Socioeconomic and Geographic Variations that Impacts the Spread of Malaria

Thilanka Munasinghe, Aayush Nitin Maheshwarkar, Karan Bhanot
Rensselaer Polytechnic Institute
munast@rpi.edu, aayushnmaheshwarkar@gmail.com, bhanotkaran22@gmail.com

Abstract

Malaria is a preventable disease that brings death to millions of people around the world. Among the victims of malaria epidemic, children under age of five are the most vulnerable. Malaria can influence where people live, work and also the family fertility decisions. In addition to the direct economic impact, Malaria can cause indirect socioeconomic impacts as it affects schooling, tourism, and sporting events. This epidemic has heavily impacted the developing countries, specially the ones that are in Sub-Sahara region of Africa. During our study, we explored the data on six indicators used by United Nations and World Heath Organization to evaluate the impact of Malaria and the Gross Domestic Product (GDP) data from chosen 34 countries. We conducted a cluster analysis on countries based on the number of confirmed deaths due to Malaria and the GDP using multiple clustering algorithms. We observe a negative relation with GDP and number of Malaria deaths.

Introduction

According to the United Nations (UN) and World Health Organization (WHO), every two minutes, a child under the age of five dies due to Malaria (WHO 2019). Every year, Malaria disease causes hundreds of thousands of deaths around the world. Malaria is vector-borne (mosquitoes are the vectors) yet preventable disease. The highest mortalities due to Malaria are reported in developing countries. We analyzed the publicly available data from the UN, UNICEF, and WHO based on several indicators specified by those organizations during this study. The work that is outlined in this paper aligns with the United Nations Sustainable Development Goals (SDG), specifically, the third goal (SDG3), which targets the "Ensure healthy lives and promote well-being for all at all ages" (United Nations 2020). Under the SDG3 targets, the sub-target 3.3 focuses on ending the epidemic of Malaria by the year 2030. *(SDG-3.3: By 2030, ending the epidemics of AIDS, Tuberculosis, Malaria and neglected tropical diseases and combat hepatitis, water-borne diseases and other communicable diseases).*

AAAI Fall 2020 Symposium on AI for Social Good.

Several research studies have been conducted on Malaria impact on Gross Domestic Product (GDP) in several counties (Orem et al. 2012; Sachs and Malaney 2002), especially in developing nations where the severity of Malaria is high. There are several studies conducted on Sub-Saharan African countries and their impact on Malaria morbidity on GDP (Orem et al. 2012). According to Sachs and Malaney (Sachs and Malaney 2002), the effects of Malaria are not short term, and it can last for generations and can have a long-term impact on the growth of an economy. During their study, they have explained how counties with active Malaria cases affect both affluent and poor communities in an equal manner (Sachs and Malaney 2002; Orem et al. 2012). Several countries have successfully controlled the Malaria epidemic. Before 1940, there were many Malaria cases in the southern United States and its territories. From 1947 to 1951, the United States conducted a successful Malaria eradication program. (Center for Disease Control and Prevention July 2018). Sri Lanka, an island nation with tropical weather, has successfully eliminated the Malaria epidemic through nationwide AntiMalaria Campaigns (AMC) (Wijesundere and Ramasamy 2017).

Malaria Intervention Techniques

There are several Malaria intervention techniques that have being deployed with successful results. Among those intervention techniques, usage of Long Lasting Insecticidal Nets (LLINs) also known as Insecticide-Treated mosquito Net (ITN) (Center for Disease Control and Prevention 2019b), Indoor Residual Spraying (ISR) of insecticide (Center for Disease Control and Prevention 2019a), and Seasonal Malaria Chemoprevention (SMC) (World Health Organization 2013) are the most prominent defensive mechanisms against Malaria.

According to the article "Bed Nets for Benin" published by the *Gates Notes* by Bill Gates Foundation (Bill Gates 2020) that provides philanthropic assistance to many countries to battle Malaria and other diseases around the globe states that resource management is one of the challenging tasks when it comes to distributing the Malaria intervention items such as bed nets. During their study, they found out that Benin's public health workers have been using traditional pen and paper to record-keeping of resource distribution ledgers. They also stated that, in the past, Benin's public health workers used paper ledgers as they collect data during home-to-home visits to distribute the mosquito nets. Due to the traditional pen and paper record-keeping mechanism, there were difficulties identifying and tracking the families who did not get the mosquito bed nets on time. Benin has recently moved from the traditional pen and paper approach and introduced mobile phones to enter real-time data and maintain a proper database system. This home-to-home visit data has been coupled with the satellite maps and rural population data to adequately address and identify the people who miss out on the distribution of bed nets and other intervention resources. We see an opportunity to apply Machine Learning and Artificial Intelligence (AI) techniques to analyze the collected data to identify vulnerable communities within large populations and optimize the supply chains and distribution efforts. We believe that the AI community can explore these uncharted territories to contribute, from the moment of data collection, analysis, model building, and develop Early Warning Systems (EWS) in many stages to tackle the Malaria epidemic. Similarly, those techniques can be applied to tackle other vector-borne diseases such as Dengue, Zika, and West Nile virus caused by mosquitoes worldwide.

Impact on Economy by Malaria

The Gross Domestic Product (GDP) of a country is an indicator of a country's economic strength. Countries with higher GDP have strong spending capability on their healthcare and the country's people's well-being. In general, the stronger the country's GDP, the ability to spend more financial resources to tackle any epidemic in that country is higher. Higher GDP reflects the ability of a country's spending power when it comes to mitigating the adverse conditions and keeping epidemics under control with minimum impact. We observed a negative relationship with the number of deaths due to Malaria and GDP during our study. Countries with a higher number of Malaria cases have a lower GDP.

When it comes to evaluating Malaria's impact on society, it has branched out to direct impact and indirect impact. Regardless of whether the impact is direct or not, it is possible to identify three major dimensions that get affected. According to Impact of Malaria Morbidity on Gross Domestic Product in Uganda (Orem et al. 2012), there are three dimensions: (1) Health, (2) Social, and (3) Economic . During their study, Orem states that the health dimension's impact has been described in terms of premature deaths and loss of life years.

Uncontrolled and wild spread situations of malaria can affect the tourism industry and foreign investments. International tourism can contribute to the economic growth of a country. Most of the tropical weather conditions attract international tourism. However, this favorable weather condition also attracts vector-borne infectious diseases such as Malaria. (Rosselló, Santana-Gallego, and Awan 2017)

Impact on Society by Malaria

Malaria can influence where people live, work and also the family fertility decisions. In addition to the direct economic impact, Malaria can cause indirect socioeconomic impacts as it affects schooling, travel, and sporting events. Another indirect consequence of Malaria is the impact on population growth. According to (Yamada 1985; Sachs and Malaney 2002; Council, on Population et al. 1998; Handa 2000), families that live in the regions with a high number of cases of child mortality under the age of five tend to have high fertility rates due to the concern of losing children at a young age, and parents tend to have additional number of children in their family to replace the ones that they lose. During that study, they investigated the "child-survivor hypothesis" to show that parents make their fertility decisions to ensure at least one child will survive.

Data Acquisition and Project Work Flow

As shown on the project workflow diagram in Figure 1, we obtained the data on the number of reported confirmed cases of Malaria (World Health Organization 2014a) and the number of reported deaths due to Malaria (World Health Organization 2014b) from the official UN data repository, where it contains the collected data by the WHO. The GDP data was obtained from the UN Data repository. We conducted Exploratory Data Analytics (EDA), data cleaning and pre-processing prior to work on the cluster analysis. Later, we conducted the cluster comparison. The six Malaria indicators' data was obtained from UNICEF repository (UNICEF 2020). The acquired UNICEF dataset comprised of 21 columns and 2632 rows, with the collected data canvasing 87 countries and the 7 continents. The WHO dataset consisted of 10 columns and 1092 rows, with a focus on 99 countries spanning across 13 years from 2000 to 2013.

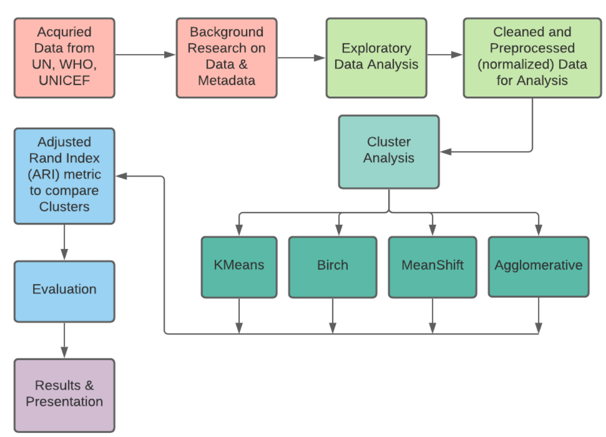

Figure 1: Data Acquisition and Project Work Flow Diagram

Exploratory Data Analysis

As will be evident from the following Exploratory Data Analysis, as the years went by, from 2005 to 2010, there were significant changes in priorities and preventive antimalarial measures (UNICEF 2020) that were deployed across various countries. Initial years from 2005 to 2009 had primarily focused to 3 major indicators, which have been listed below.

(i) **ITN use by children under 5:** Percentage of children (under age 5) who slept under an insecticide-treated mosquito net the night prior to the survey

(ii) **Households with at least one ITN:** Percentage of households with at least one insecticide-treated mosquito net (ITN)

(iii) **Pregnant women sleeping under ITN:** Percentage of pregnant women (aged 15-49 years) who slept under an insecticide-treated net the previous night

Later on, the number of indicators grew to 6, with additional three indicators that are listed below.

(iv) **IPTp for pregnant women:** Percentage of women (aged 15-49 years) who received three or more doses of intermittent preventive treatment during antenatal care visits during their last pregnancy

(v) **Febrite children under 5 who had a finger or heel stick:** Percentage of febrile children (under age 5) who had a finger or heel stick for Malaria testing

(vi) **Febrile children under 5 receiving ACT:** Percentage of febrile children (under age 5) receiving ACT (first line antimalarial drug), among those receiving any antimalarial drugs, being added with time

The indicators (iv), (v), and (vi) were introduced after the year 2010. The additional indicators added the post year 2010 could have contributed to an increase in well-being, a decrease in the number of confirmed reported cases of Malaria, and an increased focus on reducing the number of confirmed reported cases in pregnant women. The progression and changes in the deployed indicators can be seen through the graphs shown below that were implemented during the Exploratory Data Analysis (EDA). As the data for all countries is not available across all years, we explored the measures deployed by different countries across different years, focusing more on how different countries handle anti-malarial deployment measures.

As seen in Figure 2, we observe that in most countries, the most common preventive anti-malarial measure is to have at least one insecticide-treated mosquito net (ITN). Further, countries like Congo, Ethiopia, and Guinea lag behind in the deployment of all preventive measures in comparison to countries like Malawi, Senegal, etc. which have a high percentage of preventive measures deployment.

From Figure 3, we again identify that the most common preventive measure is the deployment of at least one ITN while other measures generally lag behind. Surprisingly, the approach to preventive measures is different across different countries. For example, the deployment of all three preventive measures is comparable in Benin and Cambodia, but the deployment of at least one ITN household preventive measure is significantly more focused in countries like Mali, Niger, etc. The reported confirmed Malaria cases dataset and the GDP dataset (The World Bank 2020) indicate that the internal prioritization of preventive measures between the countries could also have been a significant influencing factor in the deployment of preventive measures.

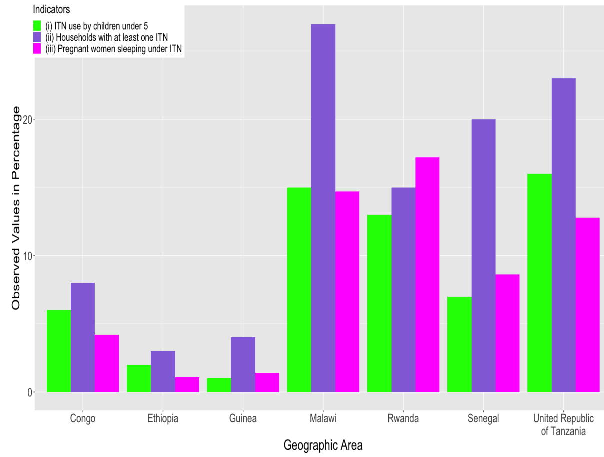

Figure 2: The observed percentage values of indicators across geographic regions for the year 2005

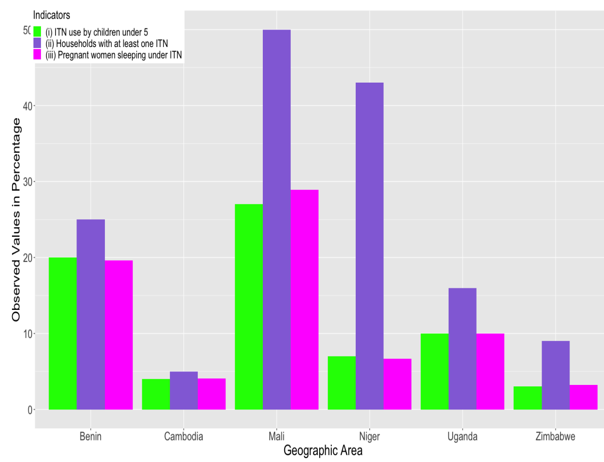

Figure 3: The observed percentage values of indicators across geographic regions for the year 2006

As shown in Figure 4, the number of preventive antimalarial measures increased starting from the year 2010. We observe that the countries shifted their anti-malarial measures to focus more towards febrile children and pregnant women. Post year 2010, the implementation of preventive anti-malarial measures and intermittent preventive treatment (ITP) were the point of major focus.

Cluster Analysis

We decided to use different clustering algorithms and compare their results as clustering can be an effective solution to facilitate the development of policies and measures for similar countries while ensuring efficient use of limited time and resources.

Clustering is a methodology that groups similar data

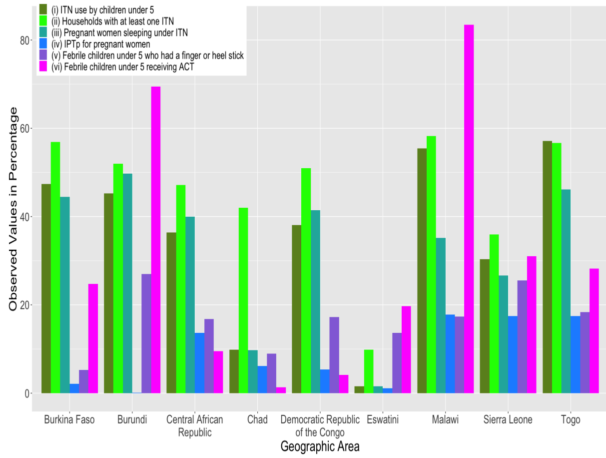

Figure 4: The observed percentage values of indicators across geographic regions for the year 2010

points together in clusters such that all data points in a cluster are more similar to each other than data points in other clusters. Unsupervised clustering is a class of clustering algorithms that groups unlabelled data points into specific groups. We applied unsupervised clustering algorithms like KMeans, Birch and other algorithms on a collection of countries to identify similarity between sparsely located countries based on any correlation between GDP values and confirmed number of deaths due to Malaria.

Table 1: Countries grouped by their continents

Continent	Countries
Africa	Angola, Burundi, Chad, Democratic Republic of the Congo, Gabon, Ghana, Madagascar, Mali, Niger, Senegal, South Africa, Sudan, United Republic of Tanzania
Asia	Azerbaijan, Bangladesh, Bhutan, Kyrgyzstan, Uzbekistan
North America	Costa Rica, Dominican Republic, El Salvador, Haiti, Mexico, Nicaragua, Panama
South America	Argentina, Bolivia (Plurinational State of), Brazil, Colombia, Ecuador, Paraguay, Peru, Suriname, Venezuela (Bolivarian Republic of)

We clustered the data based on the normalized confirmed number of deaths due to Malaria from year 2000 to 2012 for 34 countries (listed in Table 1) using KMeans (Dey 2016), Agglomerative Clustering (Kurita 1991), Birch (Zhang, Ramakrishnan, and Livny 1996) and MeanShift (Yizong Cheng 1995) algorithms with the scikit-learn Python package (Pedregosa et al. 2011). Normalization is an essential step before clustering to ensure that all features are given equal weights and noisy or redundant objects are avoided (Virmani, Shweta, and Malhotra 2015). The confirmed number

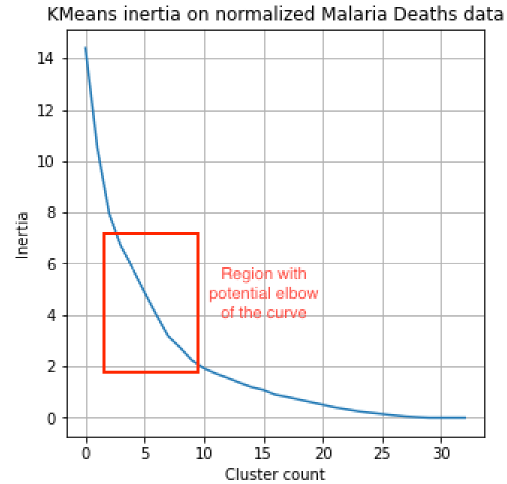

Figure 5: Elbow curve for inertia values across various clusters in KMeans clustering

of deaths due to Malaria vary a lot across years, ranging between 0 to over 21,000 deaths in a year, suggesting the need for normalization. Thus, we scaled the individual values to have unit normal using the 'L2' normalization.

We use the Elbow Method to identify the suitable number of clusters for the clustering algorithms (Kodinariya and Makwana 2013). We plot the elbow curve based on the KMeans inertia values as shown in Figure 5 to determine the number of clusters. In the elbow curve, the possible values for the elbow ranged between 3 & 7, and we decided to select 6 clusters for our analysis. The MeanShift clustering algorithm does not take the number of clusters as an argument and as a result, clustered the countries into 5 groups.

We decided to use Rand index adjusted for chance (ARI) metric to compare the clusters from the four algorithms. ARI is an evaluation metric that computes the similarity between the results of two clustering algorithms by considering all pairs in the resultant labels (Hubert and Arabie 1985). A value of 1 indicates a perfectly matching labelling while a value of 0 indicates a lack of similarity. We calculated the ARI value for every pair of clustering algorithm and the results are presented in Table 2.

Table 2: Rand index adjusted for chance (ARI) between various clustering algorithms

Algorithm 1	Algorithm 2	Score
KMeans	Agglomerative	0.753
KMeans	Birch	0.894
KMeans	MeanShift	0.264
Agglomerative	Birch	0.738
Agglomerative	MeanShift	0.248
Birch	MeanShift	0.199

On comparing the clusters from the four algorithms, we observe that KMeans, Birch and Agglomerative Clustering produced almost similar clusters, indicated by high ARI values as seen in Table 2. We also observe that the MeanShift model is dissimilar to the other algorithms. The ARI values for pairwise comparison of MeanShift with the other models (KMeans, Birch and Agglomerative Clustering) are always less than 0.3, indicating a lack of similarity. Furthermore, in the clusters generated by KMeans, Agglomerative Clustering and Birch, the two biggest clusters collectively have at least 20 countries as seen in Table 3. However, the clustering for the MeanShift model is significantly different as it has a single large cluster with more than 20 countries.

Table 3: Number of countries in each cluster for the clustering algorithms

KMeans	Agglomerative	Birch	MeanShift
8	9	12	22
7	7	8	5
7	7	7	4*
6	5	4	1
5	5	2	1
1	1	1	1

* Group of countries not clustered by MeanShift

To gain an in-depth understanding of the clusters of confirmed number of deaths due to Malaria and their correlation with Gross Domestic Product (GDP) values of a country, we deployed the KMeans algorithm on the data for the 34 countries across the years 2000-2012. Referencing the results from the elbow curve in Figure 5, we selected the number of clusters to be 6 again.

Figure 6 clusters the normalized confirmed number of deaths due to Malaria data for the 34 countries across the years 2000 to 2012. The clusters are colored in shades of red, where a darker color indicates a higher mean confirmed number of deaths due to Malaria. We observe two darkest groups in the middle of Africa, including Niger, Chad, the Democratic Republic of the Congo, etc. indicating areas with high confirmed deaths due to Malaria. On the other hand, Mexico and Kyrgyzstan have lighter shades due to lower confirmed deaths due to Malaria. Further, countries in South America like Bolivia, Brazil, and Peru have a moderate confirmed number of deaths due to Malaria when compared to other countries.

Figure 7 shows the clusters of the 34 countries based on the GDP values across the years 2000-2012. The darker the blue color shade, the higher the mean GDP value for the countries in the cluster. Brazil has the highest GDP values in South America, while other countries like Bolivia, Peru, and Argentina have relatively less GDP values. Countries in the middle and upper regions of Africa, like Mali, Niger, Chad, Sudan, etc. have the lowest GDP values across all countries.

From the clusters in Figure 6 and 7, we observe a negative correlation between the confirmed number of deaths due to Malaria and GDP values of a country. In Africa, we see that countries that have lower GDP values including Niger and Chad have higher confirmed number of deaths due to Malaria. In Brazil and Peru, the GDP values are higher but

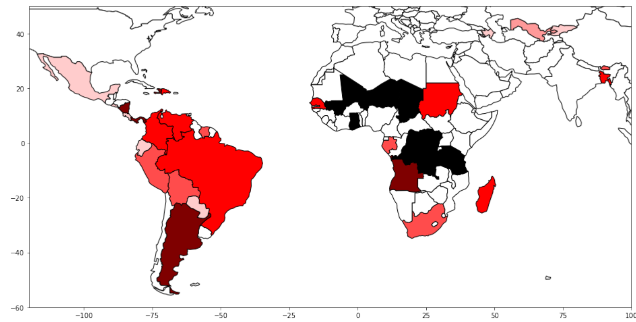

Figure 6: Countries clustered into 6 groups based on Malaria Deaths

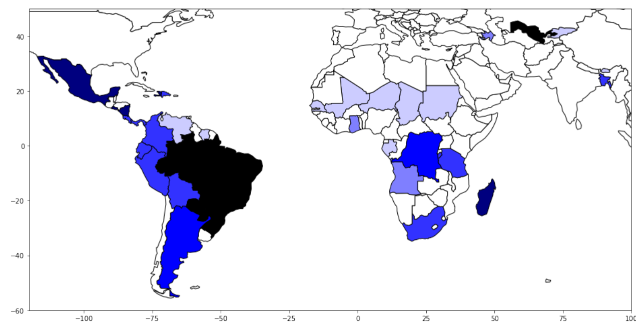

Figure 7: Countries clustered into 6 groups based on GDP

the confirmed number of deaths due to Malaria are lower. The same correlation is also observed for Mexico. However, this one-to-one direct negative correlation does not exist for all countries. For Argentina (in South America) and Democratic Republic of Congo (in Africa), the GDP values are relatively higher but the confirmed number of deaths from Malaria are still higher indicating that other factors also contribute towards the number of confirmed number of deaths due to Malaria in a country.

This cluster-wise analysis between the confirmed number of deaths due to Malaria and GDP values suggests that a strong negative correlation might exist between a country's GDP value and the number of Malaria Deaths within that country. An underlying factor can be that a country with relatively higher GDP has more economical power to allocate higher resources to combat Malaria spread and hence, curb the number of Malaria Deaths. Thus, as the GDP of a country increases, the Malaria Deaths are likely to decrease. Such a correlation and other contributing factors can be explored in depth as a future research study.

Conclusion

During this study, we observed a negative relationship with the country's GDP and the number of deaths due to malaria. This negative relationship indicates that countries with stronger economies can mitigate Malaria's adverse conditions compared to the counties with lower GDP. Among developing countries, we observe that the Sub-Sahara region of Africa had the highest number of Malaria infections and the confirmed number of deaths compared to other developing countries around the world. One of the UNICEF dataset limitations was the inconsistency in the available yearly data, where some countries did not have all the six indicators for each year, and as a result of that, we had to exclude the countries that did not contain all six indicators. Therefore, we propose and encourage data collecting authorities and agencies to keep consistent data records for the countries with the Malaria threat.

Future Work and Code Repository Access

Our code repository is available on GitHub at: https://github.com/thilankam/MalariaAnalysis We intend to keep this work as an open source project and plan on contributing to the UN's Sustainable Development Goals efforts. In the future, we are planning to implement an interactive data visualization dashboard based on these results and upcoming analysis work.

References

Bill Gates. 2020. Full coverage: Bed nets for Benin. Its bed net distribution system will help save lives from malaria and other diseases too, https://gatesnot.es/2RLHBIA, Accessed on: September 18, 2020.

Center for Disease Control and Prevention . 2019a. Indoor Residual Spraying. These mosquitoes are particularly susceptible to control through indoor residual spraying (IRS), https://bit.ly/32UnsWY, Accessed on: September 18, 2020.

Center for Disease Control and Prevention . 2019b. Insecticide-Treated Bed Nets. Insecticide-treated bed nets (ITNs) are a form of personal protection that has been shown to reduce malaria illness, https://bit.ly/33WlKEa, Accessed on: September 18, 2020.

Center for Disease Control and Prevention . July 2018. Elimination of Malaria in the United States (1947 — 1951). The term "elimination" is used when malaria transmission is no longer occurring in a specific geographic area, https://bit.ly/32TpLtx, Accessed on: September 20, 2020.

Council, N. R.; on Population, C.; et al. 1998. *From death to birth: mortality decline and reproductive change*. National Academies Press.

Dey, A. 2016. Machine learning algorithms: a review. *International Journal of Computer Science and Information Technologies* 7(3): 1174–1179.

Handa, S. 2000. The impact of education, income, and mortality on fertility in Jamaica. *World Development* 28(1): 173–186.

Hubert, L.; and Arabie, P. 1985. Comparing partitions. *Journal of Classification* 2(1): 193–218. ISSN 1432-1343. doi:10.1007/BF01908075. URL https://doi.org/10.1007/BF01908075.

Kodinariya, T.; and Makwana, P. 2013. Review on Determining of Cluster in K-means Clustering. *International Journal of Advance Research in Computer Science and Management Studies* 1: 90–95.

Kurita, T. 1991. An efficient agglomerative clustering algorithm using a heap. *Pattern Recognition* 24(3): 205 – 209. ISSN 0031-3203. doi:https://doi.org/10.1016/0031-3203(91)90062-A. URL http://www.sciencedirect.com/science/article/pii/003132039190062A.

Orem, J. N.; Kirigia, J. M.; Azairwe, R.; Kasirye, I.; and Walker, O. 2012. Impact of malaria morbidity on gross domestic product in Uganda. *International archives of medicine* 5(1): 12.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.

Rosselló, J.; Santana-Gallego, M.; and Awan, W. 2017. Infectious disease risk and international tourism demand. *Health policy and planning* 32(4): 538–548.

Sachs, J.; and Malaney, P. 2002. The economic and social burden of malaria. *Nature* 415(6872): 680–685.

The World Bank. 2020. GDP (Current US$). Data retrieved from World Bank national accounts data, and OECD National Accounts data files, https://data.worldbank.org/indicator/NY.GDP.MKTP.CD, Accessed on: September 14, 2020.

UNICEF. 2020. Cross-sector indicators. Data retrieved from UNICEF Data Warehouse, https://bit.ly/33wFcqD, Accessed on: September 14, 2020.

United Nations. 2020. SDG3: Good Health and well-being . SDG3: Ensure healthy lives and promote well-being for all at all ages, https://sdgs.un.org/goals, Accessed on: September 18, 2020.

Virmani, D.; Shweta, T.; and Malhotra, G. 2015. Normalization based K means Clustering Algorithm. *CoRR* abs/1503.00900. URL http://arxiv.org/abs/1503.00900.

WHO. 2019. Malaria. Every 2 minutes child under age 5 dies, https://bit.ly/2RLFxAk, Accessed on: September 14, 2020.

Wijesundere, D. A.; and Ramasamy, R. 2017. Analysis of historical trends and recent elimination of malaria from Sri Lanka and its applicability for malaria control in other countries. *Frontiers in public health* 5: 212.

World Health Organization. 2013. Seasonal Malaria Chemoprevention (SMC). , https://bit.ly/3kIJX7F, Accessed on: September 18, 2020.

World Health Organization. 2014a. Malaria - Number of reported confirmed cases. , https://bit.ly/302KfxZ, Accessed on: August 10, 2020.

World Health Organization. 2014b. Malaria - Number of reported deaths. , https://bit.ly/302KfxZ, Accessed on: August 10, 2020.

Yamada, T. 1985. Causal relationships between infant mortality and fertility in developed and less developed countries. *Southern Economic Journal* 364–370.

Yizong Cheng. 1995. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17(8): 790–799.

Zhang, T.; Ramakrishnan, R.; and Livny, M. 1996. BIRCH: An Efficient Data Clustering Method for Very Large Databases. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, SIGMOD '96, 103–114. New York, NY, USA: Association for Computing Machinery. ISBN 0897917944. doi:10.1145/233269.233324. URL https://doi.org/10.1145/233269.233324.