Università Politecnica delle Marche

Dipartimento di Ingegneria dell'Informazione

Facoltà di Ingegneria Informatica e dell'Automazione

RELAZIONE BIG DATA ANALYTICS E MACHINE LEARNING

Docenti:
Prof. Domenico Potena
Prof. Claudia Diamantini

Studenti:
Silvia Ciuffreda
Luca Liberatore

ANNO ACCADEMICO 2021/2022

Indice

Elenco delle figure

1 Introduzione

Lo scopo del progetto mira a comprendere i meccanismi che sono alla base della realizzazione degli SDG, analizzando le relazione tra i diversi indicatori, al fine di individuare eventuali correlazioni e tendenze comuni. In particolare, ci si concentrerà sull'identificazione di possibili sinergie o conflitti tra gli indicatori che saranno selezionati, con l'obiettivo di individuare trend e regolarità e correlare tra loro gli stessi indicatori. Infine, dopo aver ottenuto un'immagine dei Paesi sulle diverse prospettive, ci si focalizzerà l'attenzione sul loro confronto per poi illustrare a che punto siamo per il raggiungimento degli obiettivi prefissati nell'Agenda 2030.

L'analisi dei dati sarà condotta utilizzando strumenti software per l'integrazione dei dati, quale *Pentaho Data Integration*, e di elaborazione e visualizzazione dei dati tramite il linguaggio *Python*, al fine di provare ad integrare i diversi dati provenienti da Indicatori differenti e individuare eventuali pattern.

L'obiettivo, dunque, sarà quello di comprendere meglio i meccanismi che stanno alla base della realizzazione degli SDG e individuare eventuali punti di forza o debolezza nella loro implementazione.

Parte I
Parte Prima

2 Agenda 2030

L'Agenda 2030 per lo Sviluppo Sostenibile è un programma d'azione per le persone, il pianeta e la prosperità' che nasce nel settembre del 2015. L'idea venne sottoscritta da 193 Nazioni appartenenti all'ONU e si fonda su 17 Goal per lo Sviluppo Sostenibile, articolati in un programma di 169 Target. Questi obiettivi hanno impegnato, ed impegneranno, le Nazioni coinvolte per i prossimi 15 anni, da qui il nome "Agenda 2030" per l'appunto. I 17 Obiettivi possono essere generalmente suddivisi in tre macro-aree: economica, sociale ed ecologica. Scendendo nel dettaglio, i Goal si articolano in varie sezioni quali: lotta alla povertà, contrasto alla fame e al cambiamento climatico, miglioramento dell'istruzione, riduzione delle disuguaglianze, ecc. come mostrati nella Fig.1

Figura 1: Gli obiettivi per lo sviluppo sostenibile

2.1 GOAL 1 - POVERTÀ

Obiettivo (Goal 1): Porre fine alla povertà in tutte le sue forme. Il primo obiettivo dell'Agenda 2030 riguarda la sconfitta della povertà nel mondo. L'analisi parte dalla considerazione schematica di alcuni dati:

- 836 milioni di persone vivono ancora in povertà estrema
- Circa una persona su cinque nelle regioni in via sviluppo vive con meno di 1,25 dollari al giorno
- La stragrande maggioranza delle persone che vivono con meno di 1,25 dollari al giorno appartiene a due regioni: Asia meridionale e Africa subsahariana

Il *Goal 1* al suo interno si articola in vari "sotto-obiettivi", detti **Target**, che sono: *1.1, 1.2, 1.3, 1.4, 1.5, 1.a e 1.b.* Il target considerato nella relazione è l'*1.1*, il primo, che recita come riportato: "*Entro il 2030, sradicare la povertà estrema per tutte le persone in tutto il mondo, attualmente misurata sulla base di coloro che vivono con meno di 1,25$ al giorno*". Ad un livello di dettaglio maggiore vi sono gli Indicatori, nel nostro caso si prende in considerazione l'Indicatore 1.1.1, che tra l'altro è l'unico

indicatore presente nel target 1.1, il quale rappresenta la percentuale di popolazione che vive sotto la soglia internazionale di povertà, suddivisa per sesso, età, stato lavorativo e collocazione geografica. In particolare, l'Indicatore 1.1.1, a sua volta, si articola in due sotto-categorie: 1.1.1a, in cui il campione di popolazione interessato è la totalità dei cittadini, con riferimento alle varie aree geografiche; 1.1.1b, invece, prende in considerazione esclusivamente il sottoinsieme di popolazione occupata e, quindi, in età lavorativa.

METADATA[5] - 01-01-01a (popolazione totale povera)

Si procede ad analizzare i metadati dell'indicatore, con l'obiettivo di interpretare ulteriormente la loro composizione. Innanzitutto, si riporta, in lingua originale, la definizione esatta dell'indicatore stesso: "The indicator "proportion of the population below the international poverty line" is defined as the percentage of the population living on less than $ 2.15 a day at 2017 international prices". Si è dunque definita la soglia di povertà come una linea immaginaria posta a 2,15 dollari al giorno, sotto la quale si è considerati in condizioni di estrema povertà. Questi dati, che concernono i vari continenti e nazioni del mondo, sono continuamente collezionati dalla Global Working Group della Banca Mondiale e vengono periodicamente aggiornati ogni anno ad aprile, anche se non per tutti i Paesi.

Solitamente, la Banca Mondiale riceve i dati dal National Statistical Offices, oppure dall'Eurostat o dal LIS (Luxemburg Income Study). Nonostante i progressi sulla capacità di misurare la povertà, rimane comunque una sfida di alto valore, data la vasta quantità di informazione in gioco, che si affronta principalmente tramite sondaggi domestici, i quali tuttavia perdono di qualità e di consistenza quando questi vengono svolti in quei Paesi di dimensioni ridotte e relativamente stabili politicamente.

Inoltre, un ulteriore punto critico dei dati si può desumere nel momento in cui si effettuano paragoni tra valori di povertà che fanno riferimento a nazioni con differenti livelli di sviluppo. Infatti, effettuare tale confronto potrebbe non essere propriamente corretto a causa delle differenze temporali nei sondaggi, della differente preparazione degli enumeratori e delle intrinseche differenze socioeconomiche proprie di ciascun Paese.

Il livello di povertà nazionale fa riferimento alla "concezione" di povertà di quello Stato in riferimento alla valuta locale, che è differente in termini reali tra un Paese ed un altro.

In definita, nei metadati si giunge alla conclusione che i livelli di povertà proprie di uno Stato non possono essere comparati tra diverse nazioni. In aggiunta, i sondaggi non vengono effettuati per tutte le nazioni ogni anno, quindi si devono effettuare delle interpolazioni per ricavare stime della povertà utilizzando i dati di contabilità internazionale a disposizione.

Invece, ad un livello di aggregazione superiore, come quello mondiale e continentale, i dati vengono fuori come medie ponderate della popolazione.

Prima di passare all'Obiettivo successivo, cioè quello che fa riferimento alle malattie (*Goal 3*), bisogna fare una precisazione: tutti gli indicatori dell'Agenda 2030 sono classificati in tre *Tier* che si distinguono per la metodologia di sviluppo e la disponibilità dei dati al livello *globale*[3]. Sia l'Indicatore 1.1.1 che gli Indicatori 3.3.1-5, appartenenti al Target 3.3 e al corrispettivo Obiettivo, rientrano nella tipologia di *Tier*1[4], come riportato nella Fig. 2.

Figura 2: Tier classification

2.2 GOAL 3 - SALUTE E BENESSERE

Obbietivo (*Goal 3*): Assicurare la salute e il benessere per tutti e per tutte le età Una volta presentato il primo indicatore sulla povertà, ora ci si vuole soffermare sull'altro protagonista del lavoro di integrazione tra i vari indicatori messi a disposizione dalle Nazioni Unite: l'Obiettivo 3. Quest'ultimo fa riferimento alla salute e al benessere in generale rispetto la popolazione mondiale. Prima di addentrarsi nei rispettivi target ed indicatori, si vogliono esporre alcuni *dati*[8] in modo da avare ben chiaro l'argomento trattato:

- L'HIV è la causa principale di morte tra le donne in età riproduttiva in tutto il mondo;
- L'AIDS è oggi la principale causa di morte tra gli adolescenti (dai 10 ai 19 anni) in Africa e la seconda causa più comune di morte tra gli adolescenti a livello mondiale;
- Tra il 2000 e il 2015, sono state evitate più di 6,2 milioni di morti per malaria, principalmente in bambini con età inferiore ai 5 anni in Africa subsahariana. Il tasso globale di incidenza della malaria si è ridotto del 37% e il tasso di mortalità del 58%;
- Tra il 2000 e il 2013 gli interventi di prevenzione, di diagnosi e di trattamento della tubercolosi hanno salvato 37 milioni di vite. Il tasso di mortalità da tubercolosi si è ridotto del 45% e il tasso di prevalenza del 41% tra il 1990 e il 2013.

Come nel caso precedente il *Goal* si suddivide in Target, che vanno dal *3.1* al *3.9* e a cui si aggiungono il *3.a*, *3.b*, *3.c* e il *3.d*. Seguendo le orme del Goal precedente sulla povertà, anche in questo caso si sceglie un Target specifico, il 3.3: "*Entro il 2030, porre fine alle epidemie di AIDS, tubercolosi, malaria e malattie tropicali trascurate; combattere l'epatite, le malattie di origine idrica e le altre malattie trasmissibili*". Successivamente vengono presi in esame tre Indicatori: 3.3.1 che fa riferimento alla diffusione dell'HIV; 3.3.2 che concerne la tubercolosi; 3.3.3 che riguarda la malattia della malaria. Una breve descrizione dei metadati permette di comprendere la consistenza e la fruibilità dei dati raccolti sulle tre malattie.

METADATA[9] - 3.3.1 (HIV)

Per comodità, analogamente al caso precedente, si riporta la definizione originale dell'Indicatore: "The number of new HIV infections per 1,000 uninfected population, by sex, age and key populations as defined as the number of new HIV infections per 1,000 persons among the uninfected population." In altri termini, il valore assoluto che riporta tale indicatore, particolarizzato per sesso ed età, con riferimento ad una determinata area geografica, rappresenta il numero di nuove infezioni su un campione di popolazione di mille individui sani.

I valori di questo dataset sono ottenuti tramite un modello, proposto da un gruppo di epidemiologi, statistici ed esperti di salute pubblica in stretto contatto con l'UNAIDS, che stima l'andamento delle infezioni *HIV*[10] per quasi ogni Paese nel mondo.

Nella pratica, ogni Paese, tramite il proprio gruppo di esperti nel settore, si avvale di tale modello per sviluppare delle stime annue. I dati forniti al modello vengono fuori da sondaggi e da programmi di raccolta di informazioni. I risultati ottenuti vengono pubblicati ogni anno nel mese di luglio.

La validità delle stime sull'andamento delle infezioni da HIV nelle varie nazioni dipende da Paese a Paese: negli Stati ad alta concentrazione di HIV si ha un forte sorveglianza accompagnata da rilevamenti sul territorio; diversamente nelle aree geografiche meno interessate dall'infezione, in cui le stime potrebbero risultare meno accurate, si ha meno controllo.

Tuttavia, anche se le zone a maggior incidenza sono fortemente sorvegliate, effettuare confronti in diversi lassi temporali può portare a delle inesattezze di valutazione, dovute alla differente metodologia di indagine oppure ad i metodi di campionamento della popolazione stessa. Quindi, la scelta del campione di popolazione chiave può indurre il modello a produrre dei risultati sovra o sottostimati sull'incidenza dell'HIV nelle diverse areee geografiche considerate. Inoltre, l'UNAIDS attualmente si occupa solamente di nazioni con almeno 250,000 persone e, in aggiunta, esclude dal conteggio ulteriori 8 Paesi, i quali hanno un'incidenza di HIV talmente bassa da non essere presi neppure in considerazione.

METADATA[12] - 3.3.2 (TUBERCOLOSI)

Come al solito si riporta la definizione originale: "The tuberculosis incidence per 100 000 population is defined as the estimated number of new and relapse TB cases (all forms of TB, including cases in people living with HIV) arising in a given year, expressed as a rate per 100 000 population".

Diversamente dal caso precedente, si considera un campione di cento mila persone su cui si stima il numero di nuovi casi (o ricadute) in un dato anno.

I dati, al livello di Stato, vengono raccolti dalle singole nazioni e comunicati all'OMS ogni anno tra marzo e giugno, tramite un sistema di segnalazione standardizzato e gestito dall'OMS stessa. Da questi dati si ricavano delle stime per ogni singolo Paese, le quali vengono rilasciate in estate e successivamente vengono revisionate prima della pubblicazione, che solitamente ricade nel mese di ottobre.

Questo tipo di indicatore è stato utilizzato, nello svolgere analisi di tipo descrittivo, per più di un secolo come indicatore principale per valutare il carico di tubercolosi sulle popolazioni, affiancato dal calcolo della mortalità associata a questa malattia. Fortunatamente, in questo caso, la qualità dell'indicatore permette di effettuare confronti sia nel tempo all'interno della stessa nazione sia tra i diversi Paesi.

Essendo una stima dei nuovi casi, in termini numerici, sono forniti dei limiti di incertezza (bound) entro i quali la stima rimane veritiera all'interno di un range specificato. Le stime vengono fuori prendendo in esame diversi fattori: notifiche annuali di casi, indagini nazionali e segnalazioni di cause di morte per tubercolosi.

Il risulato finale è frutto di un processo consultivo ed analitico guidato dall'OMS.

METADATA[13] - 3.3.3 (MALARIA)

Definizione: "Incidence of malaria is defined as the number of new cases of malaria per 1,000 people at risk each year." Quindi, su mille persone a rischio ogni nuovo caso rappresenta l'incidenza della malaria, particolarizzata per un orizzonte temporale in anni e con una precisa collocazione geografica. Un caso di malaria, per definizione, è considerato tale quando nel paziente in questione si rilevano tracce di parassiti della malaria nel sangue tramite un test diagnostico.

Il conteggio dei casi segnalati è ottenuto tramite un sistema di sorveglianza proprio di ciascuna nazione, in cui viene incluso: il numero di casi testati, il numero di casi sospetti e il numero di casi positivi per metodo di rilevamento e per struttura sanitaria interessata. I dati sono collezionati e rilasciati ogni anno dal "Programma Nazionale per il Controllo della Malaria", in stretta collaborazione con il Ministero della Salute. L'unità di sorveglianza e di monitoraggio del "Programma Mondiale di Controllo della Malaria" è responsabile di raccogliere ed elaborare tutte le informazioni, facendo delle stime nazionali per alcuni Paesi in collaborazione anche con il "Progetto Atlante Malaria" designato dall'OMS.

Può accadere che l'incidenza stimata sia inferiore da quella riportata dal Ministero della Salute per svariati motivi, tra i quali si elencano: non tutte le strutture sanitarie hanno riferito i propri casi rilevati e quindi il numero stimato è maggiore dei casi segnalati; strutture sanitarie private che non partecipano alla segnalazione; l'indicatore viene stimato solo dove si verifica la trasmissione della malaria.

3 PENTAHO DATA INTEGRATION

In questo capitolo verrà illustrato ogni Indicatore al fine di rendere esplicita la loro composizione tabellare. Terminata la fase di presentazione dei dati, si effettuerà l'integrazione dei vari dataset a disposizione in modo da avere un quadro generale della situazione mondiale sulle condizioni di povertà e sull'andamento delle malattie considerate.

3.1 INDICATORE 1.1.1 - POVERTA'

Come software per svolgere integrazione dei dati si è scelto di utilizzare Pentaho Data Integration. I file dei dataset vengono presi dal sito ufficiale delle Nazioni Unite e caricati all'interno del software di casa Hitachi.

Il file Excel, che riguarda l'Indicatore 1.1.1, viene caricato tramite un operatore di input all'interno del software Pentaho Data Integration al fine di svolgere la fase più importante dell'analisi, ovvero l'ETL: Estrazione, Trasformazione e Caricamento dei dati. I campi del data-set sono i seguenti:

- *Goal*: 1
- *Target*: 1.1
- *Indicatore*: 1.1.1
- *Series Code*: codice identificativo per distinguere tra tutta la popolazione ("SI_POV_DAY1" -> Metadata 01-01-01a) e solamente la popolazione con un impiego lavorativo ("SI_POV_EMP1" -> Metadata 01-01-01b)
- *Series Description*: è la descrizione del precedente codice ed assume il seguente valore: "Proportion of population below international poverty line (%)" per la popolazione in generale; mentre, per il sottoinsieme di popolazione con lavoro, l'attributo riporta "Employed population below international poverty line, by sex and age (%)"

Al fine di ottenere un parallelismo rispetto alla diffusione delle malattie (Goal 3) si è considerato tutto il campione di popolazione disponibile e non solamente quello in età lavorativa (+15Y).

Quest'ultimo, a sua volta, anche se costituisce la maggior parte del dataset (essendo maggiormente particolarizzato per fasce d'età), rappresenta ad ogni modo un sottoinsieme dell'intera popolazione. Quanto detto appare evidente nella Fig. 3, in cui, tramite l'ecosistema Python, si è costruito un grafo a torta che riporta la suddivisione all'interno del dataset: i lavoratori costituiscono oltre il 91% dell'intero dataset, con un numero di righe pari a 32.396; mentre l'intera popolazione, che è la parte più piccola del dataset (8,42 %), non raggiunge le 3.000 righe.

- *Geo Area Code*[6]: indica un valore numerico associato univocamente ad un'area geografica specifica
- *Geo Area Name*: rappresenta il nome delle diverse aree geografiche
- *Time Period*: è l'anno di riferimento (assume gli stessi valori di Time Detail)
- *Value*: percentuale di popolazione che vive con meno di $2,15 al giorno ai prezzi internazionali del 2017
- *Base Period*: anno di riferimento per le conversioni monetarie
- *Source*: fonte da cui provengono i dati
- *Foot Note*: "The current extreme poverty line is set at $2,15 a day in 2017 PPP terms, Accessed September 14, 2022"
- *Age*: può assumere i seguenti valori: <15, 15-64, 65+ e tutte le età ("*ALLAGE*")
- *Location*: "ALLAREA", "RURAL" e "URBAN"
- *Nature e Reporting*: hanno sempre come valore la lettera "G" che è un codice per indicare che i dati sono di monitoraggio globale

```
In [4]: #conto quante sono le righe di tutta la popolazione e dei soli lavoratori
        data_P["SeriesCode"].value_counts()

Out[4]: SI_POV_EMP1    32396
        SI_POV_DAY1     2979
        Name: SeriesCode, dtype: int64
```

```
In [5]: #faccio una rappresentazione tramite grafo a tora
        labels = ['Lavoratori', 'Intera popolazione']
        size = data_P["SeriesCode"].value_counts()
        colors_pie = ["coral", "teal"]
        plt.figure(figsize=(8,5))
        plt.pie(size, labels=labels, autopct='%.2f%%')
        plt.axis('off')
        plt.legend()
        plt.show()
```

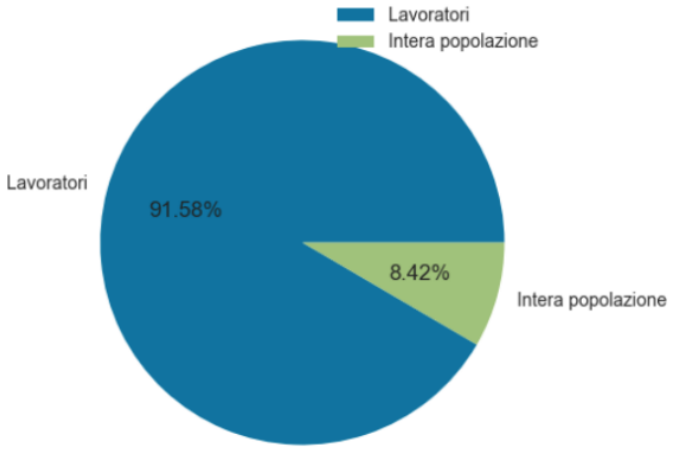

Figura 3: Il campo *Series Code* Goal 1 - Sconfiggere la povertà

- *Sex*: "BOTHSEX", "MALE" e "FEMALE"
- *Units*: "PERCENT"

Riassumendo, si sono escluse le righe inerenti i lavoratori e sono state filtrate unicamente quelle che riguardano l'intera popolazione. Questo approccio risulta ancora più convincente se si vanno a considerare i dataset relativi alle malattie (Target 3.3) in cui la distinzione per fasce d'età, nella maggior parte degli Indicatori considerati, non viene nemmeno effettuata.

Un approccio simile avviene anche nella parte dei dati dell'Indicatore 1.1.1 relativi al totale della popolazione; infatti, la maggioranza dei valori numerici percentuali fa riferimento a tutte le età ("ALLAGE"), invece che suddividersi equamente per fasce d'età (Fig.4).

```
#conto come si divide l'età
data_P["Age"].value_counts()

ALLAGE    2639
<15Y       115
15-64      113
65+        112
Name: Age, dtype: int64
```

Figura 4: Il campo *Age*

Un ragionamento analogo può essere svolto per il genere. Infatti, nei dataset delle Malattie non appare sempre una distinzione dei sessi; mentre nella tabella della Povertà la maggior parte dei valori considera entrambi i sessi insieme, senza effettuare una suddivisione "MALE/FEMALE" (Fig. 5).

Dunque, analogamente a prima, per una migliore integrazione dei dati si è deciso ulteriormente di filtrare, prima per età (ALLAGE) poi per genere (BOTHSEX).

```
#conto come si divide il sesso
data_P["Sex"].value_counts()
```

```
BOTHSEX    2757
MALE        112
FEMALE      110
Name: Sex, dtype: int64
```

Figura 5: Il campo *Sex*

Queste operazioni di caricamento dei dati e di trasformazione, propri della fase di ETL all'interno dell'ecosistema dei Big Data, sono state effettuate all'interno del software Pentaho Data Integration con l'ausilio dei vari operatori di input ("Input Excel") e di filtraggio delle righe ("filter rows").

Prima di effettuare queste operazioni però, si è fatto un conteggio dei valori nulli (Fig. 6) in modo da visualizzare la composizione del dataset.

Dall'analisi, tramite la libreria Pandas di Python, risulta chiaro come le colonne di interesse non presentino valori nulli, ma vi sono semplicemente delle colonne poco significative che sono totalmente vuote.

All'interno di Pentaho, dunque, per quanto riguarda le *features* completamente nulle e quelle prive di informazione significativa si è usato l'operatore "Select values" per rimuoverle. Questo operatore permette anche di rinominare gli attributi delle colonne di interesse e persino di assegnare il giusto tipo (int, float...) al dato in esame.

Ultima considerazione da fare, onde evitare righe duplicate e di difficile interpretazione, è necessario filtrare l'attributo Location per "ALLAREA", evitando di particolarizzarlo per distretti urbani e zone rurali, perché questa distinzione non è presente nei dataset del Target 3.3 con cui si vuole fare integrazione.

```
#faccio il conteggio dei valori nulli
data_P.isnull().sum()
```

```
Goal                    0
Target                  0
Indicator               0
SeriesCode              0
SeriesDescription       0
GeoAreaCode             0
GeoAreaName             0
TimePeriod              0
Value                   0
Time_Detail             0
TimeCoverage         2979
UpperBound           2979
LowerBound           2979
BasePeriod              0
Source                  0
GeoInfoUrl           2979
FootNote                0
Age                     0
Location                0
Nature                  0
Observation Status   2979
Reporting Type          0
Sex                     0
Units                   0
dtype: int64
```

Figura 6: Verifica valori nulli

Uno screen esplicativo sul tipo di filtraggio effettuato all'intero del dataset viene mostrato di seguito (Fig. 7) prendendo in considerazione la *Somalia*[7]: si prende "ALLAGE" per l'età, si filtra per entrambi i sessi ("BOTHSEX") e si selezionano tutte le aree ("ALLAREA"). Quindi, il valore ottenuto è 71.

GeoAreaCode	GeoAreaName	Time Period	Value	Age	Location	Sex
706	Somalia	2017	65	ALLAGE	URBAN	BOTHSEX
706	Somalia	2017	71	ALLAGE	ALLAREA	BOTHSEX
706	Somalia	2017	75	<15Y	ALLAREA	BOTHSEX
706	Somalia	2017	67	15-64	ALLAREA	BOTHSEX
706	Somalia	2017	66	65+	ALLAREA	BOTHSEX
706	Somalia	2017	70	ALLAGE	ALLAREA	FEMALE
706	Somalia	2017	71	ALLAGE	ALLAREA	MALE
706	Somalia	2017	74	ALLAGE	RURAL	BOTHSEX

Figura 7: Esempio filtraggio applicato sull'intero dataset

La sequenza di blocchi che si ottiene, all'interno di Pentaho, è riportata in Fig. 8.

Figura 8: Pentaho - Indicatore Povertà

3.2 INDICATORE 3.3.1 - HIV

Si procede come nel caso precedente caricando il data-set all'interno di Pentaho. Gli attributi che vengono fuori sono i seguenti:

- *Goal*: 3
- *Target*: 3.3
- *Indicatore*: 3.3.1
- *Series Code*: SH_HIV_INCD
- *Series Description*: Number of new HIV infections per 1,000 uninfected population, by sex and age (per 1,000 uninfected population)
- *Geo Area Code*
- *Geo Area Name*
- *Time Period*: è l'anno di riferimento (assume gli stessi valori di Time Detail)
- *Value*: numero (float con due cifre dopo la virgola) di nuove infezioni su un campione di mille persone non infetto
- *Source*: UNAIDS
- *Age*: si specializza nelle seguenti fasce si età: <15, 15-24, 15-49, 50+ e tutte le età
- *Nature*: "E" che sta per Estimated data
- *Reporting type*: "G" per Global
- *Sex*: "BOTHSEX", "MALE" e "FEMALE"
- *Units*: PER_1000_UNINFECTED_POP

Anche in questo caso, per semplicità e per favorire l'integrazione dei dati, si è optato per selezionare tutte le fasce di età, ovvero quando l'attributo "AGE" presenta il valore *"ALLAGE"*. Parimenti, essendoci una maggioranza schiacciante di valori riferiti ad entrambi i sessi (*"BOTHSEX"*) nell'Indicatore 1.1.1, si è preferito filtrare nella tabella dell'HIV per entrambi i sessi, escludendo le due categorie *"MALE"* e *"FEMALE"*. Si utilizza dunque un *filter rows* al cui interno vengono poste entrambe le condizioni legate dall'operatore logico AND[11].

Successivamente, tramite la trasformazione *Select values*, si sono selezionate unicamente le colonne di interesse, escludendo quelle con tutti valori nulli e quelle poco significative ai fini del nostro operato. Una forte assunzione riguarda il valore contenuto all'interno di ogni riga, ovvero il numero decimale (*value*) che rappresenta i nuovi casi di HIV per un campione non infetto di 1000 persone. Al fine di avere una misura di paragone con l'Indicatore 1.1.1 che fa riferimento alla percentuale di popolazione, con un'approssimazione, si è riportato il campione nell'unità delle centinaia invece che delle migliaia. Quindi, tramite l'oggetto *Formula*, si è effettuata una semplice divisione per 10 del valore dei nuovi infetti, in modo da ottenere lo stesso pool di persone della Povertà.

Infine, si è aggiunta la trasformazione *Sort rows* utile a riordinare il risultato delle operazioni precedenti per i valori dell'attributo specificato in ordine crescente. Questo passo è obbligatorio, in quanto prerequisito per l'operazione di join, tramite l'operatore *Multiway merge join*, che sarà indispensabile a fine trattazione per mettere insieme i risultati ottenuti dai diversi Indicatori.

Ad ogni modo, nel caso dell'Indicatore 3.3.1, a seguito delle trasformazioni appena elencate, si mantengono le seguenti *features*: *GeoAreaCode, AreaName, Year e Value*. Queste colonne sono analoghe a quelle filtrate per l'indicatore precedente, in cui in aggiunta vengono mantenuti i campi relativi all'età e al genere, con la finalità di esplicitare, riga per riga, come ci si riferisca alla totalità delle età e non si faccia alcuna distinzione di sesso all'interno dei dati.

Figura 9: Pentaho - HIV1

3.3 INDICATORE 3.3.2 - TUBERCOLOSI

Relativamente alla malattia della tubercolosi, la tabella presenta schematicamente i seguenti attributi:

- *Goal*: 3
- *Indicatore*: 3.3.2
- *Target*: 3.3
- *Series Code*: SH_TBS_INCD
- *Series Description*: Tuberculosis incidence (per 100,000 population)
- *Geo Area Code*
- *Geo Area Name*
- *Time Period*: è l'anno di riferimento (assume gli stessi valori di Time Detail)
- *Value*: numero stimato di nuovi casi di tubercolosi e ricaduta che si verificano in un dato anno, espresso come tasso per 100.000 abitanti
- *Upper/Lower Bound*: range di incertezza per un datore valore
- *Source*: OMS
- *Foot Note*: Data extracted as of 5 January 2022, based on data originally compiled for the 2021 WHO Global TB Report

- *Nature*: “E” che sta per Estimated data
- *tReporting type*: “G” per Global
- *Units*: PER_100000_POP

In questo caso, né l’attributo *“Age”* né quello *“Sex”* hanno valori perché sono tutti nulli, dunque si considerano implicitamente, anche in questo caso, come sesso entrambi e come età tutte le fasce di età insieme.

In aggiunta, per semplicità elaborativa, si fa un’ulteriore approssimazione eliminando i range di incertezza, cioè prendendo come unico valore quello di riferimento.

Si cancellano, analogamente a come fatto fin’ora per gli altri data-set, anche tutte quelle colonne inutili ai fini dell’analisi.

Come ultimo passo, prima di riordinare per l’applicazione dell’operatore join, si riporta il valore su un campione di 100 individui invece che 100.000, facendo una semplice divisione analoga al caso precedente dell’HIV.

In definitiva, anche in questo caso, i campi restanti sono i seguenti: *GeoAreaCode, AreaName, Year e Value*. In Fig. 10 vengono riportati i passi appena descritti all’interno della piattaforma Pentaho.

Figura 10: Pentaho - Tubercolosi

3.4 INDICATORE 3.3.3 - MALARIA

All’intero del data-set sulla malaria sono presenti queste *features*:

- *Goal*: 3
- *Target*: 3.3
- *Indicatore*: 3.3.3
- *Series Code*: SH_STA_MALR
- *Series Description*: Malaria incidence per 1,000 population at risk (per 1,000 population)
- *Geo Area Code*
- *Geo Area Name*
- *Time Period*: è l’anno di riferimento (assume gli stessi valori di Time Detail)
- *Value*: numero di nuovi casi di malaria presenti su un campione di popolazione a rischio composto da 1000 unità
- *Upper/Lower Bound*: range di incertezza per un datore valore
- *Source*: WMR
- *Foot Note*: Nature of data for some countries in this region are either (M) modelled, E(Estimated), CA (Country adjusted) or (C) Country data
- *Nature*: E, M, CA, C e NA (Data nature not available)
- *Reporting type*: “G” per Global
- *Units*: PER_1000_POP

Si procede in modo analogo all'Indicatore precedente: si riporta il campione in un *sample* di 100 persone e si filtrano le colonne che portano informazione, rinominando gli attributi e assegnando il giusto tipo al dato, trascurando le incertezze dettate dai *bound (upper/lower)*.

L'età dei soggetti interessati non è riportata e nemmeno il genere, si ricade dunque nel caso generale per cui ci si era predisposti in partenza dall'Indicatore 1.1.1 sulla povertà.

Si conclude con il solito ordinamento delle righe per codice di area geografica. Come nei casi precedenti, le colonne rimanenti sono sempre le stesse. La sequenza di "blocchetti" utilizzata per svolgere le sopracitate operazioni è riportata in Fig. 11.

Figura 11: Pentaho - Malaria

3.5 MULTIWAY MERGE JOIN

Il lavoro svolto fin'ora è stato propedeutico al fine di predisporre i dati forniti in ingresso all'operatore *Multiway Merge Join* che permette di riunire, sotto specifiche chiavi di join, i vari data-set in un unica tabella finale. Si sono utilizzate come chiavi di join il numero associato all'area geografica (*GeoAreaCode*) e l'anno di riferimento (*Time Period*). Questo operatore, inoltre, prevede che i dati in ingresso siano riordinati rispetto alle chiavi; ciò spiega l'utilizzo, in battuta finale, dell'operatore di riordinamento *Sort rows* nei precedenti dataset analizzati.

Ad ogni modo, il *Multiway Merge Join* permette di effettuare due tipi di join: *full outer*, nel quale si prendono tutte le righe di tutti i data-set e si effettua il join, eventualmente lasciando nulle quelle righe in cui, per qualche colonna di qualche indicatore, non vi è il corrispettivo valore presente; *inner join*, al contrario, effettua un'intersezione considerando solo i casi in cui l'insieme delle righe abbiano i valori per tutte le colonne, a seguito del join, escludendo quindi tutti i valori nulli.

La scelta utilizzata è ricaduta nel primo caso, cioè del "full outer join", in modo tale da non avere perdita di informazione e, laddove possibile, procedere con tecniche di imputazione per rimpiazzare eventuali valori mancanti.

La vista completa della trasformazione totale risultante, descritta dall'inizio di questo capitolo, è riportata in Fig. 12.

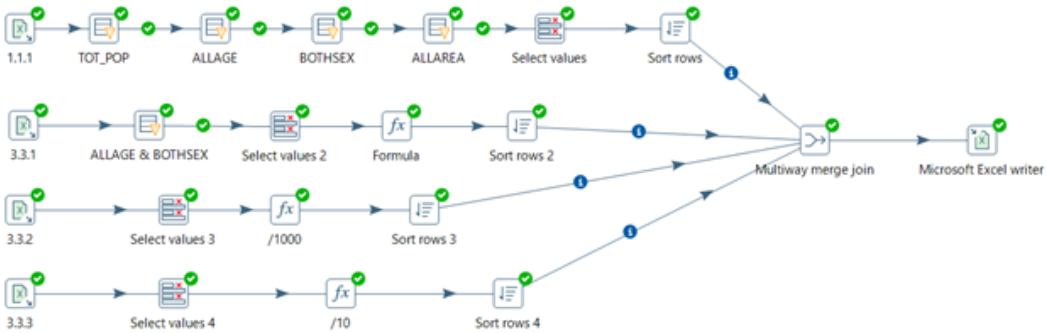

Figura 12: Pentaho - Multiway Merge Join

Il risultato finale è stato esportato sotto forma di file Excel, così da permettere, all'interno dell'ecosistema Python, varie operazioni di visualizzazione e analisi dei dati ottenuti dall'operazione di integrazione.

Innanzitutto, si vuole visualizzare la tabella finale ottenuta che, ricordiamo a seguito delle operazioni di filtraggio e integrazione, risulta essere composta dai seguenti attributi (che non sono altro che la

successione degli attributi dei singoli indicatori): *GeoAreaCode, Area Name, Year, % pop < $2.15, Age, Sex, GeoAreaCode HIV, Area Name HIV, Year HIV, Value HIV, GeoAreaCode TBS, Area Name TBS, Year TBS, Value TBS, GeoAreaCode MALR, Area Name MALR, Year MALR, Value MALR.*

Per una migliore comprensione viene riportata di seguito (Fig. 13) una tabella in cui si prende in considerazione l'area geografica "World" con i propri valori.

GeoAreaCode	Area Name	Year	% pop < $2,15	Age	Sex	GeoAreaCode_I	Area_Name_HIV	Year_HIV	ValueHIV	GeoAreaCode_TBS	Area_Name_TBS	Year_TBS	Value_TBS	GeoAreaCode	Area_Name_MALR	Year_MALR	Value_MALR
1	World	2000	29	ALLAGE	BOTHSEX	1	World	2000	0,048	1	World	2000	0,174	1	World	2000	8,109833
1	World	2001	28	ALLAGE	BOTHSEX	1	World	2001	0,045	1	World	2001	0,174	1	World	2001	8,152067
1	World	2002	27	ALLAGE	BOTHSEX	1	World	2002	0,043	1	World	2002	0,174	1	World	2002	7,845808
1	World	2003	26	ALLAGE	BOTHSEX	1	World	2003	0,041	1	World	2003	0,174	1	World	2003	7,808161
1	World	2004	24	ALLAGE	BOTHSEX	1	World	2004	0,04	1	World	2004	0,173	1	World	2004	7,756796
1	World	2005	22	ALLAGE	BOTHSEX	1	World	2005	0,038	1	World	2005	0,171	1	World	2005	7,595978
1	World	2006	21	ALLAGE	BOTHSEX	1	World	2006	0,037	1	World	2006	0,169	1	World	2006	7,316768

Figura 13: Excel - Tabella Full Outer Join

Quindi, ad esempio, se si volesse prendere come riferimento l'area geografica dell'intera Terra nell'anno 2006 si può constatare che: 21% viveva sotto la soglia di estrema povertà, su un campione di persone sane composto da 100 unità vi sono 0,037 nuovi casi di HIV, un aumento di casi stimati di tubercolosi su 100 persone dello 0,169 e, per finire, su un campione di 100 persone a rischio, si hanno nuovi casi stimati di malaria pari ad un numero di circa 7,32 persone.

Al fine di avere una maggiore leggibilità nella visualizzazione, si eliminano, per mezzo della libreria Pandas, alcune colonne ridondanti, quali quelle del luogo e dell'anno di valutazione per il Target 3.3 e si mantengono esclusivamente quelle dell'Indicatore 1.1.1 perché analoghe. Pertanto, si riordinano le colonne e si ottiene il risultato illustrato in Fig. 14.

	Year	GeoAreaCode	Area Name	% pop < $2,15	ValueHIV	Value_TBS	Value_MALR	Age	Sex
0	2000.0	1.0	World	29.0	0.048	0.174	8.109833	ALLAGE	BOTHSEX
1	2001.0	1.0	World	28.0	0.045	0.174	8.152067	ALLAGE	BOTHSEX
2	2002.0	1.0	World	27.0	0.043	0.174	7.845808	ALLAGE	BOTHSEX
3	2003.0	1.0	World	26.0	0.041	0.174	7.808161	ALLAGE	BOTHSEX
4	2004.0	1.0	World	24.0	0.040	0.173	7.756796	ALLAGE	BOTHSEX
5	2005.0	1.0	World	22.0	0.038	0.171	7.595978	ALLAGE	BOTHSEX
6	2006.0	1.0	World	21.0	0.037	0.169	7.316768	ALLAGE	BOTHSEX
7	2007.0	1.0	World	20.0	0.035	0.167	7.125073	ALLAGE	BOTHSEX
8	2008.0	1.0	World	19.0	0.034	0.164	6.990005	ALLAGE	BOTHSEX
9	2009.0	1.0	World	18.0	0.032	0.161	7.001493	ALLAGE	BOTHSEX

Figura 14: Pandas - Tabella Finale Full Outer Join

Come prima osservazione si vuole esplorare la lunghezza del Data frame e la quantità di valori nulli. In totale si registrano 4018 righe, mentre 1819 valori nulli per l'Indicatore 1.1.1, 818 e 828 rispettivamente per gli Indicatori 3.3.1 e 3.3.2 e 1666 valori mancanti per l'Indicatore 3.3.3.

```
#Lunghezza df
len(data_OUTER)
```

```
4018
```

```
#faccio il conteggio dei valori nulli
data_OUTER.isnull().sum()
```

```
GeoAreaCode      1819
Area Name        1819
Year             1819
% pop < $2,15    1819
Age              1819
Sex              1819
ValueHIV          818
Value_TBS         828
Value_MALR       1666
dtype: int64
```

Figura 15: Pandas - Lunghezza dataframe e valori nulli

Il primo approccio per gestire la grande quantità di valori nulli a disposizione si è rivolto verso l'osservazione del range di anni disponibili lungo il quale si distribuiscono i dati. Dall'analisi viene fuori come solamente un Paese (in una riga) ha i valori per l'anno 2021 e si ha un numero esiguo di righe, 20, per l'anno 2020. A causa della scarsità di dati, e quindi conseguente impossibilità di effettuare confronti, si è deciso di eliminare quelli che fanno riferimento agli anni 2020 e 2021 (Fig. 16).

```
#visualizza tutte le righe (senza puntini)
pd.set_option('display.max_rows', None)
```

```
#conto i valori degli anni
data_OUTER["Year"].value_counts(sort=False)
```

```
2000.0     85
2001.0     72
2002.0     93
2003.0     98
2004.0    110
2005.0    113
2006.0    114
2007.0    109
2008.0    114
2009.0    116
2010.0    123
2011.0    117
2012.0    124
2013.0    113
2014.0    120
2015.0    122
2016.0    118
2017.0    112
2018.0    116
2019.0     89
2020.0     20
2021.0      1
Name: Year, dtype: int64
```

```
#cancello gli anni 2020 e 2021 perché hanno pochi valori
data_OUTER.drop(data_OUTER[data_OUTER.Year >= 2020].index, inplace=True)
```

```
len(data_OUTER)
```

```
3997
```

Figura 16: Pandas - Conteggio range di anni disponibili

Il passo successivo, per gestire i *NULL*, prevede la scelta di un Indicatore di riferimento, in questo caso l'1.1.1, con il quale procedere con l'analisi. Quindi, si prelevano, tramite la funzione di Pandas *groupby*, tutti i valori "buoni" della Povertà, escludendo di fatto i valori mancanti degli altri dataset in una sorta di "left join".

A seguito dell'operazione di raggruppamento, le righe restanti diventano 3.885, coinvolgendo valori nulli solo per i rimanenti tre indicatori sulle malattie (HIV, TUBERCOLOSI e MALARIA).

A questo punto si è interessati a visualizzare quali nazioni sono presenti e in che quantità, cioè quanti anni ricoprono (senza conoscere esplicitamente l'arco temporale di appartenenza). Per ovvi motivi di visualizzazione non sono riportate tutte le nazioni, ma solo il conteggio del range di presenza (Fig. 17).

```
Africa                    20
Kyrgyzstan                20
Republic of Moldova       20
Peru                      20
Central America           20
                          ..
Tuvalu                     1
Central African Republic   1
Papua New Guinea           1
Syrian Arab Republic       1
Marshall Islands           1
```

Figura 17: Pandas - Conteggio range di anni disponibili

Giunti a tale punto, il data-set necessita di preparazione per le successive operazione di imputazione, in modo da riempire quei valori nulli causati dalla mancanza di dati.

Prima di procedere con il rimpiazzamento dei valori nulli, si è deciso di porre una soglia, sotto la quale vengono escluse quelle nazioni che abbiano un numero di righe (corrispondenti agli anni temporali che ricoprono) tali da non essere presi in considerazione per il successivo riempimento dei *NULL*. Tale

operazione risulta necessaria in quanto il metodo dell'imputazione stima i valori mancanti sulla base di quelli già presenti; pertanto se avesse a disposizione pochi dati, la valutazione risulterebbe poca accurata e, talvolta, priva di significato.

Per tale ragione, la soglia è stata stabilita uguale a 16 e successivamente sono stati eliminati tutti quei Paesi con un un numero inferiore alle 16 righe; in altre parole i territori con un numero di anni mancanti superiore a 4, nell'intervallo temporale che va dal 2000 al 2019, sono state rimosse.

Si ottengono così dati con nazioni (e continenti) che coprono un arco temporale di almeno 16 anni e fino a 20 anni.

In seguito, dall'analisi dei risultati ottenuti si è notato che gruppi di valori nulli si concentravano, per indicatore, a blocchetti in riferimento a intere nazioni. Siccome è stato preso come riferimento l'Indicatore sulla povertà, a seguito del full join, si sono riscontrati blocchi di valori nulli per intere nazioni per gli altri indicatori, non permettendo alcuna operazione di imputazione. Questa condizione rendeva impraticabile il confronto tra nazioni e/o continenti; per i suddetti motivi si è deciso di eliminare quelle righe che presentavano almeno un valore nullo: le righe risultanti sono 692. Affinché il lettore abbia ben chiaro le aree geografiche rimanenti si illustra, in Fig. 18, il nome e il numero di righe associate.

World	20
Panama	20
Kyrgyzstan	20
Northern America	20
Indonesia	20
Oceania	20
Georgia	20
Europe and Northern America	20
Europe	20
El Salvador	20
Eastern and South-Eastern Asia	20
Latin America and the Caribbean	20
Eastern Asia	20
Dominican Republic	20
Costa Rica	20
Peru	20
South-Eastern Asia	20
Sub-Saharan Africa	20
Australia and New Zealand	20
Least Developed Countries (LDCs)	20
Paraguay	19
Argentina	19
Landlocked developing countries (LLDCs)	19
Armenia	19
Honduras	19
Northern Africa	18
Northern Africa and Western Asia	18
Ecuador	18
Colombia	18
Central and Southern Asia	18
Brazil	18
Southern Asia	18
Bolivia (Plurinational State of)	18
Western Asia	18
Kazakhstan	18
Thailand	17

Figura 18: Pandas - Conteggio range di anni disponibili

E' giunto il momento di procedere con tecniche di riempimento dei valori mancanti: si calcolano gli anni delle nazioni per cui mancano i valori e si assegnano, per quegli anni, la media dei valori della rispettiva colonna. Come esempio, si consideri la Thailandia: gli anni mancanti sono il 2001, 2003 e 2005; sono state calcolate le medie per i tre indicatori (per gli anni disponibili) ed il risultato è stato assegnato alle tre righe mancanti degli anni (Fig. 19).

Area_Name	Year	% pop < $2,15	ValueHIV	Value_TBS	Value_MALR
Thailand	2000.0	4.000000	0.066000	0.241000	0.657640
Thailand	2002.0	2.000000	0.052000	0.244000	0.366474
Thailand	2004.0	1.000000	0.042000	0.232000	0.216402
Thailand	2006.0	1.000000	0.032000	0.215000	0.242573
Thailand	2007.0	1.000000	0.029000	0.205000	0.264183
Thailand	2008.0	0.000000	0.026000	0.196000	0.226291
Thailand	2009.0	0.000000	0.024000	0.188000	0.232192
Thailand	2010.0	0.000000	0.023000	0.181000	0.254727
Thailand	2011.0	0.000000	0.021000	0.176000	0.194321
Thailand	2012.0	0.000000	0.020000	0.172000	0.364302
Thailand	2013.0	0.000000	0.018000	0.170000	0.321720
Thailand	2014.0	0.000000	0.017000	0.167000	0.317381
Thailand	2015.0	0.000000	0.016000	0.163000	0.061522
Thailand	2016.0	0.000000	0.014000	0.160000	0.056754
Thailand	2017.0	0.000000	0.013000	0.156000	0.043356
Thailand	2018.0	0.000000	0.012000	0.153000	0.030946
Thailand	2019.0	0.000000	0.011000	0.150000	0.024205
Thailand	2001.0	0.529412	0.025647	0.186412	0.227941
Thailand	2003.0	0.529412	0.025647	0.186412	0.227941
Thailand	2005.0	0.529412	0.025647	0.186412	0.227941

Figura 19: Pandas - Riempimento nulli Thailandia

4 PYTHON - VISUALIZZAZIONI

Una volta aver concluso la fase di ETL e riempito i valori nulli per mezzo della tecnica di imputazione, in tale sezione ci si è concentrati sulla traduzione dei dati in rappresentazione visiva, più facilmente elaborabile per garantire precisione e dettaglio allo scopo di estrarre informazioni utili dai dati a disposizione.

4.1 NAZIONI

A seguito delle varie operazioni di ETL e delle tecniche di imputazione, si è visto come i Paesi rimanenti siano un numero esiguo. Svolgere analisi di tipo descrittivo, o addirittura di tipo predittivo, appare decisamente complicato con un numero così ristretto di informazioni. Come si può osservare dalla Figura 20, le nazioni rimanenti comprendono gran parte dell'America Latina, ad eccezione di: Messico, Nicaragua, Guatemala, Venezuela, Guyana, Cile e Uruguay. L'Europa e il Nord America sono completamente esclusi da questi valori; così come tutta l'Africa e gran parte dell'Asia, ad eccezione di Kazakistan, Kirghizistan, Armenia, Georgia, Thailandia e Indonesia. Per avere un'idea visiva delle nazioni rimanenti si osservi la Figura 20, che rappresenta la media dei valori di tubercolosi negli anni 2000-2019 per le nazioni rimaste.

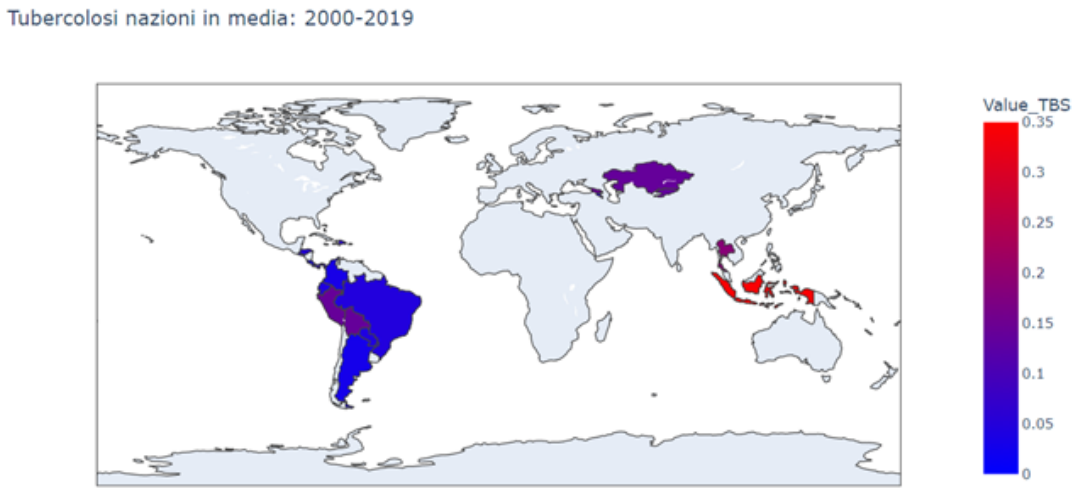

Figura 20: TBS valori medi degli anni 2000 - 2019

4.2 CONTINENTI

Le nazioni unite effettuano una macro-suddivisione delle aree continentali per il report degli indicatori SDG come riportato in Fig. 21.

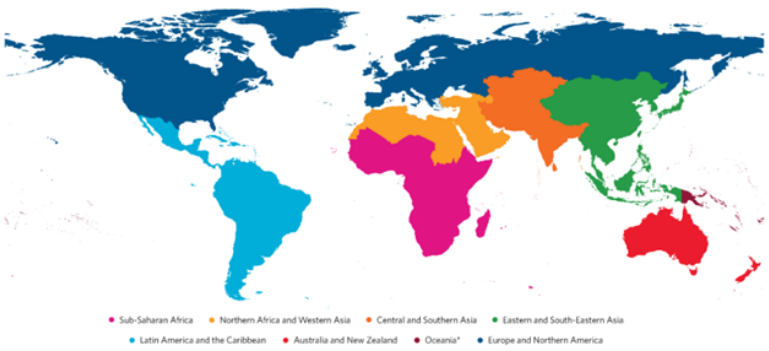

Figura 21: UN SDG Indicators - Regional groupings used in Report and Statistical Annex

Seguendo questo pattern di suddivisione sono state svolte delle visualizzazioni al fine di comprendere meglio i dati: per ognuno dei quattro indicatori vengono "plottati" dei grafi a linea che evidenziano l'andamento dei valori negli anni. Ogni "linea" rappresenta un continente e percorre l'asse delle ascisse dall'anno 2000 fino al 2019.

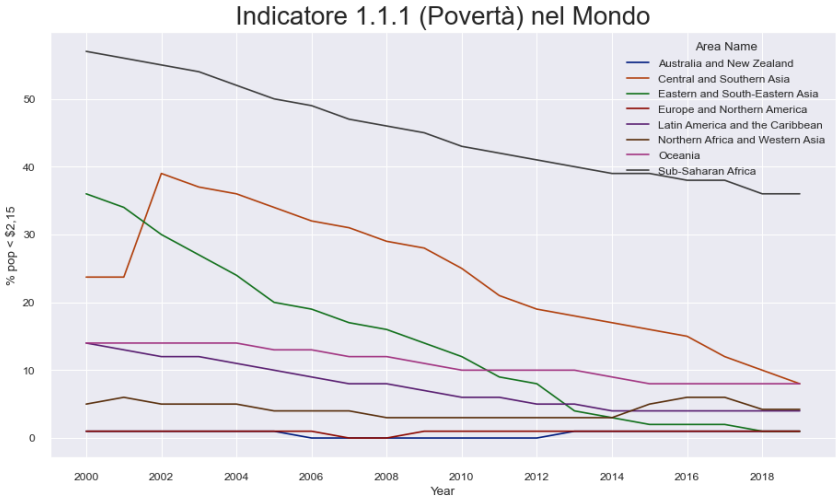

Figura 22: Python - Andamento Povertà Continenti

Si può osservare come la tendenza generale sia ad una diminuzione dei valori per tutti gli indicatori. Nella mappa sovrastante (Figura 22) si è preso in considerazione l'Indicatore 1.1.1. Complessivamente si ha, come già detto, una tendenza decrescente delle curve, ma si fa molto più accentuata per l'Asia Orientale e Sud Orientale (linea verde), che parte da valori superiori al 35% fino a giungere, per i giorni nostri, a valori prossimi allo zero, cioè vicini a quelli Europei, rimasti sempre stabili. Un comportamento analogo si osserva per l'Asia Centrale e Meridionale (linea arancione), anche se non si arriva a valori così bassi per gli anni più recenti. La situazione peggiore è rappresentata dall'Africa Subsahariana (linea nera) che, anche se ha una forte tendenza negativa, si stabilizza su valori estremamente alti per i tempi moderni, circa 35%.
Nel grafico successivo (Figura 23) si prendono in esame i valori dell'HIV. Si è dovuta escludere l'area riguardante l'Africa sub sahariana perché non permetteva una buona visualizzazione, avendo dei valori troppo elevati da schiacciare tutte le altre linee in modo compatto e indistinguibile. Per il resto si osserva una tendenza abbastanza stabile delle infezioni da HIV.

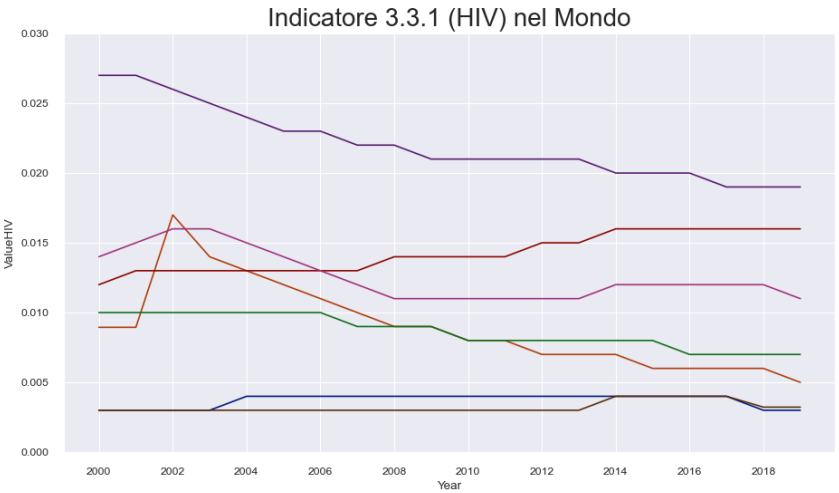

Figura 23: Python - Andamento HIV Continenti

Nella rappresentazione grafica sulla Malaria (Figura 24) si ha un discorso analogo a quella sulla HIV, ma in aggiunta viene esclusa anche l'Oceania, anche se rimane sempre sotto la curva dell'Africa Subsahariana, che quindi anche in questo Indicatore riesce a far peggio di tutti gli altri continenti. Analogamente a prima si possono notare dei comportamenti stabili per quelle aree geografiche con valori bassi ed una tendenza alla diminuzione per le aree più sofferenti.

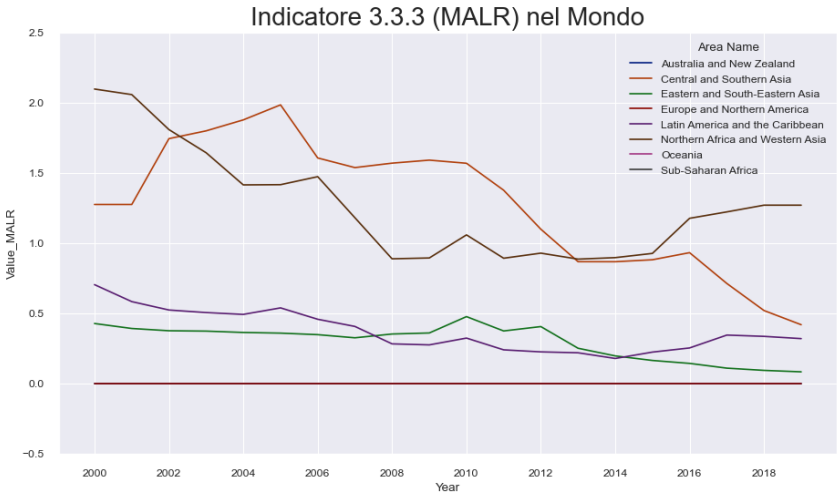

Figura 24: Python - Andamento Malaria Continenti

La figura 25, invece, riguarda l'andamento della Tubercolosi. L'Africa Subsahariana si riconferma la peggiore, anche se ha un andamento decrescente, così come l'Asia Centrale e Meridionale, che la segue per numero di casi. Per il resto si osserva un andamento abbastanza stabile.

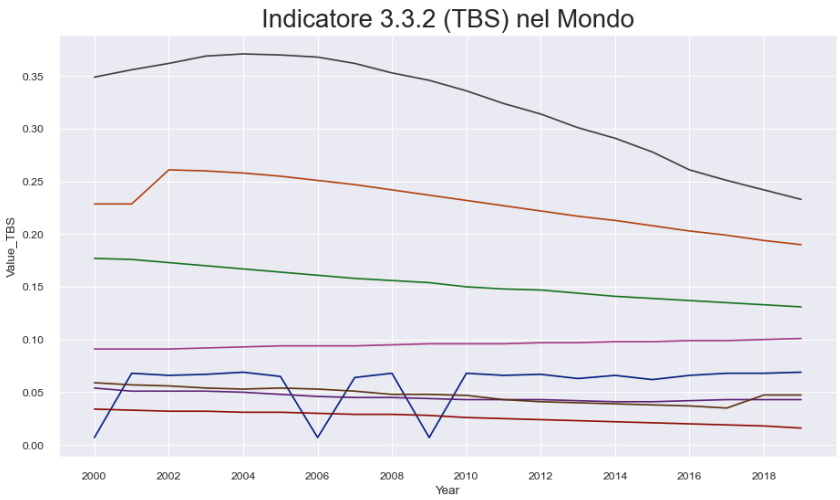

Figura 25: Python - Andamento Tubercolosi Continenti

Come ultima analisi si è cercato di individuare la correlazione tra i vari indicatori. Tramite una mappa di calore (Figura 26) si sono riscontrati i seguenti, estremamente alti, valori: l'Indicatore sulla povertà e quello sulla tubercolosi hanno una correlazione dello 0.93; le infezioni da HIV sono correlate, per lo 0.8, sempre con l'Indicatore 1.1.1 sulla povertà. Al terzo posto si osserva una correlazione dello 0.79 tra le infezioni da HIV e quelle della Malaria. Tra infezioni di HIV e Malaria si ha una correlazione dello 0.74. Infine, con una correlazione dello 0.71 si palesano le infezioni da Malaria e i valori della Povertà. Da queste considerazioni si può dedurre come, prendendo come riferimento l'Indicatore sulla povertà, ad un valore alto corrispondente, cioè ad una forte concentrazione di popolazione povertà in una determinata area geografica, si possono osservare probabilmente anche molti casi di TBS e di HIV.

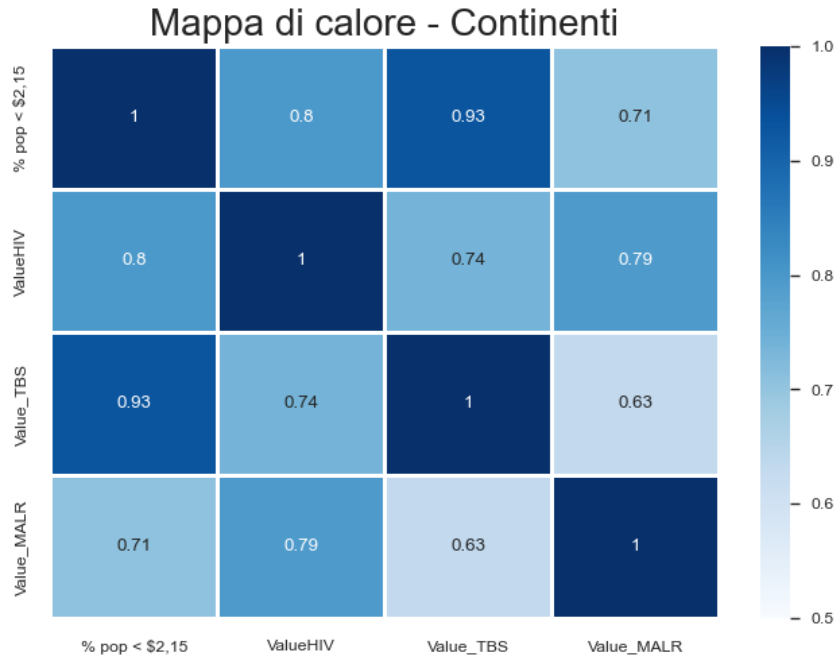

Figura 26: Python - Mappa di Calore Continenti

Parte II
Parte Seconda

5 Scelta Indicatori

Una volta aver analizzato la feature "Series Code" dell'indicatore alla povertà relativa a tutta la popolazione ed estrapolato le conclusioni discusse nei capitoli precedenti, in tale sezione, invece, ci soffermiamo sul valore univoco "Series Code" che tratta la popolazione con almeno un impiego lavorativo.

In primo luogo, prima di addentrarci nello studio di analisi della condizione di povertà cui è costretta a subire parte della popolazione nonostante fosse in possesso di un impiego lavorativo, è opportuno soffermarci sui metadati dei singoli dataset che, nelle sotto-sezioni che seguono, andremo ad elaborare. I metadati, infatti, messi a disposizione dall'Organizzazione delle Nazioni Unite, consentono una maggiore comprensione dei dati stessi al fine di avere la chiave attraverso cui abilitare più agevolmente la ricerca del fenomeno che si vuole esaminare.

Innanzitutto, si è partiti dal dataset in un certo senso preso di riferimento, ovvero la percentuale della popolazione con occupazione che vive in condizione di povertà assoluta.

5.1 Indicatore 1.1.1b - Popolazione con occupazione che vive al di sotto della soglia di povertà internazionale (%)

Definizione: Percentuale di popolazione con occupazione al di sotto della povertà internazionale di 1.90$ al giorno, è definita come la quota di persone occupate che vivono in famiglie con un consumo o un reddito pro-capite inferiore alla soglia di povertà internazionale.

Nelle sezioni precedenti dell'elaborato è stato acquisito come l'indicatore della povertà assoluta avesse due elementi univoci della colonna *Series Name*: percentuale dell'intera popolazione e percentuale della popolazione con occupazione.

Una delle motivazioni per cui si rende necessaria tale suddivisione è da ricercare nella cause profonde che determinano la condizione di miseria. Per sradicare la povertà, di fatto, è inevitabile comprenderne la radice.

A tal proposito, l'accesso ad un lavoro retribuito, spesso non è garanzia di condizioni di lavoro dignitose o di un reddito adeguato per molti dei 3,3 miliardi di lavoratori in tutto il mondo.

Nello specifico dell'Indicatore in questione, il tasso di povertà lavorativa rivela la percentuale di popolazione occupata che vive in condizioni di povertà nonostante sia occupata, il che implica che i redditi legati all'occupazione non sono sufficienti a far uscire dalla povertà gli stessi lavoratori e le loro famiglie ed, inoltre, a garantire condizioni di vita dignitose. L'adeguatezza dei guadagni, dunque, è un aspetto fondamentale della qualità del lavoro e questo deficit nella qualità del lavoro potrebbe mantenere i lavoratori e le loro famiglie in condizioni di povertà.

Il metodo di calcolo utilizzato è mostrato di seguito:

$$\textit{Tasso di povertà lavorativa (\%)} = \frac{\textit{Popolazione con occupazione povera}}{\textit{Totali occupati}} x100$$

L'agenzia che si occupa dell'elaborazione dei microdati è l'ILO (Organizzazione Internazionale del Lavoro) tant'è che viene considerata il punto di riferimento delle Nazioni Unite per le statistiche sul lavoro. Essa, infatti, compila e produce statistiche sul lavoro con l'obiettivo di diffondere serie di dati ad alta qualità comparabili a livello internazionale sul mercato del lavoro.

Purtroppo, a livello nazionale i confronti nel tempo possono essere influenzati da diversi fattori quali i cambiamenti nei tipi di indagine e nei metodi di raccolta dei dati. L'uso della parità di potere d'acquisto (PPP) (l'indice che consente di confrontare i livelli dei prezzi tra località diverse) piuttosto che dei tassi di cambio di mercato (il tasso attraverso il quale è possibile scambiare una moneta con un'altra) garantisce che si tenga conto delle differenze nei livelli di prezzo tra i vari Paesi. Tuttavia, non si può affermare categoricamente che due persone in due Paesi diversi, che vivono al di sotto di 1,90 $ al giorno, si trovino ad affrontare lo stesso grado di privazione o di bisogno.

Tra l'altro, la povertà è un concetto che si applica alle famiglie e non ai singoli individui, sulla base dell'ipotesi che le famiglie mettano a fattor comune il proprio reddito. Questo presupposto, però, potrebbe non essere sempre vero.

In aggiunta, laddove ci siano mancanza di dati per i Paesi e gli anni vengono effettuate delle stime basate su un modello di regressione multivariata.

Al fine di trovare una corrispondenza di dati con l'indicatore appena discusso per poi individuare trend e regolarità tra essi, si è fatto uso della voce "Indicatori correlati" presente nei metadati dell'Indicatore stesso. In seguito è stata effettuata una selezione degli indicatori correlati che fossero più rilevanti per il nostro scopo i cui metadati verranno approfonditi nelle sotto-sezioni che seguono, in coda all'introduzione del GOAL principale.

5.2 Goal 8 - LAVORO DIGNITOSO E CRESCITA ECONOMICA

In linea generale, l'occupazione e la crescita economica svolgono un ruolo significativo nella lotta alla povertà. La promozione di una crescita sostenibile e la creazione di sufficienti posti di lavoro dignitoso e rispettoso dei diritti umani sono di fondamentale importanza non solo per i paesi in via di sviluppo ma anche per le economie emergenti e quelle industrializzate.

Target 8.2: raggiungere livelli più elevati di produttività economica attraverso la diversificazione, l'aggiornamento tecnologico e l'innovazione, anche mirando ad un alto valore aggiunto nei settori ad alta intensità di manodopera

5.2.1 Indicatore 8.2.1: Tasso di crescita annuale del PIL reale per occupato

Il tasso di crescita annuale del Prodotto Interno Lordo (PIL) reale per occupato trasmette la variazione percentuale annuale del PIL reale per occupato.

Il PIL, infatti, è la principale misura della produzione nazionale che rappresenta il valore totale di tutti i beni e servizi finali all'interno del perimetro di produzione del Sistema dei Conti Nazionali (SNA) prodotti in un determinato paese in un determinato anno.

In particolare, con il termine *PIL reale* ci si riferisce al PIL calcolato a prezzi costanti, escludendo l'effetto dell'inflazione e favorendo il confronto tra quantità al di là delle variazioni di prezzo. Per quanto riguarda le sue stime, esse sono calcolate esprimendo i valori in termini di periodo base. Si identificano, dunque, le componenti di prezzo e quantità di un valore e si sostituisce il prezzo del periodo base con quello del periodo corrente.

E' da tenere presente che il calcolo del PIL reale fa riferimento a tutte le persone in età lavorativa che, durante un breve periodo di riferimento (almeno una settimana), sono state impegnate in una qualsiasi attività di produzione di beni e servizi o fornitura di servizi a scopo di retribuzione o profitto.

Peraltro, essendo il PIL reale per persona occupata, tale indicatore rappresenta, a tutti gli effetti, una misura della crescita della produttività del lavoro, fornendo così informazioni sull'evoluzione, l'efficacia e la qualità del capitale umano nel processo produttivo.

La crescita economica di un Paese può essere attribuita a molti fattori, tra cui l'aumento dell'occupazione e un lavoro più efficace da parte di coloro che sono occupati. Tale indicatore si focalizza proprio su quest'ultimo effetto rappresentando, quindi, una misura chiave della performance delle politiche sul mercato del lavoro al fine di monitorarne gli andamenti.

Il metodo di calcolo utilizzato è mostrato di seguito:

$$PIL\ reale\ per\ occupato\ (\%) = \frac{PIL\ a\ prezzi\ costanti}{Occupazione\ totale} x100$$

Innanzitutto, è da specificare che il numeratore e il denominatore devono riferirsi allo stesso periodo di riferimento, ad esempio lo stesso anno solare.

Le misure per la produzione utilizzate al numeratore del PIL sono ottenute dai conti nazionali e rappresentano, per quanto possibile, il PIL ai prezzi di mercato per l'economia aggregata (aggiustato per l'inflazione, a prezzi costanti appunto).

I dati sull'occupazione utilizzati al denominatore, invece, derivano da indagini sulle forze di lavoro o dalle famiglie con un modulo di occupazione. In mancanza di ciò, si possono anche utilizzare indagini sugli stabilimenti, registri amministrativi o stime ufficiali basate su fonti affidabili.

Tuttavia, nonostante i principi comuni si basano principalmente sul Sistema dei Conti Nazionali delle Nazioni Unite, esistono ancora problemi significativi di comparabilità internazionale dei dati, in particolare delle differenze metodologiche tra i vari Paesi quali le differenze nel trattamento della produzione nei settori dei servizi e le differenze nei metodi utilizzati per correggere le misure della produzione per le variazioni di prezzo. Tutto ciò si traduce in un eminente limite dell'utilizzo dei dati stessi, in quanto rappresentano un ostacolo per il raggiungimento alla coerenza internazionale delle stime dei conti nazionali.

5.2.2 Indicatore 8.5.1: Retribuzione oraria media dei lavoratori dipendenti

Come già noto dall'esperienza, le retribuzioni sono un aspetto fondamentale della qualità dell'occupazione nonchè delle condizioni di vita.

Ai fini della comparabilità internazionale, le statistiche sulle retribuzioni si riferiscono alla retribuzione lorda dei dipendenti, cioè il totale prima di qualsiasi deduzione effettuata dal datore di lavoro per quanto riguarda le tasse, i contributi dei dipendenti ai regimi previdenziali e pensionistici, i premi di assicurazione sulla vita e le quote sindacali.

Tuttavia, per una visione più accurata, il campo di applicazione delle statistiche è limitato alla copertura dell'indagine sugli stabilimenti (escludendo le piccole imprese, le imprese agricole e/o le imprese del settore informale); le indagini sugli stabilimenti, infatti, costituiscono la fonte più affidabile, data l'elevata accuratezza dei dati sulla retribuzione che ne derivano.

Il metodo di calcolo utilizzato per la retribuzione oraria media è mostrato di seguito:

$$\textit{Retribuzione oraria media} = \frac{\textit{(retribuzione oraria x ore lavorate da ciascun dipendente)}}{\textit{numero totale di ore lavorate da tutti i dipendenti}}$$

L'unità di misura è la valuta locale corrente e la ripartizione per professione si basa sull'ultima versione della Classificazione Internazionale standard delle professioni (ISCO).

Ad ogni modo, le statistiche sulle retribuzioni rappresentano una serie di complicazioni in termini di comparabilità internazionale, la maggior parte delle quali deriva dalla varietà delle possibili fonti di dati. Le varie fonti disponibili, infatti, differiscono per metodi, obiettivi e portata, il che influenza i risultati ottenuti. La copertura della fonte, inoltre, può variare in termini di aree geografiche coperte, di lavoratori coperti (ad esempio possono essere esclusi i lavoratori part-time o i lavoratori informali) e di stabilimenti coperti (ad esempio, possono essere esclusi gli stabilimenti al di sotto di una certa dimensione o di un certo settore).

Proprio per tale ragione, nei casi in cui le retribuzioni dei lavoratori esclusi dalla copertura della fonte siano significatamente diverse da quelle dei lavoratori inclusi, le statistiche non sarebbero del tutto rappresentative del Paese nel suo complesso e non sarebbe strettamente comparabili con quelle dei Paesi che utilizzano una fonte più completa.

Per concludere, dunque, non sarebbe del tutto accurato confrontare, ad esempio, le retribuzioni orarie di un'indagine sulle forze di lavoro di un Paese con le retribuzioni orarie di un'indagine sulle imprese di un altro Paese.

6 Integrazione dei dati

Terminata la fase di lettura dei metadati, in tale sezione verrà illustrato ciascun Indicatore selezionato per la correlazione in modo tale da rendere esplicita la loro composizione tabellare.

Successivamente, verrà effettuata la combinazione dei dati memorizzati nei diversi dataset, la quale restituirà, in forma filtrata, un'unica tabella al fine di avere una visione complessiva concerne il fenomeno dei lavoratori in condizioni di povertà nonchè l'andamento delle misure del PIL e della retribuzione oraria media.

Lo strumento che si è scelto di utilizzare per la combinazione dei diversi dati dati è Pentaho Data Integration, una suite di prodotti software di Business Intelligence che fornisce, appunto, ottimi servizi di integrazione dati.

6.0.1 Indicatore 1.1.1b - Lavoratori poveri

Una volta aver scaricato il file Excel dell'indicatore 1.1.1b dal sito ufficiale delle Nazione Unite, quest'ultimo viene caricato all'interno del software di casa Hitachi per mezzo di un operatore di input *"Input Excel"*.

Per quanto riguarda l'elenco dei campi del presente dataset si rimanda alla sottosezione 3.1 in cui vengono largamente descritti. Tuttavia, per una migliore leggibilità vengono riportati, di seguito, i principali campi degni di nota per la fase di preparazione, necessaria al fine di predisporre i dati da fornire in input agli operatori successivi:

- *Series Code*: codice identificativo per distinguere tra la popolazione in generale ("SI_POV_DAY1" -> Metadata 01-01-01a) e la popolazione con un impiego lavorativo ("SI_POV_EMP1" -> Metadata 01-01-01b)
- *Series Description*: descrizione del codice precedente assumendo il seguente valore: "Proportion of population below international poverty line (%)" per la popolazione in generale; mentre, per il sottoinsieme di popolazione con impiego lavorativo, l'attributo riporta "Employed population below international poverty line, by sex and age (%)"
- *Geo Area Code*: valore numerico associato univocamente ad una specifica area geografica
- *Geo Area Name*: nome delle diverse aree geografiche
- *Time Period*: anno di riferimento
- *Value*: percentuale di popolazione che vive con meno di $2,15 al giorno ai prezzi internazionali del 2017
- *Age*: assume i seguenti valori: 15+, 15-24, 25+
- *Sex*: "BOTHSEX", "MALE" e "FEMALE"

Come è consuetudine fare prima di intraprendere qualsiasi operazione di ETL all'interno di un dataset, è stato verificato il conteggio dei valori nulli in modo da visualizzare la propria composizione tabellare. Il risultato, però, mostra come le colonne di interesse non presentino valori nulli, ma vi sono semplicemente dei campi pochi significativi, i quali sono interamente vuoti. Pentaho gestisce le *features* completamente nulle e quelle prive di informazione attraverso l'operatore *"Select Values"* per rimuoverle.

Terminato il riepilogo dei campi contenuti nel dataset, come già noto dalla sezione precedentemente discussa, in questa seconda unità consideriamo solamente il sottoinsieme della popolazione con almeno un impiego lavorativo. In tal senso, quindi, si sono escluse le righe inerenti la popolazione nella sua totalità e applicato il filtraggio per tutte quelle righe che, invece, riguardassero i lavoratori. Tale azione è stata svolta mediante l'operatore *"Filter Rows"* messo a disposizione dalla stessa piattaforma Pentaho.

Per quanto riguarda l'età, essa è distribuita abbastanza uniformemente nelle fasce 15+, 15-24, 25+ come mostrato nella Fig.27. Tuttavia, la scelta si è riversata sul tipo di aggregazione ad alto livello prediligendo il valore 15+, a prova del fatto che lo standard internazionale considera età lavorativa a partire dai 15 anni.

```
15+      11036
15-24    10680
25+      10680
Name: Age, dtype: int64
```

Figura 27: Il campo *Age* - Povertà con impiego lavorativo

Un approccio simile può essere applicato anche per il genere. Tant'è vero che nei dataset del PIL e la retribuzione oraria media con i quali si vuole fare integrazione, non viene sempre inclusa una distinzione dei sessi.

Dunque, si è deciso di applicare il filtraggio, prima per età (15+) e poi per genere (BOTHSEX) in modo tale da avere una migliore combinazione dei diversi dataset e sottrarsi dagli eventuali valori ridondanti. L'operatore disponibile all'interno dello strumento software Pentaho che ha permesso ciò prende il nome di *"Filter Rows"*.

Infine, è stata inserita la trasformazione intitolata *"Sort Rows"* al fine di riordinare, in modo crescente, il risultato delle operazioni precedenti secondo i valori di una specifica feature. Tale passo è obbligatorio, in quanto prerequisito per l'operazione di join, indispensabile per unire i risultati ottenuti dai diversi Indicatori.

Alla luce delle modifiche appena elencate, si mantengono le seguenti features: *GeoAreaCode, AreaName, TimePeriod, Value.* Nella Fig. 28 viene mostrata la seguenza di trasformazioni

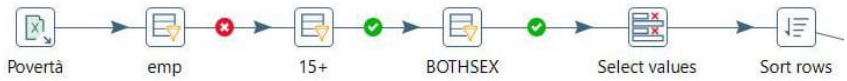

Figura 28: Pentaho - Lavoratori in condizioni di povertà

6.0.2 Indicatore 8.2.1 - Tasso di crescita annuale del PIL reale per occupato

Per semplicità e per facilitare la leggibilità del dataset in questione, vengono riportate, di seguito, soltanto le feature significative ai fini dello scopo di analisi:

- *GeoAreaName*: valore numerico associato univocamente ad una specifica area geografica
- *GeoAreaName*: nome delle diverse aree geografiche
- *TimePeriod*: anno di riferimento
- *Value*: percentuale del tasso di crescita annuale del PIL reale per occupato
- *Units*: unità di misura adottata "PERCENT"

Come anticipato nel paragrafo precedente, gli attributi *"Age"* e *"Sex"* contengono tutti valori nulli, conseguenza del fatto che il fenomeno del PIL non riconosce le distinzioni tra sesso ed età, prefissando la rappresentanza del territorio nella condizione generale in un unico dato.

Pertanto, non avendo modifiche di filtraggio e di calcolo da apportare al dataset, le uniche trasformazioni applicate sono state *Select Values* per la rimozione delle colonne inutili e *Sort Rows* per l'ordinamento crescente delle righe.

In definita, anche in questo caso, i campi restanti sono i seguenti: *GeoAreaCode, AreaName, TimePeriod, Value.* Nella Fig. 29 viene mostrata la seguenza di trasformazioni.

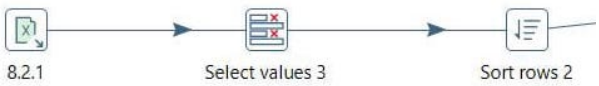

Figura 29: Pentaho - PIL reale per occupato

6.0.3 Indicatore 8.5.1: Retribuzione oraria media dei lavoratori dipendenti

Come nella descrizione del dataset precedente, vengono riportate, di seguito, soltanto le features degne di nota per gli scopi dell'elaborato:

- *GeoAreaName*: valore numerico associato univocamente ad una specifica area geografica
- *GeoAreaName*: nome delle diverse aree geografiche
- *TimePeriod*: anno di riferimento
- *Value*: valore della retribuzione oraria media, in valuta locale corrente
- *Sex*: "BOTHSEX", "MALE", "FEMALE"
- *Type of occupation*: ripartizione per professione dei lavoratori dipendenti secondo due versioni della Classificazione Internazionale standard delle professioni (ISCO), 1988 e 2008
- *Units*: valuta locale corrente

La feature che merita attenzione è quella relativa al tipo di occupazione dei dipendenti, in particolare alle due versioni della ISCO degli anni 1988 e 2008. A titolo informativo, l'ultima revisione non è altro che un esplicito chiarimento dei contenuti rispetto alla precedente, mentre resta invariato il modello concettuale alla base.

A tal proposito, la suddivisione ISCO-88 è risultata obsoleta in alcune aree, soprattutto a causa dell'impatto degli sviluppi tecnologici nell'informatica. Alcune categorie di occupazione, infatti, sono state fuse o spostate per rispondere al cambiamento tecnologico avvenuto nel mercato del lavoro.

Al contrario, l'ISCO-08 non è entrata in vigore in tutti quei territori in cui lo sviluppo tecnologico è poco evidente. Per tale ragione, ancora oggi molte aree poco sviluppate adoperano lo standard di vecchio stampo come supporto per i loro calcoli.

Appurato ciò, è chiaro come nel dataset sia presente questo lieve divario, ma per evitare la perdita di dati di diversi Paesi, è stato considerato lo standard ISCO-08 anche per quei territori che, invece, utilizzavano la versione precedente.

Al termine delle trasformazioni i campi restanti sono i seguenti: *GeoAreaCode, AreaName, TimePeriod, Value, Sex, Type of occupation*. Nella Fig. 30 viene mostrata la seguenza di operazioni di ETL applicate.

Figura 30: Pentaho - Retribuzione oraria media

6.1 Combinazione datasets - Multiway Merge Join

Le operazioni di preparazione dei dataset precedentemente descritti sono state azioni preliminari al fine di predisporre i dati forniti in ingresso all'operatore *Multiway Merge Join* il quale permette di riunire, sotto specifiche chiavi primarie, i vari dataset in un unica tabella finale. A tal proposito, anche in questo caso, sono state utilizzate come chiavi il codice associato all'area geografica, *GeoAreaCode*, e l'anno di riferimento e l'anno di riferimento, *Time Period*. Inoltre, la scelta del tipo di join è ricaduta, come nella prima parte dell'elaborato, sul *full outer join*, in modo tale da non avere perdita di informazione e, laddove possibile, procedere con tecniche di imputazione per rimpiazzare eventuali valori mancanti.

Infine, la vista completa della trasformazione totale risultante, descritta dall'inizio di questo capitolo, è riportata in Fig. 31.

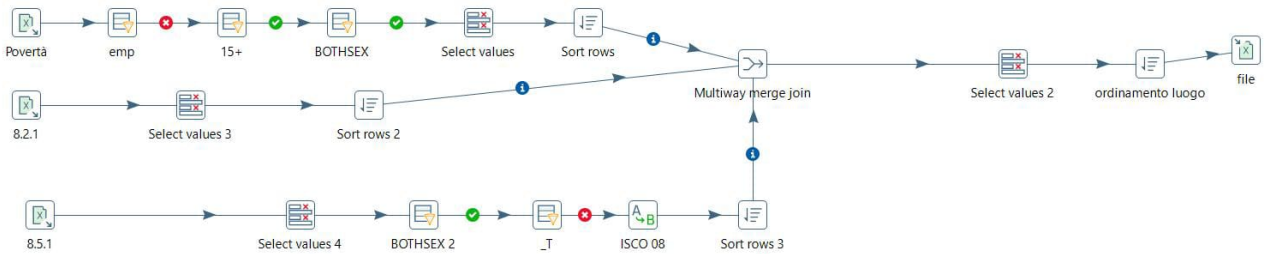

Figura 31: Pentaho - Multiway Merge Join

Ad ogni modo, il risultato finale è stato esportato sotto forma di file Excel allo scopo di esplorare, analizzare e visualizzare i dati mediante l'ecosistema Python, ottenuti dall'operazione di integrazione.

7 Esplorazione dei dati

La prima azione è sicuramente la visualizzazione della tabella finale ottenuta dalle operazioni di filtraggio e integrazione, la quale risulta essere composta dai seguenti attributi: Area Code, Luogo, Anno, % lavoratori poveri +15Y, % GDP a persona, retribuzione oraria media, tipo di occupazione, genere.

In particolare, a titolo d'esempio per una migliore comprensione, si prende in considerazione l'Afghanistan come territorio nel solo anno 2014, come riportato nella Fig. 32.

Area Code	Luogo	Anno	% poveri +15Y	% GDP a persona	retribuzione oraria media	tipo di occupazione	Genere
4	Afghanistan	2014	44.8	-2.2	67.62	Armed forces occupations	BOTHSEX
4	Afghanistan	2014	44.8	-2.2	93.61	Technicians and associate professionals	BOTHSEX
4	Afghanistan	2014	44.8	-2.2	37.95	Skilled agricultural, forestry and fishery wor...	BOTHSEX
4	Afghanistan	2014	44.8	-2.2	36.22	Craft and related trades workers	BOTHSEX
4	Afghanistan	2014	44.8	-2.2	83.86	Professionals	BOTHSEX
4	Afghanistan	2014	44.8	-2.2	59.29	Plant and machine operators, and assemblers	BOTHSEX
4	Afghanistan	2014	44.8	-2.2	57.37	Service and sales workers	BOTHSEX
4	Afghanistan	2014	44.8	-2.2	37.07	Elementary occupations	BOTHSEX
4	Afghanistan	2014	44.8	-2.2	62.28	Not elsewhere classified	BOTHSEX
4	Afghanistan	2014	44.8	-2.2	82.04	Clerical support workers	BOTHSEX
4	Afghanistan	2014	44.8	-2.2	102.93	Managers	BOTHSEX

Figura 32: Python - Afghanistan anno 2014

Come si può osservare, c'è una sostanziale ridondanza di dati per gli indicatori che riguardano la percentuale dei lavoratori poveri nonchè la percentuale del PIL, mentre i valori sulle retribuzioni orarie medie dei lavoratori si differenziano sulla base del tipo di occupazione.

Una nota di rilievo del dataframe è la mancanza di valori nulli e una lunghezza pari a 4.262 righe; ciò fa ben sperare una buona riuscita dell'analisi dati e, quindi, di ricavare informazioni per scoprire e analizzare schemi nascosti, relazioni e tendenze degli indicatori presi di riferimento.

Proseguendo con l'esplorazione dei dati, il passo successivo si è orientato verso l'ispezione del range di anni disponibili lungo il quale si distribuiscono i dati. Ciò che emerge è l'insufficienza di valori per il primo decennio e l'incremento dei Paesi aderenti al programma d'azione *Agenda 2030* negli anni successivi al primo decennio con l'obiettivo di condividere l'impegno a garantire un presente e un futuro migliore al Pianeta e alle persone che lo abitano.

Tuttavia, per gli anni 2020 e 2021 la disponibilità di dati è cominciata a retrocedere arrivando addirittura a contare un numero di righe pari a soli 29 per l'anno 2021. Molto probabilmente la causa è da attribuire all'improvvisa pandemia da COVID-19 che ha coinvolto tutte le nazioni, le quali sono state costrette a fronteggiare trascurando, così, tutti gli sforzi sostenuti per arrivare a definire una propria strategia di sviluppo sostenibile che consentisse di raggiungere gli SDGs e a rendicontare i propri risultati.

Per le ragioni discusse finora e, quindi, conseguente impossibilità di effettuare confronti data la scarsità di dati, si è deciso di eliminare tutti quei dati che fanno riferimento alla prima decade (2000-2010) e di porre l'attenzione soltanto sulla seconda decade (2011-2021). Inoltre, per le stesse ragioni, sono state rimosse le righe che riguardano l'anno 2021 (Fig. 33).

A seguito del filtraggio del range di anni, il passo consecutivo è stato quello di effettuare un conteggio sui tipi di occupazione. Ciò che si può notare è una disparità, seppur leggera nella maggior parte dei casi, tra i vari impieghi lavorativi. Per cui quello che si riesce a dedurre è che ci sia stata una mancanza nella raccolta delle fonti e nell'inserimento dei dati. Oppure ci sono dei territori in cui una determinata occupazione lavorativa viene svolta ma con un rilievo talmente basso da non venir preso in considerazione nelle indagini.

Al fine di venir meno alle problematiche nelle differenze dei diversi impieghi lavorativi e, dunque, favorire il confronto internazionale si è deciso di filtrare soltanto le occupazioni elementari come tipologia di impiego e proseguire l'analisi mettendoci nella condizione più generale.

Giunti a tale punto, si è interessati a visualizzare quali territori sono presenti e quanti anni ricoprono (senza conoscere esplicitamente l'arco temporale di appartenenza), in modo tale da rendersi subito conto dell'eventuale mancanza di dati della decade di anni 2011-2020 presa come riferimento e, nel caso, rimpiazzarli mediante l'utilizzo di tecniche di imputazione.

```
df_BOTH["Anno"].value_counts(sort=False)
```

```
Out[12]: 2014    338
         2020    242
         2018    343
         2004     61
         2005     92
         2006    144
         2008    158
         2009    196
         2010    198
         2011    208
         2012    253
         2013    295
         2017    389
         2019    400
         2015    290
         2016    287
         2007    132
         2000     22
         2001     62
         2021     29
         2002     53
         2003     70
         Name: Anno, dtype: int64
```

```
In [13]: #si tolgono gli anni prima del 2011 e il 2021
         df_BOTH.drop(df_BOTH[df_BOTH.Anno < 2011].index, inplace=True)
         df_BOTH.drop(df_BOTH[df_BOTH.Anno == 2021].index, inplace=True)
```

Figura 33: Python - Conteggio range di anni disponibili

Data la molteplicità di territori che non ricoprono tutti gli anni, si è deciso di porre una soglia, sotto la quale vengono escluse quelle nazioni che abbiano un numero di righe tali da non essere presi in considerazione per il successivo riempimento dei valori. Tale operazione risulta necessaria in quanto il metodo dell'imputazione stima i valori mancanti sulla base di quelli già presenti; pertanto se avesse a disposizione pochi dati, la valutazione risulterebbe poca accurata e, talvolta, priva di significato. Per tale ragione, la soglia è stata stabilita uguale a 3 e successivamente sono stati eliminati tutti quei Paesi con un un numero inferiore alle 3 righe; in altre parole i territori con un numero di anni mancanti superiore a 7, nell'intervallo temporale che va dal 2011 al 2019, sono state rimosse.

Si ottengono così dati con nazioni che coprono un arco temporale di almeno 3 anni e fino a 10 anni.

Affinché il lettore abbia ben chiaro le aree geografiche rimanenti si illustra, in Fig. 34, il nome e il numero di righe associate.

Viet Nam	10
Panama	10
Mongolia	10
Türkiye	10
Malaysia	10
Republic of Korea	10
Costa Rica	10
Peru	10
Honduras	9
Mauritius	9
Philippines	9
Paraguay	9
El Salvador	9
Egypt	9
Brazil	9
Armenia	8
Sri Lanka	8
Mexico	8
Pakistan	8
Argentina	8
Serbia	7
Indonesia	7
Cambodia	7
Bosnia and Herzegovina	7
State of Palestine	6
Dominican Republic	6
Thailand	6
Mali	6
Jordan	4
Jamaica	4
Chile	4
Myanmar	4
Belize	4
India	3
Ghana	3
Zambia	3

Figura 34: Python - Aree rimanenti in seguito al filtraggio della soglia

Appurato ciò, si è passati all'inserimento di tutte le righe mancanti con nazioni e anni associati, nonchè degli indicatori. Seppur in modalità grezza, l'aggiunta di righe mancanti, mediante l'utilizzo della funzione Pandas *append* è l'unico modo per ovviare all'indisponibilità delle informazioni allo scopo di tranne conclusioni utili per comprendere il fenomeno della povertà nel mondo.

Al termine della collocazione delle righe, la lunghezza del dataframe è passato a 360, contro i 264 prima dell'aggiunta di righe.

Una volta aver concluso la fase pre-elaborazione del dataframe, si può procedere con le tecniche di interpolazione dei valori mancanti.

L'interpolazione è una tecnica in Python utilizzata soprattutto nelle serie temporali per stimare ed imputare dati mancati tra due punti dati noti. Al contrario, il riempimento dei valori mancanti con la media dei dati conosciuti non è il metodo migliore in quanto ciò potrebbe influire sulla precisione dei dati.

Nel caso di studio dell'elaborato, si è fatto uso della funzione *interpolate()*, la quale, eseguita passando diversi parametri secondo i requisiti di interesse, restituisce lo stesso tipo di dati dell'input.

Delle tre tipologie di interpolazione la scelta è ricaduta su quella lineare la quale presuppone che il punto stimato si trovi sulla linea che unisce i punti più vicini a sinistra e a destra. In altre parole, riempie i valori mancanti con un ordine crescente tra i valori osservati precedenti e successivi. In particolare, la direzione di interpolazione in avanti e indietro viene eseguita per mezzo del parametro passato alla funzione interpolate(*limit_ direction = 'both'*). In Fig. 35 è mostrato il codice di Python per il calcolo dell'interpolazione lineare e l'assegnazione dei valori risultati nella rispettiva colonna.

Per una migliore comprensione del comportamento della funzione *interpolate()*, a titolo d'esempio, viene mostrata in Fig. 36 il territorio dell'Argentina. Gli anni mancanti sono il 2015 e il 2016 e l'interpolazione lineare è stata calcolata per i tre indicatori: la percentuale dei lavoratori poveri è uguale per entrambi gli anni, in quanto negli anni immediatamente precedenti e successivi i valori restano invariati a 0.3; al contrario, per quanto riguarda la percentuale del GDP e la retribuzione oraria media i valori inseriti corrispondono esattamente a due diversi segmenti appartenenti alla linea retta che collega la coppia di punti adiacenti degli anni 2014 e 2017. In generale, si nota un andamento crescente per quest'ultimi indicatori.

```python
#Interpolazione lineare in entrambe le direzioni (avanti e indietro)
values = ["% poveri +15Y", "% GDP a persona", "retribuzione oraria media"]

#per ogni nazione della lista
for i in range(len(Geo)):
    for k in range(0,3):
        #calcolo la media per l'indicatore in ogni nazione [i] e per tutti gli indicatori [k]
        imp=df_BOTH_sort.loc[df_BOTH_sort.Luogo==Geo[i], values[k]].interpolate(method='linear', limit_direction='both'
        #assegno il valore
        df_BOTH_sort.loc[df_BOTH_sort.Luogo==Geo[i],
                         values[k]] = df_BOTH_sort.loc[df_BOTH_sort.Luogo==Geo[i],
                                                       values[k]].fillna(imp)
        k+=1
    i+=1
```

Figura 35: Python - Interpolazione lineare valori mancanti

Area_code	Luogo	Anno	% poveri +15Y	% GDP a persona	retribuzione oraria media	occupazione	Genere
32	Argentina	2011	0.400000	3.200000	12.0900	Elementary occupations	BOTHSEX
32	Argentina	2012	0.400000	-1.800000	16.1800	Elementary occupations	BOTHSEX
32	Argentina	2013	0.300000	1.700000	20.4400	Elementary occupations	BOTHSEX
32	Argentina	2014	0.300000	-2.700000	27.4000	Elementary occupations	BOTHSEX
32	Argentina	2015	0.300000	-1.066667	40.6800	Elementary occupations	BOTHSEX
32	Argentina	2016	0.300000	0.566667	53.9600	Elementary occupations	BOTHSEX
32	Argentina	2017	0.300000	2.200000	67.2400	Elementary occupations	BOTHSEX
32	Argentina	2018	0.400000	-4.100000	82.0000	Elementary occupations	BOTHSEX
32	Argentina	2019	0.500000	-3.700000	110.6100	Elementary occupations	BOTHSEX
32	Argentina	2020	0.600000	-2.400000	169.5300	Elementary occupations	BOTHSEX

Figura 36: Python - Interpolazione lineare Argentina

Inoltre, per avere un riepilogo delle statistiche del dataset si applica il metodo *describe()*. Tale funzione, infatti, evidenzia la distribuzione delle variabili numeriche presenti all'interno del dataframe al fine di acquisire maggiore consapevolezza del fenomeno in questione (Fig. 37).

	Area_code	Anno	% poveri +15Y	% GDP a persona	retribuzione oraria media
count	360.000000	360.000000	360.000000	360.000000	360.000000
mean	397.305556	2015.500000	5.554722	1.940833	1514.331667
std	242.974238	2.876279	11.922172	3.382682	3899.250485
min	32.000000	2011.000000	0.000000	-11.500000	1.050000
25%	179.000000	2013.000000	0.100000	0.200000	7.192500
50%	394.000000	2015.500000	0.900000	2.000000	46.520000
75%	593.250000	2018.000000	4.800000	4.000000	496.297500
max	894.000000	2020.000000	53.900000	11.400000	28058.320000

Figura 37: Python - applicazione della funzione *describe()*

Ciò che emerge è la chiara disuguaglianza nella ripartizione della retribuzione oraria media per l'impiego lavorativo che riguardano le occupazioni elementari. Infatti, i dati presentano un minimo di 1 a un massimo che arriva addirittura a 28.058,32, il che sembrerebbe alquanto elevato come salario orario. Probabilmente, l'evidente disparità è dovuta all'utilizzo delle rispettive valute locali come misura di calcolo per la paga oraria. Pertanto, non disponendo di una valuta standard mondiale, si è deciso di escludere tale indicatore nello studio di analisi successive, in quanto avrebbe generato soltanto confusione e non avrebbe, di fatto, portato nessuna conoscenza.

8 Visualizzazioni Mappe

Una volta terminato l'elaborazione e la modellazione del dataset, il passo successivo è la *data visualization* al fine di poter tradurre le informazioni in un contesto visivo per rendere i dati più facili da comprendere e da cui trarre conclusioni.

Tuttavia, nelle sezioni precedenti, nonostante l'applicazione di tecniche di imputazione per rimpiazzare valori mancanti, ci si è resi subito conto come i territori che contenessero informazioni fossero un numero ridotto in relazione all'insieme dei Paesi mondiali. Per cui svolgere un'accurata analisi descrittiva e, dunque, mostrare l'evoluzione degli obiettivi prefissati nell'Agenda 2030 sembrerebbe alquanto complicata con un numero così esiguo di informazioni. Nonostante tali complicanze, è stato comunque provato a visualizzare, tramite una mappa, i dati dei territori di cui si hanno notificazione e a trasformarli in valore da poter essere utilizzati in sviluppi futuri.

Innanzitutto, si ricordano come i territori rimanenti, secondo la macro-suddivisione effettuata dalle nazioni unite (Fig. 21), comprendono gran parte dell'America Latina e del sud e sud-est asiatico; l'Europa e il Nord America sono esclusi da questi valori, ad eccezione della Serbia e Bosnia and Herzegovina così come l'Africa ad eccezione dell'Egitto, Mali, Zambia e Ghana.

L'obiettivo di tale sezione è quello di visualizzare 3 mappe dinamiche, ciascuna delle quali ritrae l'andamento degli indicatori scelti nelle sezioni precedenti nel range temporale che parte dal 2011 al 2020. In particolare, per una migliore visione della situazione in termini di obiettivi adottati dall'Agenda 2030 e poter, dunque, effettuare una sorta di confronto temporale, è stato deciso di mostrare le mappe degli anni 2011, 2016 e 2020.

8.1 Mappa - % lavoratori in condizioni di povertà

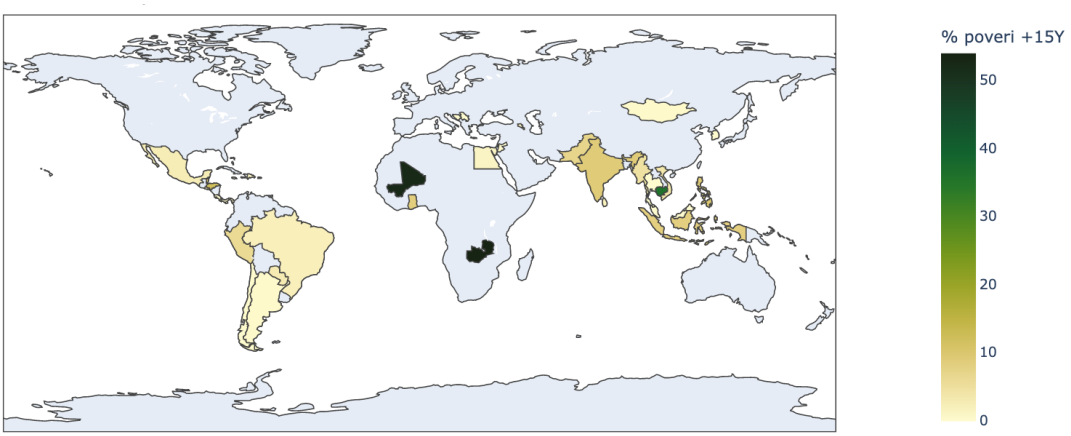

Figura 38: Mappa - lavoratori che vivono sotto la soglia di povertà anno 2011

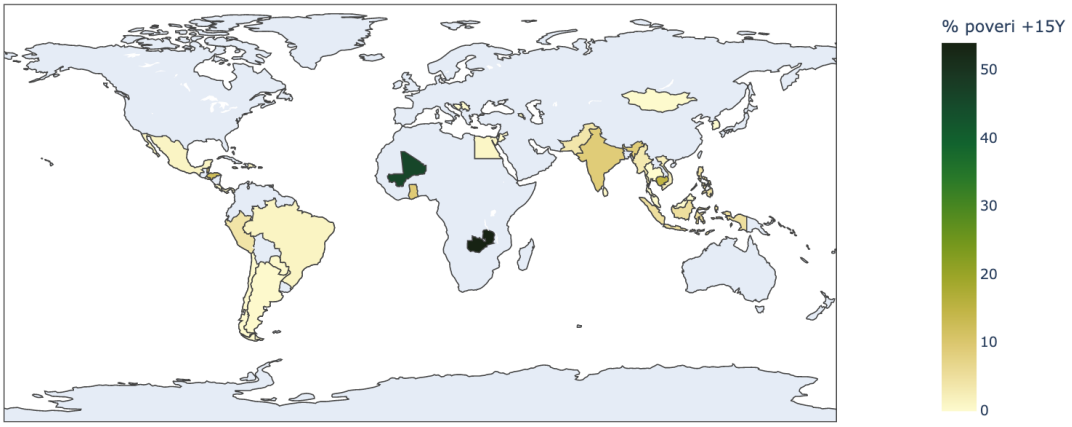

Figura 39: Mappa - lavoratori che vivono sotto la soglia di povertà anno 2016

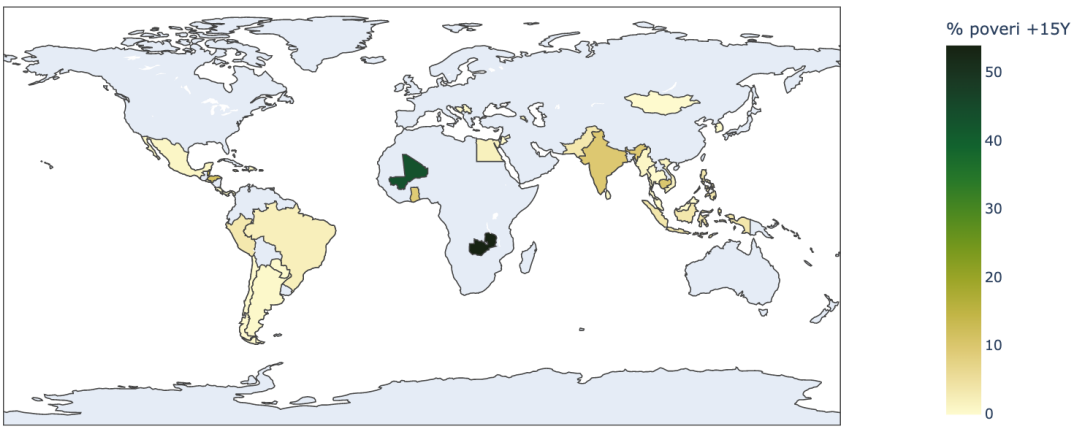

Figura 40: Mappa - lavoratori che vivono sotto la soglia di povertà anno 2020

In generale, dall'osservazione delle mappe nei tre anni presi di riferimento, non sembrano esserci sostanziali differenze tali da potersi ritenere soddisfatti del raggiungimento dei piani legati all'Agenda 2030. Tuttavia, un territorio del sud-est asiatico degno di nota è la *Cambogia*, conosciuto come uno dei Paesi più poveri al mondo e con un passato drammatico da cui tenta faticosamente di uscire. I grafici, al contrario, mostrano come tale paese asiatico abbia avuto una notevole crescita economica negli ultimi anni; questo ha fatto sì che nella decade 2011-2020 la percentuale dei lavoratori che vive al di sotto della soglia di povertà sia passata dal 35,9% al 10,1%.

Nonostante il miglioramento evidente, sono ancor diverse le sfide che la Cambogia deve ancora affrontare: quasi 3 milioni di persone soffrono di malnutrizione ed il tasso di alfabetizzazione è tra i più bassi del Sud Est asiatico, seppur con dati in crescita.

Ad ogni modo, uno dei maggiori problemi che causa il rallentamento dello sviluppo economico dell'intero Paese resta la corruzione; non va meglio sul fronte complessivo dei diritti umani e sociali.

Inoltre l'Africa, seppur i pochi territori di cui si ha informazione, resta il paese con la più alta percentuale di lavoratori che vivono sotto la soglia di povertà assoluta; il territorio della *Zambia* conta una percentuale del 53,8 nell'anno 2020, restata invariata negli anni.

Le ragioni alla base della povertà di molti paesi africani sono molte e complesse: colonialismo, processo di decolonizzazione, cause demografiche e climatiche, hanno sempre ostacolato lo sviluppo naturale delle società africane e spesso retrocedere i processi produttivi.

In linea definitiva, l'analisi ha evidenziato una situazione favorevole per l'America latina e gran parte dell'asia sud est che si distinguono per avere una percentuale di lavoratori poveri inferiore al 4%.

8.2 Mappa - % PIL a persona

Analogamente al precedente indicatore illustrato, di seguito vengono mostrate le mappe che rappresentano la percentuale del PIl a persona rispettivamente per l'anno 2011, 2016 e 2020.

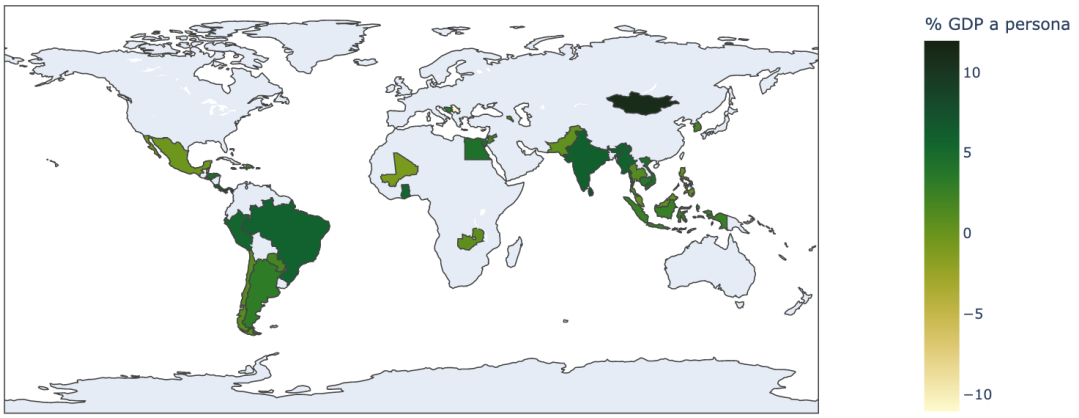

Figura 41: Mappa - PIL a persona anno 2011

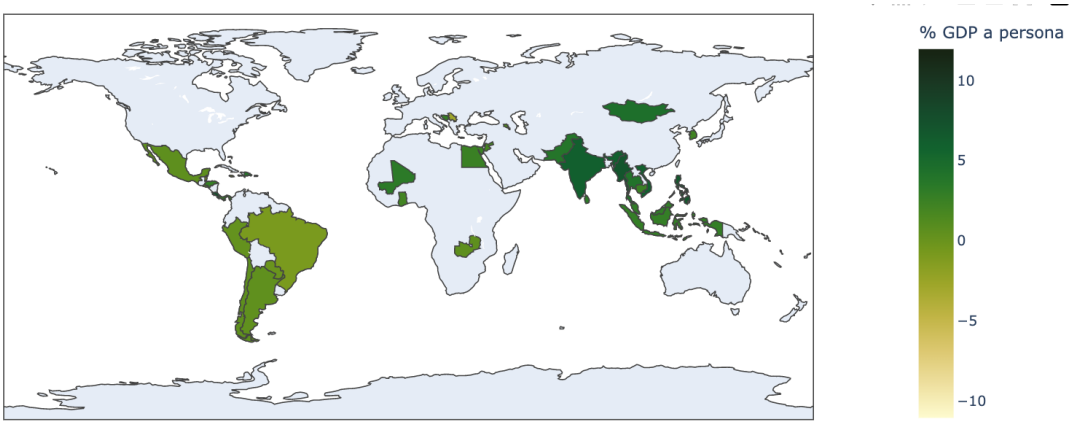

Figura 42: Mappa - PIL a persona anno 2016

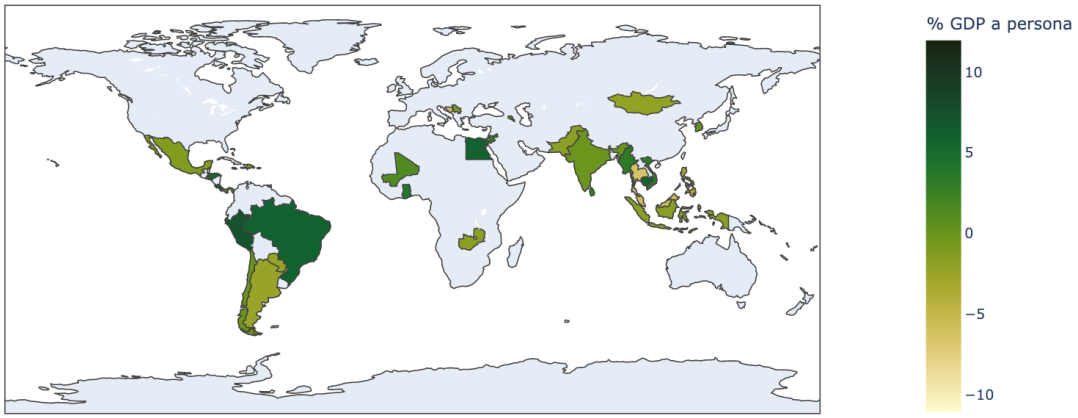

Figura 43: Mappa - PIL a persona anno 2020

Paradossalmente a quanto ci si aspetterebbe, le prospettive della crescita mondiale in linea generale sono pressochè insoddisfacenti.

In America Latina, dapprima si verifica un tasso di crescita del +5,8% e del +5,4% rispettivamente in Brasile e in Perù per poi avvisare un rallentamento già a partire dagli anni 2015-2016 toccando punte dello +0,3% e, in alcuni territori come il Brasile addirittura valori negativi del -0,8%, segnando il peggior risultato da quando è iniziata la raccolta dati. Il Brasile, infatti, si è ritrovare a fronteggiare il calo dei prezzi delle materie prime, una contrazione degli investimenti e tagli alla spesa, oltre che un sistema politico bloccato da numerosi casi di corruzione ai massimi livelli.

La conferma della fragilità di quello che era stato considerato uno dei Paesi emergenti dalle prospettive più brillanti è arrivata, dunque, con l'imminente declino dello sviluppo economico anche negli anni successivi.

Nell'anno 2020, come emerge chiaramente dalla mappa, si riscontra un aumento del tasso di crescita del 5,8%. Stando a delle ricerche, si tratterebbe probabilmente di un errore in fase di raccolta dati, in quanto, a causa della pandemia da COVID-19, ha rilevato un ulteriore rallentamento nello sviluppo economico che non va, di fatto, trascurato.

Un Paese degno di nota è rappresentato dalla Mongolia, il quale ha registrato un boom economico a partire dagli anni 2000. L'economia della Mongolia, che in passato si basava esclusivamente sulla pastorizia, a partire dalla seconda decade degli anni 2000, ha attirato l'attenzione della comunità internazionale grazie al suo tasso di crescita esponenziale e ad un nuovo attivismo internazionale. Tale Paese, infatti, gode di grandi riserve di carbone, oro e rame che hanno attratto importanti compagnie internazionali dell'industria estrattiva e, dunque, prospera grazie alla domanda di materie prime richieste dai paesi esteri, soprattutto da parte della Cina.

Tuttavia, l'incidenza dell'indicatore sul PIL è variata nel corso degli anni: da +11,1% del 2011 a +1,7% del 2017. Questo drastico calo è dipeso dal rallentamento dell'economia cinese (paese verso cui è diretta la maggior parte delle esportazioni mongole) e dalle deboli condizioni generali del mercato mondiale. Anche se negli anni successivi la produzione mineraria ha gradualmente ripreso la propria crescita, nel 2020, a causa della pandemia da COVID-19, ha purtroppo assistito a un drastico declino del -2,1%.

In generale, le considerazioni fatte per l'indicatore del PIL reale per occupato risulta in perfetta linea con quanto affermato per la percentuale di lavoratori che vivono sotto la soglia di povertà assoluta. Infatti, laddove si verifica un'alta presenza di lavoratori poveri, il PIL è estremamente basso (o addirittura in valori negativi); al contrario, con una percentuale di lavoratori poveri che si aggira intorno allo 0% si registra un alto tasso di crescita economico.

9 Discussione dei risultati ottenuti

Nella presente sezione viene effettuata una discussione dei risultati ottenuti, nonché delle problematiche riscontrate; a partire da queste ultime, in seguito, si è cercato di simulare ipotesi sui possibili miglioramenti da apportare al fine di migliorare l'accessibilità dei dati come principale risorsa in grado di assumere un valore reale. Pertanto, si concluderà sottolineando come la conoscenza estrapolata dalle esperienze possa riuscire ad aprire nuovi scenari di supporto per il raggiungimento degli obiettivi prefissati nell'Agenda 2030.

Il tentativo di analisi, ampiamente illustrato nelle sezioni precedenti, mostra l'evidente presenza di molteplici lacune già a monte nella fase di raccolta dei dati. Tali limiti, infatti, delineano un ostacolo significativo per una ricerca accurata degli indicatori SGD, il che ha portato al nostro studio una rappresentazione distorta dei progressi compiuti verso gli obiettivi prefissati. Non solo, ha ostacolato ulteriormente il monitoraggio concernente il raggiungimento dello sviluppo sostenibile indicato nell'Agenda 2030.

Diverse sono le ragioni per cui è possibile affermare la difficoltà nell'osservazione e nel costante controllo degli obiettivi a livello mondiale allo scopo di un futuro migliore. A partire dallo studio di analisi condotto nel presente elaborato, la scarsa disponibilità di dati è l'aspetto principale di cui non si può fare a meno di richiamare l'attenzione. Di fatto, nella fase di esplorazione dei dati, si è constatato come la serie temporale presentasse diversi "buchi", sia in termini di range annuali che di rappresentazioni intrinseche dei fenomeni degli indicatori per interi Paesi, confermando la realtà che le Nazioni Unite si limitano a indicare pochi dati spesso non sempre allineati temporalmente sullo stesso anno, quindi, che i dati non vengano raccolti in modo sistematico. Alcune aree geografiche non vengono neanche rappresentate nei dati disponibili e altre addirittura, pur volendo, non dispongono di sistemi reportistici in grado di raccogliere tutte le informazioni necessarie. Una delle maggiori cause è sicuramente da attribuire alla mancanza di finanziamenti e risorse: molti paesi non possiedono fondi sufficienti per implementare i sistemi di monitoraggio o non hanno le risorse umane a disposizione per gestire la raccolta dei dati.

Dunque, risulta chiaro come l'indisponibilità di informazione si traduce in una complicata comprensione dei ritmi dei progressi verso la realizzazione dell'Agenda 2030. Da ciò segue la dimostrazione che la raccolta e l'analisi dei dati diventano passaggi fondamentali del processo di realizzazione dell'agenda internazionale.

A tal proposito, come misura da noi adottata, l'utilizzo di tecniche di imputazione per rimpiazzare la mancanza di dati può essere una soluzione temporanea per ridurre l'impatto delle lacune nella raccolta dati, ma queste tecniche non sono in grado di compensare completamente la mancanza di dati, anzi potrebbero introdurre errori nella valutazione degli indicatori.

Una seconda ragione da attribuire alla scarsa riuscita dei risultati si riversa in una serie di complicazioni in termini di comparabilità internazionale, la maggior parte delle quali deriva dalla varietà delle possibili fonti di dati. Ciò significa che i dati raccolti da diverse fonti potrebbero essere discordanti o persino contrastanti, rendendo difficile valutare con precisione i progressi compiuti. Le varie fonti disponibili, infatti, differiscono per l'adozione di definizioni e metodi di calcolo differenti, obiettivi e portata, il che influenzano significativamente i risultati che vi si ottengono. Ancora, la coperture delle fonti può variare in termini di aree geografiche interessate, basti pensare all'utilizzo della propria valuta locale per la misurazione dei fenomeni riguardanti il settore economico o semplicemente alle differenze nei tipi di indagini adottati nella raccolta dei dati. In particolare, nel nostro caso specifico, non si può affermare categoricamente che due persone in due Paesi diversi, che vivono al di sotto di 1,90 $ al giorno, si trovino ad affrontare lo stesso grado di privazione o di bisogno.

In realtà, la scarsa comparabilità tra i vari Paesi risulta pressochè ragionevole se si considerano le tematiche coperte dagli indicatori a livello mondiale, dove ciascuna nazione possiede le proprie direttive politiche, sociali ed economiche. Inoltre, gli SDG coprono una così vasta gamma di obiettivi e discipline diverse, i quali spesso non hanno neanche sistemi di raccolta dati comuni.

Per le stesse ragioni appare evidente come tutto ciò ha evidenziato difficoltà nell'integrazione dati dei dataset dei diversi indicatori scelti al fine di realizzare un'immagine del paese sulle diverse prospettive trattate.

I dataset utilizzati, scaricati dal sito delle Nazioni Unite, coprono fino all'anno 2021 ma dall'analisi traspare perfettamente il declino della disponibilità di dati tant'è che nella fase di ETL abbiamo dovuto

escludere l'anno 2021 per la scarsità di informazioni. La principale causa è sicuramente da attribuire all'insorgere dell'emergenza pandemica da COVID-19 che tutti i Paesi si sono improvvisamente ritrovati a fronteggiare, motivo per cui sono state riassegnare risorse statistiche di intere nazioni togliendo perciò fondi e personale allo studio degli SDG ed, in generale, alla brusca interruzione della raccolta dati. In aggiunta, le restrizioni sulla mobilità hanno peggiorato ulteriormente gli sforzi statistici per valutare i Goal dell'Agenda 2030.

Ad ogni modo, se da un lato l'esplosione della pandemia ha purtroppo cancellato anni di lenti progressi sconvolgendo la vita come la conosciamo, dall'altra ha rappresentato un campanello d'allarme sulla necessità di rafforzare le basi statistiche internazionali al fine di fornire dati migliori. E' stata, inoltre, anche un'opportunità per sperimentare metodi innovativi di raccolta dei dati, esplorare nuovi fonti di dati e modernizzare le infrastrutture ICT per soddisfare le richieste di dati per la definizione delle politiche. Nel corso di questo processo, infatti, l'importanza di dati pienamente inclusivi è stata resa evidente. In tal senso, è stata evidenziata l'interesse verso la raccolta di dati affidabili e tempestivi per monitorare e gestire situazioni di crisi.

Non solo, la pandemia ha anche aumentato la consapevolezza dell'importanza dei dati disaggregati per garantire che i gruppi più vulnerabili siano adeguatamente rappresentati nelle statistiche e che, dunque, nessuno venga lasciato indietro. Ad esempio, i dati sulla pandemia hanno evidenziato le disuguaglianze di genere, economiche e razziali nella diffusione e nei risultati della malattia. Ciò ha portato a una maggiore attenzione alla raccolta di dati disaggregati e alla necessità di includere i gruppi vulnerabili nella misurazione degli indicatori SDG.

Per superare tutte queste sfide, dunque, potrebbe essere necessario la standardizzazione dei dati, ovvero la definizione di standard e protocolli di rilevamento uniformi per la raccolta dati, nonché la promozione della cooperazione tra paesi e organizzazioni per condividere dati e conoscenze.

L'applicazione di tale approccio integrato, inoltre, potrebbe essere un passo fondamentale per garantire la comparabilità e la coerenza dei dati raccolti da diverse fonti. Ciò significherebbe avere dati coerenti in termini di definizioni, metodi di raccolta, calcolo e presentazione, al fine di disporre dati che siano confrontabili tra loro e possano essere utilizzati per monitorare il progresso verso i singoli obiettivi.

Inoltre, l'aggiunta dell'uso di tecniche avanzate di analisi dei dati, come l'apprendimento automatico e l'intelligenza artificiale, potrebbe essere un aiuto per integrare e analizzare i dati in modo più efficace e accurato.

Alla luce di ciò, per concludere si può confermare che l'analisi effettuata nel corso del progetto ha mostrato sia segnali incoraggianti che aspetti critici, ma sono insufficienti per delineare un quadro sull'evoluzione verso la realizzazione dei 17 obiettivi sostenibili che necessitano, invece, di analisi più approfondite.

Dunque, dati tempestivi, disaggregati e di alta qualità sono più importanti che mai. Sono necessari ulteriori investimenti nelle infrastrutture di dati e informazioni, sulla base delle lezioni apprese durante la pandemia. L'obiettivo è anticipare la crisi in modo da poter attivare risposte più tempestive, anticipare i bisogni futuri e progettare le azioni urgenti necessarie per realizzare l'Agenda 2030 per lo sviluppo sostenibile.