

FACOLTÀ DI ECONOMIA E COMMERCIO
ISTITUTO DI STATISTICA E RICERCA OPERATIVA

UNIVERSITÀ DEGLI STUDI DI TRENTO

GIORGIO VITTADINI

**SULL'USO DELLE COMPONENTI PRINCIPALI COME
INPUT NELLA CLUSTER ANALYSIS AGGLOMERATIVA**

Estratto da: « *Quaderni di Statistica e Matematica applicata*
alle Scienze Economico-Sociali » - Volume XII - n. 2 - Luglio 1990 - pag. 11-22

Sull'uso delle componenti principali come input nella cluster analysis agglomerativa

GIORGIO VITTADINI
Unità predepartimentale Statistico Matematica
Facoltà di Economia e Commercio - Università di Brescia

1. Introduzione

Date p variabili osservate su n unità di una popolazione, si definisce analisi gerarchica agglomerativa dei gruppi (cluster analysis gerarchica: C.A.) l'insieme delle tecniche atte a costruire un sistema gerarchico di partizioni successive formate, rispettivamente, da $n-1$, $n-2$, ..., 2, 1 gruppi di unità omogenee rispetto alle p variabili, secondo un grado di omogeneità via via decrescente.

E' noto che i risultati che si conseguono con tali analisi dipendono dalla scelta sia delle p variabili, sia della distanza, sia del legame o algoritmo aggregativo (es. Cormack, 1971; Anderberg, 1973; Benzecri, 1973; Cailliez Pages, 1976; Hartigan, 1977; Solomon, 1977; etc.). Particolare attenzione ha trovato nella letteratura il problema dell'eccessiva numerosità delle variabili che comporta difficoltà di interpretazione dei risultati conseguiti (Sokal, 1977; Everitt, 1980, p. 49). Molti autori (fra gli altri: Leoni, 1977; Sokal, 1977; Zani, 1977, p. 56; Tomassone, 1980; Everitt, 1980, p. 49; Joliffe, 1986, p. 163), partendo dalla considerazione che le distanze euclidee calcolate sulle p variabili originarie e sulle p componenti principali (C.P.) da esse ricavate sono equivalenti (Joliffe, 1986, p. 163), avanzano la proposta di ovviare all'eccessiva numerosità applicando l'analisi gerarchica anziché a tutte le variabili, alle prime q C.P. ($q < p$) cui sia associata una quota molto elevata della varianza iniziale.

Nel presente lavoro si mette in luce che tale proposta non sempre appare adeguata agli scopi della C.A.. Infatti i risultati delle C.A. condotte sulle p variabili originarie e su q C.P. cui sia associata un'elevata quota della varianza

iniziale possono essere molto differenti, pur se si adotta la medesima distanza e il medesimo legame.

Di conseguenza, una ponderata valutazione sull'opportunità di utilizzare nelle C.A., le q C.P. in luogo delle p variabili originarie può essere data solo a posteriori, dopo l'effettuazione delle C.A. stesse, quando, con indici di natura totalmente differente dalla quota di varianza, si misuri l'effettiva dissomiglianza nella configurazione dei gruppi. Nel lavoro, a questo scopo, si propone l'utilizzo dell'indice di Fowlkes e Mallows (1983).

Dopo aver precisato i simboli utilizzati nel lavoro (par. 2), si espongono le ragioni che stanno alla base della tesi sostenuta (par. 3). Quest'ultima è suffragata (par. 4) da numerose applicazioni effettuate su dati empirici ed ottenuti attraverso simulazione. Nelle conclusioni si avanzano, infine, alcune proposte operative.

2. Notazioni

Sia Z la matrice $(n \times p)$ delle n osservazioni centrate sulle p variabili e $R = (1/n)Z'Z$ la relativa matrice delle covarianze.

Inoltre, siano L e V rispettivamente le matrici degli autovalori e dei corrispondenti autovettori di R e $Y = ZV$ i punteggi determinati dalle componenti principali; analogamente si definiscono con V_q e $Y_q = Z V_q$ le matrici, rispettivamente dei primi q autovettori ($q < p$) e dei punteggi determinati dalle prime q C.P..

Tra le distanze più comunemente usate nella C.A. (Chiandotto ed al., 1981), vi sono le distanze euclidee $D(u_i, u_j)$ tra l'unità u_i e l'unità u_j (o i loro quadrati), calcolate su tutte le p variabili originarie ed equivalenti alle analoghe distanze su tutte le p C.P., cioè:

$$D^2(u_i, u_j) = (z_i - z_j)'(z_i - z_j) = (y_i - y_j)'(y_i - y_j) \quad (1)$$

ove z e y sono le i -esime righe rispettivamente delle matrici Z e Y .

Analogamente:

$$D^2(u_i, u_j) = (y_i - y_j)'(y_i - y_j) = (z_i - z_j)'V_q V_q'(z_i - z_j) \quad (2)$$

rappresenta il quadrato delle distanze euclidee relative alle prime q C.P.

Infine, tra i legami più comunemente utilizzati nella procedura agglomerativa con la distanza euclidea si ricordano (Chiandotto ed al., 1981) il legame singolo, il legame completo, quello del centroide e il legame mediano.

3. Perdita di informazione e quota di varianza associata a q C.P. ($q < p$)

Il contributo di Robert ed Escoufier (1976) permette di interpretare il significato del passaggio da un insieme di unità definite su p variabili ad un insieme basato su q C.P..

Siano Z_1 ($n \times p$) e Z_2 ($n \times q$) le matrici delle osservazioni centrate rispettivamente su p e q variabili. Siano $C(Z_1)$ e $C(Z_2)$ gli spazi lineari generati rispettivamente da Z_1 e Z_2 . Allo scopo di misurare il grado di similarità fra i due spazi, cioè per valutare in che misura i suddetti spazi sono correlati linearmente, si utilizza il seguente indice:

$$M = \sqrt{2} \left\{ 1 - \frac{tr[cov(Z_1, Z_2)cov(Z_2, Z_1)]}{\{tr[cov(Z_1, Z_1)]^{1/2} tr[cov(Z_2, Z_2)]^{1/2}\}} \right\}^{1/2} = \sqrt{2} \{1 - r(Z_1, Z_2)\}^{1/2} \quad (3)$$

dove $r(Z_1, Z_2)$ è coefficiente normalizzato ($0 \leq r(Z_1, Z_2) \leq 1$) che assume valori tanto più elevati quanto più simili sono i due spazi.

Un caso importante è quello in cui si confrontano due spazi, l'uno generato dalle p variabili Z , l'altro da " q nuove variabili Y tra loro ortogonali, combinazioni lineari delle variabili di partenza tali che gli spazi $C(Z)$ e $C(Y)$ siano il più simili possibile" (Robert et al. 1976, p.260). In tal caso l'indice r assume la seguente forma:

$$r(Z, Y_q) = [tr(V_q' R^2 V_q) / tr R]^{1/2} = \left[\left(\sum_{j=1}^q l_j^2 \right) / \left(\sum_{j=1}^p l_j^2 \right) \right]^{1/2} \quad (4)$$

essendo V_q la matrice le cui colonne sono i primi q autovettori di R , matrice delle covarianze di Z , ed l_j ($j=1,2,\dots,p$) il j -esimo autovalore di R .

Dall'ultimo membro della (4), emerge che l'indice r è pari alla radice quadrata della somma dei quadrati dei primi q autovalori rapportata alla somma dei quadrati di tutti gli autovalori. D'altro canto si può facilmente dimostrare che l'indice r è sempre maggiore della quota di varianza, usualmente utilizzata per valutare la perdita di informazione, intendendo con il termine "informazione" il contributo alla spiegazione della variabilità del fenomeno. Per confrontare insiemi di unità definite su p variabili e q C.P. sembra quindi preferibile fare riferimento all'indice r , in quanto oltre ad avere la stessa utilità della quota di varianza per misurare la perdita di informazione, ha il significato teorico di

misurare il grado di somiglianza fra i relativi spazi lineari.

Ora, si ritiene generalmente giustificata anche nell'ambito della C.A. l'adozione della distanza euclidea su q C.P., cui sia associata una quota elevata di varianza; per quanto detto in precedenza, ciò avviene a maggior ragione se si fa riferimento all'indice r .

Tale modo di procedere tuttavia, non sembra adeguato. Da una parte, infatti, essendo lo scopo della C.A. quello della costruzione di una successione di partizioni in gruppi omogenei, il concetto di perdita di informazione deve essere innanzitutto inteso come diminuzione di "somiglianza" nella composizione dei gruppi che si vengono formando in corrispondenza dei vari passi della C.A. (Fowlkes ed al., 1983) piuttosto che come differenza globale tra le distanze iniziali. Dall'altra, come affermano diversi autori (Jolliffe, 1986, p. 164; Bock, 1987), la massimizzazione della varianza complessiva dei dati proiettati nello spazio q -dimensionale non comporta necessariamente la "conservazione" della struttura dei gruppi "naturali" a minima varianza interna nello spazio p -dimensionale.

Quanto ricordato porta a concludere che, qualunque sia il legame adottato, un elevato valore dell'indice r o della quota di varianza associati alle prime q C.P. possono in alcuni casi coesistere con una scarsa somiglianza tra la composizione dei gruppi basati su p variabili e q C.P. ai vari passi delle C.A..

Per capire il motivo di questa affermazione occorre innanzitutto ricordare che nella C.A. agglomerativa non conta soltanto il valore globale delle distanze ma anche il loro ordine. In particolare, se il legame scelto è quello completo o singolo, la composizione dei gruppi è determinata solo dall'ordine; se i legami scelti sono il centroide o il mediano, invece, sono rilevanti sia l'ordine, sia il valore delle distanze. La discordanza tra due ordinamenti di $n(n-1)/2$ distanze è misurabile attraverso il coefficiente di correlazione tra ranghi di Spearman:

$$c = 1 - [f(r^2 - 1)]^{-1} 6 \sum_{s=1}^h (s - h_s)^2 \quad f = n(n-1)/2, \quad (5)$$

dove s , h_s sono l' s -esimo ed h -esimo rango attribuiti alle distanze tra due generiche unità u_i e u_j nei suddetti ordinamenti.

Poiché non esiste relazione univoca tra differenza globale fra le distanze e discordanza nei loro ordinamenti, a valori molto elevati dell'indice r (4) e della

quota di varianza possono corrispondere valori significativamente inferiori dell'indice c (5). In questi casi, come si mostrerà nel paragrafo 4, la somiglianza fra la composizione dei gruppi può risultare assai scarsa, qualunque sia il legame adottato.

Tuttavia, neppure il fatto che il coefficiente di Spearman assuma valori vicini al suo massimo è garanzia che esista forte somiglianza fra i risultati di due C.A. Infatti, neanche tra discordanza negli ordinamenti delle distanze e dissomiglianza fra le configurazioni dei gruppi ai vari passi delle C.A. esiste una qualche relazione univoca. Perciò, pur in modo diverso, qualunque sia il legame utilizzato piccole differenze negli ordinamenti di distanze possono comportare grandi dissomiglianze nella composizione dei gruppi.

In generale, ci si attende che la somiglianza fra i gruppi sia maggiore nel caso di adozione del legame singolo e del legame del centroide che non per gli altri legami menzionati. Infatti, quando si adotta il legame singolo si fondono ad ogni passo i gruppi caratterizzati da minima distanza fra coppie di unità ad essi appartenenti. Perciò solo inversioni negli ordinamenti riguardanti le $n-1$ distanze prese in considerazione nel processo agglomerativo producono variazioni nei risultati finali: inoltre, piccole distorsioni nell'ordinamento delle distanze tendono ad avere sempre meno incidenza nel processo agglomerativo man mano che i gruppi aumentano di numerosità.

Analogamente, utilizzando il legame del centroide, eventuali discordanze nell'ordinamento delle distanze tendono ad essere riassorbite poiché le aggregazioni sono decise, ad ogni passo, rispettivamente, sulla base di medie di distanze e in base a distanze ponderate fra i valori medi.

Adottando il legame completo, invece, entrano in gioco non solo le distanze minori, ma anche quelle più elevate che ad ogni passo vengono confrontate fra loro. Ne consegue che, se ad un certo passo muta la minima distanza fra le distanze massime, viene radicalmente modificato il successivo processo aggregativo. Analogamente accade per il legame mediano, adottando il quale si aggregano i due gruppi caratterizzati dalle distanze minime fra i rispettivi centroidi. Poiché nel calcolare il nuovo centroide si prescinde dalla numerosità delle unità comprese nei gruppi, anche pochi mutamenti nell'ordinamento di distanze possono provocare forti variazioni nel processo aggregativo.

Per misurare la dissomiglianza tra i risultati ottenuti ad ogni passo di due

C.A., occorre adottare l'indice normalizzato b_k di Fowlkes e Mallows (1983). Data la k -esima ($k=1, 2, \dots, n-1$) partizione ottenuta mediante due distinte procedure aggregative A_1 e A_2 effettuate sullo stesso insieme di n unità, m_d corrisponde al numero di unità comuni al c -esimo cluster di A_1 ed al d -esimo cluster di A_2 . L'indice b_k ($0 \leq b_k \leq 1$) è così definito:

$$b_k = t_k / (v_k w_k)^{1/2} \quad (6)$$

con:

$$\begin{aligned} t_k &= \sum_d m_{cd}^2 - n; \\ m_c &= \sum_d m_{cd}; \quad m_d = \sum_c m_{cd}; \\ v_k &= \sum_c m_c^2 - n; \quad w_k = \sum_d m_d^2 - n; \end{aligned}$$

Quanto più elevato è il valore di b_k , tanto più elevata è la somiglianza fra le due partizioni; in particolare, quando b_k è uguale ad uno, si ha perfetta somiglianza; quando b_k è uguale a zero, totale dissomiglianza. Avvicinandosi agli ultimi passi della C.A. ($k=n-1$) l'indice diviene via via meno significativo in quanto il suo valore è fortemente influenzato da aggregazioni "forzate" fra unità e gruppi sicuramente molto dissimili².

4. Verifiche empiriche³

Le verifiche empiriche hanno lo scopo di mettere in luce l'esistenza di "contro-esempi" in cui in presenza di indici r (e di quote di varianza) di valore elevato corrispondono, a diversi passi delle C.A. basate su p variabili e su q C.P.

² Una possibile obiezione alle affermazioni fatte nel precedente paragrafo è che la riduzione di dimensione non è una perdita ma un guadagno di informazioni in quanto le prime C.P. individuano i gruppi naturali esistenti nella popolazione mentre le ultime C.P. rendono solo più sfumati e confusi tali gruppi (Greco 1983), ovvero mettono in luce solo un indistinto "rumore di fondo" (Marriot 1974, p. 23).

Vi sono solo due valide risposte a questa obiezione. Innanzitutto, se si ritiene superfluo il confronto fra le composizioni dei gruppi ottenuti sulla base di q C.P. e tutte le variabili a differenza di quanto accade usualmente si deve evitare anche di interpretare il significato delle C.P. e quindi dei gruppi omogenei mediante le variabili originali. Ci si condanna così a una totale impotenza interpretativa.

Secondariamente per costruire gruppi "ottimi" di unità basati sulle prime q C.P. sono più opportune della sequenza riduzione C.P. - C.A., tecniche statistiche opportunamente costruite per ottenere una rappresentazione ottimale dei dati nello spazio q -dimensionale (nel senso espresso dalla menzionata definizione di Escoffier) e contemporaneamente dar vita a un sistema di gruppi della minima varianza interna.

³ Si ringrazia il dott. Maurizio Sanarico per l'aiuto alla predisposizione dei programmi di elaborazione.

da esse estratte, indici b_k dai valori molto lontani da 1 (<0.70), qualunque sia il legame adottato.

A tal fine si sono presi in esame nove diversi esempi di cui sette sono basati su dati simulati e due su dati reali.

Le distribuzioni utilizzate nel processo di simulazione per ottenere i dati sono state scelte fra quelle che, a nostro avviso, risultano particolarmente utili come modelli per i fenomeni della vita reale. Lo studio di simulazione ha avuto come oggetto gruppi di dati formati da 60 osservazioni sia sulle tre variabili della prima distribuzione sia sulle otto variabili delle altre distribuzioni.

I risultati sono riportati in tabelle, numerate con un ordine identico al seguente:

- I) 3 variabili a "cluster distinti"
- II) Variabili normali.
- III) Variabili beta.
- IV) Variabili gamma.
- V) Variabili poisson.
- VI) Variabili lognormali.
- VII) Variabili miscuglio. [3 N(0,1), 3 Beta, 2 Gamma]

Per quanto riguarda i dati reali si sono considerate nel primo caso cinquanta osservazioni su 6 variabili e nel secondo caso 51 osservazioni su 7 variabili.

VIII) Indici di dotazione (N. Addetti/popolazione residente) per agricoltura, industria, servizi per il sistema produttivo, servizi per le famiglie, servizi di rete, servizi di destinazione collettiva neibacini del lavoro dell'Italia Centrale (Fonte: ISCOM 1989, pp. 134-136).

IX) % voti ai partiti nelle elezioni in comuni della provincia di Firenze (Fonte: Chiandotto 1981, p. 126).

Le variabili I-VIII sono centrate, le variabili IX sono standardizzate.

Per ognuna delle variabili considerate si sono calcolate le distanze euclidee su tutte le p C.P. (equivalenti alle distanze euclidee su tutte le variabili) e su q C.P. cui sia associata una quota sufficientemente elevata della varianza iniziale (≥ 0.75 per le C.P. estratte da variabili ottenute mediante simulazione, ≥ 0.80 per le C.P. estratte da variabili osservate) e si sono calcolati la quota di varianza

corrispondente alle prime $p-1, p-2, \dots, q$ C.P., l'indice r e il coefficiente c di Spearman. Si sono poi costruiti gruppi gerarchici utilizzando i legami singolo, completo, centroide e mediano, calcolando i valori dell'indice b_k a partire dal 20° passo.

Si esamina innanzitutto il caso in cui (Tab. I) i valori dell'indice r e delle quote di varianza associate alla prima C.P. (0.99 e 0.91 rispettivamente) sono molto elevati, mentre il valore del coefficiente di Spearman, relativo agli ordinamenti di distanze su tre variabili e una C.P. pur rimanendo sufficientemente elevato è chiaramente inferiore (0.80). Si osserva che effettuando l'agglomerazione gerarchica su tali dati mediante distanza euclidea con legami singolo, completo, mediano e centroide, qualunque sia il legame adottato, il valore degli indici b_k è in generale molto basso.

In un secondo insieme di casi (Tab. II-IX) sono mostrate situazioni in cui sia i valori degli indici r che quelli dei coefficienti di Spearman sono molto prossimi al loro massimo.

Nelle C.A. condotte utilizzando legami completo e mediano, gli indici b_k sono in netta prevalenza di basso valore (< 0.70); nelle C.A. condotte mediante il legame singolo l'indice b_k è talvolta di basso valore ai primi passi delle C.A. ma tende quasi sempre a diventare di valore elevato nei passi centrali e finali della C.A.; nella quasi totalità dei casi in cui si utilizza il legame del centroide l'indice b_k è molto elevato. Esistono però alcuni esempi sia per il legame singolo (Tab. II, V, VIII), che per il legame del centroide (Tab. VIII) in cui l'indice b_k rimane di basso valore in tutti i passi delle C.A..

5. Conclusioni

Le considerazioni svolte e le analisi empiriche presentate permettono di affermare che C.A. condotte su q C.P. a cui sia associata una elevata quota della varianza iniziale possono dar vita a risultati molto differenti da C.A. basate su tutte le p variabili originarie, pur se si adotta la medesima distanza euclidea e il medesimo legame. Ciò avviene, in particolare, se il coefficiente di correlazione di Spearman relativo agli ordinamenti delle distanze non è altrettanto elevato della quota di varianza o dell'indice r ; tuttavia anche quando il suo valore è prossimo a uno la somiglianza dei gruppi può essere assai scarsa, in particolare se si adottano i legami completo e mediano.

E' quindi necessario, per valutare l'effettiva variazione nei risultati, e l'opportunità del passaggio da C.A. basate sulle variabili originarie a C.A. basate su q C.P., prendere sempre in considerazione il valore di indici inerenti la somiglianza nella configurazione dei gruppi, quali ad esempio, l'indice b_k di Fowlkes e Mallows.

Bibliografia

- Anderson M.R. (1973) *C.A. for Applications*, Academic Press, New York.
- Benzecri J.P. (1973) *L'Analyse des Données*, vol. I, Dunod, Parigi.
- Bock H.H. (1987) *On the interface between C.A., principal component analysis and multidimensional scaling* in M. Bordini and A.K. Gupta (eds) *Multivariate statistical modelling and data analysis* D. Reidel, Dordrecht, pp.17-34.
- Cailliez F., Pages J.P. (1976) *Introduction à l'analyse des données*, Smash, Parigi.
- Chandonnet B., Marchetti G. (1981) *L'analisi dei gruppi: una metodologia per lo studio del comportamento elettorale*, Dip. Stat. Un. di Firenze.
- Cormack R.M. (1971) *A review of classification*, Journal of Royal Statistical Society, vol. 134, pp. 321-333.
- Everitt B.S. (1980), *Cluster Analysis*, Educ. Books, Londra.
- Fowlkes E.B., Mallows C.L. (1983), *A method for comparing two hierarchical clusterings*, Journal of the American Statistical Association, 78, pp. 553-584.
- Greco L. (1983) *Sulle proprietà delle componenti principali come metodo per la sintesi delle informazioni statistiche*, Statistica, 1, vol. 43, pp. 49-66.
- Hamigan J.A. (1977) *Clustering algorithms*, Wiley, New York.
- ISCOM (1989) *Il sistema del Terziario in Italia*, Roma.
- Jolliffe I.T. (1986) *Principal component analysis*, New York.
- Leon R. (1977), *Alcuni problemi di analisi statistica multivariata applicata al problema della localizzazione industriale*, in Atti del seminario su due temi di analisi statistica multivariata, op. cit., pp. 35-48.
- Marron F.M. (1974) *The interpretation of multiple observations*, Academic Press, Londra.
- Robert P., Escoufier Y. (1976) *A unifying tool for linear multivariate statistical methods: the R² coefficient*, Applied Statistics, vol. 25, pp. 257-265.
- Sokal R.R. (1977) *Background and current directions in cluster analysis*, in Van Ryzin J. ed., *Classification and clustering*, Academic Press, New York, pp. 1-15.
- Solomon H. (1977) *Data different clustering techniques*, in Van Ryzin J. ed., *Classification and clustering*, Academic Press, New York, pp. 155-171.
- Tomassone R. (1980) *De l'analyse des données à la modélisation*, in Atti del convegno: *Analyse des données rencontre avec l'école française*, Napoli: 30 giugno - 5 luglio, 1, pp. 21-78.
- Zani S. (1977) *L'analisi classificatoria: contributi metodologici ed impiego per l'individuazione di aree omogenee*, in A.A.V.V., *Aree di sviluppo socio-economico e comprensori in Emilia Romagna*, Istituto di statistica, Facoltà di Economia e commercio, Parma, pp. 39-108.

Tab. I - VARIABILI A "DUE CLUSTER DISTINTI" - DISTANZA EUCLIDEA SU 3 VARIABILI E SULLA PRIMA C.P.

VARIABILI	N CP.	QUOTA VAR.	r	COEFF. DI SPEARMAN	b _i					
					20	25	30	35	40	45
LEGAME SINGOLO	1	0.91	0.99	0.80	0.49	0.32	0.29	0.58	0.70	0.98
LEGAME CENTROIDE	1	0.91	0.99	0.80	0.38	0.38	0.36	0.45	0.51	0.69
LEGAME COMPLETO	1	0.91	0.99	0.80	0.36	0.29	0.30	0.39	0.49	0.64
LEGAME MEDIANO	1	0.91	0.99	0.80	0.38	0.38	0.36	0.39	0.50	0.64

Tab. II - VARIABILI NORMALI - DISTANZA EUCLIDEA SULLE 8 VARIABILI E SULLE PRIME 7, 6, 5 C.P.

VARIABILI NORMALI	N. CP.	QUOTA VAR.	r	COEFF. DI SPEARMAN	b _i							
					20	25	30	35	40	45	50	55
LEGAME SINGOLO	7	0.93	0.98	0.98	0.57	0.57	0.83	1	1	1	1	1
LEGAME SINGOLO	6	0.86	0.96	0.96	0.86	0.56	0.80	0.79	0.64	0.64	0.87	0.92
LEGAME SINGOLO	5	0.77	0.92	0.92	0.46	0.29	0.53	0.57	0.52	0.67	0.91	1
LEGAME COMPLETO	7	0.93	0.98	0.98	0.75	0.66	0.66	0.64	0.58	0.52	0.37	0.40
LEGAME COMPLETO	6	0.86	0.96	0.95	0.55	0.51	0.53	0.56	0.45	0.39	0.32	0.36
LEGAME COMPLETO	5	0.77	0.92	0.91	0.54	0.55	0.46	0.46	0.40	0.40	0.43	0.40
LEGAME CENTROIDE	7	0.93	0.98	0.98	1	0.96	1	1	0.95	1	1	1
LEGAME CENTROIDE	6	0.86	0.96	0.95	1	1	1	0.94	1	1	1	1
LEGAME CENTROIDE	5	0.77	0.92	0.91	0.56	0.90	1	0.93	0.95	0.98	1	0.96
LEGAME MEDIANO	7	0.93	0.98	0.95	0.53	0.68	0.64	0.48	0.88	0.87	0.96	0.53
LEGAME MEDIANO	6	0.86	0.96	0.95	0.44	0.76	0.61	0.54	0.44	0.66	0.98	0.44
LEGAME MEDIANO	5	0.77	0.92	0.91	0.44	0.38	0.43	0.41	0.56	0.65	0.67	0.44

Tab. III - VARIABILI BETA. DISTANZA EUCLIDEA SULLE 8 VARIABILI E SULLE PRIME 7, 6, 5 C.P.

VARIABILI BETA	N. CP.	QUOTA VAR.	r	COEFF. DI SPEARMAN	b _i							
					20	25	30	35	40	45	50	55
LEGAME SINGOLO	7	0.93	0.98	0.97	0.96	0.88	0.74	0.76	0.97	0.84	0.92	1
LEGAME SINGOLO	6	0.84	0.95	0.93	0.96	0.56	0.80	0.65	0.79	0.68	0.87	0.92
LEGAME SINGOLO	5	0.75	0.95	0.86	0.46	0.29	0.55	0.57	0.52	0.65	0.92	1
LEGAME COMPLETO	7	0.93	0.98	0.97	1	0.88	0.72	0.73	0.66	0.64	0.46	0.41
LEGAME COMPLETO	6	0.84	0.95	0.93	0.55	0.51	0.38	0.36	0.38	0.29	0.24	0.26
LEGAME COMPLETO	5	0.75	0.97	0.86	0.25	0.23	0.24	0.28	0.28	0.27	0.29	0.5
LEGAME CENTROIDE	7	0.93	0.98	0.97	1	1	0.93	1	1	1	1	1
LEGAME CENTROIDE	6	0.84	0.95	0.93	0.83	0.91	0.84	0.93	1	1	1	1
LEGAME CENTROIDE	5	0.75	0.91	0.86	0.53	0.38	0.67	0.80	0.90	0.95	0.96	1
LEGAME MEDIANO	7	0.93	0.98	0.97	0.86	0.82	0.83	0.75	0.52	0.59	0.57	0.67
LEGAME MEDIANO	6	0.84	0.95	0.93	0.55	0.38	0.38	0.54	0.46	0.46	0.68	0.85
LEGAME MEDIANO	5	0.75	0.91	0.86	0.54	0.31	0.35	0.34	0.34	0.37	0.40	0.91

Tab. IV - VARIABILI GAMMA. DISTANZA EUCLIDEA SULLE 8 VARIABILI E SULLE PRIME 7, 6, 5 C.P.

VARIABILI GAMMA	N. CP.	QUOTA VAR.	r	COEFF. DI SPEARMAN	b _i							
					20	25	30	35	40	45	50	55
LEGAME SINGOLO	7	0.94	0.98	0.98	1	0.65	0.63	1	0.92	0.91	1	1
LEGAME SINGOLO	6	0.88	0.97	0.97	0.98	0.60	0.85	0.93	0.91	0.90	1	0.96
LEGAME SINGOLO	5	0.79	0.89	0.89	0.68	0.71	0.82	0.90	0.89	0.90	0.96	0.96
LEGAME COMPLETO	7	0.94	0.98	0.98	0.52	0.70	0.71	0.60	0.49	0.53	0.41	0.52
LEGAME COMPLETO	6	0.88	0.97	0.97	0.62	0.55	0.53	0.46	0.47	0.38	0.36	0.37
LEGAME COMPLETO	5	0.79	0.89	0.89	0.49	0.46	0.49	0.46	0.35	0.30	0.49	0.59
LEGAME CENTROIDE	7	0.94	0.98	0.98	0.88	0.88	1	1	0.95	1	1	1
LEGAME CENTROIDE	6	0.88	0.97	0.97	0.63	0.83	1	0.97	0.92	0.93	1	1
LEGAME CENTROIDE	5	0.79	0.89	0.89	0.82	0.91	1	0.94	0.92	0.93	1	1
LEGAME MEDIANO	7	0.94	0.98	0.98	0.85	0.53	0.50	0.65	0.85	0.74	0.84	0.92
LEGAME MEDIANO	6	0.88	0.97	0.95	0.83	0.55	0.55	0.47	0.54	0.54	0.63	0.85
LEGAME MEDIANO	5	0.79	0.89	0.89	0.68	0.43	0.44	0.44	0.39	0.55	0.81	0.89

Tab. V - VARIABILI POISSON. DISTANZA EUCLIDEA SULLE 8 VARIABILI E SULLE PRIME 6, 5, 4 C.P.

VARIABILI POISSON	N. CP.	QUOTA VAR.	r	COEFF. DI SPEARMAN	b _i							
					20	25	30	35	40	45	50	55
LEGAME SINGOLO	7	0.94	0.98	0.98	0.76	0.56	0.85	0.97	0.94	0.95	1	1
LEGAME SINGOLO	6	0.86	0.96	0.95	0.83	0.53	0.57	0.67	0.67	0.95	0.95	0.96
LEGAME SINGOLO	5	0.77	0.93	0.91	0.61	0.64	0.88	0.86	0.88	0.97	0.96	0.95
LEGAME COMPLETO	7	0.94	0.98	0.98	0.75	0.66	0.66	0.64	0.58	0.52	0.37	0.40
LEGAME COMPLETO	6	0.86	0.96	0.95	0.55	0.51	0.53	0.56	0.45	0.39	0.32	0.36
LEGAME COMPLETO	5	0.77	0.93	0.91	0.54	0.51	0.53	0.56	0.45	0.39	0.32	0.36
LEGAME CENTROIDE	7	0.94	0.98	0.98	0.85	1	0.93	0.97	0.95	1	1	1
LEGAME CENTROIDE	6	0.86	0.96	0.95	0.55	0.85	1	0.96	0.96	0.95	1	0.98
LEGAME CENTROIDE	5	0.77	0.93	0.91	0.92	0.40	0.86	0.86	0.86	0.93	1	1
LEGAME MEDIANO	7	0.94	0.98	0.98	0.94	0.63	0.49	0.48	0.51	0.49	0.46	0.84
LEGAME MEDIANO	6	0.86	0.96	0.95	0.78	0.79	0.58	0.52	0.44	0.48	0.39	0.92
LEGAME MEDIANO	5	0.77	0.93	0.91	0.78	0.79	0.58	0.52	0.44	0.48	0.39	0.92

Tab. VI - VARIABILI LOGNORMALI. DISTANZA EUCLIDEA SULLE 8 VARIABILI E SULLE PRIME 7, 6, 5, 4 C.P.

VARIABILI LOGNORMALI	N. CP.	QUOTA VAR.	r	COEFF. DI SPEARMAN	b _i							
					20	25	30	35	40	45	50	55
LEGAME SINGOLO	7	0.97	0.99	0.99	0.49	1	0.84	0.91	0.95	1	1	1
LEGAME SINGOLO	6	0.92	0.98	0.98	0.66	0.75	0.89	0.91	0.92	0.95	1	1
LEGAME SINGOLO	5	0.85	0.97	0.97	0.59	0.66	0.88	0.91	0.89	0.86	1	1
LEGAME COMPLETO	7	0.97	0.99	0.99	0.51	0.36	0.50	0.58	0.60	0.63	0.98	1
LEGAME COMPLETO	6	0.92	0.98	0.98	0.64	0.65	0.69	0.58	0.44	0.63	0.56	1
LEGAME COMPLETO	5	0.85	0.97	0.97	0.80	0.65	0.52	0.53	0.48	0.59	0.55	1
LEGAME CENTROIDE	7	0.97	0.99	0.99	0.57	0.60	0.60	0.48	0.39	0.69	0.60	0.94
LEGAME CENTROIDE	6	0.92	0.98	0.98	0.58	0.39	0.53	0.52	0.43	0.41	0.38	0.96
LEGAME CENTROIDE	5	0.85	0.97	0.97	0.91	1	0.93	1	1	1	1	0.98
LEGAME MEDIANO	7	0.97	0.99	0.99	0.75	0.62	0.86	0.83	0.84	0.95	1	1
LEGAME MEDIANO	6	0.92	0.98	0.98	0.39	0.68	0.59	0.91	0.84	0.82	1	1
LEGAME MEDIANO	5	0.85	0.97	0.97	0.93	0.95	0.96	0.90	0.90	0.90	0.78	0.98
LEGAME MEDIANO	6	0.92	0.98	0.98	0.89	0.69	0.66	0.79	0.84	0.64	0.71	0.95
LEGAME MEDIANO	5	0.85	0.97	0.97	0.80	0.68	0.65	0.63	0.64	0.64	0.72	0.98
LEGAME MEDIANO	4	0.76	0.94	0.94	0.81	0.59	0.38	0.46	0.54	0.65	0.72	0.98

Tab. VII - VARIABILI MISCUGLIO. DISTANZA EUCLIDEA SULLE 8 VARIABILI E SULLE PRIME 7, 6, 5 C.P.

VARIABILI MISCUGLIO	N. CP	QUOTA VAR.	r	COEFF. DI SPEARMAN	b _i									
					20	25	30	35	40	45	50	55	55	55
LEGAME SINGOLO	5	0.96	0.99	0.99	0.86	0.91	0.73	0.83	1	1	1	1	1	1
LEGAME COMPLETO	4	0.89	0.98	0.97	0.36	0.29	0.56	0.97	0.88	1	1	1	1	1
LEGAME CENTROIDE	3	0.78	0.96	0.93	0.36	0.35	0.63	0.79	0.92	0.85	0.90	0.95	0.95	0.95
LEGAME MEDIANO	5	0.96	0.99	0.99	0.85	0.77	0.72	0.72	0.66	0.59	0.54	0.86	0.86	0.86
LEGAME SINGOLO	4	0.89	0.98	0.97	0.66	0.69	0.59	0.71	0.65	0.51	0.44	0.90	0.90	0.90
LEGAME COMPLETO	3	0.78	0.96	0.93	0.67	0.45	0.41	0.30	0.30	0.34	0.60	0.32	0.32	0.32
LEGAME CENTROIDE	5	0.96	0.99	0.99	0.83	0.78	0.93	0.85	1	1	1	1	1	1
LEGAME MEDIANO	4	0.89	0.98	0.97	0.65	0.65	0.92	0.83	0.97	0.97	0.98	1	1	1
LEGAME SINGOLO	3	0.78	0.96	0.93	0.39	0.56	0.68	0.61	0.84	0.90	0.90	0.84	0.84	0.84
LEGAME COMPLETO	5	0.96	0.99	0.99	0.88	0.77	0.67	0.80	0.91	0.89	0.84	0.81	0.81	0.81
LEGAME CENTROIDE	4	0.89	0.98	0.97	0.45	0.44	0.54	0.44	0.89	0.88	0.87	0.61	0.61	0.61
LEGAME MEDIANO	3	0.78	0.96	0.93	0.49	0.47	0.49	0.38	0.49	0.54	0.54	0.66	0.66	0.66

Tab. VIII - VARIABILI OCCUPAZIONALI. DISTANZA EUCLIDEA SULLE 6 VARIABILI E SULLE PRIME 2, 1 C.P.

VARIABILI "OCCUPAZIONALI"	N. CP	QUOTA VAR.	r	COEFF. DI SPEARMAN	b _i									
					20	25	30	35	40	45	50	55	55	55
LEGAME SINGOLO	2	0.92	0.99	0.98	0.58	0.60	0.74	0.85	1	0.82	0.97	0.91	0.91	0.91
LEGAME COMPLETO	1	0.80	0.98	0.89	0.42	0.60	0.68	0.63	0.66	0.50	0.52	0.98	0.98	0.98
LEGAME CENTROIDE	2	0.92	0.99	0.98	0.53	0.57	0.58	0.54	0.54	0.70	0.76	0.86	0.86	0.86
LEGAME MEDIANO	1	0.80	0.98	0.89	0.58	0.54	0.38	0.34	0.35	0.41	0.52	0.55	0.55	0.55
LEGAME SINGOLO	2	0.92	0.99	0.98	0.68	0.58	0.61	0.69	0.70	0.77	0.88	1	1	1
LEGAME COMPLETO	1	0.80	0.98	0.89	0.65	0.38	0.45	0.51	0.53	0.70	0.62	1	1	1
LEGAME CENTROIDE	2	0.92	0.99	0.98	0.66	0.69	0.80	0.61	0.76	0.88	0.92	0.68	0.68	0.68
LEGAME MEDIANO	1	0.80	0.98	0.89	0.55	0.61	0.54	0.48	0.50	0.54	0.67	0.74	0.74	0.74

Tab. IX - VARIABILI ELETTORALI. DISTANZA EUCLIDEA SULLE 8 VARIABILI E SULLE PRIME 6, 5, 4 C.P.

VARIABILI "ELETTORALI"	N. CP	QUOTA VAR.	r	COEFF. DI SPEARMAN	b _i									
					20	25	30	35	40	45	50	55	55	55
LEGAME SINGOLO	6	0.96	0.99	0.99	0.88	0.82	0.82	0.82	0.94	0.95	1	1	1	1
LEGAME COMPLETO	5	0.91	0.98	0.96	0.67	0.67	0.67	0.67	0.81	0.81	0.93	0.93	0.93	0.93
LEGAME CENTROIDE	4	0.81	0.96	0.92	0.50	0.50	0.50	0.50	0.81	0.81	0.84	0.84	0.84	0.84
LEGAME MEDIANO	6	0.96	0.99	0.99	0.71	0.71	0.71	0.71	0.60	0.68	0.86	0.86	0.86	0.86
LEGAME SINGOLO	5	0.91	0.98	0.96	0.67	0.67	0.67	0.67	0.81	0.81	0.84	0.84	0.84	0.84
LEGAME COMPLETO	4	0.81	0.96	0.92	0.50	0.50	0.50	0.50	0.81	0.81	0.84	0.84	0.84	0.84
LEGAME CENTROIDE	6	0.96	0.99	0.99	0.89	0.89	0.89	0.89	0.65	0.65	0.82	0.82	0.82	0.82
LEGAME MEDIANO	5	0.91	0.98	0.96	0.70	0.70	0.70	0.70	0.66	0.66	0.82	0.82	0.82	0.82
LEGAME SINGOLO	4	0.81	0.96	0.92	0.80	0.80	0.80	0.80	0.75	0.75	0.71	0.71	0.71	0.71
LEGAME COMPLETO	6	0.96	0.99	0.99	0.85	0.85	0.85	0.85	0.57	0.57	0.51	0.51	0.51	0.51
LEGAME CENTROIDE	5	0.91	0.98	0.96	0.77	0.77	0.77	0.77	0.62	0.62	0.63	0.63	0.63	0.63
LEGAME MEDIANO	4	0.81	0.96	0.92	0.77	0.77	0.77	0.77	0.57	0.57	0.51	0.51	0.51	0.51