VARIABILE	Media gr.	Media gen.	t_k
1 SUP. TERR.	34.031.000	29.891.875	0,83	14 DIST. A3	52	42	2,41
2 ALTITUDINE	669	476	5,08	15 DIST. SS	26	15	5,72
3 N° PIP	0,35	0,45	-0,69	16 LAUREATI	32	148	-0,83
4 SUP. TOT. PIP	11.370	102.739	-0,86	17 DIPLOMATI	274	755	-0,84
5 SUP. MIN. PIP	297	461	-0,99	18 POP. RES.	2.268	4.796	-0,99
6 SUP. MAX. PIP	865	3.921	-0,90	19 POP. ATT.	1.502	3.317	-1,00
7 N° LOTTI PIP	3,85	6,79	-0,80	20 POSTE	1,90	1,80	0,24
8 N° AI	0,45	0,73	-2,22	21 BANCHE	0,70	1,26	-0,75
9 SUP. TOT. AI	24.557	93.205	-1,83	22 U. L.	94	231	-0,99
10 MEDIA INFR.	2,05	1,92	0,30	23 ADDETTI UL	161	546	-0,94
11 DIST. CAPOL.	42	35	2,72	24 IMPRESE	89	216	-0,99
12 DIST. AEROP.	54	48	1,33	25 ADD. IMPR.	149	477	-0,94
13 DIST. FS	25	17	4,23				

Grafico V.3.3

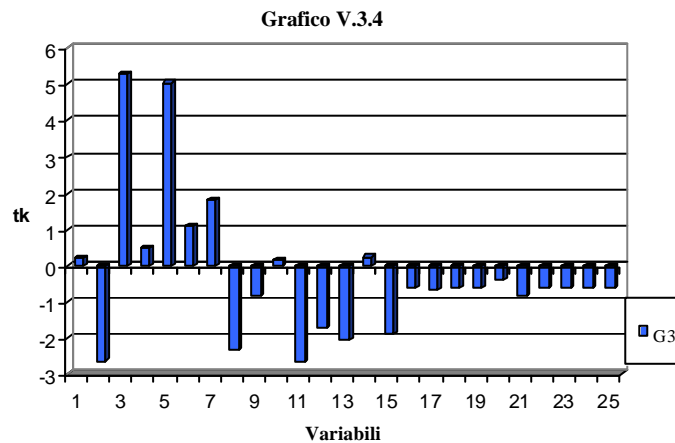


Il primo gruppo, risulta avere i valori più significativi dell'indice t_k per le variabili "Distanza tra il comune e la strada statale più vicina" (5,72), "Altitudine" (5,08), "Distanza tra il comune e la stazione ferroviaria più vicina" (4,23), "Distanza tra il comune ed il Capoluogo" (2,72), "Distanza tra il comune e l'autostrada A3" (2,41) e "Numero di Aree Industriali" (-2,22).

Confrontando le medie di gruppo con le medie generali di tali variabili, notiamo che per il gruppo considerato l'altitudine supera la media, così come i valori relativi alle distanze, mentre il numero di Aree Industriali è inferiore alla media generale; questo ci fa concludere che i valori più significativi di questo gruppo sottintendono aspetti negativi, sia per quel che riguarda le vie di comunicazione (variabili 2, 11, 13, 14 e 15), sia per quel che riguarda la presenza di aree industriali (variabile 8). Questo gruppo di comuni sembra dunque non risultare particolarmente adatto ad ospitare insediamenti industriali, in quanto non offre vantaggi né dal punto di vista economico, in relazione alla presenza di aree P.I.P. (che sono al di sotto della media), né dal punto di vista territoriale, risultando più distante dalle vie di comunicazione rispetto agli altri.

Tabella V.3.3 (G3, N=12)

VARIABILE	Media gr.	Media gen.	t _k	VARIABILE	Media gr.	Media gen.	t _k
1 SUP. TERR.	31.400.000	29.891.875	0,22	14 DIST. A3	43	42	0,24
2 ALTTUDINE	339	476	-2,62	15 DIST. SS	10	15	-1,88
3 N° PIP	1,5	0,45	5,27	16 LAUREATI	29	148	-0,62
4 SUP. TOT. PIP	176.821	102.739	0,51	17 DIPLOMATI	249	755	-0,64
5 SUP. MIN. PIP	1.602	461	5,04	18 POP.RES.	2.562	4.796	-0,64
6 SUP. MAX. PIP	8.974	3.921	1,08	19 POP.ATT.	1.746	3.317	-0,63
7 N° LOTTI PIP	16	6,79	1,83	20 POSTE	1,58	1,80	-0,37
8 N° AI	0,33	0,73	-2,30	21 BANCHE	0,42	1,26	-0,82
9 SUP. TOT. AI	49.167	93.205	-0,85	22 U. L.	112	231	-0,63
10 MEDIA INFR.	2,01	1,92	0,15	23 ADDETTI UL	206	546	-0,60
11 DIST. CAPOL.	25	35	-2,67	24 IMPRESE	106	216	-0,62
12 DIST. AEROP.	36	48	-1,73	25 ADD. IMPR.	185	477	-0,61
13 DIST. FS	12	17	-2,05				

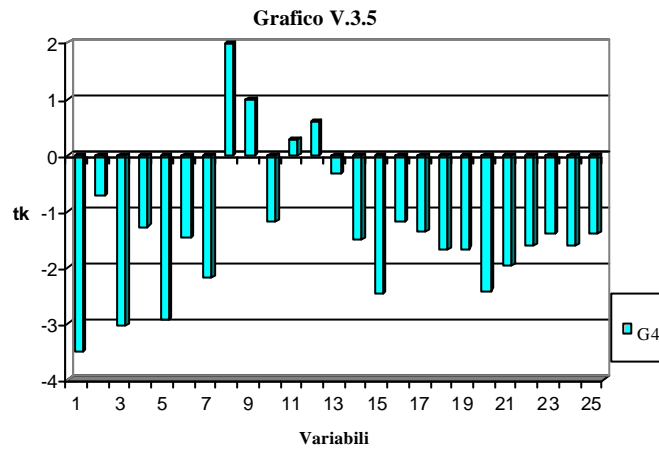


Per il terzo gruppo i valori di t_k più significativi riguardano le variabili “Numero di aree P.I.P.” (5,27), “Superficie minima dei lotti P.I.P. presenti sul territorio” (5,04), “Distanza tra il comune ed il Capoluogo” (-2,67), “Altitudine” (-2,62), “Numero di aree industriali” (-2,3), “Distanza tra il comune e la stazione ferroviaria più vicina” (-2,05). Effettuando i confronti tra le due medie, quella generale e quella di gruppo,

notiamo che il numero di aree P.I.P. presenti sul territorio (variabile 3) e la superficie minima dei lotti P.I.P. disponibili (variabile 5) superano la media, e questo risulta essere un vantaggio oggettivo dei comuni facenti parte del gruppo, mentre riscontriamo un aspetto negativo in relazione alla presenza di Aree Industriali sul territorio, il cui numero è inferiore alla media generale; sono inoltre da annoverare tra gli aspetti positivi di questo gruppo i valori assunti dalle variabili 2, 11, e 13, che si riferiscono all’altitudine ed alle misure di distanza; essi sono infatti al di sotto dei valori assunti dalle rispettive medie generali, inducendo un vantaggio dal punto di vista delle vie di comunicazione. Con riferimento a tale gruppo possiamo quindi affermare che esso presenta una certa disponibilità di zone adatte ad ospitare insediamenti industriali, che non sono lontani né dal capoluogo né dalle vie di comunicazione.

Tabella V.3.4 (G4, N=31)

VARIABILE	Media gr.	Media gen.	t _k	VARIABILE	Media gr.	Media gen.	t _k
1 SUP. TERR.	17.299.355	29.891.875	-3,50	14 DIST. A3	37	42	-1,48
2 ALTTUDINE	456	476	-0,71	15 DIST. SS	12	15	-2,46
3 N° PIP	0,13	0,45	-3,05	16 LAUREATI	28	148	-1,18
4 SUP. TOT. PIP	5.016	102.739	-1,27	17 DIPLOMATI	189	755	-1,36
5 SUP. MIN. PIP	109	461	-2,94	18 POP.RES.	1.676	4.796	-1,68
6 SUP. MAX. PIP	334	3.921	-1,45	19 POP.ATT.	1.116	3.317	-1,66
7 N° LOTTI PIP	0,97	6,79	-2,18	20 POSTE	1,06	1,80	-2,41
8 N° AI	0,90	0,73	1,98	21 BANCHE	0,19	1,26	-1,95
9 SUP. TOT. AI	120.486	93.205	1,00	22 U. L.	69	231	-1,61
10 MEDIA INFR.	1,54	1,92	-1,18	23 ADDETTI UL	135	546	-1,38
11 DIST. CAPOL.	36	35	0,26	24 IMPRESE	65	216	-1,62
12 DIST. AEROP.	50	48	0,61	25 ADD. IMPR.	127	477	-1,38
13 DIST. FS	17	17	-0,33				



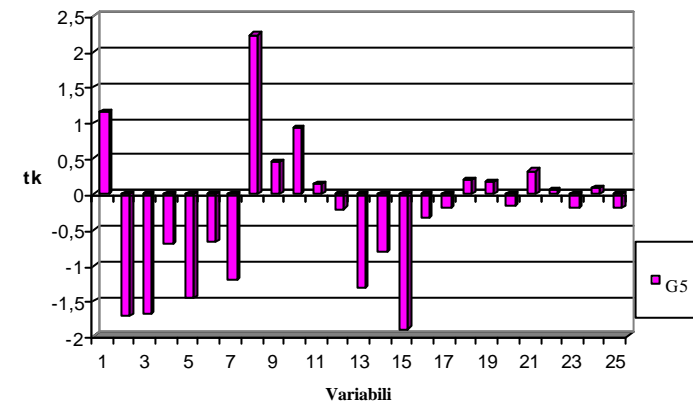
Il quarto gruppo mette in evidenza, in base ai valori assunti dall'indice t_k , le variabili "Superficie territoriale" (-3,5), "Altitudine" (-3,05), " Superficie minima dei lotti P.I.P." (-2,94), "Distanza tra il comune e la strada statale più vicina" (-2,46), "Numero di uffici postali" (-2,41), "Numero di lotti P.I.P. disponibili" (-2,18). Il valore assunto dalla variabile relativa alla superficie territoriale (variabile 1), essendo inferiore alla media, presuppone che i comuni facenti parte del gruppo non abbiano un'elevata estensione da mettere a disposizione di eventuali insediamenti; il gruppo ha però dei buoni valori per quel che riguarda l'altitudine e la distanza dalla strada statale (variabili 2 e 15), che sono al di sotto dei valori medi; le altre variabili evidenziate (variabili 5, 7 e 20) sottintendono degli aspetti negativi perché assumono dei valori al di sotto della media generale. Si tratta dunque di un gruppo composto da comuni piccoli, che presenta dei vantaggi territoriali, ma non risulta adatto dal punto di vista economico ad ospitare degli insediamenti.

Tabella V.3.5 (G5, N=14)

VARIABILE	Media gr.	Media gen.	t_k	VARIABILE	Media gr.	Media gen.	t_k		
1	SUP. TERR.	37.030.000	29.891.875	1,15	14	DIST. A3	38	42	-0,81

2	ALTITUDINE	394	476	-1,71	15	DIST. SS	11	15	-1,89
3	N° PIP	0,14	0,45	-1,69	16	LAUREATI	89	148	-0,34
4	SUP. TOT. PIP	10.000	102.739	-0,70	17	DIPLOMATI	623	755	-0,18
5	SUP. MIN. PIP	157	461	-1,47	18	POP. RES.	5.433	4.796	0,20
6	SUP. MAX. PIP	1.107	3.921	-0,66	19	POP. ATT.	3.708	3.317	0,17
7	N° LOTTI PIP	1,21	6,79	-1,21	20	POSTE	1,71	1,80	-0,16
8	N° AI	1,07	0,73	2,23	21	BANCHE	1,57	1,26	0,33
9	SUP. TOT. AI	113.921	93.205	0,44	22	U. L.	242	231	0,06
10	MEDIA INFR.	2,43	1,92	0,93	23	ADDETTI UL	452	546	-0,18
11	DIST. CAPOL.	36	35	0,14	24	IMPRESE	229	216	0,08
12	DIST. AEROP.	46	48	-0,21	25	ADD. IMPR.	393	477	-0,19
13	DIST. FS	14	17	-1,32					

Grafico V.3.6



Il quinto gruppo offre un unico valore significativo dell'indice t_k , quello relativo alla variabile "Numero di Aree Industriali" (2,23); il confronto tra la media di gruppo e la media generale fa notare che in tale gruppo la presenza di aree industriali supera la media.

Per il secondo, sesto e settimo gruppo, composti da un solo comune, riportiamo una tabella riassuntiva delle variabili, il cui confronto con le medie generali permetterà di evidenziare perché tali comuni costituiscono gruppi a sé.

Tabella V.3.6

VARIABILE	G 2	G 6	G 7	Media gen.
1 SUP. TERR.	7.650.000	111.340.000	160.240.000	29.891.875
2 ALTITUDINE	400	343	210	476
3 N° PIP	1,00	2,00	2,00	0,45
4 SUP. TOT. PIP	150.000	584.180	4.840.170	102.739
5 SUP. MIN. PIP	1.000	2.100	3.000	461
6 SUP. MAX. PIP	5.000	7.800	150.000	3.921
7 N° LOTTI PIP	45	68	114	6,79
8 N° AI	0,00	2,00	0,00	0,73
9 SUP. TOT. AI	0	1045287	0	93204,829
10 MEDIA INFR.	3	0	3,5	1,9
11 DIST. CAPOI.	29	0	25	35
12 DIST. AEROP.	64	35,6	4	48
13 DIST. FS	1	4,2	9	17
14 DIST. A3	63,8	33,3	8,5	42
15 DIST. SS	43	3,8	9	15
16 LAUREATI	171	5.924	2.638	148
17 DIPLOMATI	1.877	24.290	11.218	755
18 POP. RES.	10.805	96.975	71.754	4.796
19 POP ATT.	7.820	69.040	51.004	3.317
20 POSTE	2	17	11	1,80
21 BANCHE	5	27	22	1,26
22 U. L.	762	5.478	3.516	231
23 ADDETTI UL	1.452	16.628	9.383	546
24 IMPRESE	690	5.077	3.302	216
25 ADD. IMPR.	1.273	14.087	8.143	477

Osservando i valori della tabella V.3.6 notiamo che il secondo gruppo, costituito dal comune di Soverato, non presenta variabili che si evidenziano in particolare rispetto alle altre. Per poter quindi individuare qualche peculiarità ricorriamo all'ausilio di un'ulteriore tabella, la V.3.7., in cui riportiamo degli indici relativi ai tre gruppi in questione ed a tutte le unità.

Tabella V.3.7

	G2	G6	G7	Generale
Pop. Att./Pop. Tot.	0,723	0,712	0,711	0,691
Laur.+Dipl./Pop. Tot.	0,1895	0,3115	0,1931	0,1883
Add. UL/Pop. Tot.	0,134	0,171	0,130	0,113
S.Tot.PIP/S.Terr.	0,0196	0,0052	0,0302	0,0034

Dalla tabella V.3.7 notiamo che il gruppo due si evidenzia rispetto agli altri per quel che concerne il rapporto tra popolazione attiva e popolazione totale, cioè esso ha la porzione di popolazione attiva più elevata rispetto alla popolazione totale.

Il gruppo sei, che rappresenta il comune di Catanzaro, ha valori ottimi per quel che riguarda la popolazione e la presenza di unità economiche (variabili dalla 16 alla 25), e buoni valori per le altre variabili. Ciò si può notare sia dalla tabella V.3.6 che dalla tabella V.3.7, dalla quale si evince che questo gruppo ha la porzione più elevata di laureati e diplomati, e di addetti nelle unità locali rispetto alla popolazione totale.

L'ultimo gruppo, il settimo, rappresenta il comune di Lamezia Terme; così come per Catanzaro, troviamo dei buoni valori per le variabili relative alla popolazione ed alle unità economiche, ma tale comune risulta essere particolarmente importante per la presenza di aree P.I.P. (variabili dalla 3 alla 7), come si nota in particolare osservando nella tabella V.3.7 il rapporto tra superficie totale delle aree P.I.P. e superficie territoriale.

Questo primo tipo di analisi condotta gruppo per gruppo ci ha permesso di individuare le variabili più significative per ognuno di questi; ci consente anche di affermare che tra questi gruppi due in particolare possono essere presi in considerazione in un ambito di indagine economica: il sesto ed il settimo.

Come sappiamo questi due gruppi non si riferiscono ad aggregati di comuni, e sembrerebbe quindi lecito escluderli da un'analisi finalizzata alla ricerca delle situazioni ottimali per far nascere un distretto industriale, che come è noto richiede una certa superficie territoriale a disposizione (anche se, come risulta dai confronti proposti, le superfici territoriali dei due comuni sono notevoli); è doveroso però ricordare che un ruolo importante per la nascita di un insediamento industriale, e quindi anche di un distretto, ricoprono i cosiddetti "poli" di sviluppo, zone limitate con particolari potenzialità economiche, quali sono appunto i comuni di Catanzaro e Lamezia Terme.

Dopo quest'analisi globale dei sette gruppi considerati, abbiamo pensato alla creazione di un ausilio sintetico da mettere a disposizione di chiunque volesse avere delle indicazioni immediate sul territorio in questione. A tal proposito abbiamo costruito una serie di grafici che mettono a confronto i sette gruppi in riferimento alle seguenti variabili (che sono state considerate le più indicative in relazione all'analisi economica eseguita):

- Altitudine

- Distanze²³

Queste due variabili saranno importanti per capire se il gruppo è adatto ad un insediamento industriale dal punto di vista territoriale.

- Presenza P.I.P.²⁴
- Presenza A. I.²⁵

Questi due indicatori serviranno a valutare la capacità del territorio di ospitare zone industriali.

- Popolazione
- Numero di Unità Locali

Le ultime due variabili forniscono indicazioni sulle potenzialità già esistenti.

Grafico V.3.7

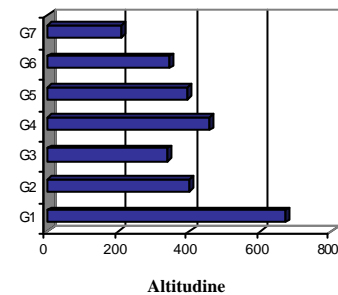


Tabella V.3.8

Gruppo	ALTITUDINE
G1	669
G2	400
G3	339
G4	456
G5	394
G6	343
G7	210

²³ Questa variabile è stata costruita a posteriori effettuando una media delle cinque variabili originarie riferite a misure di distanza.

²⁴ Vengono valutate le variabili "Numero di aree P.I.P." e "Numero di lotti P.I.P.".

²⁵ Vengono valutate le variabili "Numero di A. I." e "Superficie totale di A. I.".

Grafico V.3.8

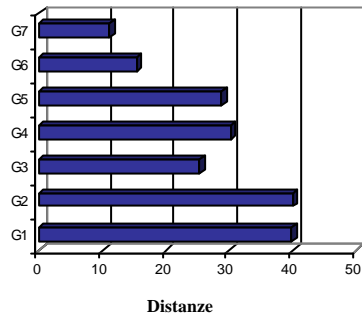


Tabella V.3.9

Gruppo	DISTANZE
G1	39,94
G2	40,16
G3	25,34
G4	30,25
G5	28,87
G6	15,38
G7	11,10

Grafico V.3.9

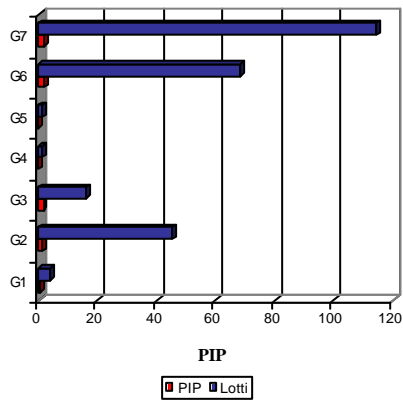


Tabella V.3.10

Gruppo	N° P.I.P.	LOTTI P.I.P.
G1	0,35	3,85
G2	1,00	45,00
G3	1,50	16,00
G4	0,13	0,97
G5	0,14	1,21
G6	2,00	68,00
G7	2,00	114,00

Grafico V.3.10

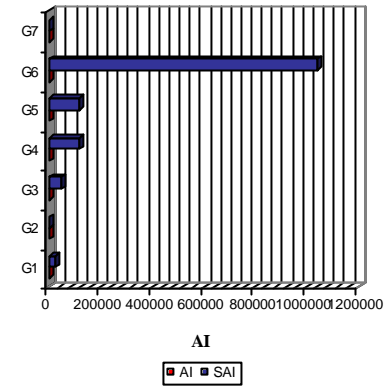


Tabella V.3.11

Gruppo	N° AI	SUP.TOT.AI
G1	0,45	24.557,40
G2	0,00	0,00
G3	0,33	49.166,67
G4	0,90	120.485,63
G5	1,07	113.921,19
G6	2,00	1.045.287,00
G7	0,00	0,00

Grafico V.3.11

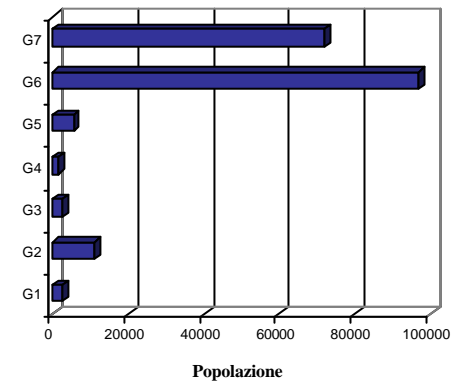


Tabella V.3.12

Gruppo	POPOLAZIONE
G1	2.268
G2	10.805
G3	2.562
G4	1.676
G5	5.433
G6	96.975
G7	71.754

Grafico V.3.12

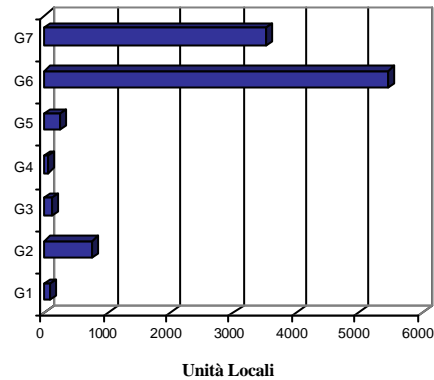


Tabella V.3.13

Gruppo	UNITA' LOCALI
G1	94
G2	762
G3	112
G4	69
G5	242
G6	5.478
G7	3.516

Osservando i grafici proposti abbiamo costruito la tabella V.3.14, nella quale vengono riportati dei giudizi sintetici sui gruppi in riferimento alle variabili prese in considerazione.

In relazione al grafico V.3.7 (Altitudine), ed alla rispettiva tabella V.3.8, consideriamo Ottimo un livello di altitudine inferiore a 250 metri, Buono inferiore a 380 metri, Discreto inferiore a 500 metri, Pessimo superiore a quest'ultimo livello. Forniamo di seguito la chiave di lettura di tutte le altre tabelle (dalla V.3.8 alla V.3.13):

	O	B	D	P
ALTITUDINE	[0 - 250]	(250 - 380]	(380 - 500]	A >500
DISTANZE	[0 - 12]	(12 - 26]	(26 - 38]	D >38
PRESENZA PIP	PIP >100	[100 - 40)	[40 - 10)	[10 - 0]
PRESENZA AI	AI >1000.000	[100.0000 - 500.000)	[500.000 - 20.000)	[20.000 - 0]
POPOLAZIONE	P >90.000	[90.000 - 60.000)	[60.000 - 10.000)	[10.000 - 0]

NUMERO U.L. UL >5.000 | [5.000 - 1.000) | [1.000 - 500) | [500 - 0] |

In base a questi criteri è stata costruita la tabella V.3.14:

Tabella V.3.14

Variabile	GRUPPI						
	G1	G2	G3	G4	G5	G6	G7
ALTITUDINE	P	D	B	D	D	B	O
DISTANZE	P	P	B	D	D	O	O
PRESENZA PIP	P	B	D	P	P	B	O
PRESENZA AI	D	P	D	D	D	O	P
POPOLAZIONE	P	D	P	P	D	O	B
UNITA' LOCALI	P	D	P	P	P	O	B

O = Ottimo	B = Buono
P = Pessimo	D = Discreto

Supponendo di attribuire ad ogni O un punteggio pari a dieci, ad ogni B sei, ad ogni D tre e ad ogni P uno, tra i sette gruppi si delinea la seguente classifica:

Gruppo	Punteggio
G6	52
G7	43
G3	20
G2	17
G5	14
G4	12
G1	8

dalla quale si evince che i gruppi più importanti sono il sesto, il settimo ed il terzo.

V.4 Conclusioni

Il lavoro che è stato svolto ha mosso i suoi passi iniziali in un ambito di osservazione del territorio. Grazie ai dati a disposizione è stato possibile formulare delle ipotesi sul tipo di analisi in relazione alle caratteristiche del territorio stesso; si è puntata quindi l'attenzione su quella particolare forma di insediamento industriale che sono i distretti e si sono ricercate le tecniche statistiche più idonee ad essere applicate ai dati; si è richiesto inoltre che i risultati delle stesse fossero utilizzabili per l'ipotesi economica avanzata, fossero cioè capaci di mettere in evidenza quegli aspetti del territorio necessari ad analizzare lo stesso (popolazione, economia, viabilità, ecc.).

Grazie quindi all'applicazione di queste tecniche statistiche, che, lo ricordiamo, sono l'Analisi in Componenti Principali e la Cluster Analysis, il territorio relativo alla provincia di Catanzaro, comprendente ottanta comuni, è stato suddiviso in sette gruppi, che rispondono a criteri di ottimalità garantiti dalle tecniche usate per costruirli. Questi gruppi sono stati poi analizzati in modo da individuare l'importanza che essi ricoprono sul territorio e per fornire, a chiunque fosse interessato, uno strumento di giudizio, che può ovviamente avere delle piccole lacune di arbitrarietà, ma che ha il pregio di essere stato costruito per mezzo di tecniche il più possibile oggettive.

Come analisi finale del lavoro svolto forniamo delle valutazioni sui gruppi individuati.

I “gruppi” di gran lunga più importanti sono il sesto ed il settimo, cioè i comuni di Catanzaro e Lamezia Terme, che presentano valori medio-alti di tutte le variabili di interesse. La distinzione che può essere fatta tra i due riguarda la presenza di Aree Industriali e di Aree P.I.P.; il comune di Catanzaro presenta valori ottimi in quanto ad insediamenti industriali e buoni per quel che riguarda le Aree P.I.P., mentre il comune di Lamezia Terme non ha Aree Industriali, ma ha valori ottimali riferiti alle Aree P.I.P.; ciò significa che Lamezia Terme presenta un vantaggio dal punto di vista operativo²⁶. Possiamo quindi affermare che questi due comuni possono essere visti come dei “poli di sviluppo” attorno ai quali costruire delle realtà economiche con buone potenzialità di crescita.

Tra gli altri gruppi quello che si fa notare maggiormente è il terzo, che non ha valori desiderabili per quel che riguarda la popolazione e la presenza in loco di unità economiche, cioè di “forze trainanti”, ma presenta delle buone caratteristiche territoriali, poiché ha dei buoni valori delle variabili relative alle distanze dalle principali vie di comunicazione, ed un valore più che buono dell'altitudine; non va poi dimenticata la presenza sul territorio di aree P.I.P., che supera la media in modo notevole. Questo gruppo si evidenzia quindi per le sue

²⁶ Ricordiamo che i Piani per gli Insediamenti Produttivi sono dei piani particolareggiati, quindi attuativi del Piano Regolatore Generale, che non hanno bisogno di ulteriori approvazioni.

“potenzialità” adatte ad essere sfruttate dal punto di vista economico. Va inoltre ricordato che la maggior parte dei comuni di questo gruppo sono prossimi ai due “poli”, cioè ai comuni di Catanzaro e Lamezia Terme, che possono quindi essere visti come dei motori di sviluppo anche per il gruppo considerato.

Tra i gruppi restanti, il primo non risulta adatto ad ospitare insediamenti industriali in quanto risulta caratterizzato da valori negativi sia in relazione alla presenza di aree P.I.P., sia in relazione alle distanze dalle principali vie di comunicazione.

Il secondo gruppo, cioè il comune di Soverato, ha nel complesso dei buoni valori, e, come abbiamo già visto, presenta dei vantaggi dal punto di vista della forza lavoro, ma insufficienti affinché possa essere considerato un “polo di sviluppo”.

Nel quarto gruppo si evidenziano dei vantaggi territoriali, che rappresentano delle buone caratteristiche da mettere a disposizione di un progetto di sviluppo, ma si nota allo stesso tempo la carenza di altre peculiarità prettamente economiche.

Il quinto gruppo, infine, presenta dei valori discreti per quel che riguarda la collocazione territoriale e la presenza di Aree Industriali, ma nel complesso è possibile attribuirgli un giudizio mediocre, che certamente non lo rende desiderabile dal punto di vista economico.

APPENDICE A: Misure di distanza

La “prossimità” di due unità statistiche è misurata, nel caso di caratteri quantitativi, per mezzo della loro “distanza”.

Definizione di distanza¹

Si definisce distanza (o metrica) tra due punti, corrispondenti ai vettori $x, y \in \mathbb{R}^p$, una funzione $d(x,y)$ che gode delle proprietà:

1) **non negatività:**

$$d(x,y) \geq 0 \quad \forall x, y \in \mathbb{R}^p$$

2) **identità:**

$$d(x,y) = 0 \quad \Leftrightarrow \quad x = y$$

3) **simmetria:**

$$d(x,y) = d(y,x) \quad \forall x, y \in \mathbb{R}^p$$

4) **disuguaglianza triangolare :**

$$d(x,y) \leq d(x,z) + d(y,z) \quad \forall x, y, z \in \mathbb{R}^p$$

Queste quattro proprietà non sono tra loro indipendenti, ma si può dimostrare che se valgono le proprietà di identità e di

¹ Zani, “Analisi dei dati statistici”, 1999.

disuguaglianza triangolare, valgono anche quelle di non negatività e simmetria.

Uno spazio con riferimento al quale si sia definita una distanza è detto “spazio metrico”.

Nel caso di una matrice di dati quantitativi, la distanza tra due generiche unità statistiche è calcolata sui vettori riga x_i e x_j , e viene indicata nel modo seguente:

$$d(x_i, x_j) = d_{ij}$$

Tra i vari tipi di distanza le più usate sono:

Distanza euclidea

Si dice distanza euclidea tra due unità statistiche i e j la norma della differenza tra i rispettivi vettori:

$$d_{ij} = \|x_i - x_j\| = \left[\sum_{s=1, p} (x_{is} - x_{js})^2 \right]^{1/2}$$

Da un punto di vista geometrico, nel caso particolare di due variabili, la distanza euclidea è rappresentata dal segmento che unisce i due punti sul piano cartesiano, essendo la radice quadrata della somma dei quadrati costruiti sui cateti. Risulta essere

influenzata dalle differenze elevate tra i valori, poiché è funzione dei quadrati delle stesse; è però invariante per trasformazioni ortogonali (rotazioni) delle variabili.

Distanza della città a blocchi

Si dice distanza della città a blocchi tra due unità i e j l'espressione:

$$d_{ij} = \sum_{s=1, p} |x_{is} - x_{js}|$$

Su un piano cartesiano bidimensionale, essa è rappresentata dalla somma dei cateti del triangolo rettangolo in cui i punti-unità si trovano ai due estremi dell'ipotenusa. A differenza della metrica euclidea, non è influenzata dalle elevate differenze tra i valori, in quanto effettua su di esse una compensazione.

Distanza di Minkowski

Si dice distanza di Minkowski di ordine k tra le unità i e j l'espressione:

$$d_{ij} = \left[\sum_{s=1, p} |x_{is} - x_{js}|^k \right]^{1/k}$$

La distanza di Minkowski costituisce una generalizzazione delle due metriche precedenti.

Distanza di Lagrange-Tchebychev

Si dice **distanza di Lagrange-Tchebychev tra due unità i e j**

l'espressione:

$$d_{ij} = \max_s |x_{is} - x_{js}|$$

E' un altro caso particolare della metrica di Minkowski, ottenuta facendo il limite, per k tendente ad infinito, dell'espressione generale d_{ij}^k .

Distanza di Camberra

Si dice distanza di Camberra tra le unità i e j l'espressione:

$$d_{ij} = \sum_{s=1,p} [|x_{is} - x_{js}| / (x_{is} + x_{js})]$$

Questa metrica fu introdotta da Lance e Williams nel 1966 e non rientra nella classe delle metriche di Minkowski. Si può interpretare tale distanza come un caso particolare della distanza della città a blocchi, e risulta essere una distanza soltanto nel caso in cui assumano valori positivi (in caso contrario non può essere sempre soddisfatta la disuguaglianza triangolare). Può essere applicata anche a variabili differenti senza ricorrere alla standardizzazione, ed è poco sensibile alla presenza di outliers.

Distanza del χ^2

Si consideri un collettivo di n unità sul quale siano state rilevate p variabili, e si supponga che tale collettivo sia stato suddiviso in H gruppi.

Si dice distanza del χ^2 tra due gruppi h e k l'espressione:

$$d_{hk}^2 = \sum_{s=1,p} [(n_{hs}/n_h - n_{ks}/n_k)^2 / (n/n_s)]$$

E' una distanza euclidea ponderata che gode delle seguenti proprietà: invarianza rispetto ai criteri di codifica delle entità; equivalenza distributiva; tiene conto anche di piccole variazioni negli scarti tra entità.

Definizione di indice di distanza²⁷

“Si dice indice di distanza tra due vettori \mathbf{x} e \mathbf{y} ? R^p una funzione che soddisfa le proprietà di non negatività, identità e simmetria”.

Un importante indice di distanza è dato dal quadrato della distanza euclidea:

$$d^2(x,y) = \|\mathbf{x}_i - \mathbf{x}_j\|^2$$

che gode dell'importante proprietà dell'additività: la somma del valore dell'indice calcolato su due sottoinsiemi p_1 e p_2 (con $p_1 + p_2 = p$) è uguale al valore dell'indice calcolato su tutte le p variabili.

²⁷ Zani, “Analisi dei dati statistici”, 1999.

APPENDICE B: Descrizione delle variabili

Variabile 1: Superficie totale del Comune

Misura l'effettiva estensione dell'area comunale.

E' di tipo quantitativo continuo. L'unità di misura è il metro quadrato.

Fonte: ISTAT.

Variabile 2: Altitudine

Altitudine: distanza verticale di un punto dal livello medio del mare.

Misura l'altitudine dei vari Comuni.

E' di tipo quantitativo continuo. L'unità di misura è il metro.

Fonte: PAGINE BLU.

Variabile 3: Numero di Aree PIP

Area PIP: area regolata dai cosiddetti "Piani per gli Insediamenti Produttivi", reperibile nelle zone destinate agli insediamenti produttivi dai PRG (Piani Regolatori Generali) o dai PdF (Piani di Fabbricazione) vigenti; il Comune utilizza tali aree per la realizzazione di impianti produttivi di carattere industriale, artigianale, commerciale o turistico. Il piano approvato ha efficacia

per dieci anni ed ha valore di Piano Particolareggiato (il Piano Particolareggiato è il mezzo di attuazione dei PRG, al quale è subordinato).

E' di tipo quantitativo discreto.

Fonte: INDAGINE T.E.A.

Variabile 4: Superficie totale di aree PIP presente nel Comune

Indica il totale di superficie adibita ad area PIP presente nel Comune.

E' di tipo quantitativo continuo. L'unità di misura è il metro quadrato.

Fonte: INDAGINE T.E.A.

Variabile 5: Superficie minima dei lotti delle aree PIP presenti nel Comune

Lotto: ciascuna delle parti in cui viene suddiviso un terreno per essere venduta, specialmente come area fabbricabile.

Indica la superficie del lotto più piccolo disponibile tra le eventuali aree PIP del Comune. E' di tipo quantitativo continuo. L'unità di misura è il metro quadrato.

Fonte: INDAGINE T.E.A.

Variabile 6: Superficie massima dei lotti delle aree PIP presenti nel Comune

Indica la superficie del lotto più grande disponibile tra le eventuali aree PIP del Comune. E' di tipo quantitativo continuo. L'unità di misura è il metro quadrato.

Fonte: INDAGINE T.E.A.

Variabile 7: Totale lotti aree PIP disponibili nel Comune

Indica il totale di lotti disponibili nelle aree PIP del Comune.

E' di tipo quantitativo discreto.

Fonte: INDAGINE T.E.A.

Variabile 8: Numero di Aree Industriali

Area Industriale: parte del territorio indicata nel PRG o nel PdF come area dotata di caratteri architettonici e funzionali omogenei, finalizzati all'accoglienza di impianti industriali.

E' di tipo quantitativo discreto.

Fonte: INDAGINE T.E.A.

Variabile 16: Superficie totale di Aree Industriali presente nel Comune

Indica il totale di superficie adibita ad area industriale presente nel Comune.

La variabile è di tipo quantitativo continuo. L'unità di misura è il metro quadrato.

Fonte: INDAGINE T.E.A.

Variabile 10: Numero medio di infrastrutture presenti nelle eventuali aree del Comune

Infrastrutture: complesso di impianti, strutture, attrezzature e simili necessarie per avviare o agevolare lo svolgimento di un'attività. Gli impianti considerati nel caso specifico sono: energia elettrica, metano, acqua, fognature, depuratore, viabilità, servizi consortili (derivano da accordi tra imprenditori).

E' di tipo quantitativo discreto.

Fonte: INDAGINE T.E.A.

Variabile 11: Distanza tra il Comune ed il Capoluogo (Catanzaro)

E' di tipo quantitativo continuo. L'unità di misura è il chilometro.

Fonte: INDAGINE T.E.A.

Variabile 12: Distanza tra il Comune e l'aeroporto più vicino

Indica la distanza il Comune e l'aeroporto più vicino; gli aeroporti interessati sono quello di Lamezia Terme e quello di Crotona.

E' di tipo quantitativo continuo. L'unità di misura è il chilometro.

Fonte: INDAGINE T.E.A.

Variabile 13: Distanza tra il Comune e la stazione ferroviaria più vicina

E' di tipo quantitativo continuo. L'unità di misura è il chilometro.

Fonte: INDAGINE T.E.A.

Variabile 14: Distanza tra il Comune e l'A3 (Salerno – Reggio Calabria)

E' di tipo quantitativo continuo. L'unità di misura è il chilometro.

Fonte: INDAGINE T.E.A.

Variabile 15: Distanza tra il Comune e la strada statale più vicina

Indica la distanza tra il Comune e la strada statale più vicina; le strade statali interessate sono la SS 280, la SS 106 e la SS 18. E' di tipo quantitativo continuo.

L'unità di misura è il chilometro.

Fonte: INDAGINE T.E.A.

Variabile 16: Numero di Laureati

Indica il numero di persone laureate tra i residenti del Comune al '91.

E' di tipo quantitativo discreto.

Fonte: ISTAT (Censimento della popolazione, 1991)

Variabile 17: Numero di Diplomi

Indica il numero di persone diplomate tra i residenti del Comune al '91.

E' di tipo quantitativo discreto.

Fonte: ISTAT (Censimento della popolazione, 1991)

Variabile 18: Popolazione residente

Popolazione residente: numero di persone aventi la propria dimora abituale nel Comune; non cessano di appartenere alla popolazione residente le persone temporaneamente dimoranti in altro Comune o all'estero.

Indica la popolazione residente di ogni Comune al 31 12 1999.

E' di tipo quantitativo discreto.

Fonte: ISTAT (Annuario Demografico)

Variabile 19: Popolazione residente compresa tra i 14 ed i 65 anni

Indica la popolazione residente in ogni Comune al 31 12 1999 avente età compresa tra i 14 ed i 65 anni.

E' di tipo quantitativo discreto.

Fonte: ISTAT (Annuario Demografico)

Variabile 20: Uffici Postali

Indica il numero degli uffici postali presenti in ogni Comune.

E' di tipo quantitativo discreto.

Fonte: PAGINE GIALLE

Variabile 21: Banche

Indica il numero delle banche presenti in ogni Comune.

E' di tipo quantitativo discreto.

Fonte: PAGINE GIALLE

Variabile 22: Numero di Unità Locali

Unità Locale: luogo variamente denominato (stabilimento, laboratorio, negozio, officina, abitazione, scuola, ecc.) in cui si realizza la produzione di beni o nel quali si organizza la prestazione di servizi destinabili o non destinabili alla vendita.

Indica il numero di Unità Locali presenti in ogni Comune.

E' di tipo quantitativo discreto.

Fonte: ISTAT (Censimento Intermedio, 1996)

Variabile 23: Numero di Addetti nelle Unità Locali

Addetto: persona dipendente e indipendente occupata (a tempo pieno o part-time o per contratto di formazione lavoro) presso l'Unità Locale ubicata sul territorio nazionale, anche se temporaneamente assente per servizio, ferie, malattia, sospensione dal lavoro, cassa integrazione guadagni, ecc.

Indica il numero di addetti nelle Unità Locali del Comune.

E' di tipo quantitativo discreto.

Fonte: ISTAT (Censimento Intermedio, 1996)

Variabile 24: Numero di Imprese

Impresa: organizzazione di un'attività economica esercitata con carattere professionale per la produzione di beni o per la prestazione di servizi destinabili alla vendita; fruisce di una certa autonomia con particolare riguardo alle scelte produttive, di vendita e di distribuzione degli utili; il responsabile è rappresentato da una o più persone fisiche, o da una o più persone giuridiche.

Indica il numero di Imprese presenti in ogni Comune.

E' di tipo quantitativo discreto.

Fonte: ISTAT (Censimento Intermedio, 1996)

Variabile 25: Numero di Addetti nelle Imprese

Indica il numero di addetti nelle Imprese del Comune.

E' di tipo quantitativo discreto.

Fonte: ISTAT (Censimento Intermedio, 1996)

APPENDICE C: Descrizione delle unità

- 1: Albi
- 2: Amaroni
- 3: Amato
- 4: Andali
- 5: Argusto
- 6: Badolato
- 7: Belcastro
- 8: Borgia
- 9: Botricello
- 10: Caraffa di Catanzaro
- 11: Cardinale
- 12: Carlopoli
- 13: Catanzaro
- 14: Cenadi
- 15: Centrache
- 16: Cerva
- 17: Chiaravalle Centrale
- 18: Cicala
- 19: Conflenti
- 20: Cortale
- 21: Cropani

- 22: Curinga
- 23: Davoli
- 24: Decollatura
- 25: Falerna
- 26: Feroletto Antico
- 27: Fossato Serralta
- 28: Gagliato
- 29: Gasperina
- 30: Gimigliano
- 31: Girifalco
- 32: Gizzeria
- 33: Guardavalle
- 34: Isca sullo Ionio
- 35: Jacurso
- 36: Lamezia Terme
- 37: Magisano
- 38: Maida
- 39: Marcedusa
- 40: Marcellinara
- 41: Martirano
- 42: Martirano Lombardo

- 43:** Miglierina
- 44:** Montauro
- 45:** Montepaone
- 46:** Motta Santa Lucia
- 47:** Nocera Tirinese
- 48:** Olivadi
- 49:** Palermiti
- 50:** Pentone
- 51:** Petrizzi
- 52:** Petronà
- 53:** Pianopoli
- 54:** Platania
- 55:** San Floro
- 56:** San Mango d'Aquino
- 57:** San Pietro a Maida
- 58:** San Pietro Apostolo
- 59:** San Sostene
- 60:** Santa Caterina dello Ionio
- 61:** Sant'Andrea Apostolo dello Ionio
- 62:** San Vito sullo Ionio
- 63:** Satriano

- 64:** Sellia
- 65:** Sellia Marina
- 66:** Serrastretta
- 67:** Sersale
- 68:** Settingiano
- 69:** Simeri Crichi
- 70:** Sorbo San Basile
- 71:** Soverato
- 72:** Soveria Mannelli
- 73:** Soveria Simeri
- 74:** Squillace
- 75:** Stalettì
- 76:** Taverna
- 77:** Tiriolo
- 78:** Torre di Ruggiero
- 79:** Vallefiorita
- 80:** Zagarise

BIBLIOGRAFIA

- Aldenderfer M. K., Blashfield R. K. (1986), *Cluster analysis*, Sage publications.
- Arrighetti A. (1982), *Piccola impresa e politica industriale*, Franco Angeli editore.
- Barnett V. (1981), *Interpreting multivariate data*, John Wiley and sons.
- Bellandi M., Russo M. (1994), *Distretti industriali e cambiamento economico locale*, Rosenberg e Sellier.
- Bioni C., Canovi L., Fornaciari E., Landi A. (1994), *Banca e impresa nei mercati finanziari locali*, Il Mulino.
- Bolasco S. (1999), *Analisi multivariata dei dati: metodi, strategie e criteri*, Carocci editore, Roma.
- Bottazzi G. (1992), *La dimensione locale*, Franco Angeli editore.
- Brasini S., Tassinari F., Tassinari G. (1996), *Marketing e pubblicità. Metodi di analisi statistica*, Il Mulino.
- Del monte A. (1977), *Politica regionale e sviluppo economico*, Franco Angeli editore.
- Del Monte A., Giannola A. (1997), *Istituzioni economiche e mezzogiorno*, La Nuova Italia Scientifica.

- Del Monte A., Raffa M. (1977), *Tecnologia e decentramento produttivo*, Rosenberg and Sellier.
- Diday E., Hayashi C., Jambu M., Ohsumi N. (1987), *Recent developments in clustering and data Analysis*, (Proceedings of the Japanese-French Scientific Seminar), Academic Press.
- Fabbris L. (1997), *Statistica multivariata. Analisi esplorativa dei dati*, McGraw-Hill, Milano.
- Fuà G., Zacchia C. (1983), *Industrializzazione senza fratture*, Il Mulino.
- Gordon A. D. (1981), *Classification*, Chapman and Hall.
- Guidicini P., Scidà G. (1983), *Tecnologie, culture e nuove ipotesi di sviluppo*, Franco Angeli editore.
- Hand D. J. (1993), *Discrimination and classification*, John Wiley and sons.
- Hartigan J. A. (1975), *Clustering algorithms*, John Wiley and sons.
- Imbriani C. (a cura di) (1991), *Commercio estero, competitività e specializzazione dell'Italia*, Franco Angeli editore.
- Jansen J., Marcotorchino J.F., Proth J. M (1991), *New trends in data analysis and applications*, North-Holland.
- Krugman P. (1991), *Geography and trade*, Luven University press, Cambridge.
- Malanima P. (1997), *Economia preindustriale*, Bruno Mondadori editore.
- Morrison D. F. (1976), *Metodi di analisi statistica multivariata*, Casa editrice ambrosiana.

- Muirhead R. J. (1982), *Aspects of multivariate statistical theory*, John Wiley and sons.
 - Rizzi A. (1985), *Analisi dei dati*, La nuova Italia Scientifica.
 - Sadocchi S. (1980), *Manuale di analisi statistica multivariata*, Franco Angeli editore.
 - Sadocchi S. (1982), *Sull'interpretazione di alcuni strumenti statistici multivariate e sulle loro connessioni*, Dipartimento Statistico Università degli studi di Firenze.
 - Sadowski W. (1971), *Statistica per economisti*, Etas Kompass.
 - Seber G. A. F. (1983), *Multivariate observations*, John Wiley and sons.
 - Silva F., Viesti G. (1989), *Il difficile sviluppo dell'industria nel mezzogiorno*, Franco Angeli editore.
 - Viesti G. (2000), *Come nascono i distretti industriali*, Laterza, Bari.
 - Zani S. (1993), *Metodi statistici per le analisi territoriali*, Franco Angeli editore.
 - Zani S. (1999), *Analisi dei dati statistici*, Giuffrè editore, Milano.
 - Zenga M. (1988), *Introduzione alla statistica descrittiva*, Vita e pensiero, Milano.
-