Dipartimento di Economia e Statistica
Corso di laurea triennale in ECOMARK

ASIMMETRIE SOCIO-ECONOMICHE E PREVISIONE DELL'INDICE DI SVILUPPO UMANO

Anno accademico2019/2020

Matilde Sciamanna
Maricola:833597

Prof. Alessandro Zini
corso di STATISTICA PER IL MARKETING

Indice

1. INTRODUZIONE

L'elaborato ha la finalità di valutare quali politiche socio-economiche possano essere più d'impatto nella riduzione delle disuguaglianze tra diversi paesi dove, come misura di tali disuguaglianze, è stato considerato l'indice di sviluppo umano.
L'indice di sviluppo umano (ISU, in inglese*: Human Development Index*, HDI) è un indicatore di sviluppo macroeconomico ed è definito come la media geometrica di tre indici di base, legati rispettivamente alla speranza di vita, al livello di istruzione e al reddito.

Ai fini dell'investigazione si è ritenuto necessario approfondire quali possano essere le principali sfumature delle tre componenti dell'indice di sviluppo umano (speranza di vita, livello di istruzione e reddito). Utilizzando delle variabili più specifiche si è cercato quindi di ottenere una suddivisione dei paesi coerente con il loro grado di sviluppo.
Successivamente si è valutato quali fattori possano essere più influenti in ottica di accrescimento di tale indice, verificando puntualmente quali delle variabili utilizzate al punto precedente influiscano in maniera significativa sull'indice.

L'analisi sopra descritta è stata condotta con un approccio scientifico, all'inizio di ogni sezione è quindi riportata una breve spiegazione teorica delle tecniche utilizzate, a seguire sono poi evidenziati e commentati i risultati empirici ottenuti attraverso un'implementazione pratica con SPSS.
Volendo rappresentare sinteticamente le analisi sviluppate, l'elaborato si può così suddividere:

- nella prima parte, al fine di stabilire se le variabili e le osservazioni selezionate fossero in grado di rappresentare efficacemente le attuali disuguaglianze tra i vari Paesi, è stata utilizzata una tecnica di analisi non supervisionata, ovvero la cluster analysis. Nello specifico, a seguito del calcolo delle componenti principali, sono state implementati sia il metodo gerarchico che quello non gerarchico (k-means)
- nella seconda parte dell'elaborato, dove l'obiettivo era invece di verificare quali variabili avessero un impatto significativo sulla crescita dell'indice di sviluppo umano, è stata utilizzata una tecnica di analisi supervisionata, ovvero la regressione lineare, avente appunto come variabile dipendente l'indice di sviluppo umano.

Il risultato finale dell'analisi è un'interpretazione ad ampio respiro dei fattori che risultano più significativi in relazione all'indice di sviluppo umano e l'individuazione di politiche macroeconomiche finalizzate appunto alla riduzione delle diseguaglianze esistenti tra i paesi.

N.B. Gli output ottenuti tramite SPSS, dato l'elevata dimensione, sono talvolta riportati in Appendice (segnalati con APPENDICE N°) o sono caricati su uno spazio cloud a cui si può accedere attraverso appositi link (segnalati nelle note a piè di pagina).

2. DESCRIZIONE DEL DATASET

Il dataset è stato costruito a partire dalle informazioni contenute nel sito Data.World[1]. Tale sito dà la possibilità di scegliere tra circa 1500 differenti indicatori sociali ed economici per più di 200 Paesi del mondo, registrati con cadenza annuale. Date le tecniche affrontate in questo corso si è ritenuto sensato concentrare l'attenzione su un'analisi cross-sectional, tralasciando quindi la componente di sviluppo temporale dei dati. Nella pratica si è quindi deciso di scaricare i dati relativi al solo 2016, anno per cui vi è un buon compromesso tra completezza e attualità dei dati.
Come si può immaginare è stata riscontrata una presenza elevatissima di dati mancanti, sia in termini di osservazioni che di variabili. Per tale motivo si è resa necessaria un'analisi esplorativa che ha permesso di selezionare 24 variabili (23 quantitative, 1 ordinale) per 120 Paesi.[2][3]
In questo studio quali-quantitativo dei dati originariamente estratti si è cercato di selezionare quelle variabili che meglio potessero rappresentare degli attributi dei tre indicatori che compongono l'indice di sviluppo umano, ovvero speranza di vita, livello di istruzione e reddito pro capite. Così facendo è stato possibile mantenere un certo tipo di legame con tale indice e al contempo approfondire i fattori che concorrono al suo calcolo.

Le variabili prese in esame sono:

1. **HDI (Human Development Index)**
   L'indice di sviluppo umano: è indice comparativo dello sviluppo dei vari Paesi calcolato tenendo conto dei diversi tassi di speranza di vita, di livello di istruzione e di reddito.

2. **Access to electricity (% of population)**
   Accesso all'elettricità: è la percentuale della popolazione con accesso all'elettricità.

3. **Birth rate, crude (per 1,000 people)**
   Il tasso di natalità grezzo: indica il numero di nascite che si verificano durante l'anno, per 1.000 abitanti stimati a metà anno.

4. **Current health expenditure per capita, current US**
   Spese sanitarie correnti pro capite espresse in dollari: spese che includono beni e servizi sanitari consumati ogni anno.

5. **Domestic general government health expenditure per capita (current US$)**
   Spesa pubblica per la salute pro capite espressa in dollari.

6. **Employment in industry (% of total employment) (modeled ILO estimate)**
   Percentuale di occupazione nel settore industriale sull'occupazione totale.

[1] https://data.world/
[2] Dataset completo: https://drive.google.com/file/d/1d3wJ95s_F0atHAo0Tj0kk31n0h455C3J/view?usp=sharing

7. **Employment in services (% of total employment) (modeled ILO estimate)**
   Percentuale di occupazione nei servizi sull'occupazione totale.

8. **Exports of goods and services (% of GDP)**
   Le esportazioni di beni e servizi: rappresentano il valore di tutti i beni e altri servizi di mercato forniti al resto del Mondo, espressi in percentuale rispetto al PIL.

9. **Foreign direct investment, net inflows (% of GDP)**
   Gli investimenti esteri diretti: sono gli afflussi netti di investimenti per acquisire una quota in un'impresa, espressi in percentuale rispetto al PIL.

10. **Individuals using the Internet (% of population)**
    Percentuale di popolazione che utilizza internet tramite: computer, telefono cellulare, TV digitale ecc.

11. **Lifetime risk of maternal death (%)**
    Percentuale di rischio di mortalità materna.

12. **People using at least basic drinking water services (% of population)**
    Percentuale di persone che utilizzano almeno i servizi idrici di base: acqua potabile, acqua convogliata, pozzi o tubi, sorgenti protette ecc.

13. **People using at least basic sanitation services (% of population)**
    Percentuale di persone che utilizzano almeno i servizi igienico-sanitari di base non condivisi.

14. **Population growth (annual %)**
    Tasso annuale di crescita della popolazione: conta tutti i residenti indipendentemente dallo status giuridico o dalla cittadinanza.

15. **Primary & secondary education, duration (years)**
    Durata, in anni, dell'istruzione primaria e secondaria.

16. **Profit tax (% of commercial profits)**
    L'imposta sul profitto: è l'ammontare delle imposte sugli utili pagati dall'azienda.

17. **Self-employed, total (% of total employment) (modeled ILO estimate)**
    Percentuale di lavoratori autonomi sui lavoratori totali: i lavoratori autonomi sono quei lavoratori la cui remunerazione dipende direttamente dagli utili derivati dai beni e servizi prodotti**.**

18. **Services, value added (% of GDP)**
    Servizi che includono il valore aggiunto nel commercio all'ingrosso e al dettaglio, trasporti e servizi pubblici, finanziario, professionali e personali, espressi in percentuale rispetto al PIL.

19. **Strength of legal rights index (0=weak to 12=strong)**
    Indice dei diritti legali: è la misura in cui le leggi in materia di garanzie e fallimenti proteggono i diritti di mutuatari e prestatori. L'indice varia da 0 a 12.

20. **Total natural resources rents (%of GDP)**
    Le rendite derivanti dalle risorse naturali sono stimate come la differenza tra il prezzo di una merce e il costo medio per produrla. Sono calcolate come la somma delle rendite del petrolio, del gas naturale, del carbone, dei minerali e delle foreste e espresse in percentuale rispetto al PIL.

21. **Unemployment, total (%of total labor force) (modeled ILO estimate)**
    La disoccupazione si riferisce alla quota della forza lavoro che è senza lavoro ma disponibile e in cerca di lavoro.

22. **Unemployment, youth total (%of total labor force ages 15-24) (modeled ILO estimate)**
    La disoccupazione giovanile si riferisce alla quota della forza lavoro di età compresa tra 15 e 24 anni senza lavoro ma disponibile e in cerca di lavoro.

23. **Urban population (%of total population)**
    Percentuale di popolazione urbana: si riferisce alle persone che vivono nelle aree urbane come definito dagli uffici statistici nazionali.

24. **Vulnerable employment, total (%of total employment) (modeled ILO estimate)**
    Percentuale di quegli occupati che hanno: salari inadeguati, bassa produttivtà e condizioni di lavoro difficili.

In linea con quanto detto nell'Introduzione, l'indice di sviluppo umano sarà utilizzato come variabile dipendente per verificare l'effettiva significatività delle altre variabili e sarà quindi introdotto nell'analisi solo nella seconda parte (regressione). Non sarà invece considerato in fase di implementazione delle tecniche non supervisionate, dove infatti il focus è quello di condurre un'analisi esplorativa sulla base di quelle che saranno poi in seguito considerate come variabili esplicative per la regressione.

3.ANALISI COMPONENTI PRINCIPALI

Cenni teorici

L'Analisi delle Componenti Principali è un metodo di analisi multivariata che consente una riduzione delle variabili investigate.

L'obbiettivo che si pone è quello di costruire delle nuove variabili, ottenute come combinazioni lineari delle variabili originarie, in modo che un numero ridotto di queste nuove variabili, tra loro incorrelate, sia in grado di spiegare una porzione rilevante della varianza totale dei dati.

La scelta del numero di componenti principali è effettuata sulla base della varianza cumulata da queste spiegata, che non deve essere inferiore ad un certo livello soglia. Bisogna quindi estrarre il minor numero di Componenti Principali, conservando quanta più significatività possibile.

Applicazione pratica in SPSS

MATRICE DI CORRELAZIONE

Attraverso la matrice di correlazione si valuta la desiderabilità di effettuare l'Analisi delle Componenti Principali.

Essa consente di osservare se tra le variabili esista un certo grado di correlazione.

Tale correlazione può essere positiva o negativa, a seconda che le due variabili investigate si modifichino nello stesso verso o in verso opposto, giungendo agli estremi ad assumere i valori rispettivamente di 1 e -1.

Dalla matrice[4] si osserva come esista un certo grado di correlazione tra le variabili, che potrebbe dare origine a problemi di multicollinearità; inoltre l'elevata dimensionalità della matrice rende difficoltosa una sua efficace analisi.

Si è quindi deciso di ricorrere al calcolo delle componenti principali (ACP) che permettono di risolvere entrambi i problemi, in quanto si può utilizzare un numero ridotto di variabili che risultano tra loro incorrelate.

Come ultima verifica, si procede implementando il Test KMO e il Test di Sfericità di Bartlett.

TEST KMO E DI SFERICITÀ DI BARTLETT

Il Test KMO indica quanta parte di varianza è spiegata da fattori comuni, mentre il Test di Bartlett opera implicitamente un confronto tra la matrice di correlazione sopra riportata e la matrice di perfetta incorrelazione, vale a dire la matrice identità.

Test di KMO e Bartlett

Misura di Kaiser-Meyer-Olkin di adeguatezza del campionamento.		,884
Test della sfericità di Bartlett	Appross. Chi-quadrato	3317,971
	gl	253
	Sign.	,000

[4] Matrice di correlazione: https://drive.google.com/file/d/1mH7CDyVgb4C1DvsP_jlRHrhx2euf6-GR/view?usp=sharing

Dato che il valore della misura di KMO è ampiamente superiore a 0,7 e che, in riferimento al test di Bartlett, l'ipotesi nulla di incorrelazione tra le variabili è ampiamente rifiutata, si può procedere con il calcolo delle componenti principali.

Per quanto riguarda l'implementazione dell'ACP su SPSS non è stata applicata nessuna rotazione e sono state considerate le sole componenti con autovalori maggiori di 1 (come restituito di default dal software).

TABELLA DI COMUNALITA'

La tabella di comunità riporta le percentuali di varianza di ciascuna variabile spiegate dalle componenti principali estratte.

Comunalità

	Iniziale	Estrazione
Access to electricity (% of population)	1,000	,884
Birth rate, crude (per 1,000 people)	1,000	,914
Current health expenditure per capita current US	1,000	,888
Domestic general government health expenditure per capita (current US$)	1,000	,902
Employment in industry (% of total employment) (modeled ILO estimate)	1,000	,802
Employment in services (% of total employment) (modeled ILO estimate)	1,000	,862
Exports of goods and services (% of GDP)	1,000	,796
Foreign direct investment, net inflows (% of GDP)	1,000	,826
Individuals using the Internet (% of population)	1,000	,890
Lifetime risk of maternal death (%)	1,000	,797
People using at least basic drinking water services (% of population)	1,000	,900
People using at least basic sanitation services (% of population)	1,000	,890
Population growth (annual %)	1,000	,753
Primary & secondary education, duration (years)	1,000	,811
Profit tax (% of commercial profits)	1,000	,686
Self-employed, total (% of total employment) (modeled ILO estimate)	1,000	,909
Services, value added (% of GDP)	1,000	,696
Strength of legal rights index (0=weak to 12=strong)	1,000	,552
Total natural resources rents (% of GDP)	1,000	,643
Unemployment, total (% of total labor force) (modeled ILO estimate)	1,000	,955
Unemployment, youth total (% of total labor force ages 15-24) (modeled ILO estimate)	1,000	,962
Urban population (% of total population)	1,000	,734
Vulnerable employment, total (% of total employment) (modeled ILO estimate)	1,000	,920

Metodo di estrazione: Analisi dei componenti principali.

Dalla tabella soprastante si evince che per 15 delle 23 variabili iniziali la percentuale di varianza spiegata è superiore all'80%, soglia decisamente elevata. Le rimanti presentano comunque una soglia superiore al 60% ad eccezione della variabile ordinale (*Strength of legal rights index*) che si attesta al 55%.

TABELLA VARIANZA SPIEGATA

Varianza totale spiegata

	Autovalori iniziali			Caricamenti somme dei quadrati di estrazione		
Componente	Totale	% di varianza	% cumulativa	Totale	% di varianza	% cumulativa
1	10,856	47,199	47,199	10,856	47,199	47,199
2	2,444	10,628	57,827	2,444	10,628	57,827
3	1,716	7,463	65,290	1,716	7,463	65,290
4	1,553	6,754	72,044	1,553	6,754	72,044
5	1,390	6,044	78,089	1,390	6,044	78,089
6	1,015	4,411	82,500	1,015	4,411	82,500
7	,747	3,248	85,748			
8	,611	2,655	88,402			
9	,517	2,247	90,649			
10	,392	1,703	92,352			
11	,333	1,447	93,799			
12	,286	1,243	95,043			
13	,252	1,094	96,136			
14	,208	,906	97,042			
15	,176	,767	97,809			
16	,146	,636	98,444			
17	,087	,378	98,823			
18	,067	,293	99,116			
19	,061	,265	99,382			
20	,059	,255	99,636			
21	,050	,219	99,855			
22	,031	,134	99,990			
23	,002	,010	100,000			

Metodo di estrazione: Analisi dei componenti principali.

L'Analisi delle Componenti Principali restituisce 23 nuove variabili, tante quante erano le va riabili originarie; qualora infatti tutte le componenti principali fossero estratte, esse spiegherebbero il 100% della varianza.

Tuttavia dato che l'obiettivo, oltre ad ottenere variabili tra loro incorrelate, è quello di operare una riduzione della dimensionalità, saranno prese in considerazione solo un numero ridotto di tali componenti.

I dati della tabella sopra riportata fanno emergere come le prime 6 componenti, associate ai maggiori autovalori, spiegano cumulativamente circa l'82% della varianza.

Volendo ridurre ulteriormente il numero di variabili, e al contempo catturare una buona parte di variabilità, si è deciso di prendere in considerazione solo le prime 4 componenti principali che spiegano cumulativamente il 72% della varianza.

Dove, nel dettaglio, la prima spiega il 47% della variabilità, la seconda il 10% e la terza e la quarta rispettivamente l'8% e il 7%.

MATRICE DEI COMPONENTI

La Matrice dei Componenti riporta nelle colonne le 6 componenti principali estratte e nelle righe le 23 variabili iniziali; nella matrice sono riportati dei valori, positivi o negativi, che indicano la correlazione di ciascuna delle componenti estratte con le variabili originarie.

Matrice dei componenti[a]

	Componente					
	1	2	3	4	5	6
Access to electricity (% of population)	,882	-,222	-,088	-,215	-,052	-,013
Birth rate, crude (per 1,000 people)	-,929	,038	-,042	,140	,102	,134
Current health expenditure per capita current US	,616	,628	,174	,097	,273	,028
Domestic general government health expenditure per capita (current US$)	,621	,630	,148	,107	,295	,005
Employment in industry (% of total employment) (modeled ILO estimate)	,571	-,254	-,446	-,208	,002	-,412
Employment in services (% of total employment) (modeled ILO estimate)	,888	,079	,056	,175	,118	,145
Exports of goods and services (% of GDP)	,453	,290	-,222	,440	-,508	-,076
Foreign direct investment, net inflows (% of GDP)	,191	,295	,060	,407	-,708	,178
Individuals using the Internet (% of population)	,913	,214	-,090	,023	,048	,033
Lifetime risk of maternal death (%)	-,826	,215	,083	,233	,070	-,052
People using at least basic drinking water services (% of population)	,920	-,162	-,067	-,152	-,016	-,026
People using at least basic sanitation services (% of population)	,917	-,123	-,099	-,136	,007	,068
Population growth (annual %)	-,607	,197	-,418	,176	,199	,316
Primary & secondary education, duration (years)	-,176	,466	,091	,289	,358	-,585
Profit tax (% of commercial profits)	-,238	,149	,502	-,460	,128	,357
Self-employed, total (% of total employment) (modeled ILO estimate)	-,937	,007	,144	-,032	-,086	-,031
Services, value added (% of GDP)	,750	,178	,199	,056	,004	,244
Strength of legal rights index (0=weak to 12=strong)	,077	,291	,480	-,258	-,285	-,289
Total natural resources rents (% of GDP)	-,525	-,055	-,492	,223	,240	,121
Unemployment, total (% of total labor force) (modeled ILO estimate)	,271	-,622	,460	,510	,151	-,006
Unemployment, youth total (% of total labor force ages 15-24) (modeled ILO estimate)	,372	-,658	,380	,468	,162	-,028
Urban population (% of total population)	,780	,095	-,170	,108	,218	,168
Vulnerable employment, total (% of total employment) (modeled ILO estimate)	-,945	,018	,132	-,036	-,085	-,042

Metodo di estrazione: Analisi dei componenti principali.

Ad esempio, il primo dato, 0,882, indica che la prima componenti principale è positivamente correlata con la prima variabile (*Access to electricity*). Nella seconda riga invece troviamo una correlazione negativa tra la prima componente principale e la seconda variabile (*Birth rate, crude*) pari a -0,929.

Ragionando in termini di intera componente, risulta chiaro come la prima componente mostri un grado di correlazione, valutato in termini assoluti, elevato con la maggioranza delle variabili.
In particolare si denota un elevato legame lineare diretto con le variabili: *Access to electricity, Employment in services, Individuals using the Internet, People using at least basic drinking water services People using at least basic sanitation services;* mentre è forte il legame lineare inverso con *Birth rate, Lifetime risk of maternal death,Self-employed, total,Vulnerable employment, total.*

Cercando quindi di fornire un'interpretazione blanda di tale componente, essa può essere sintetizzata come la combinazione lineare di quei fattori che rappresentano la disponibilità di beni/servizi di primaria utilità e delle variabili che, valorizzate negativamente, esprimono la precarietà della vita in un determinato Paese.

MATRICE DEI COEFFICIENTI DI PUNTEGGI COMPONENTI

La Matrice dei Coefficienti di Punteggi dei Componenti riporta, come in precedenza, le 6 Componenti Principali estratte nelle colonne e le 23 variabili originarie nelle righe; all'interno della matrice sono invece riportati In questo caso i pesi, associati a ciascuna variabile inziale, utilizzati nella combinazione lineare che determina le componenti.

Analizzando la tabella sottostante possiamo notare che il primo dato contenuto nella prima cella (0,081) indica che la variabile *Access to electricity* assume all'interno della prima componente principale un peso di pari entità, che seppur non preponderante (si discosta significativamente dal valore massimo assumibile[5] pari a 1) risulta comunque tra i maggiori.

[5] Quando il peso è pari a 1 la componente principale coincide con la variabile.

Matrice dei coefficienti di punteggi dei componenti

	Componente					
	1	2	3	4	5	6
Access to electricity (% of population)	,081	-,091	-,051	-,138	-,037	-,013
Birth rate, crude (per 1,000 people)	-,086	,016	-,025	,090	,073	,132
Current health expenditure per capita current US	,057	,257	,101	,063	,197	,028
Domestic general government health expenditure per capita (current US$)	,057	,258	,086	,069	,212	,005
Employment in industry (% of total employment) (modeled ILO estimate)	,053	-,104	-,260	-,134	,001	-,406
Employment in services (% of total employment) (modeled ILO estimate)	,082	,032	,033	,112	,085	,143
Exports of goods and services (% of GDP)	,042	,119	-,130	,283	-,365	-,075
Foreign direct investment, net inflows (% of GDP)	,018	,121	,035	,262	-,509	,176
Individuals using the Internet (% of population)	,084	,087	-,052	,015	,034	,033
Lifetime risk of maternal death (%)	-,076	,088	,048	,150	,050	-,051
People using at least basic drinking water services (% of population)	,085	-,066	-,039	-,098	-,011	-,026
People using at least basic sanitation services (% of population)	,085	-,050	-,058	-,088	,005	,067
Population growth (annual %)	-,056	,081	-,243	,113	,143	,311
Primary & secondary education, duration (years)	-,016	,191	,053	,186	,258	-,576
Profit tax (% of commercial profits)	-,022	,061	,293	-,296	,092	,352
Self-employed, total (% of total employment) (modeled ILO estimate)	-,086	,003	,084	-,020	-,062	-,030
Services, value added (% of GDP)	,069	,073	,116	,036	,003	,240
Strength of legal rights index (0=weak to 12=strong)	,007	,119	,279	-,166	-,205	-,285
Total natural resources rents (% of GDP)	-,048	-,022	-,287	,143	,173	,119
Unemployment, total (% of total labor force) (modeled ILO estimate)	,025	-,255	,268	,328	,109	-,006
Unemployment, youth total (% of total labor force ages 15-24) (modeled ILO estimate)	,034	-,269	,221	,301	,117	-,028
Urban population (% of total population)	,072	,039	-,099	,070	,156	,166
Vulnerable employment, total (% of total employment) (modeled ILO estimate)	-,087	,007	,077	-,023	-,061	-,042

Metodo di estrazione: Analisi dei componenti principali

Osservando più in generale i pesi associati alla prima componente principale si può notare come essi rispecchino quanto visto prima in termini di correlazione tra tale componente e le variabili originali, sia in termini di direzione che di gerarchia.

In conclusione, di questa prima fase di analisi le componenti principali sono state salvate come variabili. Pertanto la matrice dei dati originariamente di dimensioni 120 x 23 diviene ora una matrice di dimensioni 120 x 4 poiché, come prima esplicitato, le componenti principali prese in considerazione per la cluster analysis saranno le prime 4.

4. CLUSTER ANALYSIS

La cluster analysis, o analisi dei gruppi, è un insieme di tecniche di analisi multivariata dei dati volte alla selezione e al raggruppamento di unità statistiche.
Si tratta di una procedura tipicamente esplorativa il cui obiettivo è quello di riunire unità tra loro eterogenee in più sottoinsiemi omogenei, minimizzando le distanze all'interno dei gruppi e massimizzando quelle tra i gruppi.
La scelta ottimale del numero di gruppi rappresenta un compromesso tra due esigenze opposte: da un lato la volontà di giungere ad un numero il più ristretto possibile di gruppi, dall'altro la necessità di non racchiudere all'interno di uno stesso gruppo entità con caratteristiche profondamente distinte.
I metodi di classificazione più comuni sono:

- Gerarchico
- Non gerarchico

Gerarchico

Cenni teorici

Attraverso il clustering gerarchico viene costruita una gerarchia di partizioni caratterizzate da un numero crescente/decrescente di gruppi, visualizzabili mediante una rappresentazione ad albero (dendogramma), in cui sono rappresentati i passaggi si accorpamento/divisione dei gruppi.
Questo metodo consente quindi di ottenere una panoramica completa di tutti i vari raggruppamenti intervenuti per passare dall'articolazione in n gruppi a quella in un solo gruppo (metodo Top-Down o divisivo), o viceversa, nel passaggio da un unico gruppo alla suddivisione in n gruppi (metodo Bottom-Up o aggregativo).

Applicazione pratica in SPSS

L'analisi è stata affrontata utilizzando il Metodo di Ward[6] e la distanza euclidea, sulla base delle 4 Componenti Principali calcolate in precedenza.

Riepilogo elaborazione casi[a,b]

Casi					
Valido		Mancante		Totale	
N	Percentuale	N	Percentuale	N	Percentuale
120	100,0	0	,0	120	100,0

a. Distanza euclidea utilizzata

b. Legame Ward

[6] Metodo di Ward: classificazione gerarchica tramite la minimizzazione della varianza delle variabili entro ciascun gruppo.

Dal primo output ottenuto si evince che è stato possibile l'assegnazione sequenziale ai vari cluster per tutte le osservazioni.
Nella tabella sottostante vengono riportati i vari passaggi del raggruppamento: si è partiti da un numero di cluster pari a 120, uno per ciascun Paese, quindi progressivamente l'algoritmo ha provveduto ad aggregarli giungendo in ultima fase ad ottenere un unico gruppo.

Data l'elevata numerosità del dataset non è possibile mostrare output completo, verranno quindi elencati qui i primi e gli ultimi 15 valori mentre in appendice sarà possibile consultare la tabella completa (APPENDICE N°1)

Pianificazione di agglomerazione

	Combinato in cluster			Stadio prima apparizione cluster		Stadio
Stadio	Cluster 1	Cluster 2	Coefficienti	Cluster 1	Cluster 2	successivo
1	13	20	,108	0	0	80
2	58	89	,227	0	0	44
3	88	94	,384	0	0	84
4	105	116	,550	0	0	32
5	36	47	,718	0	0	24
6	52	83	,894	0	0	36
7	56	65	1,071	0	0	40
8	29	33	1,255	0	0	41
9	10	17	1,444	0	0	69
10	112	113	1,638	0	0	72
11	103	120	1,841	0	0	21
12	75	86	2,051	0	0	42
13	62	73	2,265	0	0	57
14	61	78	2,479	0	0	38
15	11	18	2,699	0	0	83
…	…	…	…	…	…	…
105	36	50	57,471	85	93	115
106	88	95	59,591	95	43	113
107	45	52	62,085	97	98	111
108	7	31	64,638	91	77	116
109	22	26	67,622	96	104	117
110	1	2	70,730	0	103	114
111	25	45	73,925	101	107	115
112	81	93	77,196	99	100	113
113	81	88	80,877	112	106	119
114	1	3	85,351	110	102	118
115	25	36	91,488	111	105	116
116	7	25	100,992	108	115	117
117	7	22	115,199	116	109	118
118	1	7	132,607	114	117	119
119	1	81	155,258	118	113	0

Le prime due colonne rappresentano le due unità statistiche che vengono raggruppate in ciascun passaggio, mentre le ultime tre colonne tengono traccia degli stadi precedenti o successivi in cui le varie unità compaiono.
La colonna "coefficienti" indica la distanza a cui vengono uniti i due cluster. A partire da essa, focalizzando l'attenzione sugli ultimi step di aggregazione, è possibile calcolare gli incrementi sia in termini assoluti che percentuali.

Stadio	Distanza	Incremento	Incremento percentuale	N°cluster
114	85,351	4,474	5,53%	6
115	91,488	6,137	7,19%	5
116	100,992	9,504	10,39%	4
117	115,199	14,207	14,07%	3
118	132,607	17,408	15,11%	2
119	155,258	22,651	17,08%	1

Gli incrementi riportati nella tabella sopra permettono di operare una scelta ottimale del numero di gruppi attraverso un criterio numerico.
Si decide, infatti, di terminare il processo di agglomerazione nel momento in cui, raggiunto un numero di cluster abbastanza ridotto in proporzione alla numerosità della popolazione in analisi, l'unione di ulteriori due gruppi coincide con un valore dell'incremento della distanza abbastanza superiore ai precedenti.
Questa scelta è operata in quanto la continuazione del processo di agglomerazione implicherebbe l'unione di gruppi con caratteristiche eterogenee tra loro, e dunque non rispetterebbe l'obbiettivo del metodo gerarchico è di mantenere le caratteristiche omogene all'interno del gruppo.

Per dare una rappresentazione più intuitiva di quanto appena descritto si può fare ricorso allo *Scree plot,* dove è sull'asse delle ascisse sono inserite le distanze di agglomerazione e sull'asse delle ordinate il numero di cluster.[7]

[7]Vista l'elevata numerosità del campione si è scelto si considerare nel grafico gli ultimi 20 cluster e le rispettive distanze, così da mostrare al meglio l'andamento della spezzata.

In figura è quindi rappresentato l'andamento della spezzata creata incrociando le distanze tra i cluster con il numero di cluster. La curva presenta una forte inclinazione iniziale e un successivo appiattimento che la porta a diventare quasi orizzontale.
In particolare, il cambio di pendenza è osservato in prossimità dei gruppi 4 e 5 gruppi dove, come riscontrato in precedenza, si verificano elevati incrementi nelle distanze di aggregazione.

Per decidere il numero effettivo di cluster da selezionare con un criterio numerico oggettivo, si è fatto ricorso agli incrementi percentuali delle distanze calcolati in precedenza, optando infine per l'utilizzo di 4 gruppi in quanto: passare da 5 a 4 gruppi comporta un incremento pari al 10,39%, passando invece da 4 a 3 si riscontra invece un incremento del 14,07% e gli incrementi successivi presentano valori in linea con quest'ultimo.
In conclusione per tutto quanto appena detto, si è deciso di optare per una divisione in 4 cluster.

Una volta stabilito il numero di gruppi è sicuramente interessante andare a verificare il contenuto di questi cluster in termini di unità statistiche.
Per fare ciò si è fatto ricorso al dendogramma[8] (APPENDICE N°2) che rappresenta una sintesi grafica del risultato ottenuto dall'analisi gerarchica; dove il "taglio" per la divisione in 4 gruppi è operato indicativamente a 15. Il dettaglio dell'appartenenza è inoltre riportato nella tabella APPENDICE N°3.

Riassumendo, la numerosità di ogni gruppo risulta:

Gruppo	N° paesi
1	21
2	56
3	16
4	27

dove:

- Il Gruppo1 contiene 21 Paesi, i quali posso essere classificati come "Paesi più sviluppati"; troviamo infatti USA, Lussemburgo, Svizzera ecc. Per questi paesi gli indicatori economici presi in considerazione assumono i valori migliori.
- Il Gruppo 3 è quello con numerosità più ristretta e contiene quei Paesi considerati sviluppati ma che nell'ultimo periodo hanno avuto delle crisi, soprattutto a livello economico; ad esempio troviamo: Italia, Grecia e Spagna
- Il Gruppo 2, è il gruppo più numeroso dove troviamo la quasi totalità dei Paesi in via di sviluppo, ad esempio Cina e India, e anche qualche paese che però presenta dei valori negli indicatori non pessimi
- Il Gruppo 4 è invece popolato dai Paesi dell'Africa centrale, i quali presentano valori disastrosi negli indicatori presi in esame e risulta infatti il più distante da tutti gli altri gruppi

[8] https://drive.google.com/file/d/1it2IrSqdwKBAJuKgAlp1gzCSDkphnbRC/view?usp=sharing

Seppur sia presente qualche anomalia nei raggruppamenti, si può ritenere soddisfacente l'esito della clusterizzazione. Infatti, gli indicatori utilizzati hanno permesso di giungere ad un raggruppamento dei vari Paesi coerente con quelle che sono le dinamiche geopolitiche attuali.
Per approfondire quanto emerso e per verificarne ulteriormente l'affidabilità si è deciso di implementare anche una tecnica di cluster analysis non gerarchica, ovvero quella che è chiamato il metodo delle k-medie.

Non gerarchico

Cenni teorici

Trattasi di un metodo di tipo partitivo in cui si cerca di determinare una partizione degli N oggetti in K gruppi che ottimizzi un criterio prefissato. Le diverse partizioni sono determinate, partendo da quella iniziale, spostando poi in successione i singoli oggetti, secondo criteri prefissati, fino a raggiunge una situazione in cui lo spostamento di singoli elementi non migliorerebbe più il valore della funzione obiettivo. A differenza dei metodi gerarchici, l'assegnazione di una oggetto a un cluster non è irrevocabile, le unità vengono riassegnate a un diverso cluster se l'allocazione inziale risulta inappropriata.
L'obiettivo, quindi, è quello di minimizzare la distanza tra ogni unità e il centroide del cluster, attraverso un'appropriata procedura iterativa.

Applicazione pratica in SPSS

Nell'analisi non gerarchica è stato utilizzato il metodo della k-means, il quale segue i passi generali di una procedura non gerarchica e che, a seguito dell'assegnazione di una nuova unità ad un gruppo, ricalcola il baricentro del nuovo e del vecchio gruppo (da cui l'unità statistica è stata rimossa).

Dovendo scegliere un numero di gruppi a priori e avendo già verificato in ambito gerarchico che 4 gruppi restituiscono un buon risultato, anche in questo caso verranno presi in considerazione 4 gruppi.

Centri cluster iniziali

	Cluster			
	1	2	3	4
PC1	1,58712	,77699	1,30308	,47643
PC2	3,00751	,17431	2,38821	-2,29982
PC3	-,46181	-4,12288	1,80010	1,83114
PC4	4,21416	-,65450	-1,16709	1,91645

Tali centri vengono progressivamente modificati man mano che l'algoritmo opera, date le continue modifiche nella composizione dei gruppi.

Cronologia delle iterazioni[a]

	Modifica nei centri del cluster			
Iterazione	1	2	3	4
1	1,503	2,927	2,771	2,543
2	,000	,210	,202	,272
3	,805	,183	,115	,150
4	,854	,100	,130	,123
5	,652	,091	,218	,060
6	,423	,104	,218	,151
7	,293	,036	,255	,075
8	,000	,150	,225	,000
9	,000	,113	,277	,146
10	,000	,091	,285	,255
11	,000	,098	,236	,000
12	,000	,000	,000	,000

a. Convergenza raggiunta grazie all'assenza o al numero limitato di modifiche nei centri del cluster. La modifica di coordinata assoluta massima per un centro è ,000. L'iterazione corrente è 12. La distanza minima tra i centri iniziali è 5,672.

In questo caso l'algoritmo opera 12 interazioni, giungendo a definire i nuovi centri sotto riportati.

Centri finali del cluster

	Cluster			
	1	2	3	4
PC1	1,10440	,17118	-1,58664	,48031
PC2	1,43721	-,39579	,51902	-1,21707
PC£	,35089	-,50912	,09922	1,00296
PC4	,20641	-,57107	,41083	,91071

I 4 gruppi ottenuti, il cui dettaglio relativo alla composizione può essere visualizzato In APPENDICEN° 4, presentano le seguenti numerosità:

Numero di casi in ciascun cluster

Cluster	1	21,000
	2	55,000
	3	26,000
	4	18,000
Valido		120,000
Mancante		,000

Volendo sintetizzare quanto ottenuto in termini di raggruppamento, si può dire che:

- Il Cluster 1 può essere identificato, anche in questo caso, come il gruppo dei Paesi sviluppati, ne fanno parte America, Lussemburgo, Svizzera ecc.
- Il Clusetr 4 rappresenta il Paesi sviluppati che hanno affrontato difficoltà economiche nell'ultimo periodo (gruppo 3 del cluster gerarchico)
- Il Cluster 2 ha sempre la numerosità più elevata e rappresenta come prima i Paesi in via di sviluppo.
- Nel Cluster 3 troviamo tutti i Paesi dell'Africa centrale (gruppo 4 del cluster gerarchico)

A conclusione del procedimento di analisi, si propone la tabella ANOVA, che fornisce informazioni riguardanti la diversa influenza che hanno le varie componenti principali nell'individuazione dei clusters. Le variabili che differenziano meglio i gruppi sono quelle che presentano un valore di F elevato.

ANOVA

	Cluster		Errore			
	Media quadratica	gl	Media quadratica	gl	F	Sign.
PC1	32,277	3	,191	116	168,890	,000
PC2	28,553	3	,287	116	99,343	,000
PC3	11,735	3	,722	116	16,245	,000
PC4	12,716	3	,697	116	18,244	,000

I test F devono essere utilizzati solo per scopi descrittivi perché i cluster sono stati scelti per massimizzare le differenze tra i casi in cluster differenti. I livelli di significatività osservati non sono corretti per tale motivo e, pertanto, non possono essere interpretati come test dell'ipotesi che le medie dei cluster siano uguali.

Nella nostra indagine possiamo affermare che sono la prima e la seconda componente principale ad aver influito maggiormente nella determinazione dei gruppi, mentre le componenti principali 3 e 4 hanno contribuito in maniera marginale. Poiché i gruppi non risultano equi-numerosi, non si procede con la valutazione delle significatività.

Da questa prima fase di analisi emerge chiaramente come le unità statistiche e le variabili selezionate, a seguito del calcolo delle componenti principali, abbiano fornito riscontri positivi in termini cluster analysis. Infatti, sia utilizzando il metodo gerarchico e che con quello delle k-medie, si ottengono dei raggruppamenti contenti Paesi con caratteristiche molto simili, anche per quanto riguarda fattori non recepiti dagli indicatori selezionati.

Essendo che l'intento di questo studio non è solo quello di cogliere le asimmetrie socio-economiche presenti tra i vari Paesi, ma quello di proporvi una soluzione, si è deciso di procedere con l'implementazione di una regressione, in cui gli indicatori finora utilizzati fungano da variabili esplicative per la previsione dell'Indice di sviluppo umano. In modo tale da poter individuare ed interpretare quelle variabili che sono più significative per la sua crescita.

5.MODELLO DI REGRESSIONE LINEARE MULTIPLA

Cenni teorici

La teoria della regressione lineare multipla risponde all'obiettivo di studiare la dipendenza di una variabile quantitativa Y da un insieme di *k* variabili esplicative, ipotizzando che tra queste ultime e la variabile dipendente esista un legame di tipo lineare.

La relazione tra le variabili esplicative e la variabile dipendente può essere scritta come: $Y=\beta 0 + \beta 1X1 + \beta 2X2 +...+ \beta kXk +\varepsilon$ (oppure in forma compatta $Y=X\beta+\varepsilon$) nella quale dovranno essere stimati i parametri β.

A tale scopo è necessario un campione di *n* osservazioni, per le quali siano registrati valori sia delle variabili esplicative che della variabile dipendente. Si parte quindi da una matrice *n* x *(k+1)* per intrepretare il legame congiunto tra le n osservazioni e la variabile risposta.
L'obiettivo di questa tecnica supervisionata è quello di prevedere i valori assunti da una variabile dipendente, a partire dalla conoscenza di quelli osservati, su più variabili indipendenti.

Applicazione pratica in SPSS per passaggi

L'intento di questa sezione è quello di implementare un modello di regressione multipla capace di spiegare la dipendenza del HDI (*Human Development Index*= Indice di sviluppo umano) con le variabili esplicative usate in precedenza per le cluster analysis.
Dato che l'utilizzo delle componenti principali comporterebbe molti problemi al momento dell'interpretazione dei coefficienti, si è deciso di ricorre al dataset originale, così come descritto nel punto **2.**
Si rende perciò necessario uno studio della possibile collinearità presente tra le variabili esplicative, poiché potrebbe portare a stime distorte dei coefficienti di regressione per le variabili troppo correlate.

A tal fine si è fatto ricorso al calcolo del *variance inflation factor* (VIF), un'indice che stabilisce, per ciascuna esplicativa, quanto la varianza della stima del coefficiente di bontà del modello sia aumentata dalla presenza di multicollinearità dovuta a quella variabile.
Solitamente un valore del VIF, associato ad una esplicativa, superiore a 10 sta a significare che quella variabile è simile ad una combinazione lineare di altre esplicative. Per tale motivo sono state eliminate, iterativamente, tutte quelle variabili esplicative che avevano VIF superiore a 10 e che erano quindi causa di multicollinearità nel modello.

Nel dettaglio, le variabili che a fine di questo procedimento risultano con in VIF < 10 sono:

Modello		Statistiche di collinearità	
		Tolleranza	VIF
1	Access to electricity (% of population)	,192	5,209
	Current health expenditure per capita current US	,321	3,117
	Employment in industry (% of total employment) (modeled ILO estimate)	,461	2,171
	Employment in services (% of total employment) (modeled ILO estimate)	,161	6,215
	Exports of goods and services (% of GDP)	,463	2,159
	Foreign direct investment, net inflows (% of GDP)	,563	1,776
	Individuals using the Internet (% of population)	,148	6,754
	Lifetime risk of maternal death (%)	,287	3,480
	Population growth (annual %)	,492	2,031
	Primary & secondary education, duration (years)	,663	1,509
	Profit tax (% of commercial profits)	,695	1,438
	Services, value added (% of GDP)	,387	2,585
	Strength of legal rights index (0=weak to 12=strong)	,798	1,254
	Total natural resources rents (% of GDP)	,512	1,955
	Unemployment, total (% of total labor force) (modeled ILO estimate)	,665	1,504
	Urban population (% of total population)	,224	4,474

a. Variabile dipendente: HDI

Queste 16 variabili permettono comunque di indagare efficacemente quali siano gli indicatori che più influiscono sulla crescita del HDI.

Si tenga conto che, a causa dell'assenza di alcuni valori appunto dell'HDI per alcune osservazioni, il campione è stato ristretto a 110 osservazioni, numerosità più che sufficiente per l'ottenimento di stime dei coefficienti consistenti.

Statistiche preliminari

Statistiche descrittive

	N	Intervallo	Minimo	Massimo	Media		Deviazione std.	Varianza	Asimmetria		Curtosi	
	Statistica	Statistica	Statistica	Statistica	Statistica	Errore standard	Statistica	Statistica	Statistica	Errore standard	Statistica	Errore standard
HDInuovo	110	58,60	36,50	95,10	73,1673	1,45300	15,23915	232,232	-,474	,230	-,784	,457
Numero di casi validi (listwise)	110											

Occorre a questo punto precisare che per rendere più agevole l'interpretazione dei coefficienti della regressione implementata di seguito, si è deciso di moltiplicare l'HDI per 100 (HIDnuovo). Tale trasformazione equivale a moltiplicare per 100 i vari parametri β del modello, senza che questo intacchi minimamente il loro livello di significatività.

Il nuovo Indice di Sviluppo Umano avrà quindi un intervallo di variazione non più tra 0 e 1, ma tra 0 e 100. In particolare esso presenta una media pari a 73,16 con una deviazione standard pari a 15,239; notiamo inoltre una leggera asimmetria negativa e una presenza minima di platicurtosi[9].

[9] Platicurtosi: allontanamento dalla normalità distributiva rispetto alla quale si verifica un maggiore appiattimento (distribuzione platicurtica)

Modello completo

Si riportano di seguito una sintesi dei principali output della regressione delle 16 variabili esplicative, individuate in precedenza, sull'HDI.

Variabili immesse/rimosse[a]

Modello	Variabili immesse	Variabili rimosse	Metodo
1	Urban population (% of total population) Strength of legal rights index (0=weak to 12=strong) Primary & secondary education, duration (years) Foreign direct investment, net inflows (% of GDP) Unemployment, total (% of total labor force) (modeled ILO estimate) , Profit tax (% of commercial profits) Total natural resources rents (% of GDP) Employment in industry (% of total employment) (modeled ILO estimate) Population growth (annual %) Current health expenditure per capita current US Exports of goods and services (% of GDP) Services, value added (% of GDP) Lifetime risk of maternal death (%) , Access to electricity (% of population) Employment in services (% of total employment) (modeled ILO estimate) , Individuals using the Internet (% of population)[b]	.	Inserisci

a. Variabile dipendente: HDInuovo

b. Sono state immesse tutte le variabili richieste.

Riepilogo del modello[b]

Modello	R	R-quadrato	R-quadrato adattato	Errore std. della stima
1	,976[a]	,953	,945	3,56494

a. Predittori: (costante), Urban population (% of total population) , Strength of legal rights index (0=weak to 12=strong) , Primary & secondary education, duration (years), Foreign direct investment, net inflows (% of GDP), Unemployment, total (% of total labor force) (modeled ILO estimate) , Profit tax (% of commercial profits), Total natural resources rents (% of GDP) , Employment in industry (% of total employment) (modeled ILO estimate) , Population growth (annual %) , Current health expenditure per capita current US, Exports of goods and services (% of GDP) , Services, value added (% of GDP) , Lifetime risk of maternal death (%) , Access to electricity (% of population) , Employment in services (% of total employment) (modeled ILO estimate) , Individuals using the Internet (% of population)

b. Variabile dipendente: HDInuovo

ANOVA[a]

Modello		Somma dei quadrati	gl	Media quadratica	F	Sign.
1	Regressione	24131,347	16	1508,209	118,675	,000[b]
	Residuo	1181,915	93	12,709		
	Totale	25313,262	109			

a. Variabile dipendente: HDInuovo

b. Predittori: (costante), Urban population (% of total population) , Strength of legal rights index (0=weak to 12=strong) , Primary & secondary education, duration (years), Foreign direct investment, net inflows (% of GDP), Unemployment, total (% of total labor force) (modeled ILO estimate) , Profit tax (% of commercial profits), Total natural resources rents (% of GDP) , Employment in industry (% of total employment) (modeled ILO estimate) , Population growth (annual %) , Current health expenditure per capita current US, Exports of goods and services (% of GDP) , Services, value added (% of GDP) , Lifetime risk of maternal death (%) , Access to electricity (% of population) , Employment in services (% of total employment) (modeled ILO estimate) , Individuals using the Internet (% of population)

Dalla seconda tabella emerge con chiarezza la bontà del modello, infatti sia R-quadrato che R-quadrato aggiustato sono al 95%. Le variabili esplicative riescono a spiegare molto bene l'indice di sviluppo umano.
Ad ulteriore conferma di ciò si può notare che nella tabella ANOVA il p-value asscoiato alla statistica-test di Fisher è pari a 0, si rifiuta quindi ampiamente l'ipotesi nulla congiunta di non significatività dei coefficienti della regressione (il modello è significativo per un α di 0,001).

Si può quindi procedere con l'analisi della significatività dei singoli coefficienti:

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati		
		B	Errore stand	Beta	t	Sign.
1	(Costante)	37,469	9,171		4,085	,000
	Access to electricity (% of population)	,102	,031	,169	3,301	,001
	Current health expenditure per capita current US	,001	,000	,144	3,643	,000
	Employment in industry (% of total employment) (modeled ILO estimate)	-,007	,065	-,003	-,105	,916
	Employment in services (% of total employment) (modeled ILO estimate)	,071	,048	,082	1,466	,146
	Exports of goods and services (% of GDP)	,026	,017	,050	1,522	,131
	Foreign direct investment, net inflows (% of GDP)	-,104	,044	-,070	-2,340	,021
	Individuals using the Internet (% of population)	,214	,031	,400	6,865	,000
	Lifetime risk of maternal death (%)	-2,365	,599	-,165	-3,952	,000
	Population growth (annual %)	-1,715	,415	-,132	-4,134	,000
	Primary & secondary education, duration (years)	,624	,639	,027	,976	,332
	Profit tax (% of commercial profits)	-,069	,049	-,038	-1,414	,161
	Services, value added (% of GDP)	,110	,051	,078	2,155	,034
	Strength of legal rights index (0=weak to 12=strong)	-,044	,131	-,009	-,339	,735
	Total natural resources rents (% of GDP)	,076	,065	,037	1,180	,241
	Unemployment, total (% of total labor force) (modeled ILO estimate)	-,070	,085	-,023	-,826	,411
	Urban population (% of total population)	,023	,032	,035	,730	,467

a. Variabile dipendente: HDInuovo

La tabella soprariportata contiene per ciascuna esplicativa: la stima del coefficiente, lo std error, il coefficiente standardizzato, la statistica-test t e il p-value ad essa associati.

Prima di procedere con l'interpretazione dei parametri stimati può essere interessante verificare le performance di un eventuale modello ristretto, che tenga in considerazione le sole variabili significative.

Per fare ciò si procede selezionando le variabili più significative del modello completo e ripetendo la regressione utilizzando solo queste ultime come esplicative, dove la variabile risposta è sempre l' HDI.
Per tenere in considerazione una maggior varietà di indicatori si è deciso di selezionare tutte quelle variabili che presentassero p-value inferiore a 0.15; eliminando dal modello quindi:

- Employment in industry (% of total empl) (modeled ILO estimate)
- Primary & secondary education, duration (years)
- Profit tax (% of commercial profits)
- Strength of legal rights index (0=weak to 12=strong)
- Total natural resources rents (% of GDP)
- Unemployment, total (% of total labor force) (modeled ILO estimate)
- Urban population (% of total population)

Modello ristretto

Si riportano di seguito i medesimi output visti in precedenza per quanto riguarda il modello completo.

Variabili immesse/rimosse[a]

Modello	Variabili immesse	Variabili rimosse	Metodo
1	Services, value added (% of GDP) Foreign direct investment, net inflows (% of GDP) Population growth (annual %) Current health expenditure per capita current US Lifetime risk of maternal death (%) Exports of goods and services (% of GDP) Employment in services (% of total employment) (modeled ILO estimate) Access to electricity (% of population) Individuals using the Internet (% of population)[b]	.	Inserisci

a. Variabile dipendente: HDInuovo

b. Sono state immesse tutte le variabili richieste.

Riepilogo del modello[b]

Modello	R	R-quadrato	R-quadrato adattato	Errore std. della stima
1	,975[a]	,950	,945	3,56520

a. Predittori: (costante), Services, value added (% of GDP) , Foreign direct investment, net inflows (% of GDP), Population growth (annual %) , Current health expenditure per capita current US, Lifetime risk of maternal death (%) , Exports of goods and services (% of GDP) , Employment in services (% of total employment) (modeled ILO estimate) , Access to electricity (% of population) , Individuals using the Internet (% of population)

b. Variabile dipendente: HDInuovo

ANOVA[a]

Modello		Somma dei quadrati	gl	Media quadratica	F	Sign.
1	Regressione	24042,198	9	2671,355	210,167	,000[b]
	Residuo	1271,064	100	12,711		
	Totale	25313,262	109			

a. Variabile dipendente: HDInuovo

b. Predittori: (costante), Services, value added (% of GDP) , Foreign direct investment, net inflows (% of GDP), Population growth (annual %) , Current health expenditure per capita current US, Lifetime risk of maternal death (%) , Exports of goods and services (% of GDP) , Employment in services (% of total employment) (modeled ILO estimate) , Access to electricity (% of population) , Individuals using the Internet (% of population)

Anche in questo caso si osserva un R-quadrato prossimo al 95% e la statistica-test F è aumentata fino ad oltre 200, rifiutando così chiaramente l'ipotesi nulla di R-quadrato pari a 0.

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sign.
		B	Errore standard	Beta		
1	(Costante)	44,397	3,069		14,469	,000
	Access to electricity (% of population)	,098	,028	,162	3,481	,001
	Current health expenditure per capita current US	,001	,000	,145	4,272	,000
	Employment in services (% of total employment) (modeled ILO estimate)	,077	,038	,090	2,052	,043
	Exports of goods and services (% of GDP)	,036	,016	,069	2,226	,028
	Foreign direct investment, net inflows (% of GDP)	-,127	,043	-,085	-2,966	,004
	Individuals using the Internet (% of population)	,240	,027	,448	9,026	,000
	Lifetime risk of maternal death (%)	-2,164	,565	-,151	-3,831	,000
	Population growth (annual %)	-1,310	,350	-,101	-3,740	,000
	Services, value added (% of GDP)	,077	,047	,054	1,645	,103

a. Variabile dipendente: HDInuovo

Per quanto riguarda i singoli coefficienti si osserva che per praticamente tutte le variabili esplicative si rifiuta l'ipotesi nulla (β=0) ad un livello di significatività del 5%, eccezion fatta per la variabile *Services, value added*.

Verifica del modello migliore

Per verificare se sia preferibile l'utilizzo del modello ristretto rispetto a quello completo si fa ricorso a quello che è definito come il test di Wald[10]. Nello specifico questo test pone a confronto le devianze residue dei due modelli, sotto l'ipotesi nulla che le esplicative non incluse nel modello ristretto non siano utili a spiegare una significativa porzione della variabilità residua, ovvero di quella variabilità della risposta non colta dal modello ristretto.

Per tale motivo si costruisce la seguente statistica test:

$$F = \frac{\frac{[DevRes]_r - [DevRes]_c}{Vc - Vr}}{\frac{[Dev\ res]c}{(n - Vc)}}$$

dove n è la dimensione del campione, Vc è il numero di variabili esplicative nel modello completo e Vr è il numero di parametri nel modello ristretto; che si distribuisce sotto l'ipotesi nulla come una F di Fisher con (Vc−Vr , n – Vc) gradi di libertà.

La statistica ha un valore di

$$F=\frac{\frac{1271,064-1181,915}{16-9}}{\frac{1181,9158}{110-16}}=1,01$$

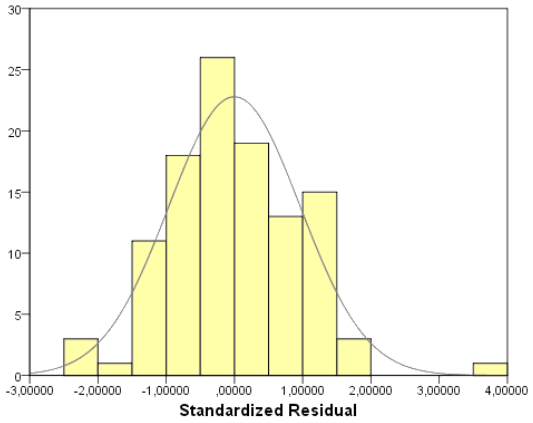
Tale valore risulta minore della soglia critica al livello di significatività α= 0.05, che per una F di fisher con (7, 94) gradi di libertà risulta pari a 2,029. Risulta perciò accettata l'ipotesi nulla del test di Wald e si può quindi concludere a favore del modello ristretto, che sarà oggetto degli approfondimenti riportati di seguito

[10] Wald test: è una prova statistica, usata tipicamente per esaminare se un effetto esiste oppure no, cioè esamina se una variabile indipendente ha un rapporto statisticamente significativo con la variabile dipendente.

Analisi dei residui

Ultimo step prima dell'interpretazione dei coefficienti e delle conclusioni finali è quello di verifica delle ipotesi sui residui, necessarie a garantire la correttezza generale del modello e dei suoi risultati.

►Normalità dei residui:



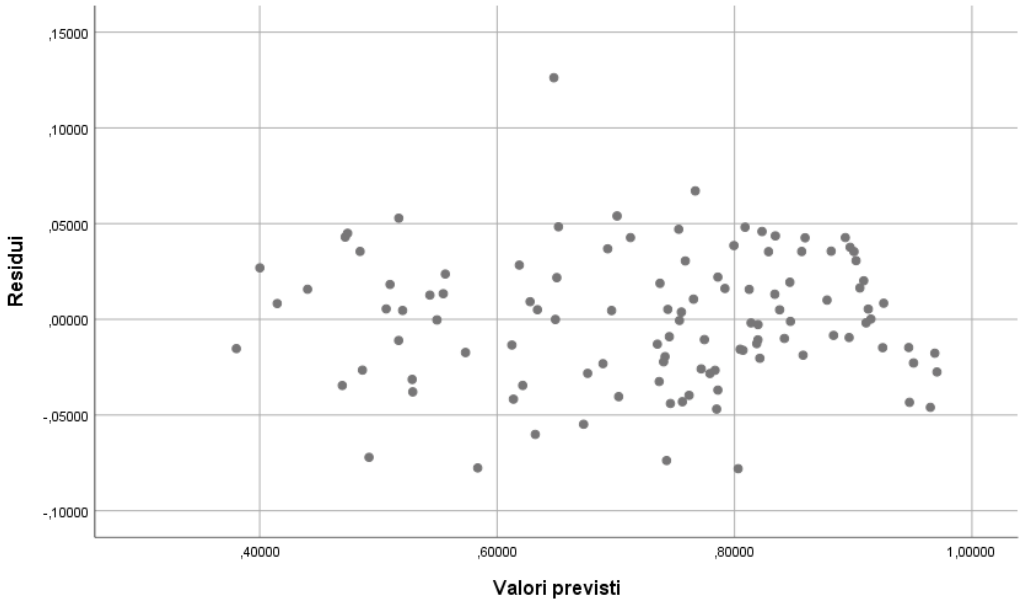
Test di normalità

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistica	gl	Sign.	Statistica	gl	Sign.
Standardized Residual	,041	110	,200*	,985	110	,249

*. Questo è un limite inferiore della significatività effettiva.
a. Correzione di significatività di Lilliefors

Sia del grafico che dall'output emerge che l'ipotesi di normalità dei residui sia rispettata, infatti l'istogramma delle frequenze dei residui standardizzati ben approssima l'andamento di una Normale standard. Per quanto riguarda i test di Kolmogorov-Smirnov e di Shapiro-Wilk, in entrambi i casi viene accettata l'ipotesi nulla di normalità dei residui standardizzati.

►Incorrelazione



Dal grafico emerge con chiarezza la distribuzione completamente casuale dei residui standardizzati, ad indicare l'assenza di legame che non sia stato catturato dal modello di regressione.

Dato che le ipotesi sui residui sembrano verificate, si può ora procedere con l'interpretazione di alcuni dei coefficienti del modello ristretto stimati in precedenza, tenendo sempre presente che l'Indice di sviluppo umano ha ora un range di variazione compreso tra 0 e 100.
Se per esempio si considera la variabile *Current health expenditure per capita current US,* il suo coefficiente pari 0,001 può essere interpretato come: un aumento di 1000 dollari della spesa pro capite annuale in ambito sanitario comporta un incremento pari a 1 dell'HDI.
Equivalentemente, nel caso di *Lifetime risk of maternal death,* avente coefficiente pari a negativo pari a -2,164: una diminuzione dell'1% nel rischio di morte materno comporta un incremento pari a 2,16 dell'HDI.
Si può nello stesso modo procedere per tutte le altre variabili selezionate per il modello ristretto.

A questo punto però, dato che il fine ultimo dell'analisi è quello di fornire indicazioni generali per lo sviluppo di politiche socio-economiche, invece di operare un'analisi puntuale dei coefficienti, si procederà con una interpretazione più generale dell'influenza delle variabili sull'HDI.
In sintesi ciò che emerge è che per accrescere l'indice di sviluppo umano è necessario:

- Garantire il più possibile l'accesso ai servizi ormai considerati di base, dall'elettricità fino ad Internet
- Aumentare i contributi in ambito sanitario, così da cercare di ridurre al minimo i tassi di mortalità
- Incentivare la crescita e l'investimento nel settore terziario, sia in termini di occupati nel settore sia in termini di vendita di tali servizi

6.CONCLUSIONI

Come già esplicitato nell'introduzione di questo elaborato, l'obiettivo di questa analisi era quello di fornire un'interpretazione ad ampio respiro dei fattori che risultano più significativi in relazione all'indice di sviluppo umano ed individuare delle politiche macroeconomiche finalizzate appunto alla riduzione delle diseguaglianze esistenti tra i paesi.

Dalle cluster analysis è emersa la conferma, se mai ve ne fosse bisogno, che i fattori che danno origine alle maggiori disuguaglianze tra i Paesi del mondo sono tanti e vari.

Dalla regressione lineare è però emerso che le variabili che influenzano in maniera più significativa l'indice di sviluppo umano possono essere ricondotte a tre macrocategorie: accesso ai servizi di base, possibilità di buone cure sanitarie e investimenti nel settore terziario.

In conclusione si ritiene che, nell'ottica di ridurre le disuguaglianze, sia compito dei vari Stati e soprattutto delle Comunità/Organizzazioni mondiali quello di mettere in atto macro-politiche volte a garantire e migliorare i tre settori sopra elencati. Tutto questo chiaramente con l'aiuto e la collaborazione di ogni singolo cittadino.

Appendice formule

Indice di correlazione di Pearson:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

Test KMO:

$$\mathrm{KMO} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} r_{ij}^2}{\left(\sum_{i=1}^{k}\sum_{j=1}^{k} r_{ij}^2 + \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}^2\right)}$$

$$\text{Con } a_{ij} = (r_{ij} \bullet 1, 2, 3, \ldots, k)$$

Test di sfericità di Bartlett:

$$\mathrm{T} = \frac{(N-k)\ln(S_p^2) - \sum_{i=1}^{k}(ni-1)\ln(S_i^2)}{1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\left(\frac{1}{ni-1}\right) - \frac{1}{N-k}\right)}$$

Metodo di Ward:

$$d_{(i,j)k} = \frac{1}{ni+nj+nk}\left[(ni+nk)d_{ik}^2 + (nj+nk)\, d_{jk}^2 - n_k d_{ij}^2\right]$$

Distanza euclidea:

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{j=1}^{d}(x_j - y_j)^2}$$

Metodo K-means

$$V(U,C) = \sum_{i=1}^{K}\sum_{X_j \in P_i} ||X_j - C_i||^2$$

Appendice tabelle

APPENDICE N°1- Pianificazione di agglomerazione

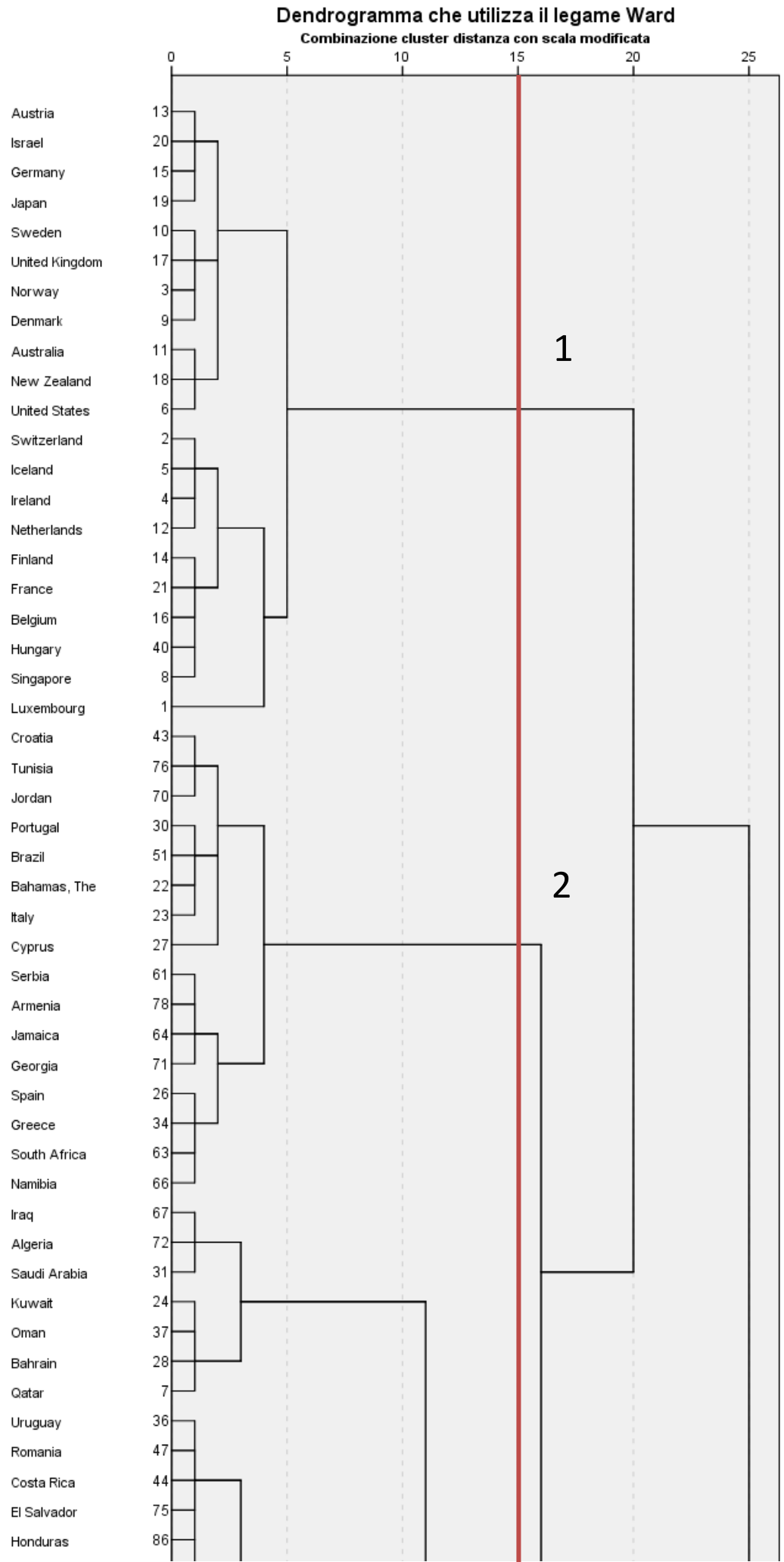
Pianificazione di agglomerazione

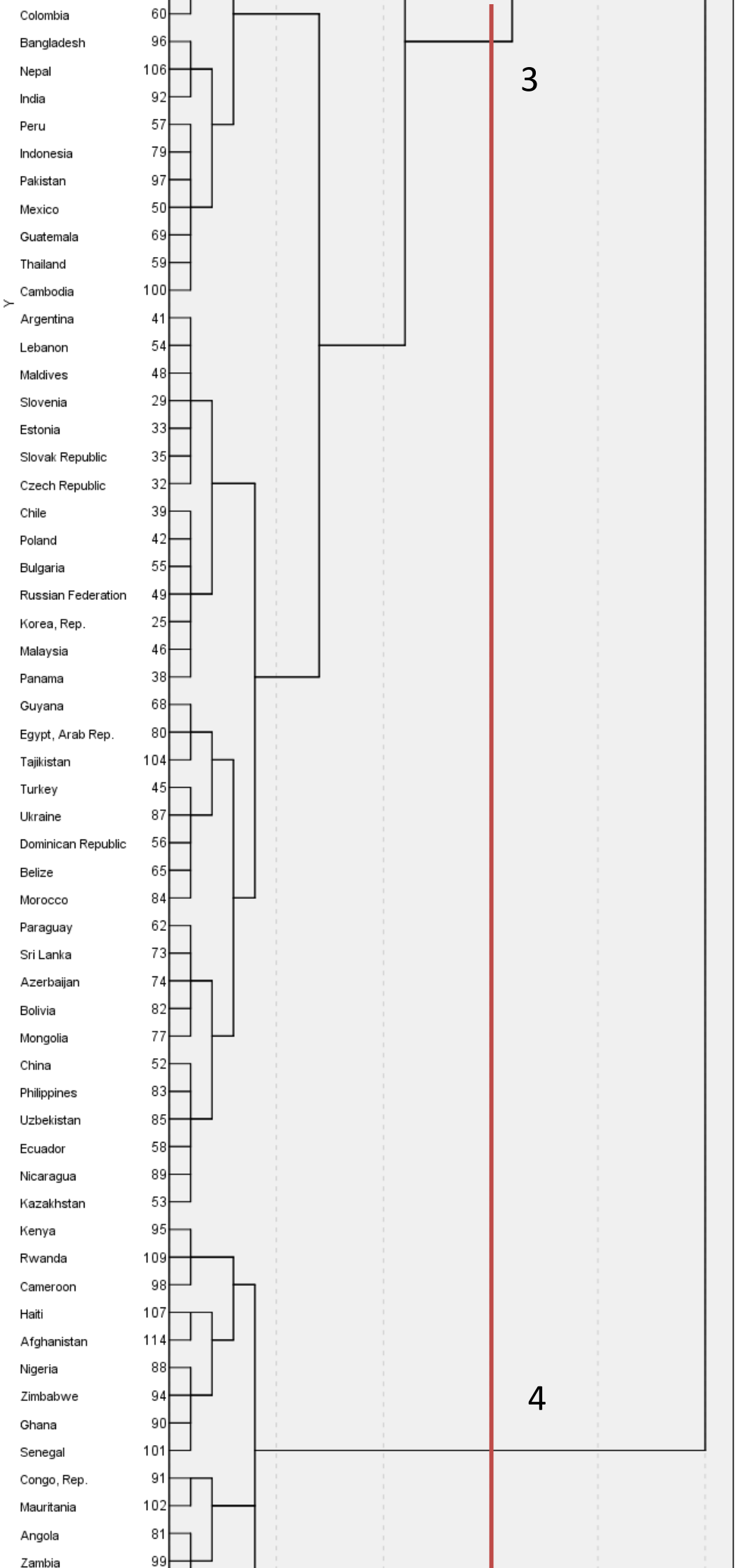
Stadio	Combinato in cluster		Coefficienti	Stadio prima apparizione cluster		Stadio successivo
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	13	20	,108	0	0	80
2	58	89	,227	0	0	44
3	88	94	,384	0	0	84
4	105	116	,550	0	0	32
5	36	47	,718	0	0	24
6	52	83	,894	0	0	36
7	56	65	1,071	0	0	40
8	29	33	1,255	0	0	41
9	10	17	1,444	0	0	69
10	112	113	1,638	0	0	72
11	103	120	1,841	0	0	21
12	75	86	2,051	0	0	42
13	62	73	2,265	0	0	57
14	61	78	2,479	0	0	38
15	11	18	2,699	0	0	83
16	16	40	2,922	0	0	75
17	57	79	3,148	0	0	50
18	39	42	3,383	0	0	31
19	45	87	3,622	0	0	76
20	95	109	3,869	0	0	43
21	103	111	4,117	11	0	47
22	117	118	4,366	0	0	61
23	91	102	4,616	0	0	99
24	36	44	4,868	5	0	85
25	43	76	5,124	0	0	73
26	3	9	5,384	0	0	69
27	41	54	5,648	0	0	63
28	30	51	5,912	0	0	81
29	74	82	6,180	0	0	57
30	25	46	6,452	0	0	53
31	39	55	6,732	18	0	52
32	93	105	7,012	0	4	61
33	50	69	7,306	0	0	55
34	68	80	7,608	0	0	48
35	96	106	7,913	0	0	62
36	52	85	8,228	6	0	74
37	22	23	8,544	0	0	81

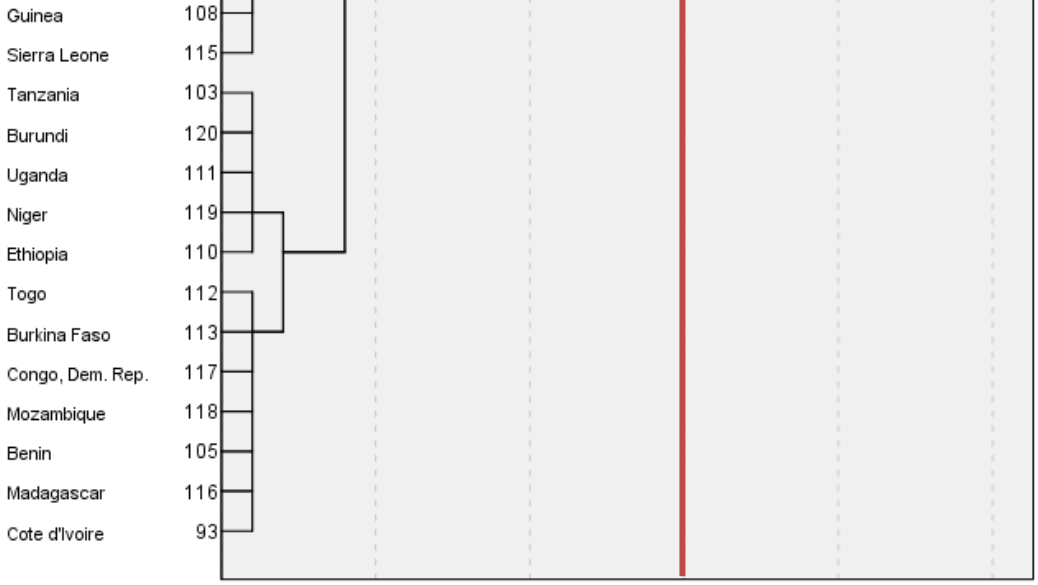
38	61	64	8,867	14	0	64
39	24	37	9,194	0	0	60
40	56	84	9,529	7	0	76
41	29	35	9,868	8	0	65
42	60	75	10,212	0	12	85
43	95	98	10,557	20	0	106
44	53	58	10,902	0	2	74
45	90	101	11,254	0	0	84
46	2	5	11,609	0	0	86
47	103	119	11,972	21	0	66
48	68	104	12,359	34	0	97
49	81	99	12,758	0	0	90
50	57	97	13,162	17	0	82
51	14	21	13,566	0	0	88
52	39	49	13,983	31	0	87
53	25	38	14,408	30	0	87
54	67	72	14,834	0	0	77
55	50	59	15,277	33	0	70
56	4	12	15,722	0	0	86
57	62	74	16,167	13	29	68
58	15	19	16,616	0	0	80
59	26	34	17,073	0	0	89
60	24	28	17,536	39	0	91
61	93	117	18,006	32	22	72
62	92	96	18,481	0	35	93
63	41	48	18,971	27	0	78
64	61	71	19,469	38	0	104
65	29	32	19,978	41	0	78
66	103	110	20,488	47	0	100
67	63	66	20,999	0	0	89
68	62	77	21,512	57	0	98
69	3	10	22,043	26	9	94
70	50	100	22,578	55	0	82
71	107	114	23,118	0	0	95
72	93	112	23,662	61	10	100
73	43	70	24,237	25	0	96
74	52	53	24,814	36	44	98
75	8	16	25,397	0	16	88
76	45	56	25,979	19	40	97
77	31	67	26,577	0	54	108
78	29	41	27,208	65	63	101
79	108	115	27,844	0	0	90
80	13	15	28,501	1	58	102

81	22	30	29,159	37	28	92
82	50	57	29,884	70	50	93
83	6	11	30,616	0	15	94
84	88	90	31,351	3	45	95
85	36	60	32,112	24	42	105
86	2	4	32,902	46	56	103
87	25	39	33,713	53	52	101
88	8	14	34,557	75	51	103
89	26	63	35,448	59	67	104
90	81	108	36,381	49	79	99
91	7	24	37,341	0	60	108
92	22	27	38,380	81	0	96
93	50	92	39,570	82	62	105
94	3	6	40,790	69	83	102
95	88	107	42,016	84	71	106
96	22	43	43,251	92	73	109
97	45	68	44,496	76	48	107
98	52	62	45,818	74	68	107
99	81	91	47,151	90	23	112
100	93	103	48,541	72	66	112
101	25	29	50,210	87	78	111
102	3	13	51,889	94	80	114
103	2	8	53,662	86	88	110
104	26	61	55,514	89	64	109
105	36	50	57,471	85	93	115
106	88	95	59,591	95	43	113
107	45	52	62,085	97	98	111
108	7	31	64,638	91	77	116
109	22	26	67,622	96	104	117
110	1	2	70,730	0	103	114
111	25	45	73,925	101	107	115
112	81	93	77,196	99	100	113
113	81	88	80,877	112	106	119
114	1	3	85,351	110	102	118
115	25	36	91,488	111	105	116
116	7	25	100,992	108	115	117
117	7	22	115,199	116	109	118
118	1	7	132,607	114	117	119
119	1	81	155,258	118	113	0

APPENDICE N°2




Y
Colombia 60
Bangladesh 96
Nepal 106
India 92
Peru 57
Indonesia 79
Pakistan 97
Mexico 50
Guatemala 69
Thailand 59
Cambodia 100
Argentina 41
Lebanon 54
Maldives 48
Slovenia 29
Estonia 33
Slovak Republic 35
Czech Republic 32
Chile 39
Poland 42
Bulgaria 55
Russian Federation 49
Korea, Rep. 25
Malaysia 46
Panama 38
Guyana 68
Egypt, Arab Rep. 80
Tajikistan 104
Turkey 45
Ukraine 87
Dominican Republic 56
Belize 65
Morocco 84
Paraguay 62
Sri Lanka 73
Azerbaijan 74
Bolivia 82
Mongolia 77
China 52
Philippines 83
Uzbekistan 85
Ecuador 58
Nicaragua 89
Kazakhstan 53
Kenya 95
Rwanda 109
Cameroon 98
Haiti 107
Afghanistan 114
Nigeria 88
Zimbabwe 94
Ghana 90
Senegal 101
Congo, Rep. 91
Mauritania 102
Angola 81
Zambia 99
3
4


Guinea 108
Sierra Leone 115
Tanzania 103
Burundi 120
Uganda 111
Niger 119
Ethiopia 110
Togo 112
Burkina Faso 113
Congo, Dem. Rep. 117
Mozambique 118
Benin 105
Madagascar 116
Cote d'Ivoire 93

APPENDICE N°3 – Appartenenza cluster Metodo Gerarchico

Appartenenza cluster

Caso	4 cluster
1:Luxembourg	1
2:Switzerland	1
3:Norway	1
4:Ireland	1
5:Iceland	1
6:United States	1
7:Qatar	2
8:Singapore	1
9:Denmark	1
10:Sweden	1
11:Australia	1
12:Netherlands	1
13:Austria	1
14:Finland	1
15:Germany	1
16:Belgium	1
17:United Kingdom	1
18:New Zealand	1
19:Japan	1
20:Israel	1
21:France	1
22:Bahamas, The	3
23:Italy	3
24:Kuwait	2
25:Korea, Rep.	2
26:Spain	3
27:Cyprus	3
28:Bahrain	2
29:Slovenia	2
30:Portugal	3
31:Saudi Arabia	2
32:Czech Republic	2
33:Estonia	2
34:Greece	3
35:Slovak Republic	2
36:Uruguay	2
37:Oman	2
38:Panama	2
39:Chile	2

40:Hungary	1
41:Argentina	2
42:Poland	2
43:Croatia	3
44:Costa Rica	2
45:Turkey	2
46:Malaysia	2
47:Romania	2
48:Maldives	2
49:Russian Federation	2
50:Mexico	2
51:Brazil	3
52:China	2
53:Kazakhstan	2
54:Lebanon	2
55:Bulgaria	2
56:Dominican Republic	2
57:Peru	2
58:Ecuador	2
59:Thailand	2
60:Colombia	2
61:Serbia	3
62:Paraguay	2
63:South Africa	3
64:Jamaica	3
65:Belize	2
66:Namibia	3
67:Iraq	2
68:Guyana	2
69:Guatemala	2
70:Jordan	3
71:Georgia	3
72:Algeria	2
73:Sri Lanka	2
74:Azerbaijan	2
75:El Salvador	2
76:Tunisia	3
77:Mongolia	2
78:Armenia	3
79:Indonesia	2
80:Egypt, Arab Rep.	2
81:Angola	4
82:Bolivia	2

83:Philippines	2
84:Morocco	2
85:Uzbekistan	2
86:Honduras	2
87:Ukraine	2
88:Nigeria	4
89:Nicaragua	2
90:Ghana	4
91:Congo, Rep.	4
92:India	2
93:Cote d'Ivoire	4
94:Zimbabwe	4
95:Kenya	4
96:Bangladesh	2
97:Pakistan	2
98:Cameroon	4
99:Zambia	4
100:Cambodia	2
101:Senegal	4
102:Mauritania	4
103:Tanzania	4
104:Tajikistan	2
105:Benin	4
106:Nepal	2
107:Haiti	4
108:Guinea	4
109:Rwanda	4
110:Ethiopia	4
111:Uganda	4
112:Togo	4
113:Burkina Faso	4
114:Afghanistan	4
115:Sierra Leone	4
116:Madagascar	4
117:Congo, Dem. Rep.	4
118:Mozambique	4
119:Niger	4
120:Burundi	4

APPENDICE N°4 – Appartenenza cluster Metodo Non Gerarchico

Appartenenza cluster

Numero di caso	Country Name	Cluster	Distanza
1	Luxembourg	1	4,407
2	Switzerland	1	,911
3	Norway	1	1,007
4	Ireland	1	1,534
5	Iceland	1	,892
6	United States	1	2,220
7	Qatar	2	3,709
8	Singapore	1	1,676
9	Denmark	1	,678
10	Sweden	1	,299
11	Australia	1	1,168
12	Netherlands	1	,693
13	Austria	1	1,168
14	Finland	1	1,058
15	Germany	1	,979
16	Belgium	1	1,010
17	United Kingdom	1	,512
18	New Zealand	1	1,518
19	Japan	1	1,769
20	Israel	1	1,104
21	France	1	1,373
22	Bahamas, The	4	1,033
23	Italy	4	,822
24	Kuwait	2	3,113
25	Korea, Rep.	2	1,045
26	Spain	4	,984
27	Cyprus	4	1,623
28	Bahrain	2	2,765
29	Slovenia	2	1,013
30	Portugal	4	1,049
31	Saudi Arabia	2	2,094
32	Czech Republic	2	1,132
33	Estonia	2	,892
34	Greece	4	1,509
35	Slovak Republic	2	1,270
36	Uruguay	2	1,189
37	Oman	2	2,702
38	Panama	2	,934
39	Chile	2	,462
40	Hungary	1	1,119

41	Argentina	2	,931
42	Poland	2	,729
43	Croatia	4	,881
44	Costa Rica	2	1,518
45	Turkey	2	,683
46	Malaysia	2	,655
47	Romania	2	1,043
48	Maldives	2	1,356
49	Russian Federation	2	,710
50	Mexico	2	1,447
51	Brazil	4	1,070
52	China	2	,652
53	Kazakhstan	2	,534
54	Lebanon	2	,880
55	Bulgaria	2	,776
56	Dominican Republic	2	,640
57	Peru	2	1,010
58	Ecuador	2	,314
59	Thailand	2	1,085
60	Colombia	2	1,599
61	Serbia	4	,495
62	Paraguay	2	,413
63	South Africa	4	1,694
64	Jamaica	4	,998
65	Belize	2	,841
66	Namibia	4	1,672
67	Iraq	2	2,162
68	Guyana	2	1,281
69	Guatemala	2	1,350
70	Jordan	4	1,606
71	Georgia	4	1,059
72	Algeria	2	1,485
73	Sri Lanka	2	,601
74	Azerbaijan	2	,807
75	El Salvador	2	1,230
76	Tunisia	4	,861
77	Mongolia	2	1,217
78	Armenia	4	,674
79	Indonesia	2	1,043
80	Egypt, Arab Rep.	4	1,322

81	Angola	3	,837
82	Bolivia	2	1,090
83	Philippines	2	,986
84	Morocco	2	,940
85	Uzbekistan	2	,850
86	Honduras	2	1,426
87	Ukraine	2	1,108
88	Nigeria	3	,926
89	Nicaragua	2	,500
90	Ghana	3	,999
91	Congo, Rep.	3	2,056
92	India	2	1,697
93	Cote d'Ivoire	3	,584
94	Zimbabwe	3	,826
95	Kenya	3	1,768
96	Bangladesh	2	1,853
97	Pakistan	2	1,093
98	Cameroon	3	1,449
99	Zambia	3	1,420
100	Cambodia	2	1,488
101	Senegal	3	,709
102	Mauritania	3	2,267
103	Tanzania	3	,717
104	Tajikistan	2	1,419
105	Benin	3	,378
106	Nepal	2	2,006
107	Haiti	4	1,831
108	Guinea	3	1,541
109	Rwanda	3	1,856
110	Ethiopia	3	,882
111	Uganda	3	,742
112	Togo	3	,811
113	Burkina Faso	3	,835
114	Afghanistan	3	1,299
115	Sierra Leone	3	,961
116	Madagascar	3	,438
117	Congo, Dem. Rep.	3	,678
118	Mozambique	3	,404
119	Niger	3	1,197
120	Burundi	3	,963

Link

Dataset

https://drive.google.com/file/d/1d3wJ95s_F0atHAo0Tj0kk31n0h455C3J/view?usp=sharing

Matrice di correlazione

https://drive.google.com/file/d/1mH7CDyVgb4C1DvsP_jlRHrhx2euf6-GR/view?usp=sharing

Dendogramma

https://drive.google.com/file/d/1it2lrSqdwKBAJuKgAlp1gzCSDkphnbRC/view?usp=sharing

Sitografia

Teoria

https://elearning.unimib.it/course/view.php?id=23581

http://host.uniroma3.it/facolta/economia/db/materiali/insegnamenti/586_5037.pdf

https://people.unica.it/lucafrigau/files/2012/04/Cap-V-_ANALISI-DEI-GRUPPI.pdf

https://www.germanorossi.it/mi/file/psico/Psic06-RegMult.pdf

Dataset iniziale

https://data.world/

SPSS

https://www.spss-tutorials.com/basics/