Quaderni di Statistica
Vol.5, 2003

Una generalizzazione dei ranghi per standardizzare i dati

Stefano Maria Pagnotta
Dipartimento Persona, Mercato, Istituzioni, Università degli Studi del Sannio
E-mail: pagnotta@unisannio.it

Summary: Ranks are often used as a device to transform data in order to reach different goals such as standardization and/or to limit the effects of outlying observations on the results of statistical analysis. However, the rank transformation has the inconvenience of wiping out all the information on the distribution since the transformed data are uniformly distributed.
In order to reduce the information loss, the present paper introduces the generalized ranks. They inherit the property of standardizing data and of resistance to outliers of ordinary ranks, but preserve some distributional information of the original data. Consequently the generalized rank transformation can be viewed as an intermediate solution between the rank-transformation and the ordinary standardization, which maintains positive features of both of them.
The new transformation is applied to a set of variables used to estimate a latent variable describing life quality; in this context the generalized ranks outperform both the ordinary rank transformation and the standardization.

Keywords: Rank, Transformation, Latent variable, Qualità della vita.

1. Introduzione

In ambito statistico, sovente, le osservazioni provenienti da una variabile casuale univariata sono trasformate in ranghi per procedere all'applicazione di tecniche non parametriche, per rendere resistente alle anomalie procedure statistiche standard o, ancora, semplicemente per standardizzare le osservazioni (Conover e Iman, 1981).

In questo lavoro consideriamo i ranghi come strumento di pre-trattamento dei dati per ottenere un triplice effetto: pervenire a graduatorie delle unità statistiche cui sono associate intensità, limitare l'effetto delle anomalie, standardizzare i dati.

Faremo riferimento ad un problema reale: l'analisi del rapporto annuale sulla qualità della vita pubblicato fin dal '90 dal quotidiano *Il Sole 24 Ore*. Come è noto, la ricerca degli indicatori diretti alla misura della qualità della vita ha suscitato viva attenzione nella comunità degli statistici fin dalla sua prima edizione. La fase iniziale di trattamento e trasformazione dei dati costituisce un elemento fondamentale in tale tipo di studio, pertanto di rilievo sono i contributi diretti a migliorare tale passaggio iniziale.

Attanasio e Capursi (1997) applicando diverse tecniche di pre-trattamento dei dati, e misurandone gli effetti sulla graduatoria finale delle province, hanno implicitamente individuato come ottimale la usuale standardizzazione in termini di media e scarto quadratico medio. Una scelta simile, ma indipendente dal lavoro citato, è stata adottata da Vitali e Merlini (1999) nell'indagine parallela sulla qualità della vita pubblicata nel 1999 sul quotidiano *Italia Oggi*.

L'ampia e flessibile classe delle trasformazioni lineari consente un trattamento omogeneo degli indicatori che portano alla misura della variabile latente *Qualità della vita*. Una trasformazione lineare conserva fedelmente la forma distributiva originaria della variabile considerata. Tale peculiarità ha un risvolto negativo quando i dati presentano una forte asimmetria naturale o dovuta ad anomalie. L'effetto risultante è in tal caso che la gran parte dei valori trasformati si comprimono in uno spazio ridotto rispetto a quello disponibile e che ha per estremi i valori trasformati del minimo e del massimo. È pratica usuale, in questi casi trasformare i dati in ranghi: ciò limita l'effetto appena descritto ma, nel contempo, la nuova variabile casuale generata ha distribuzione uniforme. La standardizzazione e la trasformazione in ranghi appaiono, rispetto a tali considerazioni, antitetiche.

In questo lavoro, partendo da queste riflessioni, insieme con quelle in Attanasio e Capursi (1997) e Vitali e Merlini (1999), proponiamo un trasformazione dei dati che si colloca come mediazione fra la stan-

dardizzazione e la trasformazione in ranghi. In questa direzione, in un precedente contributo (Pagnotta, 2002) era stata proposta un'estensione della nozione di rango che portava ad una trasformazione dei dati che risolveva solo in parte le problematiche connesse all'asimmetria dei dati originari.

Nel paragrafo due si presenta un'ulteriore estensione della nozione di rango, che riesce meglio a ridurre l'impatto delle anomalie sulla forma della distribuzione ed ha il vantaggio di essere computazionalmente meno dispendiosa. Di questa nuova estensione si illustrano, poi, dettagliatamente le sue proprietà. Nel successivo paragrafo tre la nuova proposta di trasformazione è utilizzata per rielaborare i dati del dossier 1999 sulla *Qualità della vita* de *Il Sole 24 Ore* (Cadeo, 1999) e di *Italia Oggi* (Mori, 1999). Considerazioni conclusive sono nel quarto paragrafo.

2. Una proposta di trasformazione dei dati

In un recente contributo (Pagnotta, 2002) è stata proposta una generalizzazione dei ranghi al fine di pervenire ad una sequenza di interi che conservasse sia l'ordine delle osservazioni, come nel caso dei ranghi propriamente detti (cui nel seguito ci riferiremo come *ranghi ordinari*), sia una memoria dell'informazione sulla loro distribuzione originaria. Tale generalizzazione si basa sull'individuazione di una sequenza di numeri interi ρ_i che soddisfino il criterio

$$\min_{\rho_i \in N} \sum_{\substack{i,j=1 \\ i<j}}^{n} [rank(\Delta(x_i, x_j)) - rank(\Delta(\rho_i, \rho_j))]^2, \qquad (1)$$

ove $\Delta(\cdot,\cdot)$ è un qualsiasi indice di dissimilarità compatibile con la natura dei dati; $rank(\cdot)$ è una funzione che restituisce il rango del suo argomento, e quest'ultimo appartiene ad un insieme di riferimento definito in questo caso come $\{\Delta_{i,j} = \Delta(x_i, x_j),\ \ i, j = 1, 2, \ldots, n,\ \ i < j\}$.

Se indichiamo con τ_i una permutazione tale che $x_{\tau_i} \leq x_{\tau_{i+1}}$, la sequenza cercata è derivata dalla relazione

$$\hat{\rho}_{\tau_1} = 1,\ \ \hat{\rho}_{\tau_i} = \hat{\rho}_{\tau_{i-1}} + rank(\Delta(x_{\tau_i}, x_{\tau_{i-1}})),\ \ i = 2, 3, ..., n. \qquad (2)$$

I valori ottenuti dalla (2) sono stati chiamati *Ranghi generalizzati*.

Successivi approfondimenti hanno mostrato che i *ranghi generalizzati*, pur rispondendo alle richieste che hanno guidato alla loro derivazione, hanno una *'resistenza'* trascurabile ai valori anomali, rispetto a quella posseduta dai *ranghi ordinari*. Tale considerazione ci ha spinto a rivedere la definizione dei *ranghi generalizzati* a partire dal criterio che li individua. La nuova generalizzazione che nel seguito introdurremo varrà valutata sia sul piano della conservazione parziale della forma della distribuzione originaria dei dati sia rispetto alla proprietà di *resistenza* ai valori anomali.

Come miglioramento del criterio (1), proponiamo la seguente formulazione:

$$\min_{\rho_i \in N} \sum_{\substack{i,j=1 \\ i-j=1}}^{n} [rank(\Delta(x_i, x_j)) - \Delta(\rho_i, \rho_j)]^2. \tag{3}$$

Il criterio si giustifica per più aspetti: in primo luogo il confronto, utile ad individuare la soluzione, è limitato alle sole distanze fra valori consecutivi, inoltre, viene adottato, come riferimento per calcolare i relativi ranghi, l'insieme $\{\Delta_{i,j} = \Delta(x_i, x_j),\ \ i, j = 1, 2, \ldots, n,\ \ i - j = 1\}$. Infine viene rimossa la trasformazione in ranghi delle distanze fra i valori trasformati ancora incogniti. La relazione (2) consente di pervenire ad una sequenza tale che il criterio (3) si annulli. In precedenza non era possibile raggiungere per il criterio (1) tale limite inferiore se non per basse numerosità.

A titolo di esempio consideriamo la sequenza di valori pseudo-casuali, riportati nella *Tabella 1*, generati da una normale standardizzata. Senza ledere la generalità del problema, i valori sono stati preventivamente ordinati per semplificare la lettura dei passaggi. Nella seconda riga della tabella vi sono le distanze semplici fra le osservazioni, nella successiva troviamo i corrispondenti ranghi e nella riga quattro vi sono i dati trasformati utilizzando la (2).

Tabella 1. Esempio di dati trasformati secondo il criterio (3).

x_i	-1.289	-1.107	0.022	0.476	1.171	2.159
$\Delta(x_i, x_j)$		0.182	1.129	0.454	0.695	0.988
$rank(\Delta(x_i, x_j))$		1	5	2	3	4
$\hat{\rho}_i$	1	2	7	9	12	16
$\Delta(\hat{\rho}_i, \hat{\rho}_j)$		1	5	2	3	4

Dall'ultima riga si evince come $\Delta(\hat{\rho}_i, \hat{\rho}_j)$ riproduca $rank(\Delta(x_i, x_j))$, cosicchè il criterio (3) vale zero.

I valori interi generati da (3) e (2) conservano l'ordine delle osservazioni originali. Il valore massimo che possono assumere dipende unicamente della numerosità n dei dati, essendo

$$max_i \ \hat{\rho}_i = 1 + \frac{n(n-1)}{2};$$

mentre la media aritmetica e la varianza non dipendono esplicitamente da n, come accade invece nel caso dei *ranghi ordinari*. I valori trasformati sono invarianti rispetto a preliminari trasformazioni affini dei dati originari.

Il requisito che la trasformazione proposta conservi parte dell'informazione concernente la distribuzione dei dati di origine si formalizza nello studio delle differenze nella forma della funzione di ripartizione empirica prima e dopo la trasformazione. Abbiamo assunto come distribuzione di riferimento la normale standardizzata da cui abbiamo derivato un insieme di dati Z corrispondenti a $n = 91$ percentili $z_p = \Phi^{-1}(p)$ equispaziati fra $p = 0.05$ e $p = 0.95$. Nel seguito tali valori sono indicizzati con $i = 1, 2, \ldots, 91$. Indichiamo con $\hat{\Phi}(x)$ la funzione di ripartizione empirica calcolata sui valori z_i.

Siano le $r_i^{(0)} = R^{(0)}(z_i)$ i *ranghi ordinari* delle z_i e le $r_i^{(1)} = R^{(1)}(z_i)$ i valori interi ottenuti dalla trasformazione delle z_i seguendo le (3) e (2), e con ovvia associazione consideriamo le funzioni di ripartizione empirica $\hat{F}^{(0)}(x)$ e $\hat{F}^{(1)}(x)$. Nella Figura 1 è riportato il grafico delle funzioni di ripartizione $\hat{\Phi}(x)$, $\hat{F}^{(0)}(\frac{x-\bar{r}^{(0)}}{S^{(0)}})$ e $\hat{F}^{(1)}(\frac{x-\bar{r}^{(1)}}{S^{(1)}})$, essendo $S^{(j)}$ lo scarto quadratico medio e $\bar{r}^{(j)}$ la media aritmetica delle $r_i^{(j)}$ per $j = 0, 1$.

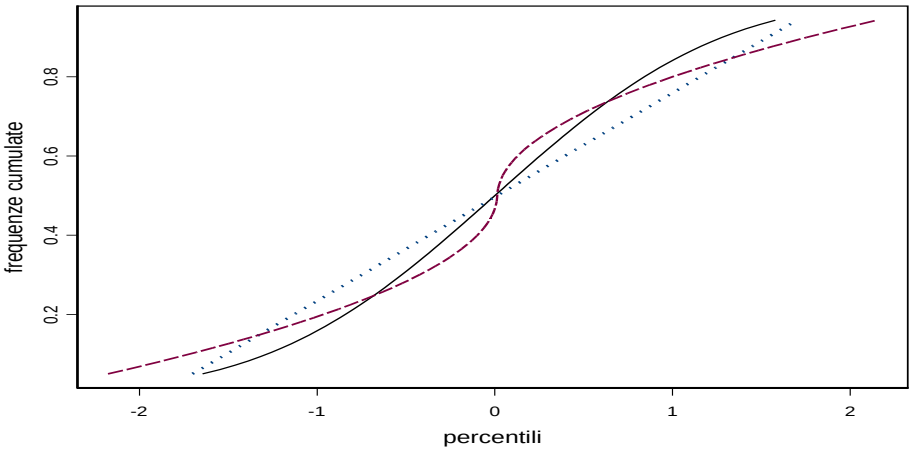

Figura. 1: *Funzioni di ripartizione empirica* $\hat{F}^{(j)}(\frac{x-\bar{r}^{(j)}}{S^{(j)}})$, $j = 0$ *(linea punteggiata) e* 1 *(linea tratteggiata), e* $\hat{\Phi}(x)$ *(linea continua).*

Il grafico relativo alla trasformazioni in ranghi (*linea punteggiata*) si presenta come una retta, indicativo del fatto che la trasformazione ha cancellato completamente la memoria della forma distributiva originaria dei dati. Malgrado ciò tale retta rimane comunque complessivamente la migliore approssimazione della funzione di ripartizione della variabile casuale normale (*linea continua*). Relativamente alla funzione di ripartizione $\hat{F}^{(1)}$ (*linea tratteggiata*) si vede invece come essa conservi la struttura campanulare della normale esaltandone la concentrazione dei valori trasformati intorno alla media. Ciò suggerisce che la trasformazione basata sul criterio (3) produca un adattamento ai dati originari migliore nel caso di dati leptocurtici. Inoltre, questa trasformazione è certamente ottimale quando i dati rivelino asimmetria, caratteristica che viene completamente cancellata dalla trasformazione in *ranghi ordinari*.

Per esplorare queste ipotesi assoggettiamo, prima, i valori in Z alla trasformazione

$$Y_h(z_i) = z_i e^{\frac{hz_i^2}{2}}. \tag{4}$$

Tale trasformazione è stata suggerita da Hoaglin (1985, p.479) per generare una famiglia di distribuzioni simmetriche con curtosi crescente al crescere di $h > 0$ e che comprenda la normale per $h = 0$. Indichiamo con $\hat{F}_h^{(j)}(x)$ le funzioni di ripartizione empirica associate ai valori $r_i^{(j)} = R^{(j)}(Y_h(z_i))$, $j = 0, 1$, e con $\hat{G}_h(x)$ la funzione di ripartizione sui valori $Y_h(z_i)$. Posto $\bar{Y}_h$ la media aritmetica delle $Y_h(z_i)$, e $S_{\bar{Y}_h}$ il corrispondente scarto quadratico medio, la distanza fra i dati originali e i loro trasformati è misurata, al netto di posizione e variabilità, dalla distanza

$$d_h^{(j)} = \frac{1}{m} \sum_k |\hat{G}_h(\frac{x_k - \bar{Y}_h}{S_{Y_h}}) - \hat{F}_h^{(j)}(\frac{x_k - \bar{r}^{(j)}}{S^{(j)}})|, j = 0, 1 \qquad (5)$$

al variare del parametro di curtosi h fra 0 e 10, essendo le x_k, $k = 1, 2, \ldots, m$, valori equispaziati fra -2.5 e 2.5. L'andamento delle $d_h^{(j)}$ al variare del parametro di curtosi h è riportata nella Figura 2. La curva

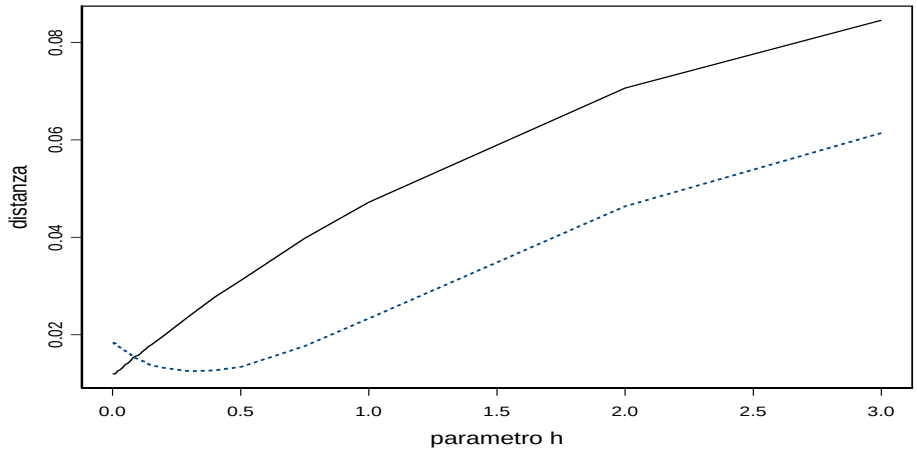

Figura. 2: *Distanze $d_h^{(j)}$ al variare del parametro di curtosi h.*

continua si riferisce ai dati trasformati in *ranghi ordinari* e si nota che per $h \approx 0$ la trasformazione $R^{(0)}$ è migliore della $R^{(1)}$ (*linea tratteggiata*).

L'ottimalità dei *ranghi ordinari* inizia a venir meno a partire dal valore $h \approx 0.08$, equivalente al valore 2.31 del momento standardizzato del quarto ordine.

Il secondo criterio di valutazione delle trasformazioni è costituito dall'asimmetria. A tal fine consideriamo la trasformazione

$$Y_g(z_i) = z_i \frac{e^{gz_i} - 1}{gz_i} \tag{6}$$

ancora suggerita da Hoaglin (1985, p.463) per generare una famiglia di distribuzioni con asimmetria controllata dal parametro g. Analogamente al caso precedente indichiamo con $\hat{F}_g^{(j)}(x)$ le funzioni di ripartizione empirica associate ai valori $r_i^{(j)} = R^{(j)}(Y_g(z_i)),\ j = 0, 1$, e con $\hat{G}_g(x)$ la funzione di ripartizione sui valori $Y_g(z_i)$. Nuovamente la distanza fra dati trasformati e originali è misurata dalla distanza (5) con ovvia riparametrizzazione rispetto al parametro di asimmetria g. Le conclusioni

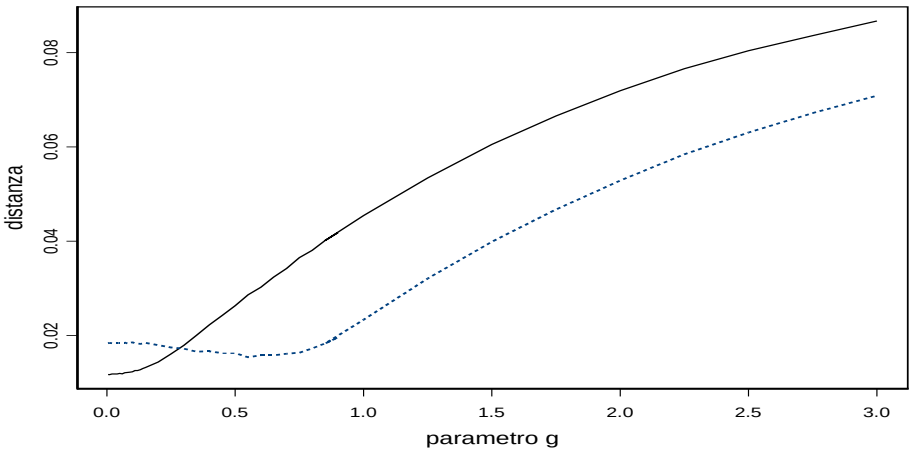

Figura. 3: *Distanze $d_g^{(j)}$ al variare del parametro di asimmetria g.*

che possono essere tratte dalla Figura 3 sono analoghe a quelle relative alla Figura 2; l'ottimalità della trasformazione $R^{(1)}$ inizia per valori di

$g \approx 0.25$ corrispondente al valore 0.36 dell'indice di asimmetria di Fisher. Il caso di asimmetria negativa è speculare a quello discusso.

La resistenza alle anomalie della trasformazione $R^{(1)}$ è investigata con la logica della funzione d'influenza empirica (Hampel *et al.*, 1986, pp.93-94) costruita con la tecnica della sostituzione.

Sia $x_k,\ k = 1, 2, \ldots, m$ la sequenza degli m valori equispaziati fra -2.5 e 2.5; per ogni $k = 1, 2, \ldots, m$ costruiamo l'insieme di valori $W_k = x_k \bigcup \{Z - z_{(0.5)}\}$, essendo $w_{i,k}$ l'i^{esimo} elemento dell'insieme W_k. Per ogni insieme W_k calcoliamo la distanza

$$\delta_k^T = \sqrt{\frac{\sum_i (z_i - \frac{T(w_{i,k}) - \overline{T(W_k)}}{S_{T(W_k)}})^2}{\sum_i (z_i^2)}}, \tag{7}$$

ove la $T(\cdot)$ rappresenta, a seconda del caso, le trasformazioni $R^{(0)}$, $R^{(1)}$, ovvero l'usuale standardizzazione Z. La Figura 4 contiene la rappresentazione grafica delle distanze $\delta_k^{(0)}$ (*linea tratteggiata*), $\delta_k^{(1)}$ (*linea punteggiata*) e la δ_k^Z (*linea continua*) al variare di x_k, che costituisce l'elemento di contaminazione e cui è associato l'insieme delle W_k. E' immediato osservare che la trasformazione $R^{(1)}$ ha una risposta alla contaminazione analoga alla trasformazione $R^{(0)}$. Le due trasformazioni si differenziano per il punto critico in cui cessa l'effetto dell'anomalia: $R^{(0)}$ non risente più di tale effetto per $|x_k| > 1.64$, $R^{(1)}$ per $|x_k| > 1.73$. Tale comportamento è indice di resistenza alle anomalie. L'usuale standardizzazione, di contro, è come noto non resistente, e, infatti, l'intensità dell'anomalia prosegue in tutto l'intervallo di analisi da -2.5 a $+2.5$.

3. L'analisi dei dati de Il Sole 24 ore

In questo paragrafo analizziamo i dati de *Il Sole 24 Ore* relativi al dossier sulla qualità della vita 1999 (Cadeo, 1999). Le analisi sono condotte secondo lo schema indicato dal quotidiano, ma sono mutate rispetto ad esso nel pre-trattamento degli indicatori statistici utilizzati per

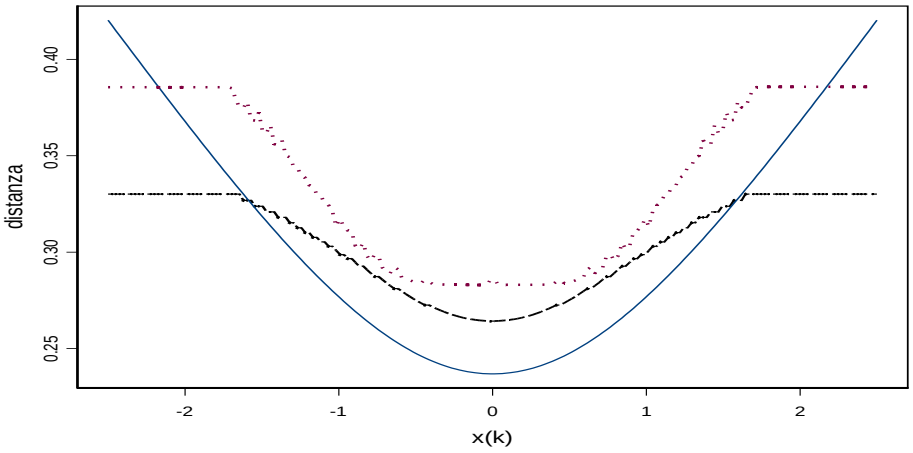

Figura. 4: *Distanze δ_k^T al variare della contaminazione x_k dell'insieme di dati di riferimento Z.*

costruire la graduatoria della qualità della vita nelle province italiane. La finalità delle analisi è di giungere ad una comparazione fra pre-trattamenti standard (ranghi ordinari e standardizzazione) e il pre-trattamento basato sulla proposta di trasformazione presentata nel precedente paragrafo.

L'insieme dei dati in esame è costituito da 36 indicatori statistici osservati sulla totalità delle 103 province italiane. Le 36 variabili sono suddivise in 6 gruppi tematici ognuno teso a cogliere una dimensione della qualità della vita.

Analisi esplorative preliminari effettuate su ognuna delle variabili hanno mostrato diversi casi di forte asimmetria e leptocurtosi. Nessuna delle variabili appare vicina al modello gaussiano. La forte asimmetria di alcune variabili induce a considerare come necessario un pre-trattamento basato su una trasformazione resistente. Tale scelta appare coerente, non tanto perchè si possa a ragione parlare di valori anomali (essendo i valori riferiti a popolazione) quanto perchè ai fini pratici le osservazioni distanti dalla maggioranza (come quelle corrispondenti alle province quali Napoli, Milano e Roma) inducono una concentrazione di valori

(in relazione, cioè, alle province medio piccole rispetto cui meglio si identifica la realtà italiana) in un intervallo molto ristretto. In alcune variabili vi sono dei dati mancanti. È ancora da evidenziare che fra gli indicatori ve ne sono alcuni che hanno un'associazione negativa con la variabile latente *Qualità della vita*, cioè un'alta intensità di quest'ultima corrisponde una bassa intensità dell'indicatore. È il caso, ad esempio, di tutte la variabili connesse all'aspetto della criminalità.

Lo schema di costruzione della graduatoria della qualità della vita proposta nel dossier (Cadeo, 1999) si basa su quattro passaggi:

1: pre-trattamento delle variabili originarie

2: assegnazione dei punteggi ad ogni provincia

3: costruzione delle variabili intermedie come media aritmetica su gruppi delle sei variabili tematiche

4: costruzione della graduatoria finale Q delle province come media aritmetica delle sei variabili intermedie.

Il punto critico dell'analisi condotta dal quotidiano sono le trasformazioni preventivamente effettuate sulle variabili prima dei passaggi di sintesi. Attanasio e Capursi (1997), Vitali e Merlini (1999), a cui rinviamo per approfondimenti, sviluppano accurate riflessioni critiche sul pre-trattamento effettuato dagli analisti del quotidiano. Per quanto concerne il secondo punto della lista precedente, i punteggi sono assegnati in modo proporzionale alle intensità partendo dal punteggio massimo 1000. Le variabili intermedie (punto 3) sono costruite come media aritmetica dei punteggi sul raggruppamento tematico delle variabili trasformate. La variabile qualità della vita (punto 4) è infine calcolata come media aritmetica calcolata sulla variabili intermedie. Nel seguito indicheremo con $Q^{(p)}$ la variabile pubblicata dal quotidiano. L'ordinamento delle province italiane rispetto a $Q^{(p)}$ restituisce la graduatoria della qualità della vita.

Le analisi da noi condotte per confrontare gli effetti delle trasformazioni incidono sul punto 1 e sul punto 2 del precedente schema. Abbiamo utilizzato come pre-trattamento dei dati, di volta in volta,

le tre trasformazioni $R^{(0)}$, $R^{(1)}$ e l'ordinaria standardizzazione Z che danno luogo alle graduatorie $Q^{(0)}$, $Q^{(1)}$ e $Q^{(Z)}$ rispettivamente. Ognuna di queste trasformazioni è corredata della corrispondente duale (Vitali e Merlini, 1999; p.23 - vedi Tabella 2) che permette di tener conto dell'associazione negativa della variabile con l'ordinamento finale.

Tabella 2. Trasformazioni duali delle trasformazioni $R^{(0)}$, $R^{(1)}$ e Z.

trasformazione	*trasformazione duale*
$R^{(0)}$	$(n+1) - R^{(0)}$
$R^{(1)}$	$(1 + \frac{n(n-1)}{2}) + 1 - R^{(1)}$
Z	$-Z$

L'assegnazione del punteggio ad ogni provincia è effettuata attraverso una trasformazione lineare che assegna valore 1 all'unità statistica con intensità più bassa e valore 1000 a quella con intensità più alta. Nessun punteggio è stato assegnato alle unità statistiche che presentano dati mancanti, ne deriva quindi che la media aritmetica del successivo punto 3 può essere riferita ad un numero di valori per unità statistica minore di 6, essendo questo il numero delle variabili che *misurano* il tema analizzato.

Alla fine del processo descritto, $Q^{(0)}$, $Q^{(1)}$ e $Q^{(Z)}$ indicano le variabili che contengono l'informazione sull'intensità della qualità della vita nelle province italiane. Insieme ad esse consideriamo anche la graduatoria $Q^{(p)}$ pubblicata dal quotidiano.

Il confronto fra le variabili segue lo schema tracciato da Attanasio e Capursi (1997). A tal fine, misuriamo, in primo luogo la concordanza fra le graduatorie derivate dalle variabili e quindi le confrontiamo rispetto ad una graduatoria di riferimento ottenuta componendo quelle disponibili.

La concordanza fra le graduatorie disponibili è basata sull'omonimo indice di Kendall (Kendall, 1962; p.95):

$$W = \frac{12S}{k^2(n^3 - n)}, \tag{8}$$

dove $k = 4$ è il numero delle graduatorie considerate e $n = 103$ è il numero delle unità statistiche. Posto

$$Q^{(\cdot)} = rank(Q^{(0)}) + rank(Q^{(1)}) + rank(Q^{(q)}) + rank(Q^{(p)}), \tag{9}$$

segue che $S = \sum_i^n (Q_i^{(\cdot)} - \frac{k(n+1)}{2})^2$.

Per i dati in esame risulta che $W = 0.9396$, intensità che legittima a considerare la graduatoria derivata da $Q^{(\cdot)}$ quale graduatoria di confronto per le altre. Calcoliamo, quindi:

$$\alpha_j = \sum_i^n |Q_{(i)}^{(\cdot)} - Q_{(i)}^{(j)}| \tag{10}$$

cioè il numero di posti in cui slittano le singole unità statistiche nella graduatoria associata a $Q^{(j)}$ rispetto alla graduatoria derivata da $Q^{(\cdot)}$, per $j = 0, 1, Z, p$. Si osservi che la (10) può essere interpretata come una distanza fra graduatorie.

I valori delle α_j, insieme al corrispondente valore dell'indice di Kendall, sono presentati nella Tabella 3. Appare subito evidente che la graduatoria $Q^{(p)}$ del quotidiano è la più lontana dalla graduatoria $Q^{(\cdot)}$. Sia la graduatoria $Q^{(Z)}$, ottenuta dai dati standardizzati, che $Q^{(0)}$, sfrutta la trasformazione in ranghi, si collocano sostanzialmente alla stessa distanza da $Q^{(\cdot)}$.

La graduatoria $Q^{(1)}$ costruita a partire dalla trasformazione $R^{(0)}$, proposta nel predente paragrafo, è quella che si rivela essere *più vicina* a $Q^{(\cdot)}$, ovvero è la graduatoria per la quale il numero di posti complessivi in cui slittano le unità statistiche è minimo tra tutte le graduatorie considerate.

Tabella 3. Confronto delle graduatorie calcolate Il Sole 24 Ore 1999.

	graduatoria	W	$Q^{(p)}$	$Q^{(Z)}$	$Q^{(1)}$	$Q^{(0)}$
1	α_j	0.9396	868	434	343	441
2	$\alpha_j^{-(p)}$	0.9877	(1136)	379	113	267
3	$\alpha_j^{-(Z)}$	0.9357	819	(600)	389	424
4	$\alpha_j^{-(1)}$	0.9289	768	475	(463)	520
5	$\alpha_j^{-(0)}$	0.9329	783	438	437	(538)

Abbiamo voluto verificare se il risultato ottenuto, considerando contemporaneamente le quattro graduatorie disponibili, fosse stabile rispetto alla presenza-assenza di una di esse. Nella righe 2-5 della Tabella 3 abbiamo calcolato di nuovo l'indice di Kendall (8) per $k = 3$ escludendo

di volta in volta una delle quattro graduatorie, come evidenziato dall'apice delle α_j. La graduatoria di riferimento è quella derivata da $Q^{(\cdot)}$ calcolata come nella (9) con esclusione di una delle quattro. Il calcolo delle α_j è rimasto invariato. Abbiamo posto fra parentesi il valore della α_j calcolata su una graduatoria che è stata esclusa dal computo della $Q^{(\cdot)}$. Osservando comparativamente le righe della tabella 3 si vede che la minimalità delle α_j è sempre in corrispondenza della graduatoria derivata dal pre-trattamente con la trasformazione $R^{(1)}$. Particolarmente interessanti sono le righe 4 e 5, in cui si vede che la graduatoria $Q^{(0)}$ è sempre ben distante dalle graduatorie $Q^{(Z)}$ e $Q^{(1)}$, ed ancora (nella riga 4) la graduatoria $Q^{(1)}$ conserva comunque la minimalità sebbene sia stata esclusa al fine della determinazione $Q^{(\cdot)}$.

La scelta del dossier 1999 de *Il Sole 24 Ore* è stata dettata dall'opportunità di confrontare l'analisi pubblicata con quella condotta parallelamente dal quotidiano *Italia Oggi* (Mori, 1999) sotto la guida scientifica di Vitali e Merlini sempre rispetto allo stesso anno di riferimento. La logica complessiva di questa seconda ricerca è uguale a quella de *Il Sole 24 Ore*. L'analisi di *Italia Oggi* si differenzia metodologicamente dalla prima per l'omogeneità del pre-trattamento degli indicatori statistici tutti trasformati secondo leggi lineari. Anche su questi dati sono state condotte le stesse analisi già presentate per i dati de *Il Sole 24 Ore*. I risultati sono sintetizzati nella Tabella 4 il cui commento è identico a quello della Tabella 3.

Tabella 4. Confronto delle graduatorie calcolate Italia Oggi 1999.

	graduatoria	W	$Q^{(p)}$	$Q^{(Z)}$	$Q^{(1)}$	$Q^{(0)}$
1	α_j	0.9137	862	611	381	494
2	$\alpha_j^{-(p)}$	0.9736	(1132)	501	260	309
3	$\alpha_j^{-(Z)}$	0.9148	803	(795)	355	472
4	$\alpha_j^{-(1)}$	0.8995	797	598	(501)	594
5	$\alpha_j^{-(0)}$	0.9053	777	625	507	(641)

4. Conclusioni

In questo lavoro è stata presentata una trasformazione che può considerarsi una generalizzazione dei ranghi infatti essa condivide con questi ultimi lo stesso criterio di minimo che dà luogo ai valori trasformati. La trasformazione proposta conserva in parte le mutue distanze fra le osservazioni e con esse parte della forma distributiva dei dati originari, proprietà non riscontrabile nell'ordinaria trasformazione in ranghi. Questa trasformazione si rivela particolarmente valida quando i dati originari presentano una forte concentrazione nell'intorno della media; inoltre è resistente alle anomalie, proprietà che condivide con i ranghi ordinari. Tenendo conto delle proprietà che sono state evidenziate essa può risultare proponibile nei casi in cui è necessaria una standardizzazione dei dati che non alteri rendendo uniforme la forma della distribuzione e che sia al contempo resistente alle anomalie.

In questo lavoro la trasformazione proposta è stata utilizzata come pre-trattamento delle variabili al fine di costruire una graduatoria derivata da una variabile latente. In tale ambito si è dimostrato che la graduatoria finale è più stabile rispetto ad altre trasformazioni standard di pre-trattamento dei dati.

La trasformazione proposta può avere campi di applicazione più ampi. Difatti sfruttando alcune caratteristiche della trasformazione: che 1) il valore massimo è legato alla numerosità dei dati, 2) le possibili n^{ple} derivabili dall'utilizzo della trasformazione sono finite e numerabili, diviene possibile investigare la possibilità di costruire nuovi test non parametrici.

Ringraziamenti: Il lavoro è stato svolto nell'ambito del progetto di ricerca *Aspetti inferenziali e computazionali delle tecniche non parametriche robuste. Applicazioni a dati territoriali.* F.A.R.2002 Università degli Studi del Sannio.

Riferimenti Bibliografici

Attanasio M., Capursi V. (1997), Graduatorie sulle qualità della vita: prime analisi di sensibilità di tecniche adottate, *Atti della XXXV Riunione Scientifica della SIEDS*, 331-342.

Cadeo R. (a cura di), Dossier sulla Qualità della vita 1999, *Il Sole 24 Ore*, n.351 del 27/12/1999

Conover W.J., Iman R.L. (1981), Rank Transformations as Bridge between Parametric and Nonparametric Statistics *The American Statistician*, 35, 3, 124-133.

Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A. (1986), *Robust Statistics: the approach based on the influence function*, John Wiley, New York.

Kendall M.G. (1962), *Rank Correlation Methods*, C.Griffin & C. ltd, London.

Mori C. (a cura di), Rapporto 1999 sulla qualità della vita in Italia, *Italia Oggi*, anno 9 n.252 del 21/12/1999

Hoaglin D.C. (1985), Summarizing Shape Numerically: the g-and-h distributions, in *Exploring Data, Tables, Trends and Shapes*, D.C.Hoaglin, F.Mosteller, J.W.Tukey eds., John Wiley, New York.

Pagnotta S.M. (2002), Generalized Rank Transformation, *Atti della XLI Riunione Scientifica della Società Italiana di Statistica*, vol.2, 609-612.

Vitali O., Merlini A. (1999), La qualità della vita: metodi e verifiche, *Rivista italiana di Economia Demografia e Statistica*, LIII, 2, 5-91.