Quaderni di Statistica
Vol. ??, 200?

Graduatorie della qualità della vita e loro sensibilità al pre-trattamento delle variabili che la definiscono: alcune critiche al Dossier de Il Sole 24 Ore.

Silvia Terzi
Dipartimento di Economia Università "Roma TRE"
E-mail: terzi@uniroma3.it

Luca Moroni
Dipartimento di Economia Università "Roma TRE"
E-mail: *moroni@uniroma3.it*

Summary: When constructing a composite indicator a choice has to be made for what concerns a link function and preliminary variable transformations. However, these choices cannot be independent. For instance, when an additive link function is used (such as the average of transformed variables) the implicit assumption is that of equal weights of simple indicators; however, if the original variables are not all linearly transformed, the mathematical relationship between the composite indicator and the original variables is not definable. This is exactly the case of the survey on the "Quality of life" in the Italian provinces, produced each year by the financial newspaper Il Sole 24 ore. The main aim of the survey is not to measure the latent variable quality of life, but rather to rank the 103 Italian provinces. In this paper we attempt to study the influence of the variable transformations adopted in the survey on the final classification, arguing in favour of linear scale- and-translation- invariant variable transformations.

Keywords: Composite indicators, Quality of life, Variable transformations.

1. Introduzione

Esistono in letteratura diversi metodi per costruire indicatori sintetici. Il comune presupposto è una assunzione teorica circa la relazione matematica che lega l'indicatore sintetico alle variabili su cui si basa; tale assunzione comporta l'individuazione di una funzione aggregatrice (media aritmetica semplice o media aritmetica ponderata; pesi individuati da "esperti" o pesi ricavati dai dati stessi), ed il pre-trattamento a cui sottoporre le variabili (eliminazione dell'unità di misura, standardizzazione, trasformazione in ranghi). Diverse combinazioni di una funzione *link* e di funzioni che permettano di trasformare le variabili di partenza in dati omogenei produrranno indicatori sintetici diversi; di conseguenza anche nell'ambito degli indicatori di *qualità della vita* sono state formulate numerose proposte (vedi, tra gli altri, Lauro (2003), Pagnotta (2003), Dossier Italia Oggi (2001)).

Nel presente lavoro, ripercorreremo la metodologia utilizzata da Il Sole 24 ore per stilare la classifica della qualità della vita nelle province italiane (Dossier Qualità della Vita 2003), per fare luce soprattutto su alcune implicazioni delle scelte effettuate nella fase di pre-trattamento delle variabili. Pertanto, senza entrare nel merito della scelta di quali variabili utilizzare né della migliore funzione aggregatrice, partiremo dalla stessa scelta de Il Sole 24 ore, ovvero i 36 indicatori da loro individuati e la media aritmetica semplice come indice di sintesi. La media aritmetica delle variabili trasformate fornisce i punteggi globali delle singole province, sulla cui base viene stilata la classifica finale. Il nostro obiettivo è valutare la sensibilità della classifica finale alle diverse trasformazioni che si possono effettuare, in modo da stimolare una riflessione critica sul metodo più coerente con i requisiti di neutralità che un indicatore sintetico di questo tipo dovrebbe avere.

Il punto di partenza della nostra critica alla metodologia utilizzata da Il Sole 24 ore è che alcune delle 36 variabili su cui si basa l'indicatore *qualità della vita* (Q.d.V.) vengono sottoposte ad una trasformazione non lineare. Non risulta quindi possibile esprimere in maniera analitica la relazione che le lega all'indicatore sintetico. Si tratta di una critica non nuova in letteratura, ben argomentata in Attanasio e Capursi (1997).

Sulla loro scia, per valutare le conseguenze della non-linearità sull'indicatore sintetico e quindi sulla classifica pubblicata da Il Sole 24 ore, lo confrontiamo con alcuni indicatori basati su trasformate lineari. Tali confronti saranno volti non tanto ad individuare la migliore stima della graduatoria della variabile latente Q.d.V. (come accade nel paper citato), bensì l'effetto che le trasformazioni non lineari hanno sulla classifica finale.

Il lavoro si articola nel seguente modo: i paragrafi 2 e 3 sono dedicati prevalentemente alle trasformazioni lineari ed hanno un carattere introduttivo al problema. Dopo aver illustrato alcune caratteristiche delle trasformazioni utilizzate nel Dossier de Il Sole 24 ore, le confrontiamo con alcune delle trasformazioni lineari più frequentemente proposte in letteratura, argomentando a favore della scelta di trasformazioni lineari invarianti per traslazione (paragrafo 2). Inoltre, esplicitando la relazione che lega l'indicatore sintetico alle variabili non trasformate (cosa possibile solo nel caso in cui si abbiano trasformazioni lineari) definiamo il contesto formale e la simbologia più appropriata per lo studio della sensibilità dell'indicatore composto Q.d.V. alle trasformazioni utilizzate (paragrafo 3). Nei paragrafi 4 e 5, entrando nel vivo del nostro contributo, studiamo la sensibilità della graduatoria finale alle diverse trasformazioni illustrate nei paragrafi precedenti. Ciò è possibile solo comparativamente, cioè attraverso il confronto tra graduatorie ottenute mediante un diverso pre-trattamento delle medesime variabili. Quindi, dopo aver introdotto alcuni strumenti necessari per individuare le cause delle maggiori differenze tra indicatori sintetici basati su trasformazioni lineari differenti, confrontiamo tra loro le classifiche che ne derivano (paragrafo 4); infine (paragrafo 5) confrontiamo tali graduatorie con la classifica de Il Sole 24 ore, sia per individuare le conseguenze della non-linearità sulla graduatoria finale, che per misurarne l'incidenza.

2.Trasformazione delle variabili

2.1 Eliminazione dell'unità di misura

Il primo passo nella costruzione di un indicatore sintetico consiste nel trasformare le variabili di partenza in indicatori semplici, adimensionali e quindi aggregabili.

Posto che si voglia dare 1 (o 1000) alla provincia con la performance migliore, posto che si sia osservata una variabile X che caratterizzi in positivo la *qualità della vita* (ovvero che sia positivamente correlata con la Q.d.V.), la trasformazione da utilizzare potrà essere:

$$t_i^+ = \frac{x_i}{\max(x)} \qquad (1)$$

dove $\max(x) = \max_{i\in\{1,\ldots,n\}}(x_i)$. Meno ovvia la trasformazione a cui sottoporre le variabili che caratterizzano in negativo la Q.d.V. (negativamente correlate). La trasformata utilizzata da Il Sole 24 ore è la seguente:

$$y_i = \frac{\min(x)}{x_i} \qquad (2)$$

e la logica sottostante é che le variabili negativamente correlate con la qualità della vita vengono prima trasformate nei loro reciproci e poi normalizzate secondo la (1) rapportandole al massimo. Tuttavia tale procedimento modifica la forma della distribuzione ed altera (non solo nel segno) la struttura di correlazione originaria delle variabili. Ci sembra quindi che la trasformazione (2) meriti una più accurata riflessione.

2.2 Il trattamento delle variabili correlate negativamente con la qualità della vita

Per valutare alcune caratteristiche della trasformazione (2), supponiamo di aver rilevato una variabile X (negativamente correlata con la Q.d.V.) su una popolazione di 10 unità e che disponendo le $x_{(i)}$ in ordine non decrescente sia risultato:

$$x_{(i)} = i, \quad i = 1,\ldots,10$$

Partendo da questo caso creiamo tre popolazioni, ottenute lasciando invariate le unità da 2 a 10 e ponendo $x_{(1)}$ uguale, rispettivamente, a 0.5, 1 e 1.5. Le tre serie trasformate (e ordinate) sono rappresentate nella Figura 1.

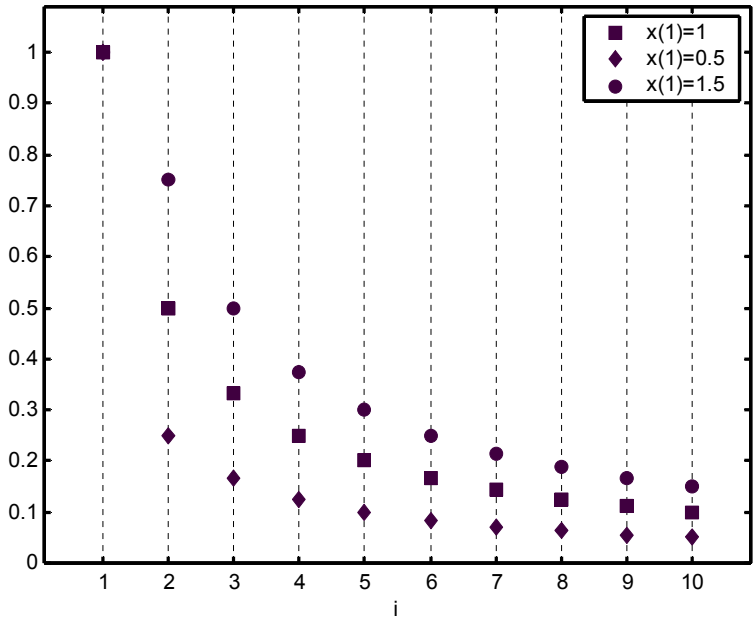

Figura 1: trasformata y al variare del minimo

È immediato fare le seguenti considerazioni:

1. La trasformazione utilizzata espande la parte alta della distribuzione e comprime la coda destra; ovvero esalta differenze anche modeste tra le prestazioni migliori, mentre riduce le differenze tra le prestazioni peggiori.

2. La trasformazione utilizzata è sensibile a piccole variazioni nel minimo della distribuzione di partenza; in particolare una diminuzione di $x_{(1)}$ ha un effetto sensibilmente superiore sulla parte alta della distribuzione che non sulla coda destra. Ciò implica che una variazione nella performance migliore si riflette in misura maggiore sulle performances buone che non su quelle meno buone.

Se riscriviamo la (2) come:

$$y_i = \frac{\min(x)}{x_i} = 1 - \frac{d\left(x_i, \min(x)\right)}{d(x_i, 0)},\ i = 1,\ldots,10$$

dove $d(\cdot\,,\cdot)$ rappresenta la distanza euclidea, possiamo notare un'altra peculiarità di questa trasformazione: una traslazione della serie delle x_i del tipo: $x_{(i)} = i + k,\ k > 0$, altera il range della y. In particolare, supponendo di avere tre popolazioni ottenute ponendo, rispettivamente, k=0, k=5, k=10, per i=1, 2,…, 10, abbiamo la seguente rappresentazione grafica, da cui è immediato constatare che a parità di range delle tre distribuzioni, il range delle y decresce al crescere di k :

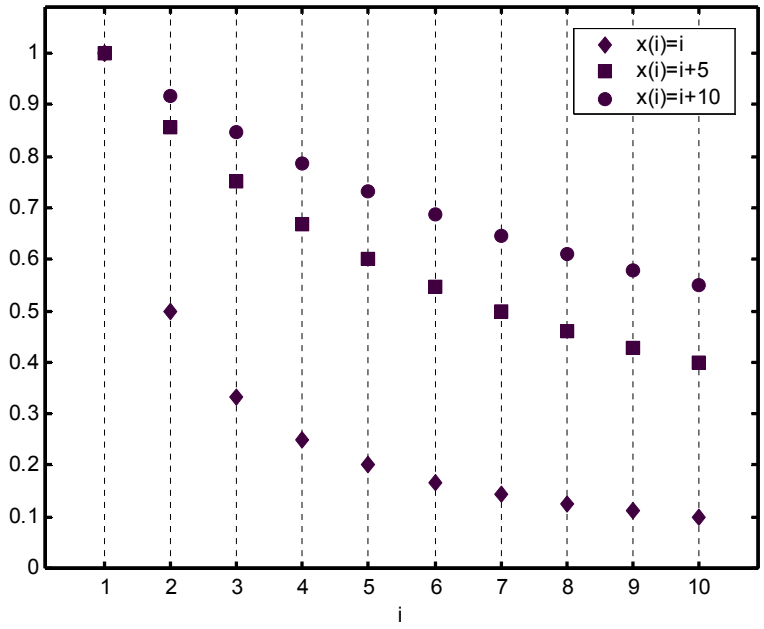

Figura 2: trasformata y di distribuzioni traslate

Per effettuare un confronto della *y* con una trasformazione lineare dei dati, scegliamo quella retta avente gli stessi minimo e massimo della *y*, ossia:

$$t_i^- = 1 - \frac{x_i}{\max(x)} + \frac{\min(x)}{\max(x)} \qquad (3)$$

La linearità della trasformazione (3) risolve il problema della disomogeneità di trattamento delle differenze tra prestazioni (ossia $x_i - x_h = c(t_i^- - t_h^-),\ c \in \mathbf{R}, \forall i,h$); inoltre modeste variazioni del minimo $x_{(1)}$, pur alterando il range della trasformata, portano a modeste variazioni nei punteggi assegnati:

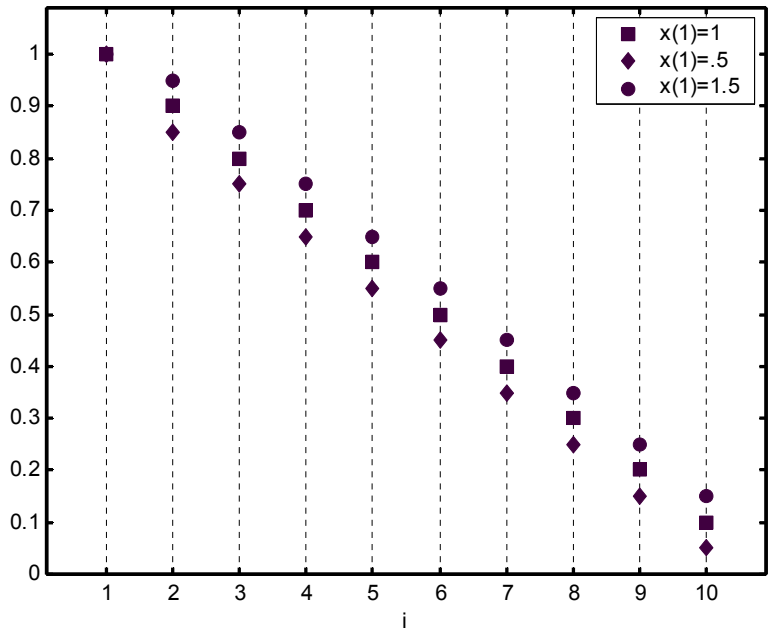

Figura 3: trasformata t^- al variare del minimo

Infine tale trasformazione ha un'importante caratteristica che l'accomuna alla (1), ovvero alla trasformazione utilizzata dal Sole24ore per le variabili positivamente correlate con la Q.di V.; infatti appartengono entrambe alla famiglia di trasformazioni:

$$T(x) = \pm\frac{x}{\max(x)} + c$$

dove c è una generica costante. I membri di tale famiglia hanno tutti – a meno del segno – lo stesso campo di variazione e gli stessi momenti.

Inoltre t^+ e t^- assumono valori nello stesso intervallo $\left[\frac{\min(x)}{\max(x)}, 1\right]$, il che rende la trasformata t^- la più naturale alternativa alla y.

Per confrontare la trasformata t^- con la trasformata y, riportiamo su uno stesso grafico le funzioni alla base delle trasformate (2) e (3), definendo:

$$\forall x \in [x_{\min}, x_{\max}] \subset \mathbf{R}^+$$

$$y(x) = \frac{x_{\min}}{x}$$

$$t^-(x) = 1 - \frac{x}{x_{\max}} + \frac{x_{\min}}{x_{\max}}$$

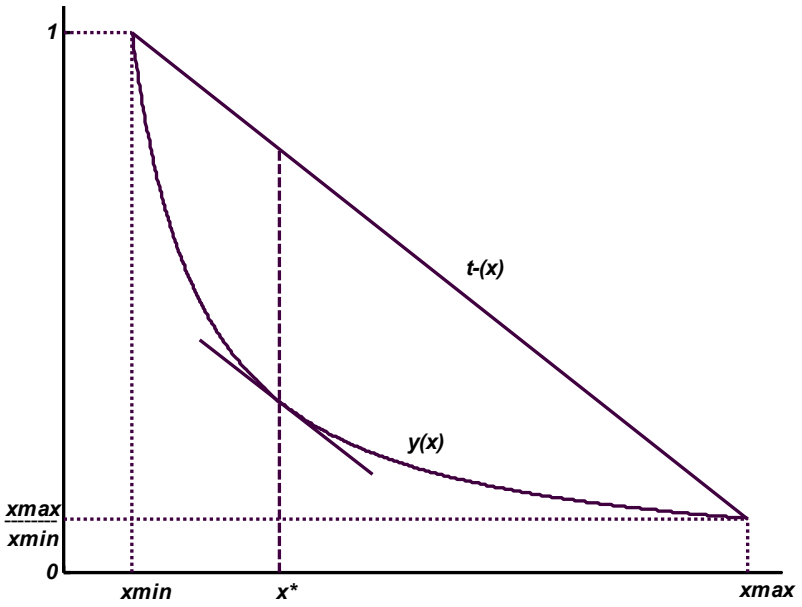

Figura 4: confronto tra le trasformate y e t -

Sia $x = x^* = \sqrt{x_{\min} x_{\max}}$; si può agevolmente verificare che nell'intervallo $[x_{\min}, x^*)$ la derivata della funzione y è sempre minore dell'inclinazione della retta t^-, mentre nell'intervallo $(x^*, x_{\max}]$ tale derivata è sempre maggiore. Ciò implica che, dati

$x_i, x_h \in [x_{\min}, x^*), x_i < x_h$, risulterà $y(x_i) - y(x_h) > t^-(x_i) - t^-(x_h)$. Viceversa per $x_h < x_i \in (x^*, x_{\max}]$. Peraltro, esisteranno anche infinite coppie di valori $x_i \in [x_{\min}, x^*), x_h \in (x^*, x_{\max}]$ per le quali sarà $y(x_i) - y(x_h) = t^-(x_i) - t^-(x_h)$.

Le principali differenze tra t^- ed y sono quindi legate al fatto che mentre t^- trasforma in maniera uniforme la distribuzione, y espande la prima parte della distribuzione e comprime la coda destra, alterando completamente la forma della distribuzione.

Vediamo con un esempio come queste caratteristiche della trasformata y influiscono sulla sintesi delle singole prestazioni, sintesi ottenuta come media aritmetica dei punteggi riportati su ciascun indicatore. Supponiamo di avere due caratteri X_j, $j = 1,2$ entrambi con distribuzione uniforme discreta in $\{1,2,\ldots,10\}$, ed indichiamo con x_{ij} le loro determinazioni. Per due unità A e B sia $x_{A1} = 1$ e $x_{B1} = x_{B2} = 2$. Vogliamo stabilire per quale valore $x_{A2} \in \{1,3,\ldots,10\}$ l'unità A occuperà una posizione in classifica migliore dell'unità B. E' facile verificare che ciò accadrà ogni volta che:

$$y_{A1} - y_{B1} > y_{B2} - y_{A2}$$

Nel nostro esempio, a causa dell'elevata differenza tra y_{A1} e y_{B1}, ciò accadrà sempre[1]; viceversa, con la trasformata t^-, l'unità A occuperà una posizione migliore o equivalente all'unità B solo se è $x_{A2} \leq 3$.

Si vede chiaramente come, a livello di sintesi, i due effetti di espansione e contrazione sopra individuati, portino un vantaggio rispetto alla trasformata t^-, alle unità che presentano al tempo stesso, ottime performances per qualche indicatore e cattive o pessime performances per qualche altro.

---

[1] In realtà, proprio per limitare tali conseguenze, Il Sole 24 Ore corregge la trasformata y ogni volta che $y_{(i)} - y_{(i+1)} > 0.25$ ponendo $y_{(i+1)} = y_{(i)} - 0.25$ e definendo di conseguenza un diverso valore min(x^*) tramite cui trasformare secondo la (2) le restanti osservazioni $x_{(i+h)}, h = 2,\ldots,n-i+1$.

2.3. Classi di trasformazioni invarianti per traslazione

Nonostante le numerose caratteristiche che accomunano gli indicatori t^+ e t^- appena definiti, essi differiscono per un aspetto importante: l'influenza degli outliers. In particolare le risposte di t^+ e t^- ad outliers corrispondenti a performance particolarmente buone (o viceversa a performance particolarmente scadenti) non sono uguali. Inoltre sia la t^+ che la t^- reagiscono in maniera asimmetrica ad un outlier eccezionalmente alto o eccezionalmente basso.

Per vedere meglio questo aspetto riscriviamo le trasformate (1) e (3) nel modo seguente:

$$t_i^+ = \frac{d(x_i, 0)}{d\big(\max(x), 0\big)} \tag{1a}$$

$$t_i^- = 1 - \frac{d\big(x_i, \min(x)\big)}{d\big(\max(x), 0\big)} \tag{3a}$$

Supponiamo di aver osservato una performance eccezionalmente scadente. A seconda del tipo di correlazione che c'è tra la X e la Q.d.V. questa si configurerà come min(x) eccezionalmente basso o come max(x) eccezionalmente alto. Tuttavia è immediato constatare che mentre un decremento in min(x) non induce alcuna variazione nelle t^+ se non per l'unità che assume tale valore, un incremento in max(x) modifica tutta la serie di valori della t^- Ciò comporta che nel primo caso risulterà penalizzata solo l'unità statistica a cui corrisponde questa pessima performance, nel secondo caso risulteranno proporzionalmente modificati tutti i valori assunti dalla t^-. Viceversa, una performance eccezionalmente buona avrà l'effetto di modificare l'intera serie t^+ (penalizzando quindi tutte le unità statistiche tranne quella con la performance migliore), mentre sulla t^- modificherà solo il valore corrispondente all'unità dal comportamento eccezionale. Inoltre, osservando le espressioni (1a) e (3a) si nota che sono entrambe sensibili

a traslazioni delle variabili di partenza ma che le traslazioni influiscono in misura diversa: nella t^- alterano solo il denominatore, nella t^+ alterano *anche* il numeratore, seppure non in misura proporzionale alla variazione del denominatore.

Poiché delle 36 variabili utilizzate nella nostra analisi più di una serie presenta outliers, e ben 17 sono negativamente correlate con la Q.d.V., sembra preferibile eliminare le asimmetrie appena viste utilizzando una trasformazione invariante per traslazione, a range costante. In particolare proponiamo:

$$w_i^+ = \frac{x_i - \min(x)}{\max(x) - \min(x)} = \frac{d(x_i, \min(x))}{d(\max(x), \min(x))}$$

per le variabili correlate positivamente e:

$$w_i^- = \frac{\max(x) - x_i}{\max(x) - \min(x)} = 1 - \frac{d(x_i, \min(x))}{d(\max(x), \min(x))}$$

per quelle correlate negativamente. Il peso di un singolo outlier risulterà attenuato ma - soprattutto - non ci sarà più asimmetria nell'effetto di outliers grandi e outliers piccoli.

Anche in questo caso si tratta di due trasformate appartenenti ad una medesima famiglia:

$$w(x) = \pm \frac{x}{\max(x) - \min(x)} + c$$

e quindi, oltre ad avere lo stesso range, hanno anche gli stessi momenti.

L'uso di tale trasformata è stato proposto – tra gli altri – da D'Esposito e Ragozini (2004) come strategia di ordinamento di osservazioni multivariate. Essi dimostrano che trasformare mediante la $w(x)$ un insieme di K variabili X_j e poi aggregarle mediante la somma, equivale a proiettare n osservazioni multivariate x_{ij}, $i = 1,2,...,n$, $j= 1,...,K$, lungo la direzione che va dal punto di coordinate pari ai minimi delle distribuzioni univariate al punto di coordinate pari ai massimi delle

distribuzioni univariate, cioè lungo una direzione significativa ai fini dell'ordinamento delle osservazioni.[2]

3. Sintesi degli indicatori

Come argomentato in Aiello e Attanasio (2004), la costruzione di un indicatore sintetico può essere vista come il risultato di due momenti distinti: individuazione di funzioni che permettano di trasformare le variabili di partenza in dati omogenei (indicatori semplici) e scelta di una funzione che applicata ai dati omogenei fornisca una misura dell'indicatore composto. Entrambi questi momenti concorrono alla definizione della relazione matematica che lega l'indicatore composto alle variabili su cui si basa. Ad esempio, la scelta di una funzione *link* lineare additiva (come la media aritmetica semplice utilizzata nel Dossier de Il Sole 24 ore) equivale all'assunzione di uguale peso degli indicatori semplici nella determinazione dell'indicatore composto; mentre il peso delle singole variabili sarà determinato dalla trasformazione utilizzata. Viceversa, qualora alcune variabili vengano sottoposte a trasformazioni non-lineari, la relazione che le lega all'indicatore sintetico non risulta definibile in termini analitici. In altri termini (vedi Aiello e Attanasio, 2004) la duplice scelta di una funzione link lineare additiva e di trasformazioni non-lineari di variabili, risulta inappropriata dal punto di vista matematico.

Scopo del presente paragrafo e' illustrare un contesto formale adeguato allo studio della sensibilità dell'indicatore composto Q.d.V. alle trasformazioni utilizzate. Esso poggia sull'esplicitazione della funzione che lega l'indicatore composto alle variabili $X_j, j = 1,\ldots,K$ su cui si basa, nell'ipotesi che siano sottoposte a trasformazioni lineari. In tal caso, prendendo come funzione *link* la media aritmetica semplice e indicando con *P* la variabile (indicatore composto) "punteggio finale", sarà:

[2] Anche i punteggi ottenuti con la trasformata *T* possono essere visti come risultato di una proiezione. In tal caso la retta su cui si proiettano le osservazioni multivariate è quella che congiunge l'origine al punto di coordinate pari ai massimi delle distribuzioni univariate.

$$P = \beta + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_K X_K \qquad (4)$$

Per esplicitare ulteriormente la (4) si indichi con *n* il numero di unità (province), con P_i il valore dell'indicatore sintetico per l'*i*-esima unità e con $N_{i,j}$ gli indicatori semplici[3]:

$$N_{i,j} = a_j x_{i,j} + b_{i,j}, i = 1,\ldots n, j = 1,\ldots K$$

Sarà quindi:

$$P_i = \sum_{j=1}^{K} \frac{N_{i,j}}{K} = \sum_{j=1}^{K} \frac{a_j x_{i,j}}{K} + \sum_{j=1}^{K} \frac{b_j}{K} = \sum_{j=1}^{K} \alpha_j x_{i,j} + \beta \qquad (5)$$

L'influenza delle variabili X_j sull'indice sintetico *P* è misurata dal coefficiente α_j, a sua volta determinato dal tipo di trasformazione lineare utilizzata. Viceversa i *K* coefficienti b_j non influiscono né sull'ordinamento delle unità né sulle loro distanze in termini dell'indice sintetico *P*. Ciò comporta la perfetta equivalenza – ai fini della graduatoria finale – tra la trasformata *t* ed una sua qualunque traslazione come anche tra la trasformata *w* e le sue traslazioni. Pertanto le trasformate lineari introdotte finora possono definirsi:

$$T(x) = \pm \frac{x}{\max(x)}$$

$$w(x) = \pm \frac{x}{range(x)}$$

I coefficienti ad esse associate sono[4]:

---

[3] Per le variabili associate negativamente con l'indicatore sintetico i coefficienti a_j assumono valore negativo. Tuttavia, per semplicità di notazione risulta più agevole sottintendere che a tali variabili viene cambiato il segno.

[4] A meno della costante sono queste le trasformate proposte, tra gli altri, da Attanasio e Capursi (1997)

$$ {}^{t}a_j = \frac{1}{\max(x_j)} $$

$$ {}^{w}a_j = \frac{1}{range(x_j)} $$

I coefficienti ${}^{t}a_j, j = 1,\ldots,K$ forniscono un indicatore sintetico P invariante per cambiamenti di scala delle variabili su cui si basa, mentre i coefficienti ${}^{w}a_j, j = 1,\ldots,K$ danno luogo ad un indicatore sintetico invariante, oltre che per cambiamento di scala, anche per traslazione delle variabili X_j.

Una possibile chiave nella scelta di a_j (e quindi delle trasformazioni da utilizzare) sta nel fatto che tale coefficiente influenza la variabilità dell'indicatore semplice, e quindi il suo peso nel determinare la variabilità dell'indice sintetico.

Il range, utilizzato nelle trasformate *w*, è un rudimentale indice di variabilità, ma ne esistono altri, quali ad esempio lo scostamento quadratico medio o la mediana degli scarti assoluti dalla mediana (MAD), i quali hanno il pregio, rispetto al range, di essere meno sensibili alle variazioni dei valori estremi della distribuzione, e quindi meno sensibili agli outliers. Conseguentemente si trovano in letteratura numerose proposte a favore dell'usuale standardizzazione:

$$ z_{i,j} = \frac{x_{i,j} - M_1(x_j)}{s.q.m._j} $$

oppure della trasformata MAD:

$$ m_{i,j} = \frac{x_{i,j} - Me(x_j)}{MAD_j}. $$

Tali trasformate danno luogo ai seguenti coefficienti a_j:

$$^{z}a_j = \frac{1}{s.q.m._j}$$

$$^{m}a_j = \frac{1}{MAD_j}$$

4. Alcuni confronti tra trasformazioni lineari

Si è appena evidenziato che le diverse trasformazioni lineari introdotte nei paragrafi precedenti assegnano diverso peso alle variabili originarie. Vogliamo ora vedere le conseguenze che questo ha sulle classifiche finali sia in termini di un'analisi delle variabili che maggiormente influiscono sulle differenze tra le classifiche stesse, sia in termini di un'analisi delle province maggiormente avvantaggiate (o penalizzate) da una scelta piuttosto che da un'altra.

Siano $P_i^w = \frac{1}{K}\sum_j w_{i,j}$, $P_i^m = \frac{1}{K}\sum_j m_{i,j}$, $P_i^z = \frac{1}{K}\sum_j z_{i,j}$; riprendendo la notazione che ha condotto alla (5) possiamo scrivere:

$$P_i^w = \frac{1}{K}\sum_j \frac{1}{range(x_j)} x_{i,j} + \beta = \frac{1}{K}\sum_j \frac{MAD_j}{range(x_j)} m_{i,j} + \gamma \qquad (6)$$

come anche:

$$P_i^w = \frac{1}{K}\sum_j \frac{1}{range(x_j)} x_{i,j} + \beta = \frac{1}{K}\sum_j \frac{s.q.m._j}{range(x_j)} z_{i,j} + \delta \qquad (7)$$

Supponiamo ora che per due unità (*A* e *B*) e due distribuzioni, sia $w_{A,1} \geq w_{B,1}$, $w_{B,2} \geq w_{A,2}$ e inoltre $w_{A,1} + w_{A,2} = w_{B,1} + w_{B,2}$; supponiamo cioè che le due unità siano equivalenti ai fini della classifica basata sulla trasformata *w*. Ciò comporta che: $w_{A,1} - w_{B,1} = w_{B,2} - w_{A,2}$ e che tali differenze sono non negative.

Utilizzando la (6), è facile verificare che quest'ultima uguaglianza equivale a:

$$\frac{MAD_1}{range_1}(m_{A,1} - m_{B,1}) = \frac{MAD_2}{range_2}(m_{B,2} - m_{A,2})$$

da cui:

$$P_A^m - P_B^m = \left(\frac{\frac{MAD_2}{range_2}}{\frac{MAD_1}{range_1}} - 1 \right)(m_{B,2} - m_{A,2}) \qquad (8)$$

Pertanto, in termini di *m* la differenza tra *A* e *B* verrà esaltata tanto più, quanto più è grande il rapporto:

$$K_{m|w} = \frac{\frac{MAD_2}{range_2}}{\frac{MAD_1}{range_1}}$$

In maniera analoga, utilizzando la (7) si può verificare che la differenza tra *A* e *B* in termini di *z* verrà esaltata tanto più, quanto più è grande il rapporto:

$$K_{z|w} = \frac{\frac{s.q.m_2}{range_2}}{\frac{s.q.m_1}{range_1}}$$

Entrambi i rapporti *MAD/range* e *s.q.m./range* decrescono in presenza di un outlier. Pertanto, supponendo che la prima distribuzione sia ottenuta dalla seconda mediante l'introduzione di un outlier, risulterà $K_{m|w} > 1$ e $K_{z|w} > 1$. Inoltre, dati $K_{m|w} > 1$ e $K_{z|w} > 1$, le differenze $P_A^m - P_B^m$ e $P_A^z - P_B^z$ saranno massime per $x_{B,2} = \max(x_2)$ e $x_{A,2} = \min(x_2)$ (il che, dato il vincolo, implica $x_{A,1} = \max(x_1)$ e $x_{B,1} = \min(x_1)$).

Per quanto riguarda il confronto tra *m* e *z* possiamo definire:

$$K_{m|z} = \frac{K_{m|w}}{K_{z|w}} = \frac{\frac{MAD_2}{s.q.m._2}}{\frac{MAD_1}{s.q.m._1}}$$

Si può notare che tale coefficiente è legato, oltre che alla presenza e all'entità di un outlier, anche alla dispersione interna alla distribuzione.

Quindi, riassumendo:

1. Se la prima distribuzione è ottenuta dalla seconda introducendo un valore estremo risulterà:

   $$K_{m|z} = \frac{s.q.m._1}{s.q.m._2} > 1$$

   ed anche $K_{m|w} > K_{z|w} > 1$. Inoltre, più è estremo tale outlier più risulta elevato $K_{m|z}$.
2. Viceversa, se la prima distribuzione è ottenuta dalla seconda spostando in una delle code una unità, si avrà $K_{m|z} > 1$ e $K_{z|w} < K_{m|w} = 1$.

E' immediato estendere questo discorso ad un confronto tra graduatorie. Infatti quando per $K_{m|w} > 1$ (oppure per $K_{z|w} > 1$) si ha $P_A^w = P_B^w$ e $P_A^m - P_B^m > 0$ (o $P_A^z - P_B^z > 0$) l'unità *A* precederà l'unità *B* nella classifica discendente basata su *m* (su *z*). Indicando con Rw_i, Rm_i ed Rz_i i ranghi della *i*-esima provincia nelle classifiche basate rispettivamente su *w*, *m*, *z*, risulterà $Rm_A < Rw_A$ ($Rz_A < Rw_A$) oppure $Rm_B > Rw_B$ ($Rz_B > Rw_B$) o entrambe. Infine, se ad elevate differenze $P_A^m - P_B^m$ corrispondono significative differenze di graduatoria, i maggiori "salti" $Rm_A - Rw_A$ oppure $Rm_B - Rw_B$ tra le due graduatorie si verificheranno quando l'unità *A* e/o l'unità *B* assumono valori estremi (ovvero $x_{A1} = \max(x_1)$ e/o $x_{B1} = \min(x_1)$).

Passando al confronto tra le classifiche della Q.d.V., studiamo le differenze tra coppie di graduatorie prendendo in considerazione solo le

differenze non inferiori a 10. Tali valori, ordinati rispetto alle differenze $Rm_i - Rw_i$ sono riportati in tabella 1.

Tabella 1 – Differenze tra coppie di graduatorie basate su w,m,z

Provincia	***Rw***	***Rz***	***Rm***	***Rm-Rw***	***Rm-Rz***	***Rz-Rw***
Aosta	25	6	1	***-24***	-5	-19
Vercelli	51	44	27	***-24***	-17	-7
Udine	28	25	8	***-20***	-17	-3
Sondrio	57	60	39	-18	***-21***	3
Savona	39	32	26	***-13***	-6	-7
Cuneo	22	22	10	***-12***	***-12***	0
Verb.Cus.Oss.	69	66	57	***-12***	-9	-3
Gorizia	17	12	6	***-11***	-6	-5
Livorno	23	19	12	***-11***	-7	-4
Teramo	43	41	32	***-11***	-9	-2
Pordenone	59	61	50	-9	***-11***	2
Bolzano	16	23	9	-7	***-14***	7
Catania	91	93	101	***10***	8	2
Varese	37	40	47	***10***	7	3
Parma	12	16	25	***13***	9	4
Latina	35	38	49	***14***	11	3
Caserta	85	85	100	***15***	***15***	0
Reggio Emilia	6	8	21	***15***	13	2
Rimini	55	59	72	***17***	13	4
Torino	53	53	71	***18***	***18***	0
Bologna	20	24	43	***23***	19	4
Prato	32	35	55	***23***	20	3
Pistoia	64	76	98	***34***	22	12
Roma	29	29	66	***37***	***37***	0
Milano	15	18	74	***59***	56	3

Nostre elaborazioni dei dati pubblicati sul sito www.ilsole24ore.com

Possiamo notare innanzitutto che le massime differenze $Rm_i - Rw_i$ sono positive. Ciò significa che nel complesso prevalgono le situazioni in cui la trasformata *m* esercita un effetto di penalizzazione degli outliers. Ciò è confermato anche dall'analisi preliminare delle 36 variabili: gli outliers sono principalmente valori legati a pessime performances piuttosto che a performances eccellenti.

Le maggiori differenze ($| Rm_i - Rw_i | > 20$) tra queste due graduatorie si verificano con salti positivi per: Milano (59), Roma (37), Pistoia (34), Prato (23) e Bologna (23); con salti negativi per Aosta (24) e Vercelli (24). Nel primo caso si tratta di province i cui valori estremi (corrispondenti a cattive performances) vengono penalizzati dalla trasformata *m*; nel secondo caso si tratta di province con performances buone che la trasformata *m* esalta.

Si può anche notare che – con l'eccezione di Sondrio, Pordenone e Bolzano – per tutte le province che saltano di almeno 10 posizioni tra una classifica e l'altra, il rango che occupano con la trasformata *z* è sempre intermedio al rango che occupano con *m* e con *w*, e che nell'insieme la trasformata *z* presenta un comportamento più simile a *w*, ossia con salti minori, che non a *m*.

Nei termini dei coefficienti *K*, calcolati ora mettendo a confronto una specifica distribuzione con una generica distribuzione regolare, ciò significa che per le distribuzioni in cui $K_{m|w}$ e $K_{z|w}$ risultano massimi (in altre parole per le distribuzioni responsabili dei maggiori salti tra la classifica basata su *w* e le altre) l'entità degli outliers è tale che anche $K_{m|z}$ risulta particolarmente elevato.

Può essere interessante citare come esempio i casi di Pistoia e di Aosta. Pistoia presenta la peggiore performance (pari a 3.35 volte quella dell'unità che la precede) sulla variabile " Indice di variazione del rapporto tra i delitti 2002 e il valore atteso in base al trend 1998-2002", la quale è anche la variabile con i più bassi rapporti *MAD/range* e s.*q.m./range* (ovvero la variabile per cui $K_{m|w}$ e $K_{z|w}$ risultano massimi). Conseguentemente, si osservano significative differenze (positive) $Rm_i - Rw_i$ e $Rz_i - Rw_i$. Tuttavia, essendo paricolarmente basso anche il rapporto *MAD/s.q.m.*, e' significativa anche la differenza $Rm_i - Rz_i$. La situazione di Aosta risulta speculare: presenta il più alto tasso di immatricolazione di auto nuove in rapporto alla popolazione maggiorenne (pari a 206,43!). E' evidente che si tratta di un valore anomalo, se non addirittura di un valore errato. Per la variabile "Immatricolazione di auto nuove in rapporto alla popolazione maggiorenne" i rapporti *MAD/range*, *s.q.m./range* e *MAD/s.q.m.*

risultano tra i più bassi e conseguentemente si hanno significative differenze (negative) $Rm_i - Rw_i$, $Rz_i - Rw_i$ e $Rm_i - Rz_i$.

5. Confronto con l'analisi de Il sole 24 ore.

Per valutare separatamente l'effetto della trasformata *y* dall'effetto della trasformata *t*, abbiamo creato una graduatoria basata sulla trasformata *t* da confrontare sia con la graduatoria de Il Sole24 ore, sia con le classifiche ottenute dalle trasformate lineari *w*, *z* e *m*. Indicando con *Rs* la graduatoria ottenuta da Il Sole24ore e con *Rt* la graduatoria ottenuta a partire dalla trasformata *t*, possiamo scrivere:

$$(Rw - Rs) = (Rt - Rs) + (Rw - Rt).$$

Le differenze *Rt – Rs* saranno da imputare esclusivamente al trattamento non lineare delle variabili correlate negativamente, in particolare ai diversi effetti della trasformata *y* rispetto alla trasformata; mentre le differenze tra *Rw - Rt* saranno da imputare ai diversi effetti delle trasformate *t* e *w*. Chiamiamo questi due effetti, rispettivamente, *effetto non- linearità* ed *effetto t*.

Calcolando le varianze di ciascuna di queste differenze:

$$var(Rw - Rs) = 189{,}47$$
$$var(Rw - Rt) = 25{,}39$$
$$var(Rt - Rs) = 194{,}45$$

possiamo innanzi tutto osservare che $var(Rt - Rs) > var(Rw - Rs)$. Ciò comporta che l'effetto non-linearità viene in parte compensato dall'effetto *t*. In altri termini, l'interazione tra effetto non-linearità ed effetto *t* determina un *effetto assorbimento*. Inoltre, analizzando i coefficienti di correlazione:

$$corr((Rw - Rs), (Rw - Rt)) = 0{,}147$$
$$corr((Rw - Rs), (Rt - Rs)) = 0{,}934$$

notiamo che l'effetto non-linearità spiega, da solo, l'86% della variabilità di (*Rw – Rs*); mentre l'effetto *t* risulta perfino più trascurabile dell'effetto assorbimento:

$$corr((Rt - Rs), (Rw - Rt)) = -0.216$$

Per quanto riguarda la graduatoria *Rz* si ha una situazione assai simile, con:

$$var(Rz - Rs) = 190{,}06$$
$$var(Rz - Rt) = 31{,}14$$

$$corr((Rz - Rs), (Rz - Rt)) = 0{,}174$$
$$corr((Rz - Rs), (Rt - Rs)) = 0{,}919$$
$$corr((Rt - Rs),(Rz - Rt)) = -0{,}228.$$

Viceversa, per quanto riguarda la classifica basata su *m* risultano:

$$var(Rm - Rs) = 289{,}86$$
$$var(Rm - Rt) = 95{,}43.$$

E inoltre:

$$corr((Rm - Rs), (Rm - Rt)) = 0{,}574$$
$$corr((Rm - Rs), (Rt - Rs)) = 0{,}819$$
$$corr((Rt - Rs), (Rm - Rt)) = -0{,}00007$$

Pertanto sono presenti e non trascurabili sia l'effetto non-linearità che l'effetto *t*, i quali, in assenza di effetto assorbimento si sommano. La conseguenza di tutto ciò è che la graduatoria *Rm* sarà la più distante dalla graduatoria *Rs*:

$$var(Rm - Rs) = 289{,}86 > var(Rt - Rs) > var(Rz - Rs) > var(Rw - Rs)$$

Per verificare tutto ciò, riportiamo in tabella 2 le province per le quali si osservano le maggiori differenze (non inferiori a 15) tra la classifica de Il Sole 24 ore, la classifica basata su *t*, e le classifiche basate su *w, z, m*. Tali differenze sono state ordinate rispetto a $Rt_i - Rs_i$.

Tabella 2 –Differenze tra coppie di graduatorie basate su s, t, w, m, z

Provincia	***Rs***	***Rt***	***Rw***	***Rz***	***Rm***	***Rt-Rs***	***Rw-Rs***	***Rz-Rs***	***Rm-Rs***
Ascoli P.	61	34	40	45	40	-27	-21	-16	-21
Lodi	59	36	38	36	31	-23	-21	-23	-28
Pescara	60	37	41	37	36	-23	-19	-23	-24
Terni	64	42	46	47	42	-22	-18	-17	-22
Chieti	50	28	31	39	37	-22	-19	-11	-13
Frosinone	83	64	67	67	63	-19	-16	-16	-20
Perugia	65	46	49	48	46	-19	-16	-17	-19
Lucca	40	26	24	28	29	-14	-16	-12	-11
Mantova	26	12	8	9	13	-14	-18	-17	-13
Viterbo	81	70	68	68	64	-11	-13	-13	-17
Pisa	34	23	19	21	18	-11	-15	-13	-16
Parma	27	16	12	16	25	-11	-15	-11	-2
Latina	58	47	35	38	49	-11	-23	-20	-9
Sassari	70	62	54	52	48	-8	-16	-18	-22
Livorno	36	30	23	19	12	-6	-13	-17	-24
La Spezia	54	49	45	46	38	-5	-9	-8	-16
Pistoia	68	65	64	76	98	-3	-4	8	30
Padova	39	39	48	50	54	0	9	11	15
Savona	44	50	39	32	26	6	-5	-12	-18
Prato	32	40	32	35	55	8	0	3	23
Caserta	76	85	85	85	100	9	9	9	24
Torino	47	58	53	53	71	11	6	6	24
Bolzano	5	17	16	23	9	12	11	18	4
Bologna	6	19	20	24	43	13	14	18	37
Aosta	7	21	25	6	1	14	18	-1	-6
Campobasso	67	81	84	82	75	14	17	15	8
Rieti	41	57	52	55	52	16	11	14	11
Genova	42	61	60	58	62	19	18	16	20
Enna	79	99	102	102	95	20	23	23	16
Isernia	62	83	87	89	86	21	25	27	24
Crotone	72	94	93	95	93	22	21	23	21
Milano	2	24	15	18	74	22	13	16	72
Rimini	38	63	55	59	72	25	17	21	34
Potenza	52	79	79	80	78	27	27	28	26
Oristano	49	77	76	74	68	28	27	25	19
Trieste	28	56	62	56	56	28	34	28	28
Belluno	24	55	61	65	58	31	37	41	34
Napoli	69	103	103	103	103	34	34	34	34
Roma	11	52	29	29	66	41	18	18	55
Sondrio	12	54	57	60	39	42	45	48	27

Nostre elaborazioni dei dati pubblicati sul sito www.ilsole24ore.com

Andiamo a studiare innanzitutto come si manifesta l'effetto non-linearità. Osservando la colonna ($Rt - Rs$), notiamo subito che le prevalgono, sia in termini di intensità che di frequenza, le differenze positive. Ovvero l'effetto non-linearità si esplica prevalentemente come sopravvalutazione (rispetto alla trasformata t^-) delle prestazioni migliori e/o schiacciamento delle differenze tra prestazioni peggiori. Infatti le 14 province che presentano differenze, sono o province le cui buone performances su qualche variabile risultano sopravvalutate dalla trasformata y, e/o province le cui cattive performances sono poco penalizzate dalla trasformata y. Viceversa, le 7 province per le quali risulta $(Rt - Rs) < -15$ sono province che non presentano comportamenti estremi; ciò è anche confermato dal fatto che nelle graduatorie Rw, Rz ed Rm occupano posizioni non molto dissimili. L'effetto che su di esse esercita la trasformata y è un effetto di "confusione" dovuto al maggiore o minore schiacciamento di alcune distribuzioni.

Per quanto riguarda l'effetto t, esso e' pressoché trascurabile sia nei confronti della trasformata w che della trasformata z: sono infatti trascurabili (con l'eccezione di Roma) le differenze $Rw_i - Rt_i$ e $Rz_i - Rt_i$. Inoltre, a conferma della presenza di un effetto assorbimento si può notare che a differenze $Rt_i - Rs_i$ positive corrispondono differenze $Rw_i - Rt_i$ negative e viceversa; ovvero l'effetto t va nella direzione di una attenuazione dell'effetto non linearità.

Possiamo infine verificare che la classifica che presenta le differenze più eclatanti con la classifica de Il sole 24 ore è la classifica basata su m. Sulle differenze $Rm - Rs$ la trasformata t agisce in misura maggiore che non sulle differenze $Rw - Rs$ (è infatti $corr\,((Rm - Rs), (Rm - Rt)) = 0{,}57$) e inoltre, essendo nullo l'effetto assorbimento, i due effetti *non-linearità* e *effetto t* risultano additivi. Pertanto, agli effetti appena individuati della trasformata y, si aggiungono le differenze dovute al diverso impatto degli outliers: la trasformata m premia le performances eccezionalmente buone e penalizza quelle pessime; viceversa, la trasformata t tratta gli outliers in maniera asimmetrica: è molto sensibile alle variazioni in $\max(x)$, insensibile a variazioni in $\min(x)$. L'uso combinato della trasformata t^+ e della y porta quindi a premiare le performance ottime e non penalizzano quelle pessime.

Riferimenti bibliografici

Lauro C. (2003), Il benessere sociale ed economico delle province italiane: una rielaborazione ed una interpretazione alternativa dei dati de Il Sole 24 Ore, Il denaro , 4 gennaio 2003.

Lauro C. (2003), La guerra degli indicatori: i risultati a confronto, Il denaro, 14 gennaio 2003.

Lauro C. (2003), La guerra delle graduatorie, Il denaro 15 gennaio 2003.

Aiello F., Attanasio M. (2004), How to transform a batch of simple indicators to make up a unique one? Atti XLII Riunione Scientifica SIS (Sessioni specializzate), CLEUP, Padova, p.327-338.

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

Cadeo R. (a cura di) Dossier sulla Qualità della vita 2003, Il Sole 24 ore, 22 dicembre 2003.

D'Esposito M.R., Ragozini G. (2004), Multivariate ordering in Performance Analysis. Atti XLII Riunione Scientifica SIS (Sessioni spontanee), CLEUP, Padova, p.51-54.

Pagnotta S.M. (2003), Una generalizzazione dei ranghi per standardizzare i dati Quaderni di statistica vol. 5 p.49-64.

Mori C.(a cura di) Rapporto 2000 sulla qualità della vita in Italia, Italia Oggi documenti, 2 gennaio 2001.