VALIDAZIONE DEI DATI E DEI MODELLI

Luigi Mussio
Politecnico di Milano – DICA
Piazza L. da Vinci, 32 – 20133 Milano
Tel. 02–2399–6501, Fax. 02–2399–6602
e–mail luigi.mussio@polimi.it

RIASSUNTO

La validazione dei dati e dei modelli si fonda su varie tecniche dell'analisi multivariata, ciascuna delle quali prende in considerazione insiemi di dati, contenenti i valori osservati di certe variabili statistiche. Queste ultime sono spesso considerate campioni estratti da universi più grandi (ad es., variabili casuali). Data la complessità delle premesse, lo studio dell'analisi multivariata si può articolare, principalmente, nei seguenti punti:

- semplificazione strutturale;
- classificazione degli oggetti;
- raggruppamento degli attributi (cluster analysis);
- analisi della connessione;
- analisi della dipendenza funzionale (regressione multipla);
- costruzione e verifica d'ipotesi (inferenza statistica);

attraversando, trasversalmente, tutta la statistica classica e, nel contempo, estendendone i contenuti.

PREMESSA

La validazione dati e dei modelli si fonda sulle varie tecniche dell'analisi multivariata, di volta in volta, studiando la variabilità e l'interdipendenza fra gli attributi, entro una classe di oggetti. Essa prende in considerazione insiemi di dati, ciascuno dei quali, relativo ad un oggetto della classe, contiene i valori osservati di certe variabili statistiche. Questi insiemi possono talvolta essere completi; mentre più spesso sono estratti da insiemi più grandi. A loro volta, le variabili statistiche osservate sono in generale campioni estratti da variabili casuali, di tipo continuo o, raramente, discreto. Da una tale generalità e complessità di premesse, lo studio dell'analisi multivariata è variamente articolato.

- Semplificazione strutturale. Gli insiemi di dati devono essere ricondotti, se possibile, in forme più semplici con cambi di variabili, in particolare, con trasformazioni capaci di sciogliere variabili connesse in variabili indipendenti.
- Classificazione degli oggetti. L'analisi dell'insieme di dati deve porre in evidenza la presenza di gruppi (clusters), ovvero di sottoinsiemi di oggetti, caratterizzati da valori preferenziali degli attributi o di parte di essi, cercando di ricondurre a poco le notevoli variabilità presenti.
- Raggruppamento degli attributi (clustering). L'analisi degli insiemi di dati deve far ricadere, per quanto possibile, differenti variabili in un unico gruppo.
- Analisi della connessione. Gli insiemi di dati devono essere studiati rispetto alla dipendenza vaga e generica o meno fra le variabili contenute (ovvero all'essere in connessione di queste ultime).

- Analisi della dipendenza funzionale. Gli insiemi di dati devono essere studiati rispetto alla dipendenza funzionale o meno fra le variabili contenute (ovvero all'essere in regressione di queste ultime), con particolare riferimento alla dipendenza lineare o correlazione (ovvero all'essere in regressione lineare o correlate).
- Costruzione e verifica d'ipotesi. Il confronto probabilistico, fra statistiche campionarie e valori teorici di riferimento, permette di formulare un giudizio critico sui risultati ottenuti nelle varie tappe dell'analisi.

Scopo di questo lavoro è cercare di addentrarsi, senza ampollosità e pretese esaustive, nei problemi operativi della "cluster analysis", della regressione multipla e dell'inferenza statistica (multivariata) tanto per stime di parametri da campioni normali, quanto per modelli non – parametrici (distribution free)[1].

PARTE I – ANALISI MULTIVARIATA

1. LA "CLUSTER ANALYSIS"[2]

La "cluster analysis" denomina varie procedure statistiche ed i corrispondenti algoritmi computazionali, capaci di classificare insiemi di dati, raggruppandoli (clustering) in gruppi o grappoli (clusters) il cui numero ed i cui centri o valori rappresentativi (cluster points) sono, in generale, inizialmente sconosciuti. Queste tecniche, spesso preliminari ad altre successive elaborazioni, aiutano a risolvere i seguenti problemi:

- esplorazione, per migliorare la conoscenza di un insieme di dati di dubbia classificazione o interpretazione;
- compattamento di un insieme di dati la cui rilevante dimensione non permette successive elaborazioni;
- classificazione, propriamente detta o stratificazione o interpretazione; da questa possono prendere avvio l'analisi della connessione e quella della dipendenza funzionale.

In ogni caso, obiettivo della "cluster analysis" è massimizzare, insieme, la compattezza interna dei gruppi identificati e la differenziazione reciproca degli stessi. Le tecniche della "cluster analysis" si suddividono in:

- tecniche gerarchiche, dove classificazione e raggruppamento procedono, da un'iterazione all'altra, tenendo conto dei risultati ottenuti in precedenza e considerati definitivi;
- tecniche non–gerarchiche, dove classificazione e raggruppamento avvengono ad ogni iterazione, se ne esiste più di una, in modo indipendente.

In entrambi i casi si hanno:

- tecniche agglomerative, dove il punto di partenza è costituito dai singoli elementi;
- tecniche divisive, dove il punto di partenza è costituito dall'intero insieme dei dati;

come pure:

[1] Un ripasso della variabile statistica doppia, benché elementare ed arcinoto, è fortemente raccomandato, perché costituisce un buon avvio allo studio dell'analisi multivariata. Infatti tutto ciò che supera le due (o al massimo tre) dimensioni non è facilmente rappresentabile (cioè non si vede!), cosa che ne complica lo studio.

[2] Questo paragrafo riprende e generalizza quanto esposto da Mattia Crespi , in seguito ad un'accurata ed approfondita ricerca bibliografica, in una relazione svolta al seminario del Gruppo di Lavoro e di Studio della SIFET *Modelli matematici per la Geodesia e la*

- tecniche di partizione completa (clustering, propriamente detto), dove il punto di arrivo è costituito da gruppi completamente disgiunti;
- tecniche di ammassamento (clumping), dove il punto di arrivo è costituito da gruppi, parzialmente o totalmente, sovrapposti.

Lo studio seguente prende in considerazione tecniche non–gerarchiche, alternativamente agglomerative, o suddivise, sempre di partizione completa. Tecniche gerarchiche, aventi le stesse caratteristiche, si possono eseguire ripetendo, ad ogni iterazione, quanto fatto per le corrispondenti tecniche non–gerarchiche ed operando, via via, su insiemi di dati, stabilmente, sempre più agglomerati o sempre più suddivisi.
Le tecniche di ammassamento (clumping), in quanto ben più complicate, dovendosi definire, nello specifico, l'entità, deterministica o probabilistica, della sovrapposizione fra i gruppi, sono considerate estranee agli scopi del presente lavoro. Infatti esse possono fare riferimento a logiche fuzzy nelle quali si intende costruire insiemi sfumati, dove l'attribuzione di un dato ad un insieme è stabilito da un funzione d'appartenza che lo ripartisce fra uno o più insiemi (e, al limite, fra tutti gli insiemi del problema in esame). Come evidente, mentre una logica booleana, presentado ipotesi in mutua esclusione, è governata da semplici coefficienti binari d'appartenenza, la presentazione delle logiche fuzzy, per la loro complessità, costituisce un argomento a sé stante cui si rimanda.

Con riferimento esclusivo allo studio proposto, la formulazione matematica di un problema di "cluster analysis" si sviluppa, allora, secondo le seguenti linee fondamentali. Scopo della "cluster analysis" è effettuare una partizione dell'insieme S contenente m oggetti (m noto) di tipo numerico o non–numerico in n cluster (n noto o n incognito, $1 \leq n \leq m$), tali che:

$$C_k \cap C_h = 0 \qquad \forall\ k,h$$

$$C_k \neq 0 \qquad \forall\ k$$

$$C_1 \cup C_2 . \cup ... \cup C_{n-1} \cup C_n = S$$

dove ciascun cluster C_j è rappresentato da un oggetto (detto cluster point) appartenente o meno all'insieme S. Un metodo generale per effettuare la "cluster analysis" è la minimizzazione di una determinata funzione obiettivo del tipo:

$$F = F(C_1,...,C_n)$$

Si noti, a riguardo, come il presente principio variazionale sia aperto a qualsiasi soluzione ammissibile, potendo variare addirittura dalla mancanza di una soluzione ammissibile, fino alla presenza di un'infinità numerabiledi tali soluzione, come pure avendo una ed una sola soluzione, oppure un certo numero finito (solitamente piccolo). Allora saranno, caso per caso, condizioni specifiche a determinarlo con precisione.

Cartografia (Bologna, 20 dicembre 1988); a lui va il ringraziamento dell'autore.

La Fig. 1.1 riporta, in forma grafica, casi e sottocasi della “cluster analysis”.

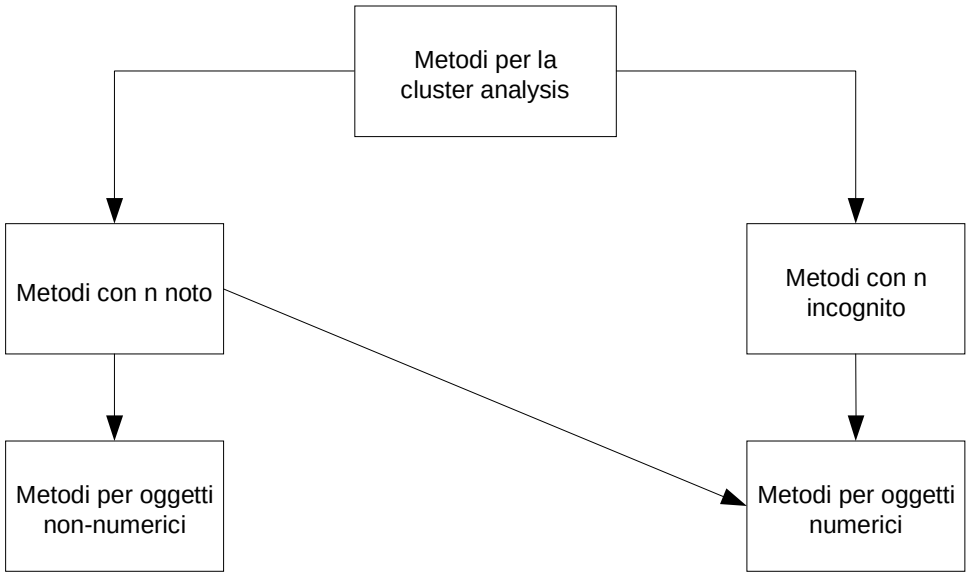


Fig. 1.1

Nel prosieguo, sono illustrate alcune funzioni obiettivo, con e senza l'introduzione dei "cluster points", articolate nei casi con n noto e n incognito e nei sottocasi, rispettivamente, per oggetti numerici e non–numerici, oppure per soli oggetti numerici. Tale elenco costituisce la quasi totalità dei casi solitamente presi in considerazione nei problemi di “cluster analysis”.
Tuttavia, con specifico riferimento ai problemi geodetici e geomatici (e, più in generale, a quelli ingegneristici), resta da sottolineare come il caso n incognito per soli oggetti numerici rappresenti quello più interessante ed importante.

a. Metodi con *n* noto

a.1 Metodi per oggetti numerici

Incognite: $u_{ik}\ (i=1,m;\ k=1,n)$

esse costituiscono i coefficienti di attribuzione degli oggetti $\underline{x}_i$ (ad una o più componenti) ai "clusters", rappresentati dai "cluster points" $\underline{y}_k$.

Funzione obiettivo con introduzione dei "cluster points":

$$\phi_{p,q,r} = \sum_{k=1}^{n} \sum_{i=1}^{m} u_{ik}{}^{r} \left\| \underline{x}_i - \underline{y}_k \right\|_p^q \qquad (1 \le p \le \infty\ ,\ q \ge 1,\ r \ge 1)$$

Funzione obiettivo senza introduzione dei "cluster points":

$$\phi_{p,q,r} = \frac{\sum_{k=1}^{n} \sum_{i=1}^{m} \sum_{j=1}^{m} u_{ik}^{r} u_{jk}^{r} \left\| \underline{x}_i - \underline{x}_j \right\|_p^q}{2 \sum_{i=1}^{m} u_{ik}^{r}} \qquad (1 \le p \le \infty,\ q \ge 1,\ r \ge 1)$$

Vincoli: $0 \le u_{ik} \le 1 \quad \forall\ i,k$ (metodi fuzzy)

$u_{ik} = 0,1 \quad \forall\ i,k$ (metodi binari)

ed, inoltre: $\sum_{k=1}^{n} u_{ik} = 1 \quad \forall\ i.$

a.2 Metodi per oggetti non–numerici

Incognite: $u_{ik}\ (i=1,m;\ k=1,n)$

esse costituiscono i coefficienti di attribuzione degli oggetti ai "clusters", rappresentati dai "cluster points".

Si osservi, a riguardo, come il caso di oggetti non–numerici limiti notevolmente le possibilità di operare, in quanto non è possibile definire fra essi alcuna nozione di distanza, non essendo ovviamente gli stessi rappresentabili tramite coordinate (ovvero con una referenza spaziale forte), ma solo tramite attributi qualitativi (ovvero con una referenza spaziale debole). In tal caso, occorre definire "ad hoc" opportuni coefficienti di dissimilarità [3], capaci di mettere in evidenza caratteristiche comuni, oppure caratteristiche diverse, fra gli oggetti presi in considerazione. D'altra parte, il problema di un referenza spaziale debole è un problema, più generale, che fa riferimento a tutto quanto, spesso di uso molto comune, non può semplicemente modellare in forma numerica.

Funzione obiettivo con introduzione dei "cluster points":

$$\phi_{q,r} = \sum_{k=1}^{n} \sum_{i=1}^{m} u_{ik}^{r} D^{q}(i,k) \qquad (q \ge 1,\ r \ge 1)$$

dove $D(i,j)$ è il coefficiente di dissimilarità tra l'oggetto i ed il cluster point k.

[3] Si chiamano coefficienti di dissimilarità gli elementi di matrici quadrate e simmetriche, a diagonale nulla, dette di dissimilarità. Essi quantificano differenze (qualitative) tra oggetti non–numerici che, non sempre, soddisfano la disuguaglianza triangolare:

$$d(i,j) \le d(i,h) + d(h,j)$$
$$d(i,j) \ge \left| d(i,h) - d(j,h) \right|$$

Essi possono essere calcolati, anche, a partire da oggetti numerici, mediante opportune funzioni.

Funzione obiettivo senza introduzione dei "cluster points":

$$\phi_{q,r} = \frac{\sum_{k=1}^{n} \sum_{i=1}^{m} \sum_{j=1}^{m} u_{ik}^{r} u_{jk}^{r} d^{q}(i,j)}{2\sum_{i=1}^{m} u_{ik}^{r}} \qquad (q \geq 1,\ r \geq 1)$$

dove $d\ (i,j)$ è il coefficiente di dissimilarità tra l'oggetto i e l'oggetto j.

Vincoli: $0 \leq u_{ik} \leq 1 \quad \forall\ i,k$ (metodi fuzzy)

$u_{ik} = 0,1 \quad \forall\ i,k$ (metodi binari)

ed, inoltre: $\sum_{k=1}^{n} u_{ik} = 1 \quad \forall\ i.$

b. Metodi con *n* incognito: per oggetti numerici

Dati aggiuntivi per il controllo delle iterazioni: s_{max} dispersione massima L_p dei "clusters";

d_{min} distanza minima L_p tra due "cluster points",

essi governano l'evoluzione di un problema di "cluster analysis" nelle sue fasi agglomerative e/o divisive.

Incognite: n (numero di clusters), ed, inoltre: $u_{ik}\ (i=1,m;\ k=1,n(j))$

esse costituiscono i coefficienti di attribuzione degli oggetti $\underline{x}_i$ (ad una o più componenti) ai "clusters", rappresentati dai "cluster points" $\underline{y}_k$ che, variando nel loro numero, costituiscono incognite aggiuntive del problema ($n(j)$ è il numero di "clusters" all'iterazione j–esima).

Funzione obiettivo all'iterazione j–esima:

$$\phi_{p,q,r} = \sum_{k=1}^{n(j)} \sum_{i=1}^{m} u_{ik}^{r} \left\| \underline{x}_i - \underline{y}_k \right\|_p^q \qquad (1 \leq p \leq \infty,\ q \geq 1,\ r \geq 1)$$

Vincoli all'iterazione j–esima: $u_{ik} = 0,1 \quad \forall\ i,k\ (k=1,\ n(j))$, ed, inoltre: $\sum_{k=1}^{n(j)} u_{ik} = 1 \quad \forall\ i$

Vincoli di controllo alla fine dell'iterazione j–esima:

$$\sum_{i=1}^{m} u_{ik}^{r} \left\| \underline{x}_i - \underline{y}_k \right\|_p^q \leq s_{max} \qquad \forall\ k\ (k=1,\ n(j))$$

$$\left\| \underline{y}_k - \underline{y}_l \right\|_p^q \geq d_{min} \qquad \forall\ k, l\ (k, l = 1, n(j))$$

pertanto, ad ogni iterazione, è aggiornato $\underline{n}(j)$, mediante opportune funzioni che ricalcolano i "cluster points", fino a quando tutti i vincoli risultano soddisfatti.

A commento della presentazione delle funzioni obiettivo, si formulano alcune osservazioni generali. Le funzioni obiettivo non–lineari nelle incognite presentano ottimi locali, in numero crescente con la non–linearità. Di conseguenza, gli algoritmi computazionali utilizzati non garantiscono la convergenza all'ottimo globale. Tuttavia è possibile cercare l'ottimo globale, utilizzando inizializzazioni differenti.
Poche funzioni obiettivo, comprese nell'elenco mostrato, sono effettivamente utilizzate per le proprietà dei metodi ad esse relative (robustezza, stabilità della soluzione rispetto all'inizializzazione, adattabilità alla non–sfericità dei clusters), oppure per le proprietà degli algoritmi computazionali (velocità di convergenza).
A proposito di questi ultimi, due tipi di algoritmi computazionali, propri della programmazione matematica, sono utilizzati, perché capaci di risolvere problemi con un decisore ed un obiettivo in ambiente deterministico:

- algoritmi evolutivi (essi pervengono alla soluzione, mediante iterazioni successive): algoritmo delle direzioni ammissibili, algoritmo "branch and bound" tipo Balas;
- algoritmi euristici (essi pervengono alla soluzione, con procedimenti basati sulle caratteristiche del problema in considerazione): algoritmo dei "medoids", algoritmo ISODATA (Questo ultimo algoritmo, particolarmente indicato per il caso n incognito per soli oggetti numerici, sarà presentato dettagliatamente e, di esso, saranno presentate alcune varianti di notevole interesse.).

Nel prosieguo, si illustrano alcuni metodi per la "cluster analysis" ed i rispettivi algoritmi computazionali, articolati negli stessi casi e sottocasi, già illustrati in precedenza.

a.1 Metodi con n noto per oggetti numerici:

- metodi fuzzy;
- problemi non–lineari vincolati (con vincoli di uguaglianza lineari);
- algoritmo delle direzioni ammissibili.

Metodo L_{1a}: p=1, q=1, r=2, con introduzione dei "cluster points":

$$\phi_{1,1,2} = \sum_{k=1}^{n} \sum_{i=1}^{m} u_{ik}^2 \sum_{h} \left| x_{ih} - y_{kh} \right| \qquad (h, \text{ indice di componente})$$

Metodo L_{1b}: p=2, q=1, r=2;
con introduzione dei "cluster points":

Metodo L_{1c}: p=2, q=1, r=2;
senza introduzione dei "cluster points":

$$\phi_{2,1,2} = \sum_{k=1}^{n} \sum_{i=1}^{m} u_{ik}^2 \sqrt{\sum_h \left(x_{ih} - y_{kh}\right)^2}$$

(*h*, indice di componente);

$$\phi_{2,1,2} = \sum_{i=1}^{m} \sum_{j=1}^{m} u_{ij}^2 \sqrt{\sum_h \left(x_{ih} - x_{jh}\right)^2}$$

(*h*, indice di componente)

Metodo L_{2a}: p=2, q=2, r=2; con introduzione dei "cluster points":

$$\phi_{2,2,2} = \sum_{k=1}^{n} \sum_{i=1}^{m} u_{ik}^2 \sum_h \left(x_{ih} - y_{kh}\right)^2$$

(*h*, indice di componente)

Si osservi come l'estensione ad un numero qualsiasi di componenti, permetta di operare con problemi ad una o più (ad es., due o tre) componenti, dando ampia generalità ai metodi proposti. La Fig. 1.2 illustra con evidenza grafica alcune importanti proprietà dei metodi con *n* noto per oggetti numerici.

robustezza	*stabilità*	*adattabilità*
$-L_{1a}$	$-L_{2a}$, L_{1c}	$-L_{1a}$
		$-L_{1b}$
$-L_{1b}$	$-L_{1b}$	
$-L_{1c}$		$-L_{1c}$
$-L_{2a}$	$-L_{1a}$	$-L_{2a}$

Fig. 1.2

a.2 Metodi con *n* noto per oggetti non–numerici:

- caratteristiche da definirsi caso per caso.

Metodo L_{1c}: q=1, r=2; senza introduzione dei "cluster points":

- metodo fuzzy; problema non–lineare vincolato (con vincoli di uguaglianza lineari);
- algoritmo delle direzioni ammissibili;

$$\phi_{1,2} = \frac{\sum_{k=1}^{n} \sum_{i=1}^{m} \sum_{j=1}^{m} u_{ik}^2 u_{jk}^2 d(i,j)}{2\sum_{i=1}^{m} u_{ik}^2}$$

Metodo L_{1d}: q=1, r=1; con introduzione dei "cluster points":

- metodo binario; problema lineare vincolato;
- algoritmo "branch and bound" tipo Balas, oppure algoritmo dei "medoids";

$$\phi_{1,1} = \sum_{k=1}^{n} \sum_{i=1}^{m} u_{ik} D(i,k)$$

Metodo L_{2a}: q=2, r=2; con introduzione dei "cluster points":

- metodo fuzzy; problema non–lineare vincolato (con vincoli di uguaglianza lineari);
- algoritmo delle direzioni ammissibili;

$$\phi_{2,2} = \sum_{k=1}^{n} \sum_{i=1}^{m} u_{ik}^{2} D^{2}(i,k)$$

La Fig. 1.3 illustra alcune importanti proprietà dei metodi con n noto per oggetti non–numerici.

$$robustezza \left| \begin{array}{ll} -L_{1c} & L_{1d} \\ -L_{2a} & \end{array} \right. \qquad stabilità \left| \begin{array}{ll} -L_{1d} & \\ -L_{2a} & L_{1c} \end{array} \right. \qquad adattabilità \left| \begin{array}{ll} -L_{1c} & L_{1d} \\ -L_{2a} & \end{array} \right.$$

Fig. 1.3

b. Metodi con n incognito per oggetti numerici:

- metodi binari; problemi lineari vincolati (iterativi);
- algoritmi "branch and bound" tipo Balas (iterativo), oppure algoritmo ISODATA.

Metodo L_{1d}: p=1, q=1, r=1; con introduzione dei "cluster points":

$$\phi_{1,1,1} = \sum_{k=1}^{n(j)} \sum_{i=1}^{m} u_{ik} \sum_{h} \left| x_{ih} - y_{kh} \right|$$ (h, indice di componente)

Metodo L_{2b}: p=2, q=2, r=1; con introduzione dei "cluster points":

$$\phi_{2,2,1} = \sum_{k=1}^{n(j)} \sum_{i=1}^{m} u_{ik} \sum_{h} \left(x_{ih} - y_{kh} \right)^{2}$$ (h, indice di componente)

Si osservi come l'estensione ad un numero qualsiasi di componenti, permetta ancora di operare con problemi ad una o più (ad es., due o tre) componenti, dando ampia generalità ai metodi proposti.

La Fig. 1.4 illustra con evidenza grafica alcune importanti proprietà dei metodi con n incognito per oggetti numerici.

$$robustezza \left| \begin{array}{l} -L_{1d} \\ -L_{2b} \end{array} \right. \qquad stabilità \left| \begin{array}{l} -L_{1d}, L_{2b} \\ \\ \end{array} \right. \qquad adattabilità \left| \begin{array}{l} -L_{1d} \\ -L_{2b} \end{array} \right.$$

Fig. 1.4

A commento della presentazione dei metodi numerici per la "cluster analysis", si formulano alcune osservazioni conclusive. I metodi con norma L_1 sono i più robusti; inoltre i metodi fuzzy, di qualsiasi tipo, riescono ad individuare gli "outliers", mediante bassi coefficienti di attribuzione relativi a tutti i "clusters". I metodi lineari (con norma L_1 o L_2) sono i più stabili nella soluzione, rispetto alla inizializzazione del calcolo; inoltre i metodi con norma L_1 sono i più adattabili alla non–sfericità dei "clusters".

Gli algoritmi dei "medoids" ed ISODATA sono descritti nel seguito; invece gli altri algoritmi, in quanto classici della Ricerca Operativa, sono considerati estranei agli scopi del presente lavoro.

L'algoritmo dei "medoids" opera nei seguenti passi:

- assegnazione di una matrice di dissimilarità tra gli oggetti;
- selezione, uno alla volta, di n oggetti come "medoids", minimizzando la dissimilarità totale:

$$D_K = \sum_{k=1}^{n} \sum_{i=1}^{m} u_{ik} D(i,k)$$

- ottimizzazione dell'insieme dei "medoids", prendendo in considerazione tutte le coppie (i,k) in cui l'oggetto k è stato selezionato, come “medoid”, e l'oggetto i non è stato selezionato ed invertendo la selezione effettuata.

L'algoritmo dei "medoids" è utilizzabile, anche, con un numero elevato di oggetti, ma non garantisce la soluzione ottima; al contrario, gli algoritmi "branch and bound" garantiscono la soluzione ottima, ma sono applicabili, solo, con un numero molto limitato di oggetti (fino a 100).

L'algoritmo ISODATA opera, invece, nei seguenti passi:

1. selezione dei "cluster points" di inizializzazione e dei valori critici per il controllo delle iterazioni: s_{max} e d_{min} (il primo verifica l'ampiezza del parametro di dispersione di un gruppo: la varianza, oppure una statistica robusta come il "median absolute value"[4]; il secondo verifica l'entità della norma fra due "clusters": la

[4] Si chiama “median absolute value” il parametro di dispersione, avente lo stesso polo, ma costruito con le stesse regole, usate nella definizione della mediana:

$$mav = \delta : \quad P(\left|x - x^{(1)}\right| < \delta) < 0.5 \ \cup \ P(\left|x - x^{(1)}\right| \geq \delta) \geq 0.5$$

Si chiama, invece, “mean absolute value” il parametro di dispersione di norma L_1, avente per polo la mediana:

distanza euclidea, oppure un'altra distanza, secondo la regola prescelta per la formazione dei "clusters");

2. attribuzione allo stesso "cluster" degli oggetti circostanti, ovvero più vicini al corrispondente "cluster point";
3. divisione in due o più "clusters" di ogni "cluster" il cui parametro di dispersione è maggiore di s_{max}: corrispondentemente, nascono due o più nuovi "cluster points" (il criterio di divisione è arbitrario: una possibile scelta, per una divisione in due, consiste nel confronto fra i parametri di dispersione, dopo aver diviso un gruppo in due nuovi gruppi la cui dimensione corrisponde, ad es., ai rapporti $1{:}3$, $2{:}2$, $3{:}1$);
4. riattribuzione degli oggetti, ai "clusters", tenendo conto dei nuovi "cluster points";
5. fusione dei "clusters" i cui "cluster points" sono più "vicini" di d_{min}: corrispondentemente, scompaiono dei "cluster points" stessi (il criterio di fusione è arbitrario: una possibile scelta consiste nella fusione, a due a due, dei gruppi di una serie, ad eccezione degli ultimi tre, fusi in uno solo, se il numero dei gruppi è dispari);
6. iterazione dal passo 2, oppure arresto, se in un certo numero di iterazioni (ad es. 2) non è variato il numero dei "cluster points" (raggiungimento di un punto di riproduzione), oppure se il numero delle iterazioni supera un certo valore limite prefissato (ad es. 25).

Il passo di attribuzione (e/o riattribuzione) ad un "cluster" degli oggetti circostanti, ovvero più vicini al corrispondente "cluster point", può essere, variamente, gestito. Nel caso di oggetti ben configurati, la norma euclidea L_2 (usuale distanza pitagorica), oppure la norma L_1 (cosidetta distanza Manhattan) o la norma L_∞ (distanza dell'estremo superiore) si prestano, bene, al raggruppamento degli elementi in "clusters", rispettivamente, di forma circolare, nel primo caso, e di forma quadrata negli altri due.
Al contrario, nel caso di oggetti concavi, oppure pluriconvessi o, addirittura, appartenenti a sottospazi (linee nel piano, linee e/o superfici nello spazio), tutte le norme mostrano evidenti difficoltà al raggruppamento degli elementi in "clusters", proprio per la particolare forma di questi ultimi. Allora particolari tecniche di "parsing" (letteralmente: analisi sintattica; metodologia della linguistica, nota e sviluppatesi a partire dalla scoperta del sanscritto, alla fine del ‘700, per l'analisi grammaticale e logica del linguaggio e della scrittura: una disciplina apparentemente molto lontata, e tuttavia applicata, da tempo, nell'ambito di quella particolare branca dell'Informatica costituita dall'Intelligenza Artificiale, con l'elaborazione di immagini, i sistemi esperti, l'apprendimento automatico e la visione robotica), note come "line following", "region growing", ecc., bene si prestano alla bisogna, fornendo i risultati attesi.

Una breve digressione su alcuni aspetti particolari delle Scienze e Tecnologie dell’Informazione e sul loro ingresso impetuoso nelle discipline del rilevamento, con il proprompere della Geomatica, viene esposta nel prosieguo. Infatti un caso interessante di tale contaminazione culturale è il riconoscimento sintattico di modelli che realizzano il confronto di parti di mappe, immagini o modelli 3D con archetipi (“parsers”). Tappa essenziale dell’intera strategia è il cosiddetto “parsing”, strumento creato in ambiente linguistico e preso a prestito dalle Scienze Cognitive. Un sistema di riconoscimento di modelli, su base sintattica, è costituito dalle seguneti fasi:

$MAV = \Delta = E(\left|x - x^{(1)}\right|)$

- analisi del modello ideale e creazione della grammatica corrispondente;
- riconoscimento del modello.

L'analisi del modello prevede la selezione delle primitive, costituenti il modello e fondamentali per la descrizione degli oggetti, e la costruzione di una grammatica, atta a descrivere una determinata classe di oggetti (modello). Il riconoscimento di modelli si attua in tre passi:

- pre–processamento dell'oggetto;
- rappresentazione del modello;
- analisi sintattica ("parsing").

Il pre–processamento è costituito dalle seguenti operazioni:

- codifica ed approssimazione del messaggio espresso;
- filtraggio e miglioramento della qualità dell'informaione contenuta.

Ad esempio, un'informazione binaria è codificata mediante una sequanza di zeri ed uni, mentre un'onda è approssimata da un'espansione in serie di Fourier od ondine ("wavelets"). Inoltre viene spesso applicata una compressione dell'informazione contenuta, mentre le procedure di filtraggio e miglioramento della qualità servono a ridurre il rumore ed il degrado dei modelli. Il processo di rappresentazione del modello consiste nella segmentazione e nell'estrazione di primitive, in modo da rappresentare il modello mediante sotto–modelli e questi ultimi in termini di primitive. Svariate sono le tecniche di segmentazione ed i tipi di elaborazione:

- "thresholding" o "clustering";
- "edge detection" (riconoscimento di spigoli);
- "region extraction" (riconoscimento di aree).

Il primo metodo si basa sull'analisi (statistica o strutturale) di una soglia per una determinata caratteristica. Il riconoscimento degli spigoli è invece una tecnica, basata sul rilevamento di discontinuità, molto usata, perché ai confini delle regioni è possibile trovare informazione. Questo metodo è molto simile alla visione vera e propria: infatti è naturale percepire discontinuità e trarre conclusioni su dimensioni, forme e tessiture.
Per decidere se una rappresentazione appartenga o meno ad una determinata classe di oggetti (modello), descritta da una determinata grammatica, cioè se sia sintatticamente corretta, si ricorre ad un analizzatore sintattico ("parser") per l'analisi sintattica ("parsing"). Esiste infatti un'analogia tra la struttura gerarchica (ad albero) dei modelli e la sintassi dei linguaggi (per alcuni linguisti, tale struttura gerarchica è addirittura parte della mente umana). I modelli sono costruiti da sotto–modelli, in svariati modi di combinazione, proprio come le frasi sono formate da parole, e i sotto–modelli da primitive, proprio come le parole sono formate dalla concatenazione di sillabe e lettere.
Le regole di combinazione delle primitive degli oggetti sono grammatiche, derivate dal linguaggio che riesce a descrivere il modello. L'approccio sintattico è tanto più valido, quanto meglio riesce a descrivere oggetti complessi mediante un modesto numero di primitive (del modello prescelto ed adottato) e di regole, applicate un certo numero di volte (secondo regole dette di ri–scrittura). Fondamentale, allo scopo, è il binomio regole

– rappresentazioni: nella mente sono rappresentate le strutture sintattiche. Di conseguenza, una struttura sintattica è una struttura gerarichica, visualizzabile da un albero (o altre rappresentazioni equivalenti).
La fig. 1.5 illustra lo schema a blocchi dell'algoritmo ISODATA.

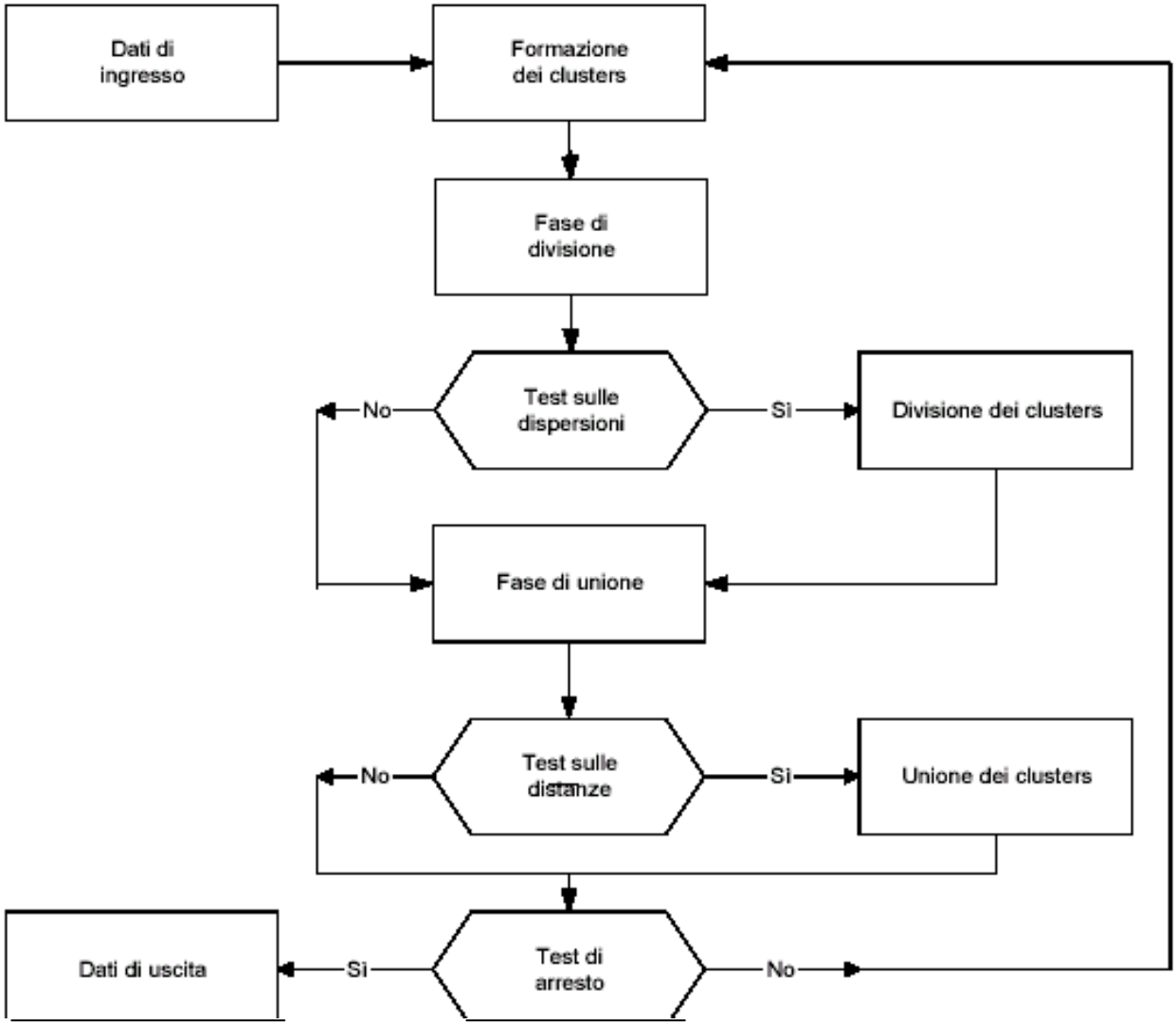


Fig 1.5

In tutti i casi suddetti (oggetti concavi, oppure pluriconvessi o, addirittura, appartenenti a sottospazi), l'attribuzione (e/o riattribuzione) ad un "cluster" degli oggetti circostanti può procedere come segue:

1. attribuzione ad ogni "cluster" dell'elemento più vicino al "cluster point";
2. attribuzione ad ogni "cluster", anche, degli eventuali elementi, appena, più lontani dal "cluster point", per tener conto di eventuali biforcazioni o biforcazioni plurime;
3. trasferimento della funzione di "cluster point" all'elemento (agli elementi), appena, selezionato (i);
4. attribuzione ad ogni "cluster" dell'elemento più vicino al "cluster point" corrente;
5. attribuzione ad ogni "cluster", anche, degli eventuali elementi, appena, più lontani del "cluster point" corrente, ancora per tener conto di eventuali biforcazioni o biforcazioni plurime;
6. ripetizione dei punti 3, 4 e 5, finché l'attribuzione di cui ai punti 4 e 5 avviene, effettivamente; al contrario, nel caso di non attribuzione ad alcun "cluster", si procede al successivo punto 7;

7. selezione di un qualsiasi eventuale elemento non, ancora, appartenente ad alcun "cluster", come nuovo "cluster point"; al contrario, nel caso di non selezione di alcun elemento, l'algoritmo ha termine;
8. esecuzione, limitatamente al solo nuovo "cluster" in formazione, dei punti del tutto analoghi ai punti 1–5 finché l'attribuzione di cui ai punti 1–2 e 4–5 avviene, effettivamente; al contrario, nel caso di non attribuzione al nuovo "cluster" in formazione, si ritorna al precedente punto 7.

Si osservi che, in assenza di "cluster points", l'algoritmo ha inizio, direttamente, dal punto 7 e procede alla formazione di "clusters", uno alla volta. Si noti, poi, come questo modo di procedere alla formazione di "clusters": in serie, anziché in parallelo, sia nei risultati più dipendente dalla casualità della partenza e meno robusto rispetto ad un'errata attribuzione (e/o riattribuzione) ad un "cluster" di oggetti, ad esso estranei.
La fig. 1.6 illustra uno schema a blocchi particolareggiato con riferimento al raggruppamento degli elementi in “clusters”.

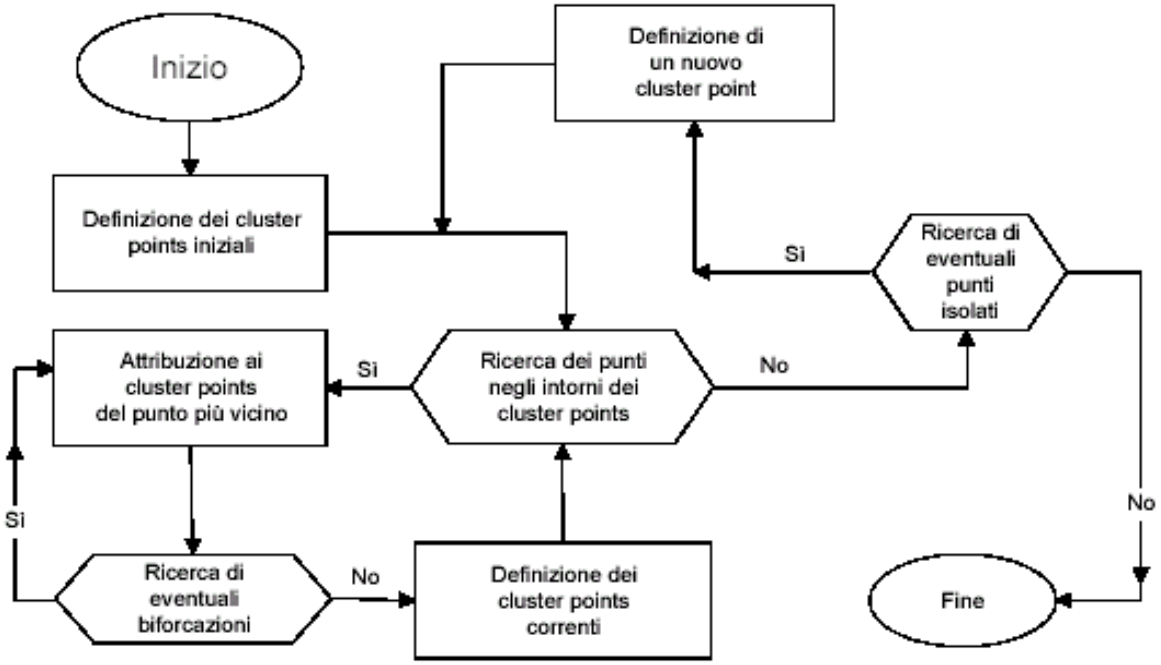


Fig. 1.6

Anche i passi di divisione in due o più "clusters" e di fusione dei "clusters" possono essere, variamente, gestiti. Ancora nel caso di oggetti ben conformati, le norme illustrate in precedenza si prestano, bene, alla definizione di indici di dispersione dei "clusters" e distanza fra i "clusters". Al contrario, nel caso di oggetti non ben conformati, le stesse norme mostrano, sempre, evidenti difficoltà alla definizione degli stessi indici. Allora divisioni e fusioni possono avvenire, ancora una volta, facendo ricorso alle particolari stesse tecniche di "parsing", adottate per l'attribuzione (e/o riattribuzione).
Si noti, a riguardo, come la divisione può avvenire, addirittura, nel corso della formulazione dei "cluster", ogni qual volta la distanza tra un "cluster point" o un "cluster point" corrente ed un elemento supera una certa opportuna soglia prefissata. Al contrario, la fusione può avvenire mediante la costruzione, con le stesse particolari tecniche di "parsing", di "clusters" di "clusters", ogni qual volta la distanza fra due qualsiasi elementi appartenenti a due "clusters" distinti non supera una certa opportuna soglia prefissata. Divisioni e

fusioni possono, poi, essere eventualmente migliorate, ripetendo una o più volte le operazioni descritte, dopo aver ridefinito le soglie sulla base di risultati, via via, ottenuti [5].

2. LA REGRESSIONE MULTIPLA[6]

La regressione lineare multipla costituisce il caso più semplice, dove è possibile dare espressione analitica al legame funzionale, evidenziato in diversi problemi, supponendo una dipendenza lineare fra le varie grandezze, da individuarsi, a partire dai dati disponibili, con la stima dei parametri della regressione stessa. La soluzione del problema è fornita dal principio dei minimi quadrati; procedure robuste sono altresì possibili con l'intento specifico di individuare ed eliminare "outliers" nella struttura dei dati (dando luogo alla cosiddetta **regressione robusta)**. La struttura dei dati per il calcolo della regressione lineare multipla è costituita da una tabella avente n colonne di vettori delle osservazioni e m ($m \geq n$) righe di elementi di ciascun vettore (Fig. 2.1). I vettori delle osservazioni, composti ciascuno dallo stesso numero di elementi ed ordinati secondo un qualsiasi parametro ordinatore (per es. il tempo di osservazione), formano una matrice rettangolare i cui elementi sono indicati con la consueta notazione x_{ki} (k=1,m; i=1,n).

Parametro ordinatore	*Vettori delle osservazioni*							
1	x_{11}	x_{12}	...	x_{1i}	...	x_{1i*}	...	x_{1n}
2	x_{21}	x_{22}	...	x_{2i}	...	x_{2i*}	...	x_{2n}
.	.	.	...	.	...	.	...	.
.	.	.	...	.	...	.	...	.
.	.	.	...	.	...	.	...	.
k	x_{k1}	x_{k2}	...	x_{ki}	...	x_{ki*}	...	x_{kn}
.	.	.	...	.	...	.	...	.
.	.	.	...	.	...	.	...	.
.	.	.	...	.	...	.	...	.
m	x_{m1}	x_{m2}	...	x_{mi}	...	x_{mi*}	...	x_{mn}
Medie per colonne	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_i$	...	$\bar{x}_{i*}$	...	$\bar{x}_n$
Varianze per colonne	$\sigma^2_{x_1}$	$\sigma^2_{x_2}$	...	$\sigma^2_{x_i}$	...	$\sigma^2_{x_{i*}}$	...	$\sigma^2_{x_n}$

Fig. 2.1

<u>Nota:</u> $i^* : x_{ki*} = y_k$ (k=1,m)

[5] L'intera procedura, oltre all'evidente adattabilità alla non–sfericità dei clusters, ha spiccate caratteristiche di robustezza, nei confronti di, sempre possibili e spesso frequenti, "outliers" di qualsiasi tipo e natura, e di stabilità della soluzione rispetto all'inizializzazione; inoltre presenta una discreta velocità di convergenza degli algoritmi computazionali impiegati. Il problema della robustezza nei confronti di “outliers” è particolarmente importante; infatti occorre tener presente che, ad esempio, in un “cluster” concavo il cosidetto centroide (ovvero il “cluster point” del “cluster” in esame) potrebbe essere occupato da un altro “cluster”, come pure pochi elementi potrebberero costituire un punto di biforcazione fra la fusione di alcuni “cluster”, in un unico “cluster” più grande, e la divisione degli stessi “cluster” in tanti “cluster” più piccoli.

[6] Questo paragrafo riprende ed aggiorna quanto esposto in: Mussio L. (1984): Il metodo della regressione multipla e le sue applicazioni per l’analisi statistica dei risutati delle compensazioni. In: Ricerche di *Geodesia, Topografia e Fotogrammetria*, n. 4, CLUP. Milano, p. 275–303.

Il metodo della regressione lineare multipla consiste nel ricalcolare uno dei vettori, scelto e fissato, a partire da altri opportuni vettori. Trasportando il problema algebrico in termini geometrici, si vuole trovare, in un iperspazio a n dimensioni, un iperpiano la cui equazione sia in forma esplicita rispetto ad una determinata dimensione.

Chiamati, allora, y_k^0 (k=1,m) gli elementi di una variabile i*, osservati in condizioni di indipendenza e tutti con uguale precisione, e y_k (k=1,m) gli elementi della stessa variabile, ricalcolati a partire dagli elementi $x_k i$ (k=1,m; i=1,n; $i \neq i$*) delle altre variabili osservate, si ha l'equazione di regressione:

$$y_k = b_0 + \sum_{(i \neq i^*)} {}_i b_i x_{ki} \qquad \forall k$$

nella quale, essendo ignota la dipendenza funzionale deterministica, compaiono tutte le variabili, anche quelle che saranno scartate, durante il procedere del calcolo, perché non significative nello spiegare la variabile y. Introdotti gli scarti residui delle equazioni di regressione v_k (k=1,m) e posto, per confronto:

$$v_k = y_k - y_k^0 = b_0 + \sum_{(i \neq i^*)} {}_i b_i x_{ki} - y_k^0 \qquad \forall k$$

si ha un sistema di equazioni agli errori.

Sostituendo, poi, gli elementi delle variabili originarie con quelli delle variabili scarto: $x'_{ki} = x_{ki} - \bar{x}_i$ e $y'_k = y_k - \bar{y}$ (dove: $\bar{x}_i$ $\forall i$ e $\bar{y}_i$ sono le medie aritmetiche dei corrispondenti vettori delle osservazioni) e, successivamente, con quelli delle variabili scarto standardizzate: $x''_{ki} = x'_{ki} / \sigma_{x_i}$ e $y''_k = y'_k / \sigma_y$ (dove: σ_{x_i} $\forall i$ e σ_y sono gli sqm dei corrispondenti vettori delle osservazioni), si ottiene un nuovo sistema di equazioni agli errori:

$$w_k = \sum_{(i \neq i^*)} {}_i a_i x''_{ki} - y''_k \qquad \forall k$$

dove:

$$a_i = b_i \sigma_{xi} / \sigma_y \qquad \forall i \qquad (i \neq i^*)$$

Tale sistema, in generale, ridondante nel numero m di equazioni rispetto al numero n–1 di incognite, deve essere risolto, al solito, con il principio dei minimi quadrati:

$$\sum_k w_k^2 = min$$

In questo modo, si ottiene il sistema normale con n–1 equazioni in n–1 incognite:

$$\sum_k w_k x''_{ki} = 0 \qquad \forall i \qquad (i \neq i^*)$$

ovvero:

$$\sum_{(i \neq i^*)} {}_i\, a_i \sum_k x''_{ki}\, x''_{kj} = \sum_k y''_k x''_{kj} \qquad \forall \mathrm{j} \qquad (j \neq i^*)$$

I coefficienti: $\sum_k x''_{ki} x''_{kj}$ e $\sum_k y''_k x''_{kj}$, della matrice quadrata simmetrica del sistema normale nelle variabili scarto standardizzate, a meno di un fattore di proporzionalità, pari a $1/m$, sono i coefficienti di correlazione lineare semplice fra i vettori delle osservazioni r_{ij} (Fig.2.2).

	x_1	x_2	...	x_j	...	x_{n-1}	y
x_1	r_{11}	r_{12}	...	r_{1j}	...	$r_{1(n-1)}$	r_{1y}
x_2	r_{21}	r_{22}	...	r_{2j}	...	$r_{2(n-1)}$	r_{2y}
.	.	.	...	.	...	.	.
.	.	.	...	.	...	.	.
x_i	r_{i1}	r_{i2}	...	r_{ij}	...	$r_{i(n-1)}$	r_{iy}
.	.	.	...	.	...	.	.
.						.	.
x_{n-1}	$r_{(n-1)1}$	$r_{(n-1)2}$	...	$r_{(n-1)j}$	...	$r_{(n-1)(n-1)}$	$r_{(n-1)y}$
y	r_{y1}	r_{y2}	...	r_{yi}	...	$r_{y(n-1)}$	r_{yy}

Fig. 2.2

I valori di questi coefficienti, corretti per il fattore di proporzionalità, variano al solito in $[-1,1]$ ed i coefficienti posti lungo la diagonale principale hanno tutti valore $+1$. Supposta una correlazione lineare fra due qualsiasi variabili, la sua intensità è misurata dal quadrato del coefficiente di correlazione lineare semplice tra le due variabili prese in esame. Costruita la matrice dei coefficienti di correlazione lineare semplice, è possibile avviare il calcolo della regressione lineare multipla. Ogni scarto totale: $y_k^0 - \bar{y}$, può essere decomposto in due parti: una detta scarto spiegato con la regressione: $y_k - \bar{y}$, l'altra detta scarto residuo, imputabile ad errori di misura o modello: $y_k^0 - y_k$. Allora, anche, la somma (quadratica) degli scarti totali: S_T^2 , può essere decomposta nella somma (quadratica) degli scarti spiegati con la regressione: S_S^2 , e nella somma (quadratica) degli scarti residui, imputabili ad errori di misura o modello: S_R^2 , essendo il doppio prodotto nullo, per il principio dei minimi quadrati:

$$S_T^2 = \sum_k (y_k^0 - \bar{y})^2 =$$
$$= \sum_k \left(y_k^0 - y_k\right)^2 + \sum_k \left(y_k - \bar{y}\right)^2 + 2\sum_k \left(y_k^0 - y_k\right)\left(y_k - \bar{y}\right) =$$
$$= S_R^2 + S_S^2 - 2\sum_k v_k \sum_i b_i x'_{ki} = S_R^2 + S_S^2 - 2\sum_i b_i \sum_k x'_{ki} v_k = S_R^2 + S_S^2$$

Innanzitutto si scelgono i livelli di significatività, per operare con la regressione lineare multipla e si incomincia il procedimento che verrà, via via, descritto, dal livello di significatività, fra quelli prescelti. Durante il procedere del calcolo, si vuole rendere minimo S_R^2 (che dimostra la falsità del modello proposto, a spiegare la realtà in esame) e rendere massimo S_S^2 (che dimostra, invece, la capacità del modello proposto, a spiegare la stessa realtà in esame). Il metodo si avvia, ponendo: $S_R^2 = S_T^2$, ovvero: $y_k = \bar{y} \quad \forall k$. In questo caso, nessuna variabile è in regressione ($n' = 0$): S_R^2 è massimo, mentre S_S^2 è addirittura nullo. Al fine di rendere minimo, ad ogni passo (l):

$$S_R^2 = \sum_k \left(y_k^0 - y\right)^2 = \left(m - 1 - n'\right)\sigma_y^2 r_{yy}^{(l)} =$$
$$= \left(m - 1 - n'\right)\sigma_y^2 \, S^2{}_R^{(l)} \big/ S_T^2$$

si costruisce una matrice, formata dalla matrice dei coefficienti di correlazione lineare semplice, orlata con i coefficienti di correlazione lineare semplice: r_{iy}, r_{yi} e r_{yy}[7]. Su questa matrice, si procede compiendo l'operazione di inversione a passi, avendo come obiettivo quello di rendere, passo dopo passo, sempre più piccola la cifra di merito $r_{yy}^{(l)}$, minimizzando S_R^2.

Al primo passo d'inversione, essendo: $r_{hh} = 1$, si sottrae da r_{yy} il termine r_{hy}^2 massimo. Si inizia, pertanto, a spiegare S_T^2 con la variabile h per la quale è massimo, in valore assoluto, il coefficiente di correlazione lineare semplice con la variabile dipendente. Ad ogni successivo passo (l) d'inversione si procede, sempre, scegliendo la variabile h tale che sia massimo:

$$r_{yh}^{(l)} r_{hy}^{(l)} \big/ r_{hh}^{(l)} = \Delta S_S^2 \big/ S_T^2$$

Si osservi che il contributo dato dalle variabili che si presentano a spiegare S_T^2, non al primo passo d'inversione (se esse non sono indipendenti dalle precedenti), è inferiore al contributo che avrebbero dato, se

[7] L'operazione di inversione, eseguita a passi, sulla matrice dei coefficienti di correlazione lineare semplice fra i vettori delle osservazioni, fa sì che questi, sin dal primo passo, cessino di essere coefficienti di correlazione lineare semplice.
Dopo ciascun passo, questa matrice, tolta l'ultima riga e l'ultima e l'ultima colonna, costituisce, a meno di un fattore di proporzionalità e tenuto conto delle varianze delle variabili indipendenti e della variabile dipendente, la matrice di varianza–covarianza dei coefficienti delle equazioni di regressione, calcolati fino a quel passo. L'ultimo termine diagonale principale, moltiplicato per la varianza della variabile dipendente, rappresentata, invece, la somma (quadratica) degli scarti residui. Lo stesso ultimo termine diagonale principale è il complemento all'unità del coefficiente di correlazione (lineare) multipla, ad un dato passo.

fossero state scelte, già, al primo passo di inversione.

Ad ogni passo (l) di inversione, dopo aver selezionato il termine pivotale h, sempre sulla diagonale principale, in quanto, ad ogni passo, si vuole operare su una sola variabile, si procede, ad es., come mostrato nella Fig. 2.3 [8]:

el. pivotale (h, h)	$r_{hh}^{(l+1)} = 1 / r_{hh}^{(l)}$	
el. della stessa riga (h,.)	$r_{hj}^{(l+1)} = r_{hj}^{(l)} / r_{hh}^{(l)}$	$r_{hy}^{(l+1)} = r_{hy}^{(l)} / r_{hh}^{(l)}$
el. della stessa colonna (.,h)	$r_{ih}^{(l+1)} = -r_{ih}^{(l)} / r_{hh}^{(l)}$	$r_{yh}^{(l+1)} = -r_{yh}^{(l)} / r_{hh}^{(l)}$
altri elementi (.,.)	$r_{ij}^{(l+1)} = r_{ij}^{(l)} - r_{ih}^{(l)} r_{hj}^{(l)} / r_{hh}^{(l)}$	$r_{iy}^{(l+1)} = r_{iy}^{(l)} - r_{ih}^{(l)} r_{hy}^{(l)} / r_{hh}^{(l)}$
		$r_{yj}^{(l+1)} = r_{yj}^{(l)} - r_{yh}^{(l)} r_{hj}^{(l)} / r_{hh}^{(l)}$
cifra di merito		$r_{yy}^{(l+1)} = r_{yy}^{(l)} - r_{yh}^{(l)} r_{hy}^{(l)} / r_{hh}^{(l)}$

Fig. 2.3

Ad ogni passo di inversione, si selezionano due variabili, rispettivamente, candidate ad entrare nel numero delle variabili in regressione e ad uscirne[9]. Ovviamente, la variabile candidata ad entrare, viene scelta fra quelle non ancora, entrate in regressione, mentre la variabile candidata ad uscire, viene scelta fra quelle già entrate in regressione. Al primo passo, si ha solo una variabile candidata ad entrare, essendo vuoto, evidentemente, l'insieme delle variabili già entrate in regressione. Applicare due volte, anche non consecutivamente, l'operazione di inversione con lo stesso termine pivotale, significa non compiere alcun passo di inversione, essendo detta operazione, come ben noto, involutoria.

Secondo il criterio di rendere $r_{yy}^{(l+1)}$ minimo, essendo: $r_{yy}^{(l+1)} = r_{yy}^{(l)} - r_{yh}^{(l)} r_{hy}^{(l)} / r_{hh}^{(l)}$, la variabile candidata ad entrare nel numero delle variabili in regressione ha: $r_{yh}^{(l)} r_{hy}^{(l)} / r_{hh}^{(l)} = \Delta S_S^2 / S_T^2$, massimo e quella candidata ad uscirne ha: $r_{yh}^{(l)} r_{hy}^{(l)} / r_{hh}^{(l)} = \Delta S_S^2 / S_T^2$, minimo.

Su queste variabili, si effettua il test statistico per il rapporto di varianze indipendenti, altrimenti detto analisi di varianza ad una via, il cui comportamento segue la distribuzione F di Fisher. Ricavato dalle tavole delle

[8] Le espressioni fornite per il calcolo dell'inversione a passi sono, proprio, quelle di eliminazione di Gauss, da eseguirsi, tuttavia, selezionando, ad ogni passo, il termine pivotale, con l'obiettivo di rendere sempre più piccola la cifra di merito, anziché percorrendo, al solito, la diagonale principale dall'alto al basso. Questo modo di procedere è detto, eliminazione di Gauss a passi selezionati; analogamente sarebbe stato possibile operare, ad es. mediante fattorizzazione di Cholesky a passi selezionati. Tuttavia, poiché il sistema normale ottenuto da una regressione lineare multipla è, sempre, molto piccolo ($n < 100$) l'innegabile semplicità dell'eliminazione di Gauss, fa optare, di solito, per questo modo di procedere, anziché per gli altri, certamente più complessi nella loro formulazione. Tutto ciò dà ragione, inoltre, della scelta di presentare il metodo della regressione lineare multipla (come tutti gli altri problemi dell'analisi multivariata), in termini algebrici, anziché in termini matriciali, come usuale nei problemi minimi quadrati.

[9] Le espressioni, variabile entrata nel numero delle variabili in regressione e variabile uscita dal numero delle variabili in regressione, significano, rispettivamente che una certa variabile dà o non da più un contributo significativo alla stima della variabile dipendente. L'espressione: variabile candidata ad entrare o ad uscire, indica che una certa variabile, sottoposta ad un opportuno test statistico, sta per dare oppure sta per non dare più un contributo significativo alla stima della variabile dipendente. L'insieme delle variabili entrate in regressione è, ovviamente, un sottoinsieme dell'insieme di dimensione $n-1$ delle variabili indipendenti, avente numerosità: $n' \leq n-1$.

distribuzioni di probabilità il valore F teorico, ad un dato livello di probabilità α, per $\nu_1 = 1$ e $\nu_2 = m - 1 - n'$ gradi di libertà (essendo n' il numero delle variabili in regressione), è possibile procedere al confronto, sia nel caso di uscita, sia nel caso di entrata, fra tale valore ed il valore F empirico, opportunamente, costruito.

Nel caso di proposta di uscita di una variabile, del numero delle variabili in regressione, si ha quale valore empirico[10]:

$$F^{(l+1)} = \upsilon_2 (-r_{yh}^{(l)} r_{hy}^{(l)} / r_{hh}^{(l)}) / r_{yy}^{(l)} =$$
$$= \upsilon_2 (\Delta S_S^2 / S_T^2) / (S_R^2 / S_T^2) = \upsilon_2 \Delta S_S^2 / S_R^2$$

Se tale valore è minore del valore F teorico, la variabile in esame esce dal numero delle variabili in regressione, si esegue un passo di inversione e si inizia il passo successivo; se, invece, è maggiore, la variabile in esame non esce dal numero delle variabili in regressione e si affronta il caso di proposta di entrata.

Nel caso di proposta di entrata di una variabile, nel numero delle variabili in regressione, si ha quale valore empirico:

$$F^{(l+1)} = (\upsilon_2 - 1)(r_{yh}^{(l)} r_{hy}^{(l)} / r_{hh}^{(l)}) / (r_{yy}^{(l)} - r_{yh}^{(l)} r_{hy}^{(l)} / r_{hh}^{(l)}) =$$
$$= (\upsilon_2 - 1)(\Delta S_S^2 / S_T^2) / (S_R^2 / S_T^2) = (\upsilon_2 - 1) \Delta S_S^2 / S_R^2$$

Se tale valore è maggiore del valore F teorico, la variabile in esame entra nel numero delle variabili in regressione, si esegue un passo di inversione e si inizia il passo successivo; se, invece, è minore, la variabile in esame non entra nel numero delle variabili in regressione e si ricomincia lo stesso procedimento sequenziale, ad un livello di probabilità inferiore.

Ogni variabile che entra nel numero delle variabili in regressione, toglie un grado di libertà al sistema; ogni variabile che ne esce, rende un grado di libertà al sistema.

I termini della diagonale principale, inizialmente, uguali a uno tendono, durante l'operazione d'inversione, a diminuire fino, al limite, zero. L'introduzione di una successiva variabile h è inibita, quando il termine $r_{hh}^{(l)}$ diventa minore, ad es., di 10^{-5}, dando luogo a notevoli imprecisioni numeriche. La variabile h è, in tal caso, fortissimamente, correlata con l'insieme delle variabili, già, in regressione.

L'inversione è bloccata:

- per il raggiungimento del minimo livello di probabilità prefissato;
- per l'esaurimento dei gradi di libertà (oppure, e meglio, la perdita dell'affidabilità delle osservazioni);

[10] Si osservi che il caso di proposta di entrata di una variabile, nel numero delle variabili in regressione, viene affrontato come il caso di proposta di uscita di una variabile, dal numero delle variabili in regressione, supponendo la stessa variabile come, già, entrata. Allora se il test di Fisher fallisce la variabile esce (cioè non–entra), mentre se passa la variabile non–esce (cioè entra).
Si noti che, infatti, l'espressione: si esegue un passo di inversione è proprio la stessa, nel caso di entrata di una variabile, nel numero delle variabili in regressione, come nel caso di uscita di una variabile, dal numero delle variabili in regressione. Infatti effettuato il test di Fisher, le successive operazioni di calcolo matriciale, necessarie per eseguire un passo d'inversione, sono, aritmeticamente, del tutto identiche nei due casi. Pertanto, come già detto, applicare due volte, anche non consecutivamente, l'operazione di inversione con lo stesso termine pivotale, significa non compiere alcun passo d'inversione.

- perché r_{yy} assume valori minori, ad es., di 10^{-30} (in tal caso, risultando, di conseguenza, S_R^2, pressoché, nullo e $S_S^2 \cong S_T^2$, non si ha, proprio, più nulla da spiegare);
- per la deriva di S_R^2 (imprecisione numerica, dilagante nella matrice, a bassi livelli di probabilità), sospettata con un determinante minore ad es., di 10^{-10}, oppure per un subitaneo abbattimento di $r_{yy}^{(l)}$, ad es., maggiore del 99%, confermato da uno scostamento, ad es., maggiore del 5%, tra S_R^2 e l'analogo valore calcolato a posteriori.

Per garantirsi, invece, contro una inopportuna spiegazione fasulla, è conveniente introdurre, fra i vettori delle osservazioni, anche, un vettore di osservazioni fittizie, costruito eseguendo estrazioni di numeri casuali; se questa variabile fasulla contribuisce, in qualche modo, a spiegare S_T^2, allora la regressione eseguita costituisce, solo, un gioco numerico.

Arrestata l'inversione della matrice al passo (l), è possibile calcolare la soluzione del sistema normale, avendo già effettuato la sostituzione in avanti, contemporaneamente alla eliminazione di Gauss, ed effettuato, ora, quella all'indietro, utilizzando il fattore triangolare superiore, ottenuto eseguendo, proprio, la eliminazione di Gauss. Si badi che entrambe le operazioni sono eseguite prendendo in considerazione, automaticamente, solo, gli elementi delle matrici corrispondenti alle variabili entrate in regressione.
La soluzione ottenuta fornisce i coefficienti a_i e gli elementi della matrice parzialmente invertita che, moltiplicati al solito per sigma zero, ovvero la media quadratica degli scarti residui delle equazioni agli errori, ne forniscono le varianze $\sigma_{a_i}^2$ e covarianze $\sigma_{a_i a_j}$. Da questi risultati, si ricavano i coefficienti della equazione di regressione, inoltre le loro varianze: $\sigma_{b_j}^2 = \sigma_{a_j}^2 \sigma_y^2 / \sigma_{x_i}^2$, e covarianze: $\sigma_{b_i b_j}^2 = \sigma_{a_i a_j}^2 \sigma_y^2 \big/ \sigma_{x_i} \sigma_{x_j}$ (Altre statistiche, ad es., coefficienti di correlazione, ridondanze locali, numero di condizione, sono calcolate, a partire dalle prime, se del caso.). Successivamente, si ottengono, nell'ordine:

(a) gli elementi del vettore ricalcolato, funzione della media degli elementi dello stesso vettore misurato e dei coefficienti di tutte le variabili entrate in regressione:

$$y_k = \bar{y} + \sum_i b_i \left(x_{ki} - \bar{x}_i \right) \qquad \forall k$$

(b) la varianza di ciascun elemento, funzione della varianza della media degli elementi dello stesso vettore misurato, inoltre della varianza e delle covarianze dei coefficienti di tutte le variabili entrate in regressione:

$$\sigma_{y_k}^2 = \sigma_{\bar{y}}^2 + \sum_i \sigma_{b_i}^2 \left(x_{ki} - \bar{x}_i \right)^2 + 2 \sum_{j>i}{}_{ij} \sigma_{b_i b_j} \left(x_{ki} - \bar{x}_i \right) \left(x_{kj} - \bar{x}_j \right) \qquad \forall k$$

(c) i corrispondenti intervalli fiduciari (Questi individuano una fascia la cui ampiezza è determinata dalla distribuzione t di Student tale che, fissato un certo livello di significatività α, nell'$1-\alpha$ dei casi un valore

osservato della variabile dipendente, per determinati valori delle variabili indipendenti, sia interno ad essa. Il valore t teorico si ricava dalle tavole delle distribuzioni di probabilità, per $\upsilon = m - 1 - n'$ gradi di libertà, essendo n' il numero delle variabili in regressione.)[11]:

$$y_k^{+/-} = y_k \pm t_{\alpha/2}\sigma_{y_k}$$

Si osservi, infine, come l'eventuale adozione di una **procedura robusta** comporti l'esecuzione di un certo numero di ripetizioni del calcolo di una regressione lineare multipla, modificando i pesi delle osservazioni, ad ogni iterazione, in funzione degli scarti residui delle equazioni agli errori, al passo precedente, e della norma prescelta. Infatti mentre la regressione lineare multipla, risolta adottando il principio dei minimi quadrati, presenta usualmente tutte le osservazioni allo stesso peso, l'adozione di una procedura robusta richiede, ad ogni iterazione, la definizione di pesi ausiliari (pur con osservazioni inizialmente equiponderate), capaci di rincondurre la singola iterazione proprio al principio dei minimi quadrati.
L'insieme dei risultati è, poi, corredato da un'opportuna documentazione statistica[12].

(a) coefficiente di correlazione (lineare) multipla:

$$R^{(1)} = S_S^2 / S_T^2 = l - S_R^2 / S_T^2 = 1 - r_{yy}^{(1)}$$

(b) indice di determinazione (I due indici variano da zero a uno e crescono con la maggiore capacità del modello proposto a spiegare S_T^2; in particolare $R^{2(1)}$ lo spiega percentualmente.): $R^{2(1)} = (R^{(1)})^2$

(c) coefficiente di correlazione parziale (Questo coefficiente, fra la variabile dipendente ed ogni variabile indipendente, a quadrato, mostra, in percentuale, il contributo offerto da ciascuna variabile h a spiegare

---

[11] Data una statistica campionaria, è possibile determinare un **intervallo fiduciario**, tale che, se la statistica teorica della popolazione è compresa nel suddetto intervallo, la statisitca campionaria può considerarsi estratta da questa popolazione, con un prefissato livello di significatività.
Di seguito, si riportano gli intervalli fiduciari relativi alle stime delle statisitche campionarie della media (per campioni numerosi e per piccoli campioni normali), della varianza e del coefficiente di correlazione (nell'ipotesi fondamentale nulla), ove ogni intervallo è ricavato a partire dal corrispondente test d'accettazione di un'ipotesi fondamentale, avendo scelto e fissato un determinato livello di significatività.

Test per media (campioni numerosi):	$\frac{\lvert\bar{x}-\mu\rvert}{\sigma/\sqrt{n}} \le z_t^{(+)}$	Intervallo fiduciario:	$\bar{x} - z_t^{(+)}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_t^{(+)}\frac{\sigma}{\sqrt{n}}$
Test per media (campioni normali):	$\frac{\lvert\bar{x}-\mu\rvert}{\hat{\sigma}/\sqrt{n}} \le t_t^{(+)}$	Intervallo fiduciario:	$\bar{x} - t_t^{(+)}\frac{\hat{\sigma}}{\sqrt{n}} \le \mu \le \bar{x} + t_t^{(+)}\frac{\hat{\sigma}}{\sqrt{n}}$
Test per la varianza:	$\chi_t^{2(-)} \le \frac{(n-1)\hat{\sigma}^2}{\sigma^2} \le \chi_t^{2(+)}$	Intervallo fiduciario:	$\frac{(n-1)\hat{\sigma}^2}{\chi_t^{2(+)}} \le \sigma^2 \le \frac{(n-1)\hat{\sigma}^2}{\chi_t^{2(-)}}$
Test per il coefficiente di correlazione:	$\frac{\lvert\hat{r}\rvert}{\sqrt{1-\hat{r}^2}/\sqrt{n-2}} \le t_t^{(+)}$	Intervallo fiduciario:	$\hat{r} - t_t^{(+)}\frac{\sqrt{1-\hat{r}^2}}{\sqrt{n-2}} \le \rho \le \hat{r} + t_t^{(+)}\frac{\sqrt{1-\hat{r}^2}}{\sqrt{n-2}}$

[12] La medesima documentazione statistica, calcolata utilizzando i risultati ottenuti, ad ogni generico passo (l) d'inversione, dà utili informazioni sulla capacità del modello proposto a spiegare, passo dopo passo, la realtà in esame.

S_T^2 , nel modello proposto.):

$$R_{yh}^{(1)} = -\Delta S_S^{2(1)} \Big/ S_R^{2\,(1)} = (-r_{yh}^{(1)} r_{hy}^{(1)} \Big/ r_{hh}^{(1)}) \Big/ r_{yy}^{(1)}$$

(d) sigma zero, o errore standard della variabile dipendente (Si osservi che, prima della regressione, lo stesso σ_0 vale: $\sigma_0 = \sqrt{S_T^2 \big/ (m-1)}$, ed è, in generale, ben maggiore. La significatività del suo abbattimento può essere, facilmente, verificata: fissato un certo livello di significatività α, eseguendo il test statistico χ^2. Il valore χ^2 teorico si ricava dalle tavole delle distribuzioni di probabilità, per $\upsilon = m-1-n'$ gradi di libertà, essendo n' il numero delle variabili in regressione.):

$$\sigma_0 = \sqrt{\frac{S_R^2}{m-1-n'}} = \sqrt{\frac{\sum_k v_k^2}{m-1-n'}}$$

Si noti come si sia, ben volutamente, scelto di dilungarsi nella presentazione dei risultati forniti e della documentazione di corredo per un duplice motivo.

- Storicamente la regressione multipla è uno dei primi metodi statistici proposti, per l'analisi multivariata; pertanto proprio nell'occasione della sua formulazione, particolare attenzione è stata prestata anche alla presentazione dei risultati forniti e della documentazione di corredo.
- Inoltre la regressione multipla ha fatto da battistrada ad altre tecniche statistiche dell'analisi multivariata, quali:

  - la "cluster nalysis" (già presentata nel paragrafo precedente del presente lavoro);
  - l'analisi di varianza (che sarà esposte nella seconda parte del presente lavoro);
  - le stime di covarianza (di cui, di certo, si troverà modo per parlare diffusamente altrove);
  - l'analisi fattoriale (di cui ancora si troverà modo per parlare brevemente altrove).

  Allora la presentazione dei risultati forniti e della documentazione di corredo della regressione multipla serve anche per illustrare quanto deve essere riportato come presentazione dei risultati forniti e della documentazione di corredo di tutte queste altre tecniche statistiche.

La Fig. 2.4 illustra uno schema a blocchi particolareggiato con riferimento all'analisi della varianza.

La **regressione non lineare multipla** costituisce un caso più generale, se non è possibile dare espressione analitica al legame funzionale evidenziato in diversi problemi, supponendo una dipendenza lineare fra le varie grandezze, perché questa dipendenza è non–lineare. Allora, occorre sempre a partire dai dati disponibili, prima scoprire il tipo di non–linearità della dipendenza da individuarsi, poi effettuare la stima dei parametri di una regressione non–lineare multipla.

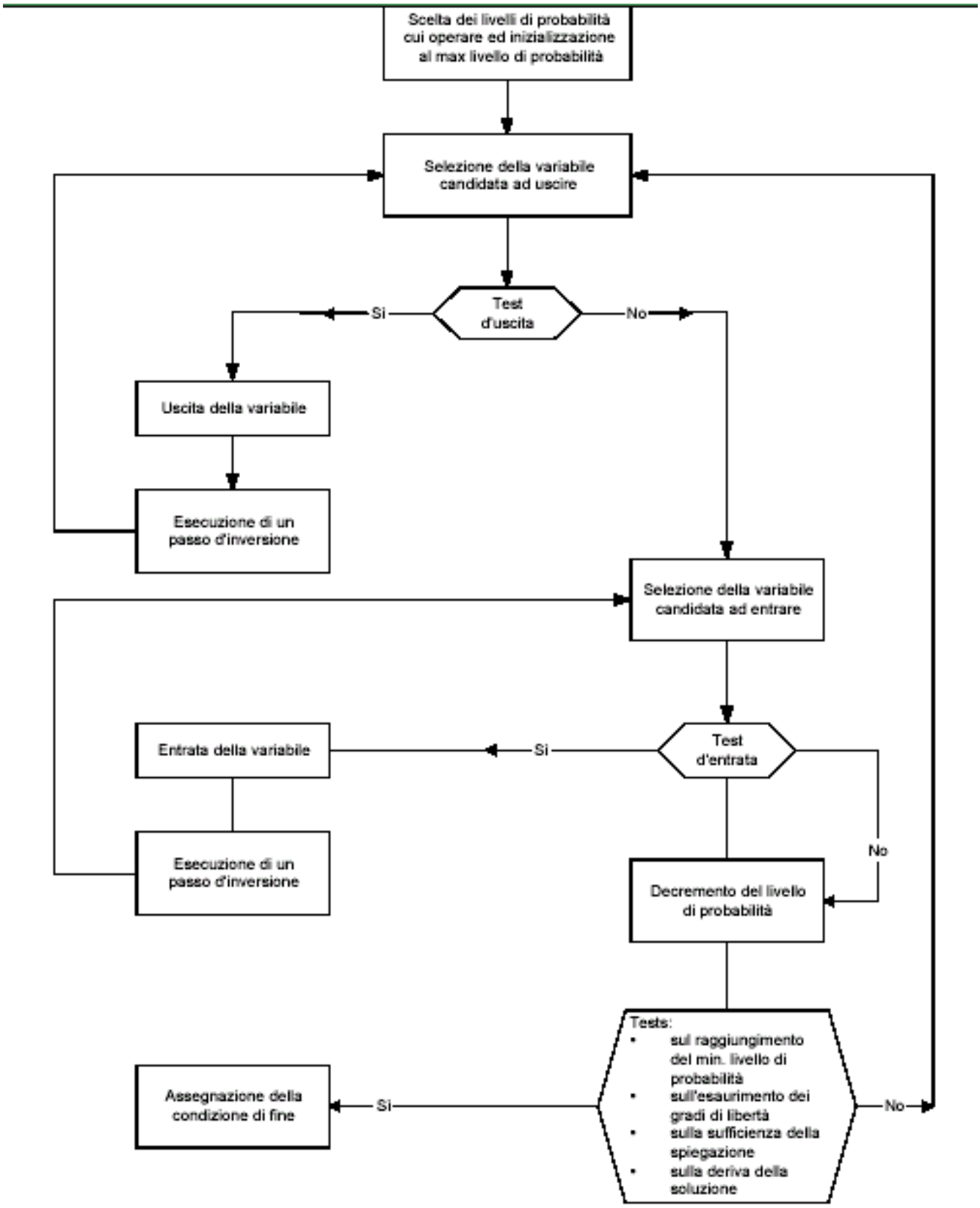


Fig. 2.4

Un'analisi preliminare della non–linearità della dipendenza per ogni coppia di variabili, ad es. x y, studia la forma di tale dipendenza: connessione, dipendenza funzionale (o regressione non–lineare), correlazione (o regressione lineare). La Fig. 2.5 illustra, a partire da una tabella a doppia entrata, la costruzione degli indici di

Pearson e di Bonferroni, nonché del coefficiente di correlazione (lineare semplice), indicatori, rispettivamente, della dipendenza funzionale (o regressione non–lineare), connessione e correlazione (o regressione lineare).

X / Y	y_1	y_2	$\cdots$	y_j	$\cdots$	y_n	*frequenze marginali*	*medie per riga*	*varianze per riga*
x_1	f_{11}	f_{12}	$\cdots$	f_{1j}	$\cdots$	f_{1n}	q_1	$\bar{y}_{x_1}$	$\sigma^2_{x_1}$
x_2	f_{21}	f_{22}	$\cdots$	f_{2j}	$\cdots$	f_{2n}	q_2	$\bar{y}_{x_2}$	$\sigma^2_{x_2}$
.	.	.	.	.	.	.	.	.	.
x_i	f_{i1}	f_{i2}	$\cdots$	f_{ij}	$\cdots$	f_{in}	q_i	$\bar{y}_{x_i}$	$\sigma^2_{x_i}$
.	.	.	.	.	.	.	.	.	.
x_m	f_{m1}	f_{m2}	$\cdots$	f_{mj}	$\cdots$	f_{mn}	q_m	$\bar{y}_{x_m}$	$\sigma^2_{x_m}$
frequenze marginali	p_1	p_2	$\cdots$	p_j	$\cdots$	p_n	N	$\sigma^2_{\bar{y}_x}$	$E_y(\sigma^2_x)$
medie per colonna	$\bar{x}_{y_1}$	$\bar{x}_{y_2}$	$\cdots$	$\bar{x}_{y_j}$	$\cdots$	$\bar{x}_{y_n}$	$\sigma^2_{\bar{x}_y}$		
varianze per colonna	σ_{y_1}	σ_{y_2}	$\cdots$	σ_{y_j}	$\cdots$	σ_{y_n}	$E(\sigma^2_y)$		

Legenda: X , Y — *variabili*

x_i , y_j — *valori argomentali*

f_{ij} , $c_{ij} = f_{ij} - q_i p_j$ — *frequenze doppie e contingenze*

Indici di Pearson:

$$\eta^2_x = \frac{\sigma^2_{\bar{x}_y}}{E_x(\sigma^2_y) + \sigma^2_{\bar{x}_y}} = \frac{\sigma^2_{\bar{x}_y}}{\sigma^2_x}$$

monolaterale: $x = g(y)$

$$\eta^2_y = \frac{\sigma^2_{\bar{y}_x}}{E_y(\sigma^2_x) + \sigma^2_{\bar{y}_x}} = \frac{\sigma^2_{\bar{y}_x}}{\sigma^2_y}$$

monolaterale: $y = h(x)$

$$\eta^2 = \frac{\sigma^2_x \eta^2_x + \sigma^2_y \eta^2_y}{\sigma^2_x + \sigma^2_y}$$

bilaterale $y = h(x)$; $x = g(x)$

Indice di Bonferroni:

$$\beta_x = \frac{1/2 \sum_{ij} |c_{ij}|}{1 - \sum_i p_i^2}$$

monolaterale: $x = g(y)$

$$\beta_y = \frac{1/2 \sum_{ij} |c_{ij}|}{1 - \sum_j q_j^2}$$

monolaterale: $y = h(x)$

$$\beta_0 = \sqrt{\beta_x . \beta_y} = \frac{1/2 \sum_{ij} |c_{ij}|}{\sqrt{(1-\sum_i p_i^2)(1-\sum_j q_j^2)}} \qquad \textit{bilaterale geometrico}$$

$$\beta_{-1} = \frac{2}{1/\beta_x + 1/\beta_y} = \frac{\sum_{ij} |c_{ij}|}{2-\sum_i p_i^2 - \sum_j q_j^2} \qquad \textit{bilaterale armonico}$$

Coefficiente di correlazione (*lineare semplice*): $\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{E(xy) - \overline{xy}}{\sigma_X \sigma_Y}$

Fig. 2.5 [13]

La struttura dei dati per il calcolo della regressione non–lineare multipla, è costituita da una tabella avente n colonne di vettori delle osservazioni e m ($m \geq n$) righe di elementi di ciascun vettore. Ancora questi dati formano una matrice rettangolare da cui, scelto e fissato uno di questi vettori che si vuole ricalcolare, a partire da altri opportuni, è possibile ricavare i coefficienti ed i termini noti di un sistema di equazioni agli errori. Da queste, dopo aver sostituito le variabili originarie con le variabili scarto e standardizzato queste ultime, utilizzando il principio dei minimi quadrati si ottiene, al solito, un sistema normale la cui corrispondente matrice ha per elementi i coefficienti di correlazione lineare semplice fra i vettori delle osservazioni. Costruita la matrice dei coefficienti di correlazione lineare semplice, è possibile iniziare il metodo della regressione lineare multipla.

Il metodo della regressione non–lineare multipla prevede una lunga sequenza di operazioni. Dapprima si ha, proprio, il calcolo di n–1 regressioni lineari multiple, per ciascuno dei vettori delle variabili indipendenti su tutti gli altri, meno il vettore della variabile dipendente che si vuole ricalcolare[14]:

$$x_{ki} = b_0^{'i} + \sum_{(j \neq i, j \neq i^*)} {}_j b_j^{'i} x_{kj} \qquad \forall k \qquad \text{dove: } i = 1,n \ (i \neq i^*: \ x_{ki^*} = y_k)$$

e di altre n–1 regressioni lineari multiple, sempre, per il vettore della variabile dipendente che si vuole ricalcolare sui vettori delle indipendenti, meno, di volta in volta, uno di questi:

$$y_k^i = b_0^{''i} + \sum_{(j \neq i, j \neq i^*)} {}_j b_j^{''i} x_{kj} \qquad \forall k \qquad \text{dove: } i = 1,n \ (i \neq i^*: \ x_{ki^*} = y_k)$$

Dopiché avendo calcolato, per ogni regressione lineare multipla, gli scarti residui delle equazioni agli errori:

13 Un'analisi preliminare della non–linearità della dipendenza per ogni ennupla di variabili, studia le forme di tale dipendenza costruendo, a partire da tabelle a multiple entrate, indici multipli di Pearson e di Bonferroni, come già detto, indicatori, rispettivamente, della dipendenza funzionale e della connessione. Tuttavia, poiché, come già detto, tutto ciò che supera le due (o al massimo tre) dimensioni non è facilmente rappresentabile, l'innegabile complicazione di questo studio lo fa, di solito, omettere.

14 Si osservi che le espressioni: variabili indipendenti e variabile dipendente, hanno significato, solo, in senso algebrico e geometrico, ma non in senso statistico: infatti tutte le variabili sono, in generale, in vario modo, correlate tra loro, come si evince, ad esempio, dalla matrice dei coefficienti di correlazione lineare semplice fra i vettori delle osservazioni.

$$v_{ki} = x_{ki} - x_{ki}^0 \qquad \forall k \qquad (i \neq i^*)$$

$$v_k^i = y_k^i - y_k^0 \qquad \forall k \qquad (i \neq i^*)$$

si devono effettuare delle trasformazioni elementari (del tipo: $x^2,\ x^3,\ e^x - 1,\ \sqrt{x},\ \sqrt[3]{x},$ $ln(x+1),\ 1/(x+1)-1,\ 1/(x^2+1)-1,\ 1/\sqrt{x+1}-1,\ \ sin\ x,\ cos\, x,$ ecc.) sugli scarti–residui delle variabili indipendenti[15]. Si calcola il coefficiente di correlazione lineare semplice fra questi e gli scarti residui della variabile dipendente, cercando, per tentativi, la trasformazione elementare che dà il coefficiente di correlazione lineare semplice massimo. Una volta trovata, si applica quella stessa trasformazione elementare a tutti gli elementi x_{ki} della variabile in esame.

La ragione di questo modo di procedere sta nel fatto che le variabili x_{ki} non sono, in generale, statisticamente indipendenti fra loro. Pertanto per trovare la trasformazione elementare che dà il massimo coefficiente di correlazione lineare semplice, fra una certa variabile dipendente e la variabile indipendente, è necessario depurare entrambe le variabili dall'influenza di tutte le altre. Di fatto, la depurazione avviene, solo, per quanto riguarda una dipendenza lineare. Infatti una depurazione comprendente, anche, dipendenze non–lineari richiederebbe, a sua volta, la stessa lunga sequenza di operazioni e, in questo modo, si potrebbe procedere senza fine.

Infine si effettua il calcolo di una regressione lineare multipla per il vettore che si vuole ricalcolare su tutti gli altri, avendo operato, sugli elementi x_{ki} di ciascun vettore, la trasformazione elementare, precedentemente, individuata per la variabile i. Al termine di quest'ultima regressione lineare multipla, si ha poi l'insieme dei risultati, corredato da un'opportuna documentazione statistica.

Si osservi, come questo modo di eseguire il calcolo di una regressione non–lineare multipla cada in difetto, quando la trasformazione di una variabile non sia esprimibile, in forma esplicita, rispetto alla variabile dipendente. In tal caso, è necessario operare con funzioni a più valori, con eventuali punti singolari, cosicché non è possibile procedere alla ricerca, per tentativi, di trasformazioni elementari delle variabili.

Allora occorre scoprire, per altra via, il tipo di non–linearità della dipendenza cercata, come, per altra via, occorre calcolare i valori approssimati dei coefficienti di un'espressione analitica linearizzata.

- L'utilizzo di adatte rappresentazioni grafiche, per lo più in ambito locale (grafi, sezioni e profili, proiezioni, viste 3D, ecc),
- il calcolo più regressioni lineari multiple (comunemente dette sotto–regressioni), dopo aver, opportunamente, effettuato la suddivisione dell'insieme dei dati in più sottoinsiemi, dove l'approssimazione lineare sia, sufficientemente, accettabile,

rispondono, positivamente, alle esigenze poste da queste elaborazioni preliminari. Successivamente dopo aver sostituito, utilizzando i valori approssimati dei coefficienti, i vettori dei dati disponibili con quelli delle loro

[15] E', altresì, possibile effettuare combinazioni algebriche di trasformazioni elementari. Si badi, tuttavia, che trasformazioni complesse, ancorché difficili da individuare, spesso non danno molta più informazione di una semplice trasformazione elementare; pertanto una trasformazione complessa deve essere costruita, solo, in base a qualche conoscenza, a priori, sulla natura del fenomeno, fra le cui grandezze si sta operando una regressione non–lineare multipla.

correzioni, basta eseguire il calcolo di una regressione lineare multipla, risolvendo un sistema linearizzato di equazioni agli errori.

3. L'INFERENZA STATISTICA MULTIVARIATA

L'inferenza statistica (multivariata) è quella parte dell'analisi multivariata dedicata alla costruzione e verifica d'ipotesi (per problemi di controllo di qualità, oppure controllo e confronto d'ipotesi di altri problemi). Infatti il confronto probabilistico permette di formulare un giudizio critico sui risultati ottenuti, nelle varie tappe dell'analisi multivariata. Così con le usuali strategie dell'inferenza statistica (Fig. 3.1), vari test multipli consentono di discriminare, tanto stime di parametri da campioni normali, quanto statisitche di modelli non – parametrici (ovvero modelli distribution free).

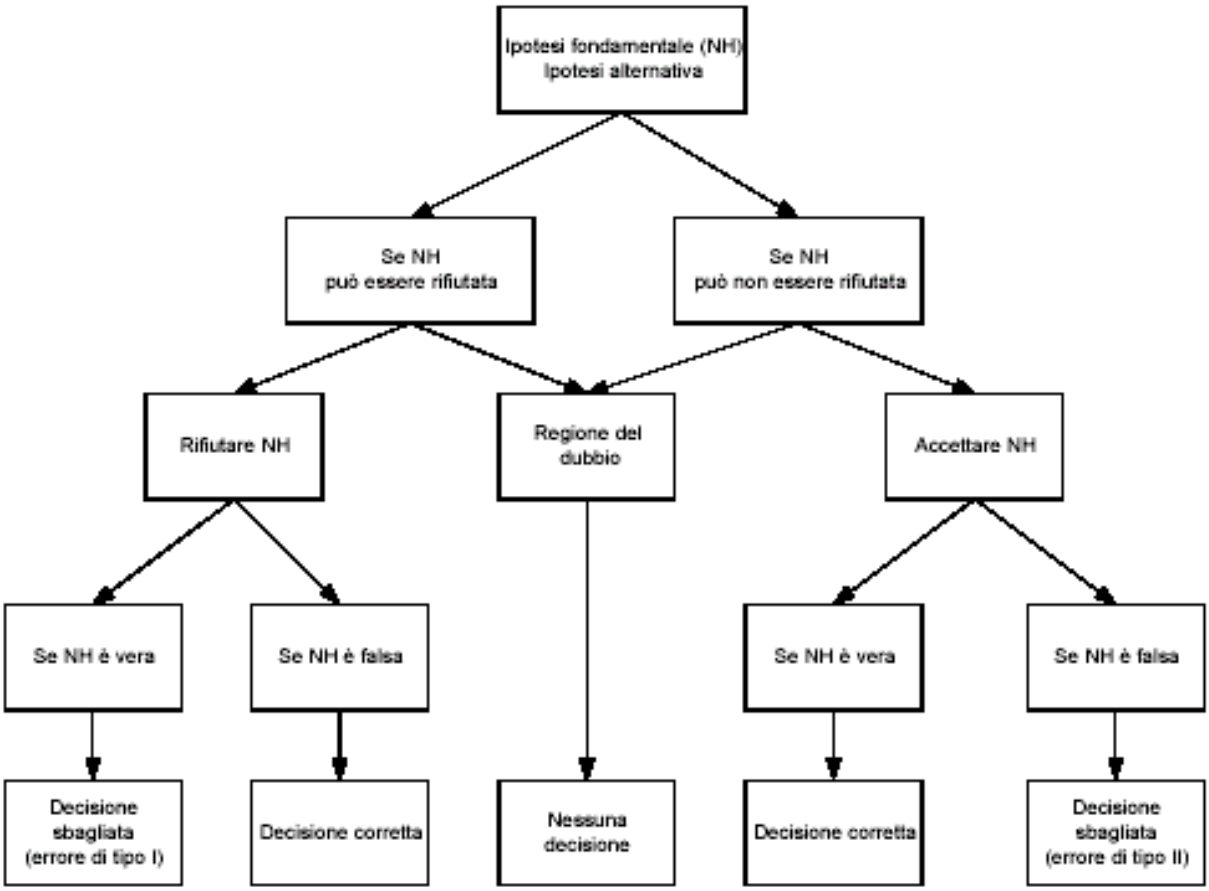


Fig. 3.1 [16].

[16] L'aggettivo non parametrico (in letteratura inglese: distribution free) qualifica un particolare gruppo di test statistici, sotto certe condizioni, sostitutivo dei test statistici classici. Infatti i test non–parametrici, rispetto ai test classici, presentano i seguenti vantaggi:

- la loro comprensione è immediata ed elementare;
- le condizioni di validità sono meno forti (più ampie);
- i calcoli necessari non presentano in generale difficoltà computazionali.

D'altra parte, i test non–parametrici presentano alcuni svantaggi:

- molta informazione viene sprecata;
- la potenza del test è bassa.

Test poco potenti tendono ad essere troppo conservativi, cioè l'ipotesi fondamentale (o nulla) è accettata anche quando dovrebbe valere l'ipotesi alternativa. Pertanto i test statistici classici sono preferibili, quando le condizioni di validità sono soddisfatte.

Gli oggetti del giudizio critico sono, come detto, i risultati ottenuti nelle varie tappe dell'analisi multivariata, in particolare: frequenze relative, contingenze, medie campionarie o altri indicatori del centro (di una popolazione), varianze campionarie o altri indicatori della dispersione, coefficienti di correlazione campionari o altri indicatori della correlazione[17].

I test multipli si differenziano per i diversi oggetti in esame, come pure per la distribuzione di appartenenza della popolazione cui i campioni si riferiscono, se normale, oppure altra (spesso sconosciuta). In generale, comunque, tutti i campioni sono supposti fra loro indipendenti, mentre se quest'ipotesi non è soddisfatta, occorre procedere, come per i modelli non – parametrici, adottando strategie assolutamente generali, ma assai poco potenti, cioè meno capaci di discriminare fra ipotesi alternative vicine, a parità di numerosità dei campioni. Nelle seguenti tabelle, si presentano parecchi test multipli, diversamente, rispondenti alla bisogna, facendo attenzione, in particolare, ai problemi pratici, legati alla loro applicazione ad esempi concreti. Per quanto riguarda, invece, i loro fondamenti teorici, questi si richiamano, in generale:

- alla definizione assiomatica di probabilità;
- alla legge dei grandi numeri ed al limite centrale della statistica;
- ai teoremi della normalità: conservazione per trasformazioni lineari, identità fra indipendenza ed incorrelazione;
- alla definizione delle variabili casuali χ^2 (chi quadrato), t (t di Student) e F (F di Fisher);
- al calcolo di probabilità estremali (ad es. di Kolmogorov–Smirnov e di Hawkins), ove necessario;
- al teorema di decomposizione ortogonale degli scarti;

e nello specifico, alla trasformazione della distribuzione di probabilità, caso per caso, per ogni determinato test multiplo. In ogni caso, poiché trattasi di problemi teorici propri del Calcolo delle Probabilità, tutto quanto riguarda i fondamenti teorici dei test multipli è considerato estraneo agli scopi del presente lavoro.

A complemento di quanto sopra esposto, la Fig. 3.2 mostra la famiglia delle curve di potenza della variabile casuale χ^2 (chi quadrato), in funzione di differenti rapporti di scala dell'ascissa a sua volta rappresentata, tanto a scala naturale, quanto in scala logaritmica, e la Fig. 3.3 mostra la famiglia degli intervalli fiduciari del coefficiente di correlazione, in funzione di differenti valori di numerosità dei campioni, avendo scelto e fissato gli usuali livelli di significatività (su due code). La ragione di queste illustrazioni sta nel fatto che, pur trattandosi di argomenti classici di statistica elementare, sono raramente riportate nella letteratura corrente.

In quest'ottica, ipotesi stringenti sulla normalità dei campioni e l'estrema specificità della grandezza da sottoporre al confronto d'ipotesi (coefficienti d'asimmetria e indici di curtosi, quantili nelle code) fanno, del test di Pearson et al. e del test di Hawkins, alcuni dei più potenti test noti proprio per questo riportati nel seguito, pur convenendo sulla loro circoscrizione a classi di problemi particolari (ad esempio, individuazione ed eliminazione di dati anomali).

Invece quando le condizioni di validità non sono soddisfatte, ad esempio se la distribuzione della popolazione non è quella normale, oppure se gli elementi della popolazione non sono statisticamente indipendenti, oppure se le varianze della popolazione sono significativamente diverse fra loro, allora i test statistici non–parametrici devono essere utilizzati. Tutto ciò vale in particolare quando i campioni sono piccoli. Infatti una delle condizioni di validità dei test classici è la dimensione grande dei campioni. L'analisi multivariata di campioni di fenomeni vari richiede spesso:

- il confronto fra medie di campioni aventi diversa varianza;
- il fra varianze di campioni non statisticamente indipendenti;
- il confronto fra contingenze di campioni con distribuzione diversa da quella normale.

In questi casi, i test statistici non–parametrici sono indispensabili.

[17] Si chiama contingenza la differenza fra una probabilità doppia (o una frequenza relativa doppia) ed il prodotto delle corrispondenti probabilità marginali (o delle corrispondenti frequenze relative marginali). La contingenza fra due variabili casuali (o due variabili statistiche) indipendenti ha valore zero; in corrispondenza ad ogni altro valore (compreso fra –1 e 1), le due variabili si dicono

L'effettiva esecuzione di un qualsiasi test statistico si attua sempre compiendo i seguenti passi:

- formulazione di una determinata ipotesi fondamentale;
- scelta del livello di significatività;
- costruzione di una statisitca campionaria, a partire da dati osservati;
- partizione della distribuzione di probabilità della statistica campionaria;
- effettuazione del confronto d'ipotesi,

a prescindere dal tipo di test realmente effettuato: semplice o multiplo, della normalità o non – parametrico (ovvero distribution – free).

a) Test di buon adattamento (goodness of fit)

Questi test verificano il buon adattamento, di una funzione densità di probabilità (o di una funzione distribuzione di probabilità) di una certa variabile casuale con l'istogramma delle frequenze relative (o quello delle frequenze cumulate) di una data variabile statistica, che si suppone un campione estratto da una popolazione costituita dalla variabile casuale stessa.

a.1) Test χ^2 (chi quadrato)

Caso monodimensionale

Dati d'ingresso: numerosità del campione N

numero di classi n

frequenze relative $f_i\,;\, i = 1,n$

probabilità delle classi $p_i\,;\, i = 1,n$

livello di significatività α

Ipotesi fondamentale: H_0: buon adattamento fra: la funzione densità di probabilità e l'istogramma delle frequenze relative.

Valore atteso: $$\chi_e^2 = N \sum_{i=1}^{n} \frac{(f_i - p_i)^2}{p_i}$$

$$\upsilon = n - 1 - h \quad gdl$$ [18]

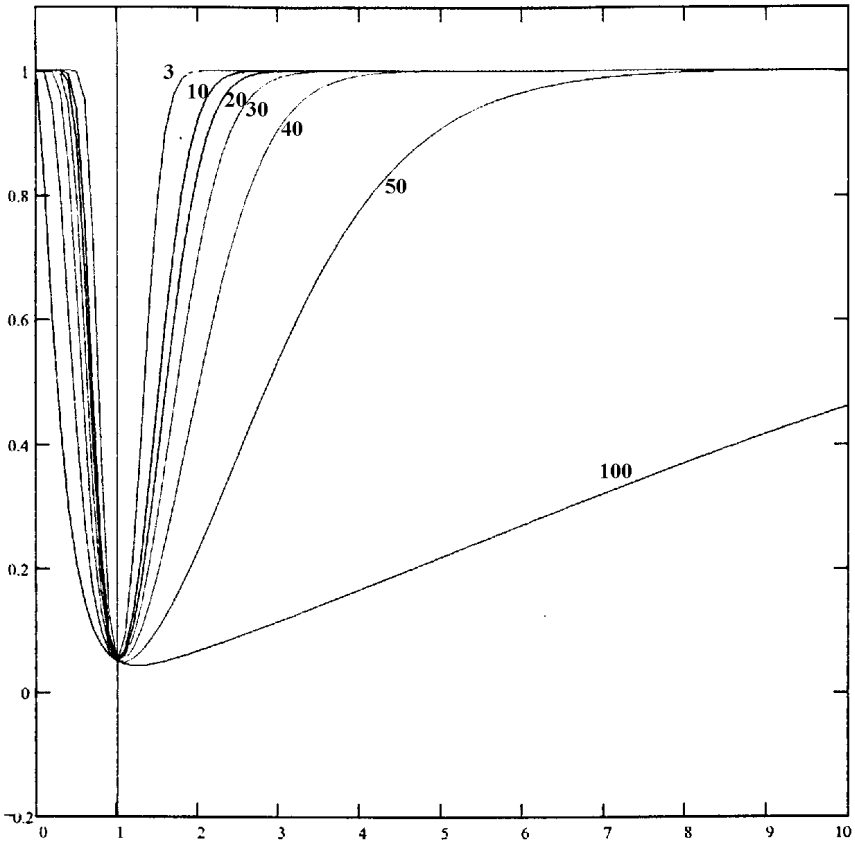
Confronto d'ipotesi: $\chi_t^{2(-)} \le \chi_e^2 \le \chi_t^{2(+)}$ [19]

---

connesse.

[18] Il numero intero h conta il numero di parametri stimati, a partire dai dati originari, ed utilizzati per l'esecuzione dei test (ad es., nel caso monodimensionale: media, e varianza; nel caso bidimensionale: medie, varianze e coefficiente di correlazione).

[19] Il confronto d'ipotesi per questo test, come pure per gli altri test di buon adattamento e per i successivi test d'indipendenza, si effettua a due code, quando (come spesso accade) sia necessario avere cautela verso spiegazioni fasulle, di qualsiasi genere e natura, altrimenti a una coda.



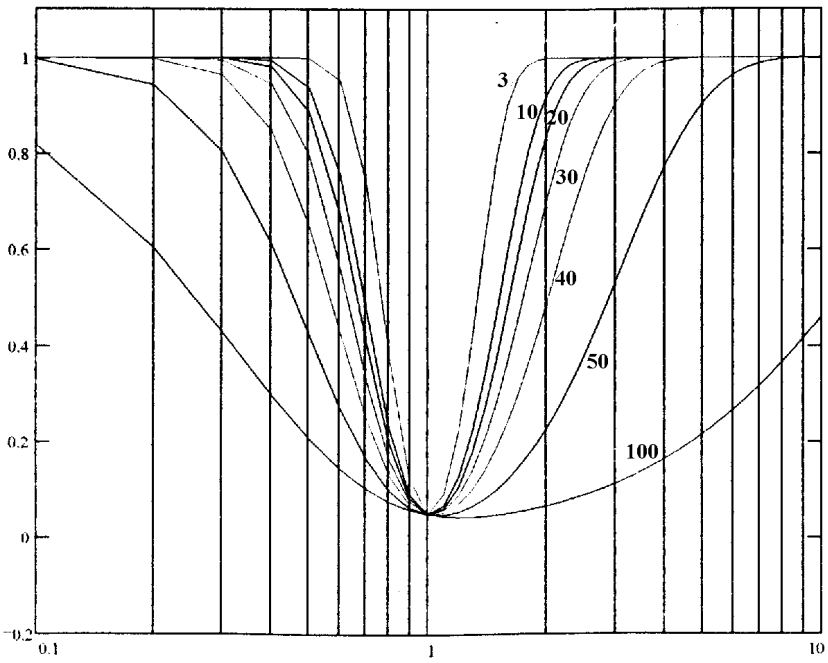


Fig. 3.2 [20]

[20] Accanto ai valori del livello di significatività del test (cioè ad un errore di prima specie, ovvero alla probabilità di accettare un'ipotesi errata) e della potenza del test (cioé ad un errore di seconda specie, ovvero alla probabilità di rifiutare un'ipotesi corretta), il p-value misura il minimo livello di significatività (standardizzato tra 0 ed 1) del test in base al quale occorre rifiutare l'ipotesi nulla, se l'ipotesi nulla è corretta. In altre parole, il p-value di un esperimento è una variabile casuale, definita sopra lo spazio campionario dell'esperimento, tale che la relativa distribuzione probabilità, sotto l'ipotesi nulla, sia uniforme nell'intervallo $[0,1]$.

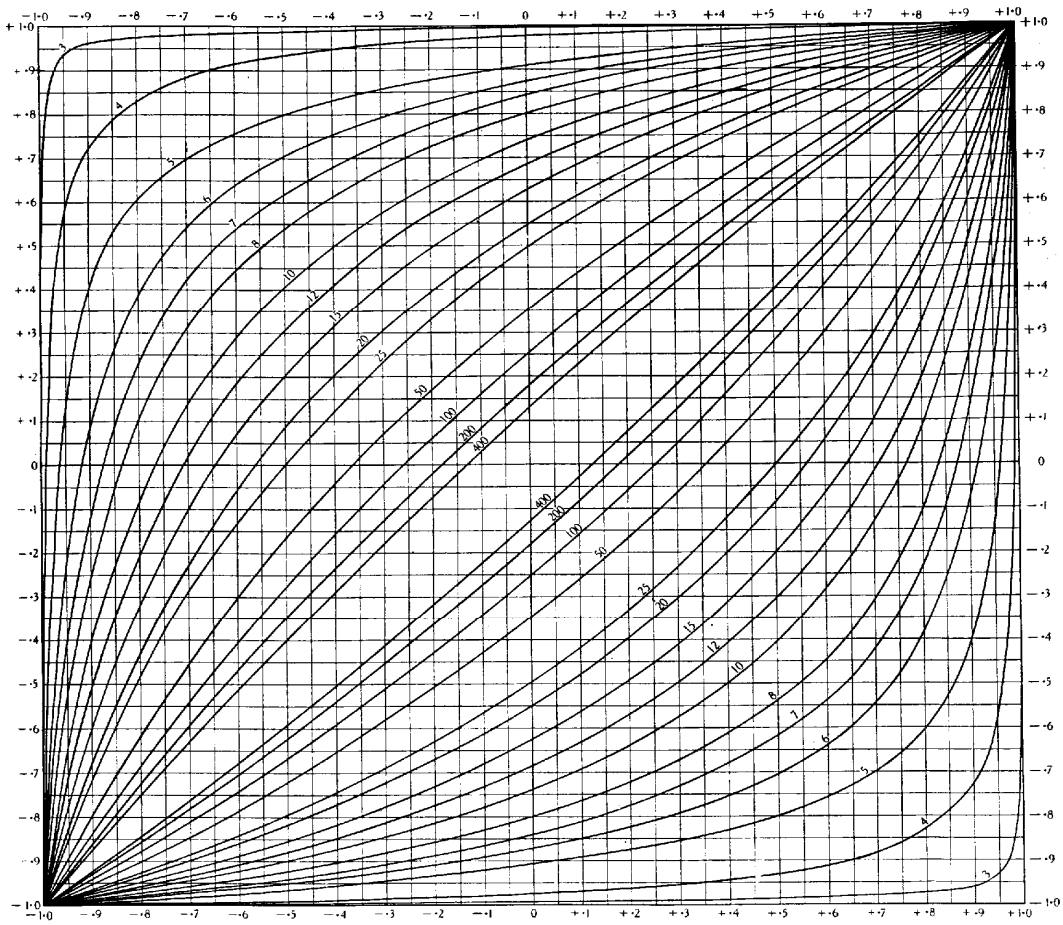


Fig. 3.3

Caso bidimensionale [21]

Dati d'ingresso: numerosità del campione N

numero di classi mn

frequenze relative f_{ij} ; $i = 1,m;\ j = 1,n$

probabilità delle classi p_{ij} ; $i = 1,m;\ j = 1,n$

livello di significatività α

Ipotesi fondamentale: H_0 : (come nel caso monodimensionale).

Valore atteso: $$\chi_e^2 = N \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(f_{ij} - p_{ij})^2}{p_{ij}}$$

$$\upsilon = mn - 1 - h \qquad gdl$$

Confronto d'ipotesi: $$\chi_t^{2(-)} \le \chi_e^2 \le \chi_t^{2(+)}$$

[21] L'estensione a casi multidimensionali è immediata, tranne per il calcolo dei gradi di libertà. In generale essi valgono il prodotto dei numeri delle classi in tutte le direzioni, diminuito di un'unità, meno il numero di parametri stimati, da valutarsi caso per caso.

a.2) Test di Kolmogorov–Smirnov

Caso monodimensionale

Dati d'ingresso: numerosità del campione N

numero di classi n

frequenze cumulate F_i ; $i = 1,n$

probabilità delle classi P_i ; $i = 1,n$

livello di significatività α

Ipotesi fondamentale: H_0 : buon adattamento fra: la funzione distribuzione di probabilità e l'istogramma delle frequenze cumulate.

Valore atteso: $D_e = \max_{i=1,n} \left(\left| F_i - P_i \right| \right)$

$\upsilon = N \qquad gdl$

D_t ascissa critica della distribuzione estremale di Kolmogorov–Smirnov[22]

Confronto d'ipotesi: $D_t^{(-)} \le D_e \le D_t^{(+)}$

Nel **caso bidimensionale** (ed anche nei **casi multidimensionali**) il test è, formalmente, identico a quello per il caso monodimensionale. Infatti il test di Kolmogorov–Smirnov non ha limitazioni, rispetto alla dimensione del problema, dovendosi, in ogni caso, individuare la massima differenza assoluta, fra una funzione distribuzione di probabilità di una certa variabile casuale ed un istogramma delle frequenze cumulate di una data variabile statistica, dove i gradi di libertà sono, sempre, uguali alla numerosità del campione.

a.3) Test Pearson et al. per la normalità

Dati d'ingresso: numerosità del campione m

coefficiente d'asimmetria γ

indice di curtosi β

livello di significatività α

---

[22] Si dicono **distribuzioni estremali**: le distribuzioni degli estremi (superiore o inferiore) di un insieme di dati o di statistiche. Le distribuzioni estremali sono ricavate dalle distribuzioni ordinarie, applicando il Teorema della Probabilità Composta in condizioni d'indipendenza, con l'ulteriore ipotesi che le distribuzioni ordinarie siano tutte uguali fra loro:

Funzione distribuzione di probabilità del valore massimo $\{y = max(x_i), i = 1,n\}$: $P_{max}(y) = [P(x)]^n$

Funzione densità di probabilità del valore massimo: $p_{max}(y) = n[P(x)]^{n-1} p(x)$

Funzione distribuzione di probabilità del valore minimo $\{y = min(x_i), i = 1,n\}$: $P_{min}(y) = 1 - [1 - P(x)]^n$

Funzione densità di probabilità del valore massimo: $p_{min}(y) = n[1 - P(x)]^{n-1} p(x)$

Ipotesi fondamentale: H_0 : normalità del campione.

Valore atteso: $$\chi_e^2 = m\left(\frac{\gamma^2}{6} + \frac{(\beta-3)^2}{24}\right) \qquad \upsilon = 2 \quad gdl$$

Confronto d'ipotesi: $$\chi_t^{2^{(-)}} \le \chi_e^2 \le \chi_t^{2^{(+)}}$$

b) Test di indipendenza

Questi test verificano l'indipendenza, delle componenti di una certa variabile casuale bidimensionale (o multidimensionale), accertando l'adattamento delle frequenze relative doppie (o multiple) di una data variabile statistica al prodotto delle frequenze marginali corrispondenti, secondo l'enunciato del teorema della probabilità composta. (Detta variabile si suppone, al solito, un campione estratto da una popolazione costituita dalla variabile casuale stessa.).

Nel caso bidimensionale[23] il test è formalmente identico a quello di buon adattamento: χ^2 (chi quadrato) o di Kolmogorov–Smirnov, bastando sostituire, fra i dati d'ingresso e nel calcolo del valore atteso, il prodotto delle frequenze marginali: $f_i f_j$; $i = 1,m;\ j = 1,n$, o delle frequenze marginali cumulate: $[f_i][f_j]$; $i = 1,m;\ j = 1,n$, alle probabilità delle classi.

Per quanto riguarda, invece, il calcolo dei gradi di libertà, essi valgono, rispettivamente, per il test χ^2 (chi quadrato):

$$\upsilon = mn-(m-1)-(n-1)-1-h = mn-m-n+1-h = (m-1)(n-1)-h \qquad gdl$$

e per il test di Kolmogorov–Smirnov: $\upsilon = N \qquad gdl.$

c) Test per l'analisi di varianza

Questi test verificano l'uguaglianza, fra la media di una certa variabile casuale e le medie di date variabili statistiche, che si suppongono campioni estratti da una popolazione costituita dalla variabile casuale stessa[24].

c.1) Test di Fisher per campioni normali (di ugual varianza o equiponderati con pesi noti)

Caso "ad una via"

Dati di ingresso: numero di campioni n

[23] Anche nei casi multidimensionali il test è, formalmente, identico a quello di buon adattamento o di Kolmogorov–Smirnov. Per quanto riguarda, invece, il calcolo dei gradi di libertà, in generale, essi valgono il prodotto dei numeri delle classi in tutte le direzioni diminuiti, ciascuno, di un'unità, meno il numero di parametri stimati da valutarsi, caso per caso, per il test di buon adattamento, mentre per il test di Kolmogorov–Smirnov essi sono, ancora una volta, uguali alla numerosità del campione.

[24] Gli stessi test verificano, anche, l'uguaglianza fra le medie di certe variabili casuali e le medie di date variabili statistiche. Infatti le differenze fra queste medie sono, ancora, medie di campioni estratti da un'unica popolazione costituente una nuova variabile casuale, avente, ovviamente, una sola media di valore nullo, in caso d'uguaglianza. In tal caso, i gradi di libertà della varianza spiegata crescono di uno, mentre quelli della varianza residua diminuiscono di uno.

numerosità di ogni campione m [25]

valori argomentali x_{ij}

medie parziali $\bar{x}_j$

media generale $\bar{x}$

varianza generale σ^2

livello di significatività α

Ipotesi fondamentale: H_0 : uguaglianza delle medie parziali fra loro e con la media generale.

Valore atteso:

$$\sigma_s^2 = m \sum_{j=1}^{n} \frac{(\bar{x}_j - \bar{x})^2}{n-1}$$

$$\sigma_r^2 = m \sum_{j=1}^{n} \sum_{i=1}^{m} \frac{(x_{ij} - \bar{x}_j)^2}{n(m-1)} = \frac{(mn-1)\sigma^2 - (n-1)\sigma_s^2}{n(m-1)}$$

$$F_e = \frac{\sigma_s^2}{\sigma_r^2}$$

$$\upsilon = n-1 \quad ; \quad \lambda = mn-1-(n-1) = n(m-1) \quad gdl$$

Confronto d'ipotesi: $F_t^{(-)} \le F_e \le F_t^{(+)}$

Caso "a due vie" senza ripetizioni nelle celle[26]

Dati d'ingresso: numero di componenti per colonne (dette: trattamenti (T)) n

numero di componenti per righe (dette: blocchi (B)) m

---

[25] L'estensione al caso di campioni avente diversa numerosità: $m_j \ (j = 1, n)$, non è complessa, in quanto, definiti:

la numerosità media $\bar{m}$

ed il peso di ciascun campione $p_j = m_j / \bar{m}$

il calcolo della varianza spiegata diviene:

$$\sigma_s^2 = \bar{m} \sum_{i=1}^{n} p_j (\bar{x}_j - \bar{x})^2 / (n-1)$$

Allora calcolata, per differenza, la varianza residua: $\sigma_r^2 = \frac{(\bar{m}n-1)\sigma^2 - (n-1)\sigma_s^2}{n(\bar{m}-1)}$, il confronto d'ipotesi procede come al solito, essendo i gradi di libertà: $\nu = n-1$; $\lambda = n(\bar{m}-1)$.

[26] L'analisi di varianza prosegue con i casi a due vie con ripetizioni nelle celle ed interazioni fra queste, a tre vie senza o con ripetizioni nelle celle ed, ancora, interazioni fra queste ecc.; in tutti questi casi, occorre:

- calcolare, direttamente, la varianza generale e le varianze spiegate;
- calcolare, per differenza, la varianza residua;
- contare tutti i corrispettivi gradi di libertà;

valori argomentali	x_{ij}
medie parziali per colonne e per righe	$\bar{x}_j$, $\bar{x}_i$
media generale	$\bar{x}$
varianza generale	σ^2
livello di significatività	α

Ipotesi fondamentale: H_0 : (come per il caso "ad una via").

Valori attesi:

$$\sigma_S^2(T) = m \sum_{j=1}^{n} \frac{(\bar{x}_j - \bar{x})^2}{n-1}$$

$$\sigma_S^2(B) = n \sum_{i=1}^{m} \frac{(\bar{x}_i - \bar{x})^2}{m-1}$$

$$\sigma_r^2 = \frac{(mn-1)\sigma^2 - (n-1)\sigma_S^2(T) - (m-1)\sigma_S^2(B)}{(m-1)(n-1)}$$

$$F_e = \frac{\sigma_S^2(T)}{\sigma_r^2} \quad \upsilon = n-1 \quad ;$$

$$\lambda = mn - 1 - (m-1) - (n-1) = (m-1)(n-1) \ gdl$$

$$F_e = \frac{\sigma_S^2(B)}{\sigma_r^2} \quad \upsilon = m-1 \quad ; \lambda = (m-1)(n-1) \qquad gdl$$

Confronto d'ipotesi: $F_t^{(-)} \leq F_e \ (T, oppure\ B) \leq F_t^{(+)}$

c.2) Test di Welch generalizzato per campioni normali (di diversa varianza)[27]

Dati di ingresso:	numero di campioni	n
	numerosità dei campioni	m_i
	medie parziali	$\bar{x}_i$
	media generale	$\bar{x}$
	varianze parziali	σ_i^2

poi calcolare i valori attesi ed eseguire, infine, i confronti d'ipotesi.

[27] Il test di Welch generalizzato, piuttosto complesso nella sua derivazione teorica, si pone in una posizione intermedia fra i test classici della normalità ed i test non – parametrici e completa la gamma dei casi possbili. Infatti esso è adatto per il confronto di medie di campioni normali, ovviamente indipendenti, di diversa varianza non equiponderati con pesi noti.
Resta da osservare come esso, benché parecchio importante, per la sua definizione e collocazione nell'insieme dei test, sia molto poco conosciuto e, per lo più, citato solo come confronto tra due varianze, anziché tra molte, così come esso è stato originariamente formulato dal suo stesso autore.

livello di significatività α

Ipotesi fondamentale: H_0 : uguaglianza delle medie parziali fra loro e con la media generale.

Valore atteso: $$F_e = \frac{\sum_{i=1}^{n} \left(\bar{x}_i - \bar{x}\right)^2}{\sum_{i=1}^{n} \frac{\sigma_i^2}{m_i}}$$

$$\upsilon = n - 1 \quad ; \quad \lambda = \frac{\left(\sum_{i=1}^{n} \frac{\sigma_i^2}{m_i}\right)^2}{\sum_{i=1}^{n} \left(\frac{\sigma_i^2}{m_i}\right)^2 \frac{1}{m_i - 1}} \quad gdl$$ [28]

Confronto d'ipotesi: $F_t^{(-)} \le F_e \le F_t^{(+)}$

c.3) Test di Hawkins contro la presenza di dati anomali

Dati d'ingresso: scarti residui v

numerosità del campione m

numero di parametri del modello n

varianze degli scarti residui $\sigma_v^2 = \sigma_0^2 \; diag \; Q_{vv}$

livello di significatività α

Ipotesi fondamentale: H_0 : assenza di dati anomali.

Valore atteso: $$H_e = max\left(\frac{v_i^2}{v\sigma_{v_i}^2}\right) \qquad \forall i$$

$$\upsilon = m - n \qquad gdl$$

H_t ascissa critica della distribuzione estremale di Hawkins[29]

Confronto d'ipotesi: $H_t^{(-)} \le H_e \le H_t^{(+)}$

---

[28] In generale, il numero dei gradi di libertà del denominatore del test di Welch generalizzato non risulta un numero intero; allora lo stesso numero è solitamente ricondotto al numero intero più prossimo.

[29] Si dicono distribuzioni estremali: le distribuzioni degli estremi (superiore o inferiore) di un insieme di dati o di statistiche. Le distribuzioni estremali sono ricavate dalle distribuzioni ordinarie, applicando il Teorema della Probabilità Composta in condizioni d'indipendenza, con l'ulteriore ipotesi che le distribuzioni ordinarie siano tutte uguali fra loro.

Funzione distribuzione di probabilità del valore massimo $\{y = max(x_i), i = 1,n\}$: $P_{max}(y) = [P(x)]^n$

Funzione densità di probabilità del valore massimo: $p_{max}(y) = n[P(x)]^{n-1} p(x)$

Funzione distribuzione di probabilità del valore minimo $\{y = min(x_i), i = 1,n\}$: $P_{min}(y) = 1 - [1 - P(x)]^n$

Funzione densità di probabilità del valore massimo: $p_{min}(y) = n[1 - P(x)]^{n-1} p(x)$

c.4) Test di Kruskal–Wallis per campioni indipendenti[30]

Dati d'ingresso: numero di campioni n

numerosità di ciascun campione m_j

numerosità totale $N = \sum_{j=l}^{n} m_j$

ranghi, ovvero numeri ordinali, in corrispondenza all'unione

di tutti i valori argomentali ordinati in modo crescente [31] r_{ij}

livello di significatività α

Ipotesi fondamentale: H_0 : (come per il test di Fisher).

Valore atteso: $\chi_e^2 = \frac{12}{N(N+1)} \sum_{j=l}^{n} \frac{R_j^2}{m_j} - 3(N+1)$ dove: $R_j = \sum_{i=l}^{m_j} r_{ij}$ $\forall j$

$\upsilon = n - 1$ gdl

Confronto d'ipotesi: $\chi_t^{2(-)} \le \chi_e^2 \le \chi_t^{2(+)}$

c.5) Test di Friedman per campioni qualsiasi[32]

Dati d'ingresso: numero di campioni n

numerosità di ciascun campione m

ranghi, in corrispondenza alle unioni, un elemento alla volta per ogni campione,

dei valori argomentali ordinati in modo crescente r_{ij}

livello di significatività α

Ipotesi fondamentale: H_0 : (come per il test di Fisher).

Valore atteso: $\chi_e^2 = \frac{12}{mn(n+1)} \sum_{j=1}^{n} R_j^2 - 3m(n+1) = \frac{6 \sum_{j=1}^{n} (R_j - \overline{R})^2}{\sum_{j=1}^{n} R_j}$

[30] Questo test generalizza per un numero di campioni qualsiasi, quanto eseguito dal test del rango di Mann–Withney, per il confronto di medie di due campioni, aventi le stesse caratteristiche.

[31] Si ponga, qui e nel prosieguo, parecchia attenzione alle diverse modalità di assegnazione dei ranghi (in questo caso specifico: alla totalità dei dati, nel prossimo test invece: per gruppi separati di dati), in quanto differenti modalità determinano differenti test e possono anche produrre differenti risultati.

[32] Questo test generalizza per un numero di campioni qualsiasi, quanto eseguito dal test del segno di Thompson, per il confronto di medie di due campioni, aventi le stesse caratteristiche.
Si ricordi, a riguardo, come il test del segno di Thompson abbia l'obiettivo di eliminare gli elementi ripetuti (con differenze nulle), valutando solo gli elementi non–ripetuti (con differenze positive e negative). Pertanto nella sua generalizzazione, con il test di Fiedman, occorre proprio assegnare i ranghi un elemento alla volta per ogni campione, assegando ranghi uguali agli elementi ripetuti (con differenze nulle) e ranghi diversi agli elementi non–ripututi (con differenze positive e negative).
In entrambi i test, l'identica numerosità per tutti i campioni è, ovviamente, condizione necessaria per il loro appaiamento.

dove: $R_j = \sum_{i=l}^{m_j} r_{ij} \qquad \forall j$

$\upsilon = n - 1 \qquad gdl$

Confronto d'ipotesi: $\chi_t^{2(-)} \le \chi_e^2 \le \chi_t^{2(+)}$

d) Test per lo studio delle componenti della varianza

Questi test verificano l'uguaglianza, fra la varianza di una certa variabile casuale e le varianze di date variabili statistiche, che si suppongono campioni estratti da una popolazione costituita dalla variabile casuale stessa[33].

d.1) Test di Bartlett per campioni normali

Dati d'ingresso: numero di campioni n

numerosità di ciascun campione m_j

numerosità totale $N = \sum_{j=1}^{n} m_j$

componenti della varianza σ_j^2

livello di significatività α

Ipotesi fondamentale: H_0: uguaglianza delle componenti della varianza fra loro e con la varianza generale.

Valore atteso: $\chi_e^2 = -2 \ln \Lambda$ dove: $\Lambda = \dfrac{\prod_{j=1}^{n} (\sigma_j^2)^{m_j/2}}{\left(\sum_{j=1}^{n} m_j \sigma_j^2 \Big/ N \right)^{N/2}}$

$\upsilon = n - 1 \qquad gdl$

Confronto d'ipotesi: $\chi_t^{2(-)} \le \chi_e^2 \le \chi_t^{2(+)}$

d.2) Test di Kruskal–Wallis per campioni indipendenti[34]

Dati d'ingresso: numero di campioni n

numerosità di ciascun campione m_j

numerosità totale $N = \sum_{j=1}^{n} m_j$

[33] Gli stessi test verificano, anche, l'uguaglianza fra le varianze di certe variabili casuali e le varianze di date variabili statistiche. Infatti i rapporti fra queste varianze sono, ancora, varianze di campioni estratti da un'unica popolazione costituente una nuova variabile casuale, avente, ovviamente, una sola varianza di valore unitario, in caso d'uguaglianza. In tal caso, i gradi di libertà del rapporto di verosimiglianza (di Fisher) crescono di uno.

[34] Questo test generalizza per un numero di campioni qualsiasi, quanto eseguito dal test del rango di Siegel–Tukey, per il confronto di

ranghi, in corrispondenza all'unione dei moduli degli scarti

(rispetto alle mediane parziali) ordinati in modo crescente[35] r_{ij}

livello di significatività α

Ipotesi fondamentale: H_0: (come per il test di Bartlett).

Valore atteso: $\chi_e^2 = \frac{12}{N(N+1)} \sum_{j=l}^{n} \frac{R_j^2}{m_j} - 3(N+1)$ dove: $R_j = \sum_{i=1}^{mj} r_{ij}$ $\forall j$

$\upsilon = n - 1$ gdl

Confronto d'ipotesi $\chi_t^{2(-)} \leq \chi_e^2 \leq \chi_t^{2(+)}$

d.3) Test di Friedman per campioni qualsiasi[36]

Dati d'ingresso: numero di campioni n

numerosità di ciascun campione m

ranghi, in corrispondenza alle unioni, un elemento alla volta per ogni campione, dei moduli degli scarti (rispetto alle mediane parziali)

ordinati in modo crescente $r_{i,j}$

livello di significatività α

Ipotesi fondamentale: H_0: (come per il test di Bartlett).

Valore atteso: $\chi_e^2 = \frac{12}{mn(n+1)} \sum_{j=1}^{n} R_j^2 - 3m(n+1) = \frac{6 \sum_{j=1}^{n} (R_j - \overline{R})^2}{\sum_{j=1}^{n} R_j}$

dove: $R_j = \sum_{i=1}^{m} r_{ij}$ $\forall j$

$\upsilon = n - 1$ gdl

Confronto d'ipotesi $\chi_t^{2(-)} \leq \chi_e^2 \leq \chi_t^{2(+)}$

---

varianze di due campioni, aventi le stesse caratteristiche.

[35] In alternativa all'ordinamento crescente dei moduli degli scarti, si può utilizzare un ordinamento a mulinello (dalle code verso il centro) degli scarti stessi e, proprio così, questa modalità è stata formulata e proposta, come nel test non – parametrico per il confronto statistico tra due popolazioni di scarti per verificare la loro uguale o diversa dispersione. Tuttavia poiché l'assegnazione dei ranghi in modo crescente, non può dipendere dal fatto che questi siano valori argomentali (tutti positivi) o moduli degli scarti (rispetto alle mediane parziali), l'assegnazione dei ranghi (con il loro segno) a mulinello diventa un po' più laboriosa e, di conseguenza, un po' meno raccomandabile.

[36] Questo test generalizza per un numero di campioni qualsiasi, quanto eseguito dal test del segno di Thompson, per il confronto di varianze di due campioni, aventi le stesse caratteristiche.

e) Test per lo studio della struttura di covarianza

Questi test verificano l'incorrelazione, delle componenti di una certa variabile casuale multidimensionale, accertando l'annullarsi di opportuni indicatori della correlazione fra le componenti di una data variabile statistica multidimensionale, che si suppone un campione estratto da una popolazione costituita dalla variabile casuale stessa[37].

e.1) Test di Hotelling per campioni normali

Dati d'ingresso: numero di componenti del campione (multidimensionale) n

numerosità di ciascun componente m

matrice di varianza–covarianza (σ_x^2 varianze degli elementi) C_{xx}

livello di significatività α

Ipotesi fondamentale: H_0 : incorrelazione fra le componenti.

Valore atteso[38]: $\chi_e^2 = -2\,ln\,\Lambda$ dove: $\Lambda = \dfrac{(det\,C_{xx})^{m/2}}{(\prod_{j=1}^{n} \sigma_{xj}^2)^{m/2}}$ [39]

---

[37] Gli stessi test verificano, anche, anche il buon adattamento di una matrice di varianza–covarianza Γ , fra le componenti di una certa variabile casuale multidimensionale, con la matrice di varianza–covarianza C, fra le componenti di una data variabile statistica multidimensionale. Infatti eseguite le decomposizioni spettrali di entrambe le matrici:

$$\Gamma = \Xi\Lambda\Xi^T \; ; \qquad C = QLQ^T$$

e calcolata la matrice:

$$H = \Lambda^{-1/2}(\Xi^T Q)L(Q^T\Xi)\Lambda^{-1/2} = \Lambda^{-1/2}\Xi^T C\Xi\Lambda^{-1/2}$$

essa è, ancora, una matrice di varianza–covarianza di un campione estratto da una popolazione costituente una nuova variabile casuale, avente, ovviamente, incorrelazione ed equiponderazione delle componenti, in caso di buon adattamento. Questo studio è noto, altresì, come estrazione delle componenti principali.

[38] Nel caso particolare, in cui lo studio della struttura di covarianza abbia come matrice di confronto la matrice identità (ovvero incorrelazione ed equiponderazione delle componenti), il test di Hotelling diventa:

$$\chi_e^2 = -2\ln\Lambda \qquad \text{dove: } \Lambda = (det \quad C_{xx})^{m/2} / (1/n \quad Tr \quad C_{xx})^{mn/2}$$

$$\upsilon = n(n+1)/2 - 1 \qquad gdl$$

[39] Sul calcolo di un determinante si veda anche: Benciolini G.B., Mussio L. (1984) Algoritmi esatti per la soluzione di grandi sistemi. In: M. Cunietti (Ed) *Ricerche di Geodesia Topografia e Fotogrammetria* , n. 4. CLUP, Milano, p. 5–46, da cui sono liberamente tratte le seguenti considerazioni.

Il calcolo di un determinante è un'operazione di statistica computazionale solitamente abbastanza complessa ed onerosa che può tuttavia essere affrontata agevolmente, nel caso in cui si sia già eseguita, per lo stesso problema in esame, una qualsiasi fattorizzazione della matrice di covarianza del problema dato o della sua inversa (ad es. una fattorizzazione di Gauss o di Cholesky), ovvero un'operazione capace di generare due matrici triangolari, riguardo alla quale si rimanda a quanto ben noto nella letteratura specialistica di statistica computazionale. Pertanto quanto qui esposto vuole solo completare l'informazione necessaria sull'argomento specifico; infatti:

- il determinante di una matrice inversa è l'inverso del determinante della matrice da cui si è ottenuta la suddetta matrice inversa:

  $$det\,C^{-1} = 1/\,det\,C$$

- il determinante del prodotto di due o più matrici è dato dal prodotto dei determinati di queste matrici:

$$\upsilon = n(n-1)/2 \qquad gdl$$

Confronto d'ipotesi: $\chi_t^{2(-)} \le \chi_e^2 \le \chi_t^{2(+)}$

e.2) Test Lawley per campioni normali

Dati d'ingresso: numero di componenti del campione (multidimensionale) n

numerosità di ciascun componente m

coefficiente di correlazione fra le componenti r_{ij}

livello di significatività α

$det(AB) = det\,A\,det\,B$ $\qquad$ $det(ABC) = det\,A\,det\,B\,det\,C$

- in particolare, il prodotto dei determinanti di due matrici triangolari fornisce il determinante della matrice da cui le stesse matrici sono state ottenute:

  $det(T^T T) = det\,T^T\ det\,T$

- il determinante di una matrice triangolare è uguale al prodotto dei suoi elementi diagonali principali:

  $det\,T = \prod_{k=1}^{n} t_{kk}$

- dato che il complemento algebrico di un elemento (di una matrice triangolare superiore) al di sopra della diagonale principale è sempre nullo:

  $T_{ij} = 0$ $\qquad$ se $\qquad j > i$

- ed anche l'elemento inverso di un elemento (di una matrice triangolare superiore) al di sopra della diagonale principale è sempre nullo:

  $\tau_{ij} = \dfrac{T_{ji}}{det\,T} = 0$ $\qquad$ se $\qquad j < i$

- come pure il complemento algebrico di un elemento (di una matrice triangolare) in diagonale principale di una matrice triangolare è uguale al prodotto di tutti gli elementi diagonali tranne se stesso:

  $T_{ii} = \prod_{k=1}^{n} t_{kk}$ $\qquad$ con $\qquad k \ne 1$

- ed inoltre l'elemento inverso di un elemento (di una matrice triangolare) in diagonale principale è sempre l'inverso dello stesso elemento:

  $\tau_{ii} = \dfrac{T_{ii}}{det\,T} = \dfrac{1}{t_{ii}}$

Per calcolare poi gli altri elementi di una matrice triangolare superiore T^{-1} (τ_{ij}, con $j > i$), conviene esplicitare l'espressione: $T^{-1}\ T = I$, da cui si ha:

$$\tau_{ii}\ t_{ii} = 1 \qquad \sum_{k=1}^{j} \tau_{ik}\ t_{kj} = 0 \qquad \text{con} \qquad j > i$$

e pertanto:

$$\tau_{ii} = \frac{1}{t_{ii}} \qquad \tau_{ij} = -\frac{1}{t_{jj}} \sum_{k=i}^{j-1} \tau_{ik}\ t_{kj} \qquad \text{con} \qquad j > i$$

Ipotesi fondamentale: H_0 : (come per il test di Hotelling).

Valore atteso: $\chi_e^2 = \left(m - 1 - \frac{2n+5}{6} \right) \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} r_{ij}^2$ con $m > \frac{2n+11}{6}$

$\nu = n(n-1)/2 \qquad gdl$

Confronto d'ipotesi: $\chi_t^{2(-)} \le \chi_e^2 \le \chi_t^{2(+)}$

e.3) Test di Wilcoxon–Wilcox modificato secondo Lawley per campioni qualsiasi[40]

Dati d'ingresso: numero di componenti del campione (multidimensionale) n

numerosità di ciascun componente m

ranghi, in corrispondenza alle unioni, un elemento alla volta per ogni campione, dei valori

argomentali ordinati in modo crescente r_{ij}

livello di significatività α

Ipotesi fondamentale: H_0 : (come per il test di Hotelling).

<u>Coefficienti di correlazione sui ranghi di Spearman:</u>

$$\rho_{jk} = 1 - \frac{6}{m(m^2-1)} \sum_{i=1}^{m} (r_{ij} - r_{ik})^2 \qquad \forall j,k(k > j)$$

$$\rho_{jj} = 1 \qquad ; \qquad \rho_{kj} = \rho_{jk}$$

Valore atteso: $\chi_e^2 = \left(m - 1 - \frac{2n+5}{6} \right) \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} r_{ij}^2 \qquad \nu = n(n-1)/2 \qquad gdl$

Confronto d'ipotesi: $\chi_t^{2(-)} \le \chi_e^2 \le \chi_t^{2(+)}$

Giova ribadire che l'inferenza statistica (multivariata) è, forse, la parte più ampia dell'analisi multivariata; pertanto innumerevoli sono i test (multipli) noti in letteratura [41]: quelli riportati hanno avuto lo scopo di illustrare un percorso, facendo particolare attenzione ai problemi pratici, legati alla loro applicazione ad esempi concreti.

[40] Lo studio della struttura di covarianza per campioni qualsiasi, effettuata con il test di Wilcoxon–Wilcox modificato secondo Lawley si basa sul calcolo dei coefficienti di correlazione sui ranghi di Spearman. Per costruzione, tuttavia, questi verificano, solo, la monotonia o meno della dipendenza funzionale e non la sua linearità. Allora uno studio più generale su una dipendenza qualsiasi (o semplice connessione), basato sul calcolo delle contingenze multiple, può costituire un'alternativa o, quantomeno, un primo approccio.

[41] Resta da osservare come l'esecuzione di test multipli, della normalità e non – parametrici, benché più invasivi relativamente ai dati in ingresso e, di conseguenza, più onerosi per il calcolo del valore atteso, non differisca sostanzialmente dai corrispondenti test semplici, bastando scegliere un livello di significatività, calcolare il valore atteso ed eseguire il confronto d'ipotesi, in base alla distribuzione di probabilità prescelta. Per questo preciso motivo, le prime due prossime appendici presentano lo studio dapprima delle variabili statistiche semplice e doppia, e successivamente di tutti i principali test statistici semplici della normalità, in quanto la statistica descrittiva prepara la conoscenza di tutte quelle statistiche che, intese come statistiche campionarie, possono essere sottoposte a test per il controllo d'ipotesi ed il confronto d'ipotesi tra due statisitche campionarie. Naturale estensione del confronto d'ipotesi tra due sono allora tutti i test multipli, appena presentati.

4. Applicazioni dell'analisi multivariata [42]

Le applicazioni della "cluster analysis" a problemi geodetici e geomatici sono di valido aiuto nell'esplorazione (preliminare), nel compattamento e nella classificazione di questi insiemi di dati. Infatti ogni qualvolta il miglioramento della loro conoscenza trova ostacolo nella dubbia classificazione o interpretazione, cosa che rende incerta e problematica la scelta di un dato modello funzionale e/o l'adozione di un certo modello stocastico, la "cluster analysis" dà allora il suo contributo, senza richiedere ipotesi iniziali troppo impegnative. Così dal riconoscimento in un insieme molto eterogeneo di dati di sottoinsiemi omogenei, per quanto riguarda, ad es., la distribuzione di probabilità, oppure lo spettro di potenza (nel dominio delle frequenze), come pure un modello funzionale, talvolta elementare (spesso una media o una regressione lineare), oppure un modello stocastico, quasi sempre elementare (ovvero una varianza e coefficienti di correlazione tutti nulli), prendono avvio successive analisi statistiche più comprensive, ma più sofisticate. Un metodo generale per l'esplorazione (preliminare) e la classificazione di insiemi di dati si articola nei seguenti punti:

- divisione dell'insieme dei dati in sottoinsiemi di dimensioni quasi uguali e calcolo dei parametri statistici di interesse;
- definizione dei "cluster points" ed esecuzione della "cluster analysis", per ogni insieme di parametri statistici;
- controllo della significatività di ogni "cluster point" ed eventuale modifica dei "clusters";
- suddivisione dell'insieme dei dati, secondo i risultati delle "cluster analysis" effettuate.

Una prima applicazione della "cluster analysis" consiste nel compattamento di un insieme di dati il cui numero è veramente eccessivo, come accade nel trattamento di osservazioni dirette geospaziali, di immagini digitali (telerilevate e non) e di mappe di dati da cartografia numerica tecnica e/o tematica, insieme alle sopraccitate immagini digitali spesso inserite in sistemi informativi territoriali (detti anche geografici, se aventi la referenza spaziale forte indotta dalle coordinate, o semplicemente territoriali, se la referenza spaziale è invece debole, come accade, ad es. con gli indirizzi, le particelle catastali e le zone censuarie). Allora il compattamento comporta l'individuazione delle zone di uguale approssimazione, senza alcun diradamento, dove prevale l'aspetto del posizionamento, e la selezione dell'informazione significativa con l'eliminazione dei "pixels" bianchi, dove prevale l'aspetto della rappresentazione.

Un'altra applicazione della "cluster analysis" è la classificazione degli oggetti, di fondamentale importanza per realizzare la modellizzazione solida degli stessi oggetti e la produzione di ortoimmagini cui superimporre, se del caso, elementi vettoriali d'interesse (quali l'altimetria, linee scheletriche e di rottura, il colore, la toponomastica, ecc.). Infatti l'accidentalità degli oggetti (in funzione dei dati da interpolare) e la funzione di autocorrelazione dei dati sono informazioni necessarie per usare rispettivamente i metodi deterministici (interpolazione polinomiale, interpolazione con funzioni splines, analisi di Fourier) ed i metodi stocastici (collocazione, frattali). Altri eventuali preprocessamenti consistono, nella divisione della regione, in esame, in zone omogenee rispetto ad alcuni parametri, per l'eccessivo numero di dati, oppure per la presenza di forti discontinuità e/o grandi singolarità.

[42] L'esecuzione di test multipli, della normalità e non – parametrici, benché più invasivi relativamente ai dati in ingresso e, di conseguenza, più onerosi per il calcolo del valore atteso, non differisce sostanzialmente dai corrispondenti test semplici, bastando scegliere un livello di significatività, calcolare il valore atteso ed eseguire il confronto d'ipotesi, in base alla distribuzione di probabilità prescelta. Per questo motivo, le prime due prossime appendici presentano lo studio dapprima delle variabili statistiche semplice e doppia, e successivamente di tutti i principali test statistici semplici della normalità, in quanto la statistica descrittiva prepara la

Le possibilità di applicazione della regressione multipla agli stessi problemi riguardano, sia la integrazione delle compensazioni, sia l'analisi statistica dei risultati delle stesse compensazioni. In entrambi i casi, si tratta di dare un'espressione analitica a fenomeni fra cui, in generale, non sono note relazioni deterministiche. Tuttavia i due casi, benché simili nelle finalità e nel contenuto, differiscono radicalmente nelle modalità di applicazione.

Nel primo caso, il calcolo della regressione multipla è inteso a integrazione di una compensazione ed è da eseguirsi contemporaneamente al calcolo di quest'ultima. Scopo di questa operazione è valutare, passo dopo passo, la significatività di pochi parametri aggiuntivi, inseriti nelle note relazioni deterministiche, solitamente impiegate in una compensazione minimi quadrati, per tener conto di qualche fenomeno collaterale inspiegato o non sufficientemente spiegato dalle relazioni stesse.

Nel secondo caso, il calcolo della regressione multipla serve per l'analisi statistica dei risultati di una compensazione ed è da eseguisi separatamente, dopo il calcolo di quest'ultima. Scopo di questa operazione è cercare di interpretare, con il metodo della regressione multipla, analizzando un insieme parecchio sovrabbondante di parametri aggiuntivi, quei legami funzionali fra diversi fenomeni, pur facilmente evidenziabili, tuttavia sono rimasti pressoché inspiegati con l'uso di note relazioni deterministiche.

Fra i problemi geodetici e geomatici cui può essere applicata la regressione multipla, sono da segnalare nelle reti geodetiche, quelli relativi alla dipendenza fra misure geodetiche eseguite e parametri meteorologici ed ambientali, ed in fotogrammetria analitica (in particolare, nella triangolazione aerea) e digitale (in particolare, nella feature extraction e nel matching di immagini), quelli relativi all'influenza sulle misure fotogrammetriche eseguite, sia di parametri metereologici ed ambientali, sia di fattori strumentali (dalle operazioni di presa, fino a quelle di restituzione).

Infatti tutti questi problemi, pur studiati con diversi esperimenti appositamente eseguiti, allo stato attuale dei fatti non sono ancora formalizzabili per mezzo di note relazioni deterministiche che esprimono i legami funzionali fra i vari fenomeni che li costituiscono. In queste condizioni, per dare espressione analitica ai legami funzionali pur evidenziati in questi problemi, non è possibile procedere altrimenti che applicando ai dati disponibili modelli interpretativi, costruiti con l'uso di vari metodi statistici.

L'uso della regressione multipla è consigliabile soprattutto nel secondo dei due casi esposti, perché computazionalmente più agevole. Così con specifico riferimento ai problemi geodetici e geomatici, segnalati in precedenza, è possibile applicare il modello interpretativo della regressione multipla per l'analisi statistica dei risultati (correzioni, variazioni, residui ed errori) delle compensazioni di reti geodetiche, oppure di blocchi fotogrammetrici, ove sia stata eseguita la triangolazione aerea, o di elaborazioni di immagini digitali. Infine resta da segnalare che il modello interpretativo della regressione multipla è classico per tutti i problemi di controllo e come esso serva per problemi, di controllo di processi, di taratura di strumenti, di collaudo di metodologie, ecc.

Una valida alternativa alla regressione multipla, specialmente utile nel secondo dei due casi presentati, è data dall'analisi di varianza alle varie vie con cui essa è definita e conosciuta (in particolare, ad una via, due vie e tre vie, le ultime due vie: senza interazioni tra le celle, oppure con le interazioni tra le celle). Infatti l'analisi di varianza è più semplice, anche se meno flessibile, presentando modelli funzionali costituiti semplicemente da medie parziali, calcolate su opportuni sottogruppi di osservazioni e talvolta anche parametri, interpretati come pseudo – osservazioni).

---

conoscenza di tutte quelle statistiche che, intese come statistiche campionarie, possono essere sottoposte a test per il controllo

Un alternativa più complessa, ma anche più ricca, alla regressione multipla è data invece dalle stime di covarianza, dove da osservazioni e talvolta anche parametri, interpretati come pseudo – osservazioni, si ricavano informazioni sulle correlazioni tra esse/i, ovvero dati ausiliari che caratterizzano loro interdipendenza lineare. Infatti benché quest'analisi sia piuttosto complessa e non esente da cadere in contraddizione, in particolare, se la numerosità dei dati è piuttosto limitata, conoscere l'interdipendenza tra osservazioni e/o parametri permette di dare il giusto valore all'informazione acquisita. Al contrario, la mancanza di queste informazioni può portare a sopravvalutare la quantità d'informazione stessa ed a supporre una bontà dei risultati scarsamente credibile.

Un'osservazione complementare sta a valutare come superare il vincolo di uguale numerosità dei campioni, ove esplicitamente richiesto, per evidenti questioni algoritmiche e/o procedurali:

- se tutti i campioni sono piccoli, è possibile ridurre alla stessa numerosità quelli eccedenti, mediando dati vicini, e/o accrescere sempre alla stessa numerosità quelli difettosi, interpolando dati vicini, anche se quest'ultima operazione è più delicata;
- se tutti i campioni sono grandi, è possibile ridurre in classi, le più piccole possibili, i dati raccolti, così da ottenere campioni di dati mediati, aventi tutti la stessa numerosità;
- se alcuni campioni sono grandi ed altri piccoli, è conveniente privilegiare i più grandi, a scapito di quelli più piccoli, specialmente se non si è ben sicuri della indipendenza dei secondi dai primi (anche se le riduzione in classi è sempre tecnicamente possibile).

Resta comunque la possibilità di procedere ad elaborazioni separate da cui ricavare valori interpolati, in numero identico per ciascun campione (ed in numero complessivo pari ai dati originari, per non creare una ridondanza fittizia), ed eseguire successive analisi statistiche su questi campioni, senza o, forse meglio, con errori accidentali simulati ad hoc. Come noto, il primo modo d'operare limita l'analisi dei dati ai soli errori di modello, mentre il secondo la estende anche agli errori di misura (accidentali e non), sempre possibili, e capaci spesso d'alterare le capacità di un dato modello, per quanto ben calibrato. In questo caso specifico, gli errori sistematici sono particolarmente insidiosi, proprio perché capaci d'alterare il modello proposto [43].

Osservazioni conclusive analizzano questioni inerenti procedure robuste (ad esempio, per l'analisi di varianza o la regressione multipla) e test non – parametrici. Infatti soluzioni di problemi mediante procedure robuste richiedono operazioni sui ranghi (cioè sui numeri d'ordine dei dati, invece che sui valori argomentali dei dati) o sottopesature dei dati anomali, oppure strategie deterministiche o statistiche diverse. Invece i test non – parametrici sono strettamente necessari quando vengono meno le ipotesi di normalità ed indipendenza dei dati. Nello specifico, i test di rango servono, ad esempio, a valle di procedure robuste, dove manca la normalità, ma è comunque garantita l'indipendenza, mentre i test del segno sono necessari, se anche l'indipendenza viene meno, come nel caso di confronto tra dati identici, diversamente elaborati, o tra parti di modelli, deterministici o stocastici, non ortogonali tra loro (e pertanto non indipendenti).

---

d'ipotesi ed il confronto tra due statisitche campionarie. Naturale estensione del confronto tra due sono i test multipli già presentati.

[43] Prescidendo dalla presenza di eventuali (pochi) errori gossolani che doveno essere identificati ed eliminati per altra via, un campionamento che incorra in buona parte degli errori sistematici (come, ad es., un campionamento che raccolga dati solo quando un fenomeno e/o un processo segue sempre un certo tipo di comportamento), non è adatto a discriminare tra modelli che pongono, in alternativa, la presenza o meno dei suddetti errori sistematici, costruendo così un modello che, di fatto, li accetta. Tutto ciò significa che, in una base di dati, prima di procedere a riempire i vuoti lasciati dai dati mancanti, occorre sempre domandarsi come si sarebbe comportato il fenomeno e/o processo, se questi dati mancanti fossero stati rilevati.

APPENDICE A – DUE ESEMPI DI STATISTICA DESCRITTIVA

La statistica descrittiva introduce allo studio della statistica ed è preliminare a qualsiasi altro studio più complesso ed avanzato, perché qualsiasi altra analisi dei dati, così come qualsiasi algoritmo e/o procedura della statistica computazionale, no può prescindere dalla conoscenza delle statistiche di base, proprie della statistica descrittiva.

In particolare, accogliendo le tradizionali ipotesi di normalità dei dati, provenienti da campioni ripuliti, lo studio delle variabili statistiche e casuali (ovvero dei dati provenienti dagli esperimenti realmente eseguiti e dei modelli astratti usati per l'interpretazione degli stessi) semplici, cioè ad una dimensione, e doppie, cioè a due dimensioni, sono basilari. D'altra parte, poiché l'ipotesi di poter lavorare con campioni ripuliti deve necessariamente essere verificata ed approvata, data la inevitabile presenza di dati anomali (da intendersi, per lo più, come errori grossolani), statistiche monodimensionali e bidimensionali della robustezza devono affiancare le suddette e più efficienti statistiche della normalità.

VARIABILE STATISTICA SEMPLICE

DATI IN INGRESSO

NUMERO DI OSSERVAZIONI = 16 NUMERO DI CLASSI = 4

VALORI OSSERVATI
-0.42 1.13 0.09 -2.01 0.77 0.00 -0.44 0.48 1.91 -0.18 -1.19 -0.24 0.56 0.98 -1.31 0.13

ELABORAZIONI E RISULTATI

MEDIA	SQM	ASIMMETRIA	CURTOSI
0.02	0.96	-0.20	2.81

AMPIEZZA DELLE CLASSI = 0.980

ISTOGRAMMA E F. DENSITA DI PROBABILITA
ISTOGRAMMA CUMULATO E .F. DISTRIBUZIONE

CENTRI DELLE CLASSI	ESTREMI DELLE CLASSI
CENTRI DELLE CLASSI STANDARD	ESTREMI DELLE CLASSI STANDARD
FREQUENZE ASSOLUTE	FREQUENZE ASSOLUTE CUMULATE
FREQUENZE RELATIVE	FREQUENZE RELATIVE CUMULATE
PROBABILITA NORMALI (SEMPLICI)	PROBABILITA NORMALI CUMULATE

-1.520	-0.540	0.440	1.420	-2.010	-1.030	-0.050	0.930	1.910
-1.604	-0.581	0.442	1.466	-2.116	-1.092	-0.069	0.954	1.977
3.	4.	6.	3.	0.	3.	7.	13.	16.
0.188	0.250	0.375	0.188	0.000	0.188	0.438	0.813	1.000
0.120	0.335	0.358	0.146	0.017	0.137	0.472	0.830	0.976

MEDIANA	MEAN ABSOLUTE VALUE	MEDIAN ABSOLUTE VALUE
0.09	0.74	0.53

VARIABILE STATISTICA DOPPIA

DATI IN INGRESSO

X / Y	1	2	3	4	Marginale(Y)	Media(Y|X	Sqm(Y|X)
1			9		9	3.0	0.0
2	3		2		5	1.8	1.0
3			4	4	8	3.5	0.5
4	1	1		7	9	3.4	1.1
Marginale(X)	4	1	15	11	31		
Media(X|Y)	2.5	4.0	1.7	3.6			
Sqm(X|Y)	0.9	0.0	0.9	0.5			

ELABORAZIONI E RISULTATI

M(X) = 2.55 M(Y) = 3.06
SQM(X) = 1.19 SQM(Y) = 0.95

INDICI DI PEARSON:

$ETA^2(X|Y) = 0.61$ $ETA^2(Y!X) = 0.39$ $ETA^2 = 0.53$

RETTE DI REGRESSIONE:

A(Y(X)) = 0.25 A(X(Y)) = 0.39
B(Y(X)) = 2.43 B(X(Y)) = 1.35 R(X,Y) = 0.31

DISTRIBUZIONE CONGIUNTA IN IPOTESI DI INDIPENDENZA

XINF = 0.5 XSUP = 4.5 YINF = 0.5 YSUP = 4.5
ZXINF= -1.7 ZXSUP= 1.6 ZYINF= -2.7 ZYSUP= 1.5

P(X)P(Y) = 0.85

INDICI DI BONFERRONI:

B(X) = 0.56 B(Y) = 0.66 B(-1) = 0.61 B(0) = 0.61

Un commento conclusivo mette in evidenza come la variabile doppia dia informazioni maggiori della variabile semplice. Infatti se la variabile semplice dà informazioni sul centro e la dispersione di una certa variabile statistica, nonché indicazioni sulla forma della variabile casuale che può fornire un modello di comportamento, la variabile doppia tratta essenzialmente della dipendenza. Questo analisi è di gran lunga più complessa e delicata, perché ha il compito di mettere in evidenza l'esistenza o meno di un legame statistico tra due variabili in esame e, nel caso della sua esistenza, sulla sua natura e la sua forza. E' altrettanto evidente come queste informazioni siano insieme, quasi sempre, di grande interesse e, altrettanto spesso, assai incerte. Infatti studiare la dipendenza è di grande interesse, perché una variabile potrebbe essere più facilmente acquisibile (talvolta è il solo dato acquisibile) di un'altra maggiormente importante, oppure tracciante o precursore di un'altra che si vuole tenere sotto controllo preventivo e così via con esempi consimili, mentre la criticità sta nel rischio di dare informazioni già con campioni non troppo grandi.

APPENDICE B– ESEMPI DI TEST SEMPLICI DELLA NORMALITA' E NON – PARAMETRICI

Come già detto in precedenza, la struttura formale dei test è sempre identica ed i vari test differiscono tra loro solo per le differenti distribuzione adottate per interpretare le diverse statistiche campionarie.
Nel prosieguo, sono svolti e commentati i principali test sulle variabili semplici e doppie, riferiti tanto al controllo d'ipotesi, quanto al confronto tra due statistiche campionarie.
La scelta di svolgere test semplici, anziché test multipli, così come essi sono stati presentati nel paragrafo sull'inferenza statistica multivariata, risiede nella loro maggiore semplicità e coincisone, fattori che giocano, quasi sempre, a vantaggio della chiarezza.

VARIABILE STATISTICA DOPPIA: X,Y

-1.66 0.52 1.85 -0.18 -0.20 1.16 0.50 -0.88 -1.64
-1.63 -1.93 -2.44 1.53 0.64 -1.04 -2.93 -0.19 -0.17

NUMEROSITA DEI CAMPIONI = 9 NUMERO DI CLASSI = 5

STATISTICHE CAMPIONARIE:

M(X)	M(Y)	SQM(X)	SQM(Y)	R(X,Y)
-0.0589	-0.9067	1.2020	1.4760	-0.3700

GAMMA(X)	GAMMA(Y)	BETA(X)	BETA(Y)
0.0333	0.1935	1.5445	1.5278

F(X)	P(X)	F(Y)	P(Y)	F(X,Y)	1	2	3	4	5
0.22	0.23	0.22	0.22	1		0.11		0.11	
0.11	0.21	0.22	0.21	2				0.11	
0.22	0.23	0.11	0.24	3					0.22
0.22	0.18	0.22	0.18	4	0.11	0.11			
0.22	0.15	0.22	0.15	5	0.11		0.11		

FC(X)	FC(Y)	P_FC(X,Y)	1	2	3	4	5
0.22	0.22	1	0.05	0.10	0.12	0.17	0.22
0.33	0.44	2	0.07	0.14	0.18	0.25	0.33
0.55	0.55	3	0.12	0.24	0.30	0.42	0.55
0.77	0.77	4	0.17	0.34	0.42	0.59	0.77
1.00	1.00	5	0.22	0.44	0.55	0.77	1.00

PC(X)	PC(Y)	C(X,Y)	1	2	3	4	5
0.23	0.22	1	-0.05	0.06	-0.02	0.06	-0.05
0.44	0.43	2	-0.02	-0.02	-0.01	0.08	-0.02
0.67	0.67	3	-0.05	-0.05	-0.02	-0.05	0.17
0.85	0.85	4	0.06	0.06	-0.02	-0.05	-0.05
1.00	1.00	5	0.06	-0.05	0.08	-0.05	-0.05

TESTS DELLA NORMALITA'

1) TEST DI BUON ADATTAMENTO (CHI QUADRATO)
A LIVELLO DI SIGNIFICATIVITA' ALFA = 0.02 (CALCOLO PROBABILITA' EFFETTUATO IN BASE AI VALORI CAMPIONARI DI MEDIA E VARIANZA):

VAL-SP(X)	VAL-SP(Y)	GDL	CHI^2-TEOR(INF) [44]	CHI^2-TEOR(SUP)
0.83	1.01	2	0.02	9.23

IPOTESI FONDAMENTALI H(0): VALIDE

2) TEST DI BUON ADATTAMENTO (DI KOLMOGOROV)
A LIVELLO DI SIGNIFICATIVITA' ALFA = 0.02:

VAL-SP(X)	VAL-SP(Y)	GDL	KS-TEOR(INF)
0.12	0.12	9	0.51

IPOTESI FONDAMENTALI H(0): VALIDE

3) TEST DI DI PEARSON ET AL. PER LA NORMALITA'
A LIVELLO DI SIGNIFICATIVITA' ALFA = 0.02:

VAL-SP(X)	GDL	VAL-SP(Y)	CHI^2-TEOR(INF)	CHI^2-TEOR(SUP)
0.80	2	0.87	0.02	9.23

IPOTESI FONDAMENTALI H(0): VALIDE

4) TEST PER IL COEF. DI CORRELAZIONE (CON IPOTESI DI INCORRELAZIONE)
A LIVELLO DI SIGNIFICATIVITA' ALFA = 0.02:

VAL-SP	GDL	T-TEOR	
1.05	7	3.00	IPOTESI FONDAMENTALE H(0): VALIDA

5) TEST PER IL CONFRONTO DI VARIANZE (DI FISHER)
A LIVELLO DI SIGNIFICATIVITA' ALFA = 0.02:

VAL-SP	GDL	F-TEOR(INF)	F-TEOR(SUP)	
1.51	8	0.17	6.03	IPOTESI FONDAMENTALE H(0): VALIDA

6) TEST PER IL CONFRONTO DI MEDIE (T DI STUDENT)
A LIVELLO DI SIGNIFICATIVITA' ALFA = 0.02:

VAL-SP	GDL	T-TEOR	
1.34	16	2.58	IPOTESI FONDAMENTALE H(0): VALIDA

[44] Nei test di buon adattamento e di indipendenza, se i campioni sono abbastanza grandi, si può omettere il confronto con il valore inferiore della distribuzione che serve a garantire contro una soluzione troppo buona, a fronte di campioni di dati particolarmente piccoli.

7) TEST PER IL CONFRONTO DI MEDIE (DI WELCH)

A LIVELLO DI SIGNIFICATIVITA' ALFA = 0.02:

VAL-SP	GDL	T-TEOR	
1.34	17	2.57	IPOTESI FONDAMENTALE H(0): VALIDA

8) TEST PER LA VARIANZA NELL'IPOTESI SQM(XUY) = 1
A LIVELLO DI SIGNIFICATIVITA' ALFA = 0.02: SQM(XUY) = 1.38

VAL-SP	GDL	CHI^2-TEOR(INF)	CHI^2-TEOR(SUP)
32.22	17	6.40	33.44

IPOTESI FONDAMENTALE H(0): VALIDA

9) TEST PER LA MEDIA NELL'IPOTESI M(XUY) = 0
A LIVELLO DI SIGNIFICATIVITA' ALFA = 0.02: M(XUY) = -0.48

VAL-SP	GDL	T-TEOR	
1.49	17	2.57	IPOTESI FONDAMENTALE H(0): VALIDA

10) TEST PER LA MEDIA - CURVA DI POTENZA ESSENDO:

H(0): M(XUY) = 0
H(1): M(XUY) = K×DELTA(M(XUY)) CON DELTA(M(XUY)) = 1.22/SQRT(2N)
A LIVELLO DI SIGNIFICATIVITA' ALFA = 0.02

0	1×DELTA(M(XUY))	2×DELTA(M(XUY))	3×DELTA(M(XUY))	
0.00	0.89	1.77	2.66	ASCISSA STANDARD
0.02	0.08	0.29	0.63	POTENZA (1 – BETA)

Un commento a conclusione dei test della normalità rileva come il test di Pearson et al. per la normalità sia più specifico, rispetto al classico test di buon adattamento chi quadrato ed anche al pur più potente test di Kolmogorov-Smirnov, e come il test per il confronto di medie di Welch non richieda l'uguaglianza delle varianze campionarie, quale precondizione, come nel test t di Student (formulato da Gosset).

<u>TESTS NON – PARAMETRICI</u>

1) TEST DI INDIPENDENZA (CHI QUADRATO)
A LIVELLO DI SIGNIFICATIVITA' ALFA = 0.02:

VAL-SP	CHI^2-TEOR(INF)	CHI^2-TEOR(SUP)	
20.25	5.80	32.03	IPOTESI FONDAMENTALE H(0): VALIDA

2) TEST DI INDIPENDENZA (DI KOLMOGOROV)
A LIVELLO DI SIGNIFICATIVITA' ALFA = 0.02:

VAL-SP	KS-TEOR	
0.19	0.51	IPOTESI FONDAMENTALE H(0): VALIDA

3) TEST SUL COEFFICIENTE DI CORRELAZIONE DI SPEARMAN (CON IPOTESI DI INCORRELAZIONE) A LIVELLO DI SIGNIFICATIVITA' ALFA = 0.02: R = -0.45

VAL-SP(X,Y)	GDL	T-TEOR	
1.33	7	3.00	IPOTESI FONDAMENTALE H(0): VALIDA

4) TEST DI RANGO DI MANN – WHITNEY (CONFRONTO DI VALORI CENTRALI)
A LIVELLO DI SIGNIFICATIVITA' ALFA = 0.02:

S(RK(X))	VAL-SP	Z-TEOR	
98.00	1.10	2.33	IPOTESI FONDAMENTALE H(0): VALIDA

5) TEST DI RANGO DI SIEGEL – TUKEY (UGUAGLIANZA DELLE DISPERSIONI)
A LIVELLO DI SIGNIFICATIVITA' ALFA = 0.02

S(RK(X))	VAL-SP	Z-TEOR	
77.00	-0.75	2.33	IPOTESI FONDAMENTALE H(0): VALIDA

6) TEST DEL SEGNO DI THOMPSON (CONFRONTO DI VALORI CENTRALI)
A LIVELLO DI SIGNIFICATIVITA' ALFA = 0.02:

F(PIU'))	VAL-SP	Z-TEOR	
0.55	0.30	2.33	IPOTESI FONDAMENTALE H(0): VALIDA

7) TEST DEL SEGNO DI THOMPSON (UGUAGLIANZA DELLE DISPERSIONI)
A LIVELLO DI SIGNIFICATIVITA' ALFA = 0.02

F(PIU')	VAL-SP	Z-TEOR	
0.55	0.30	2.33	IPOTESI FONDAMENTALE H(0): VALIDA

Un commento a conclusione test non – parametrici mette in evidenza la maggiore potenza del test di indipendenza di Kolmogorov-Smirnov, rispetto al classico test di indipendenza chi quadrato, mentre il test sul coefficiente di correlazione sui ranghi di Spearman, come noto, prova solo l'indipendenza lineare o meno (o meglio ancora la monotonicità o meno), ma non certamente l'indipendenza stocastica), inoltre la maggiore generalità dei test di segno, rispetto a quelli di rango, non essendo richiesta l'indipendenza nei primi, pur essendo i secondi più potenti. Resta da osservare, come i test di segno siamo comunque indispensabili, quando si mettano a confronto gli stessi dati, mediante algoritmi e/o procedure diverse, proprio perché nessuna indipendenza tra dati può allora essere correttamente ipotizzata.

PARTE II – COMPLEMENTI ED APPLICAZIONI DELL'ANALISI MULTIVARIATA

1. ANALISI DI VARIANZA

Un insieme di dati, stratificato per regioni (dati a referenza spaziale, o spazio–varianti) e/o per epoche (dati variabili nel tempo, o tempo–varianti), come pure per classi di valore (dati varianti in uno spazio astratto), è raggruppabile in celle bidimensionali o tridimensionali. L'informazione dentro le celle è rappresentata da una variabile statistica multidimensionale i cui elementi sono i valori degli attributi dell'informazione stessa.

Nella tecnica di analisi di varianza si considera la varianza generale come somma di due componenti, varianza spiegata e varianza residua, ortogonali tra loro. Lo scopo è quello di massimizzare la varianza spiegata (che descrive il fenomeno secondo il modello stratificato) rispetto a quella residua (che è legata, invece, al carattere statistico dei dati e, in conclusione, alla loro accidentalità).

A tal fine sono stati preparati più modelli di differente grado di complessità. Il più semplice è un modello di **analisi di varianza ad una via** nel quale si vogliono determinare i parametri $\hat{a}_i \quad \forall i = 1,....,I$ (ovvero uno per ogni classe di valore) e $\overline{a}$ (media generale dei dati). La generica equazione d'osservazione risulta essere:

$$\overline{a} + \hat{a}_i = s_{ij}^o + \hat{v}_{ij} \qquad \forall i = 1,....,I \qquad \forall j = 1,....,J$$

dove I è il numero di classi, J è la dimensione della variabile statistica. Sempre nella stessa equazione s_{ij}^o è il dato osservato nella i^{esima} classe e nel j^{esimo} elemento della variabile statistica e $\hat{v}_{ij}$ è lo scarto residuo corrispondente.

Il numero complessivo di osservazioni è $m = I \cdot J$, ed il numero di parametri è $n = I + 1$. A tali equazioni è necessario aggiungere una condizione di vincolo [45], a causa della sovraparametrizzazione dovuta ad incognite non linearmente indipendenti che determinano singolarità del sistema:

$$\hat{a}_I \equiv 0$$

Nel secondo modello, di **analisi di varianza a due vie senza interazione fra le celle**, vengono introdotti ulteriori parametri $\hat{a}_j \quad \forall j = 1,.....,J$ (ovvero uno per ogni regione), per cercare di spiegare un eventuale differente comportamento dei dati nelle varie regioni, secondo lo schema:

$$\overline{a} + \hat{a}_i + \hat{a}_j = s_{ijk}^o + \hat{v}_{ijk} \qquad \forall i = 1,....,I \qquad \forall j = 1,....,J \qquad \forall k = 1,....,K$$

dove I è il numero di classi, J è il numero di regioni e K è la dimensione della variabile statistica. Sempre nella stessa equazione s_{ijk}^o è il dato osservato nella i^{esima} classe, nella j^{esima} regione e nel k^{esimo} elemento della variabile statistica e $\hat{v}_{ijk}$ è lo scarto residuo corrispondente.

[45] Una forma più elegante d'inserimento del vincolo consiste nell'imporre somma, ovvero media, nulla ad ogni gruppo omogeneo di paramentri. Tuttavia questo modo di procedere, facile da mettere in atto per le analisi di varianza senza interazioni fra le celle, diventa invece molto laborioso ed invasivo quando si opera l'analisi di varianza con interazioni fra le celle, come può essere richiesto nell'analisi di varianza due vie e maggiormente a tre vie.

Il numero complessivo di osservazioni è $m = I \cdot J \cdot K$, ed il numero di parametri è $n = I + J + 1$. Le condizioni di vincolo sono:

$$\hat{a}_I \equiv 0, \qquad \hat{a}_J \equiv 0$$

Infine nel terzo modello, di **analisi di varianza a due vie con interazione fra le celle**, si fanno entrare in gioco anche i parametri $\hat{a}_{ij}$, uno per ogni cella, per cercare di spiegare gli effetti secondari delle interazioni fra classi e regioni. La generica equazione d'osservazione è allora:

$$\overline{a} + \hat{a}_i + \hat{a}_j + \hat{a}_{ij} = s^o_{ijk} + \hat{v}_{ijk} \qquad \forall i = 1,....,I \qquad \forall j = 1,....,J \qquad \forall k = 1,....,K$$

In questo caso il numero di parametri aumenta, diventando $n = I + J + I \cdot J + 1$, mentre il numero di osservazioni rimane, ovviamente, invariato. Le condizioni di vincolo sono:

$$\hat{a}_I \equiv 0, \qquad \hat{a}_J \equiv 0 \qquad \hat{a}_{iJ} \equiv 0 \qquad \forall i, \quad \hat{a}_{Ij} \equiv 0 \qquad \forall j \neq J$$

Si omettono lunghe elencazioni concernenti le stime dei parametri, degli scarti e delle loro varianze, in quanto i valori numerici ad esse relativi possono essere facilmente ottenuti risolvendo, a minimi quadrati, i suddetti sistemi di equazioni d'osservazione e, tramite essi, propagando, opportunamente, le varianze delle osservazioni.

Dalle stime dei parametri si ricavano, immediatamente, le varianze spiegate, relative alle classi di valore (dette, convenzionalmente, trattamenti T), alle regioni (dette, convenzionalmente, blocchi B) ed alle interazioni fra esse (la seconda, e la seconda con la terza espressione fanno esclusivo riferimento, rispettivamente, al secondo ed al terzo modello):

$$\sigma_T^2 = \frac{1}{I-1} \sum_{i=1}^{I} \hat{a}_i^2$$

$$\sigma_B^2 = \frac{1}{J-1} \sum_{j=1}^{J} \hat{a}_j^2$$

$$\sigma_{TB}^2 = \frac{1}{(I-1)(J-1)} \sum_{i=1}^{I} \sum_{j=1}^{J} \hat{a}_{ij}^2$$

e la varianza residua (rispettivamente per il primo modello e gli altri due):

$$\sigma_R^2 = \frac{1}{\nu} \sum_{i=1}^{I} \sum_{j=1}^{J} \hat{v}_{ij}^2$$

$$\sigma_R^2 = \frac{1}{\nu} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \hat{v}_{ijk}^2$$

dove il denominatore:

$$v = I \cdot J \qquad -(I-1)-1$$
$$v = I \cdot J \cdot K - (I-1) - (J-1) - 1$$
$$v = I \cdot J \cdot K - (I-1) - (J-1) - (I-1)(J-1) - 1$$

rappresenta i gradi di libertà del sistema nei tre modelli predisposti.
L'analisi di varianza prosegue, nella sua generalizzazione ed estensione, con modelli a tre vie:

- senza interazioni fra le celle;
- ad interazione dominante fra le celle;
- con interazioni fra le celle;
- completa (con interazioni fra le celle, compresa l'interazione tripla).

In questi casi, accanto alla classificazione per classi di valore (dette, convenzionalmente, trattamenti T) ed alle regioni (dette, convenzionalmente, blocchi B), si ha una terza classificazione per epoche (dette, convenzionalmente, strati S). A riguardo, si badi come, benché l'analisi di varianza a tre vie appaia molto lunga nella sua esposizione, essa sia solo apparentemente complessa ed, in realtà, contenga metodologie e procedure del tutto simili all'analisi di varianza ad una via ed a due vie, già illustrate in precedenza. Per tutte queste ragioni, l'analisi di varianza a tre vie viene esposta in forma di tabelle, così da poter contenere lo spazio necessario alla sua presentazione. Le generiche equazioni d'osservazione sono riportate e dettagliate nel prosieguo.

- **Analisi di varianza a tre vie senza interazioni fra le celle:**

$$\bar{a} + \hat{a}_i + \hat{a}_j + \hat{a}_k = s^o_{ijkl} + \hat{v}_{ijkl} \qquad \begin{aligned} &\forall i = 1,....,I \\ &\forall j = 1,....,J \\ &\forall k = 1,....,K \\ &\forall l = 1,....,L \end{aligned}$$

dove I è il numero trattamenti, J è il numero di blocchi, K è il numero di strati ed L è la dimensione della variabile statistica. Il numero complessivo di osservazioni è $m = I \cdot J \cdot K \cdot L$, ed il numero di parametri e $n = I + J + K + 1$.
A tali equazioni è necessario aggiungere alcune condizioni di vincolo, a causa della sovraparametrizzazione dovuta ad incognite non linearmente indipendenti che determinano singolarità del sistema:

$$\hat{a}_I \equiv 0, \qquad \hat{a}_J \equiv 0, \qquad \hat{a}_K \equiv 0$$

- **Analisi di varianza a tre vie ad interazione dominante fra le celle:**

$$\bar{a} + \hat{a}_i + \hat{a}_j + \hat{a}_k + \hat{a}_{ij} = s^o_{ijkl} + \hat{v}_{ijkl} \qquad \begin{aligned} &\forall i = 1,....,I \\ &\forall j = 1,....,J \\ &\forall k = 1,....,K \\ &\forall l = 1,....,L \end{aligned}$$

Il numero di parametri diventa $n = I + J + K + I \cdot J + 1$, rimanendo identico il numero di osservazioni. Le condizioni di vincolo sono:

$$\hat{a}_I \equiv 0, \quad \hat{a}_J \equiv 0, \quad \hat{a}_K \equiv 0$$
$$\hat{a}_{iJ} \equiv 0 \quad \forall i, \quad \hat{a}_{Ij} \equiv 0 \quad \forall j \neq J$$

- **Analisi di varianza a tre vie con interazioni fra le celle:**

$$\bar{a} + \hat{a}_i + \hat{a}_j + \hat{a}_k + \hat{a}_{ij} + \hat{a}_{ik} + \hat{a}_{jk} = s^o_{ijkl} + \hat{v}_{ijkl} \qquad \begin{array}{l} \forall i = 1,....,I \\ \forall j = 1,....,J \\ \forall k = 1,....,K \\ \forall l = 1,....,L \end{array}$$

Il numero di parametri diventa, allora, $n = I + J + K + I \cdot J + I \cdot K + J \cdot K + 1$, rimanendo sempre identico il numero di osservazioni. Le condizioni di vincolo sono:

$$\hat{a}_I \equiv 0, \quad \hat{a}_J \equiv 0, \quad \hat{a}_K \equiv 0$$
$$\hat{a}_{iJ} \equiv 0 \quad \forall i, \quad \hat{a}_{Ij} \equiv 0 \quad \forall j \neq J$$
$$\hat{a}_{iK} \equiv 0 \quad \forall i, \quad \hat{a}_{Ik} \equiv 0 \quad \forall k \neq K$$
$$\hat{a}_{jK} \equiv 0 \quad \forall j, \quad \hat{a}_{Jk} \equiv 0 \quad \forall k \neq K$$

- **Analisi di varianza a tre vie completa:**

$$\bar{a} + \hat{a}_i + \hat{a}_j + \hat{a}_k + \hat{a}_{ij} + \hat{a}_{ik} + \hat{a}_{jk} + \hat{a}_{ijk} = s^o_{ijkl} + \hat{v}_{ijkl} \qquad \begin{array}{l} \forall i = 1,....,I \\ \forall j = 1,....,J \\ \forall k = 1,....,K \\ \forall l = 1,....,L \end{array}$$

Il numero di parametri diventa, in questo caso, $n = I + J + K + I \cdot J + I \cdot K + J \cdot K + I \cdot J \cdot K + 1$, rimanendo ovviamente identico il numero di osservazioni. Le condizioni di vincolo sono:

$$\hat{a}_I \equiv 0, \quad \hat{a}_J \equiv 0, \quad \hat{a}_K \equiv 0$$
$$\hat{a}_{iJ} \equiv 0 \quad \forall i, \quad \hat{a}_{Ij} \equiv 0 \quad \forall j \neq J$$
$$\hat{a}_{iK} \equiv 0 \quad \forall i, \quad \hat{a}_{Ik} \equiv 0 \quad \forall k \neq K$$
$$\hat{a}_{jK} \equiv 0 \quad \forall j, \quad \hat{a}_{Jk} \equiv 0 \quad \forall k \neq K$$
$$\hat{a}_{ijK} \equiv 0 \quad \forall i \neq I, \quad \forall j \neq J$$
$$\hat{a}_{iJk} \equiv 0 \quad \forall i \neq I, \quad \forall k \neq K$$
$$\hat{a}_{IjK} \equiv 0 \quad \forall j \neq J, \quad \forall k \neq K$$
$$\hat{a}_{iJK} \equiv 0 \quad \forall i \neq I$$
$$\hat{a}_{IjK} \equiv 0 \quad \forall j \neq J$$
$$\hat{a}_{IJk} \equiv 0 \quad \forall k \neq K \qquad \hat{a}_{IJK} \equiv 0$$

I valori numerici ad esse relativi possono, ancora, essere facilmente ottenuti risolvendo, a minimi quadrati, i suddetti sistemi di equazioni d'osservazione e, tramite essi, propagando, opportunamente, le varianze delle osservazioni.

Dalle stime dei parametri si ricavano, immediatamente, le varianze spiegate, relative ai trattamenti T, ai blocchi B ed agli strati S ed alle interazioni fra essi, concordemente al modello di riferimento prescelto:

$$\sigma_T^2 = \frac{1}{I-1} \sum_{i=1}^{I} \hat{a}_i^2$$

$$\sigma_B^2 = \frac{1}{J-1} \sum_{j=1}^{J} \hat{a}_j^2$$

$$\sigma_S^2 = \frac{1}{K-1} \sum_{k=1}^{K} \hat{a}_k^2$$

$$\sigma_{TB}^2 = \frac{1}{(I-1)(J-1)} \sum_{i=1}^{I} \sum_{j=1}^{J} \hat{a}_{ij}^2$$

$$\sigma_{TS}^2 = \frac{1}{(I-1)(K-1)} \sum_{i=1}^{I} \sum_{k=1}^{K} \hat{a}_{ik}^2$$

$$\sigma_{BS}^2 = \frac{1}{(J-1)(K-1)} \sum_{j=1}^{J} \sum_{k=1}^{K} \hat{a}_{jk}^2$$

$$\sigma_{TBS}^2 = \frac{1}{(I-1)(J-1)(K-1)} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \hat{a}_{ijk}^2$$

e la varianza residua:

$$\sigma_R^2 = \frac{1}{\nu} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{l=1}^{L} \hat{v}_{ijkl}^2$$

dove il denominatore rappresenta i gradi di libertà del sistema nei quattro modelli predisposti:

$$\nu = I \cdot J \cdot K \cdot L - (I-1) - (J-1) - (K-1) - 1$$
$$\nu = I \cdot J \cdot K \cdot L - (I-1) - (J-1) - (K-1) - (I-1) \cdot (J-1) - 1$$
$$\nu = I \cdot J \cdot K \cdot L - (I-1) - (J-1) - (K-1) - (I-1) \cdot (J-1) - (I-1) \cdot (K-1) - (J-1) \cdot (K-1) - 1$$
$$\nu = I \cdot J \cdot K \cdot L - (I-1) - (J-1) - (K-1) - (I-1) \cdot (J-1) - (I-1) \cdot (K-1) - (J-1) \cdot (K-1) + \\ - (I-1)(J-1)(K-1) - 1$$

Un giudizio sull'analisi di varianza e sulla validità dei modelli adottati discende dalla comparazione fra la varianza spiegata e residua. In questa sede, si utilizzano i seguenti schemi:

- lo schema rigido di comparazione confronta direttamente la varianza spiegata e con la varianza residua;
- uno schema libero di comparazione (applicabile nel caso di analisi di varianza a due vie e a tre vie, con

interazione fra le celle) confronta prima le varianze spiegate di trattamenti, blocchi e strati con le varianze spiegate delle interazioni fra questi e poi le varianze spiegate delle interazioni con la varianza residua;

- gli schemi misti di comparazione combinano, opportunamente, lo schema rigido e lo schema libero.

Benché solo lo schema rigido dia garanzie di assoluto rigore, la scelta dello schema di comparazione dipende solitamente dall'esigenza di mettere in evidenza o meno, alcune forme di dipendenza rispetto ad altre. In ogni caso, l'impiego di test non – parametrici di segno permette, come noto, di superare le limitazioni dovute all'eventuale (ma spesso presente) non – indipendenza delle variabili.
Da ultimo, si osservi che, così come rilevato per la regressione multipla, può essere effettuata anche un'**analisi di varianza robusta**. Allora oltre a sviluppare il calcolo iterando soluzioni a minimi quadrati, mediante la definizione dei pesi ausiliari, mav spiegati e mav residui possono convenientemente sostituire varianze spiegate e varianze residui nella misura dell'efficienza dell'analisi di varianza effettuata, ed i test non – parametrici di rango bene si prestano a formulare i giudizi secondo le regole dell'inferenza statistica.

2. IL CAMPIONAMENTO[46]

Nei lavori di controllo di qualità e, in generale, controllo e confronto d'ipotesi si pone spesso il problema di eseguire campionamenti su una base di dati che, per dimensione, eccede la possibilità di indagini esaustive. Le procedure di campionamento permettono l'estrazione di campioni, numericamente e statisticamente significativi, all'interno delle basi di dati da sottoporre ad indagini statistiche. Si ricorda, a riguardo, che la popolazione può essere finita (essa è costituita da un numero determinato di elementi ed il campionamento è eseguito in blocco) od infinita (quando si osservi, ad esempio, il valore di una variabile continua, oppure se si campiona con ripetizione) e che il piano di campionamento di una popolazione può essere: semplice o sistematico, stratificato o a cluster. Dette procedure suggeriscono di effettuare la stratificazione o clusterizzazione delle basi di dati, al fine di ottenere una miglior spiegazione dei fenomeni e processi in istudio riducendo, nel contempo, il rumore residuo, insito nei dati stessi e nelle metodologie del loro rilevamento.
Una **strategia logistica**, per il campionamento delle basi di dati oggetto di studio, prende in considerazione il numero di ripetizioni dello stesso tipo ed adotta, quale misura di campionamento:

- la totalità, quando la numerosità delle stesse è inferiore ad un valore di soglia prefissato;
- un certo valore percentile, quando la numerosità supera il suddetto valore di soglia, ma non eccede, nel contempo, un secondo valore di soglia prefissato;
- un dato valore (assoluto), quando è superiore al secondo valore di soglia.

Lo scopo di questa strategia è prevenire, insieme, un campionamento poco significativo di campioni poco numerosi ed un campionamento eccessivo di campioni enormemente grandi. La curva logistica è una funzione algebrica, ben nota in demografia, per lo studio dell'accrescimento di popolazioni in ambienti limitati.
Questa curva è caratterizzabile tramite tre parametri, aventi il significato geometrico descritto per presentare la voluta approssimazione lineare. Si riportano, di seguito, alcune espressioni analitiche del suo studio di funzione:

[46] Questa appendice riprende ed aggiorna quanto esposto in: Bellone T., Mussio L., Nocera R. (1998): Campionamento di una base di dati stratificata, avente dimensione di uno o più strati non definita. In: Atti della Conferenza *Rilevamento, rappresentazione e gestione*

$$y = a + (b-a)e^{\frac{-4(b-a)}{e^2cx}} \qquad lim_{x\to 0+}\, y = a \qquad lim_{x\to\infty}\, y = b$$

$$y^I = (b-a)e^{\frac{-4(b-a)}{e^2cx}}\,\frac{4(b-a)}{e^2cx^2}$$

$$y^{II} = (b-a)c^{\frac{-4(b-a)}{e^2cx}}\,\frac{b(b-a)^2}{e^4c^2x^4}\left(1-\frac{e^2cx}{2(b-a)}\right) \qquad y^{II} = 0 \quad \text{per} \quad x = \frac{2(b-a)}{e^2c} = K$$

$$y(K) = a + (b-a)e^{-2} \qquad y^I(K) = c$$

E' chiaro che più il campione è grande, più le stime sono attendibili. D'altra parte, al crescere delle dimensioni del campione aumentano i costi e le difficoltà tecniche ed organizzative, oltretutto con rischi, non troppo remoti, di non riuscire più a garantire l'indipendenza fra le osservazioni e pertanto la bernoullianità del campione ottenuto. L'esistenza di un limite superiore va nella direzione della ricerca di un equilibrio tra la diminuzione della varianza e la complessità dei problemi.

Per delineare i termini del problema, si fa riferimento alla **stima di frequenze**, in particolare alla ricerca dei limiti di confidenza di una frequenza osservata, per una popolazione distribuita binomialmente. A riguardo, si ricordi che, essendo p e q le probabilità associate ai due valori argomantali 1 e 0, la media della variabile semplice è p e la varianza pq, la media della variabile somma è np e la varianza npq, mentre la media della variabile media è p e la varianza pq/n. Data la distribuzione binomiale, approssimata da una normale: $z = (x - n\hat{p}) / \sqrt{n\hat{p}q}$, essa viene corretta per continuità:

$$z_{inf} = \frac{x_{inf} + 0.5 - n\hat{p}}{\sqrt{n\hat{p}q}}$$

$$z_{sup} = \frac{x_{sup} - 0.5 - n\hat{p}}{\sqrt{n\hat{p}q}}$$

ovvero:

$$z_{inf}\sqrt{\frac{\hat{p}q}{n}} = \frac{x_{inf}}{n} + \frac{1}{2n} - \hat{p} = \pi f_{inf} + \frac{1}{2n} - \hat{p}$$

$$z_{sup}\sqrt{\frac{\hat{p}q}{n}} = \frac{x_{sup}}{n} - \frac{1}{2n} - \hat{p} = \pi f_{sup} - \frac{1}{2n} - \hat{p}$$

ottenendo i seguenti risultati, in cui $z_\alpha = z_{sup} = -z_{inf}$:

$$f_{inf} = \left(\hat{p} - \frac{1}{2n}\right) - z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad f_{sup} = \left(\hat{p} + \frac{1}{2n}\right) + z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Tali equazioni riguardano il campionamento con riposizionamento; in caso di campionamento senza riposizionamento, ovvero con correzione per la popolazione finita, si apportano correzioni adeguate. Come noto, la correzione per la popolazione finita ha espressione: $\sqrt{(N-n)/(N-1)}$, dove N è la numerosità della popolazione e n la numerosità del campione; pertanto si ha:

$$f_{inf} = \left(\hat{p} - \frac{1}{2n}\right) - z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}\frac{N-n}{N-1}}$$

$$f_{sup} = \left(\hat{p} + \frac{1}{2n}\right) + z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}\frac{N-n}{N-1}}$$

A questo punto, espresso l'intervallo di confidenza:

$$f_{sup} - f_{inf} = \frac{1}{n} + 2\,z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}\frac{N-n}{N-1}} = 2a$$

e sostituendo, al posto di N–1, il valore approssimato N:

$$2\,z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}\frac{N-n}{N}} = 2a - \frac{1}{n} \qquad ; \qquad 4z_\alpha^2 n\left[\hat{p}(1-\hat{p})\frac{N-n}{N}\right] = 4a^2n^2 + 1 - 4an$$

$$4\left[a^2 + z_\alpha^2\,\hat{p}\frac{(1-\hat{p})}{N}\right]n^2 - 4\left[a + z_\alpha^2\,\hat{p}(1-\hat{p})\right]n + 1 = 0$$

si ottiene il valore atteso per la numerosità del campione e, nel caso in cui il livello di significatività α valga 5%, la variabile normale standardizzata $1.96 \cong 2$ e la probabilità elementare (nel caso più sfavorevole) p sia posta uguale a 0.5, trascurando $1/N$ e sviluppando in serie binomiale $(a+1)^{1/2}$, lo stesso valore diviene:

$$n = \frac{\left[a + z_\alpha^2\,\hat{p}(1-\hat{p})\right] \pm \sqrt{z_\alpha^4\,\hat{p}^2(1-\hat{p})^2 + z_{\alpha\alpha}^2\,\hat{p}(1-\hat{p})\left(2a - \frac{1}{N}\right)}}{2\left(a + z_\alpha^2\,\hat{p}\frac{(1-\hat{p})}{N}\right)}$$

$$n = \frac{(a+1) \pm \sqrt{1 + 1\left(2a - \frac{1}{N}\right)}}{2\left(a + \frac{1}{N}\right)} = \frac{N(a+1) \pm \sqrt{1+2a}}{2\left(a + \frac{1}{N}\right)} = \frac{N\left[(a+1) \pm (a+1)\right]}{2\left(1 + Na^2\right)}$$

Trascurando la soluzione $n = 0$, si ha:

$$n = \frac{N(a+1)}{1+Na^2} = \frac{N}{1+Na^2}$$

in cui si è trascurato a, trattandosi di valore relativamente piccolo.[47]

Tale risultato può venire trasformato nel caso in cui si eseguano altri tipi di test, ad esempio, **test per la distribuzione χ^2**. Infatti nel caso di campionamento con ripetizione, ricordando che la varianza della varianza (se le osservazioni sono normali) ha espressione: $\hat{\sigma}^2_{\hat{\sigma}^2} = \hat{\sigma}^4/2n$, si ottiene:

$$\hat{\sigma}^2 + \chi^2_{sup}\frac{\hat{\sigma}^2}{\sqrt{2n}} - \left(\hat{\sigma}^2 - \chi^2_{inf}\frac{\hat{\sigma}^2}{\sqrt{2n}}\right) = 2a \qquad ; \qquad \left(\chi^2_{sup} + \chi^2_{inf}\right)\frac{\hat{\sigma}^2}{\sqrt{2n}} = 2a$$

$$n = \frac{\left(\chi^2_{sup} + \chi^2_{inf}\right)^2\hat{\sigma}^4}{8a^2}$$

Dopodiché utilizzando la correzione per la popolazione finita, ricordando che:

$$\hat{\sigma}^2_{\hat{\sigma}^2} = \frac{\hat{\sigma}^4}{2n}\frac{N-n}{N-1}$$

e sostituendo N a N–1, si ottiene:

$$\hat{\sigma}^2 + \chi^2_{sup}\frac{\hat{\sigma}^2}{\sqrt{2n}}\frac{\sqrt{N-n}}{\sqrt{N}} - \left(\hat{\sigma}^2 - \chi^2_{inf}\frac{\hat{\sigma}^2}{\sqrt{2n}}\frac{\sqrt{N-n}}{\sqrt{N}}\right) = 2a$$

$$\left(\chi^2_{sup} + \chi^2_{inf}\right)\frac{\hat{\sigma}^2}{\sqrt{2n}}\frac{\sqrt{N-n}}{\sqrt{N}} = 2a \qquad ; \qquad 2a\sqrt{N}\sqrt{2n} = \hat{\sigma}^2\left(\chi^2_{sup} + \chi^2_{inf}\right)\sqrt{N-n}$$

$$n = \frac{\hat{\sigma}^4\left(\chi^2_{sup} + \chi^2_{inf}\right)^2 N}{8a^2N + \hat{\sigma}^4\left(\chi^2_{sup} + \chi^2_{inf}\right)^2}$$

Introducendo i gradi di libertà ν e ricordando che: $E(\chi^2) = \nu$ per cui: $\chi^2_{sup} + \chi^2_{inf} \cong 2\nu$ si ottiene

[47] Allo stesso risultato, si poteva giungere, operando in ambito lineare, a partire dall'espressione dell'intervallo di confidenza nella quale trascurare il termine $1/2n$:

$$2z_{\alpha\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}\frac{N-n}{N}} = 2a \qquad \text{al quadrato:} \qquad z^2_{\alpha\alpha}\hat{p}(1-\hat{p})\frac{N-n}{N} = a^2n \qquad \text{ovvero:} \qquad n = \frac{Nz^2_{\alpha\alpha}\hat{p}(1-\hat{p})}{z^2_{\alpha\alpha}\hat{p}(1-\hat{p}) + a^2N}$$

Adottando le stesse convenzioni numeriche, già assunte in precedenza, si ottiene: $n = N/1 + a^2N,$, risultato, come già detto, identico a quello ottenuto, operando in ambito quadratico.

un'espressione, dalla struttura formale uguale alla precedente:

$$n = \frac{\nu^2 N}{\nu^2 + 2a^2 N}$$

la quale fornisce il valore atteso per la numerosità del campione, nel caso di test per la distribuzione χ^2.

La tripartizione cui si è accennato nella strategia di campionamento proposta costituisce l'approssimazione lineare di una curva logistica, avente due asintoti orizzontali pari ai due valori di soglia ed un punto intermedio di flesso a tangente inclinata, dove il coefficiente angolare della retta tangente coincide con il suddetto valore percentile. Sarebbe desiderabile poter presentare valori numerici generali per i parametri rappresentativi della curva logistica: ciò presenta alcune difficoltà, per la variabilità, da caso a caso, di tali parametri. Tuttavia sono evidenti alcune costanti, all'interno di ragionevoli intervalli di precisione richiesti.

Si può osservare l'esistenza di un limite inferiore per la numerosità del campione, corrispondente ad un valore al di sotto del quale è necessario sottoporre a test tutti gli elementi della popolazione, e di un limite superiore, corrispondente ad un valore al di sopra del quale il numero degli elementi della popolazione da sottoporre a test è approssimativamente lo stesso. Si osservi tuttavia come aumentino, all'aumentare della precisione richiesta, sia il limite inferiore, sia quello superiore. Per quanto riguarda il coefficiente angolare c, si può procedere per tentativi, a partire da un valore iniziale ottenuto dal confronto tra il valore di K, ascissa della curva logistica nel punto di flesso, ed il valore di N, corrispondente a $n=y(K)$.

Da ultimo, ricollegandosi al problema delle stime di covarianza e alle tecniche di ammassamento o clumping, appare chiaro, come e perché la prima e più importane proprietà richiesta ad un campionamento sia la sua Bernoullianità, ovvero l'indipendenza fra dati campionati. Infatti la loro eventuale non – indipendenza invalida il noto decremento della varianza di campionamento, in funzione della numerosità del campione, per effetto delle correlazioni fra i dati stessi (limitatamente all'ambito lineare).

D'altra parte, l'assoluta garanzia della Bernoullianità di un campionamento è cosa assai difficile da essere assicurata, specialmente per grandi campioni, quando i dati sono raccolti in modo denso e/o fitto. Tutto ciò porta inevitabilmente a dover accettare un limite inferiore alla varianza delle stime delle statistiche, oggetto di studio, rendendo praticamente inutili campionamenti troppo estesi. Infatti per quanto raffinate siano metodologie e procedure di campionamento, innumerevoli sono le cause di connessione / correlazione fra le osservazioni di un campione (ambiente, tarature, altre strutture latenti).

Nei limiti di queste precisazioni, ricollegandosi specificamente al problema dell'analisi di varianza, un'utilissima strategia per ridurre, il più possibile, la varianza di campionamento è data dal campionamento stratificato o da quello a cluster. Infatti la stratificazione o clusterizzazione delle osservazioni, grazie al teorema di decompisizione ortogonale della varianza, permette di ridurre la varianza residua, mettendo in evidenza una o più varianze spiegate, frutto del campionamento a strati o a cluster.

Nel prosieguo, verrà illustrata, a titolo d'esempio, la riduzione della varianza residua della media campionaria. Infatti data la numerosità n del campione, acquisito con un campionamento semplice (o sistematico) di Bernoulli, non è possibile ridurre la varianza della media campionaria: $\sigma_\mu^2 = \sigma_x^2 / n$, se non aumentando la numerosità stessa.

Un'alternativa vantaggiosa, capace di diminuire detta varianza senza accrescere n, è data dallo schema di campionamento stratificato (o clusterizzato) di Poisson, dove la suddivisione in m strati (o cluster) risponde a caratteristiche di omogeneità di determinati attributi, opportunamente prescelti per la classificazione della popolazione in esame,

Nel caso di campionamento in blocco di una popolazione finita, essendo N_i la numerosità di ciascuno strato (e pertanto: $N = \sum N_i$, la numerosità della popolazione), il peso di ciascun strato (o cluster) è dato dal rapporto: $p_i = N_i / N$. La scelta più elementare è effettuare, strato per strato (o cluster per cluster) , il campionamento con una numerosità n_i direttamente proporzionale alla numerosità N_i dello strato (o del cluster): $n_i = nN_i / N = np_i$, avendo preliminarmente scelto e fissato la numerosità del campione: $n = \sum n_i$.

Come noto, il teorema di decomposizione ortogonale della varianza fissa l'identità fra la varianza generale σ_x^2 e la somma della varianza spiegata (dalla stratificazione o clusterizzazione) $\sigma_{\bar{x}}^2$ e della varianza residua $\bar{\sigma}_x^2$. Pertanto essendo la varianza residua sempre minore della varianza generale, la varianza della media campionaria: $\sigma_\mu^2 = \bar{\sigma}_x^2 / n$, di un campionamento stratificato (o clusterizzato) risulta certamente minore a quella ottenuta dal campionamento semplice (o sistematico). Infatti essendo, nell'ordine, la varianza generale, la varianza spiegata e la varianza residua:

$$\sigma_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \sigma_{\bar{x}}^2 = \frac{m}{m-1}\sum_{i=1}^{m} p_i(\bar{x}_i - \bar{x})^2 \qquad \bar{\sigma}_x^2 = \frac{(n-1)\sigma_x^2 - (m-1)\sigma_{\bar{x}}^2}{n-m}$$

la varianza della media campionaria (nell'ipotesi ragionevole che si abbia: $n >> m$) risulta:

$$\sigma_\mu^2 = \frac{\bar{\sigma}_x^2}{n} = \frac{(n-1)\sigma_x^2}{n(n-m)} - \frac{(m-1)\sigma_{\bar{x}}^2}{n(n-m)} < \frac{\sigma_x^2}{n}$$

e, tenuto conto del sopraccitato fattore di correzione per il campionamento finito, diventa:

$$\sigma_\mu^2 = \frac{\bar{\sigma}_x^2}{n}\frac{N-n}{N-1} = \left(\frac{(n-1)\sigma_x^2}{n(n-m)} - \frac{(m-1)\sigma_{\bar{x}}^2}{n(n-m)}\right)\frac{N-n}{N-1} < \frac{\sigma_x^2}{n}\frac{N-n}{N-1}$$

Nel caso di campionamento con ripetizione di una popolazione finita o infinita, il peso di ciascuno strato (o cluster) è dato dal rapporto: $p_i = \sigma_i / \sigma_0$, essendo σ_i la misura della dispersione del generico strato (o cluster), ovvero il suo sqm, $\sigma_0 = \sum p_i \sigma_i$, la media ponderata degli sqm (e: $\bar{\sigma}_x^2 = \sum p_i \sigma_i^2$, la media ponderata delle varianze).

Questo campionamento è da preferirsi anche per il caso di campionamento in blocco di una popolazione finita. Infatti il campionamento ottimale fa aumentare o diminuire la numerosità dello strato (o del cluster), dove più ampia o ristretta è la dispersione dei dati, mentre il campionamento proporzionale prende in considerazioni solo le numerosità degli strati (o dei cluster), prescindendo dalla loro dispersione. Di consegfuenza, una scelta più attenta è effettuare, strato per strato (o cluster per cluster), il campionamento con una numerosità n_i direttamente proporzionale alla dispersione σ_i dello strato (o del cluster), oltreché alla sua numerosità: $n_i = nN_i\sigma_i /(N\sigma_0) = np_i$, avendo preliminarmente scelto e fissato la numerosità del campione: $n = \sum n_i$ (nel caso di campionamento con ripetizione di una popolazione finita o infinita, il rapporto: N_i / N, vale: $1/m$). Introducendo la varianza degli sqm:

$$\sigma_\sigma^2 = \sum_{i=1}^{m} p_i(\sigma_i - \sigma_0)^2 = \sum_{i=1}^{m} p_i\sigma_i^2 - \sigma_0^2 = \overline{\sigma}_x^2 - \sigma_0^2$$

e supponendo una popolazione sufficientemente grande, la varianza della media campionaria risulta:

$$\sigma_\mu^2 = \frac{\sigma_\sigma^2}{n} = \frac{\overline{\sigma}_x^2}{n} - \frac{\sigma_0^2}{n} < \frac{\overline{\sigma}_x^2}{n} < \frac{\sigma_x^2}{n}$$

e, tenuto conto del sopraccitato fattore di correzione per il campionamento finito, diventa:

$$\sigma_\mu^2 = \frac{\sigma_\sigma^2}{n}\frac{N-n}{N-1} = \left(\frac{\overline{\sigma}_x^2}{n} - \frac{\sigma_0^2}{n}\right)\frac{N-n}{N-1} < \frac{\overline{\sigma}_x^2}{n}\frac{N-n}{N-1} < \frac{\sigma_x^2}{n}\frac{N-n}{N-1}$$

Tanto più grande la varianza degli sqm, quanto più elevato è l'ulteriore guadagno, ottenuto con il campionamento ottimale, rispetto al campionamento proporzionale, mentre i due campionamenti coincidono nel caso in cui ogni strato (o cluster) ha la stessa dispersione.

3. VALIDAZIONE DEI MODELLI[48]

I test di validazione dei modelli permettono di sottoporre a verifica, mediante opportuni controlli e confronti d'ipotesi, le stime effettuate come, del resto, tutti i risultati ottenuti nell'ambito della statistica. Al solito, si possono avere errori nel modello deterministico: presenza di errori grossolani nelle osservazioni, ed errori nel modello stocastico: presenza di errori sistematici nelle osservazioni, ovvero cattiva conoscenza delle varianze delle osservazioni e/o delle eventuali covarianze fra le osservazioni stesse. Un'opportuna sequenza di test permette di districarsi fra le varie cause d'errore.

3.1 Autoconsistenza e crossvalidazione

Un giudizio sui risultati può essere espresso in termini numerici e statistici. I controlli di tipo numerico

[48] Questa appendice riprende ed aggiorna quanto esposto in: Crippa B., Mussio L. (1990): Approccio non – deterministico alla

rispondono a problemi di condizionamento ed affidabilità che comunemente accompagnano e seguono il metodo dei minimi quadrati, comprensivo delle sue estensioni e generalizzazioni; pertanto tutto quanto riguarda i controlli di tipo numerico è considerato estraneo agli scopi del presente lavoro. I secondi comprendono i test statistici per la valutazione di osservazioni e parametri, della loro dispersione e, se del caso, della loro dipendenza. Per quanto riguarda le osservazioni, la validazione avviene in termini di entità degli scarti – residui (oltreché numericamente in termini di affidabilità delle osservazioni all'interno dello schema di misura) e consiste essenzialmente nell'individuazione ed eliminazione degli errori grossolani cui si rimanda.

Per quanto riguarda i parametri, la validazione riguarda la loro significatività oppure la loro corrispondenza o meno a valori nominali di riferimento, nonché il confronto (come nel caso dei cosiddetti problemi di controllo) tra due serie identiche degli stessi ricavate da due insiemi indipendenti di dati (oltreché numericamente in termini di buon condizionamento dei parametri nella configurazione delle osservazioni) e si esplica con test globali, parziali e locali. Ciò significa analizzare rispettivamente il rapporto fra la/e varianza/e spiegata/e dai parametri e la varianza residua (sigma zero quadrato), oppure un singolo parametro standardizzato con il proprio sqm, se del caso, avendo sottratto il/i corrispondente/i valore/i nominale/i di riferimento[49].

Questo tipo di controllo prende il nome di **autoconsistenza**, in quanto opera sull'insieme/i dei dati che sono stati immessi nel calcolo. Infatti, come noto, una prima verifica può ottenersi già raffrontando il quadrato della/e stima/e di sigma zero con il quadrato del valore di riferimento dello stesso, assunto a priori, oppure raffrontando tra loro le due stime effettuate nelle condizioni specificate più oltre. Il test ordinario sulla varianza, facente uso della distribuzione *chi quadrato* con gradi di libertà pari al numero di osservazioni meno il numero dei parametri, permette di valutare la rispondenza o meno della stima effettuata all'ipotesi fondamentale assunta, mentre il test ordinario sul rapporto di varianze, facente uso della distribuzione *F di Fisher* con gradi di libertà sempre pari al numero delle osservazioni meno il numero dei parametri, permette il confronto fra due stime effettuate a partire da insiemi indipendenti di dati[50].

---

costruzione di un modello digitale delle altezze. Ricerche di *Geodesia, Topografia e Fotogrammetria*, n. 6. CLUP, Milano, p. 6–91.

[49] Test globale su tutti i parametri per la loro significatività, la loro rispondenza a valori nominali di riferimento ed il confronto tra i valori di due identiche serie di parametri ricavati da due insiemi indipendenti di dati:

$$F_{\nu_1=n,\nu_2=m-n}=\frac{\hat{x}^T Q_{xx}^{-1}\hat{x}/n}{\hat{\sigma}_0^2} \qquad F_{\nu_1=n,\nu_2=m-n}=\frac{(\hat{x}-x)^T Q_{xx}^{-1}(\hat{x}-x)/n}{\hat{\sigma}_0^2}$$

$$F_{\nu_1=n,\nu_2=2(m-n)}=\frac{(\hat{x}_A-\hat{x}_B)^T Q_{xx}^{-1}(\hat{x}_A-\hat{x}_B)/n}{\hat{\sigma}_{0_A}^2+\hat{\sigma}_{0_B}^2}$$

Test parziale su una parte dei parametri per la loro significatività, la loro rispondenza a valori nominali di riferimento ed il confronto tra i valori due identiche serie di parametri ricavati da due insiemi indipendenti di dati:

$$F_{\nu_1=n'<n,\nu_2=m-n}=\frac{\hat{x}^T(\Pi Q_{xx}\Pi^T)^{-1}\hat{x}/n'}{\hat{\sigma}_0^2} \qquad F_{\nu_1=n'<n,\nu_2=m-n}=\frac{(\hat{x}-x)^T(\Pi Q_{xx}\Pi^T)^{-1}(\hat{x}-x)/n'}{\hat{\sigma}_0^2}$$

$$F_{\nu_1=n'<n,\nu_2=2(m-n)}=\frac{(\hat{x}_A-\hat{x}_B)^T(\Pi Q_{xx}\Pi^T)^{-1}(\hat{x}_A-\hat{x}_B)/n'}{\hat{\sigma}_{0_A}^2+\hat{\sigma}_{0_B}^2}$$

Test locale su un parametro per la sua significatività, la sua rispondenza ad un valore nominale di riferimento ed il confronto tra due valori di uno stesso parametro ricavati da due insiemi indipendenti di dati:

$$\hat{t}_{\nu=m-n}=\frac{\hat{x}}{\sigma_{\hat{x}}} \qquad \hat{t}_{\nu=m-n}=\frac{\hat{x}-x}{\sigma_{\hat{x}}} \qquad \hat{t}_{\nu=2(m-n)}=\frac{\hat{x}_A-\hat{x}_B}{\sqrt{\sigma_{\hat{x}_A}^2+\sigma_{\hat{x}_B}^2}}$$

[50] Test su sigma zero per la sua rispondenza ad un valore nominale di riferimento e sul confronto tra due valori di sigma zero ricavati da due insiemi indipendenti di dati

Si noti invece come, nel caso di un filtraggio, le correlazioni presenti fra la stima a posteriori della varianza del rumore residuo e la varianza del rumore, nota a priori (da considerare, a sua volta, una stima estratta da una variabile casuale), impediscano di operare allo stesso modo. Pertanto solo un confronto qualitativo fra l'entità della radice del valore quadratico medio dei dati $RMS(v)$ con l'entità della radice del valore quadratico medio del rumore residuo $RMS\ (n)$ fornirà un primo elemento di giudizio. Tanto più sarà stato possibile far risaltare il segnale stocastico riducendo così il rumore accidentale, quanto più la seconda sarà piccola rispetto alla prima. Il test del segno (di Thompson), per campioni qualsiasi (e pertanto anche non indipendenti) sulla dispersione dei dati, permette di valutare il comportamento del rumore residuo rispetto ai dati d'ingresso, valutandone l'auspicato "miglioramento" (come nel caso dei cosiddetti studi “prima e dopo”, cioè quando si osserva due volte lo stesso campione).

Un controllo più rigoroso si può essere effettuato non utilizzando tutti i dati a disposizione, bensì riservandone alcuni (ad es., da qualche per cento fino ad un massimo del 10%, per insiemi di dati veramente molto grandi) per una verifica separata, detta **crossvalidazione**. Infatti effettuata l'estrapolazione e/o la predizione per tutte queste osservazioni e calcolata la differenza fra il dato di partenza ed il valore estrapolato e/o il segnale predetto, si ottiene un insieme di discrepanze, opportunamente piccole, se il precedente trattamento delle osservazioni ha fornito buoni risultati.

Date le differenze fra i dati di partenza ed i valori estrapolati a fini di controllo: $\hat{\delta} = y_e^0 - \hat{y}_e$, è possibile valutare la loro entità, sottoponendole al test ordinario sulla media, facente uso della distribuzione *t di Student* con gradi di libertà pari al numero delle osservazioni più uno meno il numero dei parametri:

$$\hat{t}_{\nu=m+1-n} = \frac{\hat{\delta}}{\sigma_{\hat{\delta}}} \qquad \text{essendo:} \qquad \sigma_{\hat{\delta}}^2 = \sigma_{y_e^0}^2 + \sigma_{\hat{y}_e}^2$$

Il test ordinario sulla varianza di queste differenze, facente uso della distribuzione *chi quadrato* con gradi di libertà pari al numero delle differenze estrapolate, permette di giudicare, in senso statistico, la bontà della estrapolazione (crossvalidando così anche l’interpolazione e/o compensazione effettuata), mentre il test ordinario sul rapporto di varianze, facente uso della distribuzione *F di Fisher* con gradi di libertà sempre pari al numero delle differenze estrapolate, permette il confronto fra due stime effettuate a partire da insiemi indipendenti di dati:

$$\hat{\chi}^2_{\nu=m'} = m' \frac{\hat{\sigma}_{\delta}^2}{\sigma_0^2} \qquad F_{\nu_A=m'_A,\nu_B=m'_B} = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_B^2}$$

Nel caso di un filtraggio, date invece le differenze fra i dati di partenza ed i valori predetti a fini di controllo: $\hat{\delta} = v_p^0 - \hat{s}_p$, le correlazioni presenti fra la stima a posteriori della varianza del rumore residuo e la varianza

---

$$\hat{\chi}^2_{\nu=m-n} = (m-n)\hat{\sigma}_0^2 / \sigma_0^2 \qquad F_{\nu_A=m_A-n_A,\nu_B=m_B-n_B} = \hat{\sigma}_A^2 / \hat{\sigma}_B^2$$

del rumore, nota a priori (da considerare, a sua volta, una stima estratta da una variabile casuale), impediscono di operare allo stesso modo. Pertanto solo un confronto qualitativo fra l'entità della radice del valore quadratico medio dei dati cosiddetti di controllo $RMS(v_p)$ con l'entità della radice del valore quadratico medio delle differenze predette $RMS(n_p)$ fornirà un primo elemento di giudizio. Tanto più sarà stato possibile far risaltare il segnale stocastico predetto riducendo così le differenze predette, quanto più la seconda sarà piccola rispetto alla prima. Il test del segno (di Thompson), per campioni qualsiasi (e pertanto anche non indipendenti) sulla dispersione dei dati, permette di valutare il comportamento delle differenze predette rispetto ai dati d'ingresso, valutandone l'auspicato "miglioramento" (come nel caso dei cosiddetti studi “prima e dopo”, cioè quando si osserva due volte lo stesso campione).
Di seguito, considerando nota da informazioni a priori la varianza del rumore, viene esposto un test “ad hoc” sulla varianza di queste differenze che permette di giudicare, in senso statistico, la bontà della predizione (crossvalidando così anche il filtraggio effettuato). Data la norma quadratica delle differenze predette, definita dalla matrice inversa della loro matrice di covarianza[51]:

$$E\left(\delta^T\, Q_{\delta\delta}^{-1}\, \delta\right) = Tr\left(Q_{\delta\delta}^{-1}\, E\left(\delta\,\delta^T\right)\right) = \sigma_n^2\; Tr\left(Q_{\delta\delta}^{-1}\, Q_{\delta\delta}\right) = \sigma_n^2\; Tr\, I_{(m)} = m'\,\sigma_n^2$$

una sua stima corretta e consistente si ottiene, come usuale, sostituendo la stima delle differenze ai loro valori teorici: $\hat{\sigma}_n^2 = \hat{\delta}^T Q_{\delta\delta}^{-1}\, \hat{\delta} / m'$.

Per agevolare il calcolo della matrice inversa della suddetta matrice di covarianza, si ponga attenzione alle seguenti considerazioni. La matrice di covarianza di queste differenze si ottiene applicando la legge di propagazione della covarianza all’espressione della stima delle differenze predette:

$$C_{\hat{\delta}\hat{\delta}} = C_{v_p v_p} + C_{s_p s}\, C_{vv}^{-1}\, C_{ss_p} - 2\, C_{s_p s}\, C_{vv}^{-1}\, C_{ss_p} = C_{v_p v_p} - C_{s_p s}\, C_{vv}^{-1}\, C_{ss_p}$$

Se le differenze sono ottenute negli stessi punti di filtraggio, utilizzando una seconda serie di dati, oppure altri punti in numero minore o uguale ai punti di filtraggio e non molto lontani da questi (ad es., inferiore al 20% della lunghezza di correlazione[52]), allora si può convenientemente effettuare la seguente approssimazione:

$$\begin{aligned} C_{\delta\delta} &= C_{v_p v_p} - C_{s_p s}\, C_{vv}^{-1}\, C_{ss_p} \cong C_{vv} - C_{ss}\, C_{vv}^{-1}\, C_{ss} = C_{vv} - \left(C_{ss} - C_{ee}\right) = \\ &= \left(C_{ss} + \sigma_n^2\, I\right) - \left(C_{ss} - \sigma_n^2\left(I - \sigma_n^2\, C_{vv}^{-1}\right)\right) = \sigma_n^2\, C_{vv}^{-1}\left(2C_{vv} - \sigma_n^2\, I\right) \cong 2\sigma_n^2\, I \end{aligned}$$

---

[51] Si chiama norma quadratica di un vettore lo scalare: $\sqrt{v^T Q\, v} = \sqrt{\Sigma_i\, q_{ii} v_i^2 + 2\,\Sigma_i \Sigma_j q_{ij} v_i v_j}$, essendo Q una qualsiasi matrice definita positiva che, una volta scelta, definisce il tipo di norma. Si ricordi che si definisce matrice definita positiva una matrice quadrata Q che soddisfi le relazioni:

$u^T Q u \geq 0 \quad \forall\, u$

$u^T Q u = \rightarrow u = 0$

Si chiama norma euclidea di un vettore v lo scalare: $\sqrt{v^T v} = \sqrt{\Sigma_i\, v_i^2}$, indicata con il simbolo $\| v \|$.

[52] Si chiama lunghezza di correlazione la distanza, corrispondente all'ascissa della funzione di covarianza, dove la covarianza del segnale risulta uguale alla metà della varianza del processo.

In questa espressione, la prima approssimazione è fatta scambiando, in termini di varianze e covarianze, il segnale predetto con quello filtrato coincidente o più vicino[53]. Di conseguenza, la stima della suddetta norma quadratica ha espressione:

$$\hat{\sigma}_n^2 = \frac{\hat{\delta}^T \hat{\delta}}{2m'} \qquad \text{da cui ancora:} \qquad \hat{\chi}^2_{\nu=m'} = m' \frac{\hat{\sigma}_\delta^2}{\sigma_0^2}$$

per l'esecuzione del sopraccitato test sulla varianza, facente uso della distribuzione *chi quadrato* con gradi di libertà pari al numero delle differenze predette.

Da ultimo, si badi che, nel caso di confronti fra metodologie, procedure ed algoritmi, da effettuarsi ovviamente con gli stessi insiemi di dati, data la ovvia non indipendenza fra i dati, sia necessario effettuare la comparazione dei risultati ottenuti con i test del segno (di Thompson), per campioni qualsiasi, sui valori centrali e/o sulla dispersione. Si noti, come questi confronti non siano infrequenti; al contrario, essi sono richiesti ogniqualvolta si voglia valutare l'effettivo miglioramento ottenibile con un cambio di metodologie, procedure od algoritmi. In questo caso infatti, ragioni di continuità delle serie storiche, oltre alla necessaria cautela nell'accettazione del rinnovamento proposto, richiedono di affiancare, per un certo periodo, due soluzioni comparandone i risultati, al fine di ottenere elementi di giudizio.

3.2 Identificazione ed eliminazione dei dati anomali

Accade talvolta che si abbiano molti dati anomali la cui presenza fa sì che qualsiasi operazione preliminare (ad es., il calcolo di valori approssimati dei parametri, la valutazione di correzioni da apportare alle osservazioni, ove queste si rendano necessarie, le stime empiriche delle eventuali funzioni di covarianza, in particolare, le covarianze calcolate per i dati più vicini rispetto alla varianza del segnale[54]) risulti fortemente alterata. Per superare questo inconveniente, è necessario mettere a punto, caso per caso, una strategia che consenta **procedure robuste di esplorazione preliminare**. Infatti una strategia generale non è nota e la sua definizione cozza con l'elevata non – linearità del problema.

Una semplificazione, a riguardo, può ottenersi se si hanno informazioni, a priori, su l'entità della varianza delle osservazioni (o del loro rumore, se si suppone che esse contengano anche un segnale stocastico). In questo caso specifico, dopo aver individuato un opportuno intorno r_1, la differenza fra lo scarto – residuo di un'osservazione qualsiasi v_i e la media degli scarti – residui delle osservazioni circostanti v_j, siti nel relativo intorno:

$$\delta_i \cong v_i - \frac{1}{m_i} \sum_{j=1}^{m_i} v_j \qquad j \neq i \qquad \text{e} \qquad 0 \leq d_{ij} \leq r_1 \qquad \forall\, i$$

[53] La seconda approssimazione ha invece il suo fondamento dell'essere, in generale: $2\,Tr\left(C_{vv}\right) = 2\,m\sigma_v^2 >> m\sigma_n^2 = \sigma_n^2\,Tr\left(I\right)$, ovvero: $\sigma_n^2 << \sigma_s^2 + \sigma_n^2 = \sigma_v^2 < 2\,\sigma_v^2$. Pertanto essendo: $\sigma_\delta^2 \cong 2\sigma_n^2$, e risultando tutte le differenze predette pressoché indipendenti fra loro, il calcolo della matrice inversa della loro matrice di covarianza non presenta ovviamente alcun problema.

[54] Con specifico riferimento alla stima empirica delle funzioni di covarianza, come accade ad esempio nello studio dei modelli digitali delle altezze, è la forma della funzione di covarianza a risultare fortemente depressa.

costituisce una stima, anche se grossolanamente approssimata, della discrepanza δ_i (o del rumore n_i) di quella data osservazione.

Si osservi innanzitutto come l'individuazione di un opportuno intorno sia agevole quando si operi con dati intensivi (cioè puntuali) a referenza spaziale forte (indotta dalle coordinate[55]), mentre diventi molto più complessa in tutti gli altri casi. Si noti poi che gli scarti – residui delle osservazioni sono solitamente un prodotto finale del trattamento delle osservazioni, pertanto la loro richiesta, all'inizio dello stesso, richiede il calcolo di soluzioni approssimate e/o parziali e/o locali, per poter disporre da subito degli stessi.

Si ricordi che alternative, maggiormente robuste e pertanto più sicure, alla media delle osservazioni circostanti sono la mediana o la media potata (in questo secondo caso, con un opportuna definizione dei pesi delle osservazioni, mettedo a zero le osservazioni anomale che si sospettano affette da errori grossolani) delle stesse, cosa che permette un'individuazione e rimozione di eventuali errori grossolani, presenti al loro interno, per poter poi concludere, senza effetti di mascheramento, la sopraccitata operazione di media mobile delle osservazioni circostanti una data osservazione.

Se i dati di partenza sono distribuiti normalmente, allora anche la stima appena fatta della discrepanza (o del rumore) è distribuita normalmente con media nulla e varianza nota, assunta da informazioni a priori. Infatti come noto, qualsiasi combinazione lineare di variabili casuali normali è una variabile casuale normale. Così il test ordinario sulla media per campioni numerosi, nell'ipotesi fondamentale nulla, consente di mettere in evidenza, un'osservazione alla volta, tutti quei dati significativamente anomali.

Inoltre capita spesso che, fra i dati utilizzati, come fra quelli riservati per il controllo, si abbiano alcuni dati anomali (e talvolta anche in quantità abbastanza rilevante), da interpretarsi spesso come errori grossolani. L'identificazione e l'eliminazione (con una procedura comunemente detta **data snooping**) di questi che, per loro relativa piccolezza, non inficiano certamente le procedure di esplorazione preliminare, per lo più, è abbastanza agevole. Infatti dati gli scarti – residui delle equazioni d'osservazione, standardizzati per il relativo sqm, essi possono essere sottoposti alla verifica d'ipotesi, globalmente con il test multiplo di Pearson et al. e singolarmente con un test "ad hoc" per uno scarto – residuo.

Il test multiplo di Pearson et al., eseguito su un'opportuna combinazione quadratica dei coefficienti di asimmetria (γ) e di curtosi (β)[56], aventi ciascuno distribuzione asintoticamente normale, rispettivamente con media: $\mu_\gamma = 0$, $\mu_\beta = 3$, e varianza: $\sigma_\gamma^2 = 6/m$, $\sigma_\beta^2 = 24/m$, dove m è la dimensione del campione in esame, ed essendo gli stessi coefficienti stocasticamente indipendenti fra loro, fa uso della distribuzione *chi quadrato* con due gradi di libertà:

$$\hat{\chi}^2_{\nu=2} = m\left(\frac{\hat{\gamma}^2}{6} + \frac{\left(\hat{\beta}-3\right)^2}{24}\right)$$

[55] Una referenza spaziale debole può essere indotta dagli indirizzi o altri indicatori di località (o posizione), non aventi le proprietà metriche delle coordinate.

[56] Coefficienti di asimmetria e di curtosi:

$$\hat{\gamma} = \sum_{i=1}^{m}\frac{(x^0-\mu)^3}{m\sigma^3} = \sum_{i=1}^{m}\frac{\hat{v}^3}{m\sigma^3} = \sum_{i=1}^{m}\frac{\hat{z}^3}{m} \qquad \beta = \sum_{i=1}^{m}\frac{(x-\mu)^4}{m\sigma^4} = \sum_{i=1}^{m}\frac{\hat{v}^4}{m\sigma^4} = \sum_{i=1}^{m}\frac{\hat{z}^4}{m}$$

Un test “ad hoc” per uno scarto residuo, facente uso della distribuzione *tau di Thompson* con gradi di libertà pari al numero delle osservazioni “effettive” meno il numero dei parametri, consente di mettere in evidenza, un'osservazione alla volta, tutti quei dati significativamente anomali:

$$\hat{\tau}_{\substack{\nu=m'-n \\ (m'\leq m)}} = \frac{v}{\sigma_{\hat{v}}} \qquad \text{essendo:} \qquad \tau_\nu = \frac{\sqrt{\nu}\, t_{v-1}}{\sqrt{(\nu-1)+t_{v-1}^2}}$$

L’uso più appropriato di questo test è la cosiddetta **selezione all’indietro con minimi quadrati in cascata**, a partire dallo scarto residuo standardizzato maggiore, aggiornando poi, ogni volta, la soluzione e selezionando via, via tutti gli scarti residui standardizzati, in ordine decrescente di valore assoluto, fintantoché il massimo scarto residuo standardizzato corrente risulti accettabile. Si badi come questa procedura non sia affatto robusta, in quanto gli effetti di mascheramento, propri dei minimi quadrati, non permettono di evidenziare i veri dati anomali, distinguendoli da quelli apparenti (ma, in effetti, privi d’errore), specialmente se il numero dei dati anomali e/o l’entità delle anomalie è elevato/a.

Un’alternativa, particolarmente vantaggiosa, è costituita dalla cosiddetta **selezione in avanti dopo (opportune) procedure robuste**, a partire dallo scarto residuo standardizzato minore, fra quelli preventivamente esclusi in base alle suddette procedure, capaci di separare il grosso buono dei dati dai dati anomali (veri o presunti). Dopo la riammissione della corrispondente osservazione, nell’insieme costituito dal grosso buono dei dati, un test estremale, facente uso della distribuzione *H di Hawkins* con gradi di libertà pari al numero delle osservazioni “effettive” meno il numero dei parametri, consente di mettere in evidenza, un'osservazione alla volta (in particolare, valutando il massimo scarto residuo standardizzato fra quelli delle osservazioni “effettive”), tutti quei dati significativamente anomali, con maggior sicurezza (Tutto ciò dà ulteriore ragione alla proposta strategia mista, consistente nel far precedere opportune procedure robuste ad una selezione mirata in avanti, condotta in modo sequenziale con il metodo dei minimi quadrati, fino ad ottenere a minimi quadrati il risultato atteso.):

$$\hat{H}_{\substack{\nu=m'-n \\ (m'\leq m)}} = \max_{i=1,m'\leq m}\left(\frac{\hat{v}_i^2}{\nu\sigma_{\hat{v}_i}^2}\right) \qquad \text{essendo:} \qquad H_v = \frac{max\left(\chi_1^2\right)}{\chi_\nu^2}$$

Nel caso di un filtraggio, basta standardizzare rispettivamente il rumore o la differenza fra i dati di partenza e segnale predetto nei punti di controllo, avendone preventivamente calcolato il relativo sqm:

$$\sigma_{\bar{n}}^2 = \sigma_n^4\, diag\, C_{vv}^{-1} \qquad \text{e:} \qquad \sigma_{\bar{\delta}}^2 \cong 2\,\sigma_n^2$$

dove l'approssimazione è fatta scambiando, in termini di varianze e covarianze, ogni punto di controllo con quello di filtraggio coincidente o più vicino. Dopodiché il test ordinario sulla media per campioni numerosi, nell'ipotesi fondamentale nulla, permette di decidere, in senso statistico, un punto alla volta, l'eliminazione dei dati anomali.

3.3 Partizione in insiemi omogenei

I dati sono, talvolta, troppo poco omogenei fra loro, ovvero opportunamente suddivisi in insiemi, alcuni insiemi hanno varianze (o altri opportuni indicatori della dispersione) significativamente diverse dagli insiemi circostanti o comunque prossimi[57]. Allora è opportuno frazionare l'insieme dei dati, in tanti insiemi omogenei, quante sono le unioni connesse degli insiemi aventi pressoché le stesse varianze e le stesse scale delle eventuali funzioni di covarianza.

Nei casi più semplici, è consigliabile costruire gli insiemi mediante quadrettatura, particolarmente elementare se i dati sono disposti regolarmente o quasi. Dopodiché assunta una qualsiasi suddivisione di comodo delle classi di varianza, è immediato stabilire quali insiemi abbiano le stesse varianze e quali varianze significativamente diverse. Si osservi che problemi d'incertezza nell'attribuzione di un insieme ad una classe, dovuti all'arbitrarietà del posizionamento delle classi, sono quasi ininfluenti sui risultati.

L'attribuzione di ciascun insieme, ad una determinata classe, produce gli insiemi omogenei cercati. A tale scopo, tecniche gerarchiche ed agglomerative di cluster analysis (ad es., line – following o region growing) possono essere vantaggiosamente impiegate. Questa procedura si basa innanzitutto sull'assunzione che la varianza del segnale costituisca buona parte della varianza generale, essendo prevalente sulla varianza del rumore, considerata maggiormente costante per tutti gli insiemi, in quanto legata anche alla precisione di osservazione.

Alcuni test statistici permettono di discriminare ipotesi alternative nella partizione in insiemi omogenei. Infatti il test per l'analisi di varianza (di Fisher) ed il test per lo studio delle componenti della varianza (di Bartlett) permettono di giudicare l'omogeneità o meno, rispettivamente delle medie e delle varianze, per campioni indipendenti e normali, mentre i test del rango (di Kruskal – Wallis) permettono di giudicare altri indicatori di posizione e di dispersione, per campioni comunque indipendenti, anche se non necessariamente normali. Inoltre nel caso di eventuali classificazioni di coefficienti di correlazione (oppure di coefficienti di correlazione sui ranghi di Spearman), altri test statistici permettono di discriminare ipotesi alternative nella partizione in insiemi omogenei. Infatti il test per lo studio di struttura di covarianza (di Hotelling, in particolare con l'approssimazione di Lawley) permette di giudicare l'omogeneità o meno dei coefficienti di correlazione, per campioni normali, mentre il test di rango (di Wilcoxon – Willcox, sempre con l'approssimazione di Lawley) permette di giudicare i coefficienti di correlazione sui ranghi di Spearman, per campioni non necessariamente normali.

Si osservi comunque come sia spesso possibile effettuare il passaggio da dati non normali a dati normali, facendo uso di note funzioni di trasformazione della distribuzione di probabilità. Si ricordi infatti che, nel caso di **eventi rari** la cui distribuzione può essere assunta approssimativamente log – normale (mentre la distribuzione esatta è data, a rigore, dalla distribuzione Gamma di Erlang), è possibile passare a dati aventi distribuzione normale, calcolando il logaritmo degli stessi dati. Un analogo passaggio a dati aventi distribuzione normale, si ottiene nel caso delle **parti piccole** la cui distribuzione può essere assunta rettangolare, calcolando la funzione inversa dalla funzione trascendente (non elementare) erf.

3.4 Tecniche multilivello: eventuali rimozione della tendenza, prefiltraggio e secondo filtraggio[58]

[57] Con specifico riferimento ai modelli digitali delle altezze, si trovano zone non – omogenee, ad es., quando si opera in zone pedemontane o in talune fasce costiere.

[58] Le tecniche multilivello permettono di affrontare lo studio di vari fenomeni, aggredendoli un po' alla volta, così come la spiegazione viene costruendosi, passo dopo passo, ovvero livello dopo livello, con l'aumento della risoluzione (o della granularità, oppure

Le analisi statistiche più raffinate hanno bisogno, di dati a media nulla e privi di qualsiasi sistematismo generale (detto: tendenza; dall'inglese: trend) che, se presente[59], deve essere preventivamente rimosso, in generale con una (semplice) interpolazione polinomiale.

Inoltre, nel caso di un filtraggio, l'idea assurda di un filtraggio prima dello stesso indica un modo di procedere vantaggioso, al fine di ridurre la persistenza di correlazione[60]. Infatti un prefiltraggio è un'operazione del tutto simile al filtraggio, anche se meno efficiente, che fa dipendere il segnale in un punto qualsiasi principalmente dai valori osservati nei punti circostanti in un intorno prefissato. Ciò significa modellare una parte del segnale con procedimenti interpolativi locali (ad es., con il metodo degli elementi finiti). Si badi che il pre – filtraggio, proprio perché meno efficiente, in generale, non abolisce affatto il successivo filtraggio.

Infine, sempre nel caso di un filtraggio, accade talvolta che il segnale sia modellabile tramite due funzioni di covarianza: una atta a evidenziare correlazioni più lunghe (per un'entità maggiore di segnale) e l'altra atta a evidenziare correlazioni più corte (per un entità minore di segnale)[61]. In tal caso, il filtraggio separa solo il segnale maggiore, lasciando nel rumore – residuo anche il segnale minore. Allora calcolata una nuova funzione di covarianza, essa risulta colorata (ovvero mostra correlazione nel rumore – residuo). Un secondo filtraggio, in base alla nuova funzione di covarianza, separa anche il segnale minore dal rumore – residuo la cui funzione di covarianza risulta bianca (ovvero non mostra correlazione nel nuovo e più piccolo rumore – residuo).[62]

3.5 Individuazione di elementi caratteristici

Informazioni sulla dislocazione di elementi caratteristici, oltre a quelle già fornite dalla analisi degli scarti – residui standardizzati (o del rumore – residuo) standardizzato, possono essere ottenute calcolando anche le differenze finite prime e seconde delle stesse quantità. Infatti in un campo 2D, queste grandezze rappresentano rispettivamente la pendenza e la curvatura medie dell'errore nei dati e/o di modello nell'intorno del dato, di volta in volta, in esame.

Si osservi come, in campo 3D, lo stesso studio possa essere facilmente effettuato sui tre piani ortogonali, riferiti agli assi cartesiani, in cui lo spazio tridimensionale può essere scomposto, mentre lo studio dell'intero dominio a quattro dimensioni sia innegabilmente più complesso ed ovviamente non rappresentabile visivamente.

Il calcolo delle differenze finite è facile. Infatti indicato con z_{ij} la quantità standardizzata nel punto P_{ij} di una

---

dell'ordine) insieme al miglioramento della precisone ed accuratezza ed alla riduzione della zona (o intervallo, oppure altro dominio del problema) in esame. Tutto ciò permette di fornire il tipo di spiegazione, richiesto ad un certo livello, adottando il modello più adeguato (per il dato il livello d'elaborazione) facendo uso di un numero ben adeguato di dati, con il campionamento sequenziale: progressivo e selettivo degli stessi per campi di punti (oppure con la progettazione ed ottimizzazione interattiva della configurazione e dello schema di misura, nel caso di strutture reticolari).

[59] Con specifico riferimento ai modelli digitali delle altezze, si trova una tendenza nei dati, ad es., quando si opera su piccole aree (una montagna, una collina, un versante, una valle, ecc.).

[60] Si definisce persistenza di correlazione il numero di passi della funzione di covarianza, prima che questa assuma definitivamente valori pressochè trascurabili (ad es., inferiori al $5\% \div 10\%$ della varianza generale). Si ricordi inoltre che si chiamano lunghezza di correlazione la distanza alla quale la funzione di covarianza vale (eventualmente una prima volta) la metà della varianza generale e punto di annullamento la distanza alla quale la funzione di covarianza si annulla, se ciò avviene (se ciò avviene più volte, allora la prima volta).

[61] Con specifico riferimento ai modelli digitali delle altezze, occorre procedere ad un secondo filtraggio, ad es., quando si opera su con dati molto fitti, capaci di mettere in evidenza tanto un segnale maggiore, avente una lunghezza di correlazione relativamente lunga, quanto un segnale minore, avente una lunghezza di correlazione molto più corta (anche se il suo contributo alla modellazione del fenomeno non è trascurabile). Infatti in tal caso, omettere questo secondo passo di correlazione significa lasciare il segnale minore indistinto dal rumore – residuo, cosa che ne amplifica il valore e ne cambia la natura.

[62] Sui risultati ottenuti è ancora possibile esprimere un giudizio, in termini numerici e statistici. I controlli di tipo numerico rispondono, anche in questo caso, a problemi di condizionamento ed affidabilità che comunemente accompagnano e seguono il metodo dei minimi quadrati, comprensivo delle sue estensioni e generalizzazioni. A loro volta, i secondi comprendono sempre i test statistici per la

griglia regolare a maglie quadrate di lato l[63], le differenze finite prime della stessa si calcolano con le seguenti espressioni:

$$\Delta z_{i,j/i} = \frac{z_{i+1,j} - z_{i-1,j}}{2l} \qquad \Delta z_{i,j/j} = \frac{z_{i,j+1} - z_{i,j-1}}{2l}$$

e le differenze finite seconde con le successive:

$$\Delta^2 z_{i,j/ii} = \frac{z_{i+1,j} - 2z_{ij} + z_{i-1,j}}{l^2} \qquad \Delta^2 z_{i,j/jj} = \frac{z_{i,j+1} - 2z_{ij} + z_{i,j-1}}{l^2}$$

$$\Delta^2 z_{i,j/ij} = \Delta^2 z_{i,j/ji} = \frac{z_{i+1,j+1} - z_{i+1\,j-1} - z_{i-1\,j+1} + z_{i-1,j-1}}{l^2}$$

Con specifico riferimento ai vari tipi di elementi caratteristici in un campo 2D, dove gli elementi caratteristici sono linee caratteristiche, la figura 3.1 mostra l'andamento della quantità standardizzata, della sua pendenza e della sua curvatura lungo una sezione ortogonale alla linea caratteristica oggetto di studio[64].
Le due differenze prime costituiscono le componenti del vettore Jacobiano, ovvero del vettore gradiente:

$$J = \vec{g} = \begin{bmatrix} \Delta z_{ij/i} \\ \Delta z_{ij/j} \end{bmatrix}$$

avente modulo, ovvero norma:

$$\|g\| = |g| = \sqrt{\left(\Delta_{ij/i}\right)^2 + \left(\Delta_{ij/j}\right)^2}$$

essendo la direzione di massima pendenza e la direzione della tangente alla curva di livello (isoipsa):

$$\vartheta_{max} = arc\,tan \frac{\Delta z_{ij/i}}{\Delta z_{ij/j}} \qquad \vartheta_0 = \vartheta_{max} + \pi/2$$

---

valutazione di osservazioni e parametri, della loro dispersione e, se del caso, della loro dipendenza.

[63] Si osservi che si hanno sempre punti disposti su una griglia (quasi) quadrata, quando si esegue la scansione di immagini, non invece quando si procede con la digitalizzazione di mappe. Tuttavia poiché, nel secondo caso, i dati di partenza sono ottenuti mediante discretizzazione delle curve di livello; a fronte di un lavoro di produzione così ingente, la discretizzazione anche delle linee caratteristiche di rottura della superficie, solitamente riportate sulle mappe, non accresce molto il lavoro di produzione. Inoltre, anche in caso di mancanza di queste, è comunque possibile avere punti disposti su una griglia regolare a maglie quadrate, interpolando opportunamente il rumore standardizzato di tutti i punti posti in un opportuno intorno. Si noti a proposito come l'interpolazione con un piano sia adatta allo studio della pendenza, mentre l'interpolazione con una quadrica di tipo funzione (ad es., paraboloide ellittico o paraboloide iperbolico o cilindrico parabolico) è indicata allo studio della curvatura. Infine nel caso, per altro assai poco frequente di punti sparsi e radi, nessuna informazione sulle linee caratteristiche di rottura della superficie può essere fornita.

[64] Con specifico riferimento alla costruzione dei modelli digitali delle altezze, può accadere che si abbiano parecchi dati anomali, non sufficienti ad inficiare le varie e diverse operazioni preliminari, ma certamente troppo abbondanti e troppo "curiosamente" distribuiti per essere considerati errori grossolani. A proposito, occorre tenere presente che il terreno ha, per sua propria natura, certe linee (punti) caratteristiche: verso l'alto (picchi, creste, ecc.), verso il basso (fosse, valli, impluvi, ecc.) di transizione alto – basso (balze, scarpate, ecc.). Come evidente, si tratta di linee di rottura della continuità o della graduale variabilità della superficie del terreno che, come tali, non possono essere ricostruite con un qualsiasi metodo di interpolazione.

Le quattro differenze seconde costituiscono gli elementi della matrice Hessiana:

$$H = \begin{bmatrix} \Delta^2 z_{ij/ii} & \Delta^2 z_{ij/ij} \\ \Delta^2 z_{ij/ji} & \Delta^2 z_{ij/jj} \end{bmatrix}$$

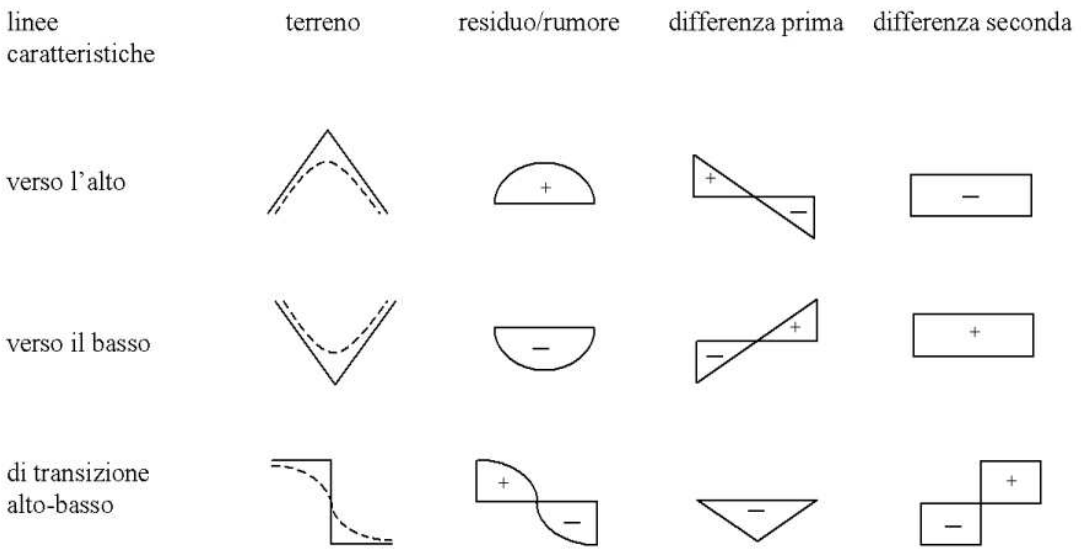


Fig. 3.4

Gli autovalori massimo e minimo di questa matrice, facilmente ottenibili calcolando il suo determinante, ovvero risolvendo l'equazione caratteristica: $Det(H - \lambda I) = 0$,

$$\lambda_{max}, \lambda_{min} = \frac{\Delta^2 z_{ij/ii} + \Delta^2 z_{ij/jj}}{2} \pm \frac{1}{2}\sqrt{\left(\Delta^2 z_{ij/ii} - \Delta^2 z_{ij/jj}\right)^2 + 4\left(\Delta^2 z_{ij/ij}\right)^2}$$

rappresentano rispettivamente la curvatura massima e la curvatura minima. Si osservi, a riguardo, che il determinante Hessiano della matrice omonima: $Det\ H = \Delta^2 z_{ij/ii}\Delta^2 z_{ij/jj} - \left(\Delta^2 z_{ij/ij}\right)^2$, può essere minore di zero, uguale a zero e maggiore di zero, rispettivamente in un punto iperbolico, in un punto parabolico e in un punto ellittico. Nel primo caso, se in quel punto la pendenza è nulla, si ha una sella, mentre nel terzo caso, se in quel punto la pendenza è nulla, si ha un massimo, se $\Delta^2 z_{ij/ii}, \Delta^2 z_{ij/jj} < 0$, e un minimo, se $\Delta^2 z_{ij/ii}, \Delta^2 z_{ij/jj} > 0$. Nei punti ellittici, la curvatura Gaussiana è data dalla media geometrica delle curvature massima e minima:

$$\lambda_{med} = \sqrt{\lambda_{max}\lambda_{min}}$$

e indica la curvatura di una sfera che meglio aderisce alla superficie in quel punto.
Un altro importante funzionale della matrice Hessiana è il Laplaciano, ovvero la traccia della matrice omonima:

$$\nabla^2 z = Tr\ H = \Delta^2 z_{ij/ii} + \Delta^2 z_{ij/jj}$$

che rappresenta una dilatazione o contrazione, mentre:

$$\gamma = \sqrt{\left(\Delta^2 z_{ij/ii} - \Delta^2 z_{ij/jj}\right)^2 + 4\left(\Delta^2 z_{ij/ij}\right)^2}$$

è il massimo fattore di taglio.

4. TEST SEQUENZIALI

Nonostante ad oggi solo pochi autori abbiano dedicato la loro attenzione ai test sequenziali, questo tipo di analisi dei dati sperimentali non è certo un tema nuovo per la statistica. Infatti in letteratura, si trovano alcuni esempi di test sequenziali, tuttavia le applicazioni sono elaborate solo per una piccola parte dei test più comunemente utilizzati.

Poiché l'inferenza statistica nasce come strumento utile per la verifica di ipotesi espresse sull'adattamento di dati a distribuzioni di probabilità, su parametri o, più in generale, sulla bontà dei modelli d'interpretazione delle osservazioni, l'applicazione di un test necessita un campione di dati significativo e rappresentativo della popolazione da cui è considerato estratto. Questo fatto può comportare l'esigenza di collezionare un numero elevato di estrazioni, prima di poter espressionere ipotesi sensate e soprattutto svolgere il test. Il vantaggio dei test sequenziali è quello di permettere la loro esecuzione già con pochissimi dati (in alcuni casi già con tre dati), arrivando a soluzione con un numero minore di estrazioni rispetto ai test tradizionali.

Infatti il numero di osservazioni su cui condurre il test non è fissato a priori, ma è determinato nel corso dell'esperimento. Il test è così effettuato dopo ogni osservazione (o gruppo di osservazioni) sull'insieme dei dati accumulati, fino a quel momento, e prosegue fino a quando non è possibile decidere quale ipotesi accettare.

Poiché lo scopo di questo tipo di analisi è quello di arrivare a scegliere, tra ipotesi alternative, con il minimo numero di osservazioni, i test sequenziali sono costruiti in modo da poter rappresentare graficamente, passo dopo passo, i risultati delle osservazioni, in funzione del numero di prove effettuate. Benché le funzioni di merito che si possono costruire siano numerose, si è scelto di utilizzare l'approccio conosciuto come test sequenziale del rapporto di verosimiglianza, poiché esso è applicabile a tutte le tipologie di test senza nessun tipo di adattamento.

L'idea è ottenere un grafico su cui siano riconoscibili due linee di confine che lo suddividano in tre aree, nel caso comune di una sola ipotesi alternativa:

- la regione di accettazione dell'ipotesi fondamentale H_0
- la regione (intermedia) del dubbio
- la regione di accettazione dell'ipotesi alternativa H_1

Le linee di confine λ_0 e λ_1 sono ricavate in funzione dell'entità dei rischi α (livello di significatività od errore di prima specie) e β (potenza del test od errore di seconda specie):

$$\lambda_0 = (1-\alpha)/\beta \qquad\qquad \lambda_1 = \alpha/(1-\beta)$$

La funzione λ che interpreta i risultati delle prove, è detta rapporto di verosimiglianza di Fisher e vale:

$$\lambda = \frac{P(\underline{x} \Rightarrow se\ H_0\ è\ vera)}{P(\underline{x} \Rightarrow se\ H_1\ è\ vera)}$$

Se la funzione di distribuzione di probabilità è continua, il rapporto di verosimiglianza si fa tra le densità di probabilità composte, ovvero fra i prodotti delle probabilità elementari per campioni Bernoulliani:

$$\lambda = \prod_{i=1}^{n} p_0(x_i) \Big/ \prod_{i=1}^{n} p_1(x_i) \qquad\qquad ln\ \lambda = \sum_{i=1}^{n} ln(p_0(x_i)) - \sum_{i=1}^{n} ln(p_1(x_i))$$

dove la forma logaritmica è particolarmente utile, se la funzione di distribuzione di probabilità adottata è di classe esponenziale.

Di norma, la funzione si muove inizialmente nella regione del dubbio, per poi dirigersi in una delle due regioni di accettazione: a questo punto, il test può essere interrotto, perché è arrivato ad una soluzione e, in probabilità, sarà accettata l'ipotesi relativa alla regione interessata (un esempio in figura 4.1)

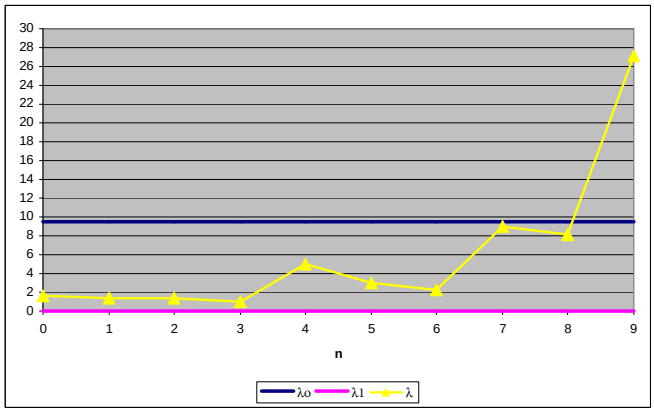


Fig. 4.1 – Un esempio di test sequenziale

Nel prosieguo, test per test (o per gruppi di test omogenei), è presentata l'applicazione dei test sequenziali, partendo dall'espressione del test tradizionale, fino ad arrivare ad ottenere il rapporto di verosimiglianza, avendo scelto e costruito due ipotesi alternative.

<u>Test su campioni numerosi</u>

I test parametrici semplici su campioni numerosi sottopongono a verifica d'ipotesi un solo parametro: la media o la differenza fra due medie, a varianza/e nota/e od incognita/e (considerata/e comunque stima/e corretta/e e consistente/i della/e varianza/e teorica/he, data la numerosità del/i campione/i). L'applicazione dei test sequenziali è molto semplice, in quanto utilizza l'espressione del test tradizionale, di volta in volta, inserendo la/e media/e ipotizzata/e nei due casi alternativi e calcolando, ad ogni passo di campionamento, la variabile casuale z nelle ipotesi H_0 e H_1. Dopodiché calcolate, passo dopo passo, le funzioni densità di probabilità normali nei due casi, il rapporto di verosimiglianza λ è rappresentato graficamente, in modo da

poter essere messo a confronto direttamente con le rette di confine λ_0 e λ_1, finché non si arriva ad una soluzione:

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \qquad z_1 = \frac{\bar{x} - \mu_1}{\sigma / \sqrt{n}} \qquad \text{da cui} \qquad \lambda = e^{-\frac{1}{2}\sum_{i=1}^{n} z_{i,0}^2} \Big/ e^{-\frac{1}{2}\sum_{i=1}^{n} z_{i,1}^2}$$

$$z_0 = \frac{\Delta\bar{x} - \Delta\mu_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad z_1 = \frac{\Delta\bar{x} - \Delta\mu_1}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad \text{da cui} \qquad \lambda = e^{-\frac{1}{2}\sum_{i=1}^{n} z_{i,0}^2} \Big/ e^{-\frac{1}{2}\sum_{i=1}^{n} z_{i,1}^2}$$

essendo solitamente nell'ipotesi fondamentale: $\Delta\mu_0 = 0$.

<u>Test su campioni normali</u>

Anche per questi test, effettuati su media, varianza e coefficiente di correlazione, si considerano due casi:

- test eseguito su un campione, verificando la rispondenza del parametro prescelto ad un valore di riferimento scelto e fissato;
- test eseguito su due campioni, verificando la corrispondenza (solitamente l'uguaglianza) dei rispettivi parametri.

Nel primo caso, è semplice costruire l'ipotesi fondamentale H_0 e quella alternativa H_1, poiché si impone, in entrambi i casi, l'uguaglianza del parametro ad un valore di riferimento scelto e fissato. Pertanto il test sequenziale è costruito, passo dopo passo,

- andando a sostituire nella espressione del test il valore del parametro nelle due ipotesi;
- ricavando le variabili casuali t di Student per la media ($t = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$) e χ^2 per la varianza ($\chi^2 = \frac{\nu\hat{\sigma}^2}{\hat{\sigma}^2}$);
- calcolando le rispettive funzioni densità di probabilità;
- aggiornando il rapporto di verosimiglianza λ, come aggiornamento di quello calcolato al passo precedente;
- facendo il confronto rette di confine λ_0 e λ_1.

Nel secondo caso, occorre modificare l'espressione dei test, evidenziando anche l'ipotesi alternativa, ovvero nei test sul confronto di medie, classico (di Gosset) con uguale varianza e di Welch con diversa varianza, si ha:

$$H_0: \quad \Delta\mu_0 = 0 \qquad t_0 = \frac{\Delta\bar{x}}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\Delta\bar{x}}{\sqrt{\frac{n_1 + n_2}{n_1 n_2} \frac{(n_1 - 1)\sigma_1^2 + (n_1 - 1)\sigma_1^2}{n_1 + n_2 - 2}}} \quad t_0 = \frac{\Delta\bar{x}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$H_1: \quad \Delta\mu_1 = \Delta \qquad t_1 = \frac{\Delta\bar{x} - \Delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\Delta\bar{x} - \Delta}{\sqrt{\frac{n_1 + n_2}{n_1 n_2}\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}}} \quad t_1 = \frac{\Delta\bar{x} - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

dove nell'ipotesi fondamentale si è posto: $\Delta = 0$, eseguendo poi un test sequenziale parametrico per campioni indipendenti e normali, facendo uso della funzione densità di probabilità t di Student:

$$p(t) = f_0\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \qquad \text{con:} \qquad \nu = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\left(\frac{\sigma_1^2}{n_1}\right)^2 \frac{1}{n_1 + 1} + \left(\frac{\sigma_2^2}{n_2}\right)^2 \frac{1}{n_2 + 1}} - 2 \qquad \text{per il test di Welch}$$

Come già detto in precedenza, non occorre aver accumulato un numero significativo di osservazioni, perché il test può essere eseguito già avendo due dati per campione. Infatti ad ogni passo n, si calcolano le medie dei due campioni e si confronta la loro differenza con il valore Δ scelto e fissato.

Nel test sul confronto di varianze, si ha invece:

$$H_0: \quad \sigma_0^2 = \sigma^2 \qquad \chi_{0,1}^2 = \frac{\nu_1 \hat{\sigma}_1^2}{\sigma^2} \qquad \chi_{0,2}^2 = \frac{\nu_2 \hat{\sigma}_2^2}{\sigma^2} \qquad F_0 = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$$

$$H_1: \quad \sigma_1^2 = k\sigma^2 \qquad \chi_{1,1}^2 = \frac{\nu_1 \hat{\sigma}_1^2}{\sigma^2} \qquad \chi_{1,2}^2 = \frac{\nu_2 \hat{\sigma}_2^2}{k\sigma^2} \qquad F_1 = \frac{k\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$$

dove nell'ipotesi fondamentale si è posto: $k = 1$, eseguendo poi un test sequenziale parametrico per campioni indipendenti e normali, ricordando che le funzioni densità di probabilità delle variabili casuali χ^2 e F di Fisher sono rispettivamente:

$$p(\chi^2) = f_0 (\chi^2)^{\frac{\nu}{2}-1} e^{-\frac{\chi^2}{2}} \qquad p(F) = f_0\left(\nu_2 F^{\frac{\nu_1}{2}-1} + \nu_1 F^{-\left(\frac{\nu_2}{2}+1\right)}\right)$$

Nel test sul confronto del coefficiente di correlazione, si ha:

$$H_0: \quad \rho_0 \qquad z_0 = \frac{Z - Z_0}{\sigma_Z} \qquad \text{e} \qquad H_1: \quad \rho_1 \qquad z_1 = \frac{Z - Z_1}{\sigma_Z}$$

essendo la trasformata Z di Fisher (per il coefficiente di correlazione):

$$Z = \frac{1}{2} ln \frac{1 + \hat{r}}{1 - \hat{r}} \qquad Z_0 = \frac{1}{2} ln \frac{1 + \rho_0}{1 - \rho_0} \qquad Z_1 = \frac{1}{2} ln \frac{1 + \rho_1}{1 - \rho_1} \qquad \sigma_Z^2 = \frac{1}{n - 3}$$

dove nell'ipotesi fondamentale si pone spesso: $\rho_0 = 0$, eseguendo poi un test sequenziale parametrico per campioni normali, ricordando che anche la funzione densità di probabilità della variabile casuale z è normale.

Il comportamento del rapporto di verosimiglianza λ, al crescere del numero dei dati n, è del tutto libero (infatti l'esempio illustrato in figura 4.2 mostra la conferma dell'ipotesi alternativa, contro quella fondamentale precedentemente ipotizzata), così come un test tradizionale è aperto a qualsiasi risultato. Inoltre anche tutte le ipotesi assunte sono proprio le stesse adottate per i corrispondenti test tradizionali.

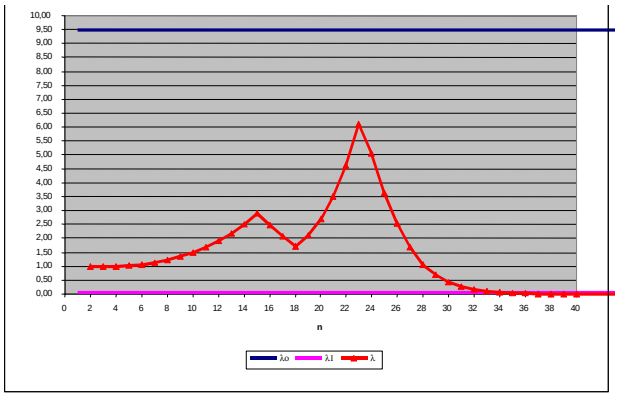


Fig. 4.2 – Un altro esempio di test sequenziale

<u>Test non – parametrici</u>

I test di rango di Mann – Whitney (per i valori centrali) e Siegel – Tuckey (per la dispersione) sono utilizzati rispettivamente per il confronto dei valori centrali o della dispersione di due campioni indipendenti, non necessariamente normali. Essi operano confrontando la somma dei ranghi di uno qualsiasi dei due campioni con la media teorica dei ranghi, standardizzata grazie alla varianza teorica dei ranghi stessi.

- Se i due campioni hanno valori centrali comparabili, il valore stimato $\hat{R}$ è vicino alla media teorica dei ranghi; in caso contrario, differisce per difetto od eccesso (e parimenti, in senso opposto, la somma dei ranghi dell'altro campione).
- Allo stesso modo, se hanno dispersione comparabile, il valore stimato $\hat{R}$ (a partire dalle differenze tra i valori argomentali di ogni campione con la rispettiva mediana) è prossimo alla media teorica dei ranghi.

Partendo da queste considerazioni e tenendo conto del limite centrale della statistica (cui si può fare riferimento trattandosi di somme di valori – i ranghi – indipendenti ed equiponderati, benché di distribuzione incognita), si arriva a costruire l'espressione che permette di eseguire i test sequenziali con le due ipotesi alternative, modificando il denominatore della media teorica dei ranghi.

$$z_0 = \frac{\hat{R} - \overline{R}_0}{\sigma_R} \qquad z_1 = \frac{\hat{R} - \overline{R}}{\sigma_R} \qquad \text{essendo:} \qquad R = \frac{n_1(n_1 + n_2 + 1)}{A} \qquad \text{con } A = 2 \text{ in } H_0$$

Infatti nell'ipotesi alternativa, la media teorica è costruita come se si sapesse a priori che i ranghi del campione prescelto siano sistematicamente più piccoli o più grandi di quelli dell'altro.

I test di segno di Thompson (per i valori centrali e per la dispersione) sono utilizzati rispettivamente per il confronto dei valori centrali o della dispersione di due campioni qualsiasi. Con ragionamenti analoghi a quelli esposti per i test di rango, si arriva a costruire anche l'espressione che permette di eseguire i test sequenziali con le due ipotesi alternative:

$$z_0 = \frac{\hat{f}^{(+)} - 0.5}{\frac{0.5}{\sqrt{\hat{f}^{(+)} + \hat{f}^{(-)}}}} \qquad z_1 = \frac{\hat{f}^{(+)} - A}{\frac{0.5}{\sqrt{\hat{f}^{(+)} + \hat{f}^{(-)}}}}$$

Infatti nell'ipotesi alternativa, la media teorica è ancora costruita come se si sapesse a priori che i ranghi del campione prescelto siano sistematicamente più piccoli o più grandi di quelli dell'altro.

Il test sul coefficiente di correlazione sui ranghi di Spearman si sviluppa invece proprio come il test classico sul coefficiente di correlazione, calcolato sui valori argomentali, in quanto identica è la distribuzione normale di comportamento della variabile casuale z, ottenuta dalla trasformata Z di Fisher dei suddetti coefficienti:

$$\hat{r}_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} \Delta_i^2 \qquad Z = \frac{1}{2} ln \frac{1+\hat{r}_S}{1\text{-}\hat{r}_S} \qquad z_0 = \frac{Z - Z_0}{\sigma_Z} \qquad z_1 = \frac{Z - Z_1}{\sigma_Z}$$

dove tutti gli addendi Δ_i sono ottenuti come differenze puntuali fra i ranghi, assegnati separatamente alle due componenti del campione dato, ed i termini (Z_0 Z_1 σ_Z^2) gli stessi già definiti in precedenza.

<u>Verso i test multipli</u>

I test di buon adattamento sono eseguiti per verificare, se i dati di un campione possono essere interpretati come estratti o meno da una popolazione di distribuzione nota; i test di indipendenza per verificare, se le componenti di un campione sono indipendenti fra loro o no.

Innanzitutto si presentano i test per una frequenza (p_0 o p_1), due frequenze ($\Delta p_0 = 0$ e $\Delta p_1 = \Delta$) ed una contingenza ($c_0 = f - pq$ e $c_1 = f - kpq$), precisando che le frequenze possono essere rilevate in un dominio di qualsiasi dimensione e la contingenza può essere calcolata utilizzando una frequenza doppia (come nella relazione presentata), oppure multipla:

$$z = \frac{\hat{f} - p}{\sqrt{\frac{p(1-p)}{n}}} \qquad z = \frac{\Delta\hat{f} - \Delta p}{\sqrt{\frac{f_1(1-f_1)}{n_1} + \frac{f_2(1-f_2)}{n_2}}} \qquad z = \frac{\hat{c}}{\sqrt{\frac{pq(1-pq)}{n}}}$$

Dopodiché i test di buon adattamento ed indipendenza consistono nel confrontare le frequenze del campione rispettivamente con la funzione densità di probabilità di una popolazione nota ed il prodotto delle frequenze marginali, dove l'ipotesi alternativa è formulata, classe per classe, come esposto appena sopra:

$$\chi^2_{m-h-1} = n\sum_{i=1}^{m}\frac{\left(\hat{f}_i - p_i\right)^2}{p_i} = \sum_{i=1}^{m}\frac{\left(\hat{F}_i - np_i\right)^2}{np_i}$$

$$\chi^2_{(n-1)(m-1)} = n\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\frac{\left(\hat{f}_i - \hat{p}_i\hat{q}_j\right)^2}{p_i q_j} = \frac{1}{n}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\frac{\left(n\hat{F}_{ij} - \hat{P}_i\hat{Q}_j\right)^2}{\hat{P}_i\hat{Q}_j}$$

essendo h il numero dei parametri di disturbo ed i termini (p_k q_l) gli stessi già definiti in precedenza.

Analogamente nel test di Kolmogorov – Smirnov (di buon adattamento ed indipendenza) e nel test di Pearson et al. (per la normalità), più potenti, occorre mettere in evidenza le ipotesi alternative rispettivamente sulla distribuzione di una popolazione (P_0 o P_1) dalla quale dati campionari possono essere interpretati come estratti o meno e sull'indipendenza o no delle componenti di un dato campione ($([p][q])_0$ .o $([p][q])_1$:

$$D_n = max\left|\hat{t}_i - P_i\right| \qquad D_n = max\left|\hat{t}_{ij} - [p_i][q_j]\right| \qquad \chi^2_2 = \frac{(\hat{\gamma} - A)^2}{6/N} + \frac{(\hat{\beta} - B)^2}{24/N}$$

dove nell'ipotesi fondamentale, per il test di Pearson et al., si impone ($A = 0$ e $B = 3$), in conformità alla forma simmetrica e normocurtica della distribuzione normale.

<u>Test multipli su campioni normali</u>

I test multipli (su campioni normali) per l'analisi di varianza verificano l'ipotesi di uguaglianza delle medie parziali fra loro e con la media generale, con il test di Fisher nell'ipotesi di uguale varianza fra i campioni raccolti e con il test di Welch nell'ipotesi di diversa varianza:

$$F_{\nu_s\nu_r} = \frac{k\sigma_s^2}{\sigma_r^2} \qquad \upsilon_s = n-1 \qquad \nu_r = n(m-1) \qquad \text{(1D)}$$

$$\upsilon_s = n-1 \qquad \nu_r = (m-1)(n-1) \qquad \text{(2D)}$$

$$\upsilon_s = n-1 \qquad \nu_r = \left(\sum_{j=1}^{n}\frac{\sigma_j^2}{m_j}\right)^2 \Bigg/ \sum_{j=1}^{n}\left(\frac{\sigma_j^2}{m_j}\right)^2 \frac{1}{m_j - 1} \qquad \text{per il test di Welch}$$

Inoltre il test per lo studio delle componenti della varianza (di Bartlett) verifica l'uguaglianza fra la varianza di una certa variabile casuale e le varianze di date variabili statistiche che si suppongono campioni estratti da una popolazione costituita dalla variabile casuale stessa:

$$\chi^2_{\nu=n-1} = -2\,ln\,\Lambda \qquad \Lambda = \frac{k\prod_{j=1}^{n}(\sigma_j^2)^{m_j/2}}{\left(\sum_{j=1}^{n} m_j\sigma_j^2 \Big/ \sum_{j=1}^{n} m_j\right)^{\sum_{j=1}^{n} m_j/2}}$$

Analogamente il test per lo studio della struttura di covarianza (di Hotelling) verifica l'incorrelazione delle componenti di una certa variabile casuale multidimensionale, accertando l'annullarsi di opportuni indicatori della correlazione fra le componenti di una data variabile statistica multidimensionale che si suppone un campione estratto da una popolazione costituita dalla variabile casuale stessa:

$$\chi^2_{\nu=n(n-1)/2} = -2\,ln\,\Lambda \qquad \Lambda = k(det\,C_{xx})^{m/2} \Big/ (\prod_{j=1}^{n} \sigma^2_{x_j})^{m/2}$$

dove nell'ipotesi fondamentale si impone sempre: $k = 1$.

Test multipli non – parametrici

I test multipli non – parametrici eseguono l'analisi di varianza e lo studio delle componenti della varianza, con il test di Kruskal – Wallis per campioni indipendenti e con il test di Friedman per campioni qualsiasi; mentre lo studio della struttura di covarianza è eseguito con il test di test di Wilcoxon – Wilcox modificato secondo Lawley:

$$\chi^2_{\nu=n-1} = \frac{A}{N(N+1)} \sum_{j=l}^{n} \frac{R_j^2}{m_j} - 3(N+1)$$ test di Kruskal – Wallis

$$\chi^2_{\nu=n-1} = \frac{A}{mn(n+1)} \sum_{j=1}^{n} R_j^2 - 3m(n+1) = \frac{B \sum_{j=1}^{n} (R_j - \overline{R})^2}{\sum_{j=1}^{n} R_j}$$ test di Friedman

$$\chi^2_{\nu=n(n-1)/2} = \left(m - 1 - \frac{2n+5}{6} \right) \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(r_{ij} - C \right)^2$$ test Wilcoxon – Wilcox / Lawley

dove nell'ipotesi fondamentale si impone ($A = 12$, $B = 6$ e $C = 0$). Infine si noti, come il test di Lawley per campioni normali possa essere utilizzato in sostituzione del test di Hotelling.

Un commento conclusivo di questa lunga collezione di test sequenziali, semplici e multipli, parametrici (o della normalità) e non – parametrici, rileva come la maggiore difficoltà, di trasformare un test tradizionale in un test sequenziale, stia spesso nella modalità adottata per inserire l'informazione legata all'ipotesi alternativa. Pertanto nei casi in cui non sia possibile costruire l'ipotesi alternativa con una semplice traslazione, una soluzione ragionevole è inserire un parametro, come una sorta di peso, con lo scopo di amplificare o smorzare il quoziente dello stimatore prescelto di cui calcolare la probabilità composta. Come già detto in precedenza, proprio questa probabilità composta, volutamente alterata rispetto all'ipotesi fondamentale, messa a quoziente con la probabilità composta, calcolata secondo la suddetta ipotesi fondamentale, dà il rapporto di verosimiglianza utilizzato per concludere il test sequenziale, mediante il confronto fra il valore atteso e le rette di confine, dipendenti dal livello di significatività e dalla potenza del test.

Test per i minimi quadrati

I test per i minimi quadrati sono strumenti importanti, utili alla validazione dei dati e dei modelli. Infatti alcuni di questi test accolgono o respingono insiemi di parametri e li confrontano con valori di riferimento o con valori precedenti. Altri test sono invece rivolti alla individuazione ed eliminazione dei dati anomali, in particolare errori accidentali, grossolani e sistematici che si possono commettere in relazione a varie cause, oppure possono dipendere da svariate sorgenti. Altri errori, ancora più sofisticati, sono detti errori di modello; fra questi: difetti di linearizzazione, malcondizionamento della configurazione dei parametri, inaffidabilità dello schema di misura, cattiva conoscenza dei pesi delle osservazioni, presenza di correlazioni fra le stesse. Allora riuscire a valutare correttamente, sulla base dei dati, se e quanto un modello costruito sia accurato,

preciso, affidabile e robusto, è uno strumento indispensabile e fondamentale per condurre a buon fine la suddetta validazione.

In particolare, i test globali di autoconsistenza e di crossvalidazione sono comunque test parametrici di verifica di ipotesi sulla varianza. Infatti come per questi ultimi, l'applicazione dei test sequenziali consiste, nel caso di confronto di ipotesi con un valore di riferimento, nell'esecuzione del test con le due ipotesi fondamentale ed alternativa, mentre nel caso di test sul rapporto di varianze, si introduce nel quoziente un fattore k che permette di esprimere l'ipotesi alternativa. Modalità analoghe sono altresì adottate nei test globali, parziali e locali di significatività dai parametri.

Infine nei test per l'identificazione e l'eliminazione dei dati anomali, ricordando che il test globale è dato dal test di Pearson et al. (cfr. verso i test multipli), tanto il test di Thompson per una selezione all'indietro, cosiddetto "data snooping", quanto il test estremale di Hawkins per una selezione in avanti ($H_e = max\left(k\hat{v}_i^2 \big/ \nu\sigma_{v_i}^2\right)$), dopo l'applicazione di procedure robuste, richiedono l'introduzione di un "peso" k, per formulare correttamente anche l'ipotesi alternativa, mentre nell'ipotesi fondamentale si impone sempre: $k = 1$. Alcune considerazioni conclusive sull'utilizzo dei test sequenziali rilevano:

- La rapidità con cui si arriva a determinare una soluzione è fortemente dipendente dal parametro stimato: in generale, si arriva a soluzione in un numero minore di passi, se si verifica un parametro di centro, rispetto ad un valore di dispersione, oppure un indice di dipendenza, in accordo con la diversa consistenza delle stime.
- Detta rapidità dipende, anche da come si scelgono le ipotesi fondamentale ed alternativa: quanto più vicine fra loro sono le due ipotesi, tanto meno velocemente si arriva ad una soluzione (del resto, è ben noto come sia fortemente sconsigliato verificare in alternativa ipotesi vicine anche con i test tradizionali).
- Inoltre la scelta dell'ipotesi alternativa deve avvenire dopo l'esame del comportamento delle osservazioni, in modo da fissare un'alternativa sensata che renda credibile il risultato. In altre parole, un'ipotesi alternativa che contraddica il comportamento delle osservazioni, se queste si discostano da quanto previsto con l'ipotesi fondamentale, porterebbe comunque ad accettare l'ipotesi fondamentale, anche se poco plausibile.

Infine è bene risottolineare, come un test sequenziale giunga a soluzione con un numero di osservazioni sempre inferiore rispetto a quelle necessarie per passare un test tradizionale, come mostrato in figura 4.3.

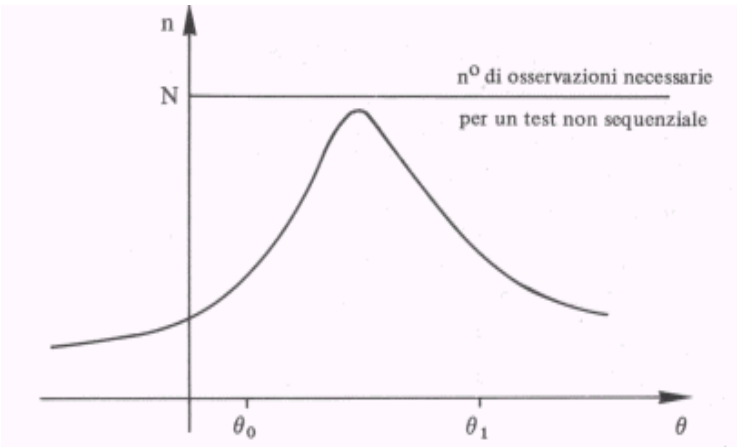


Fig. 4.3 – Andamento del numero medio di osservazioni richieste da un test sequenziale

APPENDICE C – APPLICAZIONE DELLA REGRESSIONE MULTIPLA

Inversione termica e nocività ambientale

L'ambiente, da posto a posto, per le sue condizioni fisiche e geo – ambientali, può assorbire in modo diverso l'immissione di inquinanti. Nella pianura padana, proprio per la sua locazione sottoposta a frequenti inversioni termiche, responsabili di nebbie invernali e di afa estiva, è necessario per la tutela della salute limitare al massimo la produzione di sostanze inquinanti da insediamenti industriali, motori e riscaldamento. Infatti l'aria che si respira, nelle grandi città, è sempre più sporca e ciò costituisce un problema del quale anche i mezzi di comunicazione di massa parlano da anni. Fin dal primo momento, l'attenzione è stata rivolta alle sorgenti dell'inquinamento: certi combustibili sono più inquinanti di altri, certe industrie, necessariamente inquinanti, possono rendere i loro scarichi meno nocivi con opportuni filtri e così via. Tuttavia non sempre, si è chiarito a sufficienza un altro importante aspetto del problema: la capacità dell'ambiente di assorbire l'immissione di inquinanti è diversa da posto a posto; pertanto di diverse sono le conseguenze cui la stessa quantità di inquinante può dare luogo. Questo discorso rientra nell'ambito della tematica, molto più ampia, della corretta gestione del territorio, cioé una serie di fenomeni e realtà fisiche, dove è ovviamente impossibile influire, generano vincoli, oppure comunque delle soglie di sicurezza, cui la natura chiede di sottostare. Il continuo progredire di alcune tecnologie può contribuire ad elevare tali soglie di sicurezza, senza tuttavia sopprimerle, perché intrinseche di un mondo di dimensioni finite. Non sottostare a tale limitazione significherebbe creare un vero e proprio inquinamento i cui effetti cumulati, a lungo, potrebbero essere di una gravita imprevedibile.

Un buon campanello d'allarme dell'alternarsi dell'equilibrio ambientale, a causa di un eccessivo inquinamento, è dato dall'aumentare dei casi di malattie alle vie respiratorie (cioè di malattie causate anche proprio dall'inquinamento atmosferico). Scopo di questo lavoro è stato di dimostrare come, con l'avvicinarsi delle soglie naturali di sicurezza è possibile accertare una relazione che leghi la variabilità dei casi di malattie respiratorie alla variabilità di concentrazione di inquinanti nell'atmosfera. L'area dell'hinterland milanese costituisce un tipico esempio di incompatibilità tra le caratteristiche geografiche ambientali e lo sregolato sviluppo di grandi insediamenti umani. Infatti a Milano l'inquinamento ha un effetto particolarmente pernicioso, perché esistono condizioni praticamente uniche in Europa, nella bassa – media Val Padana, tali da esaltare un fenomeno metereologico, l'inversione termica, che altrove ha una importanza molto minore. Proprio questa situazione ha un'influenza diretta e talvolta immediata sulla salute di molti milanesi.

E' possibile dare una veste rigorosa a queste affermazioni, sfruttando procedimenti statistici per spiegare il fenomeno sull'ipotesi di relazioni di causa – effetto, tra eventi meteorologici ed inquinamento e, tra questi, ed un certo tipo di malattie. E' chiaro che, operando così, non solo si analizza una particolare questione inerente l'inquinamento atmosferico di Milano, ma si vuole anche proporre un tipo di approccio numerico al problema del rapporto uomo – natura. E' necessario chiarire che cosa si intenda per inversione termica. E' noto che in condizioni normali, la temperatura dell'aria diminuisce all'aumentare della quota, in ragione di circa $6 \div 10°C$, per ogni chilometro di altezza. In queste condizioni, oppure in condizioni ancora più favorevoli, una particella inquinante, emessa ad una temperatura maggiore di quella ambiente, tende a salire raffreddandosi. Dopodiché l'aria attorno alla particella si raffredda maggiormente ed il moto verso l'alto di questa non si arresta, se non ad altissima quota. Allora la dispersione dei fiumi, purché emessi a temperatura abbastanza elevata e con energia sufficiente è assicurata.

In condizioni particolari, può tuttavia accadere che la temperatura in quota sia maggiore di quella al suolo, trovandosi allora in presenza di una inversione termica. In questa situazione, una particella di inquinante

emessa a temperatura maggiore di quella ambiente in un primo momento tende a salire, ma salendo si raffredda mentre, per contro, la temperatura dell'aria circostante aumenta. Nel giro di pochi metri, la particella si trova ad essere più fredda dell'aria che la circonda, si ferma e comincia a riabbassarsi, cosicché l'inquinamento è trattenuto nelle immediate adiacenze del luogo di emissione, formando una specie di bolla d'aria relativamente calda e sporca che solo una precipitazione o un forte vento potranno sciogliere.

Una inversione termica (Fig. C.1) può essere di due tipi:

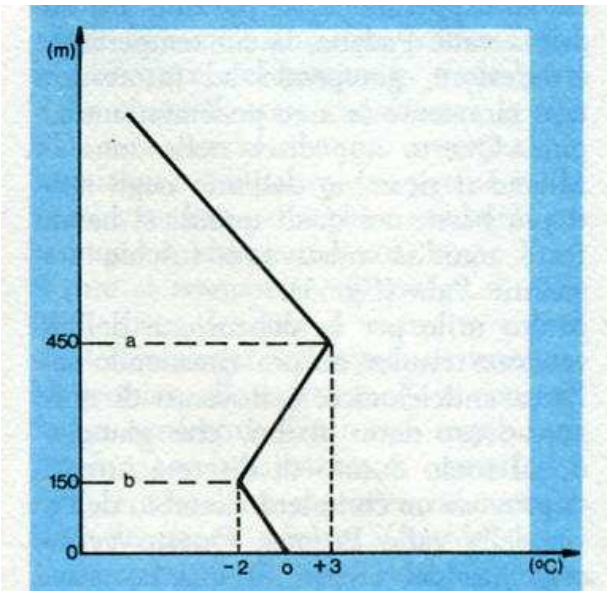


Fig. C.1 – Profilo termico verticale dell'atmosfera in condizioni di inversione termica (in ascissa è riportata la temperatura dell'aria in gradi centigradi ed in ordinata la quota in metri: la riga a costituisce la sommità dello strato di inversione e quella b base dello strato di inversione)

- Da irraggiamento: prodotta nelle notti serene con poco vento, è dovuta alla trasmissione di calore dal suolo, originariamente più caldo, verso il ciclo notevolmente più freddo. Allora il suolo e gli strati d'aria a contatto si raffreddano più rapidamente di quanto non avvenga per gli strati d'aria più in alto. Generalmente le inversioni di questo tipo si sciolgono poche ore dopo l'alba, allorquando il sole fa alzare nuovamente la temperatura del suolo. Le inversioni termiche da irraggiamento possono tuttavia avere una certa importanza, in quanto si formano proprio nelle ore in cui gli impianti di riscaldamento sono avviati dopo la sosta notturna, emettendo in brevissimo tempo una notevole quantità di inquinanti. Si tenga inoltre presente che d'inverno lo spostamento casa – posto di lavoro (o scuola) è effettuato da moltissime persone proprio in queste ore.
- Da subsidenza: dovuta alla presena dell'anticiclone (regime di alte pressioni, cielo sereno o quasi). In tali condizioni gli strati superiori dell'aria tendono a muoversi secondo una spirale discendente riscaldandosi, per compressione, e portandosi ad una temperatura maggiore del suolo e degli strati d'aria a contatto con esso. Questo tipo di inversione resiste anche nelle ore più calde del giorno e può perdurare per parecchio tempo (giorni ed anche intere settimane). Le inversioni termiche da subsidenza sono particolarmente gravose a Milano e nel suo hinterland. Infatti la valle Padana è un'ampia valle ad U, delimitata a nord dalla catena delle Alpi (quota media circa $3000\,m$) ed a Sud dalla catena degli Appennini settentrionali (quota media $1000 \div 1500\,m$). Queste due catene montuose fanno sì che l'intera Valle Padana (quote comprese tra $50 \div 200\,m$) possa considerarsi come una grande depressione.

Una delle prime conseguenze meteorologiche di una tale situazione geografica è costituita dal particolare andamento del profilo verticale del vento. Infatti prescindendo dai rari venti da est e sud – est tutti gli altri venti, prima di interessare la Valle Padana, devono superare catene montuose la cui altitudine media è notevolmente superiore a quella della valle stessa; ne consegue che raramente un vento forte scende a toccare la pianura, mantenendosi generalmente a quote superiori. Dal suolo fino ad una altezza di oltre $1000\,m$, si osservano velocità medie del vento estremamente basse (raramente superiori a $10\,km/h$ ed incrementi e decrementi ancor più contenuti (spesso non superiori a $0.05 \div 0.25\,m/s$ per diverse centinaia di metri in altezza) ; sempre limitatissime sono pure le punte massime, rappresentate dalle raffiche di vento (raramente superiori a $50\,km/h$). E' chiaro che, una volta formata una inversione termica, è ben difficile che questa si sciolga in poche ore. La formazione dell'inversione termica da subsidenza su Milano può essere schematizzata nei modi seguenti.

- In inverno: nell'area anticiclonica i venti freddi, di debole spessore ed intensità, e di notevole frequenza, provenienti da nord, attraversano le Alpi quasi a raso, si riscaldano a contatto con il versante meridionale delle montagne, molto esposto al sole, e discendono riscaldandosi ulteriormente, senza raggiungere al suolo il centro della valle padana la cui temperatura media rimane notevolmente inferiore. Questo fenomeno provoca al suolo nell'area di Milano temperature basse ed il permanere a lungo di banchi di nebbia molto fitti ed estesi (Fig. C.2).
- In estate: nell'area anticiclonica i venti caldi, spesso di notevole spessore ed intensità, ma non di elevata frequenza, provenienti da sud, attraversano gli Appennini e si portano alti sopra la valle Padana la cui temperatura è inferiore, giungendo ad interessare non raramente le aree pedemontane alpine. Questo fenomeno impedisce nella zona di Milano il ricambio dell'aria negli strati più bassi nei quali si hanno forte umidità relativa e foschie stagnanti, cioè afa (Fig. C.3).

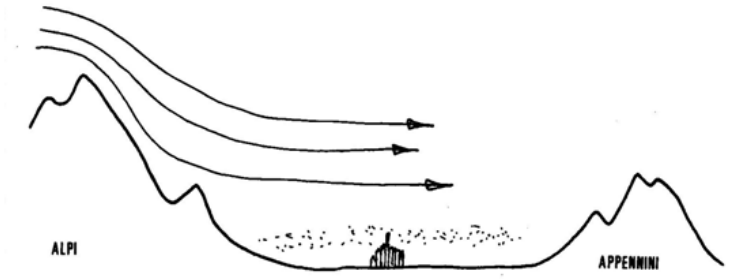


Fig. C.2 – Andamento delle correnti d'aria che determinano inversione termica nel periodo invernale

Molto utile per la distruzione dell'inversione termica, ancora rimanendo nell'area anticiclonica, è il vento di nord nord – ovest, detto foehn, che giungendo al suolo dotato di discreta intensità provoca un completo ricambio dell'aria della Valle Padana. Questo vento, originato dal crearsi di una eccessiva differenza di pressione tra la valle Padana ed il versante nord delle Alpi, provoca un traboccamento dell'aria da nord a sud lungo i pendii delle Alpi. Purtroppo questo fenomeno, abbastanza frequente nell'area pedemontana alpina, interessa la città di Milano e, in generale, il centro della valle Padana solo sporadicamente.

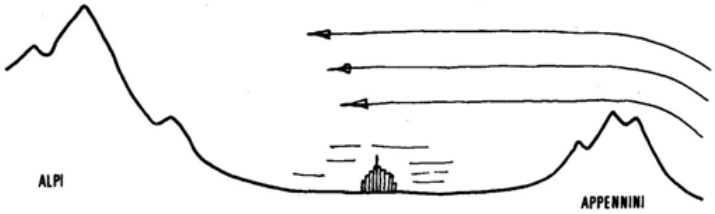


Fig. C.3 – Andamento delle correnti d'aria che determinano inversione termica nel periodo estivo

La regressione lineare multipla

L'approccio numerico cui si è già accennato consiste nel misurare due o più fenomeni che evolvono nel tempo, per ricercare se esistono analogie, tra gli andamenti dei fenomeni stessi, e per attribuire a tali analogie, secondo un criterio logico – statistico, un modello matematico che dica se sia lecito ipotizzare un rapporto di causa – effetto. Dato un certo numero di variabili, il procedimento adottato è in grado di precisare in che misura ciascuna variabile può essere considerata come dipendente da una combinazione lineare delle altre; ossia più semplicemente, per quanti giorni all'interno dell'intervallo di tempo considerato l'andamento di una variabile può essere ritenuto direttamente o inversamente proporzionale all'andamento delle altre. Questo procedimento è comunemente detto regressione lineare multipla.

La regressione lineare multipla si presta all'esame di un fenomeno fisico complesso, quale la permanenza degli inquinanti nell'atmosfera con le sue cause e conseguenze, perché tra un gran numero di variabili è in grado di scegliere quali siano le più idonee ad interpretare il fenomeno stesso. Tutto ciò consente di proporre una quantità di variabili, anche apparentemente illogiche o contraddittorie: il procedimento di calcolo provvede poi opportunamente ad accettare o scartare le variabili stesse ed il modello finale sarà costituito dalla combinazione lineare delle sole variabili significative.

Affinché si possa procedere, tutti i fenomeni considerati devono essere trasformati in numeri. Questo passaggio, immediato per le grandezze misurabili direttamente, quali temperatura, la velocità del vento, umidità relativa o precipitazioni, diventa invece alquanto meno semplice per fenomeni compositi quali inversione termica, inquinamento atmosferico, o peggio ancora, aggravarsi della situazione sanitaria di una intera città. Pertanto pur trattandosi di fatti precisi, la loro trasformazione in cifre comporta l'adozione di schemi semplificativi che vale la pena di analizzare più a fondo.

Le variabili usate

Innanzitutto è stato necessario fissare un periodo del quale fosse agevole, ma soprattutto possibile, reperire tutti i dati necessari. Sono stati scelti i 151 giorni compresi fa il 1/11/72 ed il 31/3/73, dal momento che il fenomeno, come si vedrà in seguito, ha le sue conseguenze più gravi nei mesi invernali.

Per quanto riguarda le variabili metereologiche, desunte dalle misure dell'Osservatorio Metereologico. di Brera, non si è dovuto procedere ad alcuna schematizzazione particolare. Tuttavia si sono create delle variabili un po' particolari che rappresentassero certe condizioni climatiche di particolare interesse. Si è così pensato che sulla salute dei cittadini può influire non solo la temperatura media del giorno o la minima, ma anche l'escursione termica che si riscontra nelle 24 ore; inoltre è pensabile che un improvviso colpo di freddo in una giornata abbastanza mite è più pericoloso dell'ulteriore abbassamento di una temperatura già rigida. E' nata così una variabile costituita dal prodotto dell'escursione termica giornaliera per la temperatura media.

L'andamento di questa variabile dovrebbe rispecchiare i due fenomeni suddetti. Con criteri analoghi si è ritenuto che sulla concentrazione di inquinanti nell'atmosfera può influire non solo la velocità del vento, ma anche la sua variabilità (colpi di vento). E' nata così un'altra variabile che rappresenta l'effetto di rimescolamento dell'atmosfera dovuto proprio alla variabilità della velocità del vento costituita da una funzione del rapporto tra la velocità massima e media del vento.
Il gruppo più numeroso di variabili è quello sulla inversione termica vera e propria. Dire quanto vale una inversione termica non è impresa facile; di conseguenza, si sono fornite ben dodici espressioni legate all'inversione, lasciando al metodo il compito di scegliere la più indicativa. Il dato di partenza per formare le variabili in questione è sostanzialmente il gradiente verticale, rilevato due volte al giorno, dall'Aeronautica Militare all'aeroporto di Linate. Dal gradiente termico, si ricavano immediatamente lo spessore dello strato di aria calda che genera l'inversione, la distanza dal suolo del suo limite inferiore, la differenza di temperatura fra base e sommità dello strato stesso. Queste tre grandezze sono state combinate in diversi modi, tenendo conto delle seguenti ipotesi:

- un'inversione è tanto più pericolosa, quanto più la sua base è prossima al suolo;
- un'inversione è tanto più pericolosa, quanto più lo strato d'aria calda che la genera è spesso;
- un'inversione è tanto più pericolosa, quanto maggiore è la differenza di temperatura tra la base e la sommità dello strato stesso.

Il calcolo ha dimostrato che le variabili esprimenti il verificarsi contemporaneo delle prime due ipotesi meglio si prestano a qualificare l'inversione termica.
Abbastanza semplice è stato invece tradurre in cifre l'inquinamento atmosferico. Infatti il Comune di Milano fornisce le concentrazioni di anidride solforosa (SO_2) nell'atmosfera rilevate nelle varie ore del giorno, in varie parti della città. E' stata considerata la sola SO_2; non perché sia una sostanza più inquinante delle altre, ma perché è stata rilevata con particolare assiduità; comunque è chiaro che così come l'ambiente si comporta nei confronti della SO_2, disperdendola o meno a seconda dell'intensità dell'eventuale inversione termica, allo stesso modo si comporta nei confronti di tutti gli altri inquinanti.
Producono inquinamento atmosferico tutti i veicoli a motore endotermico (monossido di carbonio: CO e, in minor misura, SO_2 ed idrocarburi incombusti), il riscaldamento domestico (SO_2 nel caso di nafta, gasolio e cherosene, e CO nel caso, più raro, di carbone), le industrie (vario a seconda della produzione: SO_2, CO, acido solfidrico: H_2S, anidride solforica: SO_3, mercaptani, diossine [65]). Inoltre si produce indirettamente inquinamento atmosferico con immissione di calore che genera ossidi di azoto (NO_X) e con la dispersione di energia elettrica che genera ozono (O_3). Va notato che la SO_2 è prodotta prevalentemente con il riscaldamento domestico e pertanto la sua concentrazione nell'atmosfera, alta nei mesi freddi diventa insignificante negli altri mesi dell'anno. Oltre alla concentrazione media e alla massima, riscontrate giorno per giorno, si sono introdotte nel calcolo anche variabili ricavate da quelle originarie, come la radice quadrata della concentrazione media, oppure il suo logaritmo, per meglio adattare il modello matematico alla realtà fisica.
A questo punto, occorre tradurre in cifre l'influenza dell'inquinamento sulla salute dei cittadini milanesi. Si sarebbe potuto svolgere una ricerca molto dettagliata; considerare il numero dei degenti per malattie polmonari in

[65] Le diossine, allora sconosciute, sono diventate rilevanti dopo l'incidente alla fabbrica ICMESA di Seveso del maggio 1976: un crimine di pace! Dopodiché esse sono state quasi sempre ritrovate in tutti le combustioni di materiali plastici, assieme all'acido cloridrico (HCl), responsabile principale delle piogge acide.

Milano e nel suo hinterland, studiarne l'evoluzione nel tempo, arrivando a concludere che è perniciosa non tanto la singola punta giornaliera di inquinamento, quanto piuttosto il permanere per lungo tempo di un inquinamento medio notevole. Tuttavia questa strada è stata scartata per l'oggettiva difficoltà di reperimento dei dati clinici necessari. Infatti alcuni degli ospedali interpellati si sono dimostrati molto restii a collaborare, mentre altri a causa dei complicati procedimenti burocratici che accompagnano sempre l'ingresso di un malato, si sono dichiarati nella impossibilità materiale di fornire, giorno per giorno, il numero dei ricoveri suddivisi per malattie.

Pertanto si è ritenuto che, per rappresentare l'effetto delle punte giornaliere d'inquinamento sulla salute pubblica sarebbe stato sufficiente introdurre una variabile campione costituita dal numero di ricoveri per malattie alle vie respiratorie verificatesi nell'Ospedale Fatebenefratelli di Milano. Infatti tale ospedale, , disponendo di un efficiente servizio elettrocontabile, ha potuto fornire, per ogni ricovero compreso fra il 1/2/72 ed il 31/3/73, la diagnosi di uscita (cioè quella più attendibile) ed i giorni di degenza (solitamente compresi tra un giorno e una settimana), oltre ad altri dati di importanza secondaria. In tal modo, è stato possibile ricostruire anche il numero di degenti nei reparti di medicina per ciascun giorno di quel periodo.

I risultati

Il calcolo ha dato risultati interessanti. Anzitutto occorre rilevare che, per annullare l'effetto di disturbo dovuto alla presenza predominante di variabili logicamente poco significative, si è proceduto anche alla formazione di modelli parziali nei quali sono state prese in considerazione solo le variabili di un certo gruppo, per poter osservare quale tipo di variabili logicamente più significative esse mascherano. Tracciare la cronistoria dei vari risultati ottenuti sarebbe un lavoro complesso e, forse inutile; pertanto sarà opportuno riassumere i modelli più interessanti usciti dal calcolo. Un primo modello globale considera come dipendente la variabile ricoveri; tale variabile, nel 28% dei giorni considerati, con una probabilità del 95% , risulta spiegata da una combinazione lineare di altre variabili tra le quali compare il tempo (numero di giorni trascorsi dall'inizio delle osservazioni).

Per spiegare nel modello la presenza del tempo (che era stato introdotto solo per controllo), si è costruito a parte un secondo modello che avesse come variabile dipendente il tempo. Questo modello parziale spiega l'80% delle osservazioni, con una probabilità del 95% , dimostrando che la variabile tempo maschera in realtà altre variabili più o meno legate all'andamento delle stagioni e, tra esse, oltre alla escursione termica che in primavera aumenta ed alla quantità di SO_2 nell'atmosfera che in primavera diminuisce (perché diminuisce il riscaldamento domestico), assume un particolare significato la variabile degenti che raggiunge valori più alti in primavera, influenzando negativamente la variabile ricoveri. Questo risultato numerico indica che l'insufficiente disponibilità di posti letto porta alla saturazione dei medesimi e ad una compressione forzosa dei ricoveri, dando ulteriore conferma ad un fatto ampiamente, noto a chi lavora negli ospedali.

Un terzo modello è stato costruito per verificare isolatamente le cause fisiche che favoriscono la concentrazione di SO_2 nell'atmosfera. In questo modello parziale, la variabile SO_2, posta come variabile dipendente, risulta nel 67% delle osservazioni, con probabilità del 95% , direttamente proporzionale all'inversione termica e inversamente alla temperatura e alla velocità del vento. Questo risultato numerico indica che la concentrazione della SO_2, come degli altri inquinanti, nell'atmosfera, dipende quasi esclusivamente da una serie di fenomeni e realtà fisiche su cui è ovviamente impossibile intervenire. Pertanto si può ottenere, una diminuzione di inquinanti nell'atmosfera, solo con un drastico contenimento delle immissioni.

Infine l'ultimo modello globale considera come variabile dipendente ancora i ricoveri, tuttavia lasciando fuori dal calcolo la variabile tempo. In questo modello il 31% dei ricoveri è spiegato, con una probabilità del 95% da una combinazione lineare delle variabili: inversione termica, SO_2, precipitazioni e degenti. Questo un risultato può essere considerato conclusivo.

Lo schema grafico di Fig. C.4 comprende con un'unica denominazione gruppi di variabili significanti un unico concetto e fornisce un quadro sintetico dei risultati, così come sono stati ottenuti dalle varie regressioni lineari multiple.

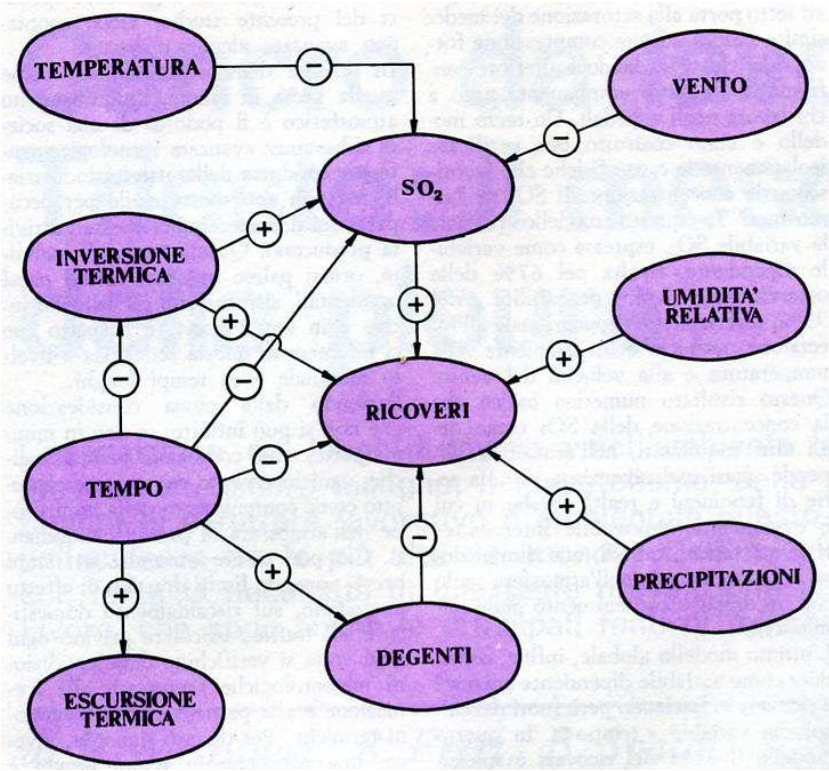


Fig. C.4 – Schema grafico della sintesi dei risultati. Il segno più o meno significa che fra le grandezze in esame esiste, con probabilità 95%, una relazione di proporzionalità diretta o inversa; le frecce significano solo il verso logico della relazione di causaeffetto supposta fra i fenomeni in esame, con la variabile tempo si vuole indicare il numero di giorni trascorsi dell'inizio delle osservazioni

Sintesi dei risultati

Volendo sintetizzare al massimo i risultati esposti nello schema grafico, si può evidenziare i dati più interessanti:

- limitatamente alle malattie considerate, oltre un terzo dei ricoveri sono dovuti a crisi provocate direttamente da l'inquinamento atmosferico;
- l'inquinamento atmosferico è aggravato dalle condizioni geoambientali *e* atmosferiche particolarissime di Milano e del suo hinterland;
- la disponibilità di posti letto negli ospedali è insufficiente;
- l'attendibilità di tutte queste informazioni è sempre superiore al 95% .

Appare chiaro come la situazione sanitaria di Milano sia fortemente influenzata dal problema meteorologico e, nonostante che ciò esuli dagli scopi specifici del presente lavoro, diventa spontaneo avanzare alcune proposte.

In un'area densamente abitata, come quella presa in esame, l'inquihamento atmosferico è il prodotto di una società abbastanza avanzata tecnologicamente per condurre delle attività industriali, ma non abbastanza civile per occuparsi dei danni ecologici che tali attività producono. Questa forma di inciviltà, ormai palese in quasi tutti i paesi occidentali, si manifesta sia a livello locale ed in tempi brevi (ed è l'aspetto che maggiormente interessa in questa sede), sia a livello mondiale ed in tempi lunghi.

Partendo dalla ovvia considerazione che non si può influire, se non in minima parte, sulle condizioni meteorologiche, ogni intervento deve essere concepito come contenimento della immissione nell'atmosfera di prodotti inquinanti. Tutto ciò può essere ottenuto, in tempi brevi, ponendo limiti drastici, di effetto immediato, sul riscaldamento domestico e sul traffico veicolare privato ogni qua! volta si verifichino delle condizioni meteorologiche favorevoli alla formazione e alla permanenza di inversioni termiche.

Per quanto riguarda invece, una soluzione in tempi lunghi è evidente l'importanza di un'oculata programmazione territoriale che indirizzi lo sviluppo di ciascun comprensorio verso quella destinazione che appare la più consona alle particolari situazioni geombientali del territorio stesso. Sarebbe opportuno pianificare il territorio milanese:

- mantenendo e realizzando basse densità di insediamento (così da evitare la costruzione di grandi centrali per il riscaldamento domestico)
- organizzando il riscaldamento domestico a gas metano (piuttosto che a nafta, gasolio e carbone);
- salvaguardando e creando ampie zone a verde (che funzionani da veri e propri polmoni dell'ambiente);
- non permettendo l'insediamento sregolato di industrie pesanti (come raffinerie, acciaierie, cementifici, industrie chimiche, ecc.);
- garantendo efficienti trasporti pubblici, per non affidare la mobilità, regionale e locale, al solo traffico privato (Fig. C.5) [66].

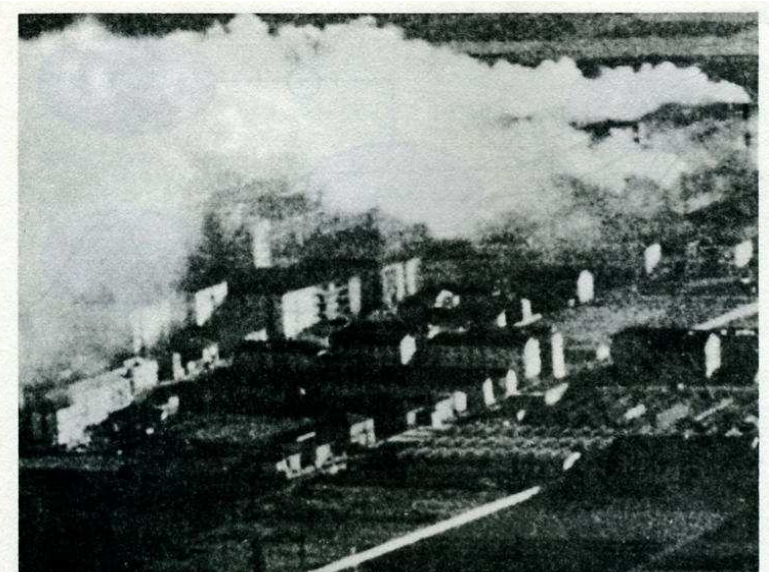


Fig. C.5 – Un 'immagine tipica dell'inquinamento da fumi negli anni '50, '60 e '70

[66] La presente appendice è ricavata dall'articolo: Bernardi S., Mandelli F., Mussio L. (1976): Inversione termica e nocività ambientale. Sapere, n. 793 (Ed. Dedalo) e ha basato i suoi fondamenti scientifici su quanto esposto in:

Bennett K.A., Franklin N.L. (1963): Statistical Analysis in Chemistry and the Chemical Industry. J. Wiley & Sons, New York.
Mc Intosh D.H., Thom A.S. (1969): Essentials of Meteorology, Wykeham Publ. Ltd., Londra.

APPENDICE D – APPLICAZIONI DELLA CLUSTER ANALYSIS

Classificazione mediante una tabella di prossimità

L'evoluzione è importante per le discipline del rilevamento, poiché anch'esse sono parte integrante del pensiero umano. Di conseguenza, esse devono prendere in considerazione gli enormi cambiamenti dovuti alle tecnologie dell'informazione, avvicinandosi ad altre discipline differenti. Un punto di partenza può essere costituito dall'individuazione dei possibili rapporti tra gruppi di lavoro di differenti associazioni scientifiche internazionali, aventi termini di riferimento simili tra loro. Pertanto con un certo grado d'originalità, una tabella di prossimità può studiare queste vicinanze, basate sulla fusione di teorie e tecniche, nonché su un accordo sistematico tra la volontà di modellare il mondo ed il desiderio di capirlo.

Il metodo sperimentale definisce gli insiemi complessi di possibili dialoghi con la natura e la storia. Infatti natura e storia possono contraddire le ipotesi fatte, nel contesto di una lingua specifica, ed anche le lingue sono costantemente in evoluzione, in rapporto ai bisogni di una società e delle sue culture. Attualmente le varie molteplicità chimico – fisiche, così come sociologiche ed economiche, mostrano fenomeni e processi nei qual si alternano i ruoli del caso e della necessità. Ad esempio, un collegamento, dato dai confini del tempo (oltreché dello spazio), è evidente tra la dinamica (cioè la scienza delle forze e delle traiettorie) e la biologia (cioè la scienza della complessità e del divenire dei processi viventi).

Infatti le attuali visoni del mondo fisico e/o della civiltà umana hanno abbandonato la concezione classica che tutte le descrizioni scientifiche debbano essere coerenti con la possibilità di supporre l'esistenza di un osservatore esterno, al mondo stesso (fisico ed antropico), per accogliere punti di vista particolari, capacità di accogliere insieme complessità e contraddizioni. Pertanto compito delle informazioni è solo cercare di aggiungere qualcosa alla quantità e qualità di conoscenze già disponibili, in una determinata comunità, e la stessa conoscenza scientifica non deve tanto essere immagazzinata in polverose biblioteche, quanto piuttosto vivere nei laboratori per pi uscirne, perché le informazioni sono un processo e non un archivio.

Il significato della comunicazione

Nel mondo occidentale oggigiorno, prevale l'idea che solo la forza della tecnologica avanzata permetta l'avanzamento della conoscenza, rendendo fertile il nucleo stesso della scienza occidentale solo grazie ad un'alimentazione finanziaria enorme. Al contrario oggigiorno, la scienza e le tecniche devono sapere / potere uscire dalle università e dai laboratori, per accettare e promuovere il loro carattere interdisciplinare, come il fuoco del progresso, perché lo scontro delle teorie è un'occasione e non una calamità. Una lezione salutare è riconoscere come la realtà sia incredibilmente più ricca della lingua che può esprimerla. Infatti una qualsiasi lingua, per quanto complessa, può esprimere soltanto una parte delle relazioni presenti.

Allora tutte le lingue e tutti i punti di vista sono complementari ed è inutile cercare un punto di vista unico che dovrebbe descrivere la totalità della realtà. Un tal sviluppo è molto importante anche per gli studi sulle discipline del rilevamento. Di conseguenza, esse dovrebbero prendere in considerazione i cambiamenti, enormi e molto profondi, dovuti alle scienze dell'informazione, avvicinarsi a discipline umanistiche abbastanza lontane (come linguistica, psicologia, sociologia della comunicazione, ecc.) ed approfondire notevolmente le conoscenze nell'ambito della matematica (di cui, in passato, sono state tanta parte, assieme all'astronomia), soprattutto dopo la conquista dello spazio e la nascita della geodesia spaziale.

Cluster analysis

Per compiere, un passo iniziale per un'analisi mirante ad esaminare qualcosa in comune tra varie associazioni scientifiche internazionali e/o differenti gruppi di lavoro attivi, si è provato a confrontare l'ISPRS (Società Internazionale di Fotogrammetria e Telerilevamento) e l'OEEPE (Organizzazione Europea per gli Studi di Fotogrammetria Sperimentale). Con un certo grado d'innovazione ed originalità, queste relazioni possono essere studiate da una tabella di prossimità la cui validità generale, per accertare somiglianze e dissomiglianze tra dati ed altre informazioni, è ben nota. Pertanto una tabella di prossimità, come indicato nella Fig. D.1, è stata compilata, per mostrare collegamenti esistenti tra due famiglie o gruppi.

- Le commissioni tecniche che conducono il lavoro scientifico e tecnico dell'ISPRS sono 7; ogni commissione si articola poi in vari gruppi di lavoro.
- Nell'OEEPE la struttura è un po' più complessa e prevede 9 commissioni scientifiche e 9 gruppi di lavoro tecnici, separati tra loro.

Pertanto una tabella a doppia entrata mostra 42 righe e 18 colonne. I valori degli elementi di ciascuna casella (0, 1, 2 e 3) sono indicatori della prossimità, forniti secondo una scala crescente di prossimità. Nella Fig. D.2, la stessa tabella di prossimità è riproposta, avendo scelto di mettere in evidenza solo due valori prossimità (2 e 3), in quanto più elevati. Nella Fig. D.3, solo il valore 3 è stato preso in considerazione, per evidenziare la somiglianza più alta. Le tabelle delle figure D.1 e D.2 sono poi state riordinate (si vedano, a riguardo, le Figure D.4 e D.5), affinché la cluster analysis potesse mettere in evidenza campi comuni di ricerca. Al contrario, la tabella della Fig. D.3 è troppo vuota per essere elaborata in modo significativo.
Dalle tabelle corrispondenti alle figure D.1 e D.4 (come mostrato dalla Fig. D.6), tre grandi cluster sono individuati che corrispondono alla principali tematiche:

- elaborazione di immagini digitali;
- applicazioni del GIS alla fotogrammetria;
- stazioni di lavoro fotogrammetriche e cartografiche.

Dalle tabelle corrispondenti alle figure D.2 e D.5 (come mostrato dalla Fig. D.7), sei cluster più piccoli (e più specifici) sono individuati; essi i fanno riferimenti ai seguenti soggetti:

- basi dei dati e gestioni di sistemi;
- GIS e stazioni di lavoro fotogrammetriche e cartografiche;
- GPS e laser scanning;
- GIS e DEM;
- elaborazione di immagini digitali;
- Telerilevamento e SAR.

A tale proposito è interessante rilevare, come nella Fig. D.6, si siano separati due cluster: GIS e DEM e GPS laser scanning, vicendevolmente molto vicini. Infatti l'uso del GPS e del laser scanning, per la produzione automatica di orto-immagini digitali, sta rapidamente sviluppandosi dappertutto.

ISPRS / OEEPE

			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
I	1	1			3				3												2
	2	2			2								1	1			2		1	1	6
	3	3		1					2	1		1					1	3	3		7
	4	4				1							1				2				3
	5	5							1	3								1	3		4
II	1	6	2						1	2			2			1			3		6
	2	7		1			1	2			2		1		1					2	7
	3	8			1		1							2	2					3	5
	4	9				2										1	3				3
	5	10					2				2			1	3					2	5
	6	11	1			2					1		1			2		1			6
	7	12					1				3		1	1						2	5
	8	13	3		2						1					3			2		5
III	1	14	3						1	3						3			3		5
	2	15		1				2				3					1	1	1		6
	3	16	2			3										1					3
	4	17	1			3		1								2					4
	5	18				2		1											1		3
	6	19				1						2				1	3				4
IV III	1	20					2	2					1	2	1						5
IV III	2	21					3	1					2	2	2	1					6
IV	1	22					1							3	1					3	4
	2	23		2				3			1	3									4
	3	24		1		1		1		1	1		3							2	7
	4	25				2											2	3			3
	5	26	1			2						1					1				4
	6	27										2	1				1				3
V	1	28	1		2				2		1					1					5
	2	29		1	1			1			2	1							2	1	7
	3	30		1				2			2	1							1		5
	4	31		1				2				1					1		2		5
	5	32		1				2				2					1		1		5
V	III	33	1										1			1					3
V	S P	34		1				2				2									3
VI	1	35		1				1		1							1		1		5
	2	36	1				1				2	1			2						5
	3	37		1			1			1		1		1				1		1	7
	4	38											3	2						3	3
V II	1	39				1	1										2		1		4
	2	40						1					1				1				3
	3	41				2						2					2	1	1		5
	4	42			1	2	1				2		1		1					1	7
	5	43				2	1					2					2				4
	6	44				1						1					3				3
	7	45						1					1				1				3
			10	12	7	15	12	16	6	7	12	16	15	9	8	11	18	7	15	11	

(207 WG's)

Fig. D.1

ISPRS / OEEPE

			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
I	1	1			3				3												2
	2	2																			2
	3	3															3	3			3
	4	4																			1
	5	5								3								3			2
II	1	6	2							2			2					3			4
	2	7						2			2									2	3
	3	8												2	2					3	3
	4	9				2											3				2
	5	10					2				2				3					2	4
	6	11				2										2					2
	7	12									3									2	2
	8	13	3		2											3			2		4
III	1	14	3							3						3			3		4
	2	15						2				3									2
	3	16	2			3															2
	4	17				3										2					2
	5	18				2															1
	6	19										2					3				2
IV III	1	20					2	2						2							3
IV III	2	21					3						2	2	2						4
IV	1	22												3						3	2
	2	23		2				3				3									3
	3	24											3							2	2
	4	25				2											2	3			3
	5	26				2															1
	6	27										2									1
V	1	28			2				2												2
	2	29									2								2		2
	3	30						2			2										2
	4	31						2											2		2
	5	32						2				2									2
V	III	33																			0
V	S P	34						2				2									2
VI	1	35																			0
	2	36									2				2						2
	3	37																			0
	4	38											3	2						3	3
V II	1	39															2				1
	2	40																			0
	3	41				2						2					2				3
	4	42				2					2										2
	5	43				2						2					2				3
	6	44															3				1
	7	45																			0
			4	1	4	10	3	8	3	3	7	8	4	5	4	4	9	2	7	7	93

(93 WG's)

Fig.D.2

ISPRS / OEEPE

			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
I	1	1			3				3												2
	2	2																			0
	3	3															3	3			2
	4	4																			0
	5	5								3								3			2
II	1	6																3			1
	2	7																			0
	3	8																	3		1
	4	9														3					1
	5	10													3						1
	6	11																			0
	7	12									3										1
	8	13	3												3						2
III	1	14	3							3					3			3			4
	2	15										3									1
	3	16				3															1
	4	17				3															1
	5	18																			0
	6	19																			0
IV III	1	20																			0
IV III	2	21					3														1
IV	1	22											3						3		2
	2	23						3				3									2
	3	24										3									1
	4	25															3				1
	5	26																			0
	6	27																			0
V	1	28																			0
	2	29																			0
	3	30																			0
	4	31																			0
	5	32																			0
V	III	33																			0
V	S P	34																			0
VI	1	35																			0
	2	36																			0
	3	37																			0

| | | |

Altri esempi

Le applicazioni della cluster analysis sono davvero molteplici e, a mo' d'esempio, nell'immediato prosieguo, sono mostrati tre esempi di cluster analysis applicata all'elaborazione d'immagini, in campi parecchio diversi tra loro. Infatti il primo esempio fa riferimento ad immagini astrometriche, dove mappe di stelle sono rappresentate in una carta derivata delle velocità, misurate in secondi d'arco per anno, nelle due coordinate della sfera celeste.

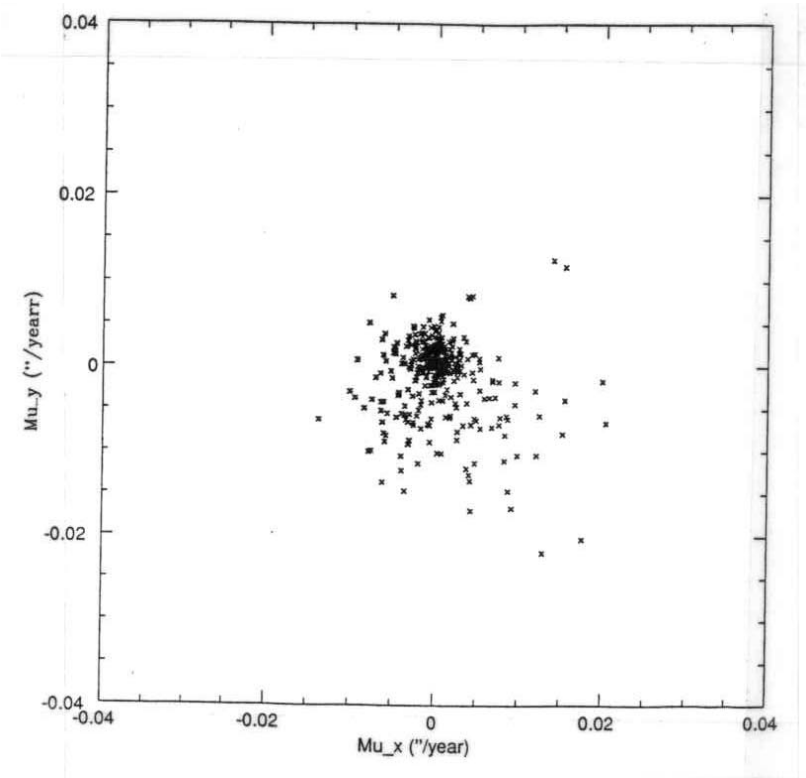


Fig. D.8 – Una nebulosa sopra un ammasso di stelle

Secondo un parere dato da alcuni astonomi, la prima clusterizzazione permette di riconoscere una nebulasa, pressoché ferma, sopra un ammasso di stelle (Fig. D.8), mentre la seconda due nebulose, una pressocché ferma ed un'altra in movimento, sempre sopra un ammasso di stelle (Fig. D.9).

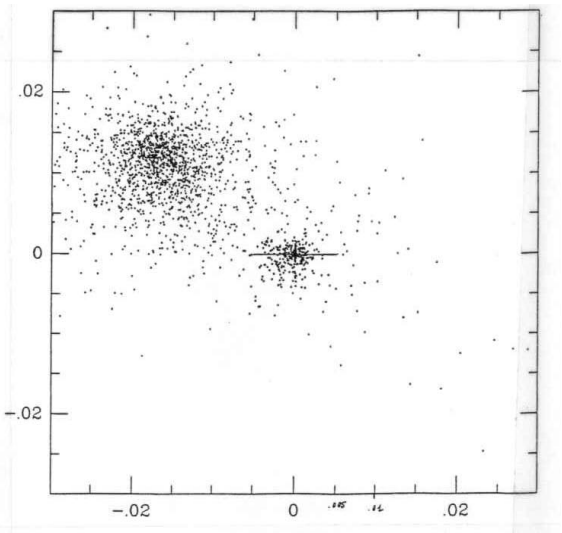


Fig. D.9 – Due nebulose sopra un ammasso di stelle

Il secondo esempio permette di completare utilmente una mosaicatura di immagini telerilevate, tramite un satellite della gamma NOAA, con l'intento di rilevare la copertura nevosa dell'arco alpino in un dato periodo. E' allora evidente, come la copertura nuvolosa, abbastanza spesso, copra porzioni, più o meno grandi, della suddetta zona, impedendo un rilevamento completo. Pertanto classificare l'innevamento rilevato, secondo altezza dei rilievi (Fig. D.10) ed esposizione dei versanti (Fig. D.11), permette di fare ipotesi probabilisticamente fondate su dove la neve avrebbe dovuto esserci (Fig. D.12), ma non è stato possibile rilevare.

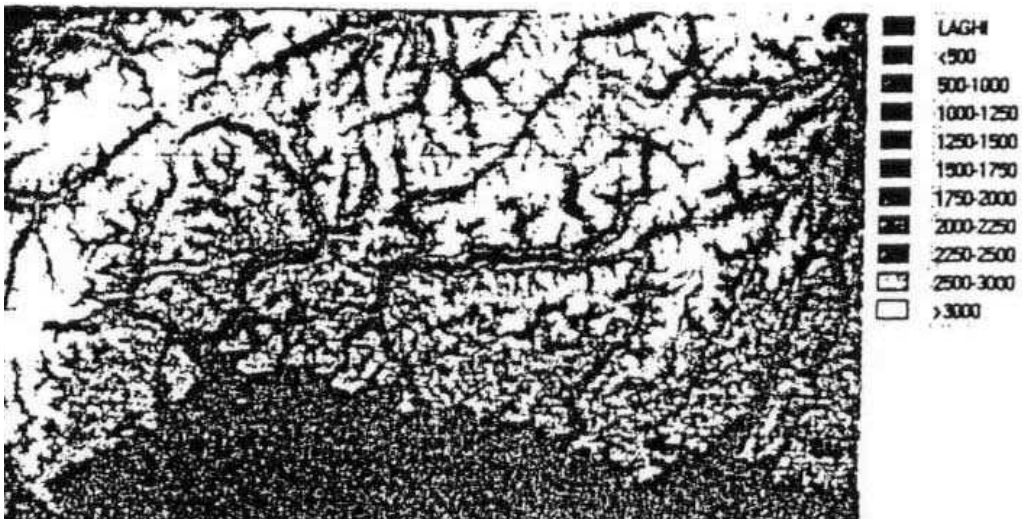


Fig. D.10 – Classificazione dell'altitudine dei rilievi dell'arco alpino

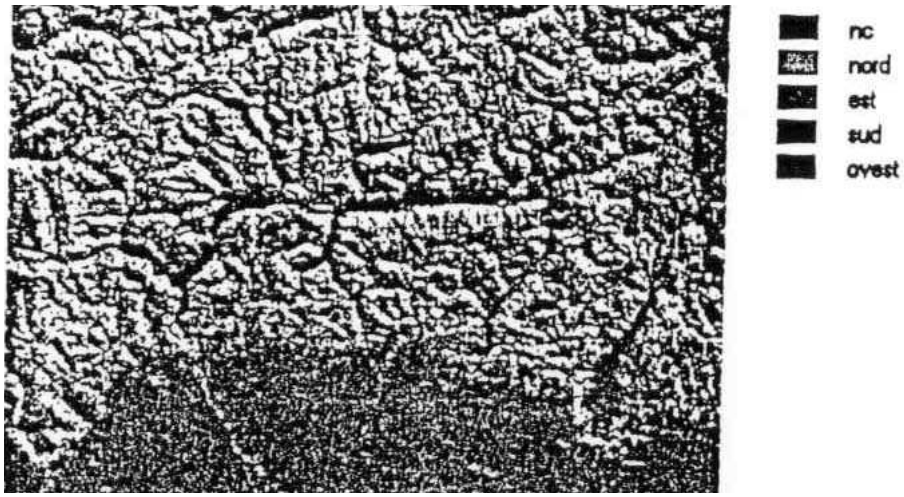


Fig. D.11 – Classificazione dell'esposizione dei versanti dell'arco alpino

A conferma della correttezza della lavoro, di classificazione, clusterizzazione ed inferenza, svolto insieme da esperti di telerilevamento e nivologia, sta la constatazione che la neve è più abbondante ad alta quota (per la minore temperatura) e nei verstanti esposti a nord (per il minore soleggiamento), mentre è più scarsa a bassa quota e nei versanti esposti a sud (rispettivamente per la maggiore temperatura ed il maggiore soleggiamento).

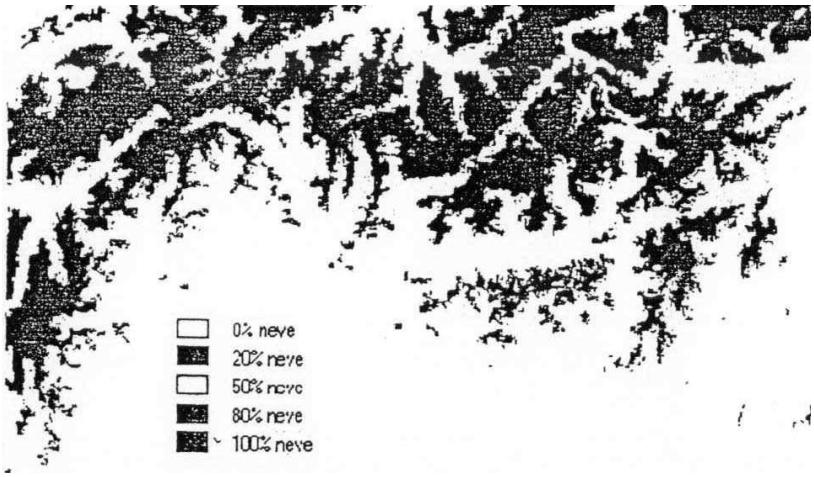


Fig. D.12 – Previsione dell'innevamento nell'arco alpino

Il terzo esempio prende in considerazione un'immagine aerea, per cartografia tecnica a media scala. Scopo ultimo del presente lavoro è l'estrazione automatica del reticolo stradale, ricavato applicando tecniche note d'elaborazione d'immagini e di fotogrammertia digitale. A tal fine, dapprima si è provveduto a costruire l'immagine gradiente, a partire dall'immagine originale, applicando poi ad essa un filtro mediano, per eliminare eventuali dati spuri. La Fig. D.13 riporta l'immagine aerea originale con superimposte le linee derivate.

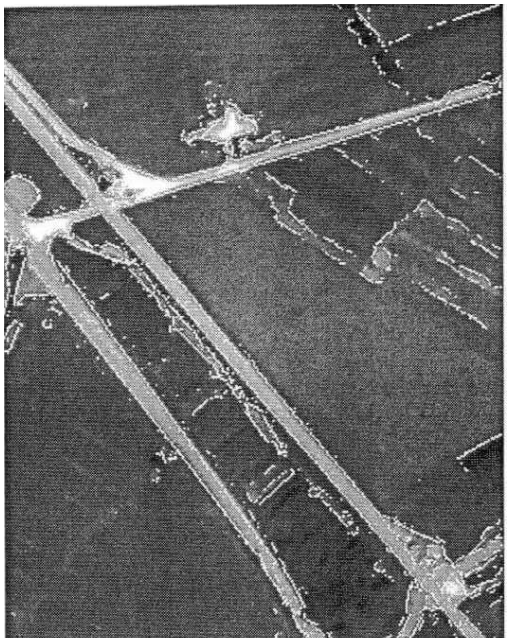


Fig. D.13 – Immagine aerea con superimposte le linee derivate da un'immagine gradiente

Successivamente l'immagine ottenuta è stata sogliata, con un filtro passa alto, eliminando le linee la cui intensità di tono di grigio è apparsa troppo bassa (come mostrato in Fig. D.14). Tuttavia poiché tale operazione, spinta agli estemi, porta a cancellare linee interessanti (lasciandone invece di inutili), si è eseguita una clusterizzazione,

pixel a pixel, con la procedura dell'intorno più vicino, ed un raggruppamento morfologico, per individuare lunghi segmenti rettilinei ed altri segmenti, anche più corti, ma paralleli ai primi (come mostrato in Fig. D.15).

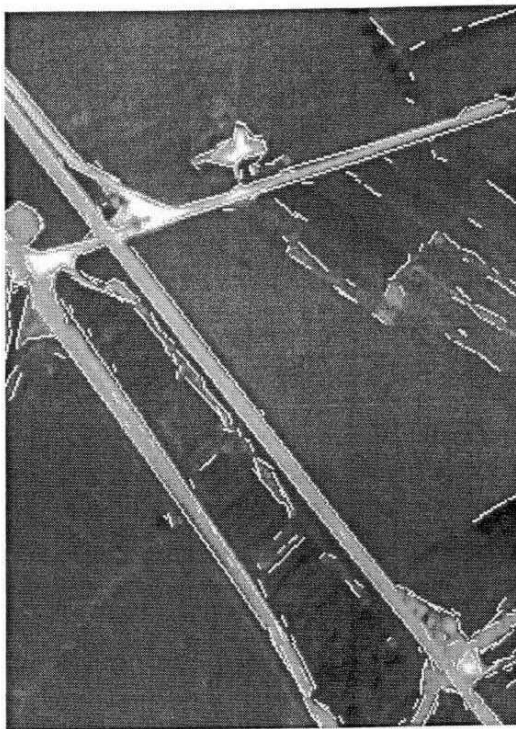


Fig. D.14 – Immagine aerea con superimposte linee dopo sogliatura in base all'intensità del tono di grigio

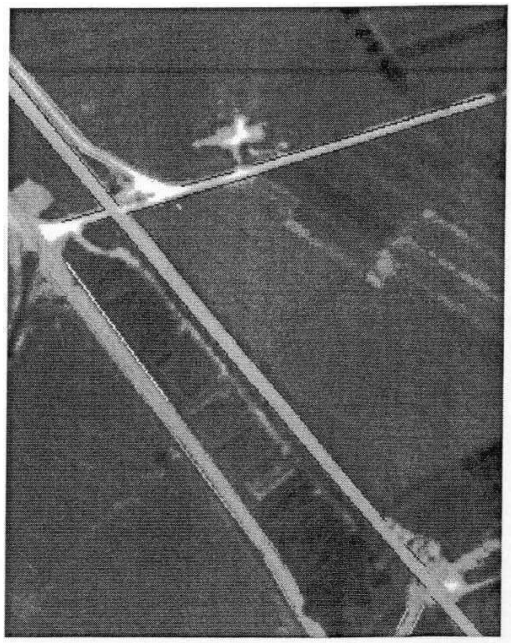


Fig. D.15 – Immagine aerea con superimposte linee dopo loro clusterizzazione e raggruppamento morfologico

BIBLIOGRAFIA DI RIFERIMENTO [67]

1. Acheson J.D. (1952): Quality Control and Industrial Statistics. R.D. Irwin, Homewood (Illinois).
2. Agresti A. (1984): Analysis of Ordinal Categorical Data. Wiley & Sons, New York.
3. Andersen E.B. (1980): Discrete Statistical Models with Social Science Applications. North Holland Publishing Company, Amsterdam.
4. Ball G.H., Hall D.J. (1967): A Clustering Technique for Summarizing Multivariate Data. In: *Behavioral Science*, 12 (1967), p. 153–155.
5. Barnett V. (1973): Comparative Statistical Inference. John Wiley & Sons, London.
6. Barry A.B. (1978): Errors in Practical Measurement in Science, Engineering, and Technology. John Wiley & Sons, New York.
7. Beyer W. H, (1966): CRC Handbook of Tables for Probability and Statistics. CRC Press, Boca Raton (Florida).
8. Bury K.V. (1975): Statistical Models in Applied Science. Wiley & Sons, New York.
9. Chiswick B.R., Chiswick S.J. (1975): Statistics and Econometrics. A Problem–solving Text. Uninersity Park Press, Baltimore.
10. Cochran W.G. (1977): Sampling Techniques. John Wiley & Sons, New York.
11. Cox D.R., Hinkley D.V. (1974): Theoretical Statistics. Chapman and Hall, Londra.
12. Cramer H. (1946): Mathematical Methods of Statistics. Princeton University Press, Princeton.
13. Cressie N.A. (1991): Statistics for Spatial Data. John Wiley and Sons, New York.
14. David F.N. (1954): Tables of The Ordinates and Probability Integral of the Distribution of the Correlation Coefficient in Small Samples. Cambridge University Press, Cambridge.
15. Davies O.L. (1960): The Design and Analysis of Industrial Experiments. Oliver and Boyd, Londra.
16. Davies O.L. (1961): Statistical Methods in Research and Production. Oliver and Boyd, Londra.
17. Draper N.R., Smith H. (1981): Applied Regression Analysis. Wiley and Sons, New York.
18. Federer W.T. (1973): Statistics and Society. Data Collection and Interpretation. M. Dekker, New York.
19. Fischer F.E. (1973): Fundamental Statistical Concepts. Canfield Press, San Francisco.
20. Franwaert E. (1987): L1 in Fuzzy Clustering. In: Y. Dodge (Ed), *Statistical Data Analysis Based on the L*1*–norm and Related Methods*, North Holland, p. 417–426.
21. Guttman I., Wilks S.S., Hunter S. J. (1965): Introductory Engineering Statistics, John Wiley & Sons, New York.
22. Hawkins D.M. (1982): Topics in Applied Multivariate Analysis. Cambridge University Press, London.
23. Jobson J.D. (1991/2): Applied Multivariate Data Analysis: vol. I & II. Springer, New York.
24. Kaufman L., Rousseeuw P.J. (1987): Clustering by Means of Medoids. In: Y. Dodge (Ed), *Statistical Data Analysis Based on the L*1*–norm and Related Methods*, North Holland, p. 405–416.
25. Kaufman L.,Rousseeuw P.J. (1990): Finding Groups in Data. Wiley & Sons, New York.
26. Kendall M. (1980): Multivariate Analysis. Charles Griffin & Company LTD, London.
27. Lauritzen S.L. (1998): Graphical Models. Clarendon Press, Oxford.
28. McPherson G. (1990): Statistics in Scientific Investigation. Springer, New York.
29. Middlebroks E.J. (1976): Statistical Calculation, How to Solve Statistical Problems. Ann Arbor Science, Ann Arbor (Michigan).
30. Mood A.M., Graybill F.A., Boes D.C. (1974): Introduction to the Theory of Statistics. Mc Graw–Hill Book Company, Londra.
31. Papoulis A. (1965): Probability Random Variables and Stochastic Processes. McGraw–Hill, New York.
32. Pfeiffer P.E. (1978): Concepts of Probability Theory. Dover Publications, New York.
33. Rao C.R. (1965): Linear Statistica] Inference and Its Applications. Wiley & Sons, New York.
34. Sachs L. (1984): Applied Statistics. A handbook of Techniques. Springer, New York.
35. Savage L.J. (1954): The Foundations of Statistics. John Wiley & Sons, New York.
36. Spath H. (1987): Using the L1 norm within clauster analysis. In: Y. Dodge (Ed), *Statistical Data Analysis Based on the L*1*–norm and Related Methods*, North Holland, p. 427–433.
37. Tanner M.A. (1996): Tools for Statistical Inference. Springer, New York.
38. Wald A. (1947): Sequential Analysis. Wiley & Sons, New York.
39. Watson G.N. (1948): A Treatise on the Theory of Bessel Functions. Cambridge University Press, Cambridge, Mac Millan Company, New York.
40. Whittle P. (1990): Risk–sensitive Optimal Control. John Wiley & Sons, Chichester (West Sussex).

[67] A complemento della bibliografia fornita, resta da osservare come pressoché sterminata sia la bibliografia relativamente all'inferenza ed all'analisi multivariata.