Questioning the World Happiness Report's Methodology: A Critical Analysis

Luis Irisarri Galera[*]
Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB),
Campus UIB, 07122 Palma de Mallorca, Spain.

(Dated: February 14, 2024)

This study conducts a critical analysis of the World Happiness Report's (WHR) methodology, focusing on the interdependencies among key variables. Our analysis reveals a significant dependence between GDP and Healthy Life Expectancy, indicating redundancy in the data. Additionally, we highlight the limitations of the Generosity variable in capturing the full spectrum of altruistic behaviour. Furthermore, our evaluation of well-being measures uncovers their inadequacy in reflecting relationships with anxiety-related mental disorders. These findings underscore the need for methodological improvements and a reevaluation of variables to accurately measure national happiness.

I. INTRODUCTION

Throughout history, the pursuit of happiness has been a central theme in philosophy, engaging thinkers from diverse cultures in a quest to define and achieve well-being. From the ethical virtues of Aristotle and the contemplative enlightenment of Siddhartha Gautama to the utilitarian principles of John Stuart Mill and Jeremy Bentham, the discourse on happiness has traversed many different paths [1]. Despite the apparent similarities in seeking a good life, the divergent theories underscore the complexity of happiness as a concept. This lack of convergence highlights not only the multifaceted nature of happiness but also the influence of cultural contexts in shaping its understanding, making it an enduring subject of human inquiry and debate [2].

In the quest for a unified theory of happiness, science has made significant strides through the development of *Positive Psychology*, often referred to as the "Science of Happiness". This field, pioneered by Martin Seligman and Christopher Peterson [3], delves into the essential elements that make life worth living, aiming to uncover the factors that foster human flourishing. Building on the foundational works of scholars like Daniel Kahneman and Ed Diener, Positive Psychology has coalesced around a comprehensive view of **happiness** that integrates the global sense of well-being with the balance of positive versus negative emotions[1] [5, 6]. This approach represents a pivotal shift towards understanding happiness not just as a fleeting state, but as a sustainable condition of well-being enriched by positive experiences.

One of the central challenges in the science of happiness is the question of measurement. Researchers employ a variety of methods, including **Observational Experience Sampling** studies, which systematically gather self-reports of behaviors, emotions, or experiences in natural settings [7]. **Cross-sectional** studies which analyze data from a population at a single point in time to identify correlations between variables [8]. **Longitudinal** studies extend this by observing changes over extended periods, though they are more ambitious and costly [9]. Then, in order to pinpoint causal relationships, researchers use **Experiments** to manipulate variables and observe their effects. Although these methods provide valuable insights, they are often limited by the subjective nature of self-reports and the complexity of human emotions. In fact, there are also innovative methods that analice behavioural indicators, such as facial muscle activity and brain chemistry, offering nuanced insights into the fleeting nature of happiness [4].

In this context, the *World Happiness Report* (WHR) [10] posits even a more ambitious question which is to quantify happiness at the national level[2]. Annually, since 2012 the WHR, ranks countries based on their citizens self-reported well-being. Then, researchers try to understand the variability of happiness between countries by considering variables such as income, social support, life expectancy, freedom to make life choices, generosity, and perceptions of corruption. The WHR has garnered widespread attention for its comprehensive approach to measuring happiness, offering a valuable resource for policymakers, researchers, and the general public [10]. However, the WHR has also faced criticism for its methodology, with some scholars questioning the validity and reliability of its measures. This has sparked a lively debate about the role of the WHR in shaping public policy and the need for more rigorous approaches to measuring happiness [12, 13].

The aim of this study is to critically examine the WHR from a data analysis standpoint. While acknowledging the WHR's valuable focus on human well-being, our analysis seeks to interrogate the report's methodology and data integrity. Specifically, this research endeavours to:

[*] luis.irisarri1@estudiant.uib.es

[1] This approach might seem superficial but there are tons of studies that show how significant this measures are [4].

[2] Historically, the concept of quantifying happiness at a national level was pioneered by Bhutan in the early 1970s. The kingdom adopted the Gross National Happiness (GNH) metric as a groundbreaking alternative to conventional economic measures, like GDP, seeking to capture a broader spectrum of national prosperity beyond financial indicators [11].

- Assess the Independence and Relevance of WHR Variables: we investigate the interrelations among the WHR variables to determine their independence and relevance in explaining happiness across nations. In particular, we question the independence between *Log GDP per capita* and *Healthy life expectancy at birth*. We also scrutinize the relevance of the *Generosity* variable, as defined by the WHR, in understanding national happiness levels.

- Evaluate the Reliability and Robustness of Well-being Measures: the study subjects the WHR's well-being quantifiers—such as the *ladder score*, *positive*, and *negative affects*—to rigorous statistical tests to evaluate their reliability and robustness as indicators of happiness.

Through these analyses, we aim to contribute the following claims:

Claim 1. *The variables used in order to understand the variability of Happiness should be independent and relevant.*

This emphasizes that variables must be non-redundant and encapsulate distinct dimensions of well-being. Such differentiation is vital for elucidating each variable's unique impact on national happiness levels and for mitigating the influence of potential confounders.

Claim 2. *Valid measures of happiness should exhibit positive correlations with established indicators of well-being and negative correlations with indicators of suffering, reflecting their intrinsic relationship with the broader construct of well-being.*

In particular, research supports that well-being measures should positively correlate with *Social support*, *Healthy life expectancy at birth*, and *Freedom to make life choices*, while negatively correlating with *Perceptions of corruption*, *Depression*, and *Anxiety* [14–16].

This critical analysis aims not only to scrutinize the WHR's methodological underpinnings but also to enrich the dialogue on measuring happiness, advocating for a nuanced approach that captures the complexity of well-being.

II. METHODS

This section outlines our study's methodology, starting with the data collection from key sources and followed by an overview of the statistical and computational techniques used to analyze the dataset.

A. Data Collection

This study utilizes data from two principal sources: the World Happiness Report (WHR) [10] and Our World in Data (OWID) [17]. The integration of these sources yields a comprehensive dataset wherein each row represents a country-year observation, and columns denote various well-being and socioeconomic indicators. Specifically, the dataset includes the following variables: *Life Ladder* (LL), *Log GDP per capita* (GDP), *Social support* (SS), *Healthy life expectancy at birth* (HLE), *Freedom to make life choices* (FMC), *Generosity* (G), *Perceptions of corruption* (PC), *Positive affect* (PA), and *Negative affect* (NA) from the WHR, complemented by *Depression* (D) and *Anxiety* (A) from OWID. From now on, we refer to LL, PA and NA as the **well-being variables**.

The temporal scope of the dataset covers the years 2005 to 2022, albeit with sporadic gaps in country data across different years. In particular, the years 2010, 2015 and 2019 are the only years without inferred data. Among these, 2019 boasts the largest sample size, making it the primary focus for single-year analyses conducted in this study. Further details regarding data definitions, sources, and sample sizes are delineated in appendix A.

B. Data Analysis

We assessed the **pairwise dependencies**[3] among variables in 2019. To this end, we calculated *Pearson's* (r), *Spearman's* (ρ), and the *Mutual Information* (I) as outlined in eqs. (B1) to (B3), respectively. Pearson's r quantifies linear correlations, Spearman's ρ assesses monotonic relationships, and Mutual Information I measures the general dependence between variables. This multifaceted approach enables a thorough evaluation of variable dependencies, crucial for understanding their impact on national happiness levels. We extended our analysis from 2006 to 2022, examining the temporal robustness of these dependencies.

To ascertain the statistical significance of our findings, we conducted *hypothesis testing* to challenge the *null hypothesis* (H_0) that dependencies ($\mathcal{D}$) are non-existent:

$$\mathrm{H}_0 : \mathcal{D} = 0, \qquad \mathrm{H}_1 : \mathcal{D} \neq 0. \tag{1}$$

Since not all of the variables satisfy the normality assumption as shown in fig. 1, we employed *permutation tests* for this purpose, offering a robust statistical validation of dependencies that is independent of the data's distribution.

[3] We use the term "dependency" to encompass correlations and mutual information, extending beyond its strict mathematical definition related to the separability of joint probability distributions, i.e. $P(X,Y) = P(X)P(Y)$.

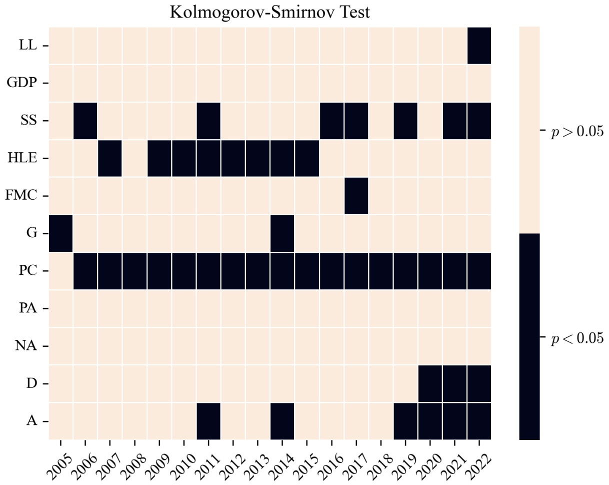


Figure 1: Kolmogorov-Smirnov test for the variables in the dataset over all the years. The p-values are calculated using the `kstest` function from the `scipy` library. $p > 0.05$ indicates that the null hypothesis that the two samples are drawn from the same distribution cannot be rejected, while $p < 0.05$ indicates that the null hypothesis can be rejected.

In our analysis, hierarchical clustering was utilized to systematically group variables based on their similarities, enabling a visual depiction of their interrelations. This process began by transforming the correlation matrix into a distance metric, as outlined in Equation B5. This transformation ensured that variables with higher correlations were positioned closer together, indicating similar behaviors, while those with lower correlations were placed further apart, signifying dissimilarity.

For the agglomerative clustering, we adopted the WPGMA algorithm. This method starts with each variable as an individual cluster and merges them based on the weighted average distance between clusters, ensuring a gradual and systematic aggregation into a unified structure. The resulting grouping is depicted in a dendrogram, visually mapping the hierarchical relationships among variables.

To validate the accuracy of the hierarchical clustering and the integrity of the dendrogram, we employed the *Cophenetic correlation coefficient*. This measure quantitatively evaluates how well the dendrogram preserves the original pairwise distances between variables, offering a critical assessment of the clustering algorithm's fidelity. The use of the Cophenetic correlation coefficient in our methodology provided a robust means to confirm the reliability of the hierarchical structure derived from our analysis [18].

For a more detailed explanation of the statistical methods used in this study, we direct the reader to appendix B.

C. Computational Specifics

The data analysis for this study was conducted using the `Python` programming language, leveraging several key libraries for data manipulation and analysis. Specifically, we utilized `pandas` for data manipulation, along with `numpy`, `scipy`, and `scikit-learn` for data analysis tasks. Correlation coefficients were calculated using the `pandas` `.corr()` method and `scipy`'s `.pearsonr()` and `.spearmanr()` functions, while for the mutual information, we employed `mutual_info_regression()` from `sklearn.feature_selection`. Hierarchical clustering analysis was performed with the `linkage()` function from `scipy`. For hypothesis testing, a custom permutation test function was developed and implemented. The source code for this analysis is available at `https://github.com/liris8`.

III. RESULTS

This section presents our results, following the methodology outlined in Section II. Initially, we focus on analyzing pairwise dependencies and hierarchical clustering for the variables in our dataset for the year 2019. Subsequently, we broaden our analysis to include data spanning from 2006 to 2022.

A. Pairwise Dependencies

We first consider the pairwise dependencies in 2019, the primary findings are illustrated in fig. 2. In light of claim 1, we observed that the correlations and mutual information between GDP and HLE are statistically significant, suggesting a pronounced interdependence between these variables. Specifically, tab. (I) highlights the strong link between economic prosperity and health, questioning the independence needed for cross-correlation analysis in happiness research.

	$r(\mathrm{GDP}, \mathrm{HLE})$	$\rho(\mathrm{GDP}, \mathrm{HLE})$	$I(\mathrm{GDP}, \mathrm{HLE})$
$\mathcal{D}$	0.81	0.85	0.69
p-value	$3.28 \cdot 10^{-34}$	$5.01 \cdot 10^{-40}$	0.00

Table I: GDP dependencies with HLE. Pearson's r, Spearman's ρ, and the mutual information I are given by eqs. (B1) to (B3). The p-values are calculated using permutation tests. The null hypothesis is given by eq. (1). $p_I = 0.00$ suggests computational limitations in handling extremely small numbers.

Conversely, p-value matrices (fig. 2) reveal minimal statistical significance for the *Generosity* variable. Neither LL nor NA show significant correlations with G, as indicated by the Pearson and Spearman coefficients. The PA has a ($r = 0.17$, $p = 0.044$), suggesting a weak relationship. On the other hand, the mutual information manages to capture some dependency between LL and G ($I = 0.21$, $p = 0.0010$), while failing to capture any with PA and NA.

	$r(\mathrm{G}, \mathrm{LL})$	$r(\mathrm{G}, \mathrm{PA})$	$r(\mathrm{G}, \mathrm{NA})$	$I(\mathrm{G}, \mathrm{LL})$
$\mathcal{D}$	0.02	0.17	0.05	0.21
p-value	0.79	0.05	0.55	0.001

Table II: Exploration of Generosity dependencies with well-being variables, employing Pearson's r correlation and mutual information as detailed in eqs. (B1) and (B3). Spearman's ρ is omitted, as it does not provide additional insight in this context. P-values are derived from permutation tests, testing the null hypothesis eq. (1).

We now evaluate claim 2 by examining each variable's relationships:

- **Life Ladder**: Significant correlations are observed with all variables except G. In contrast, the mutual information shows significance dependencies for all. Correlations and mutual information patterns align, particularly strong with GDP, SS, and HLE. FMC and PA follow in significance. NA exhibits the most substantial negative correlation and significant mutual information. PC and D also correlate negatively, while A correlates positively.

- **Positive affect**: Displays weaker statistical significance compared to LL, lacking significant Pearson correlation with GDP, HLE, and A. Significant mutual information is found with LL, FMC, and D. We observe notable positive correlations excluding PC, NA, and D. The most significant positive correlation is with FMC, followed by SS. NA, PC, and D are negatively correlated, while mutual information highlights dependencies with FMC and D, but not with SS, NA, or A.

- **Negative affect**: Shows significant correlations with all variables except G and A, while the mutual information is significant for LL and SS only. All the significant correlations exhibit the same pattern. D and PC emerge as positive correlations, whereas SS, GDP, HLE, and FMC are negatively correlated. Mutual information underscores a strong dependency with SS.

For a more robust analysis, we extend our study from 2006 to 2022, with results depicted in figs. 3 to 6. These longitudinal findings consistently align with those observed in 2019, thereby reinforcing claim 1 and claim 2.

Regarding claim 1, we have obtained fig. 3. Throughout the examined years, GDP and HLE exhibit high correlation and mutual information, underscoring their strong interdependence. In contrast, G shows persistently low correlation and mutual information, indicating its limited statistical significance over time[4].

In evaluating claim 2, LL consistently shows significant correlations, echoing the patterns observed in 2019. SS and HLE emerge as the most robust positive correlates, followed by FMC and PA. Notably, A continues to show a positive correlation. Intriguingly, as time progresses, NA and D gain increasing significance.

For Positive Affect (PA), trends similar to 2019 persist, with distinctions in positive and negative correlations remaining stable. Yet, it's notable that in years such as 2006, 2007, 2009, 2011, and 2016, A correlates more positively with PA than HLE does, underscoring fluctuating patterns in emotional well-being indicators.

The analysis of NA unveils unique temporal patterns. Its correlations, unlike those of LL and PA, exhibit less significance initially, with a clear delineation of positive and negative correlations becoming more established post-2013. From this point, the trends mirror those observed in 2019, with A notably correlating positively with NA, suggesting nuanced dynamics in how negative emotions interact with other well-being factors over time.

B. Clustering

Having analyzed pairwise dependencies, we now explore how variables cluster. fig. 7 presents hierarchical clustering performed for 2019 using Pearson's distance matrix. Despite variations introduced by different algorithms, a consistent pattern emerges across all clustering analyses, regardless of the algorithm or dependency measure used. HLE and GDP are invariably the most closely associated variables, often grouped together, while G frequently appears as an outlier. LL typically clusters with either the HLE-GDP clade or SS. PA commonly associates with FMC, whereas A and PC are generally distant from other variables.

[4] As evidenced by fig. 2, correlations below 0.2 and mutual information values under 0.13 generally yield statistically insignificant results. This threshold has been consistently observed across the study period (2006-2022).

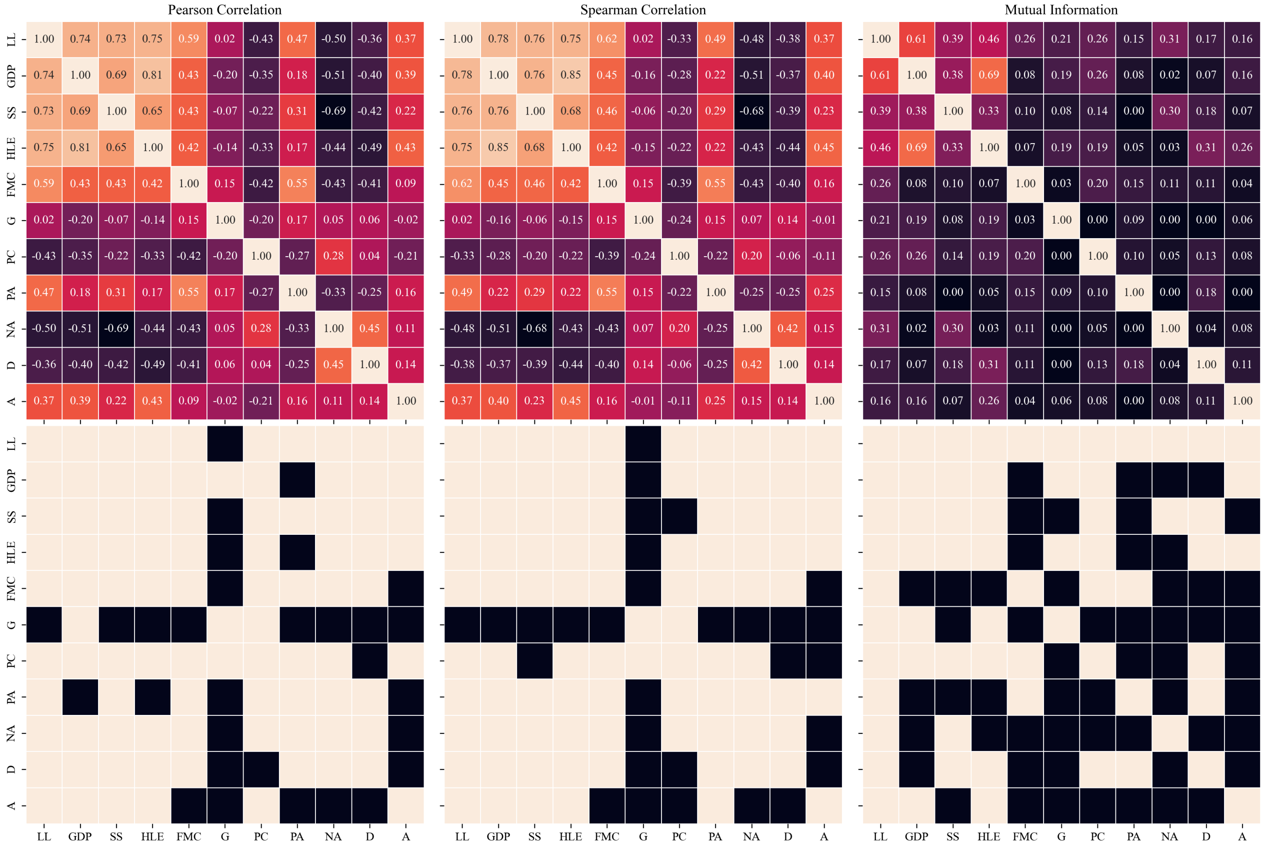


Figure 2: Correlation, mutual information, and p-value matrices for 2019. The top section presents correlation and mutual information matrices, while the bottom section displays binary p-value matrices from permutation tests. Light squared signify that the null hypothesis H_0 is rejected ($p < 0.05$), while dark squares indicate that H_0 cannot be rejected ($p > 0.05$).

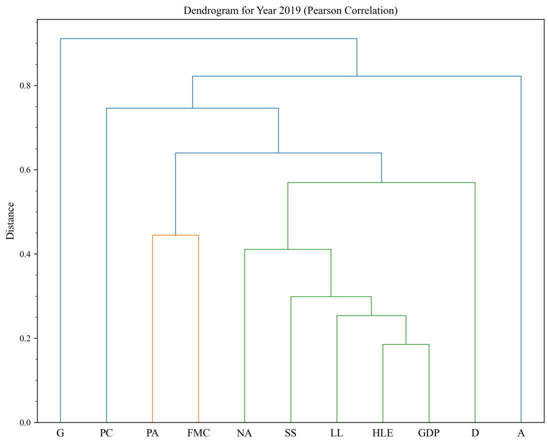


Figure 7: Hierarchical clustering of 2019 variables using the `'weighted'` method with Pearson's distance matrix eq. (B5). Clustering executed with `scipy`'s `linkage` function. The Cophenetic correlation coefficient of 0.88 indicates a high level of agreement between the dendrogram and the original distances.

Extended clustering analysis for the years 2006-2022 confirms these patterns. The Cophenetic correlation coefficient, ranging from 0.8 to 0.9, validates the clustering's accuracy, suggesting a reliable representation of the variable relationships.

IV. DISCUSSION

Before delving into our findings, it's imperative to address a fundamental limitation of our dataset: its relatively small sample sizes. Even in scenarios where the sample size reaches 3,000 individuals, for countries like the United States in 2022, this represents merely 0.0009% of its population. This proportion raises significant questions regarding the representativeness of the data for accurately reflecting a nation's well-being.

Our findings clearly demonstrate a significant interdependence between *Log GDP per capita* and *Healthy life expectancy at birth*, underscored by strong correlations and mutual information across various years. As detailed in tab. (I) for 2019, the remarkably low p-values decisively allow us to reject the null hypothesis of no depen-

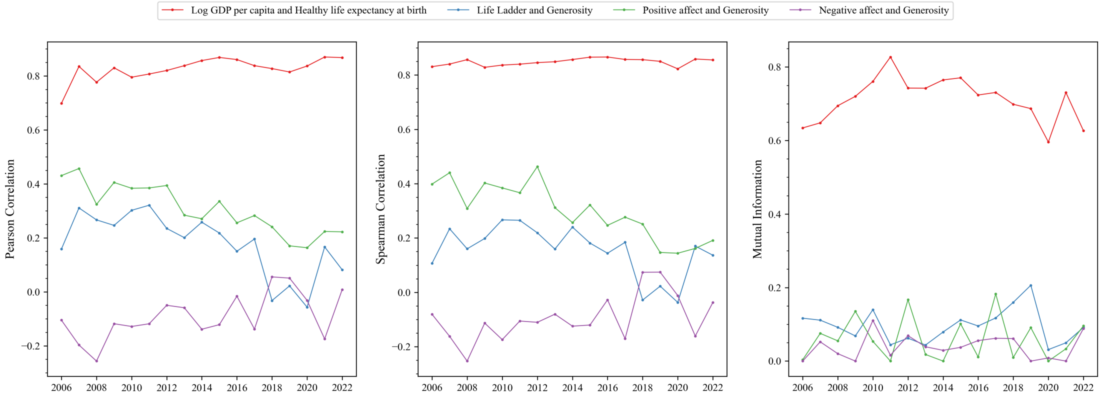


Figure 3: Time evolution of the pairwise dependencies regarding claim 1. In particular we plot GDP vs HLE and G vs LL, PA, and NA.

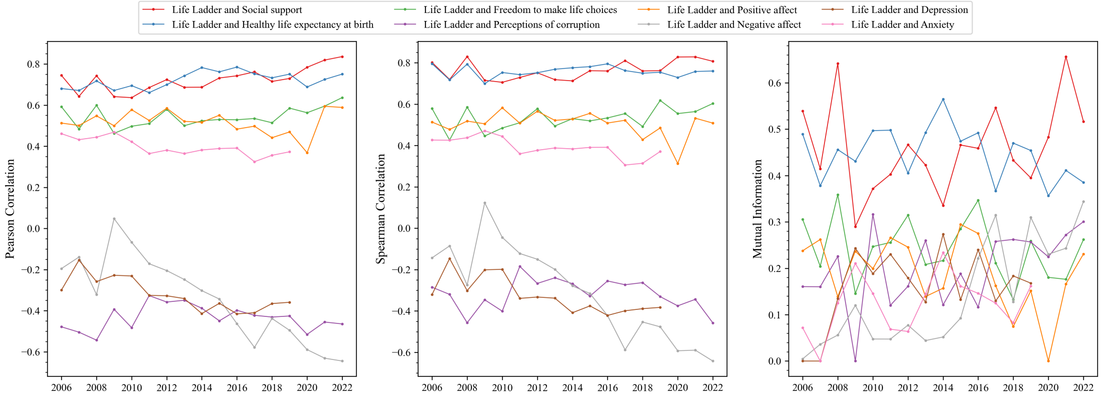


Figure 4: Time evolution of the pairwise dependencies regarding claim 2. In particular we plot the LL vs SS, HLE, FMC, PC, PA, NA, D, A.

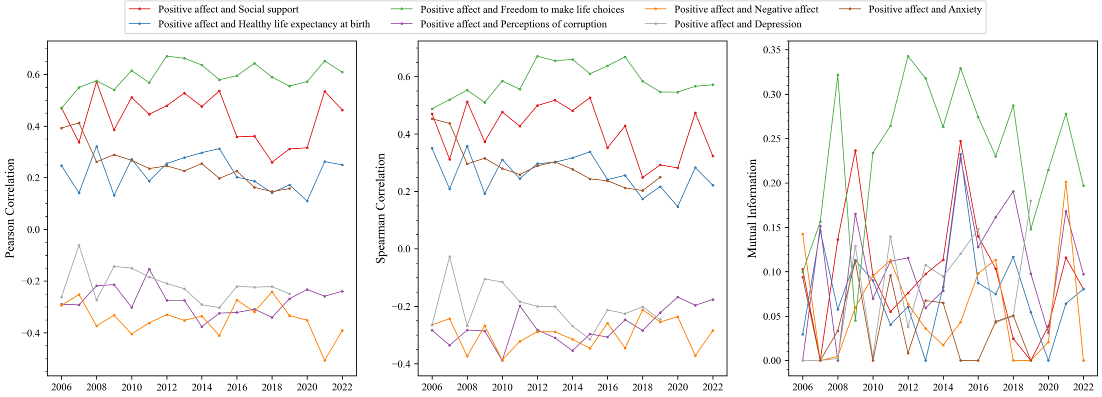


Figure 5: Time evolution of the pairwise dependencies regarding claim 2. In particular we plot the PA vs SS, HLE, FMC, PC, NA, D, A.

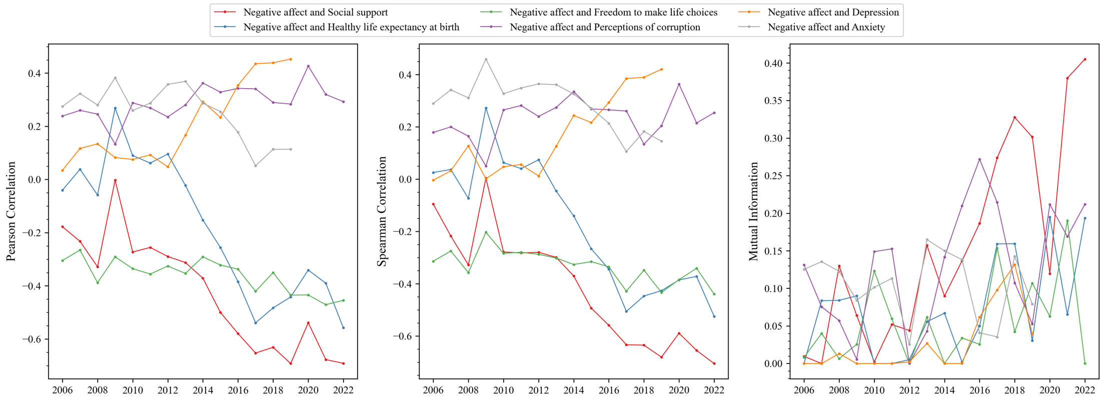


Figure 6: Time evolution of the pairwise dependencies regarding claim 2. In particular we plot the NA vs SS, HLE, FMC, PC, D, A.

dependency between these variables. Their correlations across time (fig. 3) consistently fall within the intervals of $r, \rho \in [0.80, 0.90]$ and mutual informations of $I \in [0.60, 0.85]$, marking them as among the most interdependent variables in our dataset. This pattern is reinforced by hierarchical clustering analysis, which consistently groups these variables as the most similar ones, with a Cophenetic correlation coefficient of 0.88 confirming a high fidelity between the clustering and the actual distances. Such consistent statistical significance across the board strongly suggests a great interdependence. This relationship implies a potential redundancy in utilizing both measures for assessing national well-being.

In contrast, as shown in figs. 2, 3 and 8, our findings indicate low statistical significance for the dependencies between the *Generosity* variable and well-being indicators. Tab. (II) underscores G's limited correlation with well-being, highlighting its significant mutual information with the *Life Ladder* as an anomaly rather than a consistent pattern (fig. 3). Despite a mutual information value of 0.21 hinting at some level of dependency, the practical significance of this measure in encapsulating national well-being remains debatable. Hierarchical clustering analysis further distinguishes *Generosity* as an outlier, emphasizing its unique position relative to other well-being variables. Moreover, fig. 9 illustrates the variable's erratic annual fluctuations across countries, casting doubt on its reliability and raising questions about the current methodology and definition of *Generosity* in the WHR. Such fluctuations suggest a misalignment with the genuine expression of altruism, especially considering the flawed logic that equates financial incapacity with selfishness[5]. This calls for a critical reevaluation of how *Generosity* is conceptualized and measured in subsequent reports.

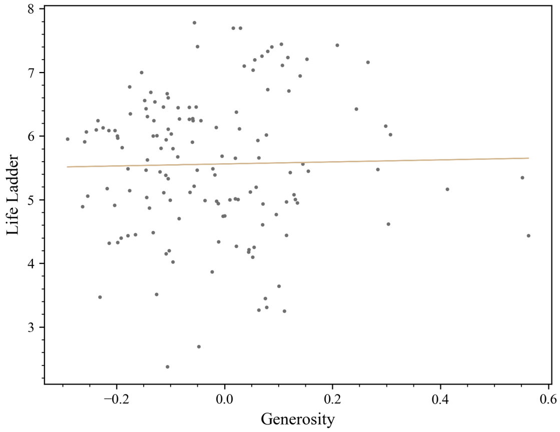
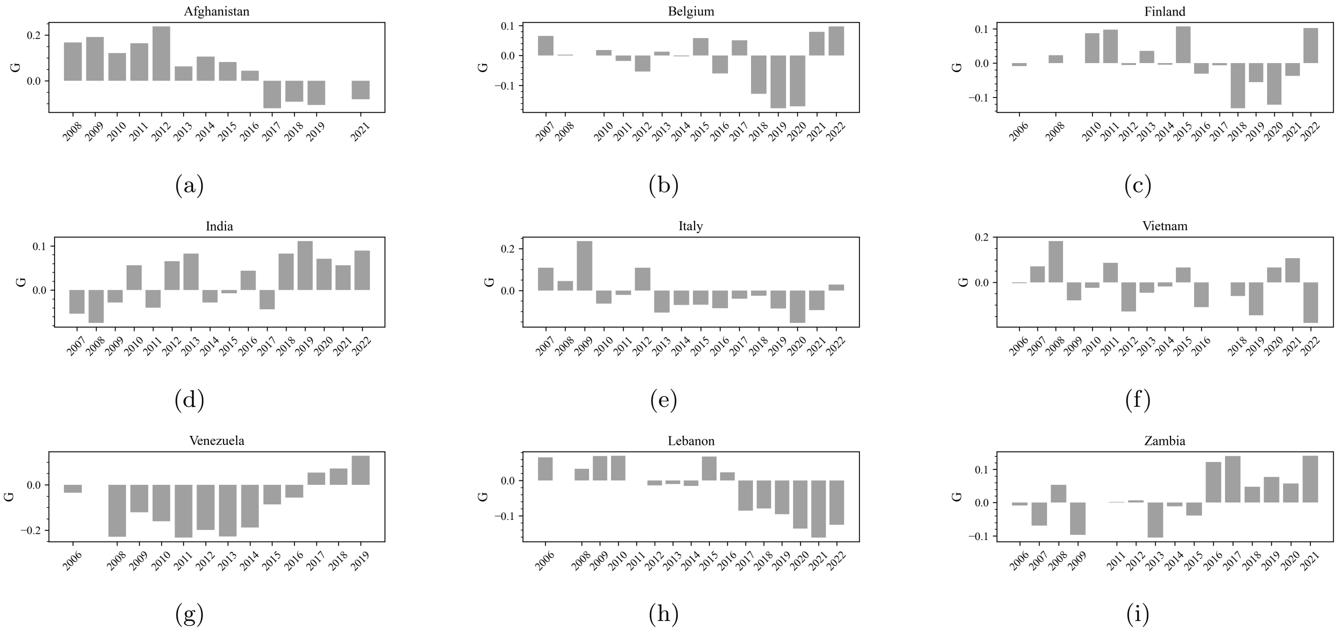


Figure 8: Scatter plot of the *Life Ladder* vs the *Generosity* for 2019. The red line denotes the linear fit. The Pearson's r correlation is 0.02 and the p-value is 0.79. The Spearman's ρ correlation is 0.02 and the p-value is 0.79. The mutual information I is 0.21 and the p-value is 0.001.

Our analysis confirms that the well-being variables align with our hypotheses, reinforcing their validity in capturing the essence of national happiness. Notably, hierarchical clustering analysis consistently groups the *Life Ladder* variable closely with *Social support* and *Healthy life expectancy at birth*, and *Positive affect* with *Freedom to make life choices.* This alignment corroborates the literature on the Science of Happiness, which identifies social support, health, and freedom as pivotal to happiness [14–16]. However, an intriguing anomaly arises with the *Anxiety* variable, which exhibits a positive correlation

[5] The current definition implies that an individual with no disposable income would, by default, be considered selfish, which is a fundamentally flawed conclusion.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

Figure 9: Time series of the *Generosity* scores for selected countries, highlighting the variable's erratic fluctuations over time. Criteria for country selection include geographic diversity and economic variability.

with both the *Happiness index* and *Positive affect*, diverging from expectations and lacking a significant correlation with *Negative affect*. This discrepancy from established happiness science prompts further scrutiny, especially given that *Anxiety* consistently emerges as an outlier in hierarchical clustering analysis, indicating its distinct behaviour from the other well-being metrics. This unexpected finding underlines the complexity of measuring well-being and the need for deeper investigation into how these variables reflect actual happiness and mental health conditions. Furthermore, these findings underscore the importance of a multifaceted approach to policy-making that considers a broad spectrum of well-being indicators. The anomalous behaviour of the *Anxiety* variable specifically suggests that interventions targeting national happiness need to address mental health complexities more directly.

V. CONCLUSIONS & FUTURE WORK

In conclusion, we emphasize the critical importance of sample size in the reliability of national happiness assessments. While we recognize the challenges associated with increasing sample sizes, particularly in terms of cost and logistical complexity, the benefits of a larger, more representative sample cannot be overstated. A more extensive sample would not only enhance the reliability of the results but also allow for a nuanced understanding of happiness across different demographic and cultural contexts. Furthermore, we advocate for a rigorous reevaluation of the measures and definitions used in quantifying well-being. This includes a thorough consideration of cultural influences, which can significantly affect the interpretation and reporting of well-being indicators [12, 13]. Addressing these measurement concerns is essential for advancing the science of happiness and ensuring that policy recommendations derived from such research are both accurate and culturally sensitive.

Building on our analysis, we highlight the pronounced interdependence between GDP and HLE. Given that HLE is derived and extrapolated, potentially introducing redundancy when considered alongside GDP, we suggest prioritizing GDP in future models. This approach not only simplifies the analysis by reducing the model's complexity but also mitigates the risk of multicollinearity, thereby enhancing the interpretability and reliability of happiness metrics. Adopting a more parsimonious model would facilitate clearer insights into the economic determinants of well-being, allowing for more focused and effective policy interventions.

Furthermore, our findings necessitate a reevaluation of the *Generosity* variable within the context of national happiness assessments. Recognizing the limitations of the current definition, which primarily hinges on financial donations, we advocate for a broader, more inclusive conceptualization of generosity. Alternative measures could include the rate of voluntarism, blood donation frequency, organ donation rates, and other forms of non-monetary contributions that reflect altruistic behaviours. These metrics offer a more holistic view of generosity, untethered from a country's economic wealth and potentially providing a more accurate reflection of societal well-being. Such an expanded definition would align more closely with the diverse expressions of altruism across cultures, contributing to a richer and more nuanced understanding of its impact on national happiness.

In summary, the well-being measures utilized in our analysis generally align with established findings within the *Science of Happiness* literature, validating their relevance and utility in understanding national happiness. Nevertheless, an exception arises with the *Anxiety* variable, whose relationship with other measures of well-being deviates from expected patterns. This inconsistency merits further investigation to uncover underlying factors and to refine our understanding of anxiety's impact on happiness. Delving deeper into this anomaly could shed light on complex dynamics between mental health conditions and perceived well-being, potentially guiding more effective strategies for enhancing national happiness.

Looking ahead, we identify two pivotal areas for future research to deepen our understanding of national happiness dynamics. First, a comprehensive time series analysis, incorporating both auto-correlation and cross-correlation techniques, promises to unravel the temporal patterns and relationships among the well-being variables. Despite the current limitations posed by sparse temporal data and the lack of significant correlations, this direction holds substantial potential for elucidating the dynamics of national happiness over time. Secondly, employing meta-analysis and machine learning methodologies to analyze the data could uncover new patterns and insights. Techniques such as pairwise correlations and clustering offer sophisticated tools for exploring the complex interplay between various factors influencing happiness. By leveraging these advanced analytical approaches, future studies can enhance our understanding of the underlying mechanisms of happiness and well-being on a national scale.

[1] D. M. McMahon, *Happiness: A History*, Happiness: A History (Atlantic Monthly Press) pp. xvi, 544.

[2] Y. Uchida and Y. Ogihara, Personal or Interpersonal Construal of Happiness: A Cultural Psychological Perspective, **2**.

[3] C. Peterson and M. E. P. Seligman, *Character Strengths and Virtues: A Handbook and Classification*, Character Strengths and Virtues: A Handbook and Classification (American Psychological Association) pp. xiv, 800.

[4] D. Keltner and E. Simon-Thomas, The Science of Happiness - BerkeleyX GG101x.

[5] *Well-Being: The Foundations of Hedonic Psychology*, Well-Being: The Foundations of Hedonic Psychology (Russell Sage Foundation) pp. xii, 593.

[6] E. Diener, Subjective well-being, **95**, 542.

[7] T. Chen, Experience Sampling Method.

[8] X. Wang and Z. Cheng, Cross-Sectional Studies: Strengths, Weaknesses, and Recommendations, An Overview of Study Design and Statistical Considerations, **158**, S65.

[9] E. J. Caruana, M. Roman, J. Hernández-Sánchez, and P. Solli, Longitudinal studies, **7**, E537, 26716051.

[10] J. F. Helliwell, R. Layard, J. D. Sachs, L. B. Aknin, J.-E. De Neve, and S. Wang, *World Happiness Report 2023*, 11th ed. (Sustainable Development Solutions Network).

[11] Bhutan's Gross National Happiness Index — OPHI.

[12] P. Dolan, L. Kudrna, and A. Stone, The Measure Matters: An Investigation of Evaluative and Experience-Based Measures of Wellbeing in Time Use Data, **134**, 57, 28983145.

[13] E. Simon-Thomas, Are World Happiness Rankings Culturally Biased?

[14] J. F. Helliwell and R. D. Putnam, The social context of well-being., **359**, 1435, 15347534.

[15] J. F. Helliwell, H. Huang, and S. Wang, Social Capital and Well-Being in Times of Crisis, **15**, 145 ().

[16] L. Mineo, Good genes are nice, but joy is better.

[17] Institute for Health Metrics and Evaluation (IHME). Global Burden of Disease (2020) – processed by Our World in Data.

[18] A. Gere, Recommendations for validating hierarchical clustering in consumer sensory projects, **6**, 100522.

[19] J. F. Helliwell, H. Huang, M. Norton, S. Wang, and L. Goff, Statistical Appendix for "World happiness, trust and social connections in times of crisis," Chapter 2 of World Happiness Report 2023.

[20] World Health Organization (WHO).

[21] G. Inc, How Does the Gallup World Poll Work?

[22] Correlation.

[23] Y. Dodge, *The Concise Encyclopedia of Statistics*, 1st ed., Springer Reference (Springer).

[24] T. M. Cover and J. A. Thomas, *ELEMENTS OF INFORMATION THEORY*.

[25] L. Paninski, Estimation of Entropy and Mutual Information, **15**, 1191.

[26] J. De Gregorio, D. Sanchez, and R. Toral, Entropy estimators for Markovian sequences: A comparative analysis, **26**, 79, 2310.07547.

[27] A. Levina, V. Priesemann, and J. Zierenberg, Tackling the subsampling problem to infer collective properties from limited data, **4**, 770.

Appendix A: Databases

In this research, we leverage data from two principal sources to construct a comprehensive dataset for our analysis: the **World Happiness Report** and **Our World in Data**. Below, we detail the datasets extracted from these sources, and the variables included in our study.

1. World Happiness Report

The World Happiness Report (WHR) is an annual publication from the United Nations Sustainable Development Solutions Network. It ranks countries worldwide by their happiness levels, assessed through various well-being indicators. For our study, we utilized the "Data for Table 2.1" from the World Happiness Report 2023, which encompasses a comprehensive assessment of national happiness based on survey results and socioeconomic data.

The dataset is organized into rows and columns, where each row represents a specific country-year observation, facilitating a temporal analysis of happiness trends. The first column identifies the country, and the second specifies the year of observation. The rest of the columns are the following:

- **Life Ladder** or **Happiness Score**: this variable represents the national average response to the Cantril Ladder question, serving as a quantifier of Subjective Well-Being (SWB) across countries. Individuals are asked: "*Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?*". This approach to measuring SWB is based on data from the Gallup World Poll (GWP), offering a direct subjective assessment of personal well-being.

- **Log GDP per capita**: denotes the natural logarithm of the Gross Domestic Product (GDP) per capita. The data sources are the World Development Indicators (WDI) and the Penn World Table (PWT) 10.01 for Taiwan, Syria, Palestine, Venezuela, Djibouti and Yemen. Values for 2022 are forecasted. The values of the countries from the PWT 10.01 end at 2019. The 2020-2022 GDP values are based on the 2019 values and the projected growth rates if they are available (see [19] for details).

- **Social support**: is the national average of the binary responses (either 0 or 1) to the GWP question "*If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?*". This variable highlights the importance of interpersonal relationships and community in contributing to individual happiness.

- **Healthy life expectancy at birth**: estimates the average number of years a newborn is expected to live in "full health", considering the impact of diseases and injuries [20]. The data source is the World Health Organization's (WHO) Global Health Observatory data repository. Data is available for years 2000, 2010, 2015 and 2019. To match this report's sample period (2005-2021), interpolation and extrapolation are used. This metric underscores the critical relationship between health and well-being.

- **Freedom to make life choices**: is the national average of the binary responses to the GWP question "*Are you satisfied or dissatisfied with your freedom to choose what you do with your life?*".

- **Generosity**: is the residual of regressing national average of the response to the GWP question "*Have you donated money to a charity in the past month?*" on GDP per capita. Notice that the generosity measure is not the survey response to the question, but the residual of the regression of the survey response on GDP per capita. This gives rise to negative values for generosity, which means that the survey response is lower than what would be expected given the country's GDP per capita.

- **Perceptions of corruption**: measures the national average of the survey responses to two questions in the GWP: "*Is corruption widespread throughout the government or not?*" and "*Is corruption widespread within businesses or not?*". The overall perception is just the average of the two binary responses. In case the perception of government corruption is missing, we use the perception of business corruption as the overall perception. The corruption perception at the national level is just the average response of the overall perception at the individual level.

- **Positive affect**: is the average of three positive affect measures in the GWP: laugh, enjoyment and doing interesting things. These measures are the responses to the following three questions, respectively: "*Did you smile or laugh a lot yesterday?*", and "*Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Enjoyment?*", "*Did you learn or do something interesting yesterday?*". This composite measure provides a multifaceted view of positive emotional experiences, contributing to a broader understanding of happiness.

- **Negative affect**: is the average of three negative affect measures in the GWP: worry, sadness and anger. These measures are the responses to the following three questions, respectively: "*Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Worry?*", "*Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Sadness?*", "*Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Anger?*". It serves as an indicator of the prevalence of negative emotional states, offering a counterbalance to the Positive Affect measure in assessing overall well-being.

The World Happiness Report incorporates the Gallup World Poll's (GWP) country coding scheme, WP5, which includes several sub-country territories to ensure a comprehensive global analysis. The GWP's methodology targets the adult population, specifically those aged 15 and older, reflecting a broad demographic scope. The typical sample size ranges from 500 to 2000 respondents per country each year (see Table 1 from [19] for further details). For a comprehensive understanding of the sampling methodology, refer to [21].

2. Our World in Data

Our World in Data (OWID) serves as an invaluable resource for global statistical analysis, offering accessible, comprehensive datasets on a wide range of topics, including health, education, and environmental factors. Among its extensive collections, the Mental Health section provides critical insights into the prevalence and impact of mental health disorders worldwide, contributing to our understanding of their relationship with overall happiness and well-being.

For the purposes of this study, we have utilized data from the second chart in the Mental Health section, as documented by the [17]. This particular dataset offers a detailed overview of the prevalence rates of various mental health disorders across the globe. To align with the temporal scope of the World Happiness Report data, our analysis focuses on the period from 2005 to 2019. Specifically, we have concentrated on two major conditions:

- **Depression**: defined according to the Diagnostic and Statistical Manual of Mental Disorders (DSM) and the International Classification of Diseases (ICD), depression is estimated as the total number of individuals diagnosed with depressive disorders, expressed as a proportion of the national population. This metric provides insight into the burden of depression on societies.

- **Anxiety**: similarly based on DSM and ICD definitions, anxiety is quantified as the total number of individuals diagnosed with anxiety disorders relative to the country's population. This measure sheds light on the prevalence of anxiety disorders.

These variables, Depression and Anxiety, are critical for our analysis as they offer a direct measure of mental health challenges faced by populations worldwide. By incorporating these metrics into our study, we aim to explore the intricate relationships between mental health disorders and subjective well-being, thereby enriching our investigation into the determinants of happiness.

Appendix B: Statistical Methods

In this section, we outline the statistical methods applied to our study, including dependency analysis through correlation and mutual information, hierarchical clustering to group variables by similarity, and hypothesis testing to assess statistical significance.

a. Dependency Analysis

In this report, we performed a dependency analysis to elucidate the relationships among the random variables under investigation. **Correlation**, quantifies the relationship's strength and direction between two random variables X, Y, irrespective of causality. To quantify these relationships, we calculated *Correlation Coefficients* for our dataset $\{(x_i, y_i) \mid i \in 1, \ldots, n\}$, where n represents the sample size [22]. This study focuses on several types of Correlation Coefficients, providing a comprehensive view of the interdependencies among variables. In this work we have considered the following *Correlation Coefficients*:

- **Pearson's correlation coefficient** (r) quantifies the linear relationship between two variables. It has a value between -1 and 1, where 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation. The sample correlation coefficient is given by:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{B1}$$

  where $\bar{x}$ and $\bar{y}$ are the sample means of x and y, respectively [23].

- **Spearman rank correlation coefficient** (ρ) evaluates the strength and direction of association between two ranked variables. This non-parametric metric quantifies statistical dependence by determining how well a monotonic function can describe the relationship between two variables. The sample Spearman rank correlation coefficient is given by:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}, \tag{B2}$$

  where d_i is the difference between the ranks of corresponding variables x and y [23].

To detect more nuanced dependencies, we also employed the **Mutual Information** (I), which is a measure of the amount of information that one random variable contains about another random variable [24].

$$I(X;Y) = \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} p(x,y) \log\left[\frac{p(x,y)}{p(x)p(y)}\right], \tag{B3}$$

where $p(x,y)$ represents the joint probability distribution of X and Y, and $p(x)$ and $p(y)$ are their respective marginal probability distributions. $\mathcal{X}$ and $\mathcal{Y}$ denote the sets of possible values for X and Y. Unlike correlation coefficients, mutual information is unbounded but constrained within:

$$0 \leq I(X;Y) \leq \min\{H(X), H(Y)\}, \tag{B4}$$

where $H(X)$ and $H(Y)$ are the **Entropies** of X and Y, respectively.

In practice, the challenge often arises from a lack of knowledge about the probability distributions of random variables X, Y, due to either mathematical complexities or insufficient understanding of the underlying experimental details. When this occurs, direct computation of mutual information using eq. (B3) becomes unfeasible, necessitating the use of approximate numerical methods. These computational approaches are known to introduce bias [25]. However, as a general guideline, when the dataset size significantly exceeds the number of possible outcomes, accurate estimation of I becomes more straightforward. In such contexts, most popular estimators perform adequately [26, 27].

b. Clustering Analysis

Clustering analysis is pivotal when our interest shifts from examining pairwise dependencies to understanding how variables group together based on similarity. This process involves categorizing objects so that those within the same cluster exhibit greater resemblance to each other than to those in different clusters.

Among various clustering techniques, Hierarchical Clustering has been our method of choice, specifically its Agglomerative approach. This bottom-up strategy begins with each observation as a separate cluster, progressively merging pairs of clusters to ascend the hierarchy. As *Distance Matrix* D we have considered:

$$D = \mathbb{I} - |C|, \tag{B5}$$

where $\mathbb{I}$ is the identity matrix and $|C|$ is the absolute value of the correlation matrix.

The choice of clustering algorithm significantly influences the analysis outcome. In our case, WPGMA algorithm was selected for its suitability to our data's characteristics. This method leverages the weighted average distance between clusters during the merging process. Furthermore, we employed the Cophenetic correlation coefficient to evaluate the dendrogram's accuracy in reflecting the original data points' pairwise distances. A coefficient value closer to 1 indicates a high fidelity of the dendrogram to the original distances, providing a robust measure of the clustering quality.

c. Hypothesis Testing

Hypothesis testing is a cornerstone of statistical analysis, offering a structured approach to evaluate theories about the world through data. This formal method helps in making informed decisions by testing the significance of statistical measures such as correlations and mutual information within a dataset.

Central to our analysis is the calculation of the *p-value* for both correlations and mutual information. The **p-value** quantifies the likelihood that the observed statistical measures could arise by chance alone. A p-value below a predetermined significance threshold indicates strong evidence against the null hypothesis, suggesting that the observed relationship is statistically significant. In this study, we adopted a significance level of 0.05, aligning with common statistical practice.

The calculation of the p-value often employs the t-distribution, predicated on the assumption that the data adhere to a normal distribution. This presumption is underpinned by the *Central Limit Theorem*, which holds for sufficiently large sample sizes. To confirm this normality assumption, we utilized the *Kolmogorov-Smirnov Test*, a statistical procedure designed to evaluate the null hypothesis that a sample is drawn from a normal distribution. This test provides a rigorous assessment of the sample's conformity to normality.

However, our analysis revealed that not all of our variables followed a normal distribution, compelling us to apply the *Permutation Test* for evaluating the significance of observed dependencies. Unlike parametric tests that assume a specific distribution, the Permutation Test is *non-parametric*, abstaining from normality assumptions. This makes it exceptionally flexible for analyzing various data distributions, albeit at the cost of computational efficiency due to its reliance on the random reshuffling of observed data to test the hypothesis.