EDS 222 Blog Post

Stephanie Copeland

Omitted-variable bias and the ability to predict Neglected Tropical Disease burden on a global scale by potentially correlative, changing environmental variables. Neglected tropical diseases (NTDs) are a group of ~20 conditions. They can infect the spectra of the human population with infection typically coming from the surrounding natural environment; they account for ~17% of infectious diseases in humans[1,2].

Many of these diseases require at the very least a host environment or a non-human reservoir host for permeation; several, in addition, are vector-borne diseases[2,3]. Their reliance on natural ecosystems makes them highly sensitive to anthropogenic change. Anthropogenic change can affect a landscape indirectly, e.g., via climate change or directly through things such as urbanization, resource exploitation, deforestation, etc[4]. Understanding infection prevalence levels furthers the analyses of how these landscape changes may be associated with the health of residing human populations.

Deforestation is an anthropogenic landscape phenomenon that has affected 10% of forested areas in the last decade since 2010, with a total of 411 Mha global tree cover loss[5]. Forests are essential to protecting and sustaining global biodiversity, with rainforests in particular accounting for 50% of the globe's biodiversity[6].

It has been argued that high levels of biodiversity would help protect against successful pathogen permeation into new areas, protect against the introduction of novel pathogens, and perhaps be associated with decreasing infections in human populations. This argument, however, is still out for debate as others have argued and supported the reverse to this[7,8]. Like the debate noted above, studies researching the link between deforestation and pathogen or pathogen-host prevalence have found mixed results[2].

Are levels of deforestation associated with the infection prevalence of neglected tropical diseases? I utilized one simple and two multiple linear regression models to answer these questions using data collected on a global scale, with the data separated into totals and then subsequently proportions of NTD burden in the form of DALYs (Disability Adjusted Life Years) and the loss in hectares of forest coverage, by country. I hypothesized that countries' with higher levels of total forest loss would also tend to have higher rates of NTDs burden. I predicted this trend would be challenging to determine globally due to disease-landscape level dynamics' natural complexities.

[1]

[1] Fun Fact: The Rod of Asclepius (2nd image in panel above) is the correct symbol for the medical field although it is often replaced with the symbol of the Caduceus. It has been typically thought to represent the staff of the Greek God of healing, Asclepius. But it has also been theorized to be the symbol for the traditional treatment of dracunculiasis, the Guinea worm disease, now considered a neglected tropical disease. The worm bursts through the skin of the host through a painful ulcerous blister and is then pulled out and wrapped around a stick over a period ranging from hours to weeks.

The Datasets *Deforestation* data was collected from the organization Global Forest Watch (www.globalforestwatch.org). This dataset encompasses forest loss related to human enterprise and natural causes such as fire, storms, or other natural disasters. It measures the area of tree cover loss at ~30x30 meter resolution. The data were then totaled to represent total tree cover loss (hectares) yearly by country since 2001. Total forest loss per country does not reflect what proportion of the country area lost forest area; therefore, a country area dataset from the World Bank was collected to attain such proportion. Data from the year 2017 was used in the analysis.

Disease data was collected using disease burden measured in DALYs (Disability Adjusted Life Years’) from the organization Our World in Data (www.ourworldindata.org). It contained total DALYs data per country on communicable, neonatal, maternal, and nutritional disease burden since 1990. The data sub-section to this set titled “neglected tropical diseases” was utilized. Total DALYs per country do not reflect the proportion of the country suffering from disease-related effects; population data from the World Bank was collected to attain such proportion. This dataset included Malaria in its tabulation of NTDs. Historically Malaria was considered an NTD; however, it is currently not listed by major world health organizations as an NTD. Data from the year 2017 was used in the analysis.

GDP per capita data for each country was obtained via a published dataset by the World Bank (https://data.worldbank.org/). GDP per capita data is one of the most common forms of financial data available at a country level. However, this data may omit specific political and economic issues of each country in this proportion. The country’s wealth is proportioned over all its citizens in this data, whereas in truth, a large portion of a country’s wealth may be retained by a select few. Data from 2017 was used in the analysis.

Part 1: *Is there a trend between forest loss and NTD burden?* Initially, a simple linear regression model utilizing Ordinary Least Squares (OLS) fit was applied to understand the relationship between the trend of deforestation and NTD burden. While a correlation calculation could have been utilized, I wanted the additional information provided by OLS, which would enable me to interpret the variance in disease burden per country related to their deforestation levels and the likelihood that deforestation levels have a significant trend with NTD burden. The model results showed an insignificant association between country level deforestation and NTDs burden (p = 0.156, α = 0.05) (R^2 = 0.0117).

$$NTD_i = \beta_0 + \beta_1 Deforestation_i + \mu_i$$

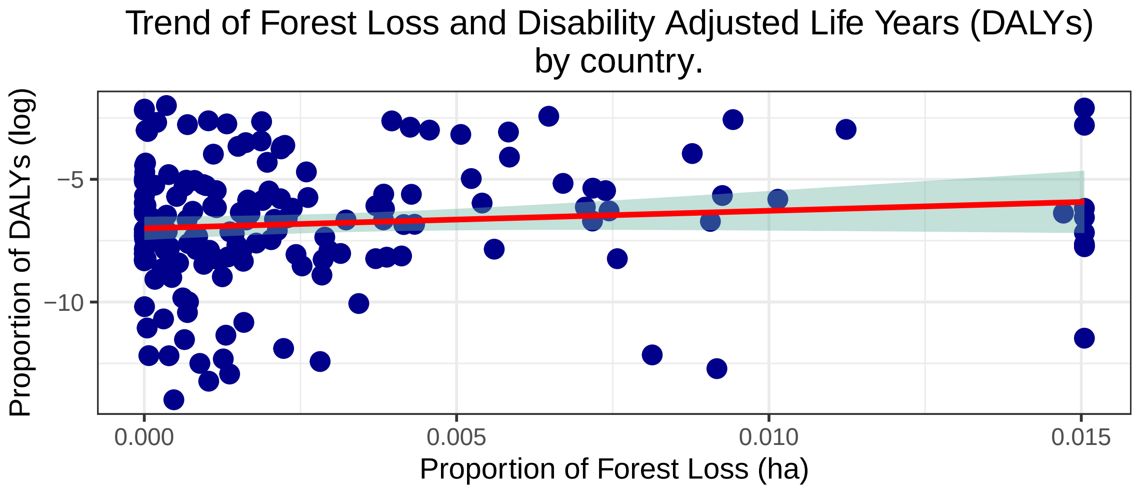

Figure 1: Scatter plot representing the trend of increasing proportions of deforestation per country (x−axis) and increasing proportions of the log of country DALY burden (y−axis). Each blue dot represents a country (n = 176). A predictive linear trend between the two variables (red line) represents the simple linear regression model with standard errors (light blue). There is an insignificant trend between increasing deforestation and NTD burden (p = 0.156), with the slope of the regression line close to zero.

The data set is heavily weighted to deforestation levels between 0% and .5%. The standard error reflects this clustering, increasing past .5% forest loss where there is less data. This outcome was relatively expected. Pathogen prevalence and disease burden are regularly linked to multiple environmental variables past just land-use change and often tend to be significantly related to socio-economic variables. Both sets of variables can differ considerably across countries. Thus, this simple linear regression model has an omitted-variables bias issue. Other country-level variables correlate with deforestation levels and influence country rates of NTD burden. With the omitted-variable bias (OVB) of this model, it is likely violating the exogeneity assumption of OLS. The model also violates the lowest variance assumption of homoskedasticity seen when plotting the residuals (Supplement, GitHub).

To aid in model fit and at least assuage some of the omitted-variables bias issues, I wanted to add a variable to the model that I knew showed a significant trend between that variable and NTD burden. Economic status and, in particular poverty levels, have been found to be an incredibly significant variable to NTD burden levels. Before adding this to the model, this relationship can be noticed by just adding GDP per capita as a color variation to the scatter plot trend of deforestation related to the NTD burden (Figure 2).

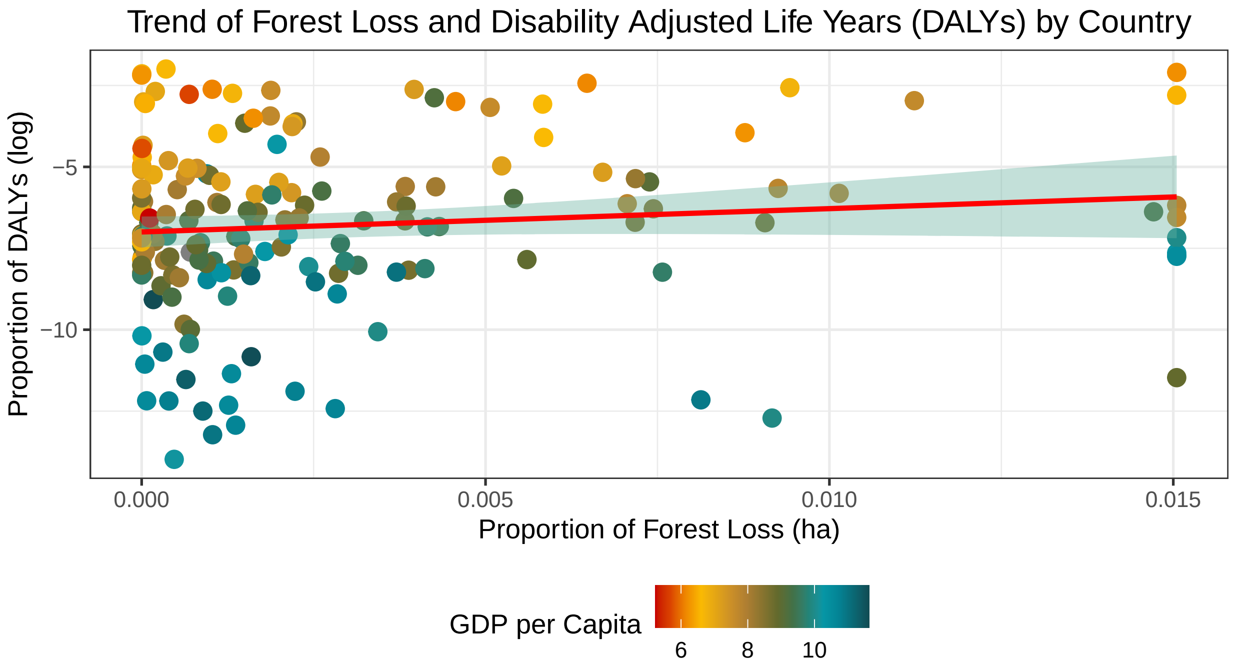

Figure 2: Is the same as Figure 1 (above), with the addition of a color gradient representing GDP per capita for each country (n = 176). Countries with higher GDPs' per capita appear to have a lower incidence of NTD burden (darker colored dots seen lower on the y−axis), whereas countries with lower GDPs' per capita appear to have a higher incidence of NTD burden (lighter colored dots seen higher on the y−axis)

Part 2: ***When the model includes an economic variable, is there a trend between deforestation levels and NTD burden?*** This time a multiple linear regression model with an OLS fit was applied to understand the strength of the relationship between trends in deforestation levels and trends in NTD burden when GDP per capita is held constant. Including an additional variable aims to decrease the issues with OVB and improve the understanding of what is an identifier to the variation seen in NTD burdens across the globe.

The model results showed a significant relationship between GDPs per capita and NTD burden ($p = <0.001$, $\alpha = 0.05$) (Figure 3). In this model, there continued to be an insignificant relationship between the level of deforestation and the level of NTD burden within a country ($p = 0.145$, $\alpha = 0.05$) (Figure 3). The model fit and share of NTD burden variation explained by the model increased to 39% compared to the simple linear regression model in part 1 (R^2adj = 03885).

$$NTD_i = \beta_0 + \beta_1 Deforestation_i + \beta_2 GDPcapita_i + \mu_i$$

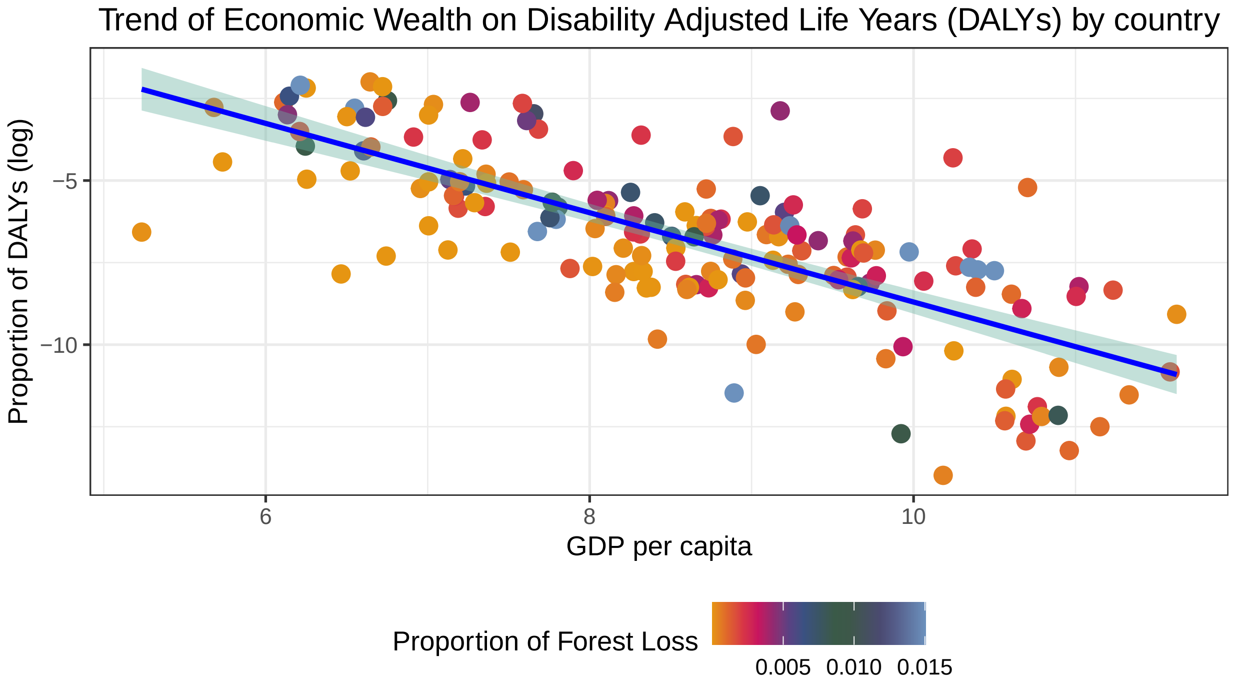

Figure 3: Scatter plot representing the increasing country GDP per capita (x−axis) and increasing proportion of the log country DALY burden (y−axis). Each dot represents a country (n=176). The color represents deforestation (low levels − light color, high levels − dark color). A predictive linear trend between the two variables (blue line) represents the multiple linear regression model with standard errors (light blue). There is a significant trend between countries' GDP per capita and NTD burden (*p* = <0.001). There is an insignificant trend between increasing deforestation and NTD burden (*p* = 0.145), with the slope of the regression line being a negative relationship between the GDP per capita and NTD burden. Increasing GDP tends to decrease the NTD burden of the country.

The addition of another variable does improve model fit and does show the significance of GDP per capita on the level of NTD burden. The relationship between forest loss as a predictor for NTD burden again was not significant. There is still likely to be an omitted variables bias and exogeneity issue which I cannot test for, but I can assume that issue still exists under the general knowledge about the complexities in determining NTD pathogen prevalence and NTD burden. The model again violates the lowest variance assumption of OLS when the residuals are plotted against GDP, showing clumped, heteroskedasticity (Supplement, GitHub).

To address the OLS assumption of lowest variance, I created a multiple linear regression model focusing on a subset of the data containing countries known to be climatically (tropical) and economically (low GDP per capita) at risk for high NTD burden.

Part 3: ***Is there a trend between deforestation levels and NTD burden when the model includes an economic variable and is selective to only countries that have low GPDs per capita and are in tropical climates?*** This multiple linear regression model with an OLS fit was applied to understand the strength of the relationship between trends in deforestation and trends in NTD burden when GDP per capita is held constant. This model only included a subset of the data from tropical, emerging nations. The focus on low-income nations may lessen the significance of GDP per capita, and these nations tend to have high levels of NTD burden. The predictiveness of deforestation on NTDs could be more apparent in this subset than on the global scale.

Even with solely low-income countries the model results showed a significant relationship between GDPs per capita and NTD burden (p = <0.001, α= 0.05) (Figure 3). There continued to be an insignificant relationship between the level of deforestation and the level of NTD burden within a country (p = 0.986, α= 0.05)(R^2 adj = 0.397)(Figure 4).

$$NTD, Trop, LI_i = \beta_0 + \beta_1 Deforestation_i + \beta_2 GDPcapita_i + \mu_i$$

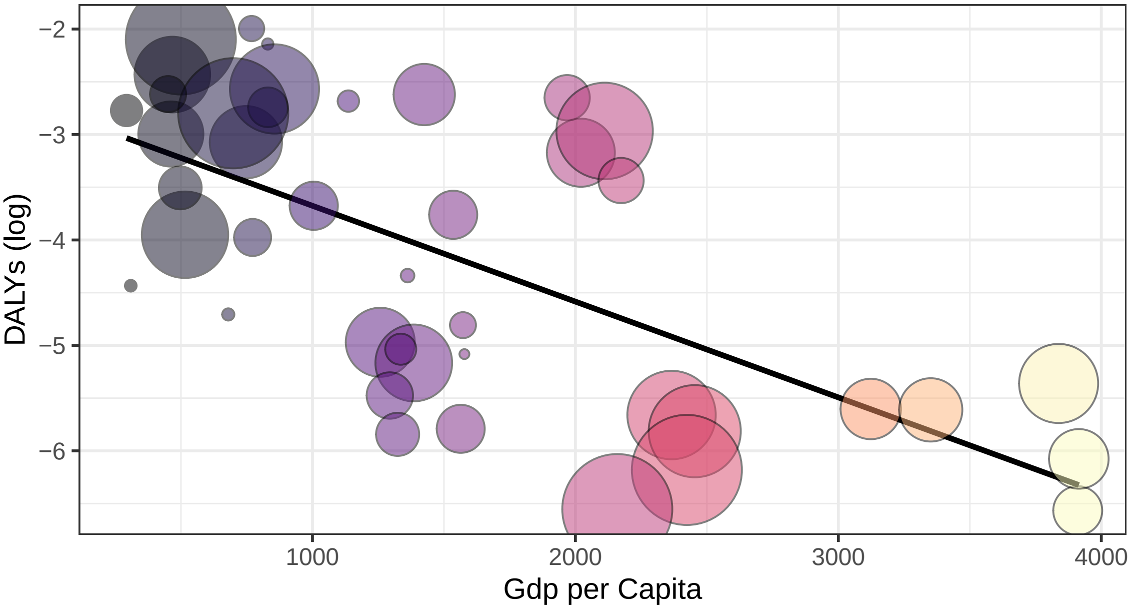

Figure 4: Scatter plot representing the increasing country GDP per capita (x−axis) and increasing proportions of the log of country DALY burden (y−axis). Each dot represents a country (n = 42). The color represents four GDP groups of GDP per capita. The dot size represents the country's level of forest loss (small −− low levels, large −− high levels). A predictive linear trend between the two variables (black line) represents the multiple linear regression model.

What is next? − ***Do the reverse with higher resolution data.*** I would start at the local scale, collecting NTD burdens from health clinics and hospitals and quantifying the local levels of "hard" forest edge. Hard-edge is anthropogenically altered land that directly abuts mature forest, compared to a natural edge that is a gradient of successional forest growth. Forest edge type and amount may be more critical to NTD burden than just deforestation. These hard edges that have drastically changed the native landscape may provide the right mixed conditions next to human populations to be a predictive mechanism along with socio-economic variables for NTD burden. If this data is magnified to a regional and country scale, I could then use predictive heat-mapping to get the temporal and spatial scale data this analysis lacks to quantify current NTD hotspots and areas at risk of becoming hotspots in the future.

As a budding (1st year) ecologist, this work was an important reminder that in complex systems such as disease ecology, many variables might be outside of my control (and well outside my expertise). An excellent reminder that seeking colleagues and collaborators outside of my field would be an incredible benefit to research and perhaps improve the understanding of these systems.

References and Data Sharing [1] WHO Health Topics (*webpage*) [2] Burkett-Cadena & Vittor (2018): *Journal of Applied and Basic Ecology* [3] Feasey et al. (2010): *British Medical Bulletin* [4] M. Booth (2018): *Advances in Parasitology* [5] Global Forest Watch (*webpage*) [6] C. Nunez (2019): *National Geographic Society* [7] Patz et al. (2004): *Environmental Health Perspectives* [8] Wood et al. (2017): *Phil. Tran. R. Soc. B* + Color palette courtesy of An Bui (calecopal package) + Full data and code (https://github.com/sjcopeland8/EDS_Final_Proj.git)