Mortalità nel mondo: analisi descrittiva e statistica dei fattori che la influenzano

Ergys Dervishaj, Luca Fagiolo
Dipartimento di Ingegneria
Perugia, Italia
ergys.dervishaj@studenti.unipg.it, luca.fagiolo@studenti.unipg.it

***Abstract*—In questo studio si ha l'obiettivo di analizzare la situazione in termini di mortalità ed i vari fattori che la influenzano nei vari continenti del mondo utilizzando il dataset "Global Health and Development (2012-2021)" [2]. L'analisi è stata condotta utilizzando il linguaggio R mediante strumenti statistici e plot dei risultati ottenuti. In particolare sono state applicate statistiche di tipo descrittivo, test di normalità (Shapiro-Wilk [4]), confronto tra gruppi (Kruskal-Wallis [3]) ed analisi post-hoc con correzioni di Bonferroni [1]. In base ai risultati si nota un forte contributo positivo rispetto alla mortalità dato da fattori come l'accessibilità all'acqua potabile, investimenti nella sanità, tasso di immunizzazione.**

I. Introduzione

Il dataset, disponibile su Kaggle, raccoglie una vasta gamma di indicatori socio-sanitari e demografici provenienti da fonti autorevoli come la World Health Organization (WHO) [9] [10] [11] e la World Bank [8] [6] [5] [7]. I dati coprono un periodo di dieci anni e includono variabili relative alla salute pubblica (come aspettativa di vita, mortalità infantile, accesso all'acqua potabile) e allo sviluppo economico (PIL pro capite, spesa sanitaria sul PIL, tasso di disoccupazione), offrendo una panoramica globale sull'evoluzione del benessere nei diversi Paesi del mondo. La World Health Organization, agenzia specializzata delle Nazioni Unite, è il principale riferimento internazionale per le politiche sanitarie, mentre la World Bank si occupa di promuovere lo sviluppo economico attraverso il sostegno finanziario e tecnico ai Paesi in via di sviluppo. L'integrazione dei dati di queste due istituzioni consente un'analisi completa delle relazioni tra salute e sviluppo, utile per identificare disuguaglianze, monitorare i progressi verso gli Obiettivi di Sviluppo Sostenibile (SDGs) e supportare le decisioni politiche basate sull'evidenza. In particolare, le principali metriche sulla salute, come l'aspettativa di vita e la mortalità, sono tratte dal Global Health Observatory della WHO, mentre le statistiche sulla popolazione, l'economia e l'ambiente provengono dal World Development Indicators (WDI) della World Bank. In questo progetto verrà condotta un'analisi statistica esplorativa e inferenziale sui dati disponibili, con l'obiettivo di indagare le principali correlazioni tra indicatori sanitari e socio-economici.

II. Materiali e Metodi

A. Dataset

Il dataset contiene osservazioni raccolte dal 2012 al 2021, sono presenti quindi 10 occorrenze per ogni stato. Le features principali sono: nome dello stato (*Country*), anno (*Year*), aspettativa di vita (*Life_Expectancy*), aspettativa di vita femminile (*Life_Expectancy_Female*), aspettativa di vita maschile (*Life_Expectancy_Male*), PIL pro capite (*GDP_Per_Capita*), tasso di obesità (*Obesity_Rate_Percent*), tasso di sottopeso (*Underweight_Rate_Percent*), tasso di accesso all'acqua potabile (*Water_Access_Percent*), inquinamento dell'aria (*Air_Pollution*), immunizzazione (*Immunization_Rate*), tasso di consumo di alcool (*Alcohol_Consumption_Per_Capita*).

B. Strumenti

Le analisi sono state condotte in R, utilizzando i pacchetti `tidyverse` per la manipolazione dei dati e `ggpubr` per la visualizzazione. `corrplot` ha supportato l'analisi delle correlazioni, `dunn.test` per i confronti post-hoc non parametrici, e `forecast` per l'analisi delle serie storiche. Sono stati inoltre impiegati `gridExtra` e `scales` per la gestione grafica, `countrycode` per la conversione dei codici paese, `fmsb` per i grafici radar e `reshape2` per la ristrutturazione dei dati.

III. Analisi Esplorativa

A. Statistiche descrittive

Per l'analisi esplorativa dei dati sono state calcolate statistiche descrittive (media, mediana, minimo, massimo) relative alle features prese in esame in rapporto al tasso di mortalità. In particolare, sono state realizzate rappresentazioni di tipo scatter plot, istogrammi, spider plot e violin plot. Lo scatter plot (Fig. 1) è utilizzato per mostrare il rapporto tra aspettativa di vita maschile e femminile. Gli istogrammi (Fig. 5) sono utilizzati per graficare la distribuzione dell'accesso medio all'acqua potabile tra i vari continenti che sono ordinati rispetto al tasso. Lo spider plot (Fig. 6-10) è utilizzato per mostrare in modo riassuntivo, per ogni continente, quali tra utilizzo di alcolici, accesso all'acqua potabile, immunizzazione, investimenti in ambito sanitario e qualità dell'aria, sono i fattori che maggiormente influenzano l'aspettativa di vita. Il violin plot (Fig. 14), vantaggioso perché combina le caratteristiche di un boxplot alle curve di densità.

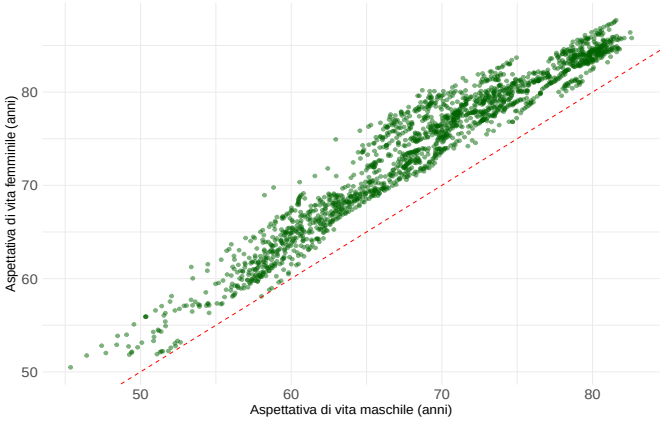

Fig. 1. Confronto tra aspettativa di vita maschile e aspettativa di vita femminile

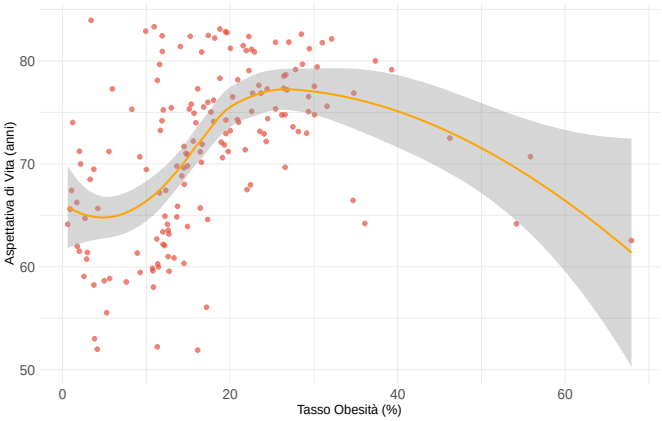

Fig. 4. Relazione tra tasso obesità e aspettativa di vita

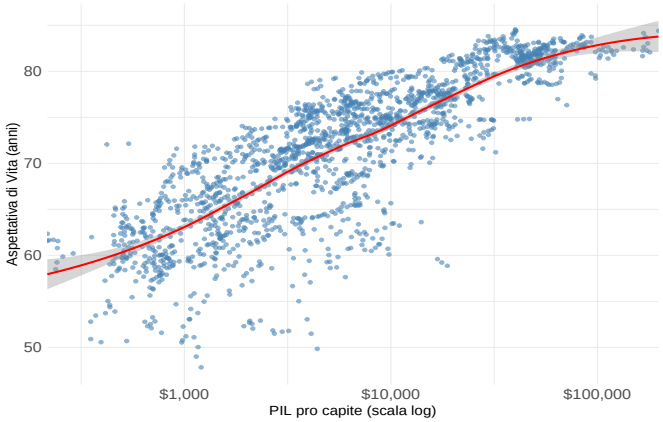

Fig. 2. Relazione tra PIL Pro capite e aspettativa di vita

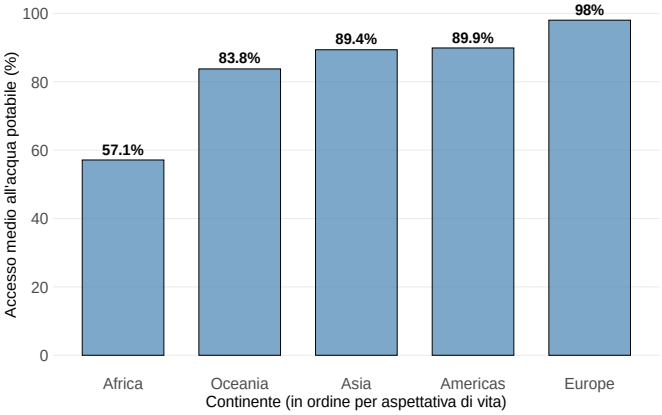

Fig. 5. Accesso all'acqua potabile tra i vari continenti

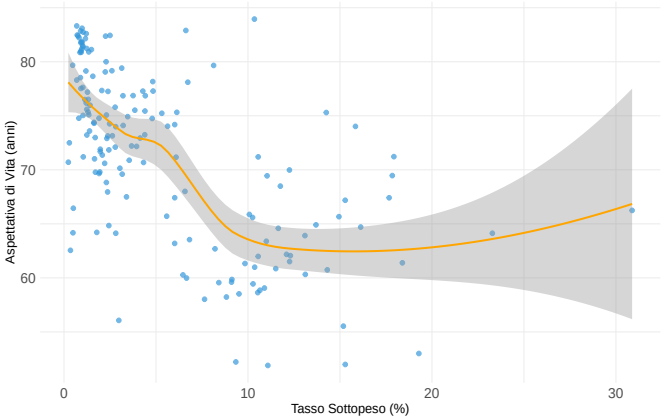

Fig. 3. Relazione tra tasso sottopeso e aspettativa di vita

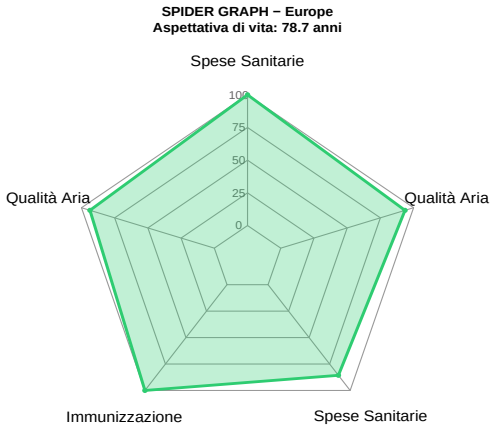

Fig. 6. Spider graph per il continente: Europe

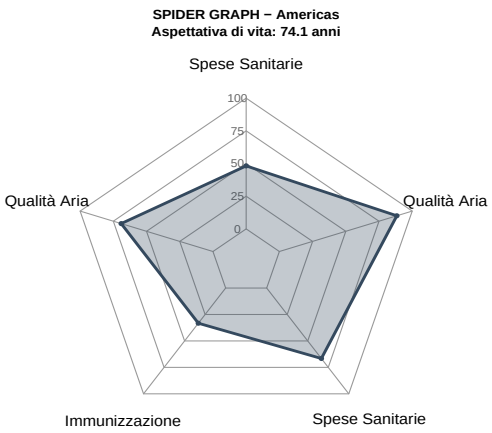

Fig. 7. Spider graph per il continente: Americas

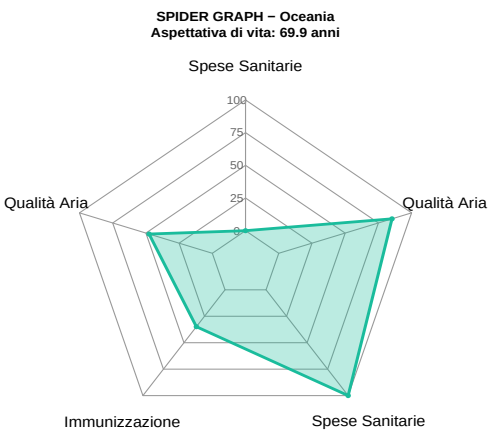

Fig. 9. Spider graph per il continente: Oceania

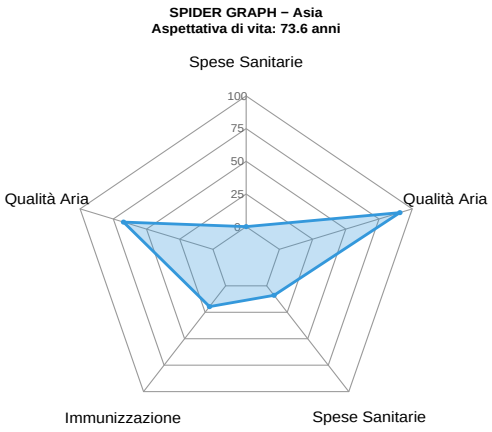

Fig. 8. Spider graph per il continente: Asia

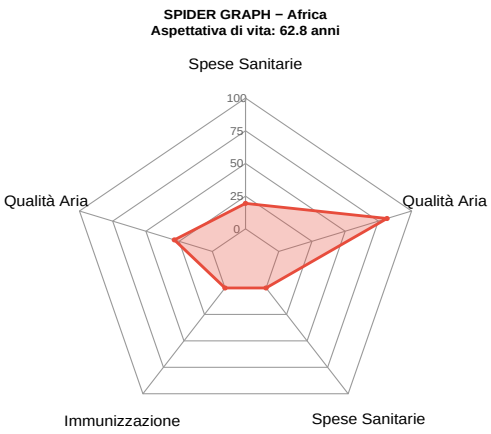

Fig. 10. Spider graph per il continente: Africa

IV. Analisi Statistica

A. *Test di normalità*

La variabile *Life Expectancy* è stata inizialmente sottoposta al test di normalità di Shapiro-Wilk [4], ottenendo un p-value inferiore alla soglia di significatività ($p < 0.001$), suggerendo una deviazione significativa dalla distribuzione normale. Per verificare visivamente questa ipotesi, è stato prodotto un *Q-Q plot* (Fig. 11), che conferma la non normalità della distribuzione.

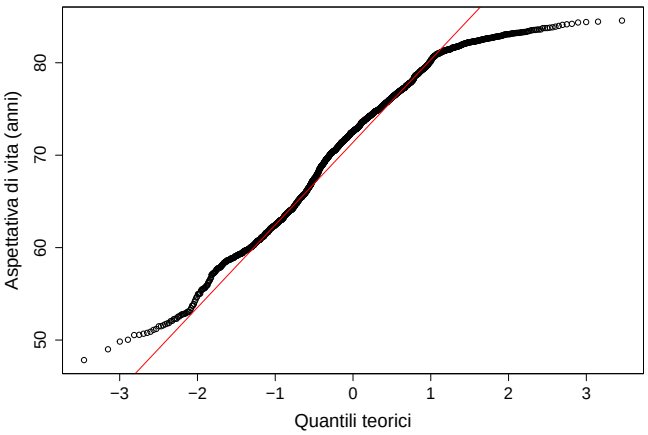

Fig. 11. Grafico Q-Q per valutare normalità

Successivamente è stata applicata una trasformazione di Box-Cox. Il valore ottimale di λ è stato stimato utilizzando la funzione `BoxCox.lambda()`. Anche dopo la trasformazione, il test di Shapiro-Wilk [4] ha evidenziato una deviazione significativa dalla normalità ($p < 0.001$), giustificando l'utilizzo di test non parametrici nelle analisi successive.

TABLE I
P-VALUE DEL TEST DI SHAPIRO-WILK SULL'ASPETTATIVA DI VITA

	Pre-BoxCox	**Post-BoxCox**
P-value	1.86×10^{-20}	4.48×10^{-19}

B. Analisi di correlazione

È stata condotta un'analisi di correlazione di Spearman tra l'aspettativa di vita e un insieme di variabili socio-sanitarie. Le correlazioni più elevate (in valore assoluto) sono state osservate con:

- *Spese sanitari per capita*,
- *GDP per capita*,
- *Accesso all'acqua potabile*,

Una matrice di correlazione colorata (Fig. 12) è stata prodotta per facilitare la lettura delle relazioni tra variabili. Ai fini del progetto, è poi stata valutata la significatività statistica delle correlazioni tra l'aspettativa di vita e alcune feature come si può osservare dai risultati nella seguente tabella.

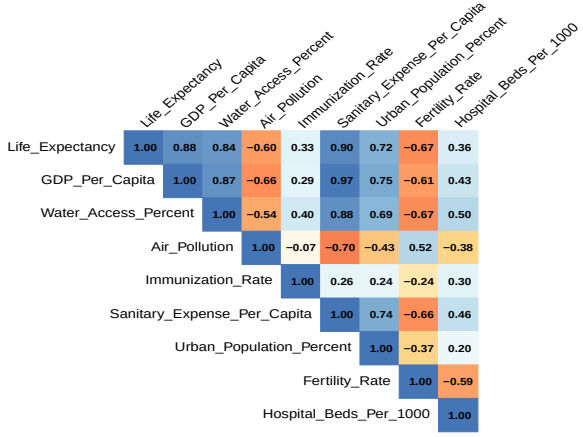

Fig. 12. Matrice di correlazione di Spearman tra le variabili numeriche.

TABLE II
CORRELAZIONI DI SPEARMAN CON L'ASPETTATIVA DI VITA

Variabile	**Correlazione**	**P-value**
GDP_Per_Capita	0.879	9.44×10^{-276}
Water_Access_Percent	0.840	3.20×10^{-227}
Air_Pollution	-0.596	6.32×10^{-83}
Immunization_Rate	0.328	9.67×10^{-23}
Sanitary_Expense_Per_Capita	0.904	1.69×10^{-315}
Urban_Population_Percent	0.720	7.17×10^{-137}
Fertility_Rate	-0.667	1.48×10^{-110}
Hospital_Beds_Per_1000	0.362	8.34×10^{-28}

C. Confronto tra continenti (test di Kruskal-Wallis)

È stato applicato il test non parametrico di Kruskal-Wallis [3] per confrontare le distribuzioni dell'aspettativa di vita tra i vari continenti. Il test ha restituito un valore di $\chi^2 = 154.37$ con $p < 0.001$, indicando differenze statisticamente significative tra almeno un paio di gruppi.

Per identificare le differenze specifiche tra continenti, è stato effettuato un test post-hoc di Dunn con correzione di Bonferroni. I risultati sono stati visualizzati tramite una heatmap dei p-value corretti. Le coppie di continenti che mostrano differenze significative ($p < 0.001$) sono evidenziate tramite una colorazione differente.

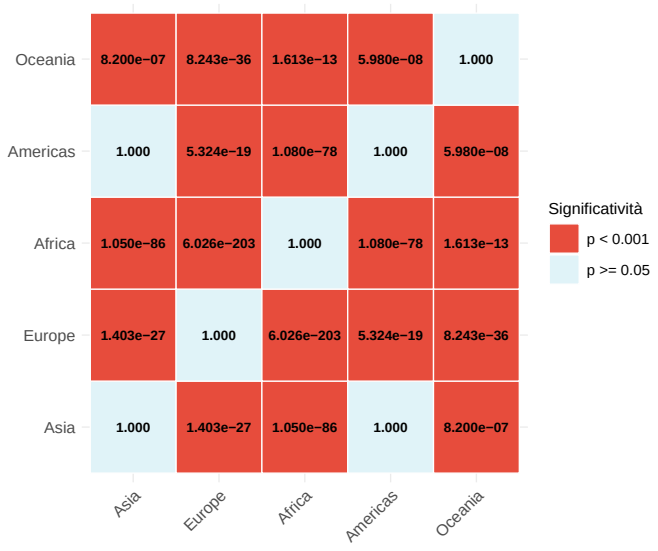

Fig. 13. Heatmap del test post-hoc di Dunn (Bonferroni).

D. Visualizzazione dei risultati

Per evidenziare le differenze tra continenti è stato utilizzato un Violin plot (Fig. 14) che mostra la distribuzione dell'aspettativa di vita per continente, evidenziando la mediana con un punto rosso. Il grafico supporta l'esito del test statistico, mostrando differenze marcate tra i continenti, in particolare tra quelli con livelli socio-economici più elevati e quelli con accesso ridotto ai servizi sanitari.

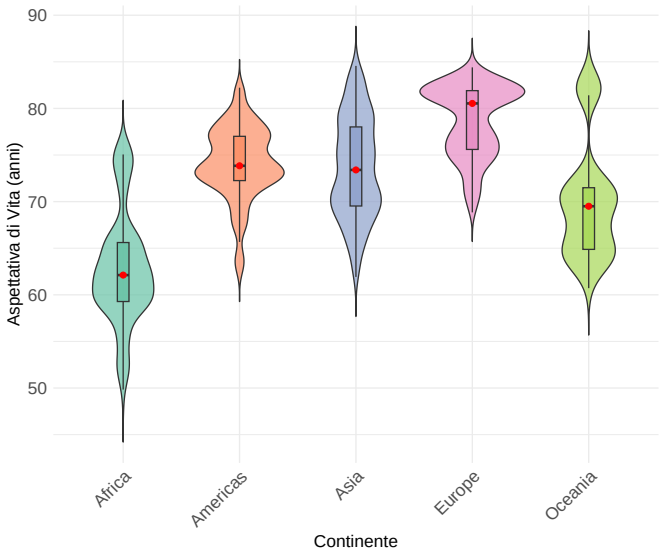

Fig. 14. Violin plot dell'aspettativa di vita per continente.

V. Conclusione

L'analisi condotta ha evidenziato come l'aspettativa di vita sia fortemente influenzata da fattori socio-economici e sanitari, tra cui la spesa sanitaria pro capite, l'accesso all'acqua potabile e il PIL. Le tecniche non parametriche impiegate si sono dimostrate adeguate in presenza di dati non normalmente distribuiti, consentendo confronti significativi tra aree geografiche. Le differenze osservate tra i continenti sottolineano il ruolo cruciale dell'investimento in salute pubblica nel miglioramento della qualità e durata della vita. L'integrazione tra analisi statistica e visualizzazione grafica ha permesso di restituire una rappresentazione chiara e approfondita del fenomeno, utile a fini descrittivi ed esplorativi.

References

[1] C. BONFERRONI. "Teoria statistica delle classi e calcolo delle probabilita". In: *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze* 8 (1936), pp. 3–62. URL: https://cir.nii.ac.jp/crid/1570009749360424576.

[2] *Global Health and Development (2012-2021)*. URL: https://www.kaggle.com/datasets/martinagalasso/global-health-and-development-2012-2021/data.

[3] William H. Kruskal and W. Allen Wallis. "Use of Ranks in One-Criterion Variance Analysis". In: *Journal of the American Statistical Association* 47.260 (1952), pp. 583–621.

[4] S. S. SHAPIRO and M. B. WILK. "An analysis of variance test for normality (complete samples)†". In: *Biometrika* 52.3-4 (Dec. 1965), pp. 591–611.

[5] World Bank. *Annual alcohol consumption in liters per person*. URL: https://data.worldbank.org/indicator/SH.ALC.PCAP.LI.

[6] World Bank. *Fertility rate, total (births per woman)*. URL: https://data.worldbank.org/indicator/SP.DYN.TFRT.IN.

[7] World Bank. *Immunization (percentage of total population)*. URL: https://data.worldbank.org/indicator/SH.IMM.IDPT.

[8] World Bank. *Population, total*. URL: https://data.worldbank.org/indicator/SP.POP.TOTL.

[9] World Health Organization. *Life expectancy at birth (years)*. https://data.who.int/indicators/i/A21CFC2/90E2E48.

[10] World Health Organization. *Obesity among adults, BMI $\geq$ 30, prevalence (age-standardized estimate)*. URL: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/prevalence-of-obesity-among-adults-bmi-=-30-(age-standardized-estimate)-(-).

[11] World Health Organization. *Overweight among adults, BMI $\geq$ 25, prevalence*. URL: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/prevalence-of-overweight-among-adults-bmi-=-25-(age-standardized-estimate)-(-).