***S**tatistica & **A**pplicazioni*
Vol. II, n. 1, 2004
pp. 73-88

Una generalizzazione della distanza di Cayley per l'analisi dei ranghi

Aride Mazzali[§]

Summary: *In this paper we consider the Cayley's distance for rank data and a generalization that we are going to introduce. The Cayley's distance is drawn by the number of the inversions in the permutation function derived composing ranking and ordering functions, while the generalized distance comes from a monotonic non decreasing transformation of the previous one. The generalized distance is then used to define the center and the spread of a data set. The important concepts of total and relative disorder are also introduced as measures that seem useful to cluster and to analyze the respondents' behaviour in either full and partial ranking. An application to the Diaconis' data-set is shown.*

Keywords: *Ranking, Ordering, Inversions, Cayley's distance, Disorder.*

1. Introduzione

Nella rilevazione delle preferenze si chiede spesso al rispondente di assegnare ranghi, cioè punteggi da 1 a *k* ad altrettanti oggetti o stimoli, convenendo che il numero 1 corrisponda all'oggetto tra tutti preferito, il 2 a quello preferito subito dopo e così via. Assumendo che non sia consentito di assegnare lo stesso punteggio ad oggetti distinti, cioè che non siano ammessi *ties*, il profilo dell'individuo *i* risulta definito dal vettore $\boldsymbol{\pi}_i = [\pi_{i1}\ \pi_{i2} \ldots \pi_{ik}]$, elemento dell'insieme Π_k delle permutazioni dei primi *k* numeri naturali.

Alla matrice dei dati $\mathbf{X}_{(nxk)}$, ottenuta dalla rilevazione di una popolazione *I* di dimensione *n*, si possono associare le distribuzioni di frequenze assolute e

[§] Dipartimento Metodi Quantitativi – Università degli Studi di Brescia – Cda Santa Chiara, 50, 25122 BRESCIA (e-mail: mazzali@eco.unibs.it).

relative dei vettori $\boldsymbol{\pi}_i$ che rappresentiamo nell'ordine con le coppie (Π_k, **n**) e (Π_k, **p**), essendo $\mathbf{n}_{(k!x1)}$ e $\mathbf{p}_{(k!x1)}$ i vettori delle frequenze assolute e relative.

I ranghi registrati nel contesto descritto sono dati ordinali e la loro analisi statistica può essere sviluppata seguendo impostazioni diverse (Agresti, 1984; Marden, 1995; Fligner *et al.*, 1993; Johnson *et. al.*, 1999; D'Elia A., 2001); in ogni caso, gioca un ruolo fondamentale la definizione della distanza tra vettori di ranghi.

Nel presente lavoro la distanza di Cayley e la sua generalizzazione che introdurremo - entrambe le distanze risultando totalmente coerenti con la natura ordinale dei dati in oggetto - saranno utilizzate sia per dedurre indici di posizione e misure di disordine, sia a fini della classificazione ed interpretazione del comportamento dei rispondenti.

La distanza di Cayley è introdotta nel paragrafo 2 seguendo un percorso formalizzato che ci consente di evidenziare i legami tra i concetti di assegnazione dei ranghi, ordinamento degli oggetti e permutazioni. Nel paragrafo 3 segue la proposta della generalizzazione di cui sopra. Indici di posizione, e misure di dispersione (o disordine medio) sono gli argomenti dei paragrafi 4 e 5, mentre l'utilizzo del concetto di disordine in ambito classificatorio è illustrato nel paragrafo 6 anche con un'applicazione al *data set* riportato in Diaconis P. (1989). Le conclusioni completano il lavoro.

2. La distanza di Cayley

La distanza di Cayley, definita come: *"numero minimo di scambi necessari per portare il vettore* $\boldsymbol{\pi}_i$ *a coincidere con* $\boldsymbol{\pi}_j$*"* (Marden,1995), in questo paragrafo sarà ricavata seguendo un percorso originale che ha il pregio di illustrare come, combinando una coppia di funzione di assegnazione di ranghi, si pervenga ad una sostituzione (permutazione) che sintetizza le relazioni tra le funzioni componenti.

Il problema dell'attribuzione dei ranghi 1, 2, .. *k* ad altrettanti oggetti O_1, O_2, ... O_k può essere tradotto in termini formali affermando che l'individuo $i \in I$ definisce una funzione bijettiva di assegnazione α_i: $\boldsymbol{O} \rightarrow \mathbf{N}_k$ il cui dominio è l'insieme degli oggetti $\boldsymbol{O}$ ed il codominio l'insieme $\mathbf{N}_k$ dei primi numeri *k* naturali. Le funzioni α_i ammettono inversa bilatera unica e possono essere convenientemente espresse nella forma canonica la quale prevede che gli elementi del dominio abbiano l'ordine fisso: O_1, O_2, ... O_k. Si perviene così alla rappresentazione unica per α_i :

$$\begin{pmatrix} O_1 & O_2 & .. & . & O_k \\ \alpha_i(O_1) & \alpha_i(O_2) & .. & . & \alpha_i(O_k) \end{pmatrix} \quad (1)$$

ove $\alpha_i(O_r)$ altro non è che il rango π_{ir} che il soggetto *i* associa allo stimolo O_r, conseguentemente, $\alpha_i(\boldsymbol{O}) = \boldsymbol{\pi}_i$.

La funzione inversa α_i^{-1} nella sua rappresentazione canonica descrive l'ordinamento degli oggetti in quanto comporta che questi siano disposti secondo la sequenza ordinata 1, 2, ... *k*:

$$\begin{pmatrix} 1 & 2 & .. & . & k \\ \alpha_i^{-1}(1) & \alpha_i^{-1}(2) & .. & . & \alpha_i^{-1}(k) \end{pmatrix} \quad (2)$$

essendo $\alpha_i^{-1}(r)$ uno degli oggetti inclusi in $\boldsymbol{O}$.

Ad esempio, alla funzione di assegnazione:
$\alpha_i = \begin{pmatrix} O_1 & O_2 & O_3 & O_4 & O_5 \\ 5 & 2 & 1 & 3 & 4 \end{pmatrix}$ corrisponde l'inversa canonica:

$\alpha_i^{-1} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ O_3 & O_2 & O_4 & O_5 & O_1 \end{pmatrix}$, da cui l'ordinamento degli oggetti:
$[O_3\ O_2\ O_4\ O_5\ O_1]$.

Uno scambio (o inversione) nella funzione di assegnazione si manifesta quando, considerando due oggetti O_r ed O_m, con $r<m$, risulta: $\alpha_i(O_r) > \alpha_i(O_m)$, mentre per l'ordinamento si ha che, essendo al numeratore $j < h$, ($j, h \in \mathbf{N}_k$), risulta $\alpha_i^{-1}(j) = O_m$, $\alpha_i^{-1}(h) = O_r$ e $m>r$. Ovviamente, la presenza di uno o più scambi nella funzione di assegnazione implica una presenza corrispondente nella funzione inversa.

Assegnazione dei ranghi e ordinamento degli oggetti appaiono, dunque, come due facce della stessa medaglia, la loro connessione essendo evidenziata dal fatto di conservare, *mutatis mutandis*, le medesime inversioni.

Si considerino ora due individui *i* e *j* con le relative funzioni di assegnazione α_i ed α_j. La funzione bijettiva: $\sigma_{ij}: \mathbf{N}_k \rightarrow \mathbf{N}_k$, risultante dalla composizione: $\sigma_{ij} = \alpha_j \circ \alpha_i^{-1}$ è una sostituzione (o permutazione) su $\mathbf{N}_k$, cui corrisponde la rappresentazione canonica:

$$\begin{pmatrix} 1 & 2 & .. & . & k \\ \sigma_{ij}(1) & \sigma_{ij}(2) & .. & . & \sigma_{ij}(k) \end{pmatrix} \qquad (3)$$

ove, ancora una volta la sequenza dei numeri $\sigma_{ij}(r)$, r =1, 2, … k, forma una permutazione su $\mathbf{N}_k$ (Barlotti *et al.*, 1975).

Ad esempio, i due vettori di ranghi: $\boldsymbol{\pi}_i$ =[3 1 4 2 5] e $\boldsymbol{\pi}_j$ =[1 3 5 4 2] sono le immagini delle funzioni di assegnazione:

$\alpha_i = \begin{pmatrix} O_1 & O_2 & O_3 & O_4 & O_5 \\ 3 & 1 & 4 & 2 & 5 \end{pmatrix}$; $\alpha_j = \begin{pmatrix} O_1 & O_2 & O_3 & O_4 & O_5 \\ 1 & 3 & 5 & 4 & 2 \end{pmatrix}$ e le loro composizioni generano le sostituzioni:

$$\sigma_{ij} = \alpha_j \circ \alpha_i^{-1} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 4 & 1 & 5 & 2 \end{pmatrix}; \ \sigma_{ji} = \alpha_i \circ \alpha_j^{-1} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 5 & 1 & 2 & 4 \end{pmatrix}.$$

Una qualunque sostituzione su $\mathbf{N}_k$, in quanto combinazione di due assegnazioni, esprime un legame tra i profili espressi dai due individui. Il numero degli scambi rilevabili in una sostituzione è sempre una sintesi di quelli presenti nelle assegnazioni componenti, vale infatti la proposizione:

PROPOSIZIONE 1

Se le assegnazioni α_i ed α_j comportano $c+q$ e $d+q$ scambi rispettivamente e tra questi q sono in comune, nella composizione (3) *questi ultimi si elidono per cui il numero totale di scambi nella sostituzione risultante sarà $c+d$. Tale numero, inoltre, è invariante rispetto alle composizioni alternative: $\sigma_{ij} = \alpha_j \circ \alpha_i^{-1}$ e $\sigma_{ji} = \alpha_i \circ \alpha_j^{-1}$.*

DIMOSTRAZIONE:

E' sufficiente considerare il caso in cui q=1. Siano allora $s_1, s_2 \in \mathbf{N}_k$ e O_{s1}, $O_{s2} \in \boldsymbol{O}$. Per le proprietà dell'inversa, l'affermazione è senz'altro vera quando $i=j$.

Nel caso $i \neq j$, lo scambio in α_j comporta che $O_{s1} \rightarrow s_2$, e $O_{s2} \rightarrow s_1$, mentre per α_i^{-1} si ha: $s_1 \rightarrow O_{s2}$ e $s_2 \rightarrow O_{s1}$. La composizione induce la sequenza: $s_1 \rightarrow O_{s2} \rightarrow s_1$; $s_2 \rightarrow O_{s1} \rightarrow s_2$, per cui nella sostituzione corrispondente lo scambio in questione risulta eliminato.

Ripercorrendo il ragionamento precedente con riferimento a σ_{ji} si trovano sempre gli stessi scambi per cui vale anche la seconda affermazione ◆

Se τ è la funzione che conta il numero di scambi presenti in una sostituzione, si può concludere che vale sempre la relazione: $\tau(\sigma_{ij}) = \tau(\sigma_{ji})$.

Con riferimento all'esempio precedente, si ha: $\tau(\sigma_{ij}) = \tau(\sigma_{ji}) = 5$.

PROPOSIZIONE 2

Siano $\boldsymbol{\pi}_i$ *e* $\boldsymbol{\pi}_j$ *i profili di due individui generati dalle assegnazioni* α_i *ed* α_j*, allora la funzione* d_C *definita sul prodotto cartesiano* $\Pi_k \times \Pi_k$ *che per ogni coppia di vettori di ranghi* $(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j)$ *assume il valore* $\tau(\sigma_{ij})$ *è una distanza, cioè:*

$$d_C : \Pi_k \times \Pi_k \rightarrow \mathbf{N};$$
$$d_C(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j) = \tau(\sigma_{ij})$$

DIMOSTRAZIONE:

Si tratta di dimostrare che d_C verifica tutte le condizioni di una distanza, e cioè:

1. è non negativa: $d_C(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j) \geq 0$ con $d_C(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j) = 0$ se e solo se $\boldsymbol{\pi}_i = \boldsymbol{\pi}_j$ per ogni *i* e *j*;
2. è simmetrica: $d_C(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j) = d_C(\boldsymbol{\pi}_j, \boldsymbol{\pi}_i)$, per ogni *i* e *j*;
3. soddisfa la disuguaglianza triangolare:
$d_C(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j) \leq d_C(\boldsymbol{\pi}_i, \boldsymbol{\pi}_u) + d_C(\boldsymbol{\pi}_u, \boldsymbol{\pi}_j)$, per ogni *i, j, u*.

La distanza d_C essendo definita come numero di scambi presenti nella sostituzione σ_{ij} è certamente un numero non negativo. Se due individui hanno lo stesso profilo, cioè $\boldsymbol{\pi}_i = \boldsymbol{\pi}_j$, la composizione $\alpha_j \circ \alpha_i^{-1}$ genera la sostituzione identica che è l'unica con zero scambi. D'altra parte se la composizione $\alpha_j \circ \alpha_i^{-1}$ genera la sostituzione identica allora, per quanto visto in precedenza, le immagini $\boldsymbol{\pi}_i$ di α_i e $\boldsymbol{\pi}_j$ di α_j devono coincidere.

Poiché, poi, la funzione τ che conta il numero di scambi presenti nella sostituzione σ_{ij} è invariante rispetto alle operazioni: $\alpha_j \circ \alpha_i^{-1}$ e $\alpha_i \circ \alpha_j^{-1}$, risulta verificata la simmetria per d_C.

Da ultimo, l'asserto della proposizione 1 ci permette di dimostrare che vale la proprietà triangolare. Preliminarmente osserviamo che denotando con $\overline{\alpha}$

l'assegnazione (unica) senza scambi, l'affermazione della proposizione 1 può essere posta nella forma:

$$\tau(\alpha_j \circ \alpha_i^{-1}) \leq \tau(\alpha_i \circ \overline{\alpha}^{-1}) + \tau(\alpha_j \circ \overline{\alpha}^{-1}) \qquad (*)$$

che si legge: *il numero di scambi risultanti dalla composizione* $\alpha_j \circ \alpha_i^{-1}$ *è sempre minore o uguale alla somma del numero di scambi delle assegnazioni componenti*. Inoltre, tenendo presente che la disuguaglianza triangolare di cui sopra in termini della funzione τ si può scrivere:

$$\tau(\alpha_i \circ \alpha_j^{-1}) \leq \tau(\alpha_i \circ \alpha_u^{-1}) + \tau(\alpha_u \circ \alpha_j^{-1}) \qquad (**)$$

consideriamo le due seguenti situazioni rilevanti:

a. *le tre assegnazioni* α_i, α_j *ed* α_u *non hanno scambi in comune;*
b. *ci sono scambi in comune sia tra* α_i *e* α_j*, sia tra* α_u *e* α_i *e tra* α_u *e* α_j.

Nel caso *a*, applicando lo sviluppo (*) ai due membri della disuguaglianza triangolare (**), troviamo:

$$\tau(\alpha_i \circ \overline{\alpha}^{-1}) + \tau(\alpha_j \circ \overline{\alpha}^{-1}) \leq \tau(\alpha_i \circ \overline{\alpha}^{-1}) + 2\tau(\alpha_u \circ \overline{\alpha}^{-1}) + \tau(\alpha_j \circ \overline{\alpha}^{-1})$$

che ci assicura la validità della disuguaglianza triangolare per d_C, dato che $2\tau(\alpha_u \circ \overline{\alpha}^{-1}) \geq 0$.

Nel caso *b*, siano $c+q$ e $d+q$ gli scambi presenti in α_i ed α_j, essendo q il numero di quelli in comune, siano poi $r + v$ gli scambi presenti in α_u ove degli r, r_1 sono in comune con α_i ed r_2 in comune con α_j mentre v sono comuni a tutte e tre le assegnazioni; ovviamente, deve essere $v \leq q$.
Sostituendo nella (**) le posizioni fatte, si ricava:

$(c+d) \leq (c+q - r_1 + r_2 - v) + (d+q + r_1 - r_2 - v) =$

$= (c+d) + ((r_2 - r_1) + (q-v)) + ((r_1 - r_2) + (q-v)) = (c+d) + 2(q-v)$

ma $v \leq q$ per cui la disuguaglianza è vera, valendo il segno di uguale solo se $q = v$.
$d_C(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j)$ essendo identificata con $\tau(\alpha_j \circ \alpha_i^{-1})$, verifica tutte le proprietà previste e, pertanto, è una distanza ◆

Riprendendo l'esempio precedente, si trova:

$$d_C\,([3\ 1\ 4\ 2\ 5],[1\ 3\ 5\ 4\ 2]) = \tau\left(\begin{pmatrix}1 & 2 & 3 & 4 & 5\\ 3 & 4 & 1 & 5 & 2\end{pmatrix}\right) = 5.$$

La distanza in esame è classificata tra le misure di disordine, in contrapposto alle distanze spaziali e gode di alcune proprietà, tra cui la *bi-invarianza*, che appare di rilievo nel contesto di un'analisi coerente dei ranghi, considerati come dati ordinali. La *bi-invarianza* si riferisce all'invarianza sia rispetto allo scambio degli oggetti (*label invariance*), sia rispetto ad una permutazione dei ranghi (*rank invariance*), applicata alla coppia di vettori considerati (Marden, 1995).

3. Una generalizzazione della distanza di Cayley

La distanza di Cayley cresce linearmente con il numero di scambi che intercorrono tra i profili di due individui, assume il valore zero quando i due hanno lo stesso profilo ed aumenta fino al suo massimo: $d_{C\,\max} = k(k\text{-}1)/2$.
In molte circostanze può essere utile considerare una distanza che valuti in modo non lineare il numero di scambi di cui sopra.

La natura ordinale dei nostri dati ci assicura che se d_C è una distanza, tale è anche qualunque sua trasformazione tramite una funzione f monotona non decrescente con $f(0)=0$. La classe di funzioni, tra le infinite possibili, che proponiamo è quella esponenziale che ci porta a definire la distanza generalizzata:

$$d_c^{\lambda} = d_C \exp(\lambda\, d_C) \qquad (4)$$

ove λ è un parametro di amplificazione non negativo che può essere tarato in relazione alle esigenze del ricercatore.
La nuova distanza, a motivo della monotonicità della trasformazione con la quale è stata ricavata, conserva le proprietà di *bi-invarianza* ma, può essere più sensibile al numero di scambi presenti nei vettori dei ranghi considerati.

Per $\lambda=0$ si ritrova la distanza precedente mentre per valori superiori d_c^{λ} cresce più che proporzionalmente al numero di scambi, ciò consente di incrementare l'eterogeneità tra profili diversi con conseguenze sui risultati delle elaborazioni. Nel seguito useremo il simbolo unificato d_c^{λ} per denotare gli elementi della famiglia, essendo sottinteso che d_c^{0} è la distanza di Cayley.

La tabella 1 mostra l'andamento delle distanze introdotte per alcuni valori di λ, limitatamente al caso *k*=5. Osserviamo che alle 120 possibili diverse sequenze di ranghi corrispondono solo 11 valori diversi delle distanze con le frequenze riportate nell'ultima colonna.

Tabella 1. *Andamento delle distanza di Cayley generalizzata per alcuni valori di* λ *nel caso k=5.*

d_c^0	$d_c^{.01}$	$d_c^{.05}$	$d_c^{.1}$	$d_c^{.5}$	n_i
0	0	0	0	0	1
1	1.01	1.05	1.11	1.65	4
2	2.04	2.21	2.44	5.44	9
3	3.09	3.49	4.05	13.44	15
4	4.16	4.89	5.97	29.56	20
5	5.26	6.42	8.24	60.91	22
6	6.37	8.10	10.93	120.51	20
7	7.51	9.93	14.10	231.81	15
8	8.67	11.93	17.80	436.78	9
9	9.85	14.11	22.14	810.15	4
10	11.05	16.49	27.18	1484.13	1

4. Indici di posizione

Un vettore di posizione o centro per la popolazione di riferimento può essere introdotto in molti modi. Così, il vettore di ranghi $\boldsymbol{\pi}^0$ cui corrisponde, quando esiste, la frequenza più elevata può essere un valido candidato. Tuttavia, nell'ambito delle metriche considerate, la via naturale per definire un indice di posizione è quella di identificare il profilo $\overline{\boldsymbol{\pi}}_\lambda$ cui corrisponde la distanza media minima. Introducendo il simbolo unificato $\mathbf{d}_{\overline{\pi}}^{\lambda} = [d_c^{\lambda}(\boldsymbol{\pi}_s, \overline{\boldsymbol{\pi}}_\lambda)]$, *s*=1, 2, ...*k!* per il vettore delle distanze rispetto al profilo incognito $\overline{\boldsymbol{\pi}}_\lambda$ ed impiegando il vettore **p** delle frequenze relative, si può scrivere:

$$\overline{\boldsymbol{\pi}}_\lambda = arg(min\ (\mathbf{d}_{\overline{\pi}}^{\lambda}\text{'}\mathbf{p})) \qquad (5)$$

Osserviamo che nel caso di popolazione uniforme, non esiste un vettore di posizione nel senso che ognuno dei *k!* vettori di ranghi può essere assunto come centro, mentre, se la popolazione non è uniforme la (5) può essere soddisfatta da uno solo o da una pluralità di vettori.
Sempre nel caso di popolazione non uniforme, poi, per la distanza generalizzata è possibile individuare un valore minimo del parametro λ,

cioè λ_m tale che per ogni $\lambda > \lambda_m$ il vettore di posizione resta invariato. Ciò accade quando i valori del parametro di amplificazione sono sufficientemente elevati per cui sulla (5) l'influenza delle distanze è preponderante.

Dal punto di vista del calcolo, si tratta di porre iterativamente $\overline{\boldsymbol{\pi}}_\lambda = \boldsymbol{\pi}_s$, ($s$ =1, 2, ... *k!*), di calcolare le distanze medie usando i pesi p_i. Il minimo delle suddette medie identifica il vettore cercato.

Usando ill *data set* della tabella 3, riportiamo nella tabella 2 i vettori di posizione in corrispondenza ad alcuni valori di λ. Il vettore che minimizza la media delle distanze di Cayley è: $\overline{\boldsymbol{\pi}}_0$=[1 5 2 4 3], in corrispondenza: *min* $E(d_c^0)$= 4.699, mentre $\overline{\boldsymbol{\pi}}_{.5}$= [1 5 4 2 3] minimizza la media della distanza generalizzata con λ=.5, cui corrisponde: *min* $E(d_c^{.5})$ = 107.48; quest'ultimo vettore, poi, coincide con il vettore invariante essendo λ_m =.46.

Tabella 2. *Vettori di posizione in funzione dei valori di* λ

λ	$\overline{\boldsymbol{\pi}}_\lambda$	*min*($\mathbf{d}_{\overline{\pi}}^{\lambda}$'**p**)
.00	1 5 2 4 3	4.69
.10	1 5 3 4 2	8.32
.20	1 5 3 4 2	15.15
.30	1 5 4 3 2	28.35
.40	1 5 4 3 2	54.47
.45	1 5 4 3 2	76.36
.46 *	1 5 4 2 3 *	81.74
.50	1 5 4 2 3	107.48
.60	1 5 4 2 3	216.92
.80	1 5 4 3 2	949.05
1.00	1 5 4 3 2	4516.22

5. Misure di disordine

La distanze introdotte sono indici di disordine elementare nel senso che misurano l'alterazione dell'ordine di un profilo $\boldsymbol{\pi}_i$ rispetto ad un altro; quest'ultimo, di solito, è un vettore di posizione come $\overline{\boldsymbol{\pi}}_\lambda$, ma potrebbe essere un qualunque altro $\boldsymbol{\pi}_j$.

Nel primo caso la distanza $d_c^\lambda(\boldsymbol{\pi}_i, \overline{\boldsymbol{\pi}}_\lambda)$ quantifica il disordine indotto dall'individuo *i*, appartenente all'insieme *I* dei rispondenti. Il disordine generato dall'intera popolazione, detto *disordine totale*, è la somma dei disordini individuali e può essere scritto nella forma matriciale seguente:

$$\Delta = \mathbf{d}_{\overline{\pi}}^{\lambda}\mathbf{'n} \tag{6}$$

Sostituendo nella (6) al vettore delle frequenze assolute quello delle frequenze relative **p**, si ottiene il *disordine medio* che è una misura della dispersione dei dati:

$$\overline{\Delta} = \mathbf{d}_{\overline{\pi}}^{\lambda}\mathbf{'p} \tag{7}$$

Il disordine associato ad un individuo *i* o ad un gruppo $I_\omega \subset I$, può anche essere espresso in termini relativi tramite i rapporti di composizione:

$$\delta_{\overline{\pi},i}^{\lambda} = \boldsymbol{d}_{\boldsymbol{c}}^{\lambda}(\boldsymbol{\pi}_i, \overline{\boldsymbol{\pi}}_\lambda)\Delta^{-1} \tag{8}$$

$$\delta_{\overline{\pi},\omega}^{\lambda} = \sum_{i\in I_\omega} d_c^{\lambda}(\boldsymbol{\pi}_i, \overline{\boldsymbol{\pi}}_\lambda)\Delta^{-1} \tag{9}$$

Gruppi particolari sono le *k!* classi di individui I_s , *s*=1,2,…*k!*, che hanno assegnato vettori di ranghi distinti e per i quali si dispone delle distribuzioni di frequenza. Introducendo la matrice diagonale $\mathbf{P} = diag(p_s)$ i cui elementi sono frequenze relative ed il vettore **1** con elementi unitari, è possibile definire un vettore che riporta i disordini relativi riferibili alle classi di cui sopra. Infatti, i contributi assoluti dei gruppi in esame sono dati dal prodotto $\mathbf{P}\mathbf{d}_{\overline{\pi}}^{\lambda}$, mentre i contributi relativi sono gli elementi del vettore $\mathbf{z} = [\delta_{\overline{\pi},s}^{\lambda}]$, (*s*=1, 2,…*k*!):

$$\mathbf{z} = \mathbf{P}\mathbf{d}_{\overline{\pi}}^{\lambda}(\mathbf{1'P}\,\mathbf{d}_{\overline{\pi}}^{\lambda})^{-1} \tag{10}$$

Quando non si desideri o non sia possibile considerare il riferimento $\overline{\boldsymbol{\pi}}_\lambda$, il disordine medio può essere introdotto come media delle distanze tra tutte le coppie dei vettori di ranghi. Se $\mathbf{D}_{(k!xk!)} = [d_c^{\lambda}(\boldsymbol{\pi}_r, \boldsymbol{\pi}_s)]$; (*s*, *r*= 1, 2, ..*k!*) è la matrice delle distanze tra coppie si trova:

$$\tilde{\Delta} = \mathbf{p'Dp} \tag{11}$$

Nel caso di popolazione uniformemente distribuita, la (11), dipendendo solo da *k*, si riduce a:

$$\overline{\Delta}_{\boldsymbol{u}} = k(k\text{-}1)/4 \qquad (12)$$

Il risultato (12) può essere impiegato per comparare la dispersione della popolazione data con quella di una ipotetica popolazione uniforme, in modo da evidenziare l'effetto di eventuali concentrazioni delle frequenze su alcuni profili rispetto ad altri.

6. Disordine e classificazione

La famiglia delle distanze considerata e le misure di disordine assumono un ruolo importante nei problemi di classificazione. La prima può essere impiegata per costruire gruppi omogenei di individui utilizzando le note tecniche della *cluster analysis* (Marden, 1995), le seconde consentono la misurazione del disordine generato dai gruppi stessi. In altri termini, il disordine, che tendenzialmente è tanto maggiore quanto più il punteggio o i punteggi assegnati risultano in contrasto con quelli di $\overline{\boldsymbol{\pi}}_{\lambda}$, può essere scomposto nei contributi delle singole classi, essendo il contributo di una classe interpretabile come il "*danno*" imputabile agli individui che formano la classe stessa per il fatto di non aver scelto il profilo di riferimento $\overline{\boldsymbol{\pi}}_{\lambda}$.

Nell'analisi descrittiva dei vettori di ranghi spesso si procede allo studio dei profili parziali dei rispondenti tramite classificazioni in base al rango $\pi_{.r}$ assegnato all'oggetto O_s; (r, s= 1, 2, ...k), oppure in base a coppie di ranghi ($\pi_{.r}, \pi_{.r'}$) assegnati alle coppie- ordinate o meno- di stimoli (O_s,$O_{s'}$); il processo può continuare considerando terne, quaterne e così via. Queste classificazioni inducono, in generale, più partizioni della popolazione in gruppi aventi numerosità diversa a ciascuno dei quali, in base a quanto detto, è sempre associabile una quota del disordine. Si possono così identificare gli individui che hanno maggior rilevanza nella spiegazione del disordine totale.

6.1 Applicazione

I concetti esposti sono stati applicati al *data set* riportato in Diaconis P. (1989) relativo a ranghi (completi) assegnati da parte di n = 5738 votanti a 5 candidati nell'elezione del presidente del *American Psychological Association* (APA) nel 1980.

La tabella 3 riproduce le frequenze relative p_s dei 120 profili ed i valori degli associati disordini relativi: $\delta^{0}_{\overline{\pi},s}$ e $\delta^{.5}_{\overline{\pi},s}$ calcolati rispetto ai vettori di posizione noti: $\overline{\boldsymbol{\pi}}_0$ =[1 5 2 4 3] e $\overline{\boldsymbol{\pi}}_{.5}$ = [1 5 4 2 3].

In primo luogo consideriamo la classificazione che individua i votanti che hanno assegnato il rango $\pi_{.r}$ al candidato O_s; (r,s= 1, 2, ... 5). Scaturiscono 10 diverse partizioni della popolazione dei votanti con associate scomposizioni dei disordini relativi $\delta^{0}_{\bar{\pi},\mathbf{rs}}$ e $\delta^{.5}_{\bar{\pi},\mathbf{rs}}$. Tutti questi elementi sono sistemati nella tabella a doppia entrata 4 in cui le righe sono intestate ai punteggi, le colonne ai candidati.

Le celle all'incrocio di ogni riga e colonna, riportano: la percentuale dei votanti che ha assegnato il rango $\pi_{.r}$ ad O_s (frequenze relative: $100p_{rs}$), il disordine relativo $100\,\delta^{0}_{\bar{\pi},\mathbf{rs}}$, il disordine relativo $100\,\delta^{.5}_{\bar{\pi},\mathbf{rs}}$. Evidentemente, lungo ogni riga e ogni colonna si realizza una partizione della popolazione per cui le relative somme delle percentuali ammontano a 100.

Mentre le frequenze relative nel senso delle righe mostrano come ogni punteggio è stato distribuito sui diversi candidati e nel senso delle colonne come ogni candidato ha ricevuto i diversi punteggi, i disordini relativi in entrambi i sensi di lettura misurano i contributi percentuali al disordine totale dei votanti inclusi in ciascuna cella.

Dalla prima riga emerge così che il candidato O_3 è quello che ha ottenuto la percentuale massima di prime posizioni 28.04% da parte di 1609 votanti, responsabili del disordine relativo massimo rispetto a $\delta^{.5}_{\bar{\pi},\mathbf{rs}}$ ma non rispetto a $\delta^{0}_{\bar{\pi},rs}$.

I valori minimi simultanei dei disordini relativi sono generati dai 1053 (18.35%) votanti che hanno assegnato il primo posto ad O_1. Lungo la prima colonna si legge che il candidato O_1 ha ricevuto la percentuale maggiore di seconde posizioni, mentre il disordine relativo maggiore è imputabile ai votanti che hanno dato ad O_1 il terzo posto nella metrica di Cayley, il quarto in quella generalizzata.

La seconda classificazione identifica i votanti in relazione alle coppie ordinate di ranghi (($\pi_{.r}, \pi_{.r'}$); $r \neq r'$) assegnate alle coppie non ordinate di candidati (($O_s, O_{s'}$); $s<s'$). Questa classificazione genera ben 20 partizioni della popolazione; le prime 10 risultano immediatamente in corrispondenza delle coppie (O_s, $O_{s'}$), le altre si ottengono aggregando i votanti che hanno assegnato le coppie di ranghi: ($\pi_{.r}, \pi_{.r'}$) e ($\pi_{.r'}\ \pi_{.r}$) (tabella 5).

Tabella 3. *Vettori dei ranghi, frequenze relative 100* p_s *e disordini relativi 100* $\delta^{0}_{\bar{\pi},s}$ *e 100* $\delta^{.5}_{\bar{\pi},s}$ *.*

$\boldsymbol{\pi}_s$	p_s (x100)	$\delta^{0}_{\bar{\pi},s}$ (x100)	$\delta^{.5}_{\bar{\pi},s}$ (x100)	$\boldsymbol{\pi}_s$	p_s (x100)	$\delta^{0}_{\bar{\pi},s}$ (x100)	$\delta^{.5}_{\bar{\pi},s}$ (x100)	$\boldsymbol{\pi}_s$	p_s (x100)	$\delta^{0}_{\bar{\pi},s}$ (x100)	$\delta^{.5}_{\bar{\pi},s}$ (x100)
12345	.52	.45	.30	24513	.92	.98	.05	42315	.89	1.32	1.00
12354	.49	.31	.55	24531	1.10	.93	.14	42351	.42	.53	.90
12435	.47	.50	.13	25134	1.38	.59	.17	42513	1.15	1.96	.65
12453	.51	.43	.29	25143	1.85	.39	.51	42531	1.01	1.51	1.13
12534	.61	.78	.08	25314	.37	.23	.02	43125	.61	.65	.68
12543	.59	.63	.16	25341	.70	.30	.09	43152	.66	.56	1.43
13245	1.78	1.14	.49	25413	.59	.50	.01	43215	.61	.78	.35
13254	1.66	.70	.94	25431	.61	.39	.03	43251	.52	.56	.59
13425	.61	.52	.08	31245	.59	.76	1.28	43512	1.46	2.18	.40
13452	.64	.41	.18	31254	.52	.56	2.12	43521	1.59	2.02	.90
13524	.49	.52	.02	31425	.73	1.09	.82	45123	.52	.45	.30
13542	.61	.52	.08	31452	.70	.89	1.50	45132	.66	.42	.74
14235	.78	.33	.10	31524	.59	1.01	.34	45213	.42	.45	.12
14253	1.22	.26	.34	31542	.52	.78	.59	45231	.59	.50	.34
14325	.42	.27	.02	32145	1.29	1.37	1.45	45312	.94	1.20	.12
14352	.89	.38	.11	32154	1.43	1.22	3.08	45321	.54	.57	.15
14523	.84	.71	.01	32415	1.31	1.67	.74	51234	.51	.86	3.81
14532	.91	.58	.05	32451	.59	.63	.66	51243	.19	.29	2.65
15234	.87	.19	.04	32514	1.12	1.66	.31	51324	.33	.63	1.35
15243	1.22	.00	.15	32541	.71	.91	.40	51342	.44	.74	3.28
15324	.30	.13	.00	34125	.61	.52	.35	51423	.80	1.71	1.73
15342	.63	.13	.03	34152	1.52	.97	1.70	51432	.87	1.67	3.54
15423	.61	.39	.00	34215	.49	.52	.13	52134	.87	1.30	3.54
15432	.70	.30	.01	34251	1.08	.92	.61	52143	.61	.78	4.60
21345	.70	.74	.78	34512	2.32	2.96	.29	52314	.42	.71	.90
21354	.52	.45	1.13	34521	1.86	1.98	.51	52341	.45	.68	1.84
21435	.45	.58	.26	35124	.63	.40	.17	52413	.77	1.47	.86
21453	.42	.45	.47	35142	.78	.33	.44	52431	.94	1.60	2.03
21534	.73	1.09	.20	35214	.47	.40	.06	53124	.45	.58	.98
21543	.63	.80	.36	35241	.71	.46	.20	53142	.59	.63	2.41
23145	3.00	2.55	1.70	35412	1.06	1.13	.05	53214	.38	.57	.43
23154	3.24	2.07	3.63	35421	1.24	1.05	.15	53241	.38	.49	.83
23415	.91	.96	.25	41235	.28	.41	1.13	53412	.85	1.45	.48
23451	.92	.79	.52	41253	.38	.49	2.89	53421	.99	1.48	1.11
23514	.91	1.16	.11	41325	.40	.68	.86	54123	.49	.52	.55
23541	.78	.83	.22	41352	.54	.81	2.20	54132	.75	.64	1.62
24135	1.67	1.07	.46	41523	.78	1.50	.88	54213	.42	.53	.24
24153	2.82	1.20	1.60	41532	.87	1.48	1.88	54231	.64	.69	.72
24315	.49	.41	.06	42135	.70	.89	1.50	54312	1.17	1.74	.32
24351	.77	.49	.21	42153	.91	.96	3.68	54321	.51	.65	.29

Tabella 4. *Classificazione dei votanti in relazione ai punteggi assegnati ai candidati: frequenze relative* $100p_{rs}$ *e disordini relativi* $100\,\delta^{0}_{\overline{\pi},rs}$ *e* $100\,\delta^{.5}_{\overline{\pi},rs}$

	O_1	**O_2**	**O_3**	**O_4**	**O_5**
1	18.35 *10.56* *4.15*	13.51 *20.46* *36.04*	28.04 *21.06* *37.29*	20.43 *26.96* *7.95*	19.68 *20.96* *14.58*
2	26.47 *19.96* *12.97*	18.77 *24.27* *30.79*	16.73 *12.84* *20.54*	16.94 *20.03* *12.25*	21.09 *22.91* *23.45*
3	22.88 *24.18* *17.97*	24.66 *24.13* *18.80*	13.82 *14.55* *16.51*	18.98 *19.29* *23.65*	19.66 *17.84* *23.07*
4	17.46 *22.90* *24.81*	24.68 *20.25* *10.47*	18.30 *22.06* *15.91*	20.29 *16.69* *24.82*	19.28 *18.10* *23.99*
5	14.83 *22.39* *40.10*	18.39 *10.90* *3.91*	23.11 *29.49* *15.91*	23.37 *17.03* *31.33*	20.30 *20.18* *14.91*

Anche qui i risultati sono stati organizzati in una tabella a doppia entrata, all'interno di ogni cella sono riportati nell'ordine: la dimensione relativa del gruppo (frequenza relativa), il valore del disordine relativo $100\,\delta^{0}_{\overline{\pi},rr'ss'}$ e, infine, $100\,\delta^{.5}_{\overline{\pi},rr'ss'}$.

Le frequenze relative nella tabella 5 descrivono come i votanti si sono ripartiti nell'assegnare le coppie ordinate di punteggi a ciascuna delle 10 coppie di candidati. Non diversamente dal caso precedente, l'interesse può riguardare le combinazioni in cui si sono avuti le maggiori o minori adesioni (frequenze), i valori maggiori o minori dei disordini relativi. Ad esempio, nella prima colonna 9.76% (560) votanti hanno assegnato i punteggi (2,3) ai candidati (O_1,O_2), generando disordini relativi $\delta^{0}_{\overline{\pi},23,12'}$=8.36% e $\delta^{.5}_{\overline{\pi},23,12'}$=6.44%. Per facilitare la lettura nella tabella 5 nel senso delle colonne, abbiamo usato il carattere neretto per denotare i massimi, il sottolineato per i minimi.

Osserviamo che il disordine relativo $\delta^{.5}_{\overline{\pi},rr'ss'}$ rispetto a $\delta^{0}_{\overline{\pi},rr'ss'}$, oltre ad assumere valori che esaltano in modo più netto i contrasti tra i diversi gruppi, raggiunge valori massimi e minimi in corrispondenza a gruppi diversi di votanti in quanto i valori di riferimento delle due metriche sono diversi.

Tabella 5. *Frequenze relative* $100p_{rr',ss'}$ *e valori dei disordini relativi* $100\,\delta^{0}_{\bar{\pi},rr'ss'}$ e $100\,\delta^{.5}_{\bar{\pi},rr'ss'}$ *per coppie di ranghi assegnati a coppie ordinate di candidati*

	O_1O_2	O_1O_3	O_1O_4	O_1O_5	O_2O_3	O_2O_4	O_2O_5	O_3O_4	O_3O_5	O_4O_5
1 2	3.19 3.10 1.50	7.53 2.62 .14	3.26 2.53 2.06	4.37 2.32 .45	2.47 3.36 5.98	3.64 6.62 13.89	3.94 6.37 **12.99**	3.31 3.11 1.32	4.97 3.56 1.67	7.81 **10.66** 8.34
1 3	5.79 3.81 1.78	3.24 1.66 .40	4.34 2.67 1.01	4.98 2.42 .95	2.93 4.05 **10.82**	3.71 6.10 9.60	3.21 5.23 8.97	6.03 4.90 2.42	7.20 4.30 1.92	4.27 5.89 11.23
1 4	5.05 2.53 .63	3.54 2.55 1.21	5.35 2.86 .68	4.41 2.63 1.64	3.97 6.38 8.93	3.07 4.11 8.32	3.21 4.59 8.94	8.12 6.06 2.40	**8.00** 6.15 1.83	3.66 4.74 **11.58**
1 5	4.32 1.13 .24	4.04 3.74 2.40	5.40 2.50 .40	4.58 3.20 1.11	4.13 6.66 10.31	3.08 3.63 4.24	3.15 4.26 5.14	**10.58** 6.98 1.81	7.88 7.05 2.53	4.69 5.67 6.14
2 1	3.45 4.10 3.19	**13.96** 7.87 .50	4.18 4.26 8.07.	4.88 3.73 1.21	5.80 6.52 4.46	5.65 **8.79** **17.85**	4.13 5.86 6.98	2.79 3.25 3.02	3.94 3.61 3.12	6.73 7.76 3.28
2 3	**9.76** **8.36** 6.44	3.54 2.62 1.26	5.94 4.65 2.29	**7.23** 4.33 2.99	3.19 4.00 8.41.	4.60 6.57 5.49	4.53 6.23 10.24	3.68 2.98 2.67	3.85 2.01 3.47	4.04 5.27 6.38
2 4	7.77 5.09 2.52	3.90 3.67 3.65	7.65 5.62 1.54	7.15 5.58. 5.27	4.58 6.30 8.75	4.18 4.81. 4.71.	4.93 5.98 8.46	4.88 3.12 3.89.	4.41 3.28 2.86	2.79 3.27 7.41
2 5	5.49 2.40 .83	5.07 5.80 7.57	**8.70** 5.436 1.07	7.22 6.32 3.51	5.19 7.45 9.16	4.34 4.09 2.74	5.18 6.20 5.11	5.39 3.49 2.66	4.53 3.94 2.81	3.38 3.73 3.48
3 1	3.66 5.08 6.65	6.26 4.81 1.58	6.76 **8.34** 7.19	6.20 5.96 2.55	**8.56** 7.04. 2.03	5.12 7.11 10.83	5.19 6.17 4.16	4.27 5.63 8.04	3.38 3.22 4.39	4.90 5.62 3.48
3 2	6.45 7.46 6.65	3.87 3.61 2.34	5.66 6.06 4.41	6.90 **7.06** 4.58	5.33 4.24 3.78	4.74 5.77 3.62	4.83 5.76 4.98	2.49 2.93 6.14	4.60 5.00 7.84	4.76 5.09 6.06
3 4	7.88 7.87 3.59.	5.63 6.46 4.36	4.62 4.61 3.94	4.76 5.24 6.08	4.93 5.61 5.72	7.15 6.16 2.62	7.13 5.60 6.12	3.43 3.03 6.00	2.42 2.46 7.85	4.97 4.80 3.95
3 5	4.90 3.77 1.08	7.13 9.30 9.69	5.84 5.18 2.44	5.02 5.93 4.76	5.84 7.24 7.29	7.65 5.09. 1.73	7.51 6.60 3.54	3.62 2.96 3.47	3.42 3.88 3.58	4.36 3.79 3.02
4 1	3.26 5.38 9.84.	4.06 3.93 2.63	5.47 7.89 8.34	4.67 5.70 4.01	7.86 4.91 1.09	5.80 7.15 6.27	5.96 5.66 2.48	5.49 7.19 11.10	5.30 5.94 3.57	3.75 3.66 4.52
4 2	5.07 7.18 8.87	2.81 3.19 3.77	4.44 5.88 5.41	5.14 6.66 6.77	4.64 3.25 1.73	4.72 4.65 2.14	**7.55** **7.26** 4.08	4.98 6.24 5.59	4.83 5.85 6.83	3.57 3.14 5.77
4 3	5.45 6.75 4.35	3.73 5.12 6.73	4.11 5.22 5.23	4.17 5.80 8.52	4.23 3.93 3.08	5.86 4.24 1.01	6.71 4.21 2.78	4.04 5.04 6.32	3.69 4.94 8.42	5.09 2.89 3.35
4 5	3.68 3.59 1.76	6.87 **10.65** 11.69	3.43 3.91 5.85	3.49 4.74 5.52	7.95 **8.15** 4.57	**8.30** 4.22 1.05	4.46 3.12 1.12	3.78 3.59 1.80	4.48 5.32 5,99	**7.88** 7.00 2.27
5 1	3.14 5.90 **16.36**	3.76 4.44 3.23	4.01 6.48 **13.69**	3.92 5.58 6.82	5.82 2.58 .37	3.85 3.92 2.34	4.39 3.27 .95	7.88 **10.90** **15.13**	7.06 8.19 3.50	4.30 3.92 3.30
5 2	4.06 6.53 13.77	2.53 3.43 6.00	3.57 5.56 8.67	4.67 6.87 **11.66**	4.29 1.99 .78	3.83 2.99 .90	4.78 3.52 1.40	6.15 7.75 7.49	6.69 **8.50** 7.12	4.95 4.02 3.28
5 3	3.66 5.20 6.24.	3.31 5.15 15.26	4.58 6.75 7.98	3.28 5.29 10.62	3.47 2.57 1.34	4.81 2.38 .41	5.21 2.18 1.08	5.23 6.37 5.09	4.91 6.59 9.26	6.26 3.8 2.11
5 4	3.97 4.76 3.73	5.23 9.38 **15.61**	2.67 3.60 9.76	2.96 4.654 11.01	4.81 3.76 1.42	5.89 1.61 .26	4.01 1.93 .47	3.85 4.48 3.62	4.44 6.22 **11.45**	7.86 5.30 1.06

7. Conclusioni

Nel presente lavoro, dopo aver dimostrato che la distanza di Cayley scaturisce in modo naturale nell'ambito del processo di assegnazione dei ranghi ed aver proposto una generalizzazione, che può risultare utile in particolari circostanze, abbiamo dedotto misure di valore centrale e di dispersione, queste ultime dette anche misure di disordine. Abbiamo, poi, rilevato come ad ogni distribuzione di frequenze, che scaturisca da una classificazione della popolazione dei rispondenti, sia associabile una distribuzione del disordine relativo che fornisce informazioni sulla quantità di eterogeneità introdotta dai componenti che formano le classi stesse.

Un'applicazione in campo elettorale mostra un possibile utilizzo dei concetti considerati. L'interesse applicativo, tuttavia, è molto ampio, si pensi all'analisi delle preferenze tipiche negli studi di mercato ove il disordine indotto da determinati gruppi può essere interpretato come grado di rifiuto di determinate marche o prodotti rispetto ad altri.

Tutte le considerazioni svolte in questo contributo hanno fatto riferimento ad una popolazione e sono state trattate a livello puramente descrittivo. E' evidente, tuttavia, che le problematiche proposte sono naturalmente inseribili in un contesto campionario e che, in tale ambito, risultano fondamentali le distribuzioni delle distanze, dei valori di posizione, del disordine medio e relativo. Sotto ipotesi particolari, alcuni di questi temi sono stati trattati (Marden, 1995), altri dovrebbero essere oggetto di ricerca ulteriore.

Riferimenti bibliografici

Agresti A., (1984). *Analysis of ordinal categorical data*. J. Wiley, N. York.

Barlotti A., Guidotti L., Nicoletti G. (1975). *Sistemi lineari ed elementi di geometria analitica del piano*. Cedam, Padova.

D'Elia A., (2001). A Multivariate Model for Studying Preferences Data, *New trends in Statistical Modellling*, Klein B. and Korsholn L., Odense Denmark, p. 425-428.

D'Elia A., Piccolo D., (2004). A mixture model for preferences data analysis. *Computationnal Statistics & Data Analysis* (*In Press*).

Diaconis P.,(1989). A generalization of spectral analysis with application to ranked data, *The Annals of Statistics*, Vol. 17, n. 3, p. 949-979.

Johnson V. E., Albert J. H., (1999). *Ordinal Data Modeling*, Springer, Berlin

Fligner M. A., Verducci J. S., (1993). *Probability Models and Statistical Analyses for Ranking Data*, Springer Verlag, Berlin.

Marden J. I., (1995). *Analyzing and Modeling Rank Data*, Chapman & Hall, London.