Analisi esplorativa delle nuove infezioni da HIV su dataset WHO

Tesina di Data Science for Health Systems

Tommaso Martinelli
Università degli Studi di Perugia
tommaso.martinelli@studenti.unipg.it

***Abstract*—Questo documento presenta i risultati dell'analisi esplorativa condotta su un dataset fornito dall'Organizzazione Mondiale della Sanità, relativo alle nuove infezioni da HIV. L'obiettivo dell'analisi, condotta utilizzando il linguaggio di programmazione R, è stato quello di indagare le differenze nelle nuove infezioni da HIV in base al sesso e all'area geografica, nonché di analizzare l'andamento nel corso degli anni (dal 1990 al 2022). Tramite strumenti grafici e test statistici è stato possibile notare differenze rilevanti tra i sessi e una distribuzione dei nuovi casi non uniforme nel mondo.**

***Index Terms*—EDA, WHO, HIV**

I. Introduzione

Il virus dell'imunodeficienza umana (HIV) è un'infezione che attacca il sistema immunitario dell'organismo, colpendo i globuli bianchi, in particolare i linfociti CD4, responsabili della risposta immunitaria [1]. L'HIV è un virus a RNA che appartiene alla famiglia dei retrovirus [2]. La diffusione avviene attraverso i fluidi corporei di una persona infetta, tra cui sangue, latte materno, sperma e fluidi vaginali; lo stadio più avanzato della malattia è la sindrome da immunodeficienza acquisita (AIDS) [3]. L'HIV è uno dei più grandi problemi di salute pubblica a livello globale, con una trasmissione in corso in tutte le aree del mondo. Lo studio si propone di analizzare l'andamento delle infezioni nel periodo compreso tra il 1990 e il 2022, esaminando le differenze di incidenza tra i sessi, le diverse aree geografiche e i singoli paesi.

II. Dataset

Il dataset utilizzato proviene dal sito dell'Organizzazione Mondiale della Sanità (World Health Organization) ed è chiamato "New HIV infections (per 1000 uninfected population)" o "HIV incidence rate" [4]. Contiene dati riguardo all'incidenza dell'HIV, cioè il numero di persone positive ogni 1000 individui sani. La raccolta di dati è avvenuta mediante dati longitudinali, raramente disponibili, o test diagnostici e indagini sulla popolazione o in strutture sanitarie; i risultati sono poi stati resi significativi a livello nazionale tramite modellazione effettuata con Spectrum, software supportato da UNAIDS. Il dataset originario è composto da 19206 campioni e 32 features, molte delle quali però ritenute ridondanti o inutili ai fini dell'analisi. Ciò è dovuto alla presenza di dati NaN oppure a informazioni di scarso valore, come codici identificativi o date di modifica del file.

III. Modellazione del dataset

Prima di iniziare l'analisi esplorativa sono state necessarie alcune operazioni sui dati, al fine di preparare il dataset in maniera ottimale. Le colonne contenenti esclusivamente dati Nan sono state rimosse, così come altre colonne che presentavano valori identici per ogni riga, come unità di misura, identificativi e descrizioni del dataset. Per l'Analisi Esplorativa dei Dati (EDA), sono state selezionate solo le colonne più significative, le quali sono state rinominate utilizzando nomi più esplicativi rispetto a quelli originali. Il dataset risultante da tali operazioni è composto dalle seguenti colonne:

- **geographic_area**: rappresenta il nome dell'area geografica di appartenenza secondo la suddivisione dell'Organizzazione Mondiale della Sanità.
- **state**: nazione del mondo a cui si riferisce il dato.
- **year**: anno in cui è stata effettuata la misurazione.
- **sex**: può assumere tre valori: male, female e bothsexes.
- **value**: numero di nuovi casi ogni 1000 individui sani.
- **lower_confidence_interval**: estremo inferiore dell'intervallo di confidenza.
- **upper_confidence_interval**: estremo superiore dell'intervallo di confidenza.

In seguito alla selezione delle features sono state necessarie delle ulteriori operazioni di preprocessamento, è stato infatti rimosso il valore "both sex" dalla variabile sex, per permettere una più semplice analisi delle differenze tra i due sessi; inoltre sono stati rimossi i campioni che avevano value Nan, poichè considerati inutili.

IV. Analisi esplorativa

L'analisi esplorativa dei dati (EDA) è una fase fondamentale nell'elaborazione e interpretazione delle informazioni contenute in un dataset. In questa sezione, verranno presentate in dettaglio le principali scoperte ottenute dall'EDA sul dataset relativo alle nuove infezioni da HIV. Questo processo di indagine si è articolato attraverso diverse angolazioni, con particolare attenzione alle differenze di genere, alle fluttuazioni temporali e alle variazioni geografiche.

A. Distribuzione dei valori

Per dare il via a questa esplorazione, inizialmente è stata esaminata la distribuzione dei valori delle nuove infezioni

da HIV, considerando l'intero spettro dei campioni senza applicare alcun filtro, al fine di trarre un quadro globale della distribuzione dei dati. L'istogramma rappresentato nella Figura 1 illustra questa distribuzione.

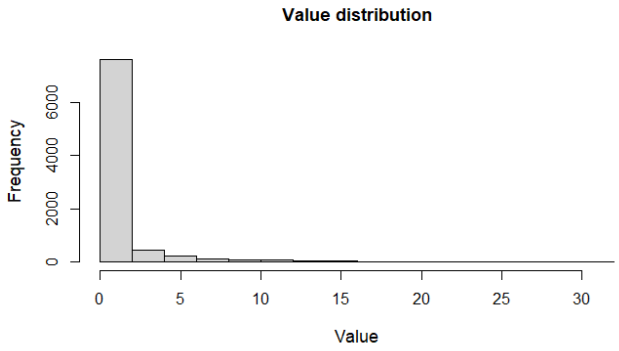

Fig. 1. Istogramma dei numeri di nuovi infetti

Si può notare che la maggior parte dei campioni presenta valori compresi tra 0 e 5 nuovi infetti ogni 1000 individui sani. Al contrario, si osservano con minor frequenza valori superiori a 5. Questa prima analisi ha fornito un'idea iniziale della prevalenza delle infezioni nel dataset.

B. Differenze tra i sessi

L'analisi è proseguita concentrandosi sulle differenze di genere. L'obiettivo di questa fase è stato quello di valutare se le distribuzioni di nuovi casi fossero simili per i due sessi. Sono state utilizzate diverse tipologie di grafici, poichè alcuni potrebbero rivelare in modo più efficiente eventuali differenze. Nei boxplot, in Figura 2, è evidente come le distribuzioni per i due sessi siano molto simili e come osservato prima, si concentrino particolarmente intorno a valori in prossimità dello 0.

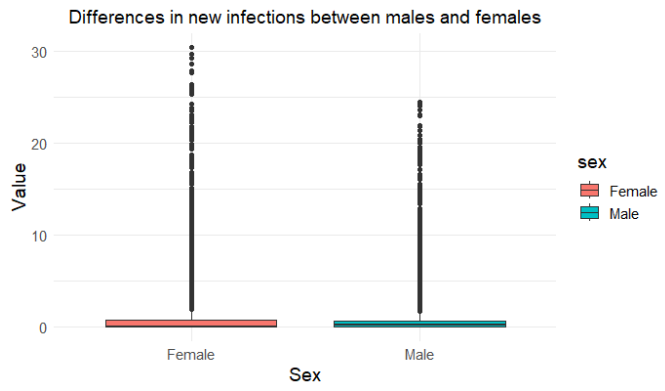

Fig. 2. Box-plot delle distribuzioni dei valori di maschi e femmine

Nonostante le distribuzioni siano simili, si è osservato che i campioni di sesso femminile presentano alcuni valori massimi più elevati rispetto a quelli di sesso maschile. Questa tendenza emerge chiaramente anche dal grafico a dispersione rappresentato nella Figura 3, il quale mette in evidenza come i valori più elevati nel dataset siano principalmente associati al sesso femminile.

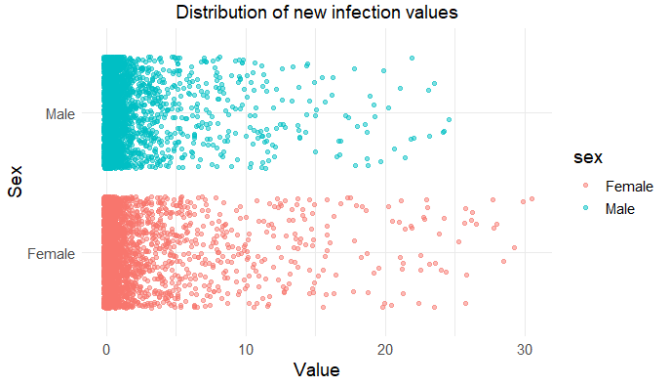

Fig. 3. Distribuzione dei nuovi valori di infezione per genere

Queste osservazioni indicano che, nonostante le somiglianze nelle distribuzioni complessive, vi sono alcune differenze nelle frequenze di valori più elevati tra i due sessi. La presenza di valori massimi superiori nel sesso femminile potrebbe essere oggetto di ulteriori approfondimenti.

C. Variazioni nel tempo

Per comprendere l'evoluzione delle nuove infezioni da HIV nel corso degli anni, è stata effettuata un'analisi temporale. Il dataset suddivide i campioni per anno, in un intervallo di tempo che va dal 1990 al 2022, per ogni anno è stata calcolata una media di nuove infezioni da HIV ogni 1000 individui sani. Per individuare eventuali tendenze in questi anni sono stati utilizzati grafici, come il grafico a barre in Figura 4.

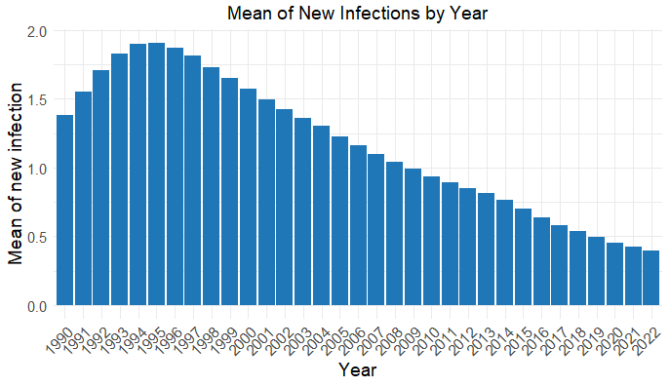

Fig. 4. Grafico a barre della variazione della media di nuove infezioni negli anni

Si può notare un chiaro andamento nelle medie delle nuove infezioni, con un aumento costante nei primi anni, fino al 1995, per poi scendere ogni anno fino al 2022. Nel dataset utilizzato non sono presenti dati ulteriori per capire le cause di questo andamento, un possibile approfondimento futuro sulle politiche e trattamenti potrebbe spiegare meglio questo trend.

Successivamente è stata analizzata la variazione di nuove infezioni nel tempo, per i due sessi separatamente, tramite l'utilizzo di un grafico a linee, riportato in Figura 5.

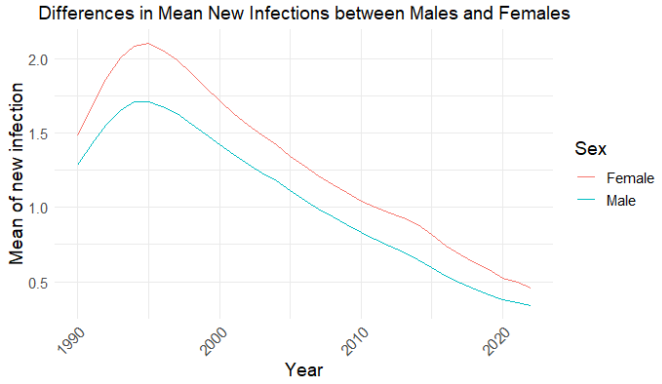

Fig. 5. Grafico a linee della variazione delle medie di nuove infezioni tra maschi e femmine durante gli anni

Il plot evidenzia come le medie di nuove infezioni tra maschi e femmine seguano un andamento simile nel tempo. È interessante notare che la media delle infezioni femminili risulta costantemente superiore a quella maschile, confermando quanto osservato in precedenza.

D. Analisi delle aree geografiche

Dopo l'analisi temporale, si sono andate a ricercare informazioni riguardo la distribuzione nel mondo delle nuove infezioni da HIV, allo scopo di individuare eventuali aree con maggiore incidenza. Ogni campione del dataset si riferisce ad una specifica nazione, inoltre ogni stato appartiene ad una delle 6 regioni con cui l'Organizzazione Mondiale della Sanità suddivide il mondo. In primo luogo quindi sono state calcolate le medie per ogni area geografica, mostrate in Figura 6.

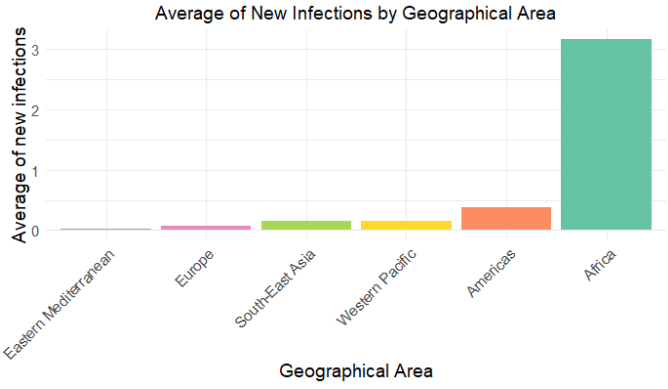

Fig. 6. Grafico a barre delle medie di nuovi infetti nelle varie aree geografiche

Il bar-plot evidenzia differenze molto marcate ed è evidente che l'area maggiormente colpita sia quella africana, che ha una media di oltre 3 nuovi infetti ogni 1000 abitanti, mentre le altre aree del mondo hanno medie tutte inferiori a 0.5. Le grandi differenze trovate potrebbero essere collegate a fattori di varia natura, che potrebbero essere approfonditi in studi futuri, incrociando i dati delle nuove infezioni da HIV a dati relativi a fattori socio economici, di prevenzione o accesso a servizi sanitari.

Viste le differenze generali tra macroaree, si è ritenuto necessario un approfondimento per ogni regione, per vedere se tutte seguono gli stessi andamenti e per andare ad analizzare le differenze tra i singoli paesi appartenenti agli stessi gruppi. Per eseguire un'analisi dettagliata delle singole aree geografiche, sono state create funzioni standard che hanno permesso di generare grafici informativi per ogni regione, analizzando l'andamento dei nuovi casi nel corso degli anni, la variazione dei casi tra i sessi, la media dei casi per paese e l'andamento temporale in ogni stato. Di seguito, verranno presentati i risultati più rilevanti per ciascuna area, insieme a considerazioni specifiche. Si è prestata maggiore attenzione alle aree con tassi di infezione più alti, mentre le altre sono riportate brevemente, in quanto non hanno un grande impatto sull'andamento generale.

*a) **Africa**:* come mostrato dalla Figura 6, la regione africana è quella maggiormente colpita dall'HIV, quindi le tendenze mondiali saranno fortemente influenzate da quelle di questa regione, che è stata analizzata in maniera più dettagliata. L'andamento dei contagi nel continente nel corso degli anni, mostrato in Figura 7, segue un andamento molto simile a quello di tutto il mondo, visto precedentemente, con valori medi più elevati.

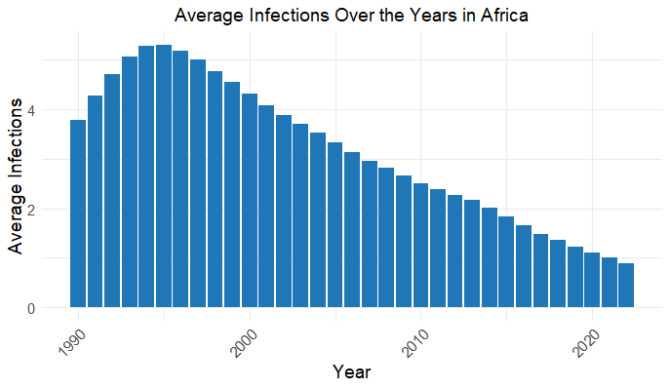

Fig. 7. Grafico a barre della variazione della media di nuove infezioni negli anni in Africa

L'andamento di crescita dura per circa 5 anni, arrivando ad un picco di oltre 5 nuovi casi ogni 1000 individui, iniziando poi a scendere fino al 2022, quando si è registrata una media di circa 1. Il grafico in Figura 8 vuole analizzare la differenza della media di nuove infezioni di uomini e donne, nel continente africano e come questa è cambiata nel corso degli anni presi in esame, mostrando che la media di nuove infezioni nella popolazione femminile risulta, anche in questo caso, essere maggiore rispetto a quella della popolazione maschile, confermando così quanto mostrato prima dall'analisi a livello globale.

Il grafico a linee mostra una differenza piuttosto marcata tra i sessi, soprattutto in corrispondenza del picco, in cui la

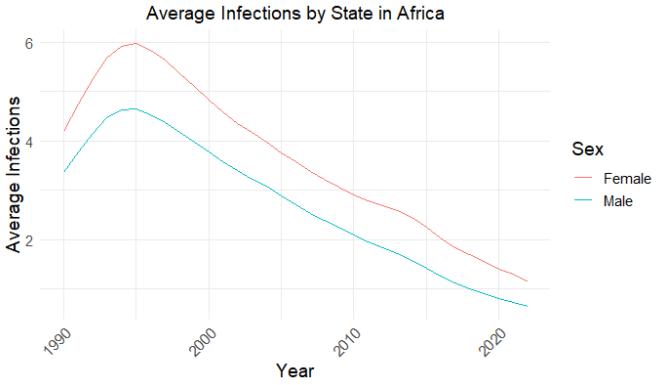

Fig. 8. Grafico a linee della variazione delle medie di nuove infezioni tra maschi e femmine durante gli anni in Africa

media delle donne arriva a un valore di 6, mentre quella degli uomini è inferiore a 5.

L'analisi prosegue andando ad osservare le differenze tra i vari paesi della regione africana, tramite il bar-plot in Figura 9, per confrontare il numero medio di nuove infezioni per ogni paese appartenente alla regione. Nel grafico sono mostrati solo i paesi con una media di infetti superiore a 1.2 su 1000 abitanti.

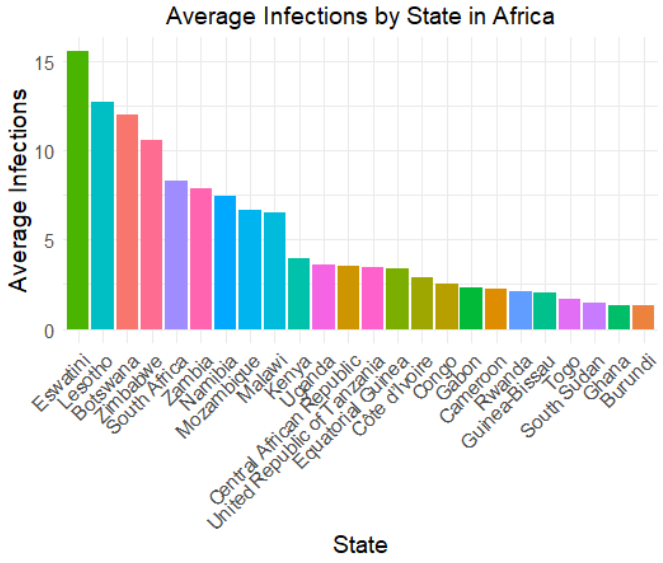

Fig. 9. Grafico a barre delle medie dei nuovi infetti nei paesi africani

Si evidenziano grandi differenze tra i vari stati della regione, Eswatini è il paese più colpito con una media di oltre 15 nuovi infetti ogni 1000 individui. In generale si è potuto notare che gli stati maggiormente colpiti sono quelli dell'Africa meridionale, con tassi di infezione superiori alla media.

b) ***America****:* l'area americana è la seconda per media di nuovi positivi all'HIV, si è dunque scelto di analizzarla attentamente per poter notare eventuali differenze rispetto all'andamento globale e a quello della regione maggiormente colpita. La sua analisi mostra una evoluzione della media nel tempo, in Figura 10, che cresce nei primi anni, fino al 1994, iniziando poi a scendere fino ad arrivare a una media quasi costante dal 2015. Il grafico non si discosta particolarmente da quelli precedentemente osservati, confermando come il periodo di maggior diffusione sia quello intorno al 1995, mentre si registrano medie più basse rispetto alla regione africana e rispetto alle medie globali.

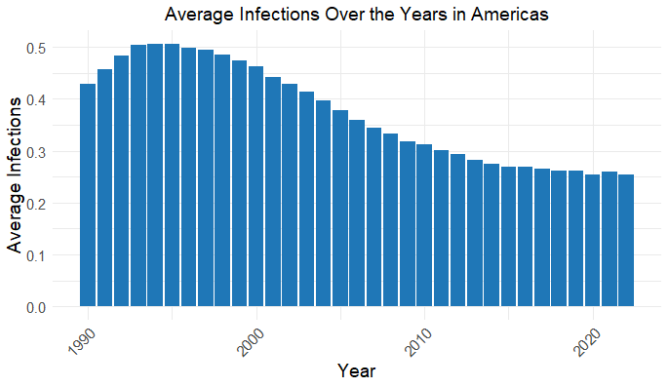

Fig. 10. Grafico a barre della variazione della media di nuove infezioni negli anni in America

Se la variazione della media nel tempo segue l'andamento generale, per quanto riguarda le differenze tra i sessi, in America si riscontrano dei risultati in contrasto con quanto visto prima. Infatti il plot in Figura 11 mostra come la media dei nuovi infetti in America sia superiore per gli uomini rispetto che per le donne, con una distanza massima di 0.2.

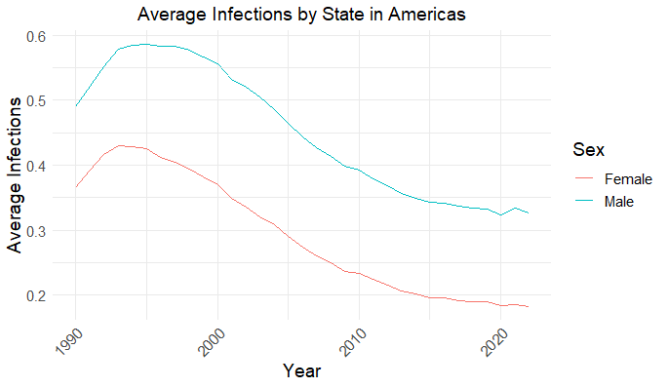

Fig. 11. Grafico a linee della variazione delle medie di nuove infezioni tra maschi e femmine durante gli anni in America

Infine sono state esaminate le medie di tutti i paesi appartenenti alla regione come mostra la Figura 12. L'area americana comprende al suo interno molti paesi diversi tra loro, molto distanti e con situazioni differenti sotto vari punti di vista, ciò può essere la causa che fa registrare grandi differenze tra gli stati. Haiti risulta essere il paese con maggior tasso di nuove infezioni, con una media di circa 1.25, seguito da altri paesi dell'area centro-americana con tasso di nuovi infetti intorno a 0.75. Nei paesi restanti si registrano invece dati con una media inferiore.

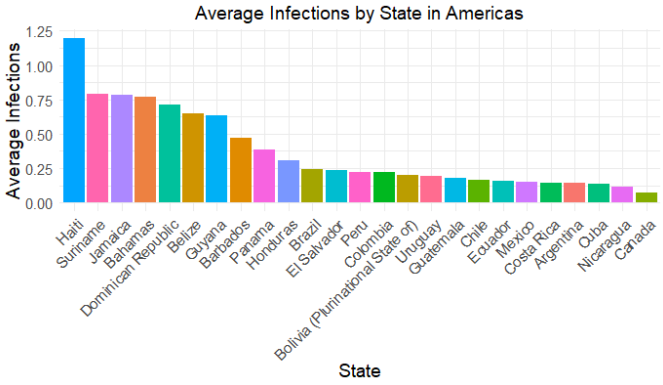

Fig. 12. Grafico a barre delle medie dei nuovi infetti nei paesi americani

c) ***Western Pacific****:* l'area in esame ha una media di infezioni minore rispetto alle aree precedentemente analizzate e alla media globale. La sua distribuzione ha un picco di 0.2 nel 1997, seguito da una fase di diminuzione, una serie di anni con tasso costante e infine un nuovo aumento. L'andamento temporale quindi non segue esattamente l'andamento globale e anche in questa area gli uomini sono maggiormente colpiti.

d) ***South-east Asia****:* la regione ha un tasso di nuove infezioni registrate piuttosto basso e una distibuzione con un massimo iniziale nel 1991 e un trend successivo quasi sempre in calo. Si riscontra un tasso di infezione superiore per il sesso maschile a differenza di quanto osservato per i dati globali.

e) ***Europa****:* il continente europeo ha una media di nuovi infetti inferiore rispetto alla media globale e un trend differente, infatti in questo caso il picco si è verificato intorno al 2011, seguito da una fase di discesa, ma una media del 2022 che indica un nuovo aumento.

f) ***Eastern Mediterranean****:* è l'area meno colpita dall'HIV, la media di nuove infezioni non supera mai gli 0.035 infetti ogni 1000 individui e ha un trend in crescita fino al 2001, seguito da un calo e un nuovo aumento dal 2019 in poi.

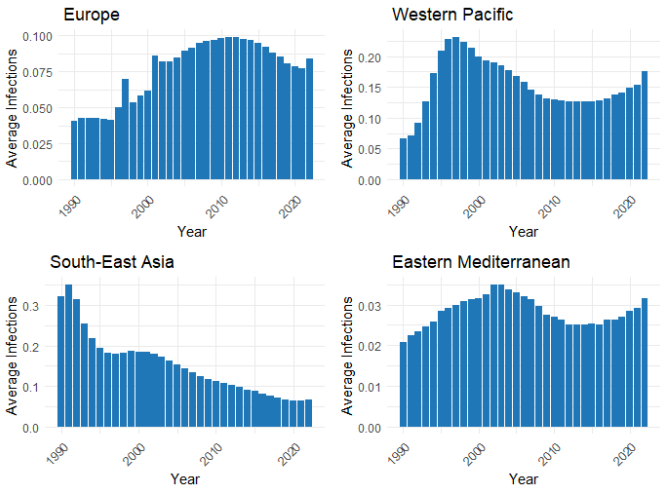

Fig. 13. Grafici a barre della variazione delle medie di nuove infezioni negli anni

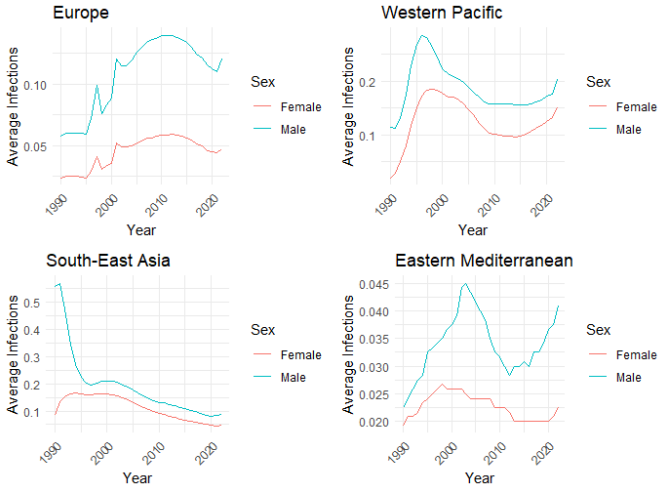

Fig. 14. Grafici a linee della variazione delle medie di nuove infezioni tra maschi e femmine durante gli anni

La Figura 13 e la Figura 14 mostrano l'andamento nel tempo delle nuove infezioni da HIV nelle 4 regioni meno colpite e le differenze tra i sessi. Gli andamenti dei grafici si discostano tutti da quelli relativi alla situazione globale e i valori risultano essere sempre inferiori. Graficamente è stato quindi possibile notare importanti differenze tra le aree, che verranno poi esaminate anche mediante uso di test statistici.

V. Test Statistici

In seguito ad una approfondita analisi esplorativa, è stato possibile estrarre alcune conclusioni parziali sui dati trattati. In questa sezione si è deciso di andare a verificare tramite l'utilizzo di test statistici alcune osservazioni fatte sui dati precedentemente.

A. Test sulla feature sex

Durante l'EDA è stato possibile notare alcune differenze nelle distribuzioni di nuovi casi tra maschi e femmine, si è dunque cercato di analizzare più attentamente la questione tramite l'applicazione di test statistici. Per prima cosa è stata verificata la normalità, poichè requisito fondamentale di molti test statistici. Già dai grafici delle distribuzioni era possibile intuire che i due sessi non avessero una distribuzione normale, ma si è deciso di utilizzare il test di Shapiro.

TABLE I
Test di Shapiro per i due sessi

Sesso	Statistica W	p-value
Maschio	0.41963	$< 2.2 \times 10^{-16}$
Femmina	0.41593	$< 2.2 \times 10^{-16}$

Il valore del p-value molto basso, mostrato nella Tabella 1, spinge a rifiutare l'ipotesi di normalità, è quindi necessario utilizzare test non parametrici per verificare se le distribuzioni di nuove infezioni nel sesso maschile e femminile abbiano differenze rilevanti dal punto di vista statistico. Si è deciso di applicare il test di Mann-Whitney per confrontare le distribuzioni dei valori tra i due gruppi, evitando così di

applicare trasformazioni ai dati che avrebbero fatto perdere l'interpretazione di questi.

TABLE II
RISULTATI DEL TEST DI MANN-WHITNEY

Gruppo	Statistica W	p-value
Maschio vs Femmina	8222544	$< 2.2 \times 10^{-16}$

Il risultato del test non parametrico applicato, riportato nella Tabella 2, suggerisce che c'è una forte evidenza contro l'ipotesi nulla, quindi significa che le distribuzioni dei due gruppi sono significativamente diverse, come si era potuto notare anche dai plot.

B. Test sulla feature area geografica

Un ulteriore test statistico può essere applicato alle differenze tra le varie aree geografiche. Dall'analisi esplorativa si notano differenze generali tra le varie zone, che potrebbero essere verificate tramite applicazione dell'ANOVA. Il test in questione richiede normalità dei dati e omoschedasticità, dalle distribuzioni è già possibile intuire che non vi è normalità ed inoltre sono stati applicati i test di Shapiro e Barnett, che hanno riportato tutti p-value prossimi allo 0. Senza i requisiti fondamentali non è dunque possibile applicare ANOVA, perciò si è deciso di utilizzare il test non parametrico di Kruskal-Wallis, per vedere se sono presenti delle differenze significative tra i gruppi.

TABLE III
RISULTATI DEL TEST DI KRUSKAL-WALLIS

Test	Statistica	p-value
Kruskal-Wallis	5009.8	$< 2.2 \times 10^{-16}$

Il risultato del test, mostrato in Tabella 3, evidenzia chiaramente che le aree differiscono tra di loro in maniera significativa ed è necessario capire se le differenze siano dovute ad una singola area oppure riguardino tutte. Si è dunque resa necessaria l'applicazione di una analisi post-hoc, tramite l'applicazione del test statistico non parametrico di Mann-Whitney, applicato a tutte le coppie di aree geografiche. Il test è stato effettuato utilizzando prima la correzione di Bonferroni, con risultati visibili in Tabella 4, e in seguito Benjamini-Hochberg, i cui risultati sono riportati nella Tabella 5. L'utilizzo di correzioni si è reso necessario per evitare che la probabilità di errore di prima specie diventi troppo alta, influenzando troppo i risultati. Ne sono state utilizzate due versioni differenti per vedere se e quanto i risultati differiscono.

La correzzione di Bonferroni è più conservativa, mentre quella di Benjamini-Hochberg è più adattabile, ma in questo caso danno risultati molto simili, infatti nonostante alcuni valori del p-value siano diversi, entrambe le versioni indicano come tutte le zone presentino differenze significative tra di loro, ad eccezione di Europa e South-East Asia.

TABLE IV
RISULTATI DEI TEST DI MANN-WHITNEY CON CORREZIONE DI BONFERRONI TRA COPPIE DI AREE GEOGRAFICHE

Area1	Area2	p-value (Bonferroni)
Americas	Africa	$< 2.2 \times 10^{-16}$
Americas	Eastern Mediterranean	$< 2.2 \times 10^{-16}$
Americas	Europe	$< 2.2 \times 10^{-16}$
Americas	South-East Asia	$< 2.2 \times 10^{-16}$
Americas	Western Pacific	$< 2.2 \times 10^{-16}$
Eastern Mediterranean	Africa	$< 2.2 \times 10^{-16}$
Eastern Mediterranean	Europe	$< 2.2 \times 10^{-16}$
Eastern Mediterranean	South-East Asia	3.1×10^{-14}
Eastern Mediterranean	Western Pacific	2.3×10^{-12}
Europe	Africa	$< 2.2 \times 10^{-16}$
Europe	South-East Asia	1
Europe	Western Pacific	2.3×10^{-6}
South-East Asia	Western Pacific	2.3×10^{-6}

TABLE V
RISULTATI DEI TEST DI MANN-WHITNEY CON CORREZIONE DI BENJAMINI-HOCHBERG (BH) TRA COPPIE DI AREE GEOGRAFICHE

Area1	Area2	p-value (BH)
Americas	Africa	$< 2.2 \times 10^{-16}$
Americas	Eastern Mediterranean	$< 2.2 \times 10^{-16}$
Americas	Europe	$< 2.2 \times 10^{-16}$
Americas	South-East Asia	$< 2.2 \times 10^{-16}$
Americas	Western Pacific	$< 2.2 \times 10^{-16}$
Eastern Mediterranean	Africa	$< 2.2 \times 10^{-16}$
Eastern Mediterranean	Europe	$< 2.2 \times 10^{-16}$
Eastern Mediterranean	South-East Asia	2.6×10^{-15}
Eastern Mediterranean	Western Pacific	1.8×10^{-13}
Europe	Africa	$< 2.2 \times 10^{-16}$
Europe	South-East Asia	0.71
Europe	Western Pacific	1.6×10^{-07}
South-East Asia	Western Pacific	1.6×10^{-07}

VI. CONCLUSIONI

Le analisi effettuate mostrano la variazione del tasso di nuove infezioni da HIV negli anni e le differenze dovute al sesso e all'area geografica. L'utilizzo di grafici e di test statistici ha permesso di osservare che il numero di nuovi infetti è fortemente influenzato dal sesso e dall'area geografica, sarebbe utile effettuare uno studio più approfondito, che possa integrare nuovi dati per capire al meglio il motivo per cui le distribuzioni sono così diverse, individuando altri eventuali fattori che influenzano il tasso di nuovi positivi. I risultati dello studio sono dunque solo indicativi e possono essere una base per studi futuri.

REFERENCES

[1] "HIV e AIDS", Ministero della Salute, URL: https://www.salute.gov.it/portale/hiv

[2] "Infezione da HIV e AIDS", Istituto Superiore di Sanità, URL: https://www.epicentro.iss.it/aids/

[3] "HIV", World Health Organization, URL: https://www.who.int/news-room/fact-sheets/detail/hiv-aids

[4] "New HIV infections (per 1000 uninfected population)", Dataset WHO, URL: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/new-hiv-infections-(per-1000-uninfected-population)