Yellow Fever risk mapping using Machine Learning

Yellow Fever (YF) is an endemic disease in Africa and the Americas. Approximately 30000 cases have been officially reported in the last 30 years, but this number relies on passive surveillance and is therefore significantly underestimated. Yellow Fever risk assessment is key to design vaccination and other intervention campaigns aimed at reducing the burden of the disease. In recent years, Artificial Intelligence (AI) and Machine Learning (ML) techniques have been proven extremely useful for disease forecast[1]. Here, we exploit the power of Machine Learning to predict Yellow Fever presence in the Americas and assess the importance of geographic and environmental factors, building on PAHO's seminal work and unique data set[2].

Results

The algorithm recalls 100% of the reported cases with 95% precision. That is, the 640 counties in the Americas that have had YF cases between 2000 and 2018 are all correctly recalled, and 31 YF-free counties are predicted to have cases. Most of these 31 counties border infected counties, as shown in Figure 2C, making them suitable candidates for the appearance of new cases.

This result was obtained using the ML algorithm AdaBoost on the subset of counties with environmental conditions suitable for YF[3], and considering the following features: precipitation, temperature, latitude, altitude, ecoregion, major habitat type, canopy loss (% of the county with tree canopy loss >30%), land use intensiveness and nonhuman primate presence.
The most relevant features for the prediction are latitude and canopy loss, followed by ecoregion classification, number of nonhuman primates, and temperature - as shown in Figure 1.

When longitude is included as variable, precision reaches 99.7% - only 2 YF-free counties are predicted to have cases. In this model, the most relevant feature is longitude, followed by the other features in the same order as above - as shown in the Appendix.

---

[1] https://link.springer.com/article/10.1186/s40504-017-0065-7
[2] http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0005897
[3] Tropical counties located between -78.41 and -35.9 latitude, where the average annual precipitation is below 3809mm, and with presence of nonhuman primates.

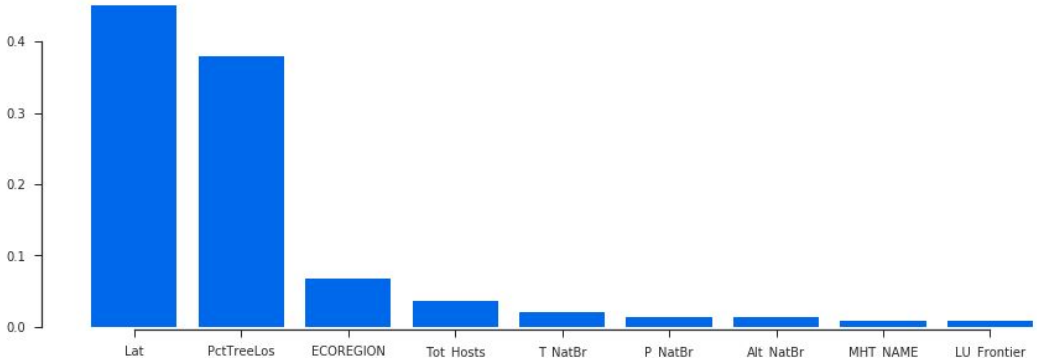

Figure 1: Relative feature importance assessed by the AdaBoost algorithm.

Next steps

Predictions are highly precise when the whole dataset is used, however results are not robust when the dataset is geographically stratified to train and test the algorithm on different subsets of counties. Hence, proposed future lines of work include:

- Further test the model robustness, possibly making use of temporal data to test if the algorithm could be used to predict new case locations, and to eventually monitor risk dynamically.
- Test the model using variables obtained from new sources of data available through Private Sector partnerships established by UNICEF Innovation. Examples include human mobility (e.g. air traffic) and school locations and their connectivity.

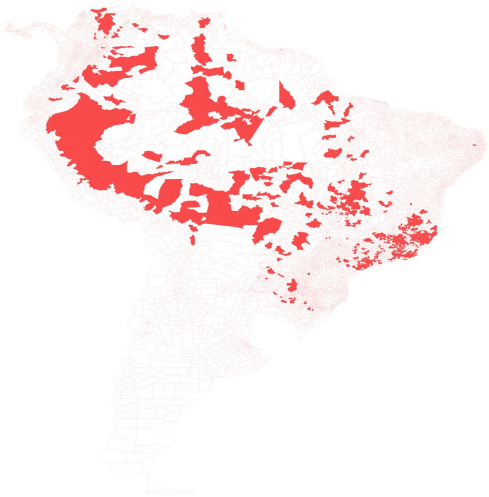

*Figure 2A: **Reported cases** of yellow fever in the Americas. Counties in red have recorded at least one case between 2000 and 2018.*

*Figure 2B: **Probability of yellow fever cases** as predicted by the algorithm. The more intense the red, the higher the probability.*

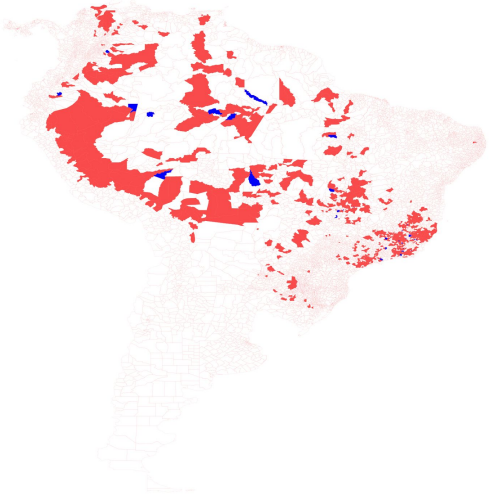

*Figure 2C: **Predicted cases**: counties where the algorithm correctly predicted yellow fever cases are coloured in red, and counties where the algorithm incorrectly predicted cases are shown in blue.*