UNIVERSITÀ POLITECNICA DELLE MARCHE

FACOLTÀ DI ECONOMIA "GIORGIO FUÁ"

Corso di Laurea Magistrale in
Data Science per l'Economia e le Imprese
LM-56 e LM-91

Indici composti e trasformazione di dati.

Composite indicators and data trasformation

Relatrice

Prof.ssa Recchioni Maria Cristina

Candidato

Piccione Mirco

Correlatrice

Dr.ssa Ciommi Mariateresa

Anno Accademico 2021-2022

Alla mia compagna Federica per avermi sempre supportato in ogni cosa. A mio nipote Alessandro, sempre nei miei pensieri a cui auguro che un giorno possa seguire i propri sogni.

Indice

Elenco delle figure

Elenco delle tabelle

Premessa

"Se torturi i dati abbastanza, alla fine confesseranno quello che vuoi".

Darrell Huff.

Prima di iniziare ad esporre questo progetto di Laurea Magistrale in *data science*, ci tengo a raccontare che, sin da quando mi sono iscritto a questo corso di Laurea, spesso mi è capitato di dover spiegare che tipo di figura fosse il *Data Scientist*.

Il *data scientist* racchiude diverse discipline, tra le quali statistica, metodi scientifici, intelligenza artificiale (AI) e analisi dei dati per estrarre valore da essi. In questa figura, dunque, si combinano un'ampia gamma di competenze disciplinari al fine di analizzare i dati raccolti dal Web, dagli smartphone, dai clienti, dai sensori[1] e da altre fonti per ottenere *insight* [2] utili (ovvero: immagini, dati gps, dati grezzi,

[1] I sensori IoT comunicano tra loro attraverso un insieme di tecnologie wireless. Ne scaturisce una grande mole di dati da elaborare e memorizzare.

[2] Traduzione in italiano: Intuizione. Riuscire a cogliere un *insight* significa capire cosa il cliente cerca. Questo porta, di conseguenza, a capire come offrire quel prodotto o servizio nel modo più accattivante e convincente.

ecc). La mansione del *data science* comprende la preparazione dei dati per l'analisi, inclusa la pulizia, l'aggregazione e la manipolazione dei dati per eseguire analisi avanzate dei dati [3].

Il *data scientist*, con lo studio dei dati, crea dei modelli utili ai manager aziendali per ottenere informazioni nascoste[4] e/o aiutarli nelle loro decisioni strategiche.

Le aziende al giorno d'oggi raccolgono grandi quantità di dati, conservati nei *data warehouse* (magazzino dei dati [5]). I dati raccolti e conservati possono offrire vantaggi in termini di trasformazione ad aziende in tutto il mondo, ma solo se siamo in grado di interpretarli.

Il *data scientist* quindi oltre a saper manipolare i dati è anche in grado di poterli interpretare. Lo studio dei dati mostra come i trend e lo studio degli *insight* crei alle aziende un nuovo strumento che gli consente un utilizzo più oggettivo nel prendere decisioni più mirate e creare prodotti e servizi più innovativi.

Uno dei numerosi vantaggi aziendali per lo studio dei dati è la creazione dei modelli di *machine learning* (ML), ovvero, l'algoritmo che viene creato è in grado di apprendere dalla grande quantità di dati con cui vengono alimentati. I dati

[3]Chiamata fase di *Pre-processing*, serve per aggiungere valori mancanti, aggiustamento di dati rumorosi, rimuovere eventuali outliers, integrazione di dataset, ecc.

[4]Chiamato processo di *Data Mining*, prevede l'insieme di tecniche e metodologie che hanno per oggetto l'estrazione di informazioni utili da grandi quantità di dati

[5]Traduzione dall'inglese *Data Warehouse* è una grande raccolta di dati aziendali per aiutare un'organizzazione a prendere decisioni

costituiscono la base dell'innovazione, ma il loro valore deriva dalle informazioni che i *data science* possono ottenere e in base alle quali possono agire.

Introduzione

Nel corso di questa tesi verrà spiegato che cos'è un indice composito o anche chiamato indice sintetico, verranno evidenziati i suoi punti di forza e di debolezza, nonché le critiche proposte dalla letteratura statistica.

Per la parte empirica di questa tesi è stato utilizzato il software statistico R[1], mentre il *dataset* deriva dal report annuale sull'*Human development index*, l'indice di sviluppo umano conosciuto anche con l'acronimo *HDI* effettuato dall'Organizzazione delle Nazioni Unite[2] per i paesi di tutto il mondo. I dati sono resi pubblici dallo stesso ente e si possono scaricare sul loro sito[3].

[1] R è un linguaggio di programmazione specifico per la statistica e la grafica computazionali, sostenuto dalla R Core Team e dalla R Foundation for Statistical RStudio è una versione più elaborata della console R

[2] United Nations Development Programme, "2022 Special Report on Human Security", *UNDP United Nations Development Programme*, 2022

[3] "undp-data center" https://hdr.undp.org/data-center/human-development-index#/indicies/HDI

Il punto di partenza di questo studio è stata l'analisi effettuata dall'ONU attraverso l'utilizzo di un indice composito[4]. È bene sapere che per poter calcolare un indice composito bisogna agire con corretti sistemi metodologici in grado di manipolare i dati al fine di normalizzare le variabili e aggregarle in modo opportuno.

Nel corso di questa tesi si riporterà al lettore le fasi stabilite dall'OECD per la creazione dell'indice composito[5]; tuttavia, l'attenzione principale sarà sulle normalizzazioni e aggregazioni delle variabili scelte per la creazione dell'indice sintetico.

L'indice composito è in grado di riassumere le variabili scelte e permette di essere facilmente compreso consentendo così di catturare l'attenzione del pubblico e dei media.

Le motivazioni che hanno spinto lo scrivente allo studio ed alla stesura della tesi sono state la complessità nel creare un indice composito robusto.

Due anni di università magistrale in *data science* e la letteratura statistica sono state la base per la stesura di questa tesi il cui obbiettivo è quello di mostrare al lettore un modo robusto nella definizione e poi nel calcolo di un indice composito.

L'analisi dei dati sarà svolta comparando diversi tipi di normalizzazione: dall'indice di Mazziotta-Pareto, fino alla normalizzazione da noi proposta attraverso la

[4] "OECD" https://www.oecd.org/sdd/42495745.pdf

[5] ibid.

media aritmetica, geometrica e quadratica.

La tesi si compone di quattro capitoli: nel primo viene spiegato che cos'è un indice composito, nel corso di questo capitolo si cercherà di dare una visione chiara sulla sua definizione, su come costruirlo, i vantaggi e gli svantaggi di un indice composito e i problemi aperti ancora discussi dalla letteratura statistica, in questo modo si cercherà di favorire la lettura ad un pubblico meno esperto.

Nel secondo capitolo si introdurranno i metodi di normalizzazione, si spiegherà nel dettaglio le normalizzazioni degli z-scores, del metodo Min-Max e del metodo Min-Max vincolato, il metodo dei numeri indici, della logistica e quello percentuale. Si affronteranno inoltre, la normalizzazione di Box-Cox e le aggregazioni delle variabili. Per concludere il secondo capitolo si spiegherà in modo teorico la proposta che si è adottata nello studio del codice.

Nel terzo capitolo si parlerà nel dettaglio dell'*Human Develompment Index*, della sua storia e le modifiche apportate dall'UNPD su come migliorare l'indice HDI. Infine, si riporteranno i risultati effettuati attraverso il nostro codice.

Il quarto capitolo è composto dalla conclusione di questa tesi.

Capitolo 1

Gli indicatori compositi

1.1 Che cos'è un indicatore composito

Secondo la definizione dell'ISTAT[1] *"Un indice composito (o indice sintetico) è una combinazione matematica (o aggregazione) di un insieme di indicatori elementari (variabili) che rappresentano le diverse componenti di un concetto multidimensionale da misurare (per es., sviluppo, qualità della vita, benessere, ecc.)."* Un indice composto in letteratura viene definito come una sintesi di tutti gli indici elementari. La metodologia riguardante la sintesi si è affermata negli ultimi anni, per via di un quantitativo di dati, sempre maggiore, con l'aumento dei dati, aumenta in modo esponenziale anche il numero di osservazioni statistiche. Le osservazioni statistiche non sono altro che una manifestazione oggetto di studio e possono essere molteplici, come ad esempio il lancio di una moneta, la temperatura, ecc.

[1]https://www4.istat.it/it/strumenti/metodi-e-strumenti-it

Tutti i dati raccolti possono essere rappresentati in forma tabellare. Ogni osservazione statistica è contenuta dentro una riga di questa tabella, e le colonne sono le variabili.

Un indice sintetico si compone di vari indici semplici: questi vengono "costruiti" durante il raccoglimento e l'elaborazione dei dati, secondo la letteratura si passa da una trasformazione di un dato chiamato "grezzo" ad un dato che costituisce un riferimento, chiamato, "feautures" o "osservazione statistica".

Per creare un indice sintetico esistono delle regole che permettono di riflettere al meglio il quadro teorico del dataset, che consente di evidenziare, (pesare) gli indicatori semplici e combinarli in modo da mantenere la struttura o le dimensioni originarie; spesso è una costruzione molto complessa che richiede delle scelte metodologiche da prendere.

Le scelte da considerare sono molteplici: bisogna porsi l'obiettivo del fenomeno da misurare inoltre deve risultare chiaro ciò che si intende misurare e si deve sempre fare riferimento al quadro teorico del dataset, senza stravolgere i gruppi delle osservazioni da cui è composto. Come prima cosa occorre stabilire se l'indicatore sintetico sia di tipo riflessivo [2] o formativo[3].

[2]L'indicatore viene visualizzato come un "effetto" dell'obiettivo posto come misurazione, dove un cambiamento dell'indicatore semplice costituisce un cambiamento dell'indicatore composito.

[3]Se l'obiettivo da misurare viene visualizzato come "causa", in questo caso un cambiamento dell'indicatore semplice, diversamente da quanto scritto sopra, non compromette il cambiamento dell'indicatore composito

Trovato l'obiettivo, occorre selezionare gli indicatori semplici e stabilirne la loro polarità, dove per polarità si intende il segno positivo o negativo della relazione tra l'indicatore elementare e il fenomeno da misurare. La polarità ricopre quindi un ruolo fondamentale nella creazione dell'indice composito[4]. Gli indicatori semplici che andranno a costituire l'indicatore sintetico dovranno essere presi in base alla loro validità, tempestività, rilevanza, ecc.

Diventa importante quindi scartare gli indicatori che potrebbero risultare ridondanti, ovvero che ripetono la stessa informazione, ed evitare di scartare quelle variabili che contengono una informazione essenziale ai fini del calcolo dell'indice composito (in questo caso l'indicatore composito perderebbe di rilevanza in quanto avremmo scartato un indicatore con un elevato quantitativo di informazione).[5]

Il numero di indicatori compositi (CI) esistenti cresce di anno in anno, testimonianza di un interesse sempre maggiore. Per avere una misura di questo interesse basta vedere il numero di articoli e pubblicazioni presenti su *Google Scholar*: dal periodo 2000 al periodo 2010 sono presenti circa 26.700 risultati di ricerche scientifiche, mentre dal 2011 al 2022 sono presenti circa 43.300 risultati quasi il doppio.

---

[4]"Indice composito" https://www4.istat.it/it/strumenti/metodi-e-strumenti-it/analisi

[5]ibid.

1.2 Come si costruisce un indicatore composito

Per calcolare un indice composito il manuale dell'OECD (Organizzazione per la cooperazione e lo sviluppo economico) suggerisce 10 passaggi step che lo stesso ente organizzativo definisce come *sequenza ideale*[6].

Per la creazione di un indice composito è importante sia il suo sviluppo teorico e sia la sua coerenza in tutto il processo di creazione. La coerenza ne diviene il punto centrale poiché ogni scelta di modifica determina un effetto a catena, tale che la modifica di un processo per la costruzione dell'indicatore andrà a modificare il risultato dell'indicatore composito stesso. Diviene così importante non solo la scelta metodologica ma anche il suo adattamento identificativo rispetto al dataset originale.

L'OECD definisce come primo step per calcolare un indice composito la definizione del quadro teorico. Questo è la base di partenza per la selezione e la combinazione delle variabili che si vogliono utilizzare per l'indicatore composito e fornisce una visione di tutta la procedura da effettuare.

Definito il quadro teorico si passa alla vera e propria selezione dei dati, essa si baserà sulla misurabilità degli indicatori scelti comprendendone tutta la parte analitica e la loro pertinenza al fenomeno scelto da analizzare. In questo step vengono messi in discussione la qualità, i punti di forza e di debolezza di ogni indicatore

[6] "OECD" (v. nota 4),pag.19

scelto per la composizione dell'indicatore composito.

Come passo successivo si effettua lo studio sulle osservazioni che si trovano all'interno degli indicatori selezionati, in particolare è necessario possedere un dataset completo, ovvero privo di valori mancanti. In caso contrario è opportuno attuare le politiche di *data mining* per sostituire i dati anomali o rimpiazzare quelli mancanti.

Il quarto step si basa sull'analisi multivariata, come cluster e metodologie similari. Questo step ha lo scopo di valutare l'intero set di dati, nel particolare la sua idoneità e getta le basi per le future aggregazioni.

La normalizzazione si trova a metà del percorso sulla costruzione dell'indice composito, questa fase ha lo scopo di "purificare" le osservazioni rendendole equiparabili. In questa fase inoltre vengono analizzati e trattati gli *outliers*.

Come sesto step abbiamo la ponderazione e la conseguente aggregazione degli indici scelti. Queste vengono definite in base al quadro teorico scelto in partenza ed in base all'analisi multivariata effettuate nel quarto step.

Al settimo step avviene l'analisi di incertezza e della sensibilità dell'indicatore composito appena creato. In questa fase si studia lo schema di normalizzazione precedentemente adottato, la possibile scelta dei cluster adeguata, la possibile rimozione o sostituzione adottata dei dati mancanti. È in questo passaggio che si cerca di identificare tutte le possibili incertezze dell'indicatore composito, e si conducono

le analisi di inferenza per determinare le influenze maggiori che portano l'incertezza dell'indice.

Nello step successivo, l'ottavo, si effettua un ritorno ai dati originari. Questo punto si rende necessario per poter ottenere una visione complessiva, e rappresenta un punto fondamentale e molto discusso della letteratura statistica cioè la trasparenza.

Il nono step consiste nel collegamento dell'indicatore composito appena costruito con altri indicatori; attraverso analisi di correlazione al fine di identificare eventuali informazioni nascoste.

Infine, come ultimo step viene effettuata la valutazione dei risultati, in modo che l'indicatore composito venga migliorato per la presentazione al pubblico: l'indicatore composito, infatti deve risultare di facile lettura ad un ampio pubblico. Risulta quindi opportuno la selezione di tutte le tecniche di visualizzazione che aiutino a rendere chiara e precisa la comunicazione delle informazioni sintetizzate dall'indice composito.

1.2.1 Pro e contro di un indicatore composito

L'indicatore composito non è esente da critiche, e ad oggi, non esiste un punto di incontro tra gli studiosi a favore della costruzione dell'indice composito e gli studiosi che si dichiarano contrari al loro utilizzo.

In questo paragrafo riporteremo i vantaggi e gli svantaggi per la creazione di un indice composito.

Gli indicatori compositi sono molto simili ai modelli matematici o computazionali. Come tale, la loro costruzione deve più alla maestria del modellatore che alle regole scientifiche universalmente accettate per la codifica[7].

I vantaggi, secondo gli studiosi, riguardanti l'indicatore composito, risultano essere la sua valutazione ad intervalli regolari, infatti, aiuta a comprendere il cambiamento nel tempo, è inoltre un ottimo strumento per le analisi economico-politiche poiché aiuta nella comprensione delle tendenze o nel monitoraggio delle prestazioni. In questo modo si viene a creare una scuola di pensiero pro-aggregazioni, in quanto un singolo numero può catturare più facilmente, l'attenzione dei media e dei *policy makers*.

[7] "OECD" (v. nota 4), pag.14

Gli svantaggi, secondo gli studiosi, risultano essere focalizzati principalmente sulla metodologia adottata per la costruzione. Non ci si chiede se l'indice sia buono o cattivo, piuttosto l'obiezione risiede nelle scelte arbitrarie riguardanti la metodologia utilizzata per la costruzione dell'indice composito.

Per una maggiore chiarezza, di seguito si riporta un riassunto riguardante i vantaggi e gli svantaggi di un indice composito.

I vantaggi dell'indicatore composto si possono riassumere in:

- Vengono considerate un grande numero di informazioni.
- Riescono a sintetizzare orizzonti complessi.
- Rendono di facile lettura le informazioni anche a persone non esperte.
- Vengono spesso utilizzati come strumento per i decision makers[8]

[8]Traduzione in italiano, responsabili delle decisioni.

Gli svantaggi dell'indicatore composito si possono riassumere in:

- Il messaggio dell'indice composito potrebbe risultare fuorviante
- La riduzione degli indicatori comporta un rischio di interpretazioni non adeguate.
- La costruzione e la metodologia scelta è totalmente arbitrale.
- La costruzione sempre maggiore di indicatori compositi non è accompagnata da uno sviluppo metodologico.

1.3 Problemi aperti

Nel corso di questo capitolo si è affrontato come secondo la letteratura statistica è possibile costruire un indice composito, descrivendone i suoi vantaggi ed i suoi svantaggi. In questo paragrafo si affronteranno i problemi ancora irrisolti relativi alla costruzione di un indice composito. Si spiegherà con una breve introduzione le problematiche storiche che hanno portato alla costruzione degli indici compositi fino a definire le problematiche dell'oggettività, della soggettività e dei problemi di normalizzazione e aggregazione.

Per creare un indice composito è opportuno saper distinguere l'oggettivo[9] dal soggettivo[10]

Riguardo il concetto di benessere lo studio si è diramato in due diverse direzioni: la prima è l'approccio della vita e la seconda è l'approccio della qualità della vita. La prima direzione, si concentra, infatti, sulle condizioni oggettive, poiché comprende variabili di indicatori sociali da statistiche che rappresentano i fatti sociali. La seconda direzione, invece, tiene conto delle variabili di indicatori soggettive.

Considerando la storia economica, sia la prima che la seconda rivoluzione industriale hanno segnato una differente economia rispetto al periodo precedente.

[9] Un fenomeno può essere considerato "oggettivo" se è osservabile e se vi è un elevato grado di accordo intersoggettivo su ciò che si osserva

[10] Viene utilizzato con riferimento alla definizione di fenomeni

Uno dei primi effetti delle rivoluzioni industriali fù la crescita delle produzioni. Il suo aumento, avvenuto intorno al diciannovesimo secolo, ha creato un bisogno di misurare l'andamento economico, vennero così introdotte delle misure come il "PIL", "percentuali dei lavoratori", "percentuali dei disoccupati", "redditi", ecc. Lo studio di queste quantità economiche non comprendeva però altre variabili che al giorno d'oggi consideriamo importanti, es. "miglioramento della qualità della vita". Solo nel tardo diciannovesimo secolo si è cominciato a dare voce anche ad altri aspetti, quali le "condizioni di lavoro" o: "periodi di riposo","ambiente", "politica", "sanità", "istruzione", ecc..

Intorno agli anni '90 cominciò un dibattito sugli indicatori multidimensionali e gli indicatori unidimensionali[11].

L'evoluzione di questo dibattito ha così portato ad uno dei problemi ancora aperti che riguardano la costruzione dell'indice composito, nello specifico, i problemi ancora irrisolti descrivono un'ampia problematica riguardante la scelta delle normalizzazioni e delle aggregazioni adottate spesso in modo arbitrario, da chi definisce l'indice composito. Entrando nel dettaglio, il dibattito in corso evidenzia la difficoltà di normalizzare e aggregare gli indicatori, poichè la metodologia di costruzione può risultare fuorviante portando così l'indicatore composito a errori interpretativi. Proprio per questo motivo le domande riguardanti l'indicatore composito si focalizzano sulla

[11]La variabile unidimensionale fornisce una sola informazione

trasparenza metodologica applicata, affidabilità e accuratezza. L'obiezione riguardo la costruzione dell'indicatore sintetico è data dal fatto lo sviluppo metodologico adottato non è sempre rigoroso.

Le metodologie applicate, devono tenere conto sia dell'eterogeneità dei dati e sia dell'omogeneità, del numero degli indicatori e dei loro rispettivi pesi.

Altre volte, la sintesi riguarda realtà più complesse e richiede l'applicazione di procedure più complesse, poichè risulta difficile integrare e aggregare i vari indicatori in modo che riflettano i dati e la struttura di partenza. L'aggregazione fa si che si assuma in modo implicito che i vari indicatori elementari siano sostituibili, infatti, sia la normalizzazione che l'aggregazione sono condizionate dal modello teorico stabilito in partenza per la costruzione dell'indice sintetico. Proprio per le problematiche appena evidenziate diventa opportuno fornire informazioni dettagliate riguardo la metodologia utilizzata, componentistica dell'indice sintetico e la sua modalità di interpretazione. Uno degli indicatori più famosi è sicuramente quello relativo all'Indice di Sviluppo Umano (ISU) e comunemente indicato con l'acronimo inglese HDI (*Human Development Index*, proposto nel 1996 dall'economista Mahbub Ul Haq. L'HDI sintetizza aspetti differenti dello sviluppo umano quali: l'aspettativa di vita la vita, l'istruzione e il reddito. Viene calcolato come la media geometrica di tre indicatori normalizzati. Tuttavia anche l'indice HDI è soggetto a limitazioni e critiche, quali: la selezione degli indicatori, la correlazione delle variabili e la

metodologia attuata.

Un altro problema aperto riguarda l'inclusione di indicatori sia oggettivi che soggettivi, la misurazione del benessere e del progresso.

Sono stati quindi proposti indicatori compositi che cercano di tener conto sia delle misure oggettive che soggettive. L'indicatore più conosciuto è "la scala della vita"[12] noto come *"Cantril's Ladder of Life Scale"*[13]

Un secondo indice che prende in considerazione le misure oggettive e soggettive è *"Happy Planet Index"*[14], esso si pone l'obiettivo di misurare il benessere sostenibile[15]

Molti indici vengono costruiti utilizzando indicatori soggettivi tramite dati raccolti dalle varie indagini statistiche sulla popolazione.

Occorre, però porsi la domanda se si sta abusando degli indicatori soggettivi, la letteratura statistica utilizza tali indicatori per facilitarne la chiarezza e catturare l'attenzione dei media, gli indicatori soggettivi vengono anche utilizzati dai politici

[12] La domanda di questo indice fa riferimento a: Immaginate una scala con gradini numerati da zero in basso a dieci in alto. Supponiamo di dire che la parte superiore della scala rappresenta la migliore vita possibile per voi e la parte inferiore della scala rappresenta la peggiore vita possibile per voi. Se il gradino superiore è 10 e il passo inferiore è 0, su quale gradino della scala ti senti personalmente stare al momento attuale?

[13] Koichiro Shiba, Richard G Cowden, Natasha Gonzalez, Matthew T Lee, Tim Lomas, Alden Yuanhong Lai, Tyler J VanderWeele, "Global trends of mean and inequality in multidimensional wellbeing: Analysis of 1.2 million individuals from 162 countries, 2009–2019", *Frontiers in public health*, 2022

[14] Simon Clarke, "Freedom and Happiness: Does Freedom Make People Happy?", *Journal of Political Science: Bulletin of Yerevan University*, 2022

[15] "Quanto soddisfatti i residenti di ogni paese dicono di sentirsi con vita complessiva, su una scala da zero a dieci

per le proprie proposte elettorali. La principale perplessità ricare nella effettiva fattibilità di creare una misura valida per tutti gli stati del mondo.

Si conclude affermando che il benessere è una misura difficile da misurare e che si divide in due macrocategorie: la prima, oggettiva, si misura il benessere con i fenomeni osservabili; la seconda, soggettiva, misura il benessere con i fenomeni non osservabili empiricamente, tali misure vengono effettuate attraverso possibili sondaggi alla popolazione.

Si riassume affermando che i principali problemi aperti nella costruzione di un indicatore composito dove la letteratura statistica è chiamata a rispondere riguardano:

i) la normalizzazione

ii) l'aggregazione.

Il lavoro di tesi, pertanto, si concentrerà su questi due aspetti.

Capitolo 2

Normalizzazione e Aggregazione

Si fornirà una breve definizione di normalizzazione e aggregazione. La normalizzazione in statistica consente di trasformare le variabili in una scala di distribuzione in modo da poter eliminare possibili influenze e consentire un metodo di confronto tra le variabili, tale confronto avviene poichè si elimina l'unità di misura appartenente alle variabili, esistono vari tipi di normalizzazione che spiegheremo nel corso di questa tesi.

A volte conviene effettuare delle aggregazioni tra variabili quando si è in presenza di un *dataset* di grandi dimensioni, può risultare complessa la gestione di questi dati poichè diviene difficile poterli sia interpretarli e sia gestire la quantità di informazione contenuta nelle variabili. L'aggregazione permette di sintetizzare le variabili in modo da poter gestire al meglio sia la sua interpretazione e sia la quantità di informazioni. Diviene infatti più semplice eseguire studi statistici attraverso l'aggregazione delle variabili. Nel corso di questa tesi spiegheremo vari tipi di aggregazione delle variabili.

2.1 I metodi di normalizzazione

Le istituzioni stanno adottando in modo più frequente l'indice composito, infatti, tale indice è in grado di fornire misure multidimensionali es. il benessere, la povertà, ecc. Un esempio di indice multidimensionale del benessere è dato dall'indice composito dello sviluppo umano, denominato con l'acronimo di HDI, l'indice viene creato dal programma di sviluppo umano (UNPD,1990;2022), l'HDI viene creato attraverso una aggregazione di variabili. La procedura per la costruzione dell'indice composito presenta vari problemi, quali la normalizzazione dei dati in modo da rendere le varie osservazioni statistiche comparabili tra loro, la normalizzazione delle variabili non deve essere influenzata dal tempo, occorre che ne risulti indipendente (Mazziotta, Pareto; 2021). L'obiettivo della normalizzazione è quello di modificare i valori delle osservazioni all'interno del set di dati per poter avere una scala comune, senza perdita di informazioni o di distorsioni, proprio per questo motivo la normalizzazione è una tecnica utilizzata spesso per il *Pre-processing* dei dati per il *Machine Learning*.

Spesso conviene normalizzare le variabili quando non si ha la piena conoscenza di come è stato costruito il dataset e spesso risulta opportuno normalizzare le variabili, perché queste possono avere unità di misura differenti, la normalizzazione delle variabili introduce quello che in statistica viene definita una scala comune (ad

esempio il nostro dataset potrebbe avere come dati di partenza una variabile con osservazioni 0-1,una variabile con valori molto bassi, ed una variabile con valori compresi tra 1.000 e 10.000, questa differenza di scala dei numeri causa dei problemi quando l'intenzione è quella di combinare i dati per modellarli in base all'algoritmo necessario).

La scelta della normalizzazione dei dati non diviene complessa soltanto per l'indipendenza dal tempo ma anche perchè nel tempo possono aggiungersi nuove variabili e nuove osservazioni statistiche. Le normalizzazioni più utilizzate sono il metodo Min-Max poichè consente di normalizzare la variabilità degli indicatori, con lo svantaggio di non utilizzare un riferimento comune per confrontare i dati, poichè di anno in anno il punto minimo ed il punto massimo di una variabile può cambiare. Risulta quindi importante la scelta dei parametri da utilizzare per la normalizzazione, tali parametri, definiti anche paletti o valore di riferimento, permettono di confrontare i dati nel tempo (Mazziotta, Pareto; 2021).

Esistono diversi metodi per normalizzare le variabili, di seguito ne vedremo le loro caratteristiche.

2.1.1 Il metodo degli z-scores

Il metodo della standardizzazione dove le osservazioni vengono distribuite secondo la media e la varianza, viene chiamata anche come calcolo degli Z-scores:

$$z_{ij} = \frac{x_{ij} - \bar{x_j}}{\sigma_j}$$

dove la media ($\bar{X}$) e la deviazione standard, rispetto alla variabile j (σ) vengono rispettivamente calcolate in questo modo:

$$\bar{x_j} = \frac{\sum_{i=1}^{n} X_{ij}}{n}$$

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{n} (x_{ij} - \bar{x_j})}{n}}$$

dove:

- z_{ij}= punteggi standardizzati
- x_{ij} = dato grezzo osservato
- $\bar{X}_j$ = media dei dati grezzi osservati, calcolata rispetto alla variabile j
- σ_j = deviazione standard (della popolazione), calcolata rispetto alla variabile j
- n = numero totale delle osservazioni

La standardizzazione viene applicata come trasformazione lineare di un insieme di dati. Il metodo prevede di collegare una variabile aleatoria con media (μ) e varianza (σ^2), ad una variabile aleatoria con distribuzione "standard", ossia di media zero e varianza uno. L'operazione permette di avere una variabile a cui è stata modificata la scala di misurazione, l'osservazione si trova così depurata dall'unità di misura, i valori di questa normalizzazione vengono chiamati punteggi z o valori standardizzati. La peculiarità di detta operazione si riscontra nel fatto che la variabile standardizzata avrà come caratteristiche una μ=0 ed una σ^2=1[1].

Il vantaggio di questo metodo è ottenere una media uguale a zero ed un errore standard uguale ad uno, lo svantaggio risulta essere dalla presenza dei valori negativi e del fatto di non avere un intervallo [a,b] fissato per il valore standardizzato

[1]"OECD" (v. nota 4)

2.1.2 Il metodo Min-Max

Se si vuole confinare il valore normalizzato in un intervallo [0,1] occorre

$$z_{ij} = \frac{x_{ij} - min(x_i)}{[MAX(x_j) - min(x_j)]}$$

Dove: $\min(x_j)$ indica il valore minimo della variabile j e $max(X)$ indica il valore massimo della variabile j.

La normalizzazione max-min riscala le variabili in modo lineare, nell'intervallo [0,1]. Come per il metodo precedente anche attraverso questo tipo di normalizzazione l'indicatore viene purificato dall'unità di misura ed il suo range sarà pari a uno[2].

Questo rappresenta il suo principale vantaggio, lo svantaggio che risulta molto sensibile agli *outliers*[3].

Esiste un secondo metodo Min-Max, denominato metodo Min-Max vincolato, questo metodo utilizza un riferimento comune degli indicatori, tale metodo permette infatti di centrare la normalizzazione, il calcolo del Min-Max vincolato viene effettuato prendendo il valore minimo e massimo che rappresenta l'intervallo massimo dell'indicatore, così facendo le osservazioni anomale sia minime che massime vengono "schiacciate" in base al valore preso di riferimento. Il metodo Min-Max

[2] "OECD" (v. nota 4)

[3] con il termine *outliers* si indicano quelle osservazioni che possiedono un valore anomalo rispetto ai valori riportati all'interno del proprio indice elementare.

vincolato utilizza una scala comune che varia da zero ad uno.

2.1.3 Il metodo numeri indici

$$I_{ij} = \frac{x_{ij}}{x^*_{oj}} 100$$

Dove:

$$x^*_{oj} = \frac{\sum_{i=1}^{n} x_{ij}}{n}$$

oppure:

$$x^*_{oj} = max_i\{x_{ij}\}$$

Nel primo caso i dati vengono normalizzati rispetto al valore medio, nel secondo caso rispetto al valore massimo. La normalizzazione per mano di numeri indici garantisce l'eliminazione dell'unità di misura e di mantenere una certa distanza relativa tra le varie unità. Nel caso in cui il denominatore sia il massimo, si otterranno valori pari o inferiori a cento[4].

Il vantaggio di questo metodo è la preservanza della variabilità originale, lo svantaggio risulta essere molto più sensibile agli *outliers*.

[4] "OECD" (v. nota 4)

2.1.4 Il metodo della logistica

$$z_{ij} = \frac{1}{1 + exp(-x_{ij})}$$

La normalizzazione attraverso la logistica viene utilizzata quando la variabile è di tipo dicotomico, ovvero, le osservazioni assumono soltanto valore zero ed uno, di solito il valore possiede una informazione dove: uno è indicato come successo e il valore pari a zero è indicato come insuccesso[5].

Il suo vantaggio è poter calcolare una normalizzazione di tipo dicotomico, lo svantaggio risiede nella sua procedura complessa.

2.1.5 Il metodo del rango

$$g_{ij} = rank\{x_{ij}\}$$

La normalizzazione attraverso il metodo del rango sostituisce il valore di partenza di ciascuna unità con il rango (numero d'ordine) con cui l'unità è stata inserita nell'elenco secondo il suo indicatore elementare di appartenenza[6].

[5] "OECD" (v. nota 4)

[6] ibid.

Il suo vantaggio è che risulta essere insensibile agli *outliers*, il suo svantaggio è assumere la stessa distanza tra i vari valori.

2.1.6 Il metodo percentuale

Un ulteriore metodo, seppur meno utilizzato è il metodo percentuale:

$$p_{ij} = \frac{x_{ij}}{x^*_{oj}} 100$$

Dove:

$$x^*_{oj} = \sum_{i=1}^{n} x_{ij}$$

Nel metodo di normalizzazione percentuale si sostituisce il dato osservato x_{ij} con la percentuale. La somma dei corrispondenti valori normalizzati dell'indice elementare sarà pari a cento[7].

Il suo vantaggio risiede nel suo ampio range di valori [0-100], lo svantaggio è la sua additività del fenomeno osservato.

[7]"OECD" (v. nota 4)

2.2 Normalizzazione Box-Cox

La normalizzazione denominata Box-Cox, prende il nome dagli studiosi George Box e Sir David Roxbee Cox (1964)[8]. Si tratta di una trasformazione di variabili dipendenti distribuiti in modo non normale in una forma normale[9]. Il metodo box-cox risulta poco usato nella costruzione di un indicatore composito[10] ma di grandi potenzialità. Infatti, si basa su trasformazioni di potenze e un numero ampio di test per assicurare la normalizzazione. L'esponente del metodo Box-Cox è il lambda (λ), dove i suoi valori variano da -5 a +5. Il metodo considera tutti i valori del λ e ne seleziona quello ottimale[11], ovvero, il valore che massimizza la funzione chiamata *Log-likelihood*[12]. Se indichiamo con y il vettore dei dati originali, la sua trasformazione ha la forma:

[8] La trasformazione di Box Cox prende il nome dagli statistici George Box e Sir David Roxbee Cox che collaborarono su un documento del 1964 e svilupparono la tecnica. Questa tecnica è ormai operativamente semplice con l'uso dei PC

[9] Spesso, i dati non sono distribuiti in modo normale

[10] Francesca Mariani, Mariateresa Ciommi, "Aggregating Composite Indicators through the Geometric Mean: A Penalization Approach", *Computation*, 2022

[11] Il "valore ottimale" è quello che si traduce nella migliore approssimazione di una curva di distribuzione normale

[12] Il *log-likelihood* è un metodo conosciuto come metodo della massima verosimiglianza, il suo scopo è quello di determinare uno stimatore attraverso un procedimento matematico

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \lambda \neq 0; \\ \\ \log(y) & \lambda = 0. \end{cases}$$

Questa normalizzazione è definita solo per i dati positivi. Tuttavia, Box e Cox hanno proposto una seconda formula che può essere utilizzata anche in presenza di valori negativi:

$$y(\lambda) = \begin{cases} \frac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1} & \lambda_1 \neq 0; \\ \\ \log(y+\lambda_2) & \lambda_1 = 0. \end{cases}$$

Il metodo di Box-Cox include dei casi speciali in base al valore di λ[25]:

[25]"Normalizzazione Box-Cox" https://www.sixsigmain.it/ebook/Capu13-4.html

λ	$x_i(\lambda)$
$\lambda = -1.0$	$y_i(\lambda) = \frac{1}{x_i}$
$\lambda = -0.5$	$y_i(\lambda) = \frac{1}{\sqrt{x_i}}$
$\lambda = 0$	$y_i(\lambda) = \ln(y_i)$
$\lambda = 0.5$	$y_i(\lambda) = \sqrt{y_i}$
$\lambda = 2.0$	$y_i(\lambda) = y_i^2$

$$L = -\frac{v}{2} * \ln S_t^2 + (\lambda - 1) * \frac{v}{n} \sum \ln (Y)$$

dove:

- L = log-likelihood;
- v = n-1 osservazioni;
- n = numero osservazioni;
- S_t^2 = varianza dei dati già trasformati precedentemente con le formule descritte sopra;
- λ = parametro di stima;
- y_i = dato originale

2.3 Aggregazione delle variabili

L'aggregazione è una combinazione di tutti gli indicatori semplici normalizzati che formano l'indice sintetico. Oltre alla scelta della funzione di aggregazione, spesso si è interessati ad introdurre un peso, che rifletta l'importanza di ciascun indicatore elementare. Si tratta del problema della ponderazione. Esistono vari metodi per poter aggregare le variabili normalizzate, il modo più semplice è quello di assegnare a tutti gli indicatori normalizzati lo stesso peso. Risolto il problema della ponderazione occorre focalizzarsi sulla normalizzazione. In particolare esistono aggregazioni compensative e non compensative: i valori di sintesi compensativi vengono calcolati come valori lineari e sono composti da indicatori elementari considerati sostituibili; i valori di sintesi non compensativi vengono calcolati come non-lineari e sono composti da indicatori elementari considerati non sostituibili. La letteratura statistica evidenzia come ci sia un alto grado di difficoltà nell'aggregare gli indicatori, in quanto ci si espone nel rischio che l'indicatore composito risulti fuorviante o che induca in errore interpretativo, uno dei casi più frequenti di difficoltà nell'aggregazione risiede nelle scelte metodologiche che si riscontra per costruire l'indice composito.

Quello che dovrebbe guidare tutte le fasi della costruzione dell'indice composito è la necessità di bilanciare la richiesta di sintesi con la realtà osservata; non è accettabile

un indice composito che effettua una compressione della realtà in base ai valori assunti da uno o più indicatori.

Nei paragrafi successivi discuteremo alcuni dei metodi di aggregazione più comunemente usati in letteratura.

2.3.1 Media potenziata di ordine r

Le medie potenziate di ordine r, dette anche media di potenze è un'ampia classe di medie che contiene al suo interno la media aritmetica, media geometrica, ecc. In generale tale classe è definita come:

$$M_r = (\frac{1}{n}\sum_{i=1}^{n} x_{ij})^{\frac{1}{r}} = \sqrt[r]{\frac{1}{n}\sum_{i=1}^{n}(x_{ij})^r}$$

Si definiscono momenti di ordine r della variabile statistica le potenze r-esime delle medie potenziate di ordine r:

$$(M_r)^r = \mu_r = \frac{1}{n}\sum_{i=1}^{n}(x_{ij})^r$$

Le proprietà dei momenti di ordine r sono le seguenti:

$$\sum_{i=1}^{n}(x_{ij}^r - \mu_r) = 0 \sum_{i=1}^{n}(x_{ij}^r - \mu_r)^2 = minimo$$

Al variare di r si ottengono differenti medie, in seguito ne vedremo alcune.

La media aritmetica (r=1)

La media aritmetica non è altro che una media potenziata di ordine r=1

La media aritmetica si definisce nel seguente modo:

$$M1 = \bar{X} = \frac{1}{k}\sum_{j=1}^{k}(z_{ij})$$

La media geometrica (r=0)

La media geometrica si definisce come una media potenziata di ordine 0, si riporta di seguito la formula per calcolarla:

$$M0 = \left(\prod_{j=1}^{k} z_{ij}\right)^{\frac{1}{k}} = \sqrt[k]{x_1 x_2 \cdots x_k}$$

$$LogM0 = \frac{1}{n}(n_1 \cdot \log x_{1j} + n_2 \cdot \log x_{2j} + ... + nk \cdot \log xk_j)$$

La media geometrica è calcolabile solo in presenza di valori positivi e questo vale per tutte le medie generalizzate.

La media quadratica (r=2)

Si definisce la media quadratica come una media potenziata di ordine r=2, si calcola nel modo seguente:

$$M2 = \sqrt{\frac{1}{k}\sum_{j=1}^{k}(x_{ij})^2}$$

2.3.2 Il Metodo tassonomico di Wroclaw

Il metodo tassonomico di Wroclaw, si basa sul concetto di "unità ideale", prendendo i migliori valori tra quelli osservati. La sintesi degli indicatori si ottiene mediante il calcolo della distanza euclidea, dove per valori crescenti:

$$T_{ij} = \frac{x_{ij} - \bar{x_j}}{\sigma_j}$$

Per i valori decrescenti:

$$T_{ij} = \frac{x_{ij} - \bar{x_j}}{\sigma_j} \cdot (-1)$$

- $\bar{x_j}$ = media
- σ_j = scostamento quadratico medio

Quando la distanza tra l'osservazione e l'unità ideale è nulla l'indice assume valore 0, l'indice assume valori maggiori di 0 quando l'osservazione e l'unità ideale sono

più distanti, maggiore è la distanza, maggiore risulterà l'indice.

Per ogni unità, si calcola la "distanza", nel modo seguente:

$$D_i = \sqrt{\sum_{j=1}^{k} (T_{ij} - max(T_{ij}))^2}$$

dove:

- $T_{ij} = max_j(T_{ij})$ se l'indicatore ha una polarità positiva
- $T_{ij} = min_j(T_{ij})$ se l'indicatore ha una polarità negativa

L'indice sintetico, per l'unità i, è dato dalla formula:

$$MTW_i = \frac{D_i}{\bar{D_o} + 2S_o}$$

dove:

- $\bar{D_o}$ = media delle distanze
- S_o = scostamento quadratico medio delle distanze

2.4 Il Problema della penalizzazione

2.4.1 La media aritmetica penalizzata: il metodo Mazziotta-Pareto

Il metodo di Mazziotta-Pareto costruito attraverso la media aritmetica penalizzata è un indice alternativo, chiamato anche MPI (Mazziotta-Pareto Index) (Mazziotta,Pareto 2016), ha come punto di partenza la media aritmetica che viene modificata, o meglio, penalizzata per mitigare l'effetto compensativo della media. La penalità ha pertanto lo scopo di "sbilanciare" i valori degli indicatori.

Le fasi per la costruzione dell'MPI sono le seguenti:

- standardizzazione degli indicatori elementari mediante scarti relativi dalla media rispetto allo scostamento quadratico medio (z-score)
- aggregazione degli indicatori mediante media aritmetica con funzione di penalità basata sulla "variabilità orizzontale". La penalità si basa sul coefficiente di variazione.

In dettaglio, siano: Mx_j la media e Sx_j lo scostamento quadratico medio del j-mo indicatore, ovvero:

$$Mx_j = \frac{\sum_{i=1}^{n} x_{ij}}{n};$$

e

$$Sx_j = \sqrt{\frac{\sum_{i=1}^{n} (X_{ij} - Mx_j)^2}{n}}$$

La matrice standardizzata $Z = z_{ij}$ è definita nel seguente modo, se il j-mo indicatore è concordante, ovvero se ha polarità positiva:

$$z_{ij} = 100 + \frac{(X_{ij} - Mx_j)}{Sx_j} \cdot 10$$

Mentre nel caso di polarità negativa si avrà:

$$z_{ij} = 100 - \frac{(X_{ij} - Mx_j)}{Sx_j} \cdot 10$$

Dove l'indice di Mazziotta-Pareto si ottiene mediante la formula:

$$MPI_i \pm = M_{\bar{z_i}} \bar{z_i} \cdot cv\bar{z_i}$$

dove:

$$M_{\bar{z_i}} = \frac{\sum_{h=1}^{m} \bar{z_{ih}}}{m}$$

$$S\bar{z_i} = \sqrt{\frac{\sum_{h=1}^{m} (\bar{z_{ih}} - M\bar{z_i})^2}{m}}$$

$$cv\bar{z_i} = \frac{S_{\bar{z_i}}}{M_{\bar{z_i}}}$$

2.4.2 La media geometrica penalizzata: il metodo Mariani-Ciommi

In analogia con l'interpretazione dell'indice di Mazziotta-Pareto, Mariani e Ciommi (Mariani,Ciommi 2022) forniscono una nuova interpretazione della media geometrica introducendo la media geometrica penalizzata. Come per Mazziotta-Pareto la costruzione ne segue alcuni step:

- stima dei minimi quadrati per i valori dei dati trasformati dalla funzione Box-Cox di ordine uno
- prodotto tra la media geometrica e un fattore di penalizzazione che dipende dalla varianza (orizzontale) degli indicatori normalizzati opportunamente scalati e trasformati tramite la funzione Box-Cox di ordine uno.

Il punto di partenza di questo nuovo indicatore è che l'indicatore media geometrica può essere visto come una stima dei minimi quadrati degli indicatori normalizzati trasformata tramite la funzione Box-Cox di ordine zero,h_0, come segue:

$$u_{0i} = h_0^{-1}(\frac{1}{m}\prod_{j=1}^{m} h0(z_{ij}))$$

dove z_{ij} sono, come al solito, gli indicatori normalizzati. pure h_0 è definita dalla seguente formula:

$$h_0(x) = \ln(x)$$

Pertanto, con questa nuova notazione la media geometrica penalizzata GMi per l'unità i può essere scritta come:

$$GM_i^{\pm} = u0_i \cdot h_0^{-1}(\pm S_{0_i}^2) = u_{0i} \exp \pm S_{0_i}^2$$

dove il segno $\pm sempre legato al concetto di polarite S_{0_i}^2$ viene calcolata nel modo seguente:

$$S_{0_i}^2 = \frac{1}{m} \sum_{j=1}^{m} h_0(\frac{z_{ij}}{u_{0i}})^2 = \frac{1}{m} \sum_{j=1}^{m} (\ln z_{ij} - \ln u_{0i})^2$$

2.5 Una nuova proposta

In questo paragrafo cercheremo di rispondere ad uno dei problemi aperti tra la letteratura statistica riguardante l'indice composito. Nel corso di questa tesi si è scritto di come il cuore del dibattito sulla creazione dell'indice composito risieda sulle scelte della normalizzazione e dell'aggregazione adottate in modo arbitrario da chi definisce l'indice composito. Nello specifico, l'obiettivo iniziale dello studio è stato quello definire un indice composito robusto, ma che richiedono il minor numero di manipolazioni possibili[13], al fine di scegliere la miglior normalizzazione attraverso le informazioni nascoste contenute nei dati. Si è partiti come base di riferimento teorica alle normalizzazioni relative alla standardizzazione z:

$$z_{ij} = \frac{x_{ij} - \bar{x_j}}{\sigma_j}$$

Normalizzazione Min-Max:

$$z_{ij} = \frac{x_{ij} - min(xi)}{[MAX(x_j) - min(x_j)]}$$

[13]Da qui la citazione utilizzata nella premessa da Darrell Huff

Normalizzazione Min-Max vincolato attraverso un parametro h:

$$z_{ijh} = A \cdot x_{ij} + B$$

dove x_{ij} sono le osservazioni grezze e A e B vengono rispettivamente calcolate:

$$A = \frac{h}{(max(x_j) - (min(x_j)))}$$

$$B = A \cdot min(x_j)$$

Per ognuna delle normalizzazioni proposte sono stati calcolati tre indicatori compositi mediante:

- indice con media aritmetica (r=1)
- indice con media geometrica (r=0)
- indice con media quadratica (r=2)

Costruiti i tre indici si è adottata la trasformazione di Box-Cox, attraverso la seguente formula:

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0; \\ \log(y) & \lambda = 0. \end{cases}$$

Il vettore risultante dalla trasformazione di Box-Cox viene successivamente sottoposta al test di Lilliefors[14].

Il λ per le normalizzazioni degli indici con base aritmetica, geometrica e quadratica è stato calcolato attraverso la creazione di un parametro impostato che varia da -6 a +6 con un intervallo di 0.5, il calcolo per ogni indice è stato ripeturo per ogni valore del parametro p, in base al test di Lilliefors è stato scelto il λ migliore.

Come fase finale dello studio sono stati confrontati gli ordinamenti indotti dall'indice, calcolando le differenze tra n posizioni di una stessa unità rispetto ai tre indici.

Il test finale ha prodotto una differenza di posizionamento dei tre indici vicina a zero, di conseguenza l'indice composito finale, calcolato tramite la comparazione dei tre indici ha prodotto un ranking molto più robusto.

[14]Il test di Lilliefors, derivato dal test di Kolmogorov-Smirnov chiamato anche *distance test*,si fonda sulla distanza massima tra la distribuzione cumulata osservata e quella cumulata attesa

2.5.1 L'indice con C0, C10, C100

Al fine di testare la robustezza del metodo, il procedimento teorico descritto nel paragrafo 2.6 è stato utilizzato con tre diversi metodi di normalizzazione:

- il primo metodo denominatoC_0
- il secondo metodo C_10
- il terzo metodo C_100

rispettivamente le normalizzazioni dei tre metodi sono:

$$C_0 = z_{ij} = \frac{x_{ij} - \bar{x_j}}{\sigma_j}$$

$$C_{10} = z_{ij} = \frac{x_{ij} - \bar{x_j}}{\sigma_j} + 10$$

$$C_{100} = z_{ij} = \frac{x_{ij} - \bar{x_j}}{\sigma_j} + 100$$

Nel particolare, durante la normalizzazione dei vettori:

nel primo caso con C_0 non si è applicata nessuna modifica della normalizzazione vettoriale.

Nel secondo caso con C_{10} si è applicata una modifica della normalizzazione vettoriale con una aggiunta di una costante pari a dieci.

Nel terzo caso con C_{100} si è applicata una modifica della normalizzazione vettoriale con una aggiunta di una costante pari a cento.

Infine sono stati costruiti gli indicatori definitivi per media aritmetica, geometrica e quadratica, normalizzati.

Per la media aritmetica:

$$inice_l = \frac{(indexAM^{\lambda} - 1)}{\lambda}$$

$$IndexAM_n = \frac{(index_l - \bar{index_l}}{\sigma^2(index_l)^{0.5}}$$

Per la media geometrica:

$$inice_l = \frac{(indexGM^{\lambda} - 1)}{\lambda}$$

$$IndexGM_n = \frac{(index_l - \bar{index_l}}{\sigma^2(index_l)^{0.5}}$$

Per la media quadratica:

$$inice_l = \frac{(indexQM^{\lambda} - 1)}{\lambda}$$

$$IndexQM_n = \frac{(index_l - \bar{index_l}}{\sigma^2(index_l)^{0.5}}$$

L'operazione è stata effettuata per il calcolo con le tre costanti diverse in base al miglior λ ottenuto tramite la trasformazione di Box-Cox.

Infine è stata effettuata una comparazione del risultato dei tre indici compositi ed è stato costruito il *ranking* finale per ogni tipo diverso di normalizzazione calcolata.

Capitolo 3

Human Develompment Index

Nell'ultimo periodo abbiamo assistito ad un interesse sempre maggiore nel definire indici compositi applicati a concetti quali il benessere.

L'indicatore composito, come si è visto nel primo capitolo, è basato su un modello matematico-statistico multidimensionale del fenomeno che si vuole misurare. Per illustrare i risultati della nostra proposta ci siamo concentrati sull'indice di sviluppo umano (HDI).

Ogni anno l'ONU elabora un report chiamato "Human Development" che si pone l'obiettivo di confrontare lo sviluppo umano dei paesi di tutto il mondo.

Il primo rapporto pubblicato risale al 1990 (UNDP,1990).

3.1 Storia dell'HDI

L'indice di sviluppo umano è stato introdotto nel 1990, attraverso il programma di sviluppo delle Nazioni Unite, conosciuto anche con l'acronimo "UNPD".

l'introduzione dell'indice, nel primo rapporto pubblicato dall'UNPD, ha introdotto un nuovo tipo di misura riguardante la situazione politico-economica.

Esso si pone l'obiettivo progressista nel misurare il benessere dell'uomo per ogni nazione, l'HDI infatti è stato fin da subito un mezzo di confronto tra le nazioni ed i vari governi. Dal 1990 al 1994, la metodologia di costruzione dell'indice ha subito tre cambiamenti: uno dei cambiamenti è stato il cambio del valore minimo e massimo dell'indice, poiché ogni anno i valori presentavano minimi e massimi differenti; quindi, era impossibile paragonare i vecchi indici e la crescita del benessere umano. Dal 1994 in poi, furono fissate delle soglie fisse che consentissero la comparazione dei vari indici HDI calcolati annualmente. Originariamente l'indice HDI veniva calcolato con la media aritmetica, basandosi su tre dimensioni:

- salute
- istruzione
- qualità della vita

Nel 2010 il calcolo dell'indice ha subito una ulteriore modifica, infatti, dal calcolo della media aritmetica si è passati all'uso della media geometrica, in modo

da rispondere a tutte le critiche dovute alla sostituibilità tra le variabili che derivano dall'uso appunto della media aritmetica, inoltre con l'utilizzo della media geometrica se una variabile diminuisce o aumenta, questa garantisce lo stesso aumento o diminuzione nella variabile aggregata.

Il valore massimo raggiungibile dall'indice HDI è uno, il valore pari a uno indica che, quel dato Paese ha raggiunto tutti gli obiettivi. Viceversa, il minimo è zero che si raggiunge quando un Paese ha il valore zero anche in una sola dimensione.

Le principali fonti di dati che l'UNPD utilizza per creare il *dataset* sono:

- UN Population Division - New York[1]
- UNESCO (United Nations Educational, Scientific and Cultural Organization) – Parigi[2]
- World Bank – Washington[3]
- FAO (Food and Agriculture Organization) – Roma[4]
- ILO (International Labor Organization) – Ginevra[5]

[1]"Population Division" https://www.un.org/development/desa/pd/

[2]"UNESCO" https://en.unesco.org/about-us/unesco-house

[3]"World Bank" https://www.worldbank.org/en/home

[4]"FAO" https://www.fao.org/contact-us/en/

[5]"ILO" https://www.ilo.org/global/lang--en/index.htm

- OECD (Organization for Economic Cooperation and Development) - Parigi[6]
- UNFPA (United Nations Fund for Population Activities) - New York[7]
- WHO (World Health Organization) - Ginevra[8]

3.2 I dati

Di seguito un estratto dell'introduzione sul report 2022 *Human Development*

"We live in a world of worry: the ongoing Covid-19 pandemic, war in Ukraine and elsewhere, record-breaking temperatures, fires, and storms. Each is a troubling manifestation of an emerging, new uncertainty complex that is unsettling lives around the world. It is driven by three novel, interacting layers of uncertainty at a global scale: the destabilized planetary systems of the Anthropocene, the pursuit of sweeping societal transformations to ease planetary pressures and widespread, intensifying polarization".[9]

I dati oggetto di studio fanno riferimento al 2022[10]

[6] "OECD" https://www.oecd.org/

[7] "UNFPA" https://www.unfpa.org/

[8] "WHO" https://www.who.int/

[9] estratto dell'introduzione dove spiega il complesso di incertezza che sta sconvolgendo le vite di tutto il mondo, https://report.hdr.undp.org/intro

[10] Si possono scaricare a questo link "undp-data center" (v. nota 3)

I dati grezzi scaricati dal sito UNPD sono rappresentati dalle variabili di seguito esposte:

- HDI rank
- Country
- Human Development Index (HDI)
- Life expectancy at birth
- Expected years of schooling
- Mean years of schooling
- Gross national income (GNI) per capita
- GNI per capita rank minus
- GNI per capita rank minus HDI rank
- HDI rank (anno precedente)

Di seguito un estratto dei dati grezzi scaricati dal sito dell'UNPD[11]:

[11] la tabella completa la si può trovare in appendice B

Country	HDI	LEI	EI	MEI	GNI	GNI HDI rank	HDI rank
Switzerland	0,962	84,0	16,5	13,9	66.933	5	3
Norway	0,961	83,2	18,2	13,0	64.660	6	1
Iceland	0,959	82,7	19,2	13,8	55.782	11	2
Hong Kong, China (SAR)	0,952	85,5	17,3	12,2	62.607	6	4
Australia	0,951	84,5	21,1	12,7	49.238	18	5
Denmark	0,948	81,4	18,7	13,0	60.365	6	5
Sweden	0,947	83,0	19,4	12,6	54.489	9	9
.....	....	...	...	...	...	...	...

Tabella 3.1: Dati grezzi

Fonte: "undp-data center"
https://hdr.undp.org/data-center/human-development-index#/indicies/HDI

3.2.1 Statistiche descrittive

Di seguito illustreremo alcune statistiche descrittive effettuate. Si specifica che essendo delle analisi iniziali per capire i dati, le variabili sono influenzate dagli outliers:

Come si evince nella figura 3.1 l'analisi effettuata attraverso l'istogramma consente di vedere la forma dei dati[12]. Scendendo in dettaglio, l'istogramma evidenzia come per il grafico appartenente alla variabile "aspettativa di vita" il suo centro risieda intorno ai settanta anni di vita degli individui e la sua distribuzione varia da un minimo di cinquant'anni ad un massimo di novant'anni. La disposizione delle barre suggerisce che i dati hanno una distribuzione normale.

variabili	skewness test	p-value skew	kurtosis test	p-value kurt
Life expectancy at birth	-0.2348	0.1751	2.3895	0.02207
Expected years of schooling	-0.021569	0.8997	2.7495	0.555
Mean years of schooling	-0.3665	0.03727	2.0277	1.028e-07
Gross national income	1.9898	2.075e-15	8.7955	1.399e-08

Tabella 3.2: Indici di distribuzione

Fonte: nostra elaborazione

Per quanto riguarda la variabile "aspettativa scolastica", si nota come il suo centro risiede in un range superiore a dieci anni, e poco minore di 15 anni. La sua distribuzione varia tra i cinque anni di aspettativa scolastica ai venti anni, anche in questo caso la disposizione delle barre suggerisce che i dati hanno una

[12] l'asse X (orizzontale) riporta i valori dei dati includendo un range di valori. L'asse Y (verticale) mostra il numero dei dati che hanno il valore compreso nel range della barra

distribuzione normale. Per la variabile "media di anni di scuola" si nota subito come la distribuzione dei dati non sia simmetrica, generanno quindi una asimmetria positiva[13]. Per la variabile "reddito", si nota come anche in questo caso la distribuzione dei dati non sia simmetria, ma inversamente da quanto descritto per la variabile "media scolastica", in questo caso si genera una asimmetria negativa[14].

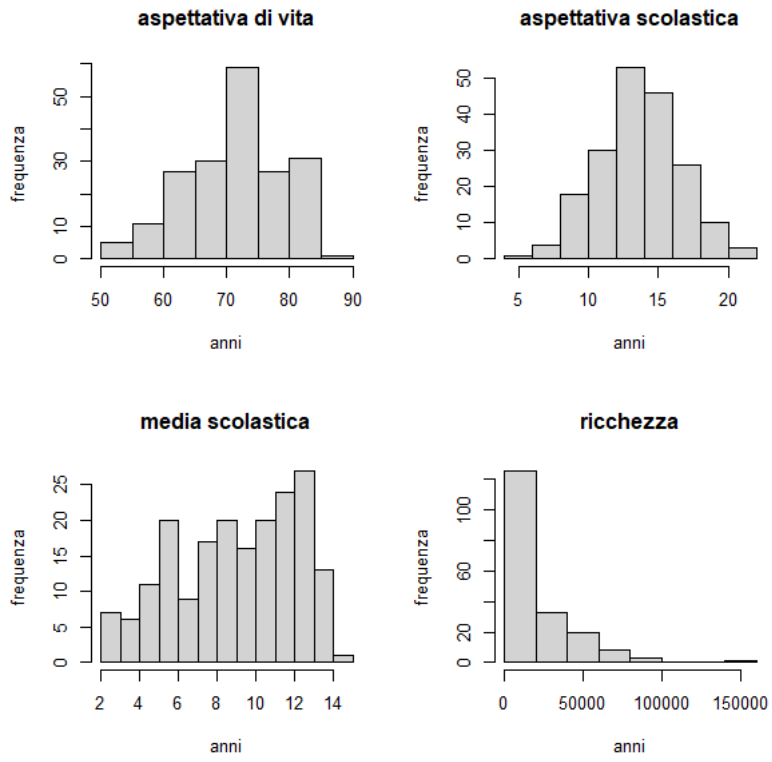

Figura 3.1: Istogramma
Fonte: nostra eleborazione

[13]Si definisce asimmetria positiva quando i dati non risultano simmetrici e il punto più alto ricade verso la parte destra del grafico

[14]Si definisce asimmetria negativa quando i dati non risultano simmetrici e il punto più alto ricade verso la parte sinistra del grafico

Una rappresentazione alternativa di quanto spiegato nel grafico 3.1

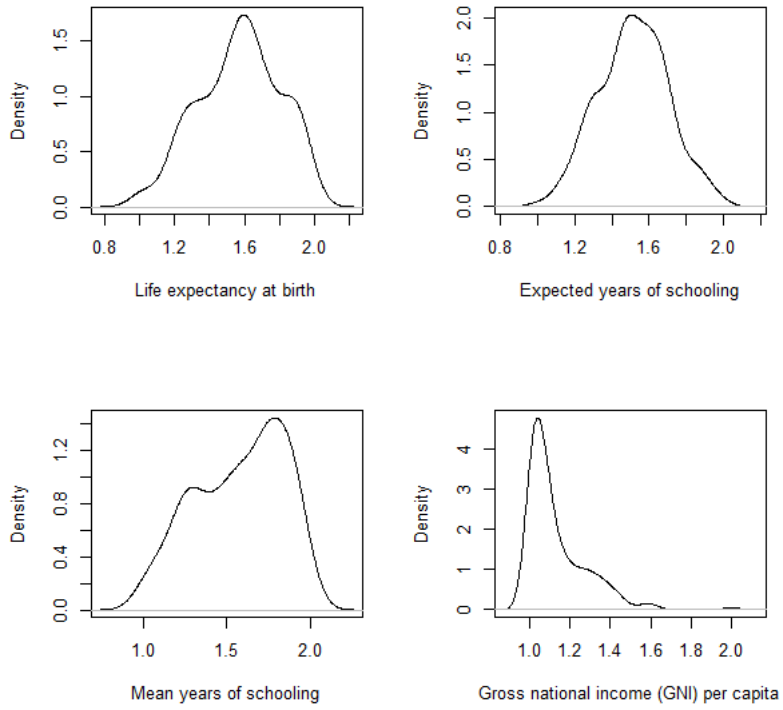

Figura 3.2: Grafico densità
Fonte: nostra eleborazione

Lo *scatterplot*[15] consente di capire se esiste una relazione tra due variabili[16]. Nel grafico 3.4 i valori delle variabili sono riportati rispettivamente sull'asse X e Y di un piano cartesiano.

Il grafico a dispersione consente quindi di poter valutare se esiste una correlazione

[15]trad. italiana: grafico di dispersione

[16]per relazione si intende la correlazione statistica

tra le variabili. I casi di relazione possono variare in relazioni positive[17], negative[18] o nulle[19].

Per il calcolo della correlazione si utilizzano i coefficienti di correlazione di Pearson, di Spearman o di Kendall.

Correlazione

La correlazione è una misura (numero) statistica che evidenzia una relazione tra due variabili, tale per cui un valore di una variabile X corrisponde un valore della variabile Y. Se entrambe le variabili variano, possiamo definire che gli indicatori hanno una relazione che può essere sia positiva che negativa[20].

Un esempio di correlazione tra variabili lo possiamo trovare tra la "temperatura media annuale" e il "disboscamento medio annuale", si nota che all'aumentare del disboscamento aumenterà anche la temperatura media annuale, questo esempio appena riportato mostra una correlazione tra variabili di tipo lineare.

L'analisi statistica del tipo di relazione viene definita come lineare e non lineare: è bene notare che la correlazione mostra il rapporto tra variabili che ci permette di

[17] situazione che si verifica quando, all'aumentare dei valori di X, aumentano in media anche i valori di Y

[18] all'aumentare dei valori di X, i valori di Y diminuiscono

[19] l'aumento dei valori di X, i valori di Y non avranno cambiamenti

[20] "Indice composito" (v. nota 4)

affermare una relazione tra di esse, ma non ci permette di misurare la causa-effetto tra le due variabili.

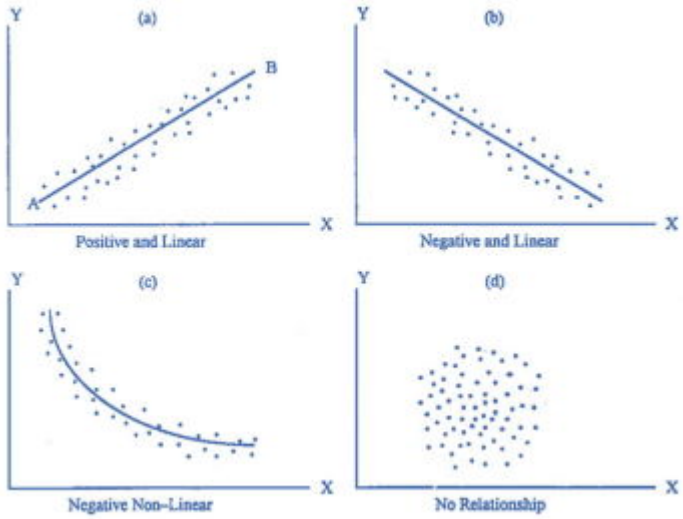

Figura 3.3: Grafico di dispersione **Fonte:** https://www.humanwareonline.com/project-management/center/diagramma-di-dispersione/

Come si evince nella figura 3.3 la dispersione delle osservazioni statistiche, nel grafico (A) e (B) formano una linea retta. Nell'esempio descritto sopra si può affermare come l'aumento del disboscamento medio annuale abbia una relazione positiva con l'aumento delle temperature medie annuali.

Nel caso di una relazione non lineare si assisterebbe a una dispersione delle osservazioni statistiche con un andamento simile ad una parabola o ad una iperbole.

Nel caso di un andamento non-lineare come si evince nella figura 3.3 nel grafico (c), si nota come al variare del valore di X attraverso livelli alti o bassi, i valori di Y non subiscono lo stesso cambiamento di valore con la stessa intensità del

cambiamento di valore della variabile X. In altre parole, tutte le relazioni che non crescono/diminuiscono sempre allo stesso ritmo, sono relazioni non lineari. Un esempio di relazione non lineare viene rappresentato tra la variabile "individui" e la variabile "glucosio nel sangue".

Come mostrato nella figura 3.3 nel riquadro (D), le due variabili X e Y non hanno nessun tipo di relazione, infatti, una variazione positiva o negativa di X non ha alcuna influenza nella variazione positiva o negativa di Y. Si può misurare la forza di una relazione in base al raggruppamento delle osservazioni statistiche, ovvero, più le osservazioni sono raggruppate intorno alla retta o alla parabola, tanto più forte è la relazione tra X e Y. Per esprimere la relazione tra le variabili X e Y si utilizza il coefficiente di correlazione: si tratta di un coefficiente standardizzato che può assumere valori compresi tra -1 (correlazione perfetta negativa) e +1 (correlazione perfetta positiva), una correlazione prossima a 0 indica che tra X e Y non vi è alcuna relazione (relazione neutra).

Nella letteratura statistica sono presenti vari tipi di coefficienti di correlazione[21], tuttavia, in questo elaborato si menzionerà solo il coefficiente r di Pearson e il coefficiente r di Spearman. Il primo ci aiuta a misurare una relazione lineare e viene

[21] Il coefficiente di correlazione è una misura specifica usata nell'analisi della correlazione per quantificare la forza della relazione lineare tra due variabili.

calcolato nel seguente modo:

$$[!ht]\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

dove:

- x_i= osservazione appartenente alla variabile X
- $\bar{x}$= media aritmetica della variabile X
- y_i= osservazione appartenente alla variabile Y
- $\bar{y}$= media aritmetica della variabile Y

Il secondo coefficiente, invece, serve a misurare la correlazione tra due variabili di tipo ordinale (ad esempio un professore può chiedere al proprio assistente di ordinare i suoi studenti per profitto, dal più bravo al meno bravo, e per socievolezza, dal più socievole al meno socievole, e vedere, se tra i due indicatori esiste una relazione). Il coefficiente rdi Spearman è una approssimazione del coefficiente di Pearson e si calcola in questo modo:

$$[!ht]\rho s = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

dove:

- d_i^2= $r_i - s_i$ essendo r_i e s_i rispettivamente il rango della prima variabile e della seconda variabile della i-esima osservazione

- n= numero totale delle osservazioni

Anche il coefficiente di Spearman avrà come dominio da +1 a -1.

Pearson	**LEI**	**EI**	**MEI**	**GNI**
LEI	**1.0000000**	0.7798626	0.7344082	0.7353677
EI	0.7798626	**1.0000000**	0.7785429	0.6390923
MEI	0.7344082	0.7785429	**1.0000000**	0.6507158
GNI	0.7353677	0.6390923	0.6507158	**1.0000000**

Tabella 3.3: Pearson

Fonte: nostra eleborazione

Spearman	**LEI**	**EI**	**MEI**	**GNI**
LEI	**1.0000000**	0.7920863	0.7279123	0.8504788
EI	0.7920863	**1.0000000**	0.7905501	0.8336870
MEI	0.7279123	0.7905501	**1.0000000**	0.8286098
GNI	0.8504788	0.8336870	0.8286098	**1.0000000**

Tabella 3.4: Spearman

Fonte: nostra eleborazione

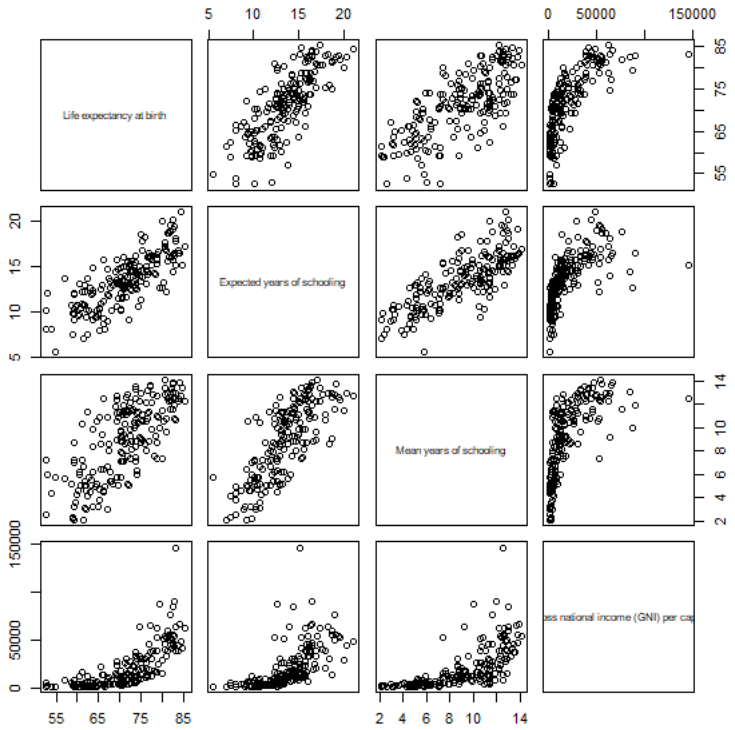

Figura 3.4: Grafico scatterplot
Fonte: nostra eleborazione

Grafici Box-plot e Violin

Con il termine *box-plot* conosciuto anche con il nome di grafico a baffi, si fa riferimento ad un grafico sugli *outliers*. Di seguito riporteremo le sue caratteristiche principali che consentiranno una migliore lettura al grafico 3.5:

- Al centro del grafico a baffi si trova la "linea centrale", la linea rappresenta la mediana[22]. In base al posizionamento della linea centrale si può capire se

[22]la mediana è un tipo di media dove se i dati superano un determinato valore, allora essi si troveranno nella metà superiore, viceversa se non lo superano, i dati si troveranno nella metà inferiore

i dati risultano simmetrici o non simmetrici, in quanto se la mediana risulta perfettamente al centro la variabile sarà simmetrica, viceversa, la variabile risulterà asimmetrica se la mediana sarà posizionata più in alto o più in basso.

- la parte inferiore della mediana rappresenta il venticinquesimo percentile (chiamato anche quartile); la parte superiore della mediana rappresenta il settantacinquesimo percentile[23].

- i baffi del grafico rappresentano la variazione dei dati. Se i dati non superano i baffi significa che i loro valori si estenderanno fino ai minimi e massimi. Se i dati superano i baffi allora la variabile possiede degli *outliers*.

Simile al grafico 3.5 sono i grafici a violino, essi rappresentano una distribuzione diversa rispetto alle "scatole" del *box-plot*. A differenza dei grafici a baffi, i grafici a violino possono essere sovrapposti i punti delle osservazioni statistiche appartenenti alla variabile.

Si riportano i grafici a baffi ed a violino per le variabili oggetto di studio:

[23] Si definisce range interquartile (IQR) la differenza tra i due percentili

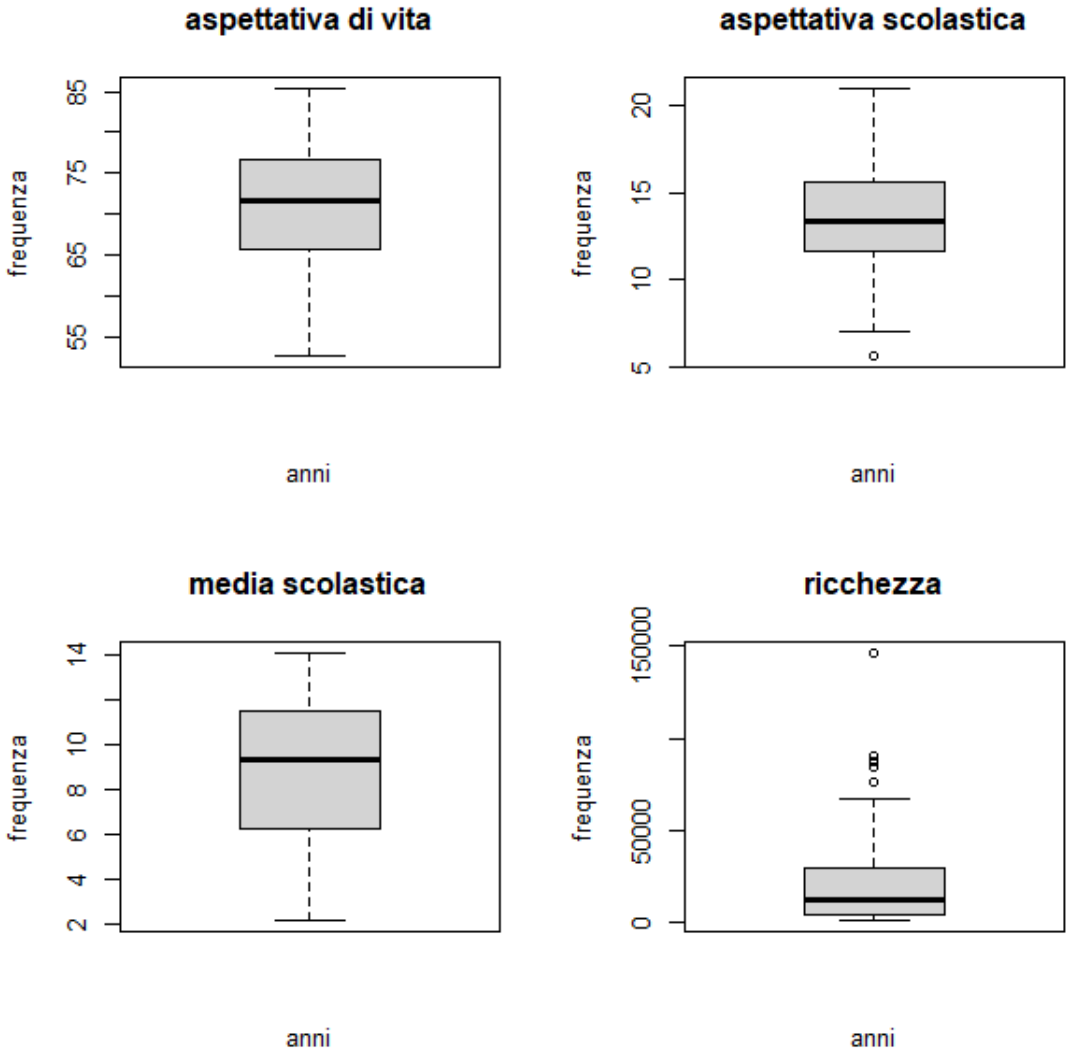

Figura 3.5: Box-plot
Fonte: nostra eleborazione

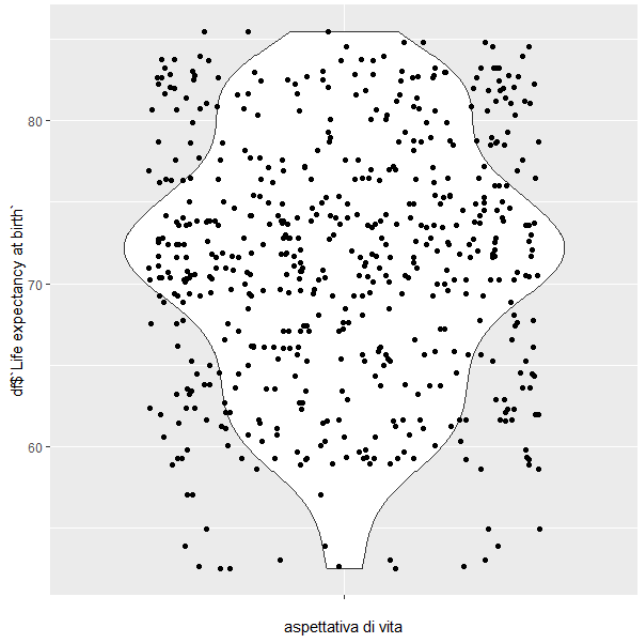

Figura 3.6: Violin aspettativa di vita
Fonte: nostra eleborazione

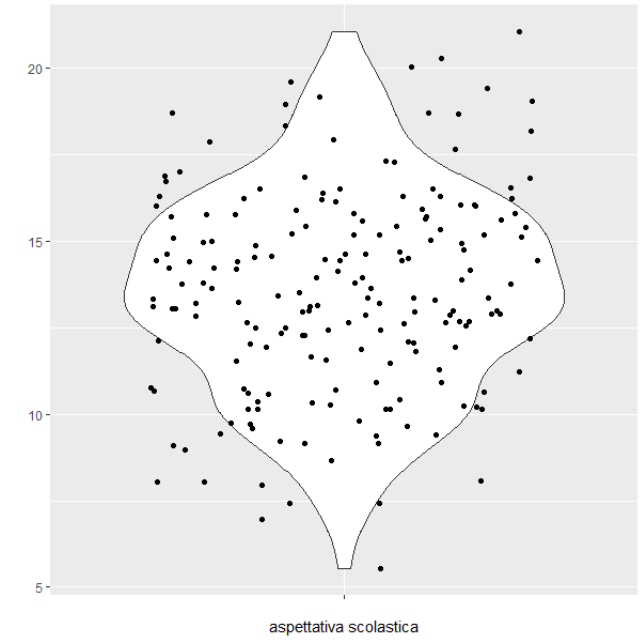

Figura 3.7: Violin aspettativa scolastica
Fonte: nostra eleborazione

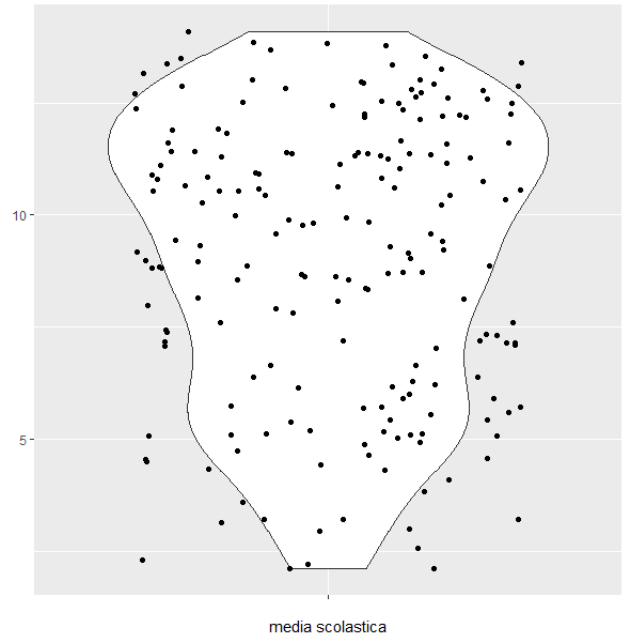

Figura 3.8: Violin media scolastica
Fonte: nostra eleborazione

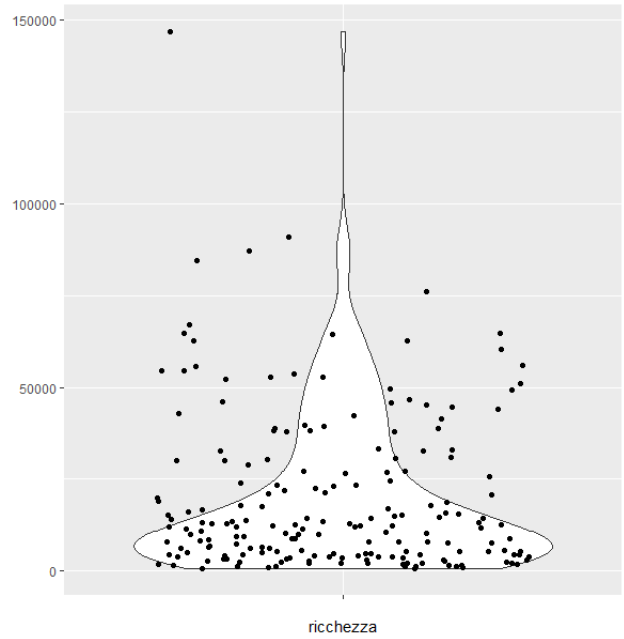

Figura 3.9: Violin ricchezza
Fonte: nostra eleborazione

3.3 Risultati

Di seguito si riporteranno i risultati dello studio sulla costruzione dell'indicatore composito HDI, calcolata con la metodologia classica e poi applicando ai dati il metodo di Mazziotta-Pareto e infine con la nostra proposta.

3.3.1 L'indice con Mazziotta Pareto

L'indice ottenuto applicando la metodologia proposta da Mazziotta-Pareto non mostra particolari segni di scostamento rispetto alla distribuzione dell'indice calcolato come media geometrica cioè quello ottenuto con la metodologia dell'HDI; l'unica divergenza che si nota è quella relativa alla figura 3.10 grafico denominato MPI, il picco mostrato in figura con coordinate X=120 e Y=0.020 risulta più simmetrico rispetto al grafico denominato HDI.

Se si passa ad analizzare i ranking prodotti dai due metodi emergono delle differenze:

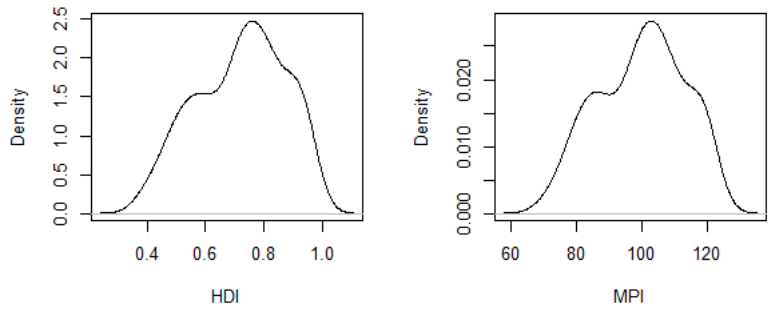

Figura 3.10: Grafico densità HDI vs. Mazziotta-Pareto
Fonte: nostra eleborazione

Country	**HDI**	**MPI**	**rHI**	**rMPI**
Switzerland	0.962	120.975	1	4
Norway	0.961	120.890	2	5
Iceland	0.959	121.338	3	2
Hong Kong, China (SAR)	0.952	120.650	4	6
Australia	0.951	122.395	5	1
Denmark	0.948	119.840	6	9
Sweden	0.947	120.629	7	7
....	...	...	...	...

Tabella 3.5: Tabella ranking HDI MPI

Fonte: nostra eleborazione

3.3.2 L'indice con C_0, C_{10}, C_{100}

Per ciascuna delle tre formulazioni, dopo aver trasformato i dati[24] attraverso la box-cox: si effettuerà una aggregazione delle variabili usando come funzione aggregativa tre medie potenziate, ovvero media aritmetica, media geometrica e media quadratica.

Come illustrato precedentemente nell'applicare la Box-Cox, faremo variare il λ da -6 a +6 con passo di 0.5 e fisseremo λ modo da normalizzare i dati, utilizzando il miglior λ scelto per il calcolo della normalizzazione di ogni variabile aggregativa

Arriveremo quindi a denominare per ognuna delle formulazioni un valore di λ diverso, ed anche un p-value[25] per ognuno delle tre forme aggregative.

Durante la normalizzazione degli indicatori aggregati si è eseguito il test di Lilliefors. Il test di Lilliefors (H.Lilliefors; 1967) conosciuto anche con il nome *distance test*, si basa attraverso la distanza massima tra quella osservata e quella attesa. Il test si Lilliefors si basa sull'ipotesi nulla e l'ipotesi alternativa, l'ipotesi nulla prevede una distribuzione di tipo normale N~(μ,σ^2) dove i vari gruppi di osservazioni devono risultare uguali a zero. L'ipotesi alternativa prevede una distribuzione di tipo normale N~(μ,σ^2) dove i vari gruppi di osservazioni non risultano uguali a zero a

[24] Si considereranno solo tre indicatori, per la costruzione dell'indice composito, ovvero: salute, istruzione e qualità della vita

[25] Il valore del p-value si utilizza nella statistica inferenziale per il test di ipotesi e possiede un valore di significatività osservata in termini percentuale

causa di una asimmetria tra i gruppi di osservazioni. Per il rifiuto dell'ipotesi nulla il calcolo dello scarto massimo deve risultare maggiore a quello riportato nella tabella di Lilliefors per la verifica della normalità di una distribuzione.

Nel caso di C0 i test di Lilliefors sono i seguenti: Si riportano gli output dei test delle tabelle per la media aritmetica 3.6, geometrica 3.7 e quadratica 3.8:

Lilliefors test	p-value
0.1479553	0.1479553

Tabella 3.6: Tabella Lilliefors e p-value con C0 aritmetica

Fonte: nostra eleborazione

Lilliefors test	p-value
0.2450587	0.2450587

Tabella 3.7: Tabella Lilliefors e p-value con C0 geometrica

Fonte: nostra eleborazione

Lilliefors test	p-value
0.08293136	0.08293136

Tabella 3.8: Tabella Lilliefors e p-value con C0 quadratica

Fonte: nostra eleborazione

Nel caso di C10 i test di Lilliefors sono i seguenti.

Si riportano gli output dei test delle tabelle per la media aritmetica 3.9, geometrica 3.10e quadratica 3.11: Nel caso di C100 i test di Lilliefors sono i seguenti.

Si riportano gli output dei test delle tabelle per la media aritmetica 3.12, geometrica 3.13e quadratica 3.14:

Lilliefors test	p-value
0.1668788	0.1668788

Tabella 3.9: Tabella Lilliefors e p-value con C10 aritmetica

Fonte: nostra eleborazione

Lilliefors test	p-value
0.1785324	0.1785324

Tabella 3.10: Tabella Lilliefors e p-value con C10 geometrica

Fonte: nostra eleborazione

Lilliefors test	p-value
0.1536727	0.1536727

Tabella 3.11: Tabella Lilliefors e p-value con C10 quadratica

Fonte: nostra eleborazione

Lilliefors test	p-value
0.1525473	0.1525473

Tabella 3.12: Tabella Lilliefors e p-value con C100 aritmetica

Fonte: nostra eleborazione

Lilliefors test	p-value
0.1538883	0.1538883

Tabella 3.13: Tabella Lilliefors e p-value con C100 geometrica

Fonte: nostra eleborazione

Lilliefors test	p-value
0.1512198	0.1512198

Tabella 3.14: Tabella Lilliefors e p-value con C100 quadratica

Fonte: nostra eleborazione

3.4 Confronti

In questa sezione si riporteranno i vari *output* dei tre indici calcolati rispettivamente con C_0, C_{10}, C_{100}.

Le tabelle 3.15 3.16 e 3.17 riportano i migliori valori di λ e i corrispettivi p-value delle tre medie ultimate e per ciascuno dei tre metodi di normalizzazione.

indicatore	**lambda**	**p-value**
media aritmetica	0.5	0.11108624
media geometrica	0.0	0.16776683
media quadratica	1.0	0.05631879

Tabella 3.15: Tabella lambda e p-value con C0

Fonte: nostra eleborazione

indicatore	**lambda**	**p-value**
media aritmetica	-3.0	0.10673676
media geometrica	-3.5	0.11241072
media quadratica	-3.0	0.09707541

Tabella 3.16: Tabella lambda e p-value con C10

Fonte: nostra eleborazione

indicatore	**lambda**	**p-value**
media aritmetica	-6.0	0.09207501
media geometrica	-5.5	0.08515199
media quadratica	-6.0	0.08944829

Tabella 3.17: Tabella lambda e p-value con C100

Fonte: nostra eleborazione

I risultati ottenuti con la normalizzazione ed aggregazioni oggetto del nostro studio quando si normalizza usando C_0 come mostrato nella tabella 3.15, il nostro metodo suggerisce di prendere valori di λ a seconda della media.

Per esempio nel caso della media geometrica è stato trovato $\lambda = 0$ che vuol dire adattare una normalizzazione logaritmica.

É stato ripetuto lo stesso procedimento di normalizzazione anche per le altre due costanti C_{10} e C_{100}, in questo caso non è risultato per nessuno dei tre indicatori in entrambi i casi un $\lambda = 0$, quindi si è proceduto con la normalizzazione descritta attraverso le formule in presenza di un λ diverso da zero.

Country	**rAMn**	**rGMn**	**rQMn**	**comp**	**RankC**
Liechtenstein	1	1	1	2,073992533	1
Australia	2	2	2	1,79521943	2
Iceland	3	3	3	1,739411856	3
Switzerland	4	6	4	1,682623122	4
Norway	5	5	6	1,6800561	5
Ireland	6	7	8	1,645193983	6
Sweden	7	8	7	1,642955679	7
Singapore	8	4	10	1,640716014	8
New Zealand	9	11	5	1,638071098	9
Denmark	10	9	9	1,61305687	10
...	...	...	...	...	...

Tabella 3.18: Tabella ranking con C_0

Fonte: nostra eleborazione

Le tabelle 3.18, 3.19 e 3.20 riportano i *ranking* per Paese di ogni indicatore aggregato normalizzato. La colonna **comp** indica il risultato della comparazione dei tre indici aggregati e normalizzati, è stata calcolata attraverso:

$$comp = (rAMn + rGMn + rQMn)/3$$

dove 3 è uguale al numero di indici utilizzati per la comparazione. La colonna **RankC** fa riferimento al ranking della comparazione calcolata precedentemente.

Country	rAMn	rGMn	rQMn	comp	RankC
Liechtenstein	1	1	1	2,050143288	1
Australia	2	2	2	1,779321898	2
Iceland	3	3	3	1,725206772	3
Switzerland	4	4	4	1,669102586	4
Norway	5	5	5	1,667578524	5
Ireland	6	6	8	1,634113084	6
Sweden	7	8	6	1,63153089	7
Singapore	8	7	9	1,629967123	8
New Zealand	9	9	7	1,62702667	9
Denmark	10	10	10	1,602430337	10
...	...	...	...	...	...

Tabella 3.19: Tabella ranking con C10

Fonte: nostra eleborazione

I risultati con la costante C_{100} mostrano una robustezza migliore rispetto a quelli evidenziati con la costante C_{10}, questo tipo di robustezza è la migliore che si possa calcolare attraverso questo tipo di normalizzazione, poiché aumentando la costante con un numero superiore a cento, il ranking non subirebbe nessun cambiamento. Rispetto alle tabelle precedenti, i risultati della tabella con la costante pari a cento, vedi tabella 3.20.

Country	rAMn	rGMn	rQMn	comp	RankC
Liechtenstein	1	1	1	2,164090455	1
Australia	2	2	2	1,856323167	2
Iceland	3	3	3	1,795484709	3
Switzerland	4,5	4	4,5	1,733624823	4
Norway	5,5	5	5,5	1,733114109	5
Ireland	7,5	6	7,5	1,690166514	6
Sweden	7,5	8	8,5	1,689145086	7
Singapore	7,5	8	7,5	1,688634372	8
New Zealand	7,5	9	7,5	1,68710223	9
Denmark	10	10	10	1,657959786	10
...	...	...	...	...	...

Tabella 3.20: Tabella ranking con C100

Fonte: nostra eleborazione

Capitolo 4

Conclusione

Questo elaborato di studio si è prefissato di tentare di rispondere ad alcune domande ancora insolute circa la costruzione di indicatori compostiti, ovvero alla domanda: come normalizzare aggregare i dati attraverso l'utilizzo del dataset pubblicato dalla ricerca *Human Development Report 2021-2022*[1]. Il lavoro ha previsto una prima fase teorica che, partendo dallo stato dell'arte della costruzione di indicatori compositi, ha portato alla formulazione di una nuova proposta di normalizzazione che sembra rendere gli ordinamenti robusti. Il nostro scopo si è focalizzato nella manipolazione degli indici semplici al fine di ottenere un indice composito meno sensibile al cambiamento di determinate aggregazioni.

Il risultato finale è coerente con la domanda che ci si è posta prima di iniziare questo studio di normalizzazione. Siamo partiti dalle criticità evidenziate nel corso dei capitoli riguardanti i problemi ancora aperti e discussi dalla letteratura statistica

[1]Programme (v. nota 2)

sulle normalizzazioni e aggregazioni attuate per la costruzione dell'indice composito. Si è stati in grado di poter manipolare le variabili applicando una trasformazione che rende normali i nostri dati. Tale trasformazione, chiamata Box-Cox, ha consentito di ottenere il miglior lambda ed il suo rispettivo p-value per tre diversi metodi di aggregazione e tre metodi aggregati.

Questa tesi di laurea ha preso spunto dai vari concetti presenti nella letteratura statistica ed ha posto una visione differente rispetto ad essi. Gli studi effettuati già in precedenza mostrano come migliorare l'indice composito della ricerca effettuata dall'UNPD.

La metodologia statistica risulta essere complessa ed in costante miglioramento con processi metodologici sempre in miglioramento, per questo ci sarà sempre un metodo diverso per ottenere lo stesso risultato. Proprio per questo motivo il nostro studio non potrà essere esente da critiche e miglioramenti.

Tra gli sviluppi futuri di questo studio c'è la possibilità di costruire intervalli di confidenza e applicare quindi metodologie di statistica inferenziale.

Pertanto, le ricerche future che avranno di base questo studio potranno effettuare una comparazione con l'indice composito calcolato nella nostra proposta con qualche altro indice composito basato sull'HDI, inoltre, potranno ampliare il nostro studio effettuando dei calcoli attraverso l'uso del nostro indice composito comparandolo con i dati storici messi a disposizione dall'UNPD.

Ringraziamenti

Ringrazio tutte le persone che hanno aiutato alla riuscita di questa tesi di laurea apportando suggerimenti, e osservazioni varie. Ma la responsabilità per ogni errore contenuto spetta a me medesimo.

La mia gratitudine va a Mirco Piccione in primis, per aver dedicato anima e corpo nel corso di questi anni, a tutti i traguardi raggiunti e a quelli futuri.

Ringrazio la mia famiglia nel particolare: Maurizio Piccione, Patrizia Lombardo, Noemi Piccione, Alessandro Migliorisi e Francesco Migliorisi, per avermi supportato in questi anni. la mia compagna Federica Bondavalli per essere sempre stata al mio fianco e per avermi spronato ad affrontare questo corso di laurea.

Un altro speciale ringraziamento va al supporto che mi è stato dato dalla famiglia della mia compagna, nel particolare: Alessandro Bondavalli, Mariangela Furnari e Francesca Bondavalli.

Desidero ringraziare anche la mia relatrice: Maria Cristina Recchioni per essere sempre stata un modello da seguire e per la sua costante dedizione all'insegnamento ed al continuo miglioramento del corso di laurea in data science, ringrazio anche la mia co-relatrice per avermi offerto lo spunto e per avermi seguito nel cammino di questa tesi di laurea: Mariateresa Ciommi.

Un caloroso ringraziamento a tutti gli studenti di data science per avermi permesso di poterli rappresentare a livello accademico, spero di aver fatto un buon lavoro, quanto meno ci ho provato.

In particolare, tra gli studenti vorrei ringraziare tutto il mio gruppo di studi con cui abbiamo condiviso ansie, gioie e nervosismi vari, in particolare: Daniele Montella per le meravigliose serate passate in compagnia di Edoardo Volponi, Luana Lemme per i continui battibecchi e "lo studio matto e disperato", e tutto il gruppo dei membri di data clown.

Vorrei ringraziare anche gli "amici di una vita": Stefano Meli, Michele Cutrale per tutte le avventure passate insieme e tutti gli amici della mia città natale.

Questo mio viaggio universitario è ormai giunto al termine.

Appendice A

codice in Rstudio

Si utilizzeranno le seguenti librerie [1]:

- library(rstudioapi) [2].
- library(openxlsx) [3].
- library(MASS) [4].
- library(quantmod) [5].
- library(nortest) [6].

```
normalize = function(x, na.rm = TRUE) {
  return((x- min(x)) /(max(x)-min(x)))}
```

[1] "CRAN R" https://cran.r-project.org

[2] Il pacchetto rstudioapi è progettato per rendere facile l'accesso condizionale alle API RStudio dai pacchetti CRAN

[3] Questo pacchetto R semplifica la lettura e la creazione di file in formato xlsx

[4] Fornisce supporto alle funzioni e manipolazioni del dataset

[5] Fornisce supporto con funzioni di analisi

[6] Fornisce supporto nel verificare l'ipotesi composta di normalità

```
zscore = function(x, na.rm = TRUE) {
  return((x- mean(x)) /(var(x)\^0.5))}
normalizeh = function(x,h, na.rm = TRUE) {
  a=h/(max(x)-min(x))
  b=-a*min(x)
  return(a*x+b)}

df_def=matrix(rep(0,ncol(df)*nrow(df)), nrow=nrow(df));
aux=NULL;
for (j in 1:ncol(df)){
aux=normalize(df[,j])+1
df_def[,j]=aux}

data_def=data.frame(df_def)
rownames(data_def)=mydata$Country
colnames(data_def)<- c("Life␣expectancy␣at␣birth",
                       "Expected␣years␣of␣schooling␣",
                       "Mean␣years␣of␣schooling",
                       "Gross␣national␣income␣(GNI)␣per␣capita")

IndexAM=(1/ncol(df_def))*rowSums(df_def)
 geometric mean
IndexGM=NULL;
for(i in 1:nrow(df_def)){
   IndexGM[i]=(prod(df_def[i,]))^(1/ncol(df_def))
}
\_quadratic mean
IndexQM=((1/ncol(df_def))*rowSums(df_def^2))^0.5

p=seq(from=-6, to=6, by=0.5)
np=3;

lambda=NULL;
pvsh=NULL;

y=IndexAM
shmax=0.0;
for (ll in p){
    print(ll)
```

```
    if(ll==0){
        yyy=log(y)}
    else{
        yyy=(y^ll-1)/ll
    }
    yy=(yyy-mean(yyy))/var(yyy)^0.5
    sh=lillie.test(yy)
    shp=sh$p.value
    if(shp>shmax){
        lambda[1]=ll
        pvsh[1]=shp
        shmax=shp
     }
  }

y=IndexGM
shmax=0.0;
for (ll in p){
  print(ll)
  if(ll==0){
    yyy=log(y)}
  else{
    yyy=(y^ll-1)/ll
  }
  yy=(yyy-mean(yyy))/var(yyy)^0.5
  sh=lillie.test(yy)
  shp=sh$p.value
  if(shp>shmax){
    lambda[2]=ll
    pvsh[2]=shp
    shmax=shp
  }
}


y=IndexQM
shmax=0.0;
for (ll in p){
  print(ll)
```

```
    if(ll==0){
      yyy=log(y)}
    else{
      yyy=(y^ll-1)/ll
    }
    yy=(yyy-mean(yyy))/var(yyy)^0.5
    sh=lillie.test(yy)
    shp=sh$p.value
    if(shp>shmax){
      lambda[3]=ll
      pvsh[3]=shp
      shmax=shp
    }
}

ll=lambda[1]
index_1=(IndexAM^ll-1)/ll
IndexAM_n=(index_1-mean(index_1))/var(index_1)^0.5
sh=lillie.test(IndexAM_n)
print(cbind(sh$p.value,pvsh[1]))

ll=lambda[2]
index_1=(IndexGM^ll-1)/ll
IndexGM_n=(index_1-mean(index_1))/var(index_1)^0.5
sh=lillie.test(IndexGM_n)
print(cbind(sh$p.value,pvsh[2]))

ll=lambda[3]
index_1=(IndexQM^ll-1)/ll
IndexQM_n=(index_1-mean(index_1))/var(index_1)^0.5
sh=lillie.test(IndexQM_n)
print(cbind(sh$p.value,pvsh[3]))

AM_data=data.frame(mydata$Country,IndexAM,IndexAM_n)
colnames(AM_data)=c("Country","AM","AM_n")
AM_data$RankAM=190-rank(IndexAM)
AM_data$RankAMn=190-rank(IndexAM_n)

GM_data=data.frame(mydata$Country,IndexGM,IndexGM_n)
```

```
colnames(GM_data)=c("Country","GM","GM_n")
GM_data$RankGM=190-rank(IndexGM)
GM_data$RankGMn=190-rank(IndexGM_n)

QM_data=data.frame(mydata$Country,IndexQM,IndexQM_n)
colnames(QM_data)=c("Country","QM","QM_n")
QM_data$RankQM=190-rank(IndexQM)
QM_data$RankQMn=190-rank(IndexQM_n)

compareC1=data.frame(mydata$Country,IndexAM,IndexGM,
                     IndexQM,IndexAM_n,IndexGM_n,IndexQM_n,
                     AM_data$RankAMn, GM_data$RankGMn,
                     QM_data$RankQMn)
colnames(compareC1)=c("Country","AM","GM","QM","AMn","GMn",
                      "QM_n","rAMn","rGMn","rQMn")

comp_comp=(IndexAM_n+IndexGM_n+IndexQM_n)/3
compareC1$comp=comp_comp
compareC1$RankC=190-rank(comp_comp)

df_def=matrix(rep(0,ncol(df)*nrow(df)), nrow=nrow(df));
aux=NULL;
cost=10
for (j in 1:ncol(df)){
  aux=normalizeh(df[,j],hh)+cost
  df_def[,j]=aux
}
data_def=data.frame(df_def)
rownames(data_def)=mydata$Country
colnames(data_def)<- c("Life␣expectancy␣at␣birth",
                       "Expected␣years␣of␣schooling␣",
                       "Mean␣years␣of␣schooling",
                       "Gross␣national␣income␣(GNI)␣per␣capita")

IndexAM=(1/ncol(df_def))*rowSums(df_def)

IndexGM=NULL;
for(i in 1:nrow(df_def)){
  IndexGM[i]=(prod(df_def[i,]))^(1/ncol(df_def))}
```

```
IndexQM=((1/ncol(df_def))*rowSums(df_def^2))^0.5

p=seq(from=-6, to=6, by=0.5)
np=3;
lambda=NULL;
pvsh=NULL;

y=IndexAM
shmax=0.0;
for (ll in p){
  print(ll)
  if(ll==0){
    yyy=log(y)}
  else{
    yyy=(y^ll-1)/ll
  }
  yy=(yyy-mean(yyy))/var(yyy)^0.5
  sh=lillie.test(yy)
  shp=sh$p.value
  if(shp>shmax){
    lambda[1]=ll
    pvsh[1]=shp
    shmax=shp
  }
}

y=IndexGM
shmax=0.0;
for (ll in p){
  print(ll)
  if(ll==0){
    yyy=log(y)}
  else{
    yyy=(y^ll-1)/ll
  }
  yy=(yyy-mean(yyy))/var(yyy)^0.5
  sh=lillie.test(yy)
  shp=sh$p.value
```

```
   if(shp>shmax){
     lambda[2]=ll
     pvsh[2]=shp
     shmax=shp
   }
}


y=IndexQM
shmax=0.0;
for (ll in p){
   print(ll)
   if(ll==0){
     yyy=log(y)}
   else{
     yyy=(y^ll-1)/ll
   }
   yy=(yyy-mean(yyy))/var(yyy)^0.5
   sh=lillie.test(yy)
   shp=sh$p.value
   if(shp>shmax){
     lambda[3]=ll
     pvsh[3]=shp
     shmax=shp
   }
}

ll=lambda[1]
index_1=(IndexAM^ll-1)/ll
IndexAM_n=(index_1-mean(index_1))/var(index_1)^0.5
sh=lillie.test(IndexAM_n)

ll=lambda[2]
index_1=(IndexGM^ll-1)/ll
IndexGM_n=(index_1-mean(index_1))/var(index_1)^0.5
sh=lillie.test(IndexGM_n)
print(cbind(sh$p.value,pvsh[2]))

ll=lambda[3]
```

```
index_1=(IndexQM^11 -1)/11
IndexQM_n=(index_1-mean(index_1))/var(index_1)^0.5
sh=lillie.test(IndexQM_n)
print(cbind(sh$p.value,pvsh[3]))

AM_data=data.frame(mydata$Country,IndexAM,IndexAM_n)
colnames(AM_data)=c("Country","AM","AM_n")
AM_data$RankAM=190-rank(IndexAM)
AM_data$RankAMn=190-rank(IndexAM_n)


GM_data=data.frame(mydata$Country,IndexGM,IndexGM_n)
colnames(GM_data)=c("Country","GM","GM_n")
GM_data$RankGM=190-rank(IndexGM)
GM_data$RankGMn=190-rank(IndexGM_n)

QM_data=data.frame(mydata$Country,IndexQM,IndexQM_n)
colnames(QM_data)=c("Country","QM","QM_n")
QM_data$RankQM=190-rank(IndexQM)
QM_data$RankQMn=190-rank(IndexQM_n)

compareC10=data.frame(mydata$Country,IndexAM,IndexGM,IndexQM,
                      IndexAM_n,IndexGM_n,IndexQM_n, AM_data$RankAMn,
                      GM_data$RankGMn,QM_data$RankQMn)
colnames(compareC10)=c("Country","AM","GM","QM","AMn","GMn",
                       "QM_n","rAMn","rGMn","rQMn")

comp_comp=(IndexAM_n+IndexGM_n+IndexQM_n)/3
compareC10$comp=comp_comp
compareC10$RankC=190-rank(comp_comp)

df_def=matrix(rep(0,ncol(df)*nrow(df)), nrow=nrow(df));
aux=NULL;
cost=100
hh=0.01
for (j in 1:ncol(df)){
  aux=normalize(df[,j])+cost
  #aux=normalizeh(df[,j],hh)
  #aux=exp(df[,j]-mean(df[,j]));
```

```
  df_def[,j]=aux
}
data_def=data.frame(df_def)
rownames(data_def)=mydata$Country
colnames(data_def)<- c("Life expectancy at birth",
                       "Expected years of schooling ",
                       "Mean years of schooling",
                       "Gross national income (GNI) per capita")

IndexAM=(1/ncol(df_def))*rowSums(df_def)

IndexGM=NULL;
for(i in 1:nrow(df_def)){
  IndexGM[i]=(prod(df_def[i,]))^(1/ncol(df_def))}

IndexQM=((1/ncol(df_def))*rowSums(df_def^2))^0.5

p=seq(from=-6, to=6, by=0.5)
np=3;

lambda=NULL;
pvsh=NULL;

y=IndexAM
shmax=0.0;
for (ll in p){
  if(ll==0){
    yyy=log(y)}
  else{
    yyy=(y^ll-1)/ll
  }
  yy=(yyy-mean(yyy))/var(yyy)^0.5
  sh=lillie.test(yy)
  shp=sh$p.value
  if(shp>shmax){
    lambda[1]=ll
    pvsh[1]=shp
    shmax=shp
  }
```

```
}

y=IndexGM
shmax=0.0;
for (ll in p){
  if(ll==0){
    yyy=log(y)}
  else{
    yyy=(y^ll-1)/ll
  }
  yy=(yyy-mean(yyy))/var(yyy)^0.5
  sh=lillie.test(yy)
  shp=sh$p.value
  if(shp>shmax){
    lambda[2]=ll
    pvsh[2]=shp
    shmax=shp
  }
}


y=IndexQM
shmax=0.0;
for (ll in p){
  if(ll==0){
    yyy=log(y)}
  else{
    yyy=(y^ll-1)/ll
  }
  yy=(yyy-mean(yyy))/var(yyy)^0.5
  sh=lillie.test(yy)
  shp=sh$p.value
  if(shp>shmax){
    lambda[3]=ll
    pvsh[3]=shp
    shmax=shp
  }
}
```

```
ll=lambda[1]
index_1=(IndexAM^ll-1)/ll
IndexAM_n=(index_1-mean(index_1))/var(index_1)^0.5
sh=lillie.test(IndexAM_n)

ll=lambda[2]
index_1=(IndexGM^ll-1)/ll
IndexGM_n=(index_1-mean(index_1))/var(index_1)^0.5
sh=lillie.test(IndexGM_n)
print(cbind(sh$p.value,pvsh[2]))

ll=lambda[3]
index_1=(IndexQM^ll-1)/ll
IndexQM_n=(index_1-mean(index_1))/var(index_1)^0.5
sh=lillie.test(IndexQM_n)
print(cbind(sh$p.value,pvsh[3]))

AM_data=data.frame(mydata$Country,IndexAM,IndexAM_n)
colnames(AM_data)=c("Country","AM","AM_n")
AM_data$RankAM=190-rank(IndexAM)
AM_data$RankAMn=190-rank(IndexAM_n)

GM_data=data.frame(mydata$Country,IndexGM,IndexGM_n)
colnames(GM_data)=c("Country","GM","GM_n")
GM_data$RankGM=190-rank(IndexGM)
GM_data$RankGMn=190-rank(IndexGM_n)

QM_data=data.frame(mydata$Country,IndexQM,IndexQM_n)
colnames(QM_data)=c("Country","QM","QM_n")
QM_data$RankQM=190-rank(IndexQM)
QM_data$RankQMn=190-rank(IndexQM_n)

compareC100=data.frame(mydata$Country,IndexAM,IndexGM,IndexQM,
                       IndexAM_n,IndexGM_n,IndexQM_n,
                       AM_data$RankAMn,GM_data$RankGMn,
                       QM_data$RankQMn)
colnames(compareC100)=c("Country","AM","GM","QM","AMn","GMn","QM_n",
                        "rAMn","rGMn","rQMn")
```

```
comp_comp=(IndexAM_n+IndexGM_n+IndexQM_n)/3
compareC100$comp=comp_comp
compareC100$RankC=190-rank(comp_comp)
```

Appendice B

Dataset originale

dove le variabili sono:

- HDI rank
- Country
- V1= Human Development Index (HDI)
- V2= Life expectancy at birth
- V3= Expected years of schooling
- V4= Mean years of schooling
- V5= Gross national income (GNI) per capita
- V6= GNI per capita rank minus HDI rank 2021
- V7 =HDI rank 2020

HDI rank	Country	V1	V2	V3	V4	V5	V6	V7
1	Switzerland	0,962	84,0	16,5	13,9	66.933	5	3
2	Norway	0,961	83,2	18,2	13,0	64.660	6	1
3	Iceland	0,959	82,7	19,2	13,8	55.782	11	2
4	Hong Kong, China (SAR)	0,952	85,5	17,3	12,2	62.607	6	4
5	Australia	0,951	84,5	21,1	12,7	49.238	18	5
6	Denmark	0,948	81,4	18,7	13,0	60.365	6	5
7	Sweden	0,947	83,0	19,4	12,6	54.489	9	9
8	Ireland	0,945	82,0	18,9	11,6	76.169	-3	8
9	Germany	0,942	80,6	17,0	14,1	54.534	6	7
10	Netherlands	0,941	81,7	18,7	12,6	55.979	3	10
11	Finland	0,940	82,0	19,1	12,9	49.452	11	12
12	Singapore	0,939	82,8	16,5	11,9	90.919	-10	10
13	Belgium	0,937	81,9	19,6	12,4	52.293	7	16
13	New Zealand	0,937	82,5	20,3	12,9	44.057	16	13
15	Canada	0,936	82,7	16,4	13,8	46.808	9	15
16	Liechtenstein	0,935	83,3	15,2	12,5	146.830	-15	14
17	Luxembourg	0,930	82,6	14,4	13,0	84.649	-13	17
18	United Kingdom	0,929	80,7	17,3	13,4	45.225	9	17
19	Japan	0,925	84,8	15,2	13,4	42.274	12	19
21	United States	0,921	77,2	16,3	13,7	64.765	-14	21
22	Israel	0,919	82,3	16,1	13,3	41.524	10	22
23	Malta	0,918	83,8	16,8	12,2	38.884	12	26
23	Slovenia	0,918	80,7	17,7	12,8	39.746	10	23
25	Austria	0,916	81,6	16,0	12,3	53.619	-8	23
26	United Arab Emirates	0,911	78,7	15,7	12,7	62.574	-15	25
27	Spain	0,905	83,0	17,9	10,6	38.354	10	27
28	France	0,903	82,5	15,8	11,6	45.937	-2	28
29	Cyprus	0,896	81,2	15,6	12,4	38.188	9	29
30	Italy	0,895	82,9	16,2	10,7	42.840	0	32
31	Estonia	0,890	77,1	15,9	13,5	38.048	8	30
32	Czechia	0,889	77,7	16,2	12,9	38.745	4	30
33	Greece	0,887	80,1	20,0	11,4	29.002	17	33
34	Poland	0,876	76,5	16,0	13,2	33.034	8	36
35	Bahrain	0,875	78,8	16,3	11,0	39.497	-1	35
35	Lithuania	0,875	73,7	16,3	13,5	37.931	5	34
35	Saudi Arabia	0,875	76,9	16,1	11,3	46.112	-10	38
38	Portugal	0,866	81,0	16,9	9,6	33.155	3	39
39	Latvia	0,863	73,6	16,2	13,3	32.803	4	37
40	Andorra	0,858	80,4	13,3	10,6	51.167	-19	45
40	Croatia	0,858	77,6	15,1	12,2	30.132	8	41
42	Chile	0,855	78,9	16,7	10,9	24.563	14	43
42	Qatar	0,855	79,3	12,6	10,0	87.134	-39	42
44	San Marino	0,853	80,9	12,3	10,8	52.654	-25	46
45	Slovakia	0,848	74,9	14,5	12,9	30.690	1	40
46	Hungary	0,846	74,5	15,0	12,2	32.789	-2	44

HDI rank	Country	V1	V2	V3	V4	V5	V6	V7
47	Argentina	0,842	75,4	17,9	11,1	20.925	17	47
48	Türkiye	0,838	76,0	18,3	8,6	31.033	-3	48
49	Montenegro	0,832	76,3	15,1	12,2	20.839	16	52
50	Kuwait	0,831	78,7	15,3	7,3	52.920	-32	54
51	Brunei Darussalam	0,829	74,6	14,0	9,2	64.490	-42	49
52	Russian Federation	0,822	69,4	15,8	12,8	27.166	-1	49
53	Romania	0,821	74,2	14,2	11,3	30.027	-4	53
54	Oman	0,816	72,5	14,6	11,7	27.054	-2	51
55	Bahamas	0,812	71,6	12,9	12,6	30.486	-8	58
56	Kazakhstan	0,811	69,4	15,8	12,3	23.943	1	59
57	Trinidad and Tobago	0,810	73,0	14,5	11,6	23.392	1	56
58	Costa Rica	0,809	77,0	16,5	8,8	19.974	8	57
58	Uruguay	0,809	75,4	16,8	9,0	21.269	5	55
60	Belarus	0,808	72,4	15,2	12,1	18.849	8	60
61	Panama	0,805	76,2	13,1	10,5	26.957	-8	67
62	Malaysia	0,803	74,9	13,3	10,6	26.658	-8	61
63	Georgia	0,802	71,7	15,6	12,8	14.664	17	64
63	Mauritius	0,802	73,6	15,2	10,4	22.025	-1	62
63	Serbia	0,802	74,2	14,4	11,4	19.123	4	62
66	Thailand	0,800	78,7	15,9	8,7	17.030	6	64
67	Albania	0,796	76,5	14,4	11,3	14.131	17	68
68	Bulgaria	0,795	71,8	13,9	11,4	23.079	-8	64
68	Grenada	0,795	74,9	18,7	9,0	13.484	18	70
70	Barbados	0,790	77,6	15,7	9,9	12.306	26	71
71	Antigua and Barbuda	0,788	78,5	14,2	9,3	16.792	2	71
72	Seychelles	0,785	71,3	13,9	10,3	25.831	-17	69
73	Sri Lanka	0,782	76,4	14,1	10,8	12.578	21	75
74	Bosnia e Herz	0,780	75,3	13,8	10,5	15.242	4	73
75	Saint Kitts and Nevis	0,777	71,7	15,4	8,7	23.358	-16	76
76	Iran	0,774	73,9	14,6	10,6	13.001	15	77
77	Ukraine	0,773	71,6	15,0	11,1	13.256	11	78
78	North Macedonia	0,770	73,8	13,6	10,2	15.918	-3	79
79	China	0,768	78,2	14,2	7,6	17.504	-8	82
80	Dominican Republic	0,767	72,6	14,5	9,3	17.990	-11	82
80	Moldova	0,767	68,8	14,4	11,8	14.875	-1	81
80	Palau	0,767	66,0	15,8	12,5	13.819	5	80
83	Cuba	0,764	73,7	14,4	12,5	7.879	37	73
84	Peru	0,762	72,4	15,4	9,9	12.246	13	85
85	Armenia	0,759	72,0	13,1	11,3	13.158	4	87
86	Mexico	0,758	70,2	14,9	9,2	17.896	-16	88
87	Brazil	0,754	72,8	15,6	8,1	14.370	-5	86
88	Colombia	0,752	72,8	14,4	8,9	14.384	-7	88
89	St. Vinc Grenadines	0,751	69,6	14,7	10,8	11.961	11	82
90	Maldives	0,747	79,9	12,6	7,3	15.448	-14	97
91	Algeria	0,745	76,4	14,6	8,1	10.800	13	96

HDI rank	Country	V1	V2	V3	V4	V5	V6	V7
91	Azerbaijan	0,745	69,4	13,5	10,5	14.257	-8	100
91	Tonga	0,745	71,0	16,0	11,4	6.822	34	90
91	Turkmenistan	0,745	69,3	13,2	11,3	13.021	-1	93
95	Ecuador	0,740	73,7	14,6	8,8	10.312	11	99
96	Mongolia	0,739	71,0	15,0	9,4	10.588	9	90
97	Egypt	0,731	70,2	13,8	9,6	11.732	4	97
97	Tunisia	0,731	73,8	15,4	7,4	10.258	10	94
99	Fiji	0,730	67,1	14,7	10,9	9.980	9	94
99	Suriname	0,730	70,3	13,0	9,8	12.672	-6	92
101	Uzbekistan	0,727	70,9	12,5	11,9	7.917	18	107
102	Dominica	0,720	72,8	13,3	8,1	11.488	0	106
102	Jordan	0,720	74,3	10,6	10,4	9.924	8	104
104	Libya	0,718	71,9	12,9	7,6	15.336	-27	117
105	Paraguay	0,717	70,3	13,0	8,9	12.349	-10	100
106	Palestine, State of	0,715	73,5	13,4	9,9	6.583	21	109
106	Saint Lucia	0,715	71,1	12,9	8,5	12.048	-7	104
108	Guyana	0,714	65,7	12,5	8,6	22.465	-47	107
109	South Africa	0,713	62,3	13,6	11,4	12.948	-17	102
110	Jamaica	0,709	70,5	13,4	9,2	8.834	4	110
111	Samoa	0,707	72,8	12,4	11,4	5.308	24	112
112	Gabon	0,706	65,8	13,0	9,4	13.367	-25	113
112	Lebanon	0,706	75,0	11,3	8,7	9.526	-1	103
114	Indonesia	0,705	67,6	13,7	8,6	11.466	-11	116
115	Viet Nam	0,703	73,6	13,0	8,4	7.867	6	113
116	Philippines	0,699	69,3	13,1	9,0	8.920	-3	113
117	Botswana	0,693	61,1	12,3	10,3	16.198	-43	110
118	Bolivia	0,692	63,6	14,9	9,8	8.111	0	119
118	Kyrgyzstan	0,692	70,0	13,2	11,4	4.566	26	121
120	Venezuela	0,691	70,6	12,8	11,1	4.811	20	118
121	Iraq	0,686	70,4	12,1	7,9	9.977	-12	122
122	Tajikistan	0,685	71,6	11,7	11,3	4.548	23	126
123	Belize	0,683	70,5	13,0	8,8	6.309	6	120
123	Morocco	0,683	74,0	14,2	5,9	7.303	1	122
125	El Salvador	0,675	70,7	12,7	7,2	8.296	-8	124
126	Nicaragua	0,667	73,8	12,6	7,1	5.625	6	129
127	Bhutan	0,666	71,8	13,2	5,2	9.438	-15	125
128	Cabo Verde	0,662	74,1	12,6	6,3	6.230	2	127
129	Bangladesh	0,661	72,4	12,4	7,4	5.472	4	128
130	Tuvalu	0,641	64,5	9,4	10,6	6.351	-2	131
131	Marshall Islands	0,639	65,3	10,2	10,9	4.620	12	131
132	India	0,633	67,2	11,9	6,7	6.590	-6	130
133	Ghana	0,632	63,8	12,0	8,3	5.745	-2	135
134	Micronesia	0,628	70,7	11,5	7,8	3.696	22	136
135	Guatemala	0,627	69,2	10,6	5,7	8.723	-20	133
136	Kiribati	0,624	67,4	11,8	8,0	4.063	14	137

HDI rank	Country	V1	V2	V3	V4	V5	V6	V7
137	Honduras	0,621	70,1	10,1	7,1	5.298	-1	138
138	Sao Tome	Principe	0,618	67,6	13,4	6,2	4.021	13
139								
139	Namibia	0,615	59,3	11,9	7,2	8.634	-23	134
140	Lao People's	0,607	68,1	10,1	5,4	7.700	-18	142
140	Timor-Leste	0,607	67,7	12,6	5,4	4.461	7	140
140	Vanuatu	0,607	70,4	11,5	7,1	3.085	23	142
143	Nepal	0,602	68,4	12,9	5,1	3.877	10	144
144	Eswatini	0,597	57,1	13,7	5,6	7.679	-21	141
145	Equatorial Guinea	0,596	60,6	9,7	5,9	12.074	-47	147
146	Cambodia 0,593	69,6	1,5	5,1		4.079	3	148
146	Zimbabwe 0,593	59,3	2,1	8,7		3.810	9	145
148	Angola 0,586	61,6	2,2	5,4		3.466	-14	149
149	Myanmar 0,585	65,7	0,9	6,4		3.851	5	145
150	Syrian Arab Republic	0,577	72,1	9,2	5,1	4.192	-2	152
151	Cameroon 0,576	60,3	3,1	6,2		3.621	6	150
152	Kenya 0,575	61,4	0,7	6,7		4.474	-6	150
153	Congo 0,571	63,5	2,3	6,2		2.889	11	153
154	Zambia 0,565	61,2	0,9	7,2		3.218	7	154
155	Solomon Islands	0,564	70,3	10,3	5,7	2.482	13	155
156	Comoros 0,558	63,4	1,9	5,1		3.142	6	156
156	Papua New Guinea	0,558	65,4	10,4	4,7	4.009	-4	157
158	Mauritania	0,556	64,4	9,4	4,9	5.075	-20	158
159	Côte d'Ivoire	0,550	58,6	10,7	5,2	5.217	-22	159
160	Tanzania	0,549	66,2	9,2	6,4	2.664	7	160
161	Pakistan	0,544	66,1	8,7	4,5	4.624	-19	161
162	Togo	0,539	61,6	13,0	5,0	2.167	12	163
163	Haiti	0,535	63,2	9,7	5,6	2.848	2	162
163	Nigeria	0,535	52,7	10,1	7,2	4.790	-22	163
165	Rwanda	0,534	66,1	11,2	4,4	2.210	6	165
166	Benin	0,525	59,8	10,8	4,3	3.409	-7	166
166	Uganda	0,525	62,7	10,1	5,7	2.181	6	166
168	Lesotho	0,514	53,1	12,0	6,0	2.700	-2	168
169	Malawi	0,512	62,9	12,7	4,5	1.466	13	169
170	Senegal	0,511	67,1	9,0	2,9	3.344	-10	170
171	Djibouti	0,509	62,3	7,4	4,1	5.025	-32	171
172	Sudan	0,508	65,3	7,9	3,8	3.575	-14	171
173	Madagascar	0,501	64,5	10,1	5,1	1.484	8	173
174	Gambia	0,500	62,1	9,4	4,6	2.172	-1	173
175	Ethiopia	0,498	65,0	9,7	3,2	2.361	-5	175
176	Eritrea	0,492	66,5	8,1	4,9	1.729	3	176
177	Guinea-Bissau	0,483	59,7	10,6	3,6	1.908	0	177
178	Liberia	0,481	60,7	10,4	5,1	1.289	7	179
179	Congo	0,479	59,2	9,8	7,0	1.076	9	180
180	Afghanistan	0,478	62,0	10,3	3,0	1.824	-2	177

HDI rank	Country	V1	V2	V3	V4	V5	V6	V7
181	Sierra Leone	0,477	60,1	9,6	4,6	1.622	-1	181
182	Guinea	0,465	58,9	9,8	2,2	2.481	-13	182
183	Yemen	0,455	63,8	9,1	3,2	1.314	1	183
184	Burkina Faso	0,449	59,3	9,1	2,1	2.118	-8	185
185	Mozambique	0,446	59,3	10,2	3,2	1.198	2	184
186	Mali	0,428	58,9	7,4	2,3	2.133	-11	186
187	Burundi	0,426	61,7	10,7	3,1	732	4	187
188	Central African	0,404	53,9	8,0	4,3	966	1	188
189	Niger	0,400	61,6	7,0	2,1	1.240	-3	189
190	Chad	0,394	52,5	8,0	2,6	1.364	-7	190
191	South Sudan	0,385	55,0	5,5	5,7	768	-1	191

Bibliografia

Agresti A, *An introduction to categorical data analysis* (John Wiley & Sons 2018).

Baldi S, "L'indice di sviluppo umano delle Nazioni Unite. Vantaggi e limiti della misurazione sintetica dello sviluppo", *Affari Sociali Internazionali n.3*, 1998.

Box GP, **Cox** DR, "An Analysis of Transformations", *Journal of the Royal Statistical Society. Series B (Methodological)*, 1964.

Clarke S, "Freedom and Happiness: Does Freedom Make People Happy?", *Journal of Political Science: Bulletin of Yerevan University*, 2022.

Cotton R, *Learning R: a step-by-step function guide to data analysis* (" O'Reilly Media, Inc" 2013).

LE RELAZIONI TVQ, "Capitolo II LE RELAZIONI TRA VARIABILI QUANTITATIVE: CORRELAZIONE E COGRADUAZIONE", *Analisi dei dati e data mining per le decisioni aziendali*, 2007.

Mariani F, **Ciommi** M, "Aggregating Composite Indicators through the Geometric Mean: A Penalization Approach", *Computation*, 2022.

Otoiu A, **Pareto** A, **Grimaccia** E, **Mazziotta** M, **Terzi** S, *Open issues in composite indicators. A starting point and a reference on some state-of-the-art issues* (vol. 3, Roma TrE-Press 2021).

Programme UND, "2022 Special Report on Human Security", *UNDP United Nations Development Programme*, 2022.

Rigon T, "Trasformazione box cox: un'analisi basata sulla verosimiglianza",

Sakia RM, "The Box-Cox Transformation Technique: A Review", *Journal of the Royal Statistical Society. Series D (The Statistician)*, 1992.

Shiba K, **Cowden** RG, **Gonzalez** N, **Lee** MT, **Lomas** T, **Lai** AY, **VanderWeele** TJ, "Global trends of mean and inequality in multidimensional wellbeing: Analysis of 1.2 million individuals from 162 countries, 2009–2019", *Frontiers in public health*, 2022.

STANTON EA, "The human development index: A history", *PERI Working Papers*, 2007.

Sitografia

"BETTER LIFE INDEX: DEFINITIONS AND METADATA" https://www.oecd.org/wise/OECD-Better-Life-Index-definitions-2021.pdf.

"CRAN R" https://cran.r-project.org.

"DATA NORMALIZATION FOR AGGREGATING TIME SERIES: THE CONSTRAINED MIN-MAX METHOD" http://www.sieds.it/wp-content/uploads/2022/01/9_01441RV_Mazziotta.pdf.

"FAO" https://www.fao.org/contact-us/en/.

"ILO" https://www.ilo.org/global/lang--en/index.htm.

"Indice composito" https://www4.istat.it/it/strumenti/metodi-e-strumenti-it/analisi.

"INSTAT" instat.it.

"La sintesi degli indicatori di qualità della vita: un approccio non compensativo" http://win.aiquav.it/QOL2010/doc/Presentazioni/2/0204_Mazziotta_Pareto.pdf.

"METODOLOGIE DI SINTESI E ANALISI DEL TERRITORIO" https://www.istat.it/it/files/2014/10/Paper_Sessione-IV_Massoli_Mazziotta_Pareto_Rinaldelli.pdf.

"Normalizzazione Box-Cox" https://www.sixsigmain.it/ebook/Capu13-4.html.

"Nota metodologica" https://i.ranker.istat.it/wr_guida_notametodologica.htm.

"OCSE" ocse.it.

"OECD" https://www.oecd.org/sdd/42495745.pdf.

"OECD" https://www.oecd.org/.

"OPEN ISSUES IN COMPOSITE INDICATORS" http://romatrepress.uniroma3.it/wp-content/uploads/2021/03/open-togmp.pdf.

"Population Division" https://www.un.org/development/desa/pd/.

"Studio di fattibilità per la costruzione di un indice dello stato di benessere dell'infanzia e dell'adolescenza nelle regioni italiane" https://www.ricercaepratica.it/r.php?v=3786&a=37709&l=347761&f=allegati/03786_2022_02/fulltext/03_Ricerca_sul_campo_indice_benessere.pdf.

"Treccani" https://www.treccani.it/enciclopedia/robustezza-statistica.

"Un indice sintetico non compensativo per la misura della dotazione infrastrutturale: un'applicazione in ambito sanitario" https://www4.istat.it/it/files/2011/12/1_2011_3.pdf.

"undp-data center" https://hdr.undp.org/data-center/human-development-index#/indicies/HDI.

"UNESCO" https://en.unesco.org/about-us/unesco-house.

"UNFPA" https://www.unfpa.org/.

"UNICEF" unicef.it.

"WHO" https://www.who.int/.

"World Bank" https://www.worldbank.org/en/home.