Climate Change and Infectious Diseases

The Effects of Climate and Human Factors on Malaria Outbreaks

Kamen Shah
CU Boulder
kash1057@colorado.edu

Steve Su
CU Boulder
susc@colorado.edu

Jimmy Young
CU Boulder
jiyo4476@colorado.edu

Jacob Reed
CU Boulder
jare4857@colorado.edu

Derek Sessions
CU Boulder
dese0351@colorado.edu

Abstract

When we set out on this project, the first thing we began to look at was climate data and how we could use a set of climate data to find novel and useful patterns. From this we started to look for areas in which we believed climate could play an interesting role. This led us into the realm of Healthcare data and ultimately on the idea of infectious diseases. From this we started to look at how climate may play a role in furthering or limiting the breadth and scope of an outbreak of an infectious disease. To limit the scope of our search into something of manageable size and magnitude we decided to narrow our focus onto malaria outbreaks in several Central American countries. As we began our analysis, we realized that there are many more factors to consider than just climate when looking at an outbreak. It was at this point we expanded our data sets to not only include climate data sets and malaria outbreak data sets but also data sets from the World Bank which contained data on human factors relative to the specific country such as gross domestic product (GDP), percent of population that is immunized, and other key attributes. Thus, we finalized the questions we were trying to answer into "Is there a relationship between malaria outbreaks and climate data and\or human data?", "Assuming climate and human factors are involved with malaria outbreaks which attributes for each play the largest roles and, if possible to determine, which main factor (climate or human) was more closely linked with malaria outbreaks?", and "Is there a pattern or relationship within the data to suggest means to limit outbreaks of malaria and, as an extension, other infectious diseases?"

We found that there is a distinct correlation between human factors and malaria outbreaks although it was not always with expected attributes. We also found there to be a distinct correlation for climate data and malaria outbreaks, based on our findings human factors were usually the better indicator for malaria outbreaks than climate factors alone.

Introduction

There were three main questions we were trying to answer as outlined in the abstract. The first is "Is there a relationship between malaria outbreaks and climate data and\or human data?" This question is simple enough in that the whole purpose of this project was to look at different sets of data and look to see if patterns and relationships between them could be found. This is important because it is the foundation of what data mining is and if there is no pattern or relationships between the data sets then we have failed to find new and interesting information. The next question we sought to answer was "Assuming climate and human factors are involved with malaria outbreaks which attributes for each play the largest roles and, if possible to determine, which main factor (climate or human) was more closely linked with malaria outbreaks?" This question is a natural progression from the first as it looks to take a comparative approach both within and between the two sets of data as they relate to malaria outbreaks. This question is important to answer because when searching for new patterns and relationships that are potentially novel and interesting it makes sense to look for the stronger correlations as these tend to be the most useful. The final question we looked to answer was "Is there a pattern or relationship within the data to suggest means to limit outbreaks of malaria and, as an extension, other infectious diseases?" This question completes the progression of mining data as you move from 1. Searching for patterns and relationships 2. Which patterns and relationships are strongest and potentially most insightful 3? How can this newly mined information be applied and made useful. This question (or type of question) is the most important to answer as it is the one that has potentially meaningful and lasting applications outside of mining data simply for the joy of it. Hence, by seeking to answer this set of questions in order we looked to apply the data mining process across these data sets to find new information to help combat outbreaks of malaria and hopefully other infectious diseases.

Related Work

There are many studies which try to formulate a model to predict the influence of climate change to infectious diseases but it is rare to find a study using data mining techniques looking a wide breadth of data. Current studies are more narrowly focused with emphasis on biology of the vector and very specific data sets (Xiao 2016). However, there was one study found which does use a data mining like approach but still not in a traditional sense (Institute of Medicine 2008). Here they tracked climate attributes with respect time to predict mosquito populations.

1. Institute of Medicine. 2008. Global Climate Change and Extreme Weather Events: Understanding the Contributions to Infectious Disease Emergence: Workshop Summary. Washington, DC: The National Academies Press.P.200 https://doi.org/10.17226/12435.

2. LuLiang, Peng Gong ,Climate change and human infectious diseases: A synthesis of research findings from global and spatio-temporal perspectives. Citation DataEnvironment international, ISSN: 1873-6750, Vol: 103, june 2017 https://www.sciencedirect.com/science/article/pii/S0160412015300489

3. Longstreth, J. D., and J. Wiseman. 1989. The potential impact of climate change on patterns of infectious disease in the United States. In The potential effects of global climate change on the United States: Appendix G Health, ed. J. B. Smith and D. A. Tirpak for the Office of Policy, Planning, and Evaluation, U.S. Environmental Protection Agency. Washington, D.C.: U.S. Environmental Protection Agency http://www.ciesin.columbia.edu/docs/001-488/001-488.html

4. XiaoxuWu, YongmeiLu, SenZhou, LifanChen, BingXu, Impact of climate change on human infectious diseases: Empirical evidence and human adaptation Environment international, ISSN: 1873-6750, Vol: 86, Page: 14-23 , January 2016 https://www.sciencedirect.com/science/article/pii/S0160412015300489

5. https://www.who.int/globalchange/climate/en/chapter6.pdf

Data Set

Weather Data (independent variable):

The climate factors came from the National Oceanic and Atmospheric Administration's (NOAA) Global Historical Climatology Network (GHCN). It is a database comprised of integrated climate summaries from various stations across the globe. The database had five core attributes: precipitation (tenths of mm), snowfall (mm), snow depth (mm), maximum temperature (tenths of degrees C), and minimum temperature (tenths of degrees C). There are dozens of other attributes that become scarcer across the different objects in the database as different stations are equipped with different equipment and are in different climates. This database contains climate data on a daily, monthly, and yearly basis including some data from every year since 1763. This database is updated daily and we had to query it by utilizing Google's BigQuery API.

LINK

https://www.kaggle.com/noaa/ghcn-d

World Bank Data (independent variable):

The human factors came from the World Bank. Human factors also play an important role in determining disease outbreaks. The data base included 1600 attributes for each country and 50 years of data. Many entries were discarded due to empty values and the number of years had to be reduced to match malaria outbreak data availability. This process was automated by creating python functions using pandas. Some examples of the human factor attributes included, GDP, percent of population receiving immunizations, foreign aid received USD, net national income USD, percent of population with high school degree. The number of attributes were vast and included many items which probably have nothing or very little to do with disease outbreaks. Some examples included, percent of labor force in military, birth rate, child employment in agriculture, tax revenue percent of GDP. However, we felt it was also important to include attributes which had low probability as factors to see if they could be chosen by the prediction model.

LINK
https://databank.worldbank.org/data/source/world-development-indicators#advancedDownloadOptions

Pan American Health Organization (dependent variable):

The malaria disease outbreak data was sourced from the Pan American Health Organization. Here we obtained data for each year and each country. This was the dependent variable were the weather and human factors data were the independent variable. We chose to analyze countries located in Central America for specific reasons. The countries in Central America are small so there is less likelihood that the independent and dependent variables vary across the country. This would not be true for a large country like Brazil were there are many climatic zones and varying socioeconomic levels. In order to analyze a large diverse country we would need data broken down into those specific regions of the country. This was not possible given the data set available to us.

LINK

http://www.paho.org/data/index.php/en/indicators/visualization.html

Main Techniques Applied

Data cleansing and preprocessing was done mostly through the pandas library. However, before we could preprocess the data we first needed to load the data into the python environment. The weather data was almost 15 gigabytes in size, which led us to the decision of pulling the data from a SQL database rather than locally loading it. We generated SQL queries that were sent to a google BigQuery database which returned an extensive output. The output would then be saved in the python environment as a pandas dataframe. For the world bank and malaria data we decided that they easiest way to load the data was to download csv files and manually load them into a pandas dataframe. Once all the data was loaded into dataframes, we removed duplicate columns that had the same attribute expressed in different units. This was done to ensure that each column represented a unique attribute. For the weather data we first removed any objects with missing data. We then down sampled the data from days to years. Once this data was completely processed we exported it to a csv file. For the world bank and malaria data we cleansed the data through removing any missing objects. We then reindexed the dataframe and ensured that the dataframe was properly formatted.

The core tool we used in this project was the OLS library within python. The OLS library allowed us to implement a multiple linear regression by modelling the relationship between independent variables and a response variable. Through the model we were able to predict the outcome of the response variable. The regression model is based on the relationship between the dependent variables to the independent variable, so the more correlation we see between the variables the better the model will fit actual data. A problem with multiple linear regressions is if the independent variables are highly correlated with each other, then the model can be affected negatively. This would be an example of multicollinearity which occurs when the independent variables are correlated and therefore not independent which is an assumption made by the model. We can see how well the independent variables can predict the dependent variable without error through the R-squared value. The R-squared value or the coefficient of determination is a metric that measures how much variation in the outcome can be explained by the variation seen in the independent variables. This means that a model with a higher R-squared value will better predict the response variable and thus is better to model the data. To run the model through the OLS function we needed to reduce the number of attributes in the model to below 20. The way we did this with the world bank data was through backwards selection. The backwards selection function calculates the SSE (measure of unexplained variation) of the attributes and removes attributes with the largest SSE value. The function continues to iterate over the attributes until there are only 9 attributes left. We found that 9 attributes resulted in a model that best fit the data. These remaining attributes are likely to be the independent variables that have the highest correlation to the response and thus are the best variables to model the response. We ran the backwards selection program on the world bank data because we needed to refine a subgroup of attributes from the original of 500 attributes. This was not the case for the weather data, which means that we used all the attributes we collected pertaining to weather.

Once we created separate models using world bank data and weather data we ran an F-test to determine if there was significance in one model being better than the other. We decided to use a p-value of 0.1 and calculated the f-value for the test. For all the tests we calculated that the model using the world bank data was significantly better than the model with weather data. This could be a result that there were more attributes in the world bank data model, however we also saw generally higher correlation between the independent variables and the dependent in this model.

Key Results

Results - human factors:

From the initial 1600 World Bank attributes only 600 were included in the analysis per country after missing data was excluded. From the 600 factors our algorithm chose the 9 attributes with the best fit to the malaria outbreak data as determined by the OLS test. In general, the results of the OLS analysis show good fit with the human factors data. See examples in the visualization section below. By examining the nine highest attributes from each country, income related attributes showed up in all the countries in one form or another. For example, in El Salvador, Panama, and Honduras, net national income was selected and in Nicaragua GNI (gross national index) was selected. Both attributes are similar but have a variation on how income is determined. For our purposes we will lump them together as one category. The attribute with the next highest frequency was life expectancy which appeared in four of the seven countries, namely, Panama, Guatemala, El Salvador and Belize. This could reasonably be combined with the attribute mortality rate which appears in Honduras and Belize. Together, these two attributes cover five of the seven countries. It makes sense that life expectancy and mortality rate would be closely related to deadly disease outbreaks. The other attributes that appeared in more than one country were cereal yield (Honduras and Costa Rica), age dependency ratio (Honduras and Belize), and HIV infection rates (Honduras and Panama).

One can imagine that as one's income rises the quality of life and resources to combat diseases are improved. So, it may not be surprising that income measures may have a connection to disease outbreaks. Next, five out of the seven countries reported life expectancy or mortality rate as a significant attribute. This is likely an example of conflating correlation and causation as a higher mortality rate does not increase one's chances of getting malaria, but instead, the number of people that get malaria drives the mortality rate higher. Nevertheless, life expectancy and mortality rate can be used to predict the number of malaria outbreaks in a given year. Among the attributes that appeared in two countries, HIV infection rates and cereal yield likely have a direct causal role in malaria outbreaks. Cereal yield is a measure of a country's production of grain products and requires workers in agricultural environments which exposes them to mosquitos and other sources of malaria. HIV is known to cause complications with malaria, increasing the number and severity of malaria cases in HIV patients, especially those with advanced immunosuppression.

We were surprised that health and education attributes such percent of population immunized, number of physicians per 1000 people, and percent of population with high school diplomas were not identified. It could be that income related attributes overshadowed the health and education attributes since there were a multitude of income attributes but with slight variations. To overcome this another analysis could be run without the income attributes.

Results - weather factors:

From the 5 core attributes of the climate data set our algorithm chose the three attributes with the best fit to the malaria outbreak data as determined by the OLS test. In general, the results of the OLS analysis show a good fit with the climate factors data but not quite as good as the fit with the human factors data. Notably, four of our five models based on climate data were relatively good at fit but notably worse than the human factors models. It is interesting to note that the three highest attributes for each country were all the same: minimum precipitation, maximum precipitation, and average temperature. Given the context of Central American countries it is not hard for one to understand why these three attributes were the ones most closely linked to malaria outbreaks as the warm temperatures of these countries and overall amount of water (i.e. high humidity) are very complimentary to disease outbreaks.

Applications

The applications of the information mined here are both significant and far reaching but not clear in direct usage. Through the conclusions drawn here it is easy to see that climate and human factors have an effect on malaria outbreaks. With this research as a stepping stone, it is feasible to think that more particular information could be mined on the biggest impacts of climate and human factors on malaria outbreaks for different countries and regions. This information could, in turn, be used to develop plans, policies, and even infrastructure on local, regional, and national levels to combat malaria outbreaks and prevent people from becoming infected. With this potential application it is not hard to imagine how the same process and information could be used to develop a similar level of understanding for outbreaks of other diseases and act against them in similar ways based on the factors found to have the largest impact on different outbreaks.

Graphs and Other Visualizations

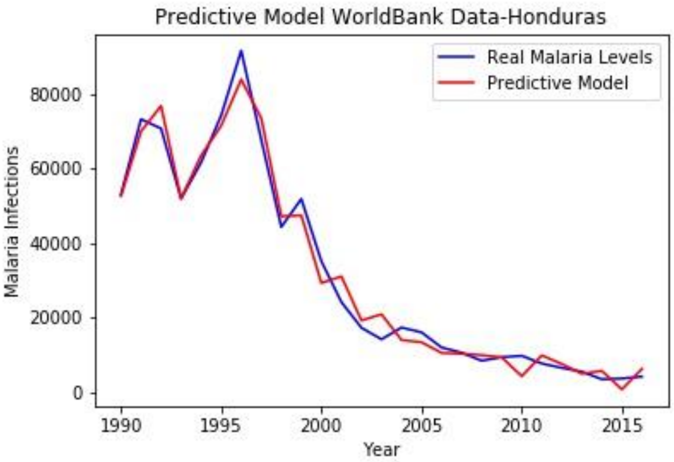


Fig. 1. Predictive model for Honduras using World Bank data with a R^2 value of 0.981

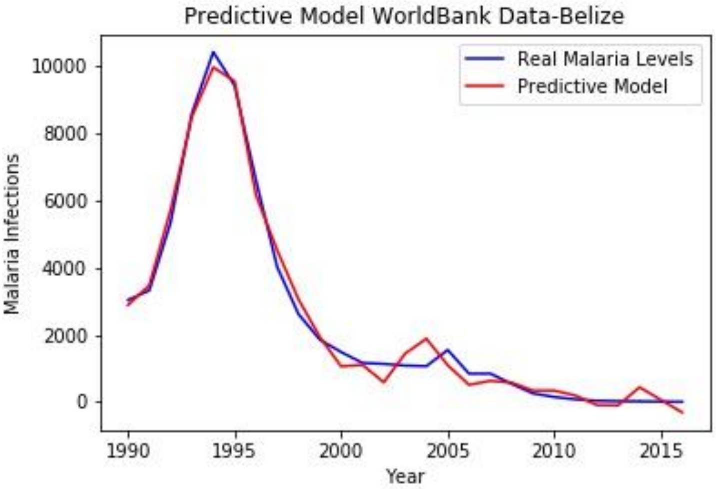


Fig. 3. Predictive model for Belize using World Bank data with a R^2 value of 0.987

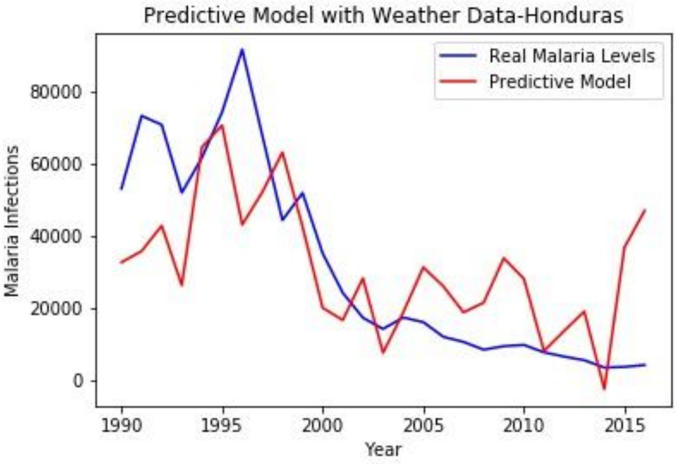


Fig. 2. Predictive model for Honduras using Weather data with a R^2 value of 0.418

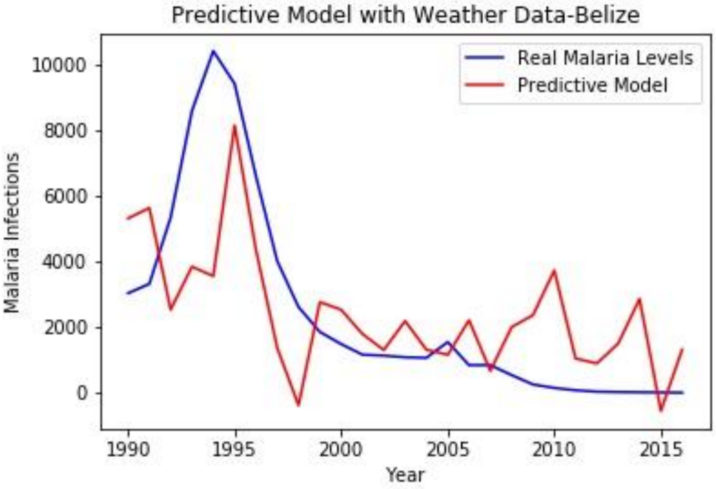


Fig. 4. Predictive model for Belize using Weather data with a R^2 value of 0.382

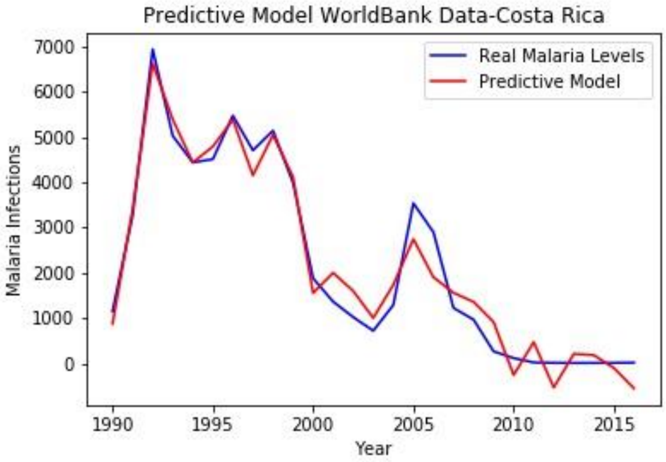


Fig. 5. Predictive model for Costa Rica using World Bank data with a R^2 value of 0.956

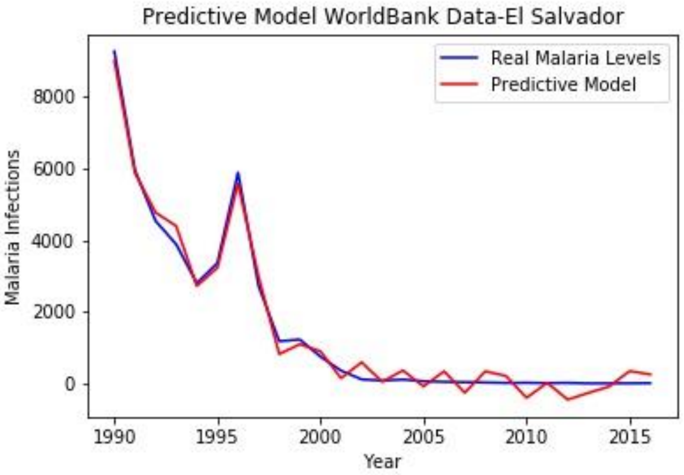


Fig. 7. Predictive model for El Salvador using World Bank data with a R^2 value of 0.987

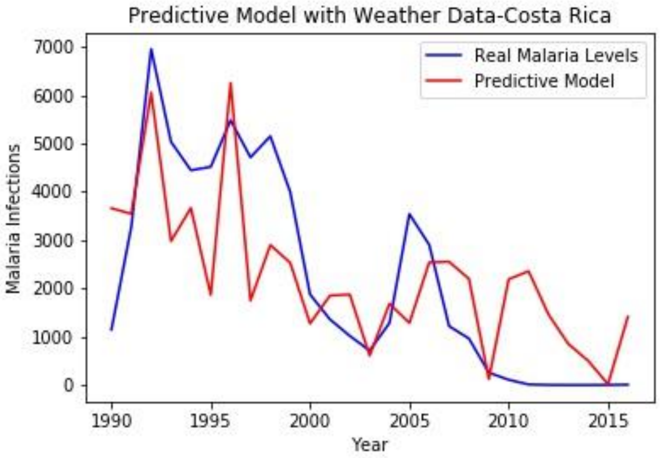


Fig. 6. Predictive model for Costa Rica using Weather data with a R^2 value of 0.492

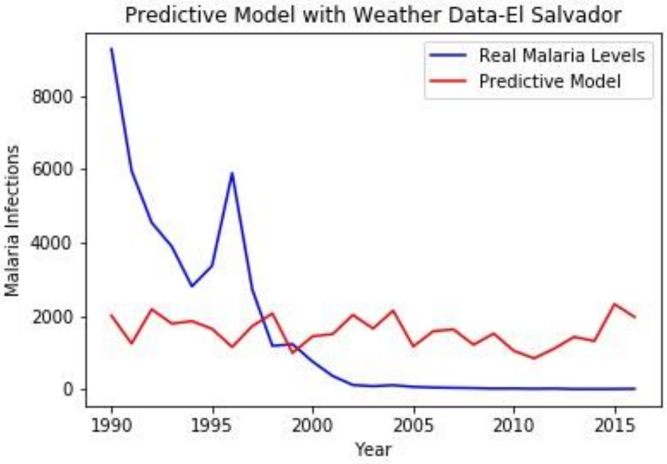


Fig. 8. Predictive model for El Salvador using Weather data with a R^2 value of 0.028

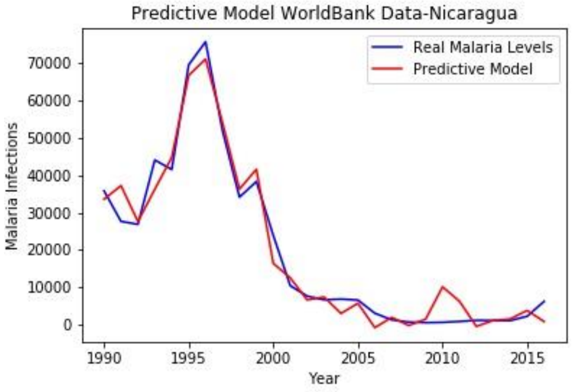


Fig. 9. Predictive model for Nicaragua using world Bank data with a R^2 value of 0.964

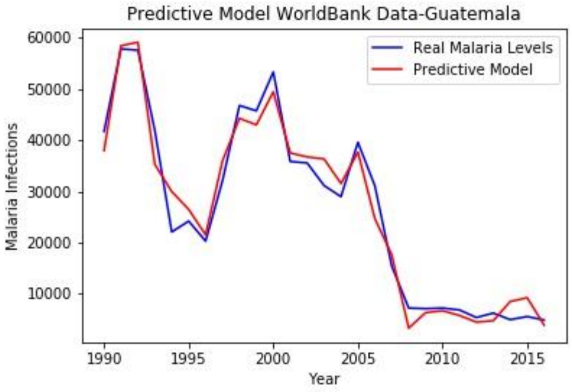


Fig. 10. Predictive model for Guatemala using World Bank data with a R^2 value of 0.963

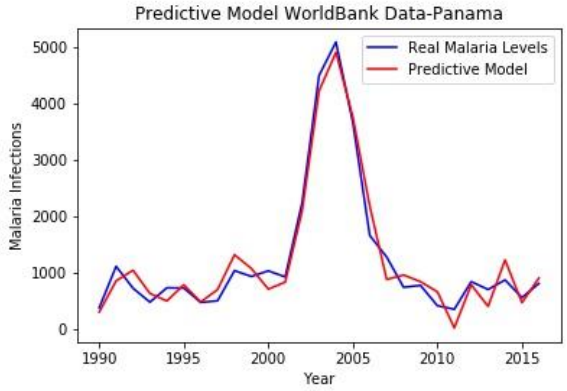
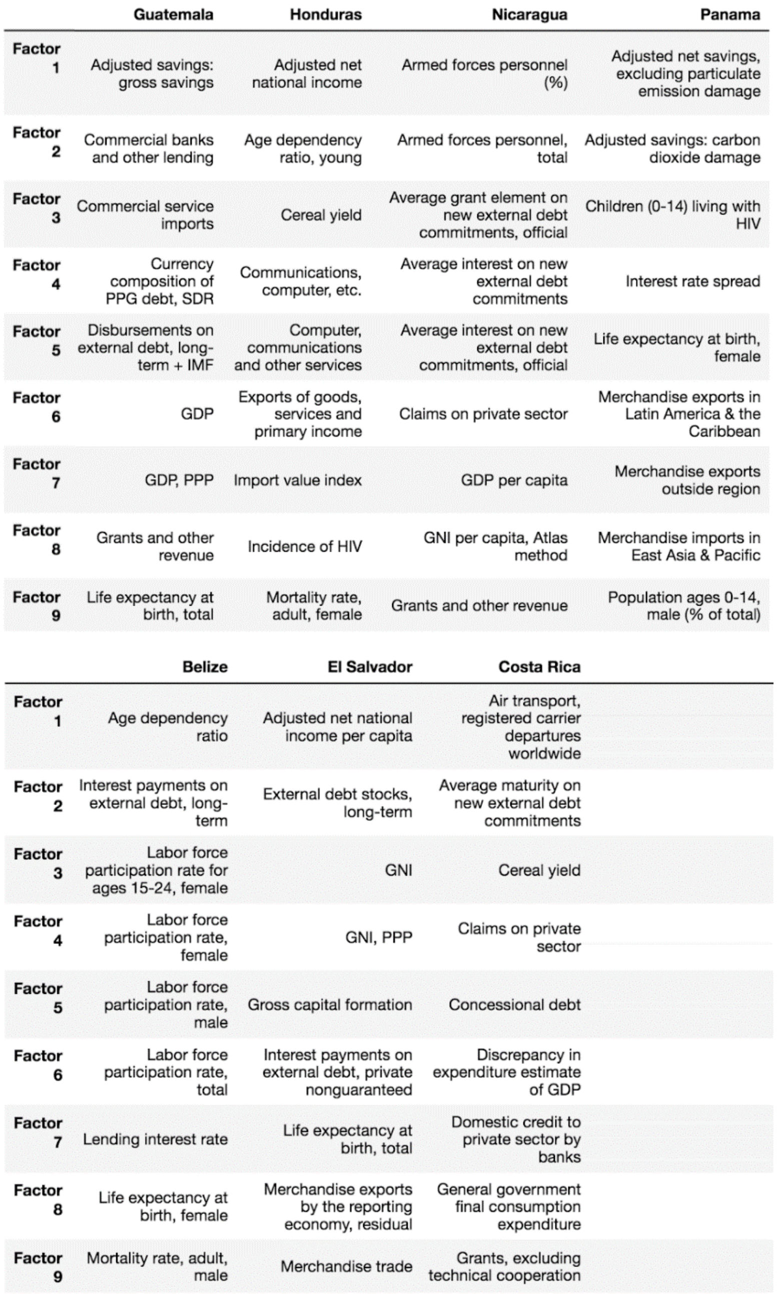


Fig. 11. Predictive model for Panama using World Bank data with a R^2 value of 0.961



	Guatemala	Honduras	Nicaragua	Panama
Factor 1	Adjusted savings: gross savings	Adjusted net national income	Armed forces personnel (%)	Adjusted net savings, excluding particulate emission damage
Factor 2	Commercial banks and other lending	Age dependency ratio, young	Armed forces personnel, total	Adjusted savings: carbon dioxide damage
Factor 3	Commercial service imports	Cereal yield	Average grant element on new external debt commitments, official	Children (0-14) living with HIV
Factor 4	Currency composition of PPG debt, SDR	Communications, computer, etc.	Average interest on new external debt commitments	Interest rate spread
Factor 5	Disbursements on external debt, long-term + IMF	Computer, communications and other services	Average interest on new external debt commitments, official	Life expectancy at birth, female
Factor 6	GDP	Exports of goods, services and primary income	Claims on private sector	Merchandise exports in Latin America & the Caribbean
Factor 7	GDP, PPP	Import value index	GDP per capita	Merchandise exports outside region
Factor 8	Grants and other revenue	Incidence of HIV	GNI per capita, Atlas method	Merchandise imports in East Asia & Pacific
Factor 9	Life expectancy at birth, total	Mortality rate, adult, female	Grants and other revenue	Population ages 0-14, male (% of total)

	Belize	El Salvador	Costa Rica
Factor 1	Age dependency ratio	Adjusted net national income per capita	Air transport, registered carrier departures worldwide
Factor 2	Interest payments on external debt, long-term	External debt stocks, long-term	Average maturity on new external debt commitments
Factor 3	Labor force participation rate for ages 15-24, female	GNI	Cereal yield
Factor 4	Labor force participation rate, female	GNI, PPP	Claims on private sector
Factor 5	Labor force participation rate, male	Gross capital formation	Concessional debt
Factor 6	Labor force participation rate, total	Interest payments on external debt, private nonguaranteed	Discrepancy in expenditure estimate of GDP
Factor 7	Lending interest rate	Life expectancy at birth, total	Domestic credit to private sector by banks
Factor 8	Life expectancy at birth, female	Merchandise exports by the reporting economy, residual	General government final consumption expenditure
Factor 9	Mortality rate, adult, male	Merchandise trade	Grants, excluding technical cooperation

Fig. 12. Table showing the nine factors for each country that were back-selected from World Bank data that best predicted Malaria outbreaks