UNIVERSITÀ DEGLI STUDI DI CATANIA
Dipartimento di Economia ed Impresa
Master's degree in Data Science for Management

Ashish Singh BISHT

Spatiotemporal patterns and correlates of psoriasis: an epidemiological analysis at the global level

Master's thesis

Supervisors: Prof. Andrea Giuseppe Maugeri
Co-supervisor: Prof. Giuliana Favara

Academic Year 2024 - 2025

कर्मण्येवाधिकारस्ते
मा फलेषु कदाचन ।
मा कर्मफलहेतुर्भूः
मा ते सङ्गोऽस्त्वकर्मणि॥

"You have the right to perform your duty,
but not to the fruits thereof.
Let not the results of your actions be your motive,
nor let your attachment be to inaction."

— Bhagavad Gita 2.47

Contents

Chapter 1

Introduction

1 Introduction and Motivation

Psoriasis is a chronic inflammatory skin condition that significantly affects the quality of life of people and poses substantial challenges to healthcare systems around the world. Understanding its global distribution and associated factors is crucial for developing effective public health strategies. According to WHO (World Health Organization), 2-3 percent of the global population is affected by this skin condition, which is approximately 125 million people worldwide, with significant variations across regions and demographics. The disease is characterized by recurrent episodes of remission and exacerbation, leading to physical discomfort, psychological distress, and reduced quality of life for those affected. Given its substantial burden on individuals and healthcare systems, understanding the spatio-temporal patterns and correlates of psoriasis is critical to guide public health strategies. During the past few decades, there has been growing recognition of the importance of analyzing long-term trends to uncover the global distribution and burden of psoriasis. Recent studies, such as those from the Global Burden of Disease initiative, have highlighted trends in psoriasis incidence and prevalence across different regions. Although some regions report a decline in incidence, others face a steady or rising burden, underscoring the need for comprehensive and detailed epidemiological analyses. This thesis examines global spatiotemporal patterns of psoriasis from 1990 to 2021, leveraging extensive datasets to explore trends and identify associated factors. By focusing on this timeframe, the study aims to provide a deeper understanding of how the burden of psoriasis has evolved over three decades, considering variations in regional and national contexts. Psoriasis is more than a medical condition; it profoundly affects the lives of those who experience it, disrupting their physical well-

being, emotional stability, and social interactions. As someone who personally knows individuals impacted by this chronic disease, I have witnessed the challenges they face, from the physical discomfort of symptoms to the stigma and misconceptions surrounding the condition. This personal connection has deepened my resolve to contribute meaningfully to the understanding and management of psoriasis. On a global scale, psoriasis remains a significant public health concern, with its burden extending far beyond the affected individuals to their families, communities, and healthcare systems. Despite its prevalence, many aspects of the disease, including its geographical distribution, risk factors, and trends over time, are not fully understood. Addressing these gaps is essential to improve outcomes and quality of life for those living with psoriasis. By analyzing global spatiotemporal patterns through measures such as Disability-Adjusted Life Years (DALYs), Prevalence, Incidence, and Years Lived with Disability (YLDs), this thesis aims to shed light on the evolution of this disease over three decades. This work aspires to provide actionable insights that can support targeted public health interventions, raise awareness, and ultimately reduce the burden of psoriasis on individuals and society.

Chapter 2

Dataset Preparation

1 Dataset Description

This study integrates two large-scale datasets to explore the global burden of psoriasis and its association with a wide range of developmental and contextual factors. The primary epidemiological data were obtained from the *Institute for Health Metrics and Evaluation (IHME)* as part of the *Global Burden of Disease (GBD)* study [1], while the contextual indicators were sourced from the *World Bank Open Data* repository [2].

The combined dataset spans the years **1990 to 2021** and covers more than 200 countries. It includes key psoriasis-related health metrics namely, Prevalence, Incidence, Years Lived with Disability (YLDs), and Disability-Adjusted Life Years (DALYs) as well as a wide array of socioeconomic, environmental, and demographic indicators.

Prior to analysis, several preprocessing steps were undertaken to ensure data quality, consistency, and compatibility across sources. These steps included temporal alignment, missing data handling, column standardization, country name harmonization, and feature selection based on correlation analysis. The following sections describe the datasets and preprocessing pipeline in detail.

2 Datasets and Preprocessing

This study utilizes two primary datasets to explore the association between the burden of psoriasis and a broad range of development indicators across countries and over time.

2.1 Dataset 1: Psoriasis Burden Indicators

The first dataset was obtained from the *Global Burden of Disease (GBD)* study [1]. It provides comprehensive epidemiological metrics for psoriasis, including:

- `measure_name` (DALYs, YLDs, Incidence, Prevalence),
- `location_name` (Country name),
- `sex_name`, `age_name`,
- `cause_name` (Psoriasis),
- `metric_name`, `year`, `val`, `upper`, `lower`.

The dataset spans the years **1990 to 2021** and includes over 200 countries. For consistency, only records corresponding to *both sexes* and *all age groups combined* were retained.

2.2 Dataset 2: Development Indicators

The second dataset was sourced from the *World Bank Open Data* repository [2], initially comprising 389 socioeconomic, environmental, and demographic indicators. Examples include:

- Access to electricity (% of population) [EG.ELC.ACCS.ZS],
- CO_2 emissions (metric tons per capita) [EN.ATM.CO2E.PC],
- Urban population growth (annual %) [SP.URB.GROW],
- Fertility rate, sanitation coverage, HIV prevalence, and water productivity indicators.

This dataset also spans the years **1990 to 2021**.

2.3 Preprocessing Steps

The following preprocessing steps were applied to harmonize and prepare the data for analysis:

1. **Temporal Alignment:** Both datasets were filtered to cover the common period from 1990 to 2021.

2. **Zero Value Treatment:** All zero values that were likely placeholders for missing data (i.e., exact zeros in integer form) were replaced with `NaN`. This was applied conditionally to avoid affecting valid decimal values such as 0.0 or actual measurements of zero.

3. **Column Standardization:** Columns such as `location_name` and `Time` were renamed to `Country Name` and `year` to ensure consistency before merging.

4. **Country Name Balancing:** Country name mismatch between the datasets (e.g., "United States" vs. "United States of America") were manually resolved. As a result, all countries were successfully retained post-merging.

5. **Data Merging:** The datasets were merged using an `inner join` on `Country Name` and `Year`. This yielded a fully aligned dataset with complete data for all overlapping countries and years.

6. **Indicator Selection:** A fast correlation analysis was performed to identify indicators strongly associated with the four psoriasis metrics (DALYs, YLDs, Incidence, Prevalence). Based on the strength of correlation and data completeness, a set of **74 development indicators** was selected for dimensionality reduction and downstream analysis.

7. **Missing Data Handling:** From the development indicators dataset, all columns with more than 25% missing values were removed to ensure analytical robustness. This reduced dimensionality while preserving information quality, resulting in leaving a final set of **67 development indicators** .

This preprocessing pipeline ensured the resulting dataset maintained wide geographic coverage including all major countries and high data integrity, making it suitable for further statistical analysis.

Chapter 3

Exploratory Data Analysis

1 Purpose and Aim

Exploratory Data Analysis (EDA) is a crucial step in understanding the structure and characteristics of a dataset before conducting advanced statistical modeling. In this study, we explore the global spatiotemporal patterns and contextual correlates of psoriasis burden using four key epidemiological measures: Prevalence, Incidence, Years Lived with Disability (YLDs), and Disability-Adjusted Life Years (DALYs).

The aim of the EDA is to uncover underlying trends, variations, and relationships within the dataset. Through summary statistics, visualizations, and comparative analyses, we identify patterns across countries and time periods. These initial insights provide critical context and help guide subsequent analyses, including dimension reduction, clustering, and regression modeling.

2 General Dataset Overview

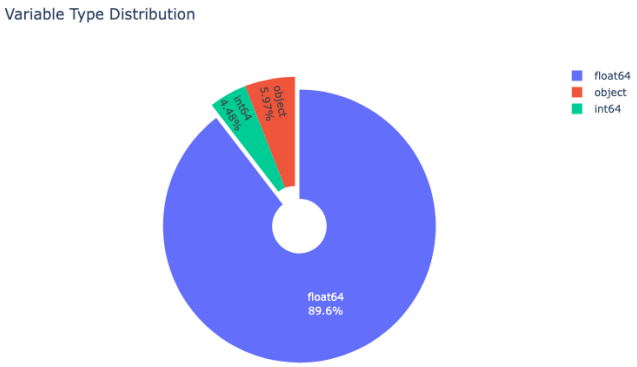
The merged dataset used in this study contains a total of **14,400** rows and **67 columns**, representing country-year observations from **200 countries** over a **32-year** period **(1990 to 2021)**. The dataset includes four primary psoriasis burden measures Prevalence, Incidence, Years Lived with Disability (YLDs), and Disability Adjusted Life Years (DALYs) as well as a broad range of contextual development indicators sourced from global databases such as the World Bank.

Among the 67 variables, **60** are numeric and reflect diverse factors such as demographic structure, access to electricity, sanitation, fertility rate, greenhouse gas emissions, and urbanization levels. The remaining variables are categorical, including country names, codes, and measure identifiers.

Missing values were present in several indicators, particularly those related to environmental metrics, sanitation, and healthcare access. For example, variables such as the tuberculosis case detection rate and energy intensity had over **2,000 missing entries**, primarily concentrated in earlier years or lower-income countries. However, core outcome variables and most key predictors were largely complete, supporting robust cross-country and longitudinal analyses.

This comprehensive dataset provides a solid foundation for subsequent exploratory, dimensional, and inferential analyses focused on understanding the structural determinants of psoriasis burden globally.

2.1 Variable Type Distribution



w

Figure 3.1: Variable type distribution in the dataset. Most variables are numeric, enabling advanced statistical and machine learning analysis.

This plot summarizes the data types of all columns in the dataset. The vast majority of variables (90 percent) are continuous numeric (float64), representing quantitative development indicators. A smaller proportion are categorical (object), including country names and codes, while a few are integers (int64), used for identifiers like year and location. This heavy numeric composition is ideal for performing statistical analyses like PCA, clustering, and regression.

2.2 Top Variables with Missing Values

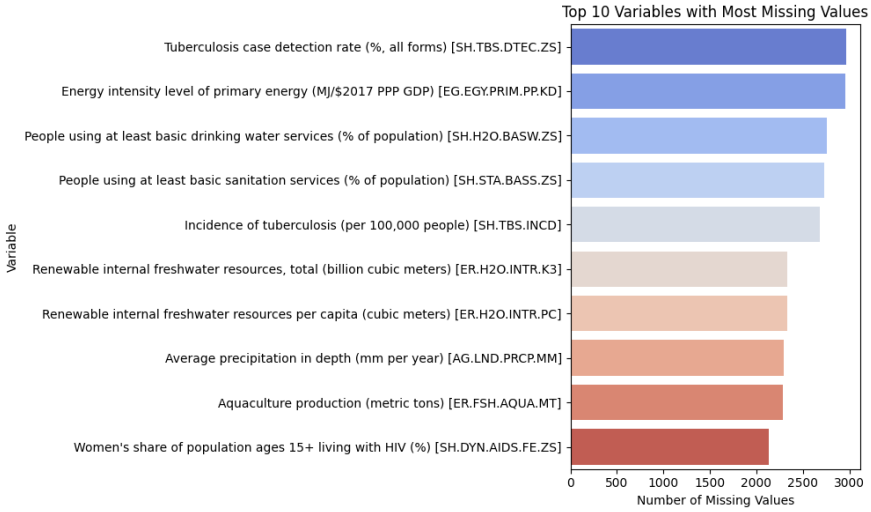


Figure 3.2: Top 10 variables with the highest number of missing values. These were considered during preprocessing and some were removed from multivariate analysis.

This horizontal bar chart displays the ten variables with the highest number of missing entries. Indicators such as tuberculosis case detection rate and energy intensity have nearly 3,000 missing values, mainly due to inconsistent country-level reporting over time. Environmental and resource-based indicators also suffer from sparsity. These insights help guide variable selection and imputation strategies in later stages of analysis.

2.3 Distribution of Psoriasis Burden Measures

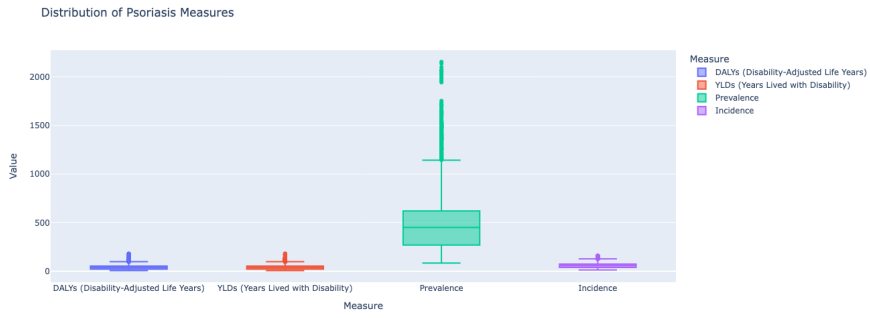


Figure 3.3: Boxplot showing the distribution of four psoriasis burden indicators. Prevalence shows a wider range compared to DALYs, YLDs, and Incidence.

This boxplot illustrates the distribution of values for the four primary psoriasis burden metrics across all countries and years. Prevalence shows the widest spread and largest outliers, indicating high variability in reported cases globally. DALYs and YLDs have more compact distributions, while Incidence falls in between. These differences highlight the importance of analyzing each metric separately, as they capture distinct dimensions of disease burden.

3 Univariate and Bivariate Analysis

Univariate analysis involves examining each variable individually to summarize and find patterns in the data. In this section, we present the distribution and descriptive statistics of both the psoriasis burden measures and selected potential influencing indicators.

3.1 Distribution of Psoriasis Burden Across Countries

To visualize the distribution of psoriasis burden across countries:

- A scatter plot (Figure 3.4) displays the burden across different countries.

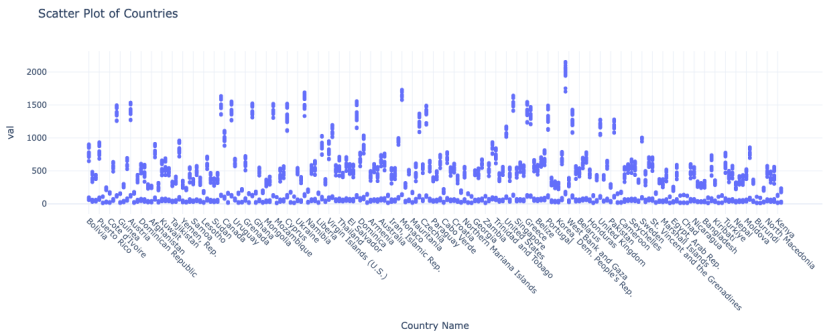
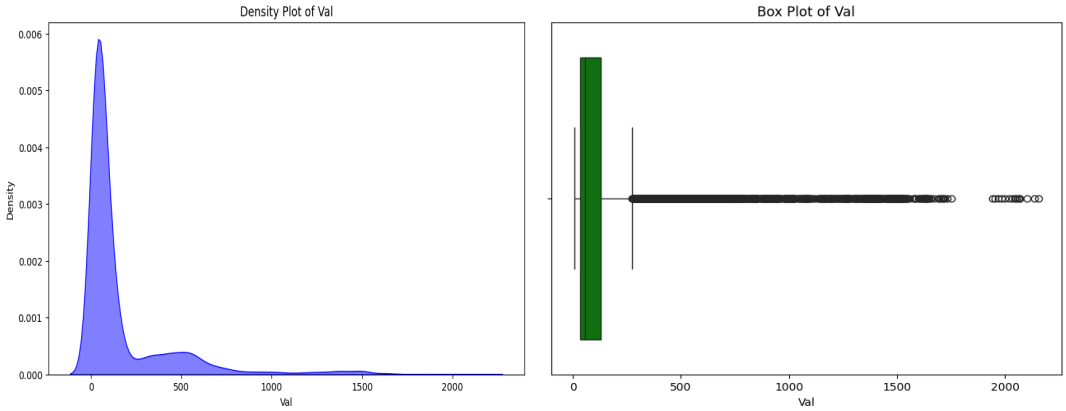


Figure 3.4: Scatter plot showing the distribution of psoriasis burden across countries. Each point represents a value for a country in a given year.

Figure 3.4 provides an overview of psoriasis-related values across countries over time. It highlights the variability in reported values between countries and years and serves as an initial check for data distribution and completeness before further analysis.

- A density plot and a box plot (Figure 3.5) reveal the distribution of burden values and help identify potential outliers.



(a) Density plot of psoriasis burden values.

(b) Boxplot of psoriasis burden highlighting outliers.

Figure 3.5: Distribution of psoriasis burden: (a) Density plot showing the spread of values; (b) Boxplot identifying extreme values and outliers.

Figure 3.5 provides a detailed view of the distribution of psoriasis burden values across countries:

- Panel (a) shows a density plot that illustrates the overall shape of the distribution. The distribution is highly right-skewed, indicating that most countries report lower burden values, while a small number report extremely high values. The presence of multiple peaks may suggest clustering of burden levels among different groups of countries.
- Panel (b) presents a boxplot that highlights the statistical spread of the values. A long upper tail and numerous outliers confirm the presence of countries with unusually high burden. The median lies closer to the lower end of the range, reinforcing the observed skewness.

Together, these plots illustrate the uneven global distribution of psoriasis burden, highlighting disparities across countries. This motivates further investigation into the underlying factors driving these differences.

3.2 Descriptive Statistics of Independent Indicators

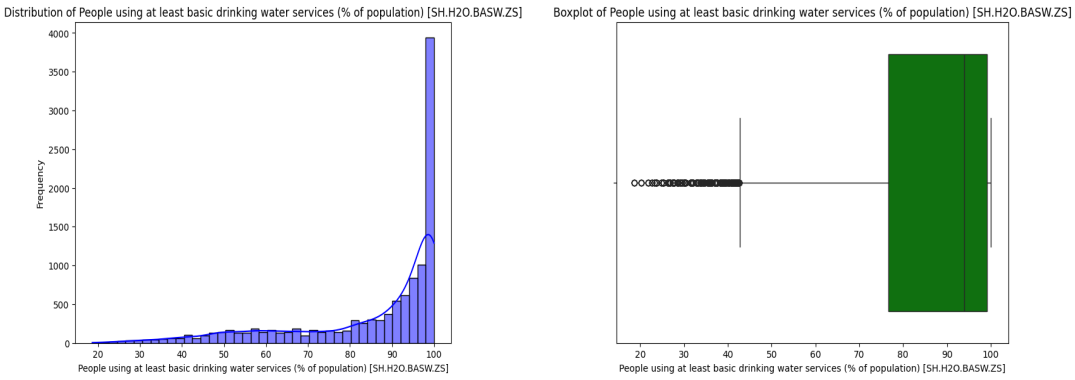
We examined several indicators to understand their distribution across countries. For each, we computed descriptive statistics and plotted histograms and boxplots.

3.3 People using at least basic drinking water services

- Mean, min, max, and standard deviation values were computed (Table 3.1).
- The distribution is visualized in Figures 3.6a and 3.6b.

Table 3.1: Descriptive statistics: Drinking water access (%)

Statistic	Value
Mean	85.214013
Median	94.011081
Min	18.682303
Max	100.000000
Std Dev	18.527754



(a) Histogram and KDE of access to drinking water.

(b) Boxplot of access to drinking water.

Figure 3.6: Distribution of access to basic drinking water services: (a) Histogram with KDE showing skewed concentration near full access; (b) Boxplot highlighting outliers and variability among countries.

Figures 3.6a and 3.6b illustrate the distribution of access to at least basic drinking water services across countries.

Panel (a) shows a strong right skew, with a large concentration of countries having access levels near 100%. However, the presence of a long left tail and several bars at lower values indicate that a number of countries still experience limited access.

Panel (b) confirms this observation, with several outliers on the lower end representing countries where access is significantly below the global

average. The median access is above 90%, but the wide interquartile range and standard deviation of 18.46 reflect variability across the dataset.

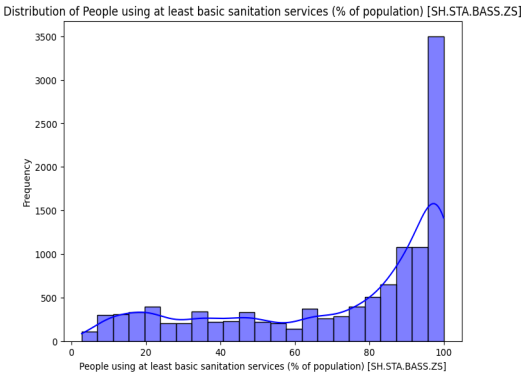
Overall, while most countries appear to have high access to drinking water, a non-negligible number still face accessibility challenges.

3.4 People using at least basic sanitation services

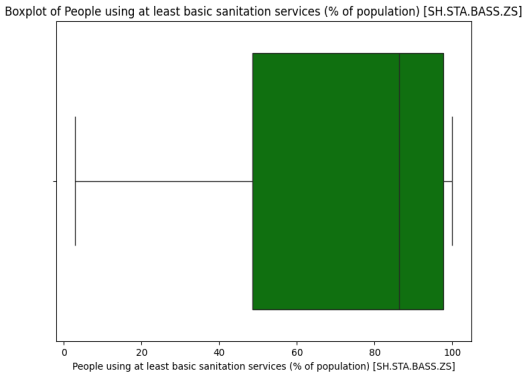
- Mean, min, max, and standard deviation values were computed (Table 3.2).
- The distribution is visualized in Figure 3.7.

Table 3.2: Descriptive statistics: Sanitation service access (%)

Statistic	Value
Mean	72.266440
Median	86.341588
Min	2.793897
Max	100.000000
Std Dev	29.820264



(a) Histogram and KDE of access to sanitation services.



(b) Boxplot of access to sanitation services.

Figure 3.7: Distribution of access to basic sanitation services: (a) Histogram with KDE showing overall distribution; (b) Boxplot showing the spread and identification of countries with low access.

Figure 3.7 presents the distribution of access to at least basic sanitation services across countries.

Panel (a) shows that the distribution is heavily skewed to the right, with a large cluster of countries having sanitation coverage near or above 90%. However, a notable portion of the data lies in the lower access range, highlighting disparities in global sanitation infrastructure.

Panel (b), the boxplot, emphasizes the wide spread in the data. The interquartile range spans from moderate to high access levels, but a significant number of outliers appear on the lower end, representing countries where access remains critically low. These countries likely face systemic challenges related to public health infrastructure and investment.

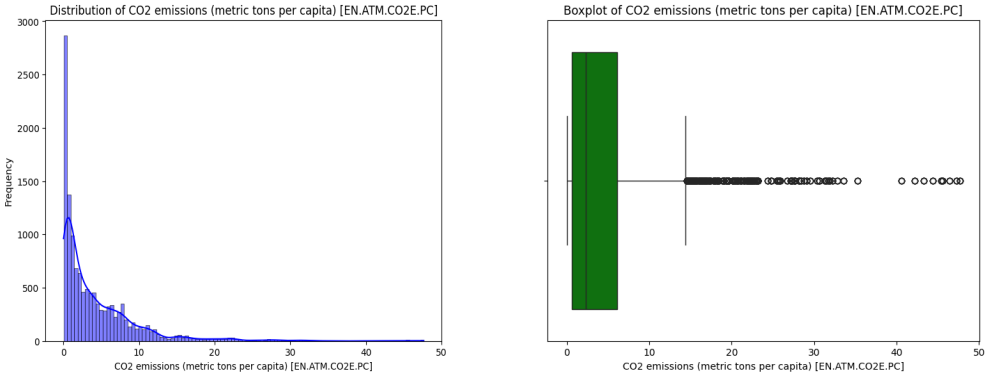
The descriptive statistics support these observations: while the median is high (86.5%), the standard deviation of 29.93% indicates a broad variability. The minimum value is as low as 2.8%, further highlighting inequality in access.

3.5 CO2 emissions (metric tons per capita)

- Mean, min, max, and standard deviation values were computed (Table 3.3).
- The distribution is visualized in Figures 3.8a and 3.8b.

Table 3.3: Descriptive statistics: CO2 emissions (%)

Statistic	Value
Mean	4.251203
Median	2.294078
Min	0.021790
Max	47.656962
Std Dev	5.462401



(a) Histogram and KDE of CO_2 emissions. **(b)** Boxplot of CO_2 emissions.

Figure 3.8: Distribution of CO_2 emissions (metric tons per capita): (a) Histogram and KDE showing a right-skewed distribution; (b) Boxplot highlighting outliers and emission variability across countries.

Figure 3.8 and Table 3.3 provide a descriptive overview of CO_2 emissions per capita across countries.

The histogram and KDE in panel (a) show a highly right-skewed distribution. Most countries emit relatively low levels of CO_2 per person, but a few countries have extremely high emissions, as indicated by the long tail to the right.

This pattern is confirmed by the boxplot (panel (b)), which reveals a compressed interquartile range and a large number of high-value outliers. These high-emitting countries significantly influence the global average, as seen in the gap between the mean (4.35) and the median (2.33) in Table 3.3. The wide standard deviation (5.62) further reflects the unequal distribution of per capita emissions worldwide.

3.6 Age dependency ratio (percentage of working-age population)

- Mean, min, max, and standard deviation values were computed (Table 3.4).
- The distribution is visualized in Figures 3.9a and 3.9b.

Table 3.4: Descriptive statistics: Age dep ratio (%)

Statistic	Value
Mean	63.531967
Median	57.369833
Min	17.282537
Max	116.874114
Std Dev	18.858794

(a) Histogram and KDE of age dependency ratio.

(b) Boxplot of age dependency ratio.

Figure 3.9: Distribution of age dependency ratio (% of working-age population): (a) Histogram with KDE showing central tendency and variation; (b) Boxplot identifying spread and potential outliers across countries.

Figure 3.9 and Table 3.4 summarize the global distribution of the age dependency ratio, which represents the proportion of dependents (young and elderly) relative to the working-age population.

Panel (a) presents the histogram and KDE. The distribution appears moderately right-skewed, with most countries clustering around 50% to 70%. The peak occurs near 55–60%, which is consistent with the median value of 56.75% reported in Table 3.4. However, the distribution includes countries with very low and very high ratios, extending from a minimum of 17.28% to a maximum of 116.14%, highlighting the wide demographic differences between regions.

Panel (b), the boxplot, shows a broad interquartile range and a fairly symmetrical spread of the data, with no extreme outliers. The central box confirms that half of the countries fall between roughly 45% and 80%, indicating moderate to high dependency burdens in many regions.

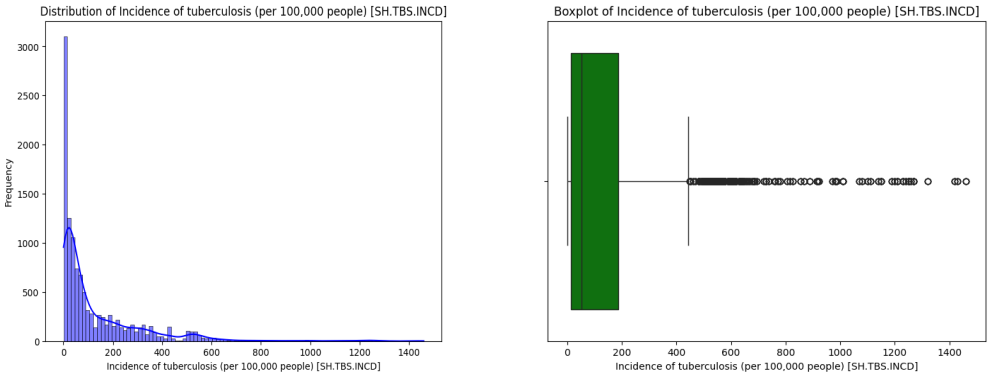
These patterns reflect variations in population structure—such as aging populations in developed countries or high youth dependency in low-income nations—making age dependency an important contextual factor for understanding national health burdens.

3.7 Incidence of tuberculosis (per 100,000 people)

- Mean, min, max, and standard deviation values were computed (Table 3.5).
- The distribution is visualized in Figures 3.10a and 3.10b.

Table 3.5: Descriptive statistics: Incidence of Tuberculosis (%)

Statistic	Value
Mean	132.731689
Median	53.000000
Min	0.410000
Max	1460.000000
Std Dev	188.118896



(a) Histogram and KDE of incidence of tuberculosis.

(b) Boxplot of incidence of tuberculosis.

Figure 3.10: Distribution of tuberculosis incidence (per 100,000 people): (a) Histogram and KDE showing a heavily right-skewed distribution; (b) Boxplot revealing a high number of outliers in countries with elevated disease burden.

Figure **??** presents the distribution of tuberculosis incidence across countries, measured per 100,000 people.

Panel (a) shows a histogram with KDE overlay, revealing a highly right-skewed distribution. While the median incidence is 52, the mean is much higher at approximately 133, indicating the influence of extreme values. A small number of countries report significantly elevated incidence rates, with the maximum reaching 1460 cases per 100,000 people.

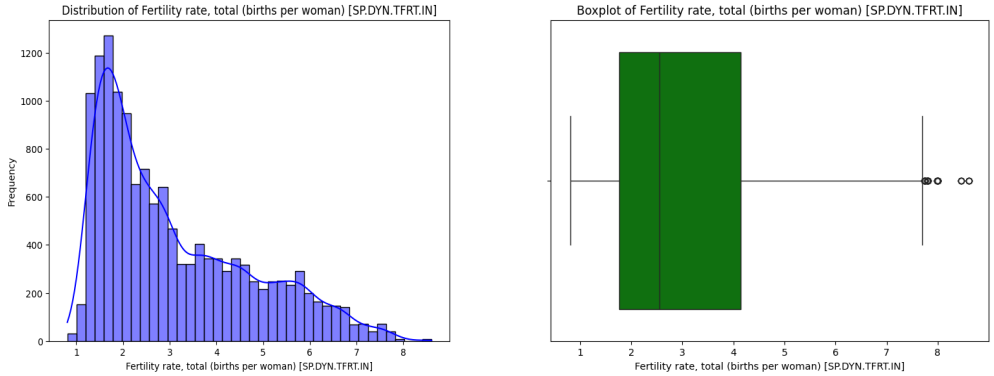
Panel (b), the boxplot, clearly highlights these high outliers, demonstrating considerable variability in tuberculosis burden across regions. This underscores the unequal global distribution of tuberculosis, with some countries still facing major public health challenges.

3.8 Fertility rate, total (births per woman)

- Mean, min, max, and standard deviation values were computed (Table **??**).
- The distribution is visualized in Figures 3.11a and 3.11b.

Table 3.6: Descriptive statistics: Fertility rate (%)

Statistic	Value
Mean	3.084060
Median	2.554000
Min	0.808000
Max	8.606000
Std Dev	1.624846



(a) Histogram and KDE of fertility rate.

(b) Boxplot of fertility rate.

Figure 3.11: Distribution of fertility rate (births per woman): (a) Histogram and KDE showing a right-skewed distribution with a concentration around 2–3 births per woman; (b) Boxplot highlighting higher fertility rate outliers in certain countries.

Figure 3.11 visualizes the distribution of fertility rate across countries, measured in total births per woman.

Panel (a) displays a histogram with a KDE overlay, showing that the distribution is right-skewed. Most countries have a fertility rate between 2 and 3, but a notable number extend beyond 5, with a few reaching 8 births per woman. This indicates high variability in fertility behavior globally, influenced by socioeconomic, cultural, and healthcare-related factors.

Panel (b) presents the corresponding boxplot, which confirms the presence of significant outliers at the higher end. These outliers suggest that while many countries have undergone demographic transition, some—particularly in lower-income regions—still experience high fertility rates.

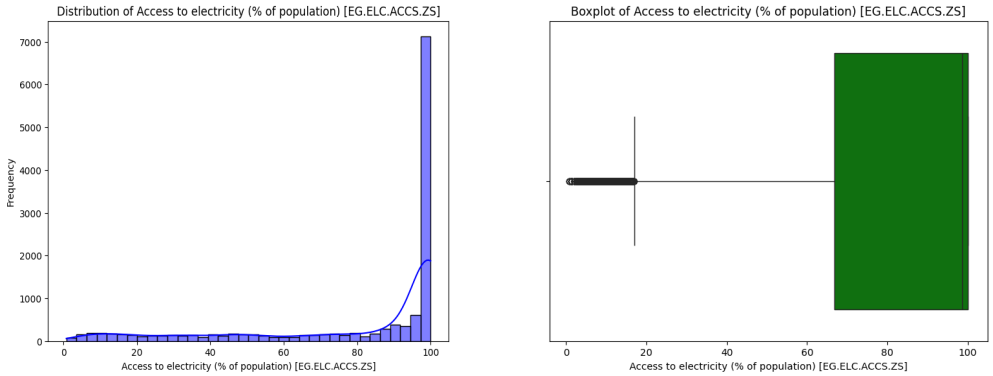
This visualization supports understanding of population growth trends and their implications for healthcare and development.

3.9 Access to electricity (percentage of population)

- Mean, min, max, and standard deviation values were computed (Table 3.7).
- The distribution is visualized in Figures 3.12a and 3.12b.

Table 3.7: Descriptive statistics: Access to electricity (%)

Statistic	Value
Mean	80.066850
Median	98.623154
Min	0.800000
Max	100.000000
Std Dev	29.775420



(a) Histogram and KDE of access to electricity.

(b) Boxplot of access to electricity.

Figure 3.12: Distribution of access to electricity (percentage of population): (a) Histogram and KDE showing most countries near full access with some having significantly lower access; (b) Boxplot highlighting inequality and extreme outliers.

Figure 3.12 illustrates the distribution of access to electricity across countries, measured as the percentage of the population with access.

Panel (a) shows a histogram with a KDE overlay, revealing a **left-skewed distribution**. The **median access** is high at 98.6%, while the **mean is lower**, approximately 79.98%, indicating that although many countries have nearly universal access, there is a tail of countries with significantly lower levels. This is also supported by the wide range, with a minimum value of just 0.8% and a standard deviation of about 30.

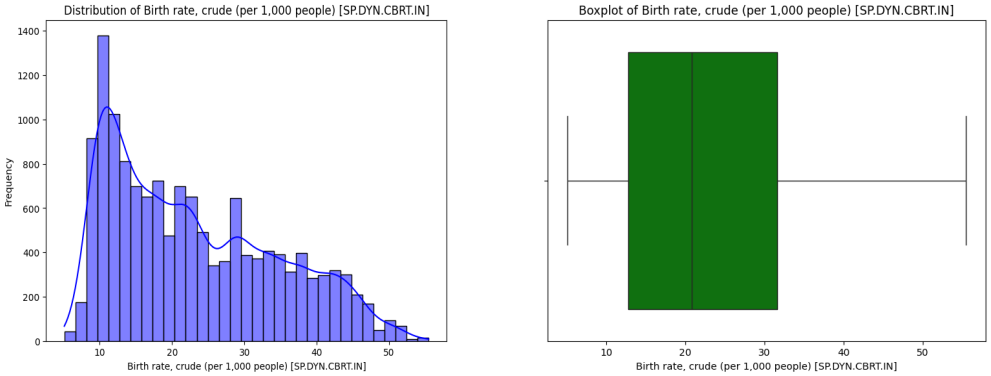
Panel (b), the boxplot, highlights these countries with **low access as outliers**, showing substantial inequality in electricity availability worldwide. The figure reflects both widespread success in electrification for some regions and ongoing infrastructure gaps in others.

3.10 Birth rate, crude (per 1,000 people)

- Mean, min, max, and standard deviation values were computed (Table 3.8).
- The distribution is visualized in Figures 3.13a and 3.13b.

Table 3.8: Descriptive statistics: Birth rate (%)

Statistic	Value
Mean	23.014072
Median	20.804000
Min	5.100000
Max	55.467000
Std Dev	11.517120



(a) Histogram and KDE of birth rate.

(b) Boxplot of birth rate.

Figure 3.13: Distribution of crude birth rate (per 1,000 people): (a) Histogram with KDE showing a right-skewed distribution and concentration around the median; (b) Boxplot reflecting overall spread and variability across countries.

Figure 3.13 illustrates the distribution of crude birth rates across countries, measured as births per 1,000 people.

Panel (a) displays a histogram with a KDE overlay, revealing a moderately **right-skewed distribution**. The **mean birth rate** is approximately 22.83, slightly higher than the **median of 20.65**, suggesting the presence of countries with notably higher birth rates pulling the average up. The wide spread, with values ranging from 5.9 to 55.47 and a standard deviation of 11.55, reflects substantial variation in fertility patterns globally.

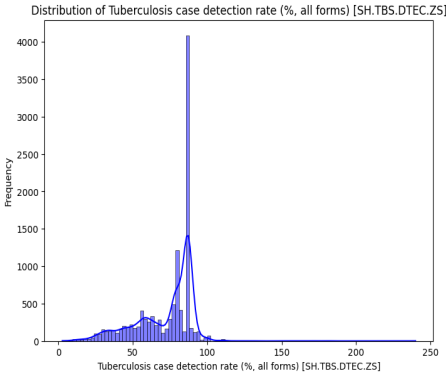
Panel (b), the boxplot, confirms this variation and shows some dispersion without prominent extreme outliers. Overall, the figures highlight the demographic diversity between countries, where some maintain high fertility rates while others have transitioned to much lower levels.

3.11 Tuberculosis case detection rate (percentage all forms)

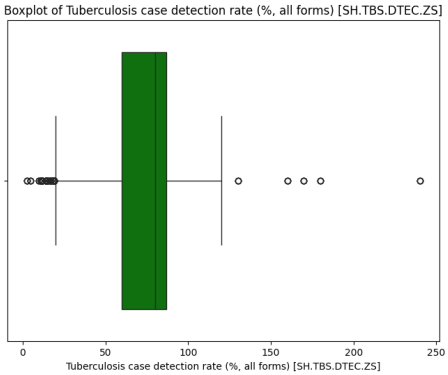
- Mean, min, max, and standard deviation values were computed (Table 3.9).
- The distribution is visualized in Figures 3.14a and 3.14b.

Table 3.9: Descriptive statistics: Tuberculosis case detection rate (%)

Statistic	Value
Mean	72.549042
Median	80.000000
Min	2.600000
Max	240.000000
Std Dev	19.344532



(a) Histogram and KDE of tuberculosis case detection rate.



(b) Boxplot of tuberculosis case detection rate.

Figure 3.14: Distribution of tuberculosis case detection rate (%, all forms): (a) Histogram with KDE indicating a narrow peak with heavy tails; (b) Boxplot highlighting mild and extreme outliers.

Figure 3.14 illustrates the distribution of tuberculosis case detection rate across countries, expressed as the percentage of all estimated cases detected.

Panel (a), the histogram with KDE overlay, shows a distribution centered around 80%, with a noticeable peak indicating that many countries cluster near this value. The presence of extreme values—ranging from as low as 2.6% to as high as 240%—creates a long-tailed, slightly right-skewed shape. The mean (72.5%) is lower than the median (80%), suggesting some downward pull from countries with limited detection capacity.

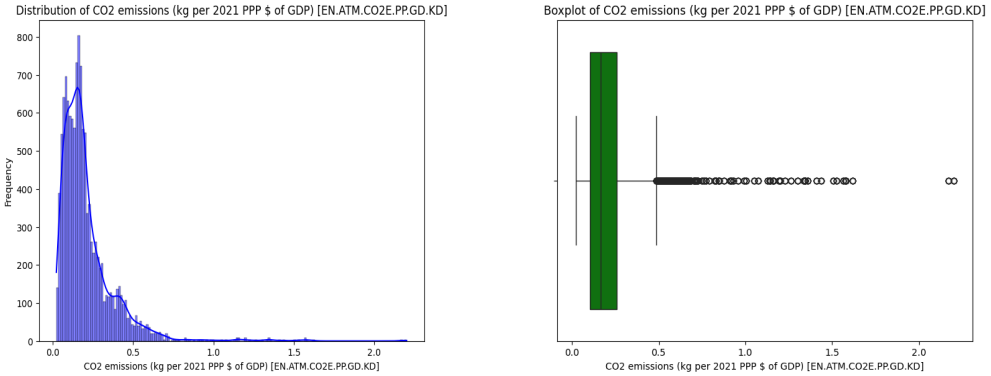
Panel (b), the boxplot, reveals the presence of both mild and extreme outliers, reflecting variability in national reporting and surveillance quality. These figures highlight the global disparity in case detection efforts, emphasizing the need for improved diagnostic reach and healthcare access in low-performing regions.

3.12 CO2 emissions (kg per 2021 PPP of GDP)

- Mean, min, max, and standard deviation values were computed (Table 3.10).
- The distribution is visualized in Figures 3.15a and 3.15b.

Table 3.10: Descriptive statistics: CO2 emissions (kg prt 2021 PPP of GDP)(%)

Statistic	Value
Mean	0.211273
Median	0.163845
Min	0.022513
Max	2.201322
Std Dev	0.183243



(a) Histogram and KDE of CO_2 emissions per unit of GDP.

(b) Boxplot of CO_2 emissions per unit of GDP.

Figure 3.15: Distribution of CO_2 emissions per unit of GDP (kg per 2021 PPP $ of GDP): (a) Histogram and KDE revealing heavy right skew; (b) Boxplot showing high emission outliers.

Figure 3.15 illustrates the distribution of CO_2 emissions per unit of GDP (in kilograms per 2021 PPP $ of GDP), capturing carbon intensity of economic activity across countries.

Panel (a), the histogram with KDE overlay, reveals a strong right-skewed distribution. Most countries emit below 0.5 kg per unit GDP, as indicated by the median of 0.16 kg. However, the mean is slightly higher (0.21 kg), suggesting that a few countries with high carbon intensity are pulling the average upward. The distribution has a long tail extending beyond 2 kg per unit GDP.

Panel (b), the boxplot, confirms the presence of substantial outliers, highlighting countries with disproportionately high CO_2 emissions relative to economic output. This suggests notable disparities in energy efficiency and the carbon footprint of production across nations.

3.13 Visual Relationship Between Indicators and Psoriasis Measures

Although correlation and regression will be addressed later, here we present scatter plots showing how each indicator relates visually to the psoriasis burden metrics (Prevalence, Incidence, DALYs, YLDs).

For each indicator, we plotted the following against the four burden measures:

- People using basic drinking water services

- People using basic sanitation services
- Access to electricity
- CO_2 emissions (per capita and per GDP)
- Age dependency ratio
- Incidence of tuberculosis
- Fertility rate, total (births per women)
- Birth rate, crude (per 1,000 people)

3.13.1 Prevalence

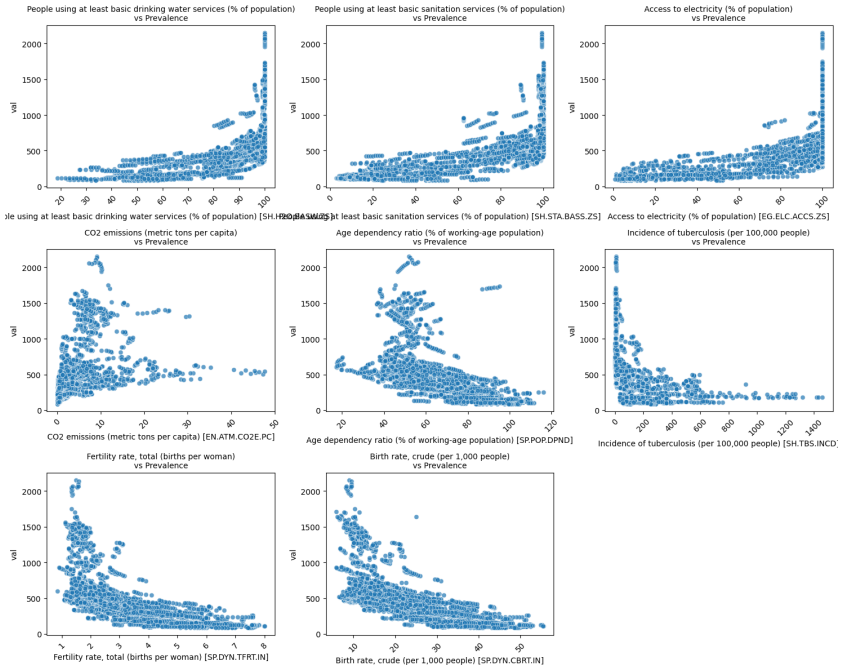


Figure 3.16: Indicators vs Prevalence

Figure 3.16 illustrates the bivariate relationships between selected indicators and the prevalence of psoriasis across countries.

A few visible patterns emerge from the scatter plots:

- There appears to be a **positive association** between prevalence and access-related indicators such as access to basic drinking water, sanitation, and electricity. Countries with higher infrastructure access tend to report higher psoriasis prevalence, potentially due to better healthcare access and diagnosis rates.
- Conversely, indicators such as **CO_2 emissions per capita**, **fertility rate**, and **birth rate** show a **negative trend**, suggesting that higher prevalence may be more common in countries with lower fertility and birth rates, typically more developed nations.
- The **age dependency ratio** and **incidence of tuberculosis** both show an inverse, non-linear relationship with prevalence. Countries with higher dependency ratios or tuberculosis burden generally report lower psoriasis prevalence, possibly due to competing health priorities or underreporting.

3.13.2 Incidence

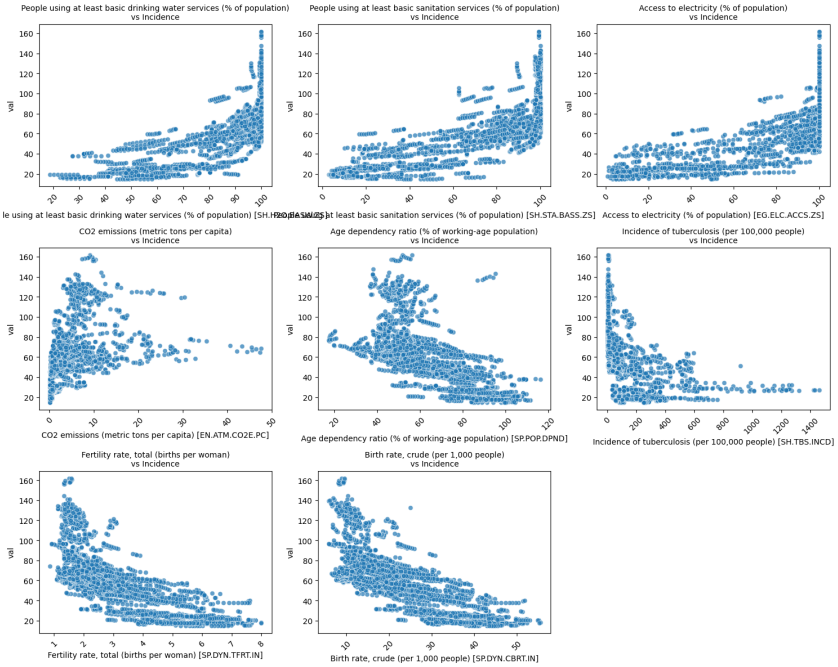


Figure 3.17: Indicators vs Incidence

Figure 3.17 displays the scatter plots of key contextual indicators against the incidence of psoriasis (measured per 100,000 people). These visualizations reveal several notable trends:

- A **positive relationship** is observed between incidence and access-related indicators (basic drinking water, sanitation, electricity). As access improves, reported incidence tends to increase possibly due to enhanced healthcare infrastructure and diagnostic capabilities in more developed regions.
- **Negative associations** are seen with indicators like **fertility rate, birth rate,** and **incidence of tuberculosis**. Countries with lower fertility and infectious disease burdens tend to have higher reported psoriasis incidence, supporting a link between development and non-communicable disease recognition.
- The relationship with **CO_2 emissions** and **age dependency ratio** appears more dispersed, though some clustering suggests that middle-income countries might experience moderate incidence values.

3.13.3 DALYs(Disability-Adjusted Life Years)

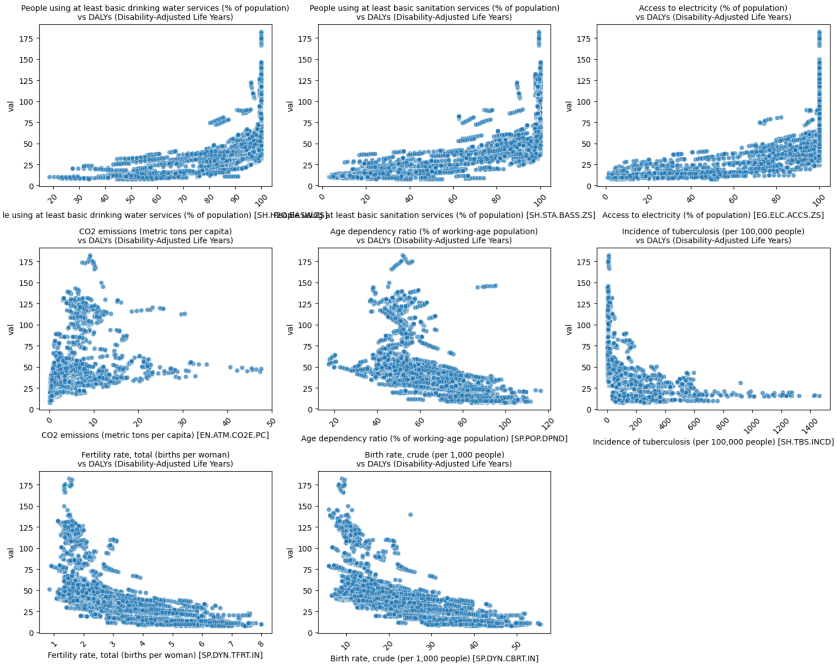


Figure 3.18: Indicators vs. DALYs

Figure 3.18 illustrates the bivariate relationships between selected contextual indicators and the DALYs (Disability-Adjusted Life Years) attributable to psoriasis across countries.

A few visible patterns emerge from the scatter plots:

- There appears to be a **positive association** between DALYs and access-related indicators such as access to basic drinking water, sanitation, and electricity. Countries with better access to these services tend to report higher DALYs, potentially reflecting stronger healthcare infrastructure and improved chronic disease reporting.

- In contrast, indicators such as **CO_2 emissions per capita**, **fertility rate**, and **birth rate** exhibit a **negative relationship** with DALYs. This pattern suggests that countries with higher environmental emissions or lower population growth rates (typically more developed) experience a greater psoriasis burden in terms of disability.

- The **age dependency ratio** and **incidence of tuberculosis** also show a non-linear inverse pattern, where countries with higher dependency

ratios or infectious disease burdens tend to report lower psoriasis DALYs possibly due to prioritization of acute health issues or diagnostic limitations.

3.13.4 YLDs(Years Lived with Disability)

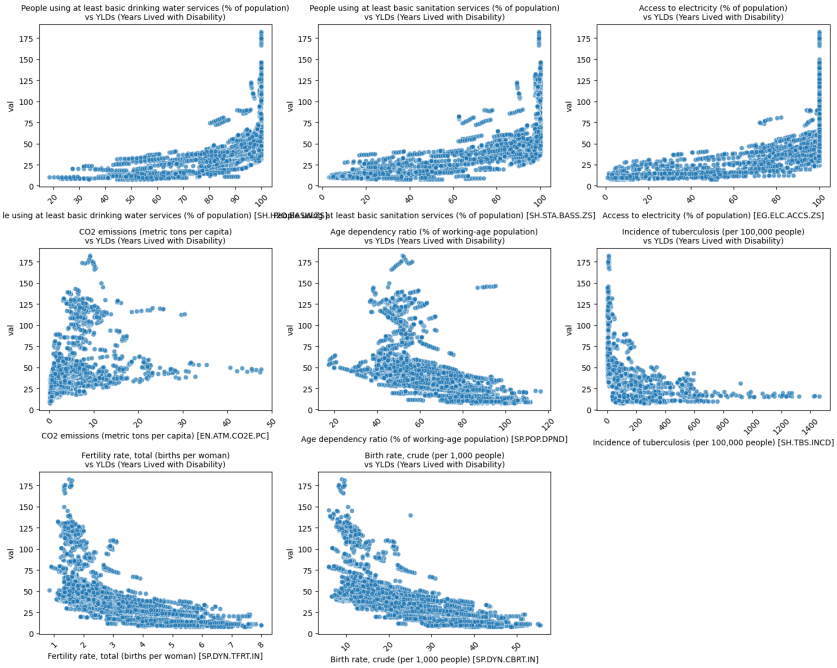


Figure 3.19: Indicators vs. YLDs

Figure 3.19 displays the scatter plots depicting the visual relationships between selected contextual indicators and YLDs (Years Lived with Disability) due to psoriasis.

Several trends are visible in these bivariate plots:

- A clear **positive association** is observed between YLDs and infrastructural indicators like access to basic drinking water, sanitation, and electricity. This suggests that countries with improved living standards and healthcare systems may report more disability cases from chronic conditions like psoriasis.

- **Negative relationships** are observed between YLDs and both **fertility** and **birth rates**. This aligns with the broader epidemiological trend where chronic non-communicable diseases tend to be more prevalent in countries with lower population growth and aging demographics.

- Indicators such as **CO_2 emissions per capita**, the **age dependency ratio**, and the **incidence of tuberculosis** show inverse or non-linear associations, possibly reflecting differences in disease prioritization, demographic structure, or reporting infrastructure.

The scatter plots presented above offer preliminary visual evidence of systematic relationships between contextual indicators and the burden of psoriasis (Prevalence, Incidence, DALYs, and YLDs).

Across all four burden metrics, a recurring pattern emerges: countries with greater access to basic services (water, sanitation, electricity) tend to report higher psoriasis burdens, likely due to better healthcare access and diagnostic capabilities. Conversely, high fertility, high birth rates, and elevated infectious disease indicators (e.g., tuberculosis incidence) tend to be associated with lower psoriasis metrics, consistent with the underreporting or lower prioritization of chronic non-communicable diseases in less developed contexts.

While these scatterplots suggest important trends, visual inspection alone cannot establish the strength or significance of these relationships. Therefore, in the next section, we conduct a more formal pca, correlation analysis followed by multivariable regression modeling. These techniques will help quantify associations, control for confounding factors, and identify the most influential predictors of psoriasis burden globally.

4 Time-Series Imputation

After the completion of univariate and bivariate analyses, some indicators still contained missing values scattered across years for a small subset of countries. To address this and preserve the temporal continuity of the dataset, a time-aware imputation strategy was applied.

- **LOCF (Last Observation Carried Forward):** For a missing value X_t, if a previous non-missing value X_{t-k} exists, then:

$$X_t = X_{t-k} \quad \text{where } k = \min\{k > 0 : X_{t-k} \text{ is observed}\}$$

- **NOCB (Next Observation Carried Backward):** If no earlier value exists for imputation (i.e., missing values at the start), the next available value X_{t+k} is used:

$$X_t = X_{t+k} \quad \text{where } k = \min\{k > 0 : X_{t+k} \text{ is observed}\}$$

The combined LOCF+NOCB approach ensures that every missing value is replaced using the nearest available observation in time:

$$X_t = \begin{cases} X_{t-k} & \text{if } \exists k > 0 \text{ such that } X_{t-k} \text{ is observed} \\ X_{t+k} & \text{otherwise, if } \exists k > 0 \text{ such that } X_{t+k} \text{ is observed} \end{cases}$$

This two-step imputation preserved the temporal structure without introducing artificial trends and ensured the dataset was complete and consistent for downstream analyses such as dimensionality reduction and regression modeling.

Chapter 4

Multivariate Analysis

Multivariate analysis refers to a collection of statistical methods used to examine the relationships among multiple variables simultaneously. It allows us to evaluate the impact of several independent variables on one or more dependent variables while controlling for potential confounding effects.

In contrast to univariate analysis which consider only one variable at a time, multivariate analysis captures the complex, real-world interplay among several factors.

In this study, multivariate analysis is applied to understand how a combination of contextual indicators (e.g., access to electricity, CO_2 emissions, fertility rate, etc.) jointly influences psoriasis burden measures (Prevalence, Incidence, DALYs, YLDs). By accounting for these indicators together, we aim to uncover the strongest predictors and clarify the independent effects of each factor.

1 Correlation Analysis

To identify the most influential contextual indicators related to the psoriasis burden, we computed both Pearson and Spearman correlation coefficients with each of the four measures: Prevalence, Incidence, DALYs, and YLDs. Indicators with an absolute correlation coefficient ($|r|$ or $|\rho|$) greater than 0.65 were considered strongly associated.

1.1 Pearson Correlation Coefficient

The Pearson correlation coefficient measures the strength and direction of the **linear relationship** between two continuous variables. It assumes

both variables are normally distributed and the relationship between them is linear.

The formula for Pearson r is:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where:

- x_i and y_i are the individual sample points
- $\bar{x}$ and $\bar{y}$ are the sample means of x and y
- n is the number of observations

1.2 Spearman Rank Correlation Coefficient

Spearman rank correlation coefficient (ρ) assesses the strength and direction of a **monotonic relationship** between two variables. It is a non-parametric test that uses the ranks of the data rather than their raw values, making it robust to outliers and skewed distributions.

The formula for Spearman ρ (in the case of no tied ranks) is:

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

Where:

- d_i is the difference between the ranks of x_i and y_i
- n is the number of observations

In practice, when tied ranks exist, Spearmans ρ is calculated by applying Pearsons formula to the ranked variables.

1.3 Threshold for Interpretation

In this study, we focused on identifying **strong correlations**, defined as those with an **absolute correlation coefficient $|\rho|$ or $|r|$ greater than 0.65**. This threshold ensures that only meaningful and potentially influential relationships are highlighted in the subsequent interpretation.

Only those indicators with $|r| > 0.65$ or $|\rho| > 0.65$ in relation to the psoriasis burden measures are retained in the results.

Instead of listing correlations separately for each outcome, we summarize the top correlating indicators across all four burden measures in Table **??**.

This avoids redundancy and highlights variables that consistently exhibit strong relationships.

Indicator	**Max Pearson $\|r\|$**	**Max Spearman $\|\rho\|$**
Age dependency ratio, young (% of working-age population)	0.78	**0.84**
Age dependency ratio, old (% of working-age population)	0.71	**0.74**
Birth rate, crude (per 1,000 people)	0.75	**0.81**
Fertility rate, total (births per woman)	0.69	**0.77**
Adolescent fertility rate (births per 1,000 women ages 15–19)	0.66	**0.70**
Age dependency ratio (total, % of working-age population)	0.66	**0.72**
Access to electricity (% of population)	–	**0.69**
Access to electricity, urban (% of urban population)	–	**0.69**
People using basic sanitation services (% of population)	–	**0.68**
Incidence of tuberculosis (per 100,000 people)	–	**0.68**
CO_2 emissions (metric tons per capita)	–	**0.67**
Rural population (% of total population)	–	**0.66**
Urban population (% of total population)	–	**0.66**

Table 4.1: Top indicators with strong correlations ($|r|$ or $|\rho| > 0.65$) with psoriasis burden metrics

Notes:

- Only indicators with $|r|$ or $|\rho|$ above 0.65 for at least one burden measure are shown.
- Bold values highlight the strongest observed correlations.
- Variables marked with '-' were not among the top Pearson correlations but showed strong Spearman associations.

This consolidated view allows us to identify age structure, fertility, sanitation, and development-related factors as potential key contributors to the global psoriasis burden. These indicators were considered for inclusion in the subsequent multivariate regression analysis.

2 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a statistical technique used for dimensionality reduction while preserving as much variability as possible in the dataset. It transforms the original correlated variables into a new set of uncorrelated variables called *principal components*, which are linear combinations of the original variables. These components are ordered such that the first few retain most of the variation present in the original dataset. PCA is widely used for data compression, visualization, and as a preprocessing step for machine learning algorithms.

2.1 Aim and Purpose

To address multicollinearity and reduce the dimensionality of the dataset, Principal Component Analysis (PCA) was performed on a subset of indicators that showed strong correlations ($|r|$ or $|\rho| > 0.65$) with psoriasis burden metrics (Prevalence, Incidence, YLDs, and DALYs). PCA transforms correlated variables into orthogonal principal components (PCs), capturing the most variation in the data with fewer dimensions.

2.2 Selected Indicators for PCA

The following 13 indicators were selected based on the strength of their associations with the disease burden:

- Access to electricity (% of population) [EG.ELC.ACCS.ZS]
- Access to electricity, urban (% of urban population) [EG.ELC.ACCS.UR.ZS]
- Adolescent fertility rate (births per 1,000 women ages 15–19) [SP.ADO.TFRT]
- Age dependency ratio (% of working-age population) [SP.POP.DPND]
- Age dependency ratio, old (%) [SP.POP.DPND.OL]
- Age dependency ratio, young (%) [SP.POP.DPND.YG]

- Birth rate, crude (per 1,000 people) [SP.DYN.CBRT.IN]
- CO_2 emissions (metric tons per capita) [EN.ATM.CO2E.PC]
- Fertility rate, total (births per woman) [SP.DYN.TFRT.IN]
- Incidence of tuberculosis (per 100,000 people) [SH.TBS.INCD]
- People using at least basic sanitation services (%) [SH.STA.BASS.ZS]
- Rural population (% of total population) [SP.RUR.TOTL.ZS]
- Urban population (% of total population) [SP.URB.TOTL.IN.ZS]

To ensure equal weighting and comparability of features before applying Principal Component Analysis, all variables were standardized using z-score normalization:

$$z = \frac{x - \mu}{\sigma}$$

where x is the original value, μ is the mean, and σ is the standard deviation of the variable.

2.3 Scree Plot and Explained Variance

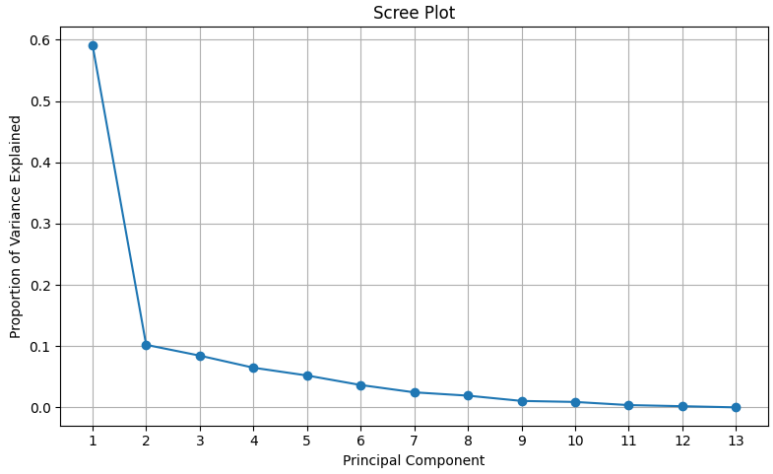


Figure 4.1: Scree plot showing variance explained by each principal component

Figure 4.1 displays the scree plot showing both the individual and cumulative variance explained by the principal components. Notably, the first component

alone accounts for over 60% of the total variance. After the first few components, the marginal contribution of additional components decreases sharply, forming a visible ”elbow” in the plot.

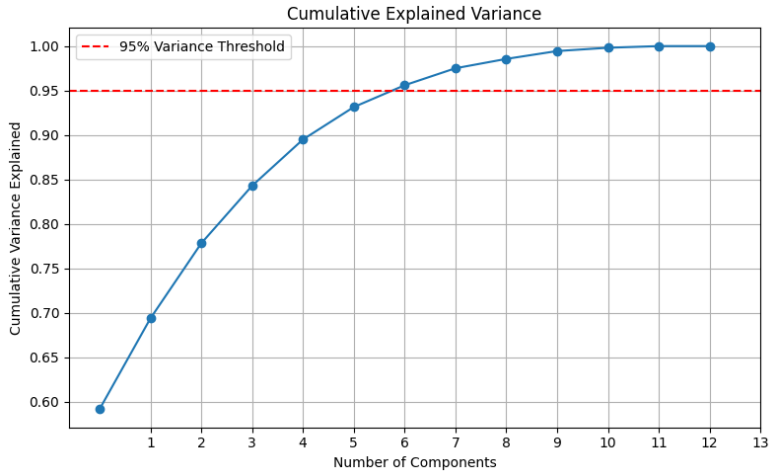


Figure 4.2: Cumulative explained variance by the principal components

Figure 4.2 shows the cumulative explained variance. The first 6 to 7 components together explain approximately 95% of the total variance, justifying their selection for dimensionality reduction in subsequent analyses.

2.4 Principal Component Loadings Interpretation

To uncover the latent structure in the indicator dataset and reduce dimensionality while preserving most of the variance, Principal Component Analysis (PCA) was conducted on the standardized indicators. These indicators were selected based on strong Spearman and Pearson correlations with the disease burden metrics (DALYs, YLDs, Prevalence, and Incidence). Figure 4.3 visualizes the top 7 contributing indicators for each of the first six principal components (PCs), categorized into themes such as *Demographic*, *Infrastructure*, *Health*, and *Environmental*.

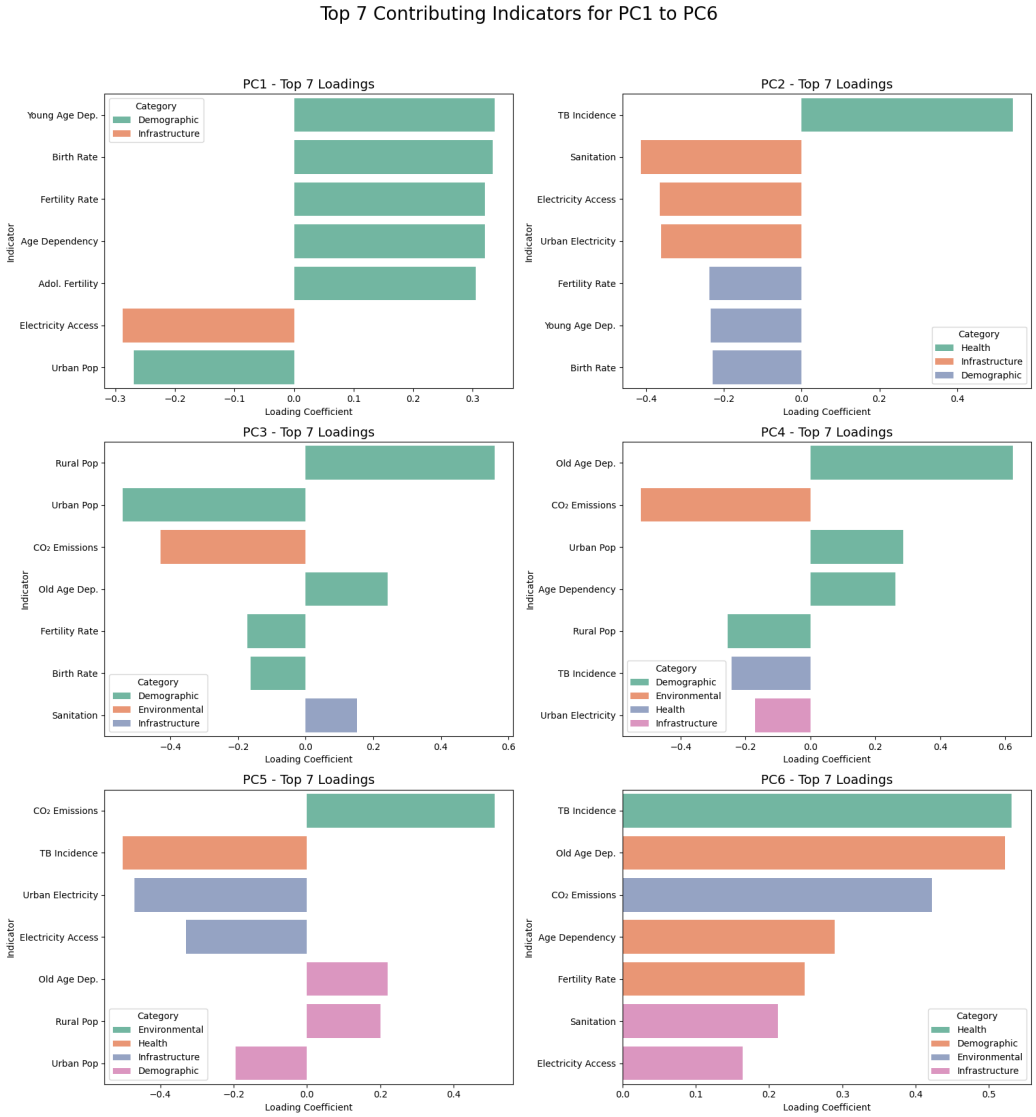


Figure 4.3: Top 7 contributing indicators to PC1 through PC6. Bar length denotes loading magnitude; color represents thematic category.

Each principal component reveals a distinct latent dimension shaped by development indicators:

- **PC1** is primarily shaped by ***Demographic*** variables, including *Young Age Dependency*, *Birth Rate*, and *Fertility Rate*. This component appears to capture a dimension of population growth and youthfulness, with inverse associations to *Electricity Access* and *Urban Population*, suggesting a rural, high-fertility demographic profile.
- **PC2** reflects contrasts between ***Health and Infrastructure***, with strong negative contributions from *Tuberculosis Incidence*, *Sanitation*, and *Electricity Access*. This axis likely represents disparities in basic health infrastructure and disease burden.
- **PC3** combines ***Demographic***, ***Environmental***, and ***Urbanization*** aspects. Key indicators such as *Rural Population*, *CO_2 Emissions*, and *Urban Population* load heavily, indicating a spectrum from low-emission rural nations to more urbanized, industrialized settings.
- **PC4** integrates indicators related to ***Ageing***, ***Pollution***, and ***Urban Development***. It is influenced by *Old Age Dependency*, *CO_2 Emissions*, and *Urban Population*, suggesting that this axis may describe transitions toward older, more developed societies with environmental burdens.
- **PC5** emphasizes ***Infrastructure and Environmental Access***. Negative contributions from *CO_2 Emissions* and *TB Incidence*, alongside positive associations with *Electricity Access* and *Urban Population*, suggest this component captures infrastructure gaps and public health challenges.
- **PC6** centers on a combination of ***Health***, ***Environmental***, and ***Demographic*** burdens, with strong positive loadings from *Tuberculosis Incidence*, *Old Age Dependency*, and *CO_2 Emissions*. This dimension may reflect compound vulnerabilities in ageing, polluted, and under-resourced settings.

These components offer a reduced representation of complex development dynamics that underlie spatial patterns in psoriasis burden, facilitating downstream clustering and regression modeling.

2.5 Scatterplot of Countries in Principal Component Space

Unsupervised dimensionality reduction was conducted using Principal Component Analysis (PCA), which transformed the multivariate dataset into a lower-dimensional space. The resulting scatterplot in Figure 4.4 presents countries

projected along the first two principal components (PC1 and PC2). This visualization captures the primary patterns of variance across countries based on the selected indicators.

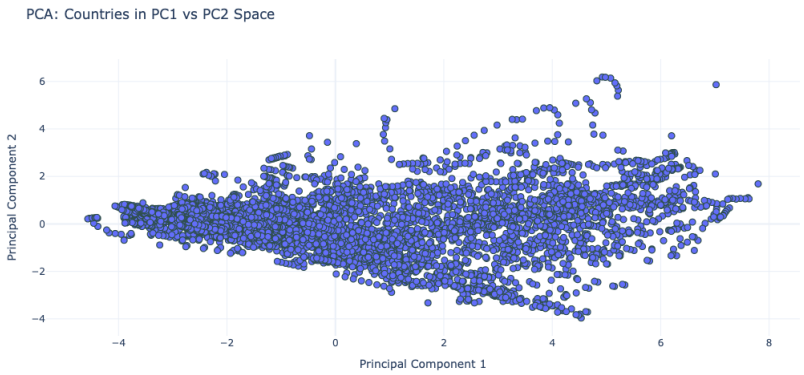


Figure 4.4: PCA map displaying the distribution of countries along PC1 and PC2 space.

To enhance interpretability, an interactive version of the PCA map was created using Plotly. This interactive feature enables the identification of individual countries, their corresponding year along with the exact pca coordinates by hovering over data points. As Shown in Figure 4.5, where a popup box displays detailed information for a selected country (Iraq).

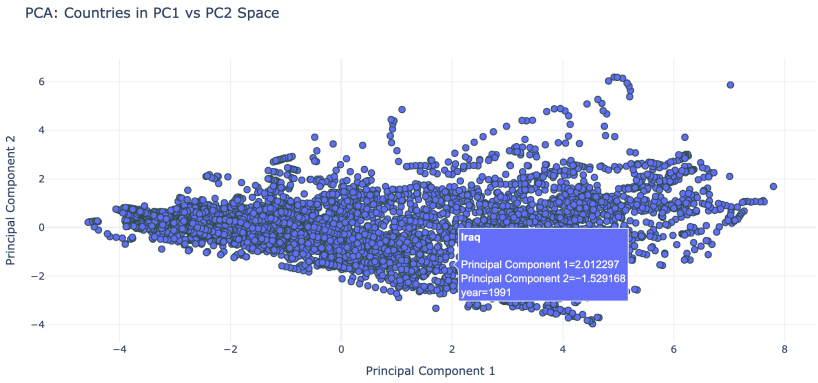


Figure 4.5: Interactive PCA map view. Country tooltip (Iraq) shown while hovering.

3 Clustering Analysis

To uncover latent groupings among countries based on key indicators, K-Means clustering was applied to the standardized dataset. The optimal number of clusters was determined using two methods:

- **Elbow Method:** This approach examines the within-cluster sum of squares (inertia) across varying values of k. A noticeable "elbow" at $k = 3$ suggests diminishing returns in clustering quality beyond this point (Figure 4.6).
- **Silhouette Score:** This metric evaluates the coherence of individual clusters by measuring how well each point fits within its assigned cluster compared to others. The silhouette score was highest at $k = 2$, but remained reasonably strong at $k = 3$, making it a suitable choice for clustering (Figure 4.7

  The silhouette score for a single sample i is defined as:

  $$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

  Where:

  - $a(i)$ is the mean intra-cluster distance (the average distance between i and all other points in the same cluster),
  - $b(i)$ is the mean nearest-cluster distance (the average distance between i and all points in the nearest different cluster),
  - $s(i) \in [-1, 1]$, where values close to 1 indicate appropriate clustering.

  The overall silhouette score for the dataset is the mean of all individual scores:

  $$S = \frac{1}{n} \sum_{i=1}^{n} s(i)$$

  Where n is the total number of data points.

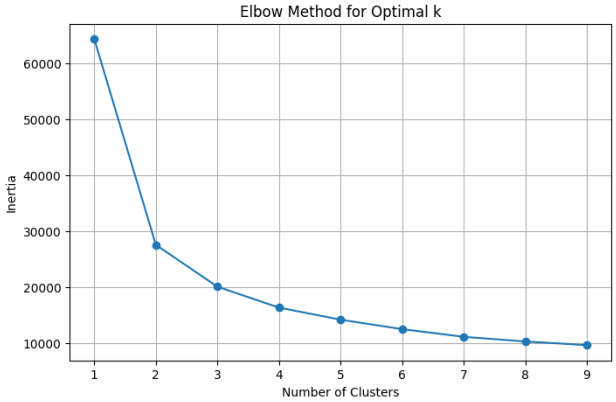


Figure 4.6: Elbow method showing optimal number of clusters

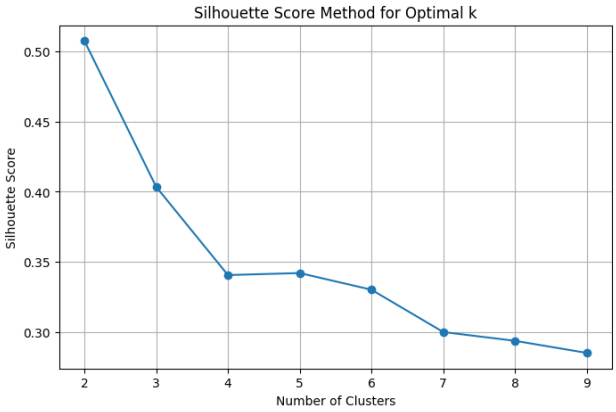


Figure 4.7: Silhouette Score method for selecting the optimal number of clusters

Based on these diagnostics, $k = 3$ was selected as the most interpretable and balanced number of clusters. The resulting groupings of countries were then visualized in a two-dimensional space using the first two principal components derived from PCA (Figure 4.8).

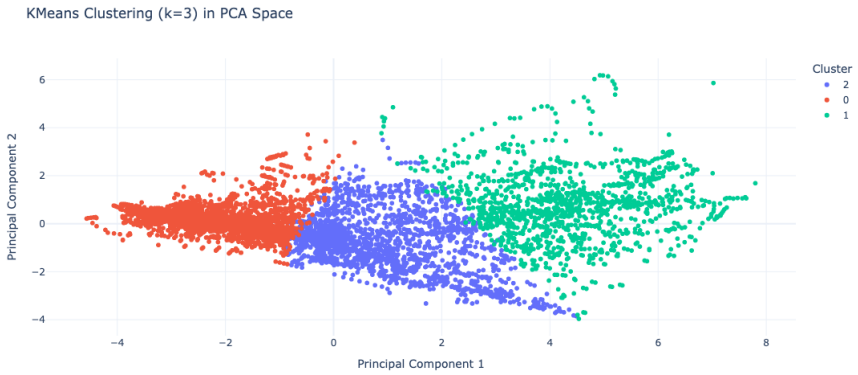


Figure 4.8: K-Means clustering results visualized using PCA components

3.1 Interpretation of Cluster Profiles

As illustrated in Figure 4.8 the application of KMeans clustering on PCA-transformed indicator scores resulted in three distinct clusters. These clusters reflect meaningful groupings based on countries, demographic, health, environmental, and infrastructure-related profiles. To interpret the nature of each cluster, we analyzed the mean values of key indicators per cluster, as shown in Table ??.

Cluster 0 (High Development): Countries in Cluster 0 exhibit high levels of infrastructure and public health development. They have nearly universal access to electricity (98.2%) and urban electricity coverage (99.3%), alongside high sanitation service usage (90.8%). These countries also show low fertility rates (2.06 births per woman), low adolescent fertility (29.7), and lower crude birth rates (14.7 per 1,000 people). Demographically, they have higher proportions of elderly populations, reflected in the higher old-age dependency ratio (17.3). Additionally, they produce the highest CO_2 emissions (6.26 metric tons per capita), characteristic of more industrialized economies.

Cluster 2 (Medium Development): Countries in Cluster 2 lie between the extremes. They show moderate infrastructure—electricity access at 86.3%, urban access at 94.7%, and sanitation at 80.6%. Fertility rates (3.98), birth rates (30.2), and dependency ratios (76.0) suggest they are undergoing demographic transition. Their CO_2 emissions (2.52) are lower than high-income nations but higher than the least developed ones. Tuberculosis incidence is also intermediate (87.8 per 100,000), pointing to improving but still vulnerable

health systems.

Cluster 1 (Low Development): Cluster 1 represents countries with significant development challenges. These nations have low access to electricity (47.8%) and sanitation services (53.8%), and a largely rural population (68.4%). Demographically, they display high birth (40.7) and fertility (5.7) rates, high adolescent fertility (120.5), and an extremely high young-age dependency ratio (85.7), all of which are typical of early-stage demographic transitions. Health indicators also point to poorer outcomes, such as a high tuberculosis incidence rate (197.2 per 100,000). These countries are generally in Sub-Saharan Africa and parts of South Asia, as seen in the 1990 cluster map.

Indicator	Cluster 0	Cluster 1	Cluster 2
Access to electricity (%)	98.21	47.78	86.34
Urban electricity access (%)	99.35	68.36	94.67
Adolescent fertility rate	29.74	120.52	77.21
Age dependency ratio (total)	50.19	91.61	75.99
Old-age dependency ratio	17.28	5.91	7.20
Young-age dependency ratio	32.91	85.70	68.79
Birth rate (per 1,000)	14.72	40.73	30.25
Fertility rate (births per woman)	2.06	5.70	3.98
CO_2 emissions (t per capita)	6.26	0.62	2.52
TB incidence (per 100,000)	50.50	197.24	87.78
Sanitation access (%)	90.75	53.81	80.58
Rural population (%)	26.06	68.36	52.36
Urban population (%)	77.28	31.64	49.13

Table 4.2: Cluster Profiles Based on Mean Indicator Values

3.2 Spatiotemporal Evolution of Country Clusters (1990-2021)

To visualize spatial patterns and their evolution over time, choropleth maps were generated for the initial (1990) and most recent (2021) years. These maps display country groupings according to K-Means clustering based on principal components derived from socio-demographic and infrastructure indicators.

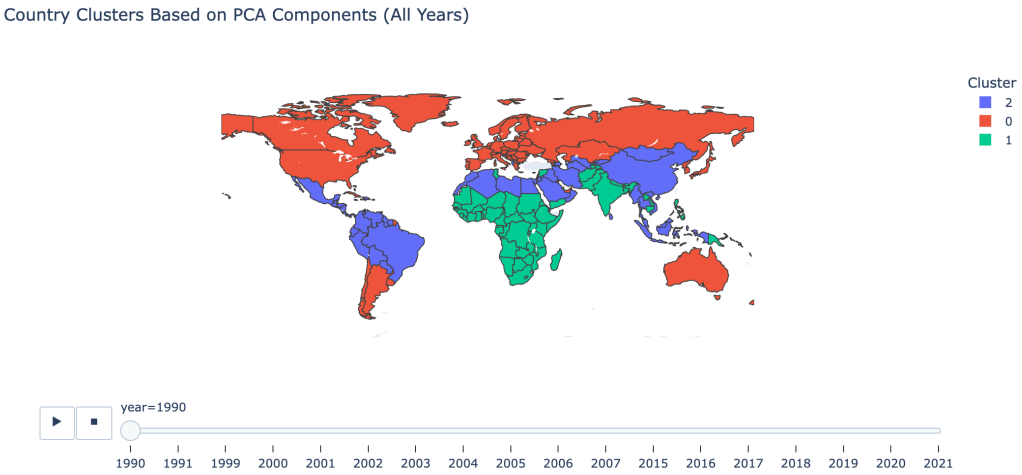


Figure 4.9: Country Clusters Based on PCA Components (Year 1990)

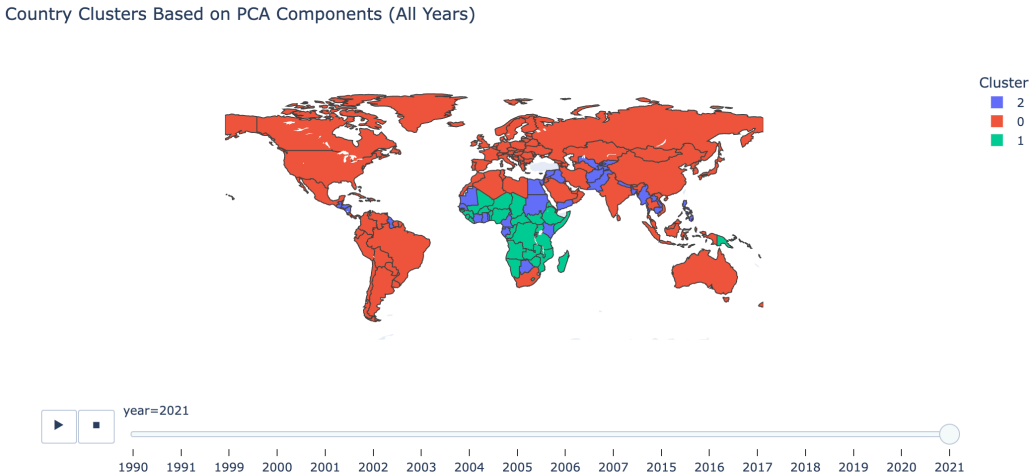


Figure 4.10: Country Clusters Based on PCA Components (Year 2021)

Over the period from (Figure 4.9) 1990 to (Figure 4.10) 2021, a total of **75 countries** changed their cluster affiliation based on principal component scores and K-means clustering. This reflects meaningful changes in socioeconomic and infrastructural indicators across decades. Some prominent shifts include:

- **China, Brazil, and Mexico** transitioned from Cluster 2 to Cluster 0, indicating significant improvements in infrastructure, demographic transition, and health indicators.
- **Bangladesh, Pakistan, and Philippines** moved from Cluster 1 to Cluster 2, suggesting moderate but consistent development progress.
- **Libya, Colombia, and Indonesia** also moved from Cluster 2 to Cluster 0, aligning more closely with countries showing higher development standards.

These spatial shifts underscore how clustering based on PCA can reveal not only groupings but also trajectories in global development and health determinants.

3.3 Income Group Distribution Across Clusters

To further contextualize the PCA-based clustering, I incorporated additional metadata from the World Bank Open Data, specifically the *Income Group* classification for each country. This data categorizes countries as Low income, Lower middle income, Upper middle income, or High income based on Gross National Income (GNI) per capita.

By merging the income group data with the clustered PCA results, we are able to examine how socioeconomic classifications align with the discovered clusters.

Figure 4.11 presents a bar plot showing the distribution of income groups across the three identified clusters.

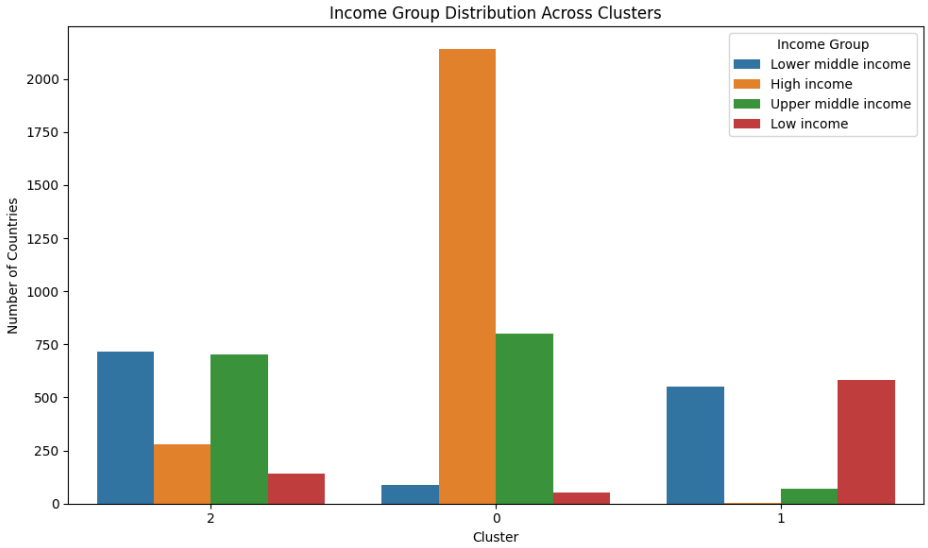


Figure 4.11: Income Group distribution across PCA-based clusters.

To better understand how the clusters relate to economic development, we compared them with World Bank income classifications. By mapping countries according to both their cluster assignment and income group (low, lower-middle, upper-middle, and high income), we explored whether countries with similar development and health characteristics also share similar income levels. This revealed patterns such as whether higher income countries tend to fall into the same cluster and highlights exceptions where a countrys health and demographic profile may not match its income category.

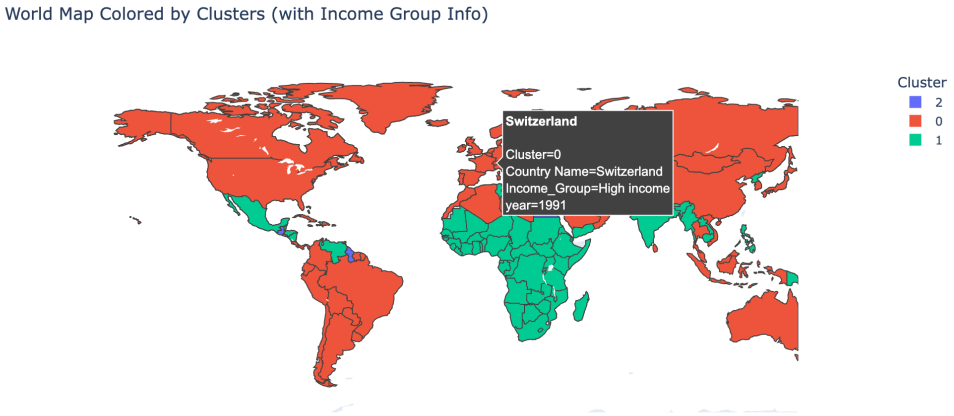


Figure 4.12: World map colored by clusters with income group information

Figure 4.12 overlays the PCA-based clusters with each country's income group classification. This visualization helps clarify how health, demographic, and infrastructure indicators correspond to economic development levels.

Cluster interpretations based on indicator profiles:

- **Cluster 0 (red)**: Includes countries with high access to electricity and sanitation, low fertility and birth rates, and higher levels of industrialization. These characteristics are consistent with high-income countries, often featuring well-developed healthcare systems and aging populations.
- **Cluster 1 (green)**: Comprises countries with very high fertility and birth rates, high dependency ratios, limited access to infrastructure services, and high burdens of communicable diseases. This cluster aligns closely with low-income or lower-middle-income countries, especially in Sub-Saharan Africa and parts of South Asia.
- **Cluster 2 (blue)**: Represents countries in transition, showing moderate values across most indicators. These include intermediate levels of fertility, infrastructure access, and disease burden, typically found in upper-middle-income countries progressing through demographic and health transitions.

This joint spatial and socioeconomic analysis confirms that unsupervised clustering based on key development indicators reveals coherent groupings aligned with global income classification systems.

4 Modeling Psoriasis Burden Using Contextual Development Indicators

4.1 Introduction and Objective

This section presents the modeling of psoriasis burden using contextual development indicators as predictors. To quantitatively assess these relationships, separate regression analyses were conducted for each psoriasis related measure Prevalence, Incidence, Years Lived with Disability (YLDs), and Disability Adjusted Life Years (DALYs) as the dependent variables. The objective is to identify significant socioeconomic, environmental, and demographic factors that explain variation in psoriasis burden across countries and over time, and to estimate the strength and direction of their associations.

4.2 Methodology

4.2.1 Data Preparation

To prepare the data for regression modeling, a subset of the merged dataset was constructed by selecting the relevant contextual indicators, the psoriasis burden values, and their corresponding metadata. The final dataset, structured as a long-format table, included the following columns: `Country Name`, `year`, `measure_name`, `val`, and the selected development indicators.

Duplicate rows, if any, were removed to avoid redundant observations. All independent (contextual) variables were standardized (z-scored) to enable comparability of regression coefficients and to reduce the impact of scale differences. The psoriasis burden metrics Prevalence, Incidence, DALYs, and YLDs were treated as continuous dependent variables.

4.2.2 Model Specification

To model the relationship between contextual development indicators and psoriasis burden, several regression techniques were applied, each offering unique strengths:

- **Variance Inflation Factor (VIF):** Computed prior to modeling to assess multicollinearity among predictors. Variables with high VIF were flagged as potential sources of instability as usually having variables with VIF more than 10 is a sign of multicollinearity.

- **Ordinary Least Squares (OLS):** Used as a baseline model to quantify the linear association between predictors and each psoriasis outcome (Prevalence, Incidence, YLDs, DALYs).
- **Ridge Regression (L2 Regularization):** Applied to mitigate multicollinearity by penalizing large coefficients and improving model generalization.
- **Random Forest Regression:** Implemented as a non-linear, ensemble-based approach to validate results from linear models and to capture potential interaction effects between predictors. Feature importance scores were used to assess the relative contribution of each variable.

Each model was applied independently to the four psoriasis outcomes, using standardized contextual indicators as predictors. Model performance was evaluated using R^2, Mean Squared Error(MSE) and out-of-sample error (where applicable). The combination of parametric and non-parametric techniques provided a robust framework for identifying and validating the key drivers of psoriasis burden.

4.2.3 Multicollinearity Check Using Variance Inflation Factor (VIF)

Before performing multivariable regression analysis, it is essential to examine multicollinearity among independent variables. High multicollinearity can distort the estimation of regression coefficients, inflate standard errors, and reduce the reliability of the model.

To assess multicollinearity, we computed the **Variance Inflation Factor (VIF)** for each predictor variable. A VIF value above 10 is commonly considered indicative of high multicollinearity, while values above 5 may suggest moderate correlation issues. A VIF of infinity (∞) indicates perfect linear dependence, which must be addressed before model fitting.

The VIF for predictor variable X_j is defined as:

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

Where:

- R_j^2 is the coefficient of determination obtained by regressing X_j on all other predictor variables,
- A higher VIF indicates greater multicollinearity. Common thresholds:
  - VIF > 5: moderate multicollinearity,
  - VIF > 10: high multicollinearity (typically problematic).

Table 4.3: Variance Inflation Factor (VIF) for Predictor Variables

Variable	**Description**	**VIF**
Access to electricity (% of population)	Proxy for infrastructure and modernization	8.25
Access to electricity, urban (%)	Urban infrastructure quality	4.95
Adolescent fertility rate	Early reproductive burden	3.93
Age dependency ratio	Total population dependency	∞
Age dependency, old (%)	Elderly support burden	∞
Age dependency, young (%)	Child support burden	∞
Birth rate, crude (per 1,000)	General fertility pressure	42.17
CO_2 emissions (t per capita)	Environmental development proxy	1.79
Fertility rate, total (births per woman)	Average reproductive load	25.54
Incidence of tuberculosis (per 100,000)	Public health infrastructure quality	1.58
Basic sanitation access (%)	Hygiene and disease prevention	4.04
Rural population (% of total)	Inverse of urbanization level	20.50
Urban population (% of total)	Urbanization proxy	20.78

As shown in Table 4.3, several variables exhibit multicollinearity:

- The **age dependency ratio** and its components (old and young) show perfect collinearity (VIF = ∞), likely due to mathematical identity relationships. Only one component was retained to avoid redundancy.
- **Birth rate**, **fertility rate**, and **urban/rural population proportions** exhibit extremely high VIFs, suggesting strong overlap. Only one representative variable from this cluster was included in the regression model.
- Variables like **CO_2 emissions** and **tuberculosis incidence** have acceptable VIFs, indicating minimal multicollinearity concerns.

To mitigate multicollinearity effects, we applied variable selection based on correlation analysis and retained only non-redundant predictors in the final regression models. This ensured model interpretability and stability in coefficient estimation.

4.2.4 Ordinary Least Squares (OLS) Regression

Ordinary Least Squares (OLS) regression was employed as the baseline linear modeling approach to explore the relationship between contextual development indicators and the burden of psoriasis. This method was selected for its simplicity, interpretability, and ability to estimate the strength and direction of linear associations between predictors and outcomes.

The OLS estimator aims to minimize the residual sum of squares between observed and predicted values. In matrix notation, the estimator for the regression coefficients is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

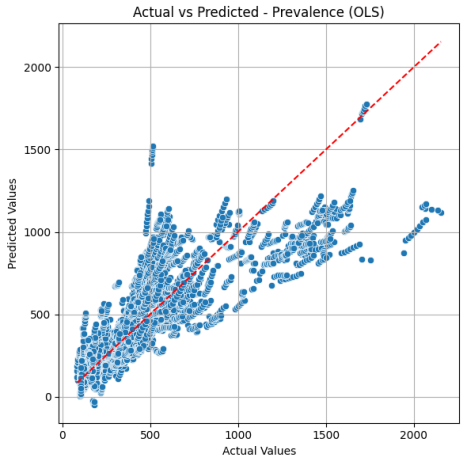
Where:

- $\mathbf{X}$ is the matrix of predictor variables,
- $\mathbf{y}$ is the response variable vector,
- $\hat{\boldsymbol{\beta}}$ is the vector of estimated coefficients,
- $\mathbf{X}^\top$ is the transpose of $\mathbf{X}$,
- $(\mathbf{X}^\top \mathbf{X})^{-1}$ is the inverse of the Gram matrix.

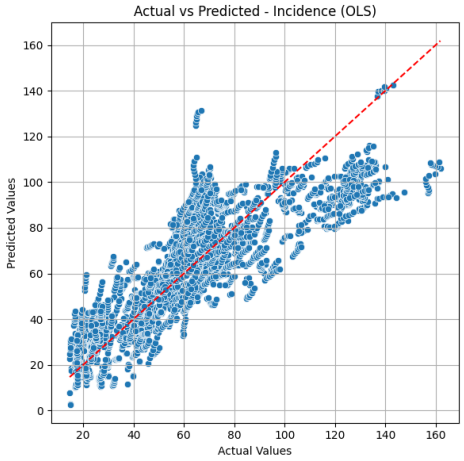
Each psoriasis burden metric **Prevalence**, **Incidence**, **Disability-Adjusted Life Years (DALYs)**, and **Years Lived with Disability (YLDs)** was modeled separately as the dependent variable. The same set of standardized contextual indicators was used as predictors across all models. These included demographic, environmental, and infrastructural variables.

Model Performance

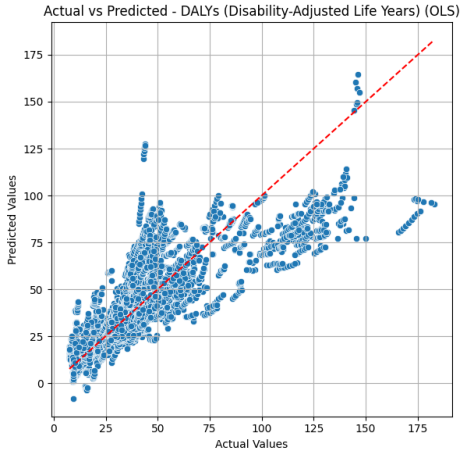
- The models achieved strong performance, with R^2 values ranging from **0.623 to 0.695**, indicating that 62-70% of the variance in psoriasis outcomes could be explained by the contextual predictors.
- Adjusted R^2 values were closely aligned with the R^2 values, reflecting model stability and suggesting that the number of predictors did not lead to overfitting.
- Prediction quality is illustrated in Figure 4.13, which shows the predicted vs. actual values for each outcome. The red diagonal line represents the ideal prediction line ($y = x$), helping assess model bias and residual dispersion.



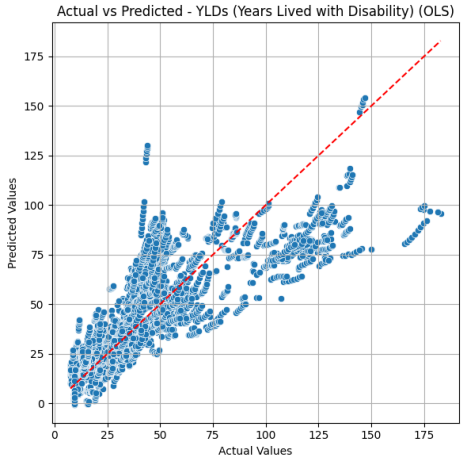
(a) Prevalence



(b) Incidence



(c) DALYs

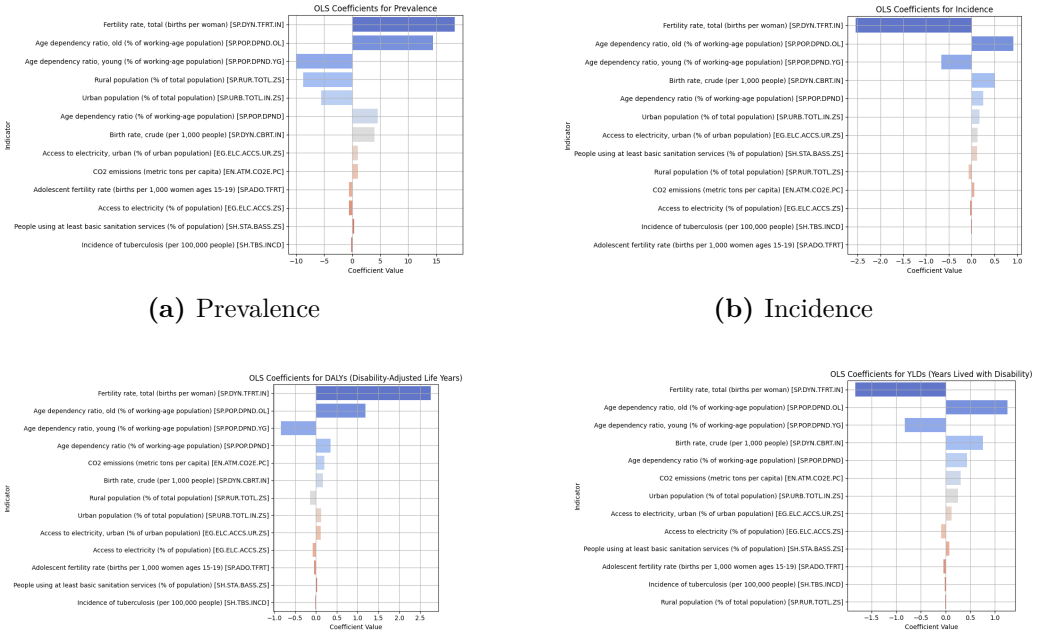


(d) YLDs

Figure 4.13: Actual vs. Predicted Values for Psoriasis Burden Outcomes (OLS Models). The red dashed line represents the ideal prediction line ($y = x$).

Significant Predictors Several predictors showed strong and consistent associations across models (Figure 4.14):

- **Urban access to electricity** and **old-age dependency ratio** were positively associated with most psoriasis outcomes, possibly reflecting stronger healthcare infrastructure and diagnosis capabilities in developed regions.
- **Young-age dependency ratio** and **tuberculosis incidence** were negatively associated with outcomes, suggesting an inverse relationship between competing disease burdens and psoriasis recognition.
- Other variables such as **fertility rate** and **CO_2 emissions** appeared significant in select models, indicating broader socio-environmental influences.



(a) Prevalence **(b)** Incidence

(c) DALYs **(d)** YLDs

Figure 4.14: OLS Coefficient Estimates for Psoriasis Outcomes. Positive and negative contributions of each contextual variable to the respective outcome.

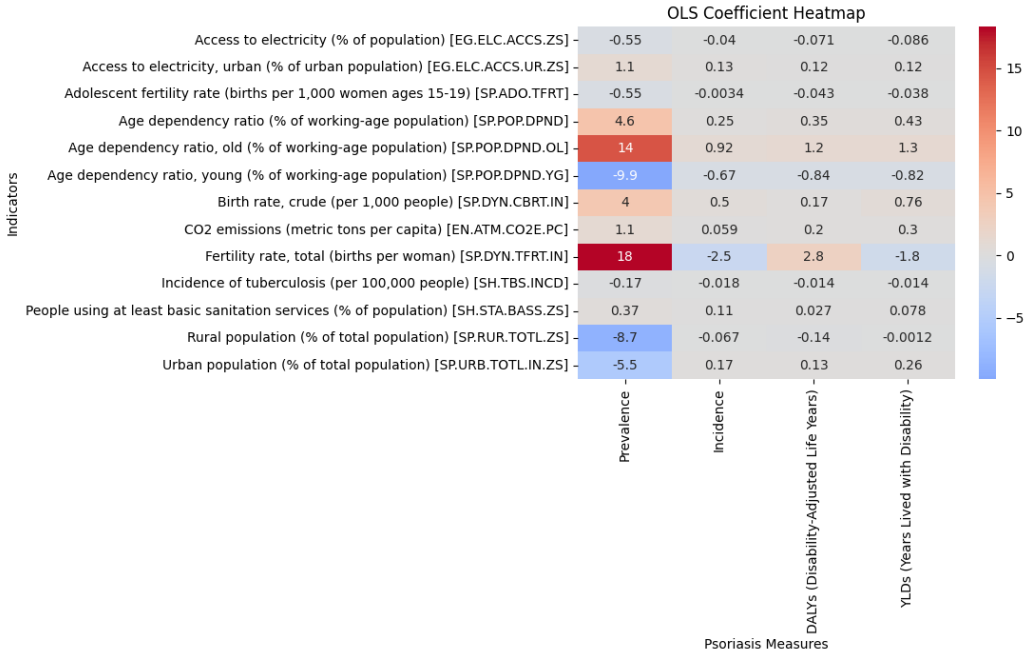


Figure 4.15: Heatmap of OLS Coefficients across Psoriasis Burden Measures. Warmer colors indicate stronger positive associations; cooler colors indicate negative associations.

Model Diagnostics

- Warnings regarding very small eigenvalues indicated potential multicollinearity in the predictor set, a common issue in development data where indicators may be structurally correlated.
- To mitigate these issues and assess robustness, regularized regression techniques (Ridge and Lasso) were applied in subsequent sections.

OLS Coefficient Summary Table To facilitate comparison across models, Table **??** provides a consolidated view of the coefficients for each outcome.

Indicator	Prevalence	Incidence	DALYs	YLDs
Access to electricity (%)	-0.55	-0.04	-0.07	-0.09
Urban electricity access (%)	1.08	0.13	0.12	0.12
Adolescent fertility rate	-0.55	-0.003	-0.04	-0.04
Age dependency ratio	4.55	0.25	0.35	0.43
Old-age dependency ratio	14.47	0.92	1.19	1.26
Young-age dependency ratio	-9.92	-0.67	-0.84	-0.82
Birth rate (crude)	3.99	0.50	0.17	0.76
CO_2 emissions	1.05	0.06	0.20	0.30
Fertility rate (total)	18.32	-2.54	2.76	-1.83
TB incidence	-0.17	-0.018	-0.014	-0.014
Basic sanitation (%)	0.37	0.11	0.03	0.08
Rural population (%)	-8.74	-0.067	-0.14	-0.001
Urban population (%)	-5.50	0.17	0.13	0.26

Table 4.4: OLS Coefficient Estimates for Psoriasis Burden Metrics

Overall, OLS modeling provided a transparent and foundational analysis of how contextual development indicators relate to the burden of psoriasis. The next sections examine whether these findings hold under regularized and non-linear regression approaches.

4.2.5 Regularized Regression Using Ridge

To address multicollinearity in the dataset, we applied **Ridge Regression**, which introduces an L_2 penalty to shrink coefficients and stabilize the estimates. This is especially useful given the strong collinearity between population structure and development indicators identified during VIF analysis.

Ridge regression modifies the ordinary least squares (OLS) loss function by adding a penalty term proportional to the square of the magnitude of coefficients:

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \mathbf{x}_i^\top \beta)^2 + \alpha \sum_{j=1}^{p} \beta_j^2 \right\} \tag{4.1}$$

where α is the regularization strength, with higher values leading to greater shrinkage of coefficients.

We used cross-validation to select the optimal regularization parameter (α) for each psoriasis burden outcome. The results are summarized in Table **??**.

Measure	Optimal α	R^2 Score	Mean Squared Error (MSE)
DALYs	23.10	0.6243	379.73
YLDs	7.56	0.6234	380.63
Prevalence	1.07	0.6275	51856.93
Incidence	4.33	0.6949	285.09

Table 4.5: Ridge Regression Performance by Outcome

Model Performance Summary Ridge regression models showed moderately strong performance across all psoriasis burden outcomes. The model for **Incidence** performed the best, achieving the highest explanatory power with an R^2 of **0.6949** and the lowest mean squared error (MSE) of **285.09**, indicating precise predictions. **Prevalence** also demonstrated reasonable predictive power ($R^2 = 0.6275$), though its MSE was substantially higher (**51,856.93**) due to the larger scale of prevalence values. Models for **DALYs** and **YLDs** yielded comparable results, each explaining approximately **62%** of the variance in outcomes. The optimal regularization strength (α) varied by outcome, reflecting differences in multicollinearity and noise sensitivity across the burden measures.

Actual vs. Predicted Interpretation Figure 4.16 presents scatter plots comparing the actual versus predicted values for each psoriasis outcome using Ridge regression. In all subplots, the red dashed line represents the ideal prediction line ($y = x$), where predictions perfectly match the actual values.

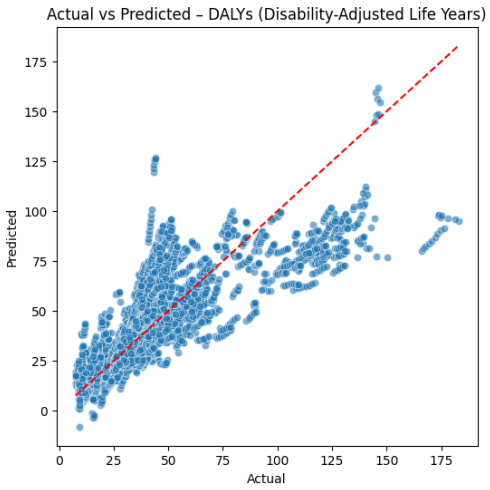
The scatter patterns indicate that Ridge regression captured the overall trends across all four outcomes, though with varying precision:

For **Incidence**, predictions closely align with the diagonal, suggesting strong agreement and low dispersion, consistent with its highest R^2 value (0.6949).

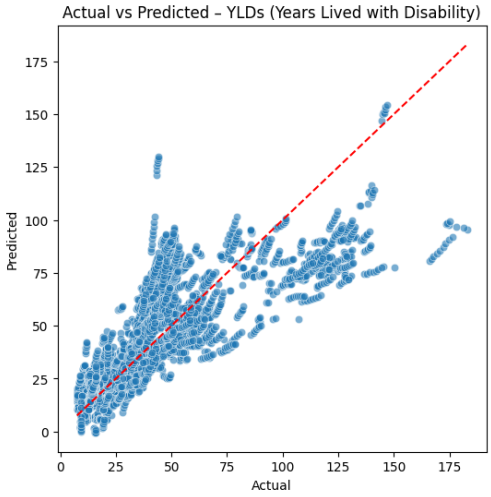
Prevalence and **DALYs** show slightly greater spread around the diagonal, particularly at higher values, indicating moderate prediction accuracy and some underestimation of extreme cases.

YLDs results mirror the pattern seen in DALYs, with Ridge effectively capturing the central trends but with more variance in the tails.

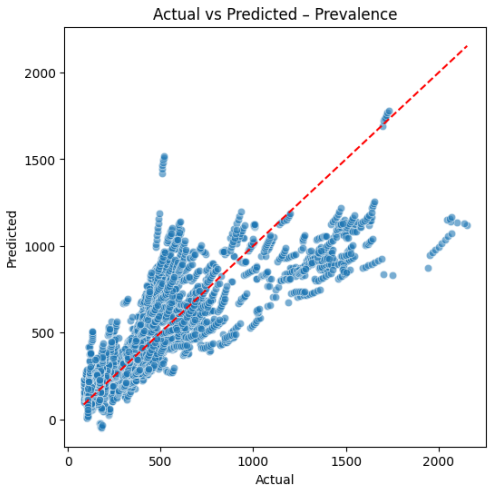
These results confirm Ridge regression's ability to generalize well and manage multicollinearity, although slight deviations at the distribution extremes suggest that some non-linear or interaction effects may remain unaccounted for.



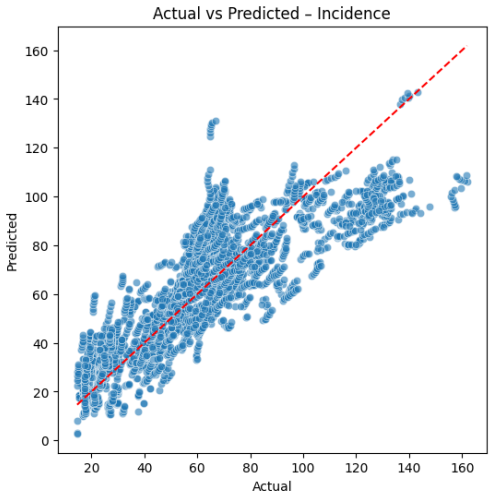
(a) Actual vs. Predicted - DALYs



(b) Actual vs. Predicted - YLDs



(c) Actual vs. Predicted - Prevalence

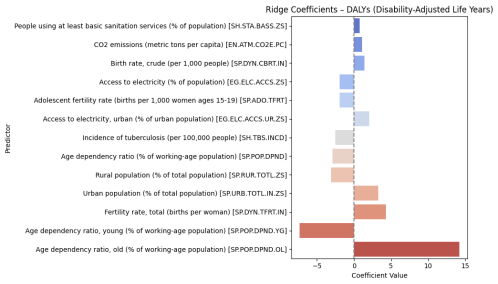


(d) Actual vs. Predicted - Incidence

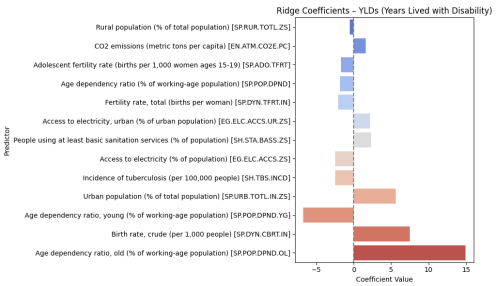
Figure 4.16: Ridge Regression: Actual vs. Predicted Values for Psoriasis Burden Outcomes. The red dashed line represents the ideal fit ($y = x$).

Coefficient Interpretation To interpret the influence of individual predictors, we visualized the standardized Ridge regression coefficients. Positive coefficients indicate a direct relationship with the burden measure, while negative values imply an inverse relationship. Figures 4.17a, 4.17b, 4.17c, and 4.17d display bar plots of the top contributing indicators per outcome.

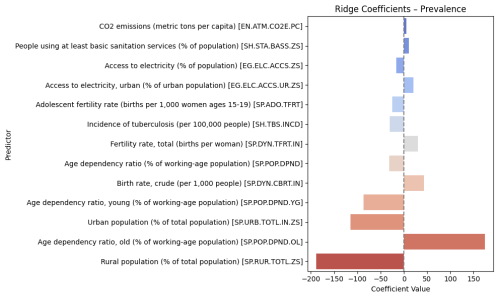
Figure 4.17 displays the standardized Ridge regression coefficients for all four psoriasis outcomes in a 2x2 layout. Each bar represents the strength and direction of the relationship between a predictor and the burden measure, enabling comparison across models.



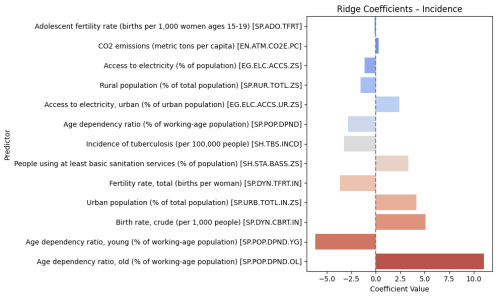
(a) DALYs (Disability-Adjusted Life Years)



(b) YLDs (Years Lived with Disability)



(c) Prevalence



(d) Incidence

Figure 4.17: Ridge Regression Coefficient Estimates for Psoriasis Burden Outcomes. Bars represent standardized coefficients indicating the strength and direction of associations between contextual predictors and each burden measure.

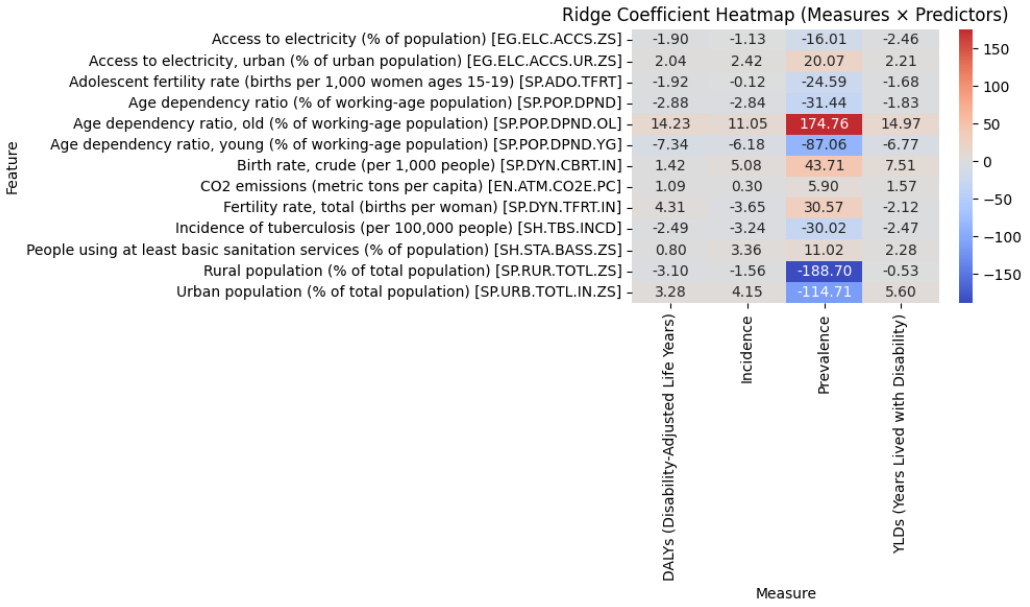


Figure 4.18: Heatmap of Ridge Coefficients Across Psoriasis Burden Measures

Key Insights

- **Old-age dependency ratio** was the most influential positive predictor across all models-especially in DALYs and Prevalence—indicating greater chronic burden in aging populations.
- **Young-age dependency ratio** showed consistently strong negative associations, suggesting that younger populations tend to have lower psoriasis burden.
- **Urban population** and **fertility rate** were positively associated with Prevalence and Incidence, likely reflecting improved reporting and healthcare access.
- **Rural population share** had strong negative coefficients, especially in Prevalence, possibly due to underdiagnosis or lower healthcare penetration.

4.2.6 Nonlinear Modeling Using Random Forest Regression

To complement linear modeling approaches and capture potential nonlinear relationships, we applied **Random Forest Regression** to predict psoriasis burden outcomes. Random Forests are ensemble models that construct multiple decision trees and aggregate their predictions, reducing overfitting

and improving generalization. This method is particularly well-suited for datasets with complex feature interactions and does not assume linearity or independence of predictors.

In Random Forest regression, the final prediction $\hat{y}$ is the average of the predictions from all individual decision trees:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(x)$$

Where:

- T is the total number of trees in the forest,
- $h_t(x)$ is the prediction from the t^{th} tree.

We trained separate Random Forest models for each outcome DALYs, YLDs, Prevalence, and Incidence . The performance metrics are summarized in Table **??**.

Measure	R^2 Score	Mean Squared Error (MSE)
DALYs	0.9543	43.67
YLDs	0.9542	43.75
Prevalence	0.9534	2883.66
Incidence	0.9628	27.26

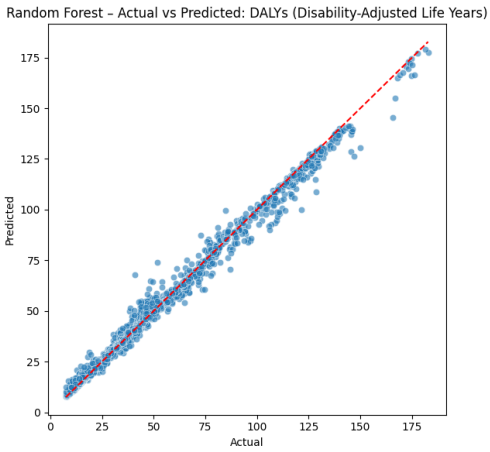
Table 4.6: Random Forest Regression Performance by Outcome

Model Performance Summary Random Forest models outperformed Ridge regression across all psoriasis burden measures, demonstrating strong predictive accuracy. The model for **Incidence** achieved the best performance ($R^2 = 0.9628$, MSE = 27.26), indicating highly precise predictions. Both **DALYs** and **YLDs** were modeled with near-identical strength ($R^2 \approx 0.954$), capturing over 95% of the variance in observed values. **Prevalence** also showed high explanatory power, though with a larger MSE (2883.66) due to its broader numeric range.

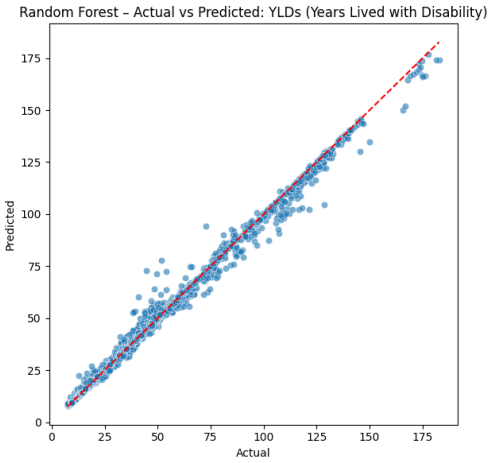
These results highlight the value of tree-based methods in modeling complex patterns and nonlinear associations among contextual predictors.

Actual vs. Predicted Interpretation Figure 4.19 displays scatter plots comparing actual and predicted values for all four outcomes. The red dashed line represents the ideal fit line ($y = x$). The close clustering of points around the diagonal line across all panels reflects the high accuracy of Random Forest predictions.

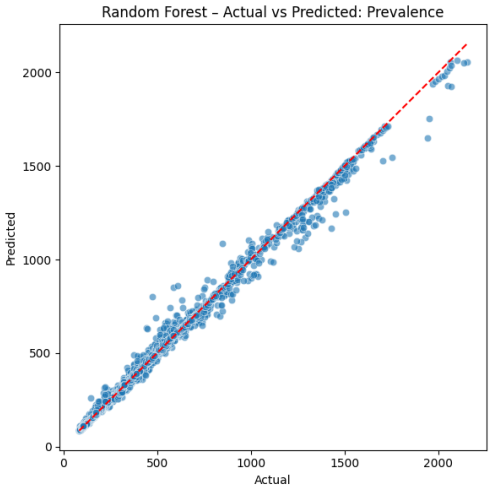
The **Incidence** model shows exceptionally tight alignment with the ideal line, matching its superior R^2. Predictions for **DALYs**, **YLDs**, and **Prevalence** also closely track true values, with only minimal dispersion observed at the distribution extremes.



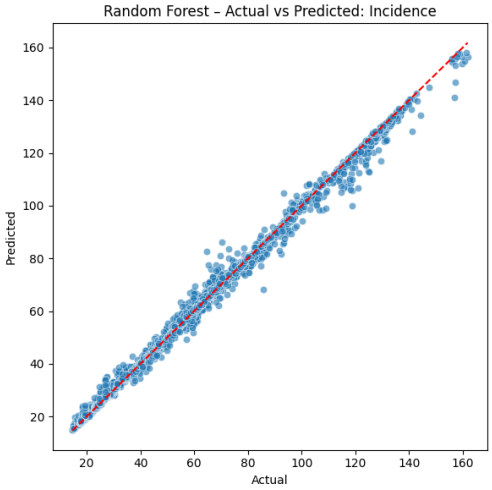
(a) Actual vs. Predicted - DALYs



(b) Actual vs. Predicted - YLDs



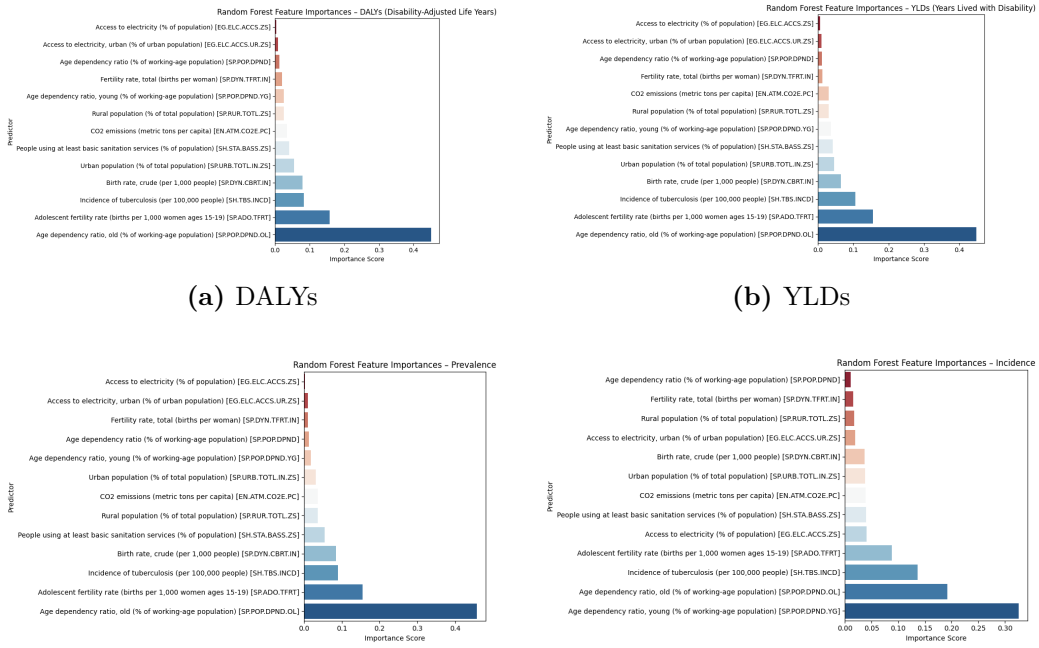
(c) Actual vs. Predicted - Prevalence



(d) Actual vs. Predicted - Incidence

Figure 4.19: Random Forest Regression: Actual vs. Predicted Values for Psoriasis Burden Outcomes. The red dashed line indicates ideal predictions ($y = x$).

Feature Importance Interpretation Figures 4.20a–4.20d show the top predictors identified by Random Forest models for each outcome. The **old-age dependency ratio** was the most influential predictor for **DALYs**, **YLDs**, and **Prevalence**, reflecting greater chronic burden in aging populations. For **Incidence**, the **young-age dependency ratio** was most important, highlighting the role of demographic structure. The **adolescent fertility rate** and **tuberculosis incidence** were also consistently relevant across outcomes, suggesting that fertility and broader health system conditions influence psoriasis burden. These patterns underscore the importance of population age structure and healthcare context in shaping psoriasis outcomes.



(a) DALYs **(b)** YLDs

(c) Prevalence **(d)** Incidence

Figure 4.20: Random Forest Feature Importances for Psoriasis Burden Outcomes. Bars indicate the relative contribution of each predictor to the model.

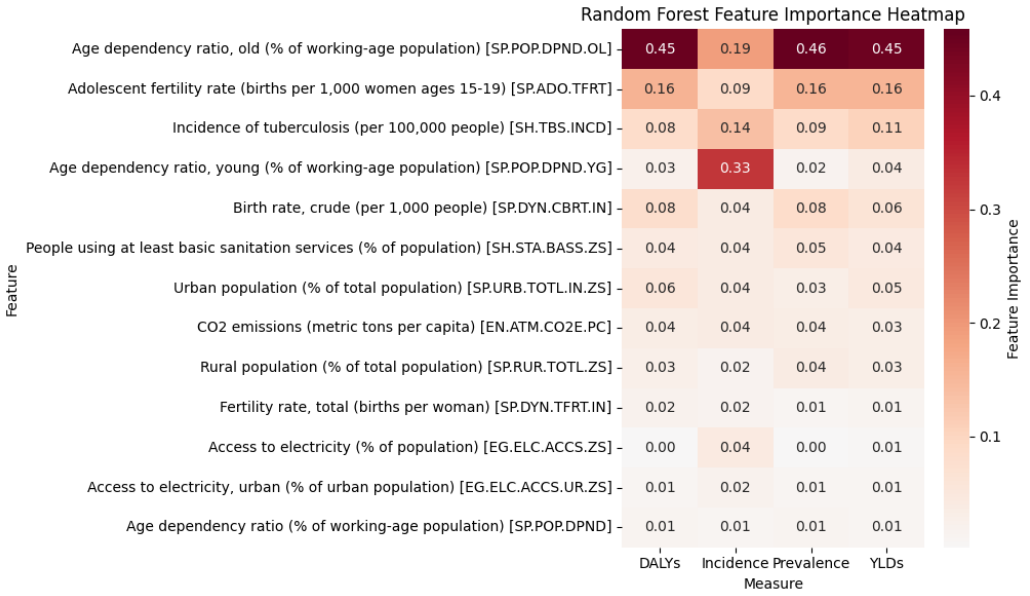


Figure 4.21: Heatmap of Random Forest Feature Importances Across Psoriasis Burden Outcomes

Key Insights

- The **old-age dependency ratio** emerged as the most important predictor for DALYs, YLDs, and Prevalence, reaffirming the role of aging populations in chronic psoriasis burden.
- **Young-age dependency ratio** was the most influential variable for predicting Incidence, likely reflecting dynamics in population structure and new case emergence.
- **Adolescent fertility rate** and **tuberculosis incidence** were important across models, possibly reflecting broader health system and demographic profiles.
- Random Forest models consistently emphasized demographic indicators over infrastructure-based ones like electricity or sanitation access.

Model Performance Comparison Table 4.7 summarizes the performance of all three regression models-Ordinary Least Squares (OLS), Ridge Regression, and Random Forest (RF)-across the four psoriasis burden outcomes. Performance was assessed using the R^2 score (explained variance) and Mean Squared Error (MSE). Random Forest consistently achieved the highest R^2 and lowest MSE

values, highlighting its superior predictive capability, particularly for non-linear patterns.

Table 4.7: Model Performance Comparison Across Regression Techniques

Measure	R^2 (OLS)	R^2 (RF)	R^2 (Ridge)	MSE (OLS)	MSE (RF)	MSE (Ridge)
DALYs	0.6243	0.9543	0.6243	379.73	43.67	379.73
Incidence	0.6949	0.9628	0.6949	285.09	27.26	285.09
Prevalence	0.6275	0.9534	0.6275	51856.93	2883.66	51856.93
YLDs	0.6234	0.9542	0.6234	380.63	43.75	380.63

4.3 Key Findings from Multivariate Analysis

The modeling analysis revealed key insights into how contextual development indicators relate to the global burden of psoriasis. Ordinary Least Squares (OLS) provided a transparent baseline and demonstrated that a substantial proportion of variance in psoriasis outcomes could be explained by demographic and infrastructural factors. However, issues of multicollinearity highlighted the need for regularization, which was effectively addressed by Ridge Regression. Ridge improved model stability without compromising interpretability, reinforcing the influence of age structure and urban development.

Random Forest Regression emerged as the most powerful approach, outperforming both OLS and Ridge in predictive accuracy across all outcomes. Its non-parametric nature allowed it to capture complex, nonlinear relationships and interactions among predictors that linear models could not fully address. Feature importance analysis consistently identified demographic indicators particularly the old-age and young-age dependency ratios as primary drivers of psoriasis burden.

Collectively, these findings underscore the value of combining linear and nonlinear modeling techniques to gain both interpretability and predictive strength. They affirm that demographic structure, reproductive health indicators, and broader development contexts play a pivotal role in shaping the burden of psoriasis at the population level.

Chapter 5

Spatiotemporal Analysis of Psoriasis Determinants

This chapter explores how the burden of psoriasis has evolved over space and time, analyzing trends across countries, development clusters, and key time points. The goal is to understand not just where the burden is concentrated, but how it shifts in relation to structural determinants such as infrastructure, healthcare access, and demographic transition.

1 Temporal Trends

To examine how psoriasis burden measures (DALYs, YLDs, Prevalence, and Incidence) evolved over time:

- We computed the yearly **median values** for each measure across all countries.
- We also plotted cluster-wise median trends over time (Cluster 0, 1, and 2), based on the PCA-based cluster assignments.

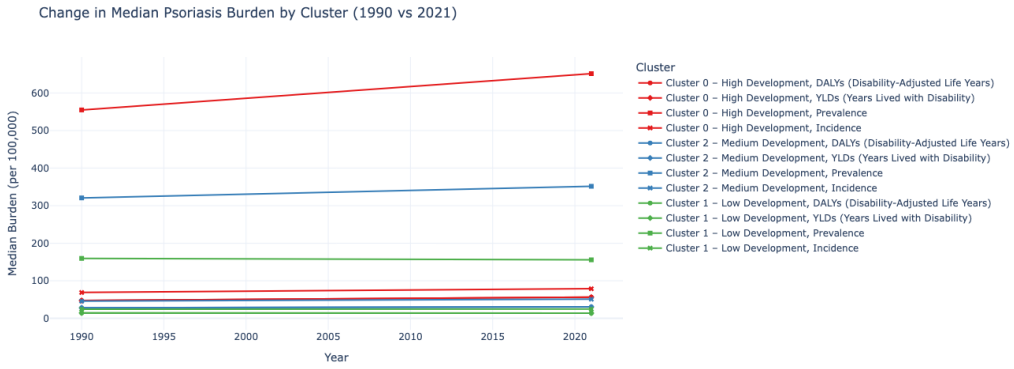


Figure 5.1: Temporal Trends of Psoriasis Burden Measures by Cluster (1990-2021)

As seen in Figure 5.1, countries in **Cluster 0** tend to show higher values for non-communicable disease burden (e.g., DALYs and Prevalence), potentially due to better healthcare access and diagnosis. In contrast, countries in **Cluster 1** report consistently lower values, possibly due to underdiagnosis or limited healthcare reporting infrastructure.

1.1 Temporal Trends in Psoriasis Burden by Cluster (1990-2021)

To understand how psoriasis burden has changed over time across development groups, we examined the temporal trends in the median values of four key indicators - **DALYs**, **YLDs**, **Prevalence**, and **Incidence** - from 1990 to 2021. These plots provide a longitudinal view of burden trajectories for each of the three clusters derived earlier:

- **Cluster 0 - High Development**
- **Cluster 1 - Low Development**
- **Cluster 2 - Medium Development**

The following plots show that Cluster 0 consistently maintains the highest median burden across all indicators. Notably, these burdens have increased slightly or remained stable over time. Clusters 1 and 2 exhibit lower median values with flatter or slightly fluctuating trends, possibly due to underdiagnosis, limited healthcare access, or reporting inconsistencies.

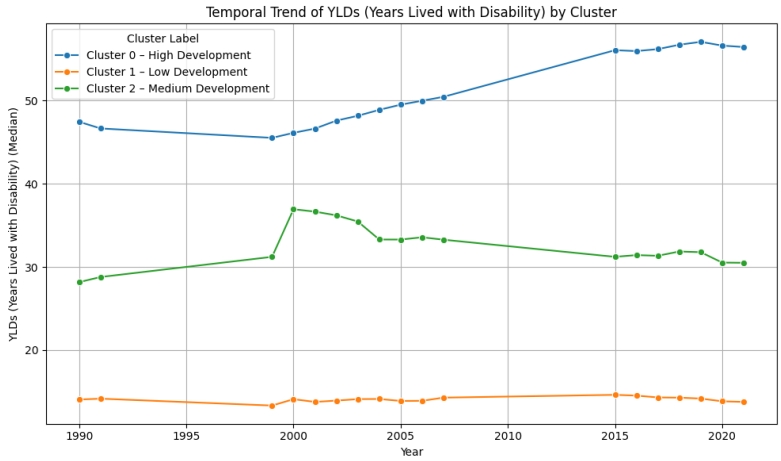


Figure 5.2: Temporal Trend of YLDs (Years Lived with Disability) by Cluster (1990-2021)

Interpretation: YLDs show a clear increasing trend in Cluster 0, suggesting greater chronic disease recognition or aging populations. Cluster 2 shows an early peak followed by stabilization, while Cluster 1 remains consistently low.

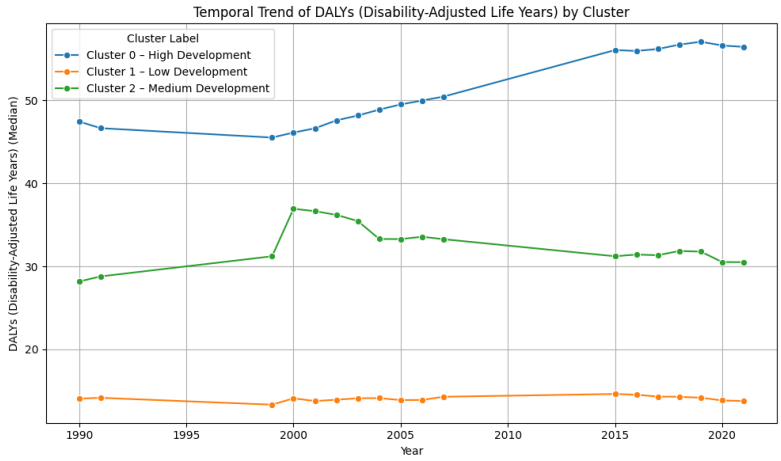


Figure 5.3: Temporal Trend of DALYs (Disability-Adjusted Life Years) by Cluster (1990-2021)

Interpretation: DALYs follow a similar pattern to YLDs, with Cluster 0 exhibiting the steepest growth. This supports the idea that high-development countries are experiencing increasing psoriasis impact, possibly due to longer lifespans and lifestyle factors.

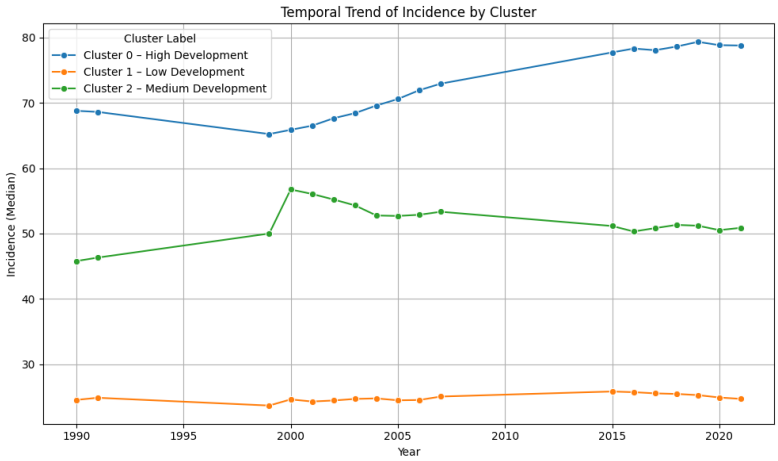


Figure 5.4: Temporal Trend of Incidence by Cluster (1990-2021)

Interpretation: Incidence rates are highest in Cluster 0 and show a gradual upward trend. Cluster 2 displays a peak around 2000 followed by decline, while Cluster 1 remains flat, indicating potential underreporting or diagnostic limitations.

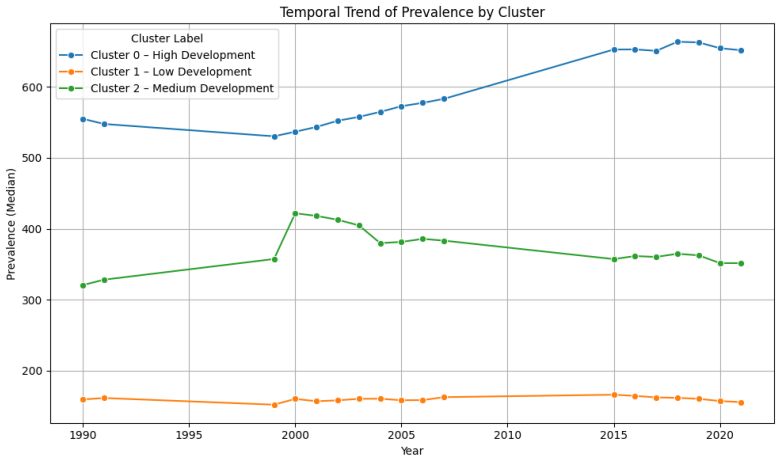


Figure 5.5: Temporal Trend of Prevalence by Cluster (1990-2021)

Interpretation: Prevalence remains significantly higher in Cluster 0 throughout the timeline, with visible increases. Cluster 2 shows moderate levels, while Cluster 1's trend remains relatively constant and low.

Summary: These temporal trends reinforce previous findings that psoriasis burden is not only higher in more developed regions but is also increasing over

time. In contrast, low-development countries show consistently lower levels, likely reflecting structural or data limitations, rather than true absence of burden.

2 Spatial Maps

To understand the global distribution and temporal evolution of psoriasis burden, we generated choropleth maps for each of the four key indicators: Prevalence, Incidence, YLDs, and DALYs. For each indicator, three maps were created: one for 1990, one for 2021, and one showing the percentage change over time. These visualizations help identify regional disparities, highlight progress or deterioration, and reveal areas with potential underreporting or increasing disease awareness. Mapping these trends provides valuable geographic context for interpreting global burden dynamics and guiding public health priorities.

2.1 Spatial Distribution of Prevalence

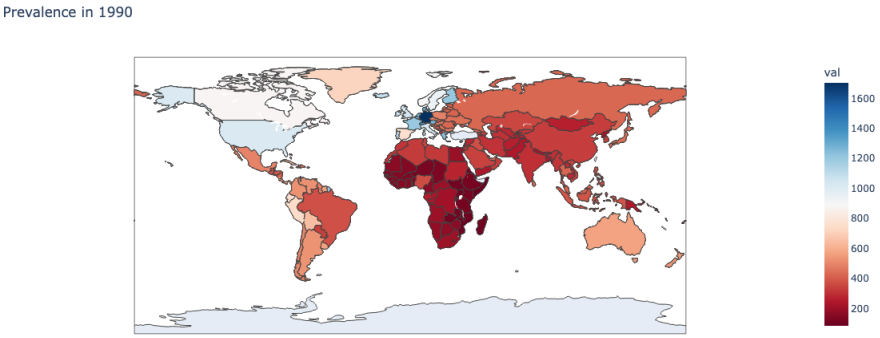


Figure 5.6: Prevalence in 1990

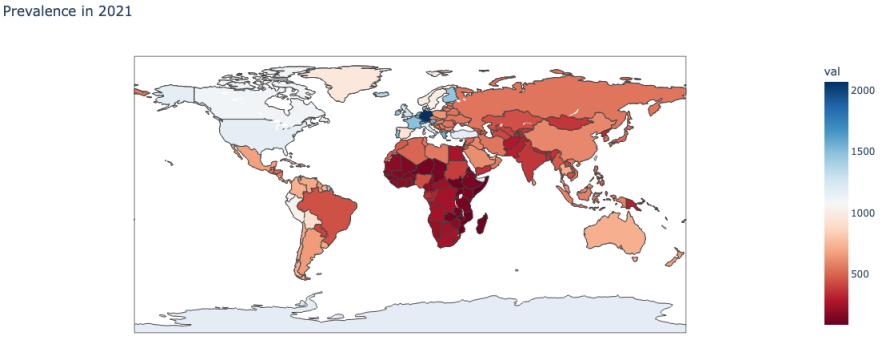


Figure 5.7: Prevalence in 2021

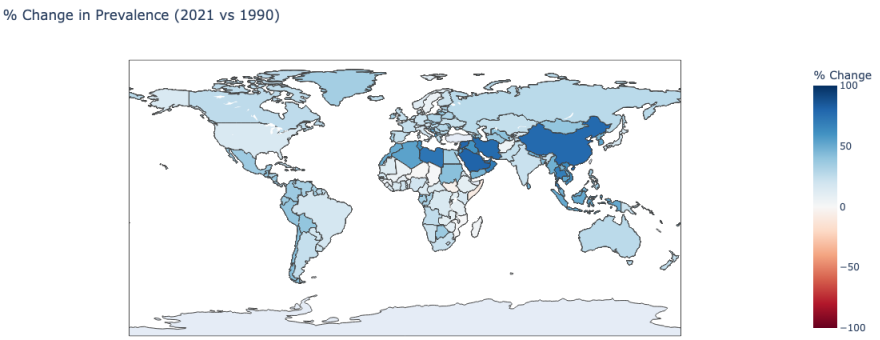


Figure 5.8: Percentage Change in Prevalence (1990-2021)

The choropleth maps indicate substantial increases in prevalence across many middle-income countries, particularly in the Middle East and Asia. Improvements in diagnostic systems and aging populations may partly explain these increases.

2.2 Spatial Distribution of Incidence

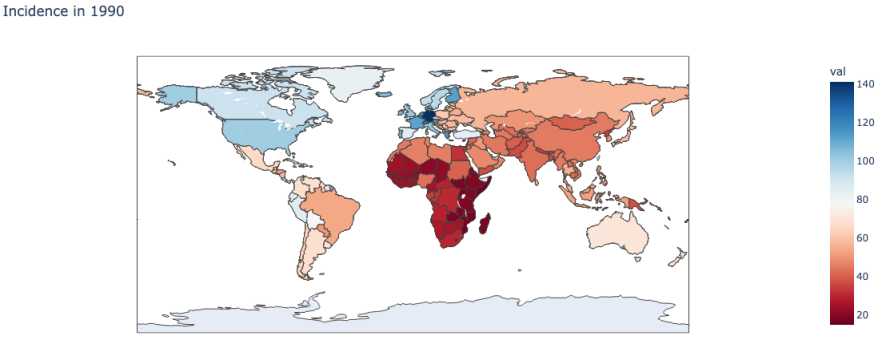


Figure 5.9: Incidence in 1990

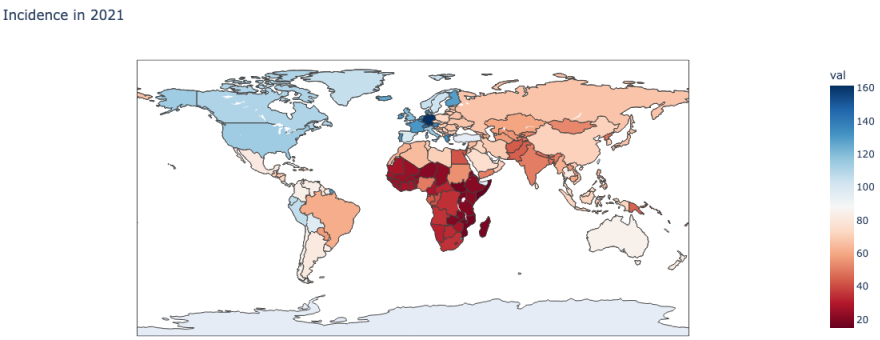


Figure 5.10: Incidence in 2021

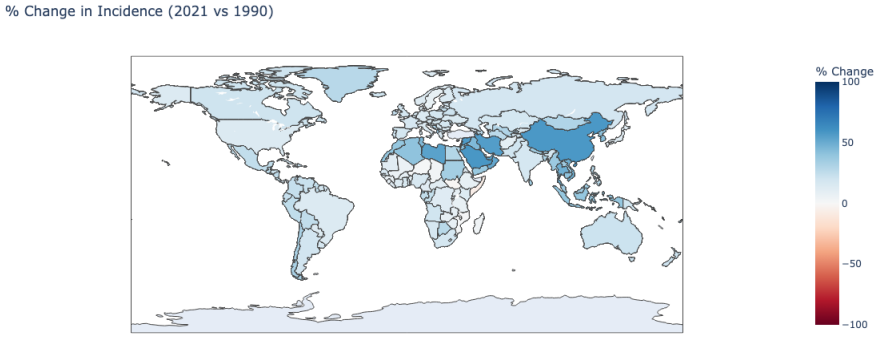


Figure 5.11: Percentage Change in Incidence (1990-2021)

The spatial distribution of incidence shows increasing detection in countries transitioning to higher development, possibly reflecting better health infrastructure and reporting mechanisms.

2.3 Spatial Distribution of DALYs

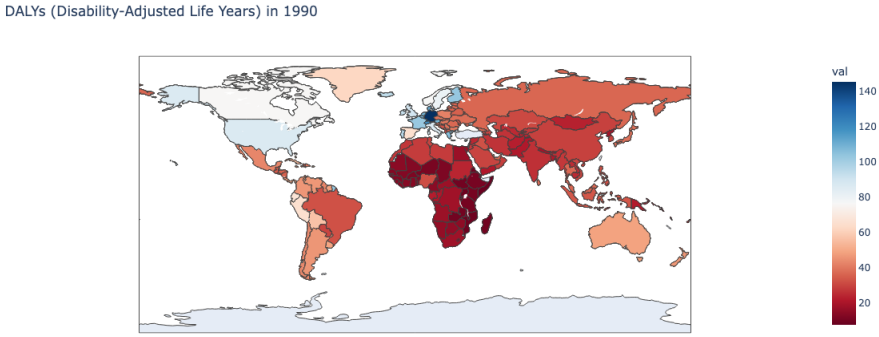


Figure 5.12: DALYs per 100,000 in 1990

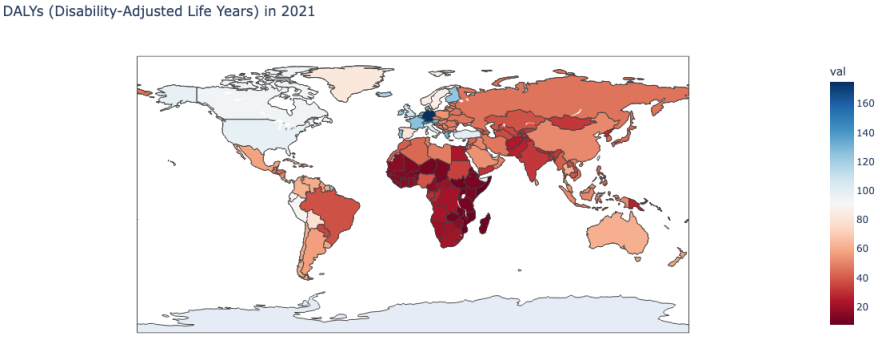


Figure 5.13: DALYs per 100,000 in 2021

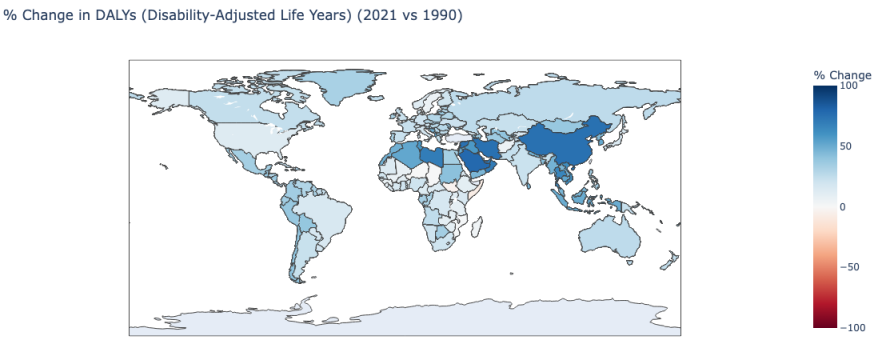


Figure 5.14: Percentage Change in DALYs (1990-2021)

Countries in the Middle East, North Africa, and South Asia exhibit sharp rises in DALYs. This reflects both improved reporting capacity and increasing burden due to demographic shifts.

2.4 Spatial Distribution of YLDs

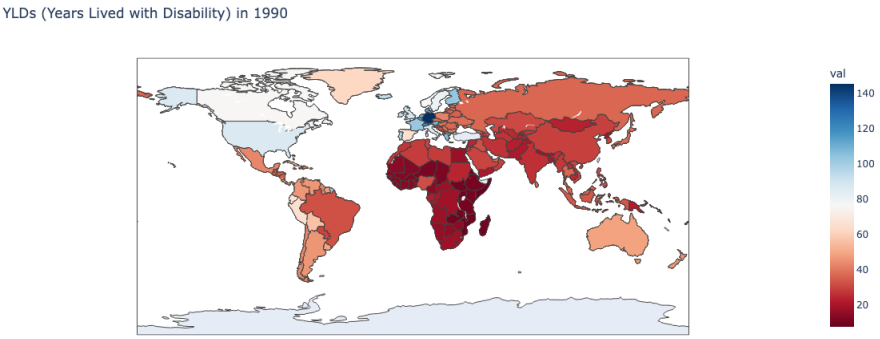


Figure 5.15: YLDs per 100,000 in 1990

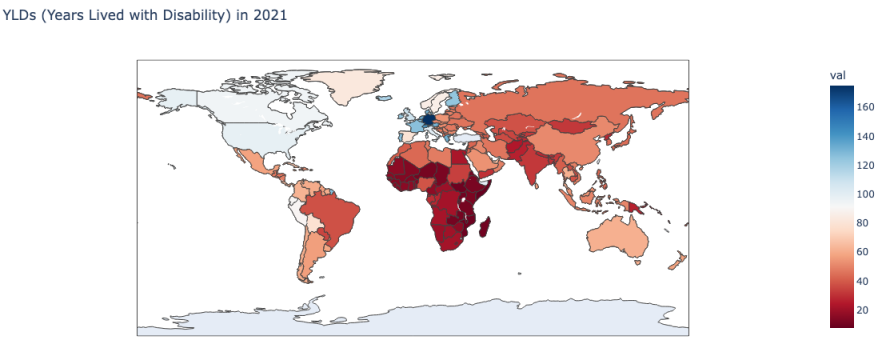


Figure 5.16: YLDs per 100,000 in 2021

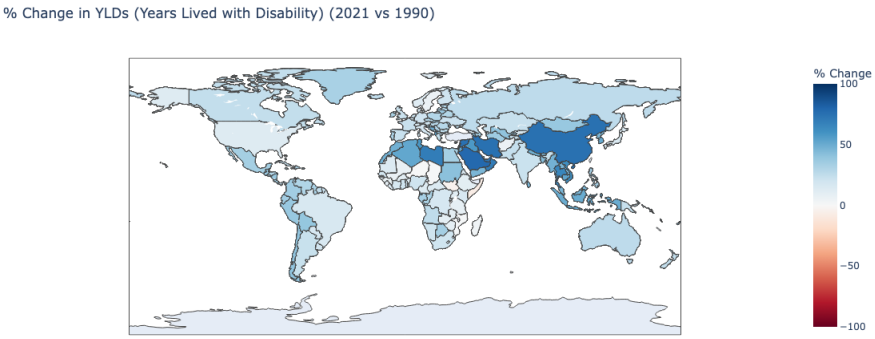


Figure 5.17: Percentage Change in YLDs (1990-2021)

The increase in YLDs is widespread, especially in emerging economies. These changes suggest improved survival and chronic disease management, contributing to longer-lived disability cases.

Table 5.1: Top 5 Countries by Percentage Change in Psoriasis Burden (1990–2021)

DALYs	**Δ Value**	**% Change**	**YLDs**	**Δ Value**	**% Change**
Maldives	22.96	86.05	United Arab Emirates	27.01	72.38
Saudi Arabia	23.43	78.52	Oman	20.26	70.23
China	22.00	75.30	Libya	19.68	69.86
Iran, Islamic Rep.	20.02	75.29	Thailand	24.28	65.79
Syrian Arab Rep.	18.31	75.19	Turkiye	26.15	64.62
Prevalence	**Δ Value**	**% Change**	**Incidence**	**Δ Value**	**% Change**
Viet Nam	216.29	62.93	Palau	23.21	44.71
Tunisia	210.46	61.38	Korea, Rep.	20.98	42.45
Albania	200.91	59.85	Cambodia	17.04	40.88
Iraq	221.26	59.69	Korea, DPR	13.89	40.82
Mauritius	257.56	59.36	Jordan	18.67	40.78

3 Spatio-Temporal Cluster Comparison

To evaluate how the burden of psoriasis differs across countries grouped by development level, we conducted a cluster-wise comparison using **PCA-based clustering** (Cluster 0: High Development, Cluster 1: Low Development, Cluster 2: Medium Development). The boxplots below visualize the distribution of psoriasis burden indicators (DALYs, YLDs, Prevalence, Incidence) across these clusters for the years 1990 and 2021.

To assess the statistical significance of differences between clusters, we applied the **Kruskal-Wallis test**, a non-parametric method for comparing distributions across more than two independent groups.

3.1 Boxplots of Burden by Cluster

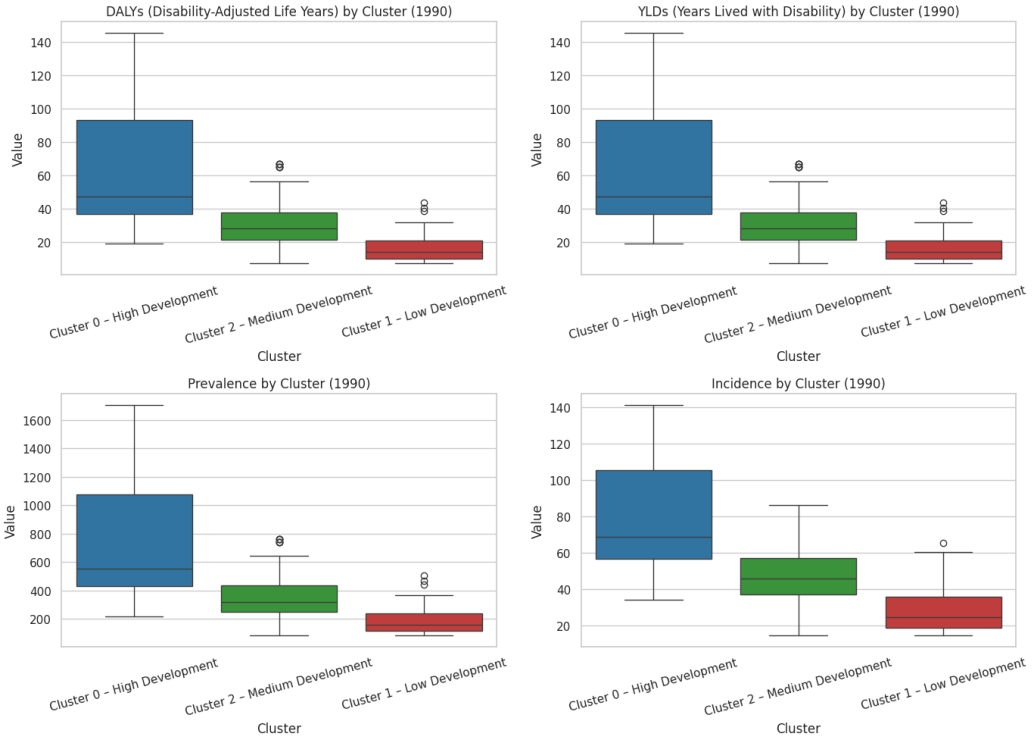


Figure 5.18: Psoriasis Burden by Cluster in 1990

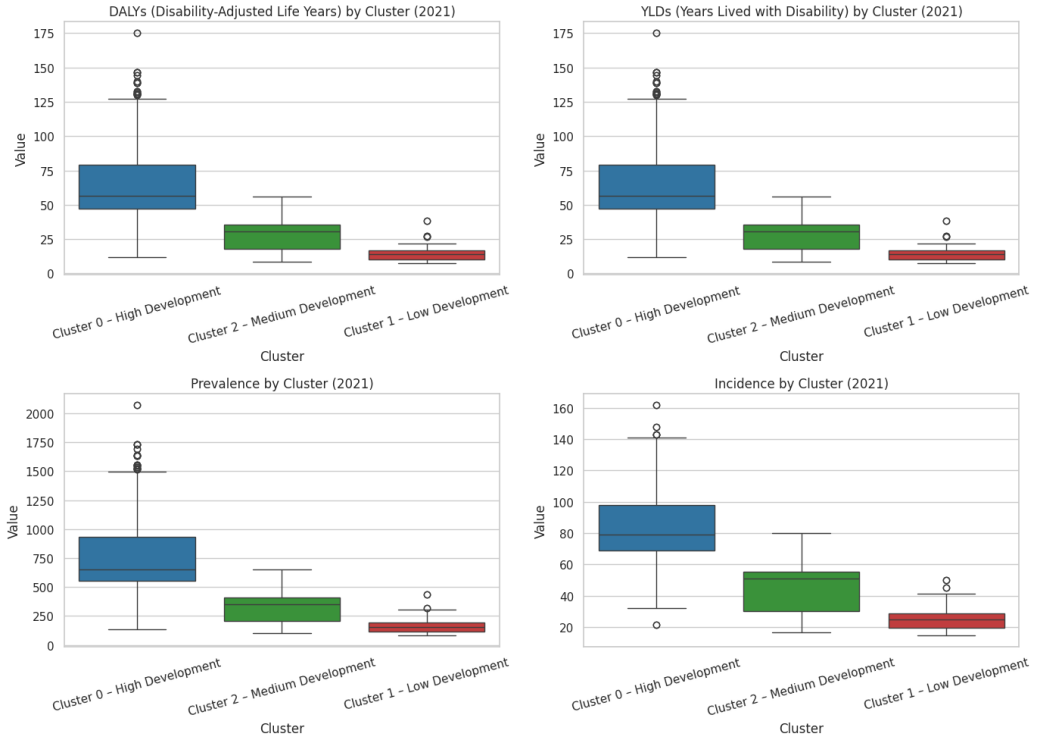


Figure 5.19: Psoriasis Burden by Cluster in 2021

The boxplots reveal persistent disparities in disease burden between development clusters. Cluster 0 (high development) consistently shows higher median values across all measures, likely reflecting better detection. Cluster 1 (low development) has the lowest values, with limited variation. Over time, Cluster 2 shows a gradual upward shift, suggesting increasing burden in developing nations.

3.2 Key Results

- Psoriasis burden measures (DALYs, YLDs, Prevalence, Incidence) **differ significantly across clusters** ($p < 0.001$ in most years).
- **Cluster 0** shows the highest median burden, reflecting better detection/reporting in more developed settings.
- **Cluster 1** shows the lowest, likely due to underdiagnosis or limited healthcare access.

3.3 Cluster-Wise Median Burden Summary (1990 vs 2021)

This subsection provides a summary of the median psoriasis burden across development-based clusters (Cluster 0 - High Development, Cluster 1 - Low Development, Cluster 2 - Medium Development) in 1990 and 2021. The values are reported as median and interquartile range (IQR) per 100,000 population. The results indicate persistent disparities across clusters, with Cluster 0 consistently showing the highest median burden across all four measures. These findings reflect the role of development in healthcare access, diagnosis, and reporting accuracy over time.

DALYs (Disability-Adjusted Life Years):

Year	Cluster 0	Cluster 1	Cluster 2
1990	47 (37-94)	14 (10-21)	28 (22-38)
2021	56 (47-79)	14 (10-17)	30 (18-35)

Table 5.2: Median (IQR) of DALYs per 100,000 by Cluster

YLDs (Years Lived with Disability):

Year	Cluster 0	Cluster 1	Cluster 2
1990	47 (37-94)	14 (10-21)	28 (22-38)
2021	56 (47-79)	14 (10-17)	30 (18-35)

Table 5.3: Median (IQR) of YLDs per 100,000 by Cluster

Prevalence:

Year	Cluster 0	Cluster 1	Cluster 2
1990	555 (429-1077)	160 (117-239)	321 (248-434)
2021	651 (552-934)	156 (118-195)	352 (205-409)

Table 5.4: Median (IQR) of Prevalence per 100,000 by Cluster

Incidence:

Year	Cluster 0	Cluster 1	Cluster 2
1990	69 (57-106)	25 (19-36)	46 (37-57)
2021	79 (69-98)	25 (19-29)	51 (30-55)

Table 5.5: Median (IQR) of Incidence per 100,000 by Cluster

4 Key Takeaways

- Psoriasis burden is consistently higher in countries with higher development profiles (Cluster 0).
- Countries progressing into higher clusters over time show increasing burden, likely due to improved healthcare systems, diagnosis rates, and aging populations.
- Lower-cluster countries may have underestimated burdens due to infrastructural limitations.

The spatiotemporal analysis revealed marked geographic and temporal disparities in the burden of psoriasis from 1990 to 2021. Developed countries consistently showed the highest levels of Prevalence, YLDs, and DALYs, suggesting a strong association between disease burden and diagnostic/reporting capabilities, healthcare access, and demographic profiles such as aging populations. In contrast, Developing and Underdeveloped countries exhibited lower reported burdens, which may partly reflect underdiagnosis and limited surveillance infrastructure.

Chapter 6

Conclusion

This thesis combined epidemiological, socioeconomic, environmental, and demographic data from over three decades to examine the worldwide burden of psoriasis from a variety of angles. Combining statistical, temporal, and spatial analyses has shown a startling trend: nations with older populations, lower fertility, and more developed infrastructure typically have higher psoriasis burdens, which is probably due to both improved diagnostic capabilities and easier access to healthcare. Conversely, in less developed areas, recorded burdens may be lower due to underreporting or conflicting public health priorities.

The study found logical country groupings that represent common development pathways and health system features using principal component analysis and clustering. In addition to providing information about regional differences, these clusters offer a framework for tracking changes in the burden of disease as countries advance. Additionally, both linear and nonlinear regression analyses revealed strong correlations between psoriasis outcomes and important contextual factors, confirming the influence of environmental exposures, healthcare infrastructure, and demographic shifts on disease burden.

This study has significant public health implications in addition to its methodological rigour. More attention should be paid to psoriasis as a sign of systemic inflammation, quality of life, and access to healthcare, as it is frequently overlooked in favour of more pressing health issues. This work emphasises the need for comprehensive, equity-oriented health policies that address the social determinants of chronic disease beyond treatment by placing psoriasis within larger development narratives. In conclusion, this thesis contributes a global, data-driven perspective to the epidemiology of psoriasis, offering a valuable foundation for future research, surveillance, and intervention strategies tailored to different developmental contexts.

Chapter 7

Acknowlegment

I would like to express my deepest gratitude to Professor Maugeri for his exceptional guidance, insightful feedback, and constant support throughout the course of this thesis. His expertise and mentorship not only shaped the direction of my research but also inspired me to think critically and pursue clarity in my work. I am truly fortunate to have had the opportunity to learn under his supervision.

I also wish to thank the faculty and staff of DMI and DEI at the University of Catania for providing an intellectually stimulating and supportive environment. Their dedication to academic excellence created the foundation upon which this work was built.

This thesis holds a profound personal meaning for me. Psoriasis is not just a research subject; it is a reality that affects people close to me. Several of my family members and friends live with this disease, and witnessing their challenges has been both humbling and motivating. Their experiences inspired me to dedicate myself to this topic with sincerity and purpose. I hope that this work contributes, in however small a way, to improving understanding and awareness of the burden they and many others endure.

I would like to thank my family and friends for their constant encouragement, patience, and belief in me. Your support during the most demanding moments of this journey has been my anchor.

I would like to thank myself for constantly challenging myself and trying new things in life and for keeping my head up and moving forward, even when things felt overwhelming.

Finally, I thank God for everything that I have in my life: friends who became my family, my individual achievements, the love and care of my parents and sister, being able to eat good food, and living a life of abundance. I am sincerely grateful.

Bibliography

[1] Theo Vos, Stephen S Lim, Cristiana Abbafati, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the global burden of disease study 2019. *The Lancet*, 396(10258):1204–1222, 2020.

[2] World Bank. World bank open data. `https://data.worldbank.org`, 2025. Accessed 2025-07-09.