Modeling Disease Outbreaks

Thomas Nash

April 19, 2016

Background

The goal of the project is to predict when and where disease outbreaks will occur. The initial approach was to create two separate models, one for each specific disease and another generic model for any disease. It was believed that there would be a significant difference between these two models, some diseases would be strongly correlated with some predictors like climate and others would be independent of these factors. After evaluating the availability of current models and gaining a better understanding of how this prediction is to be performed, the focus was switched towards visualization and modeling through maximum-entropy. The map provided by the Center for Foreign Relations (CFR) was informative to the location and impact of the outbreaks, but the visual was full of noise and hard to navigate. A new visualization which incorporated likelihood rankings was developed to supplement the current map.

Methods

Data

Climatic data is in the form of 19 variables of temperature, precipitation, and seasonality taken from WorldClim, a service which provides free climate data for ecologic and GIS use (Hijmans et. al, 2005). The data is easy to use as it comes in raster format which allows easy manipulation in R for scaling and combining data. The population density data comes from NASA's Socioeconomic Data and Applications Center (SEDAC), also in raster form but on a much finer scale (CIESIN, et. al 2005). The disease outbreak data which tracks known occurrences of vaccine-preventable outbreaks is from the CFR. This data was used for two reasons: the formatting and available latitude and longitude of the outbreak; and the fact that these diseases are preventable by vaccine, meaning if at-risk regions could be targeted the disease could be eradicated. Much filtering and pre-processing was done to correct errors in the data, be it formatting, typos, or missing data. For the sake of this project, attack data was removed and only disease outbreaks were used.

Maxent

In order to model the outbreaks, maximum-entropy modeling was adapted from its ecological application to species distribution (Phillips et. al 2004). This

method of modeling uses only presence data as we cannot say with certainty where outbreaks have not occurred. Samples are selected from the available data by filtering on years when outbreaks occurred, types of disease outbreak, and the impact scale in number affected and number of fatalities. Features of these samples are then provided as well in the form of climatic variables and population density. Using the latitude and longitude of where each outbreak occurred, the maxent attempts to find the maximum likelihood distribution that corresponds with the input data. This process is made possible using the *dismo* package in R which has maximum-entropy modeling built-in. A threshold for cross-validation (using 80/20 test, train split) was set so that outbreak parameters which had at least 40 occurrences could be used.

Results

Combining the maxent approach with the disease data in R, a Shiny application was designed which allows the user all combinations of input variables as well as visualization of the locations of the selected outbreaks before running the model. The only downside to the model is that due to the Java backend the authors used, it takes quite a long time to process. In addition to a global map of likelihoods, additional statistics on the model accuracy, variable significance, and thresholds are given.

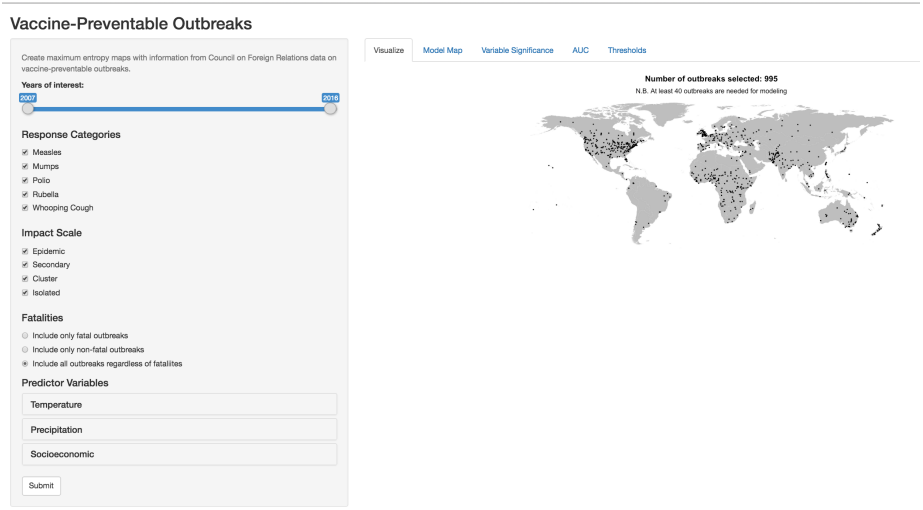

Figure 1: Current UI, changing input parameters dynamically updates the map to show the distribution to the user before model is run.

After the user has selected the desired predictors (for the example used in this paper, all predictors are used on all data points), the model is run and a map showing the likelihood of the chosen outbreak type occurring globally which is generated by maxent. Other tabs include information as to the significance of chosen predictors and the accuracy of the model.

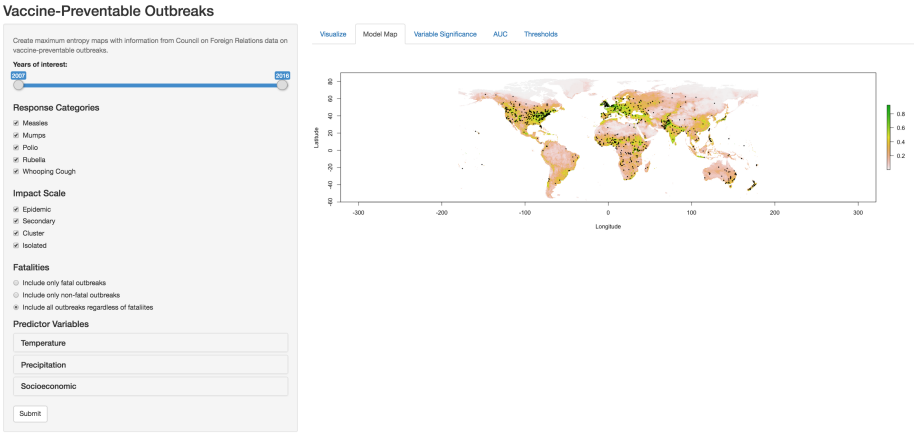

Figure 2: The maxent model now colors the globe with the likelihood of the same outbreak chosen in the inputs occurring, overlaying the instances where the outbreaks did occur.

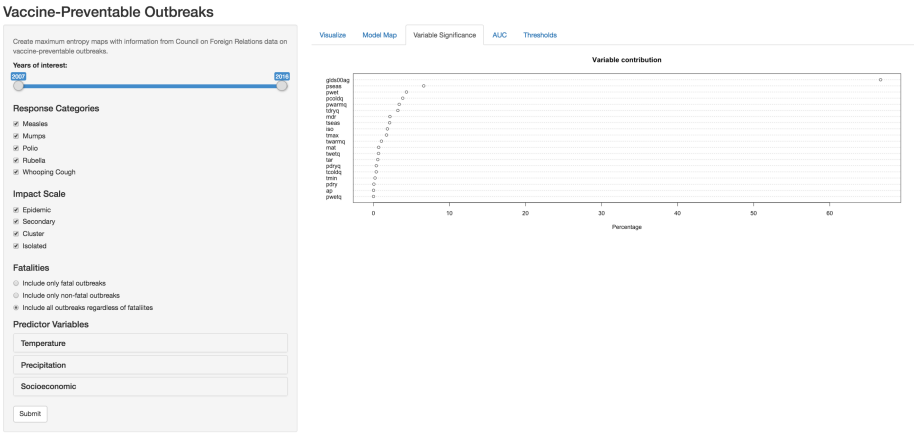

Figure 3: The importance of each variable in the model is given, population density far outweighs the climatic predictors.

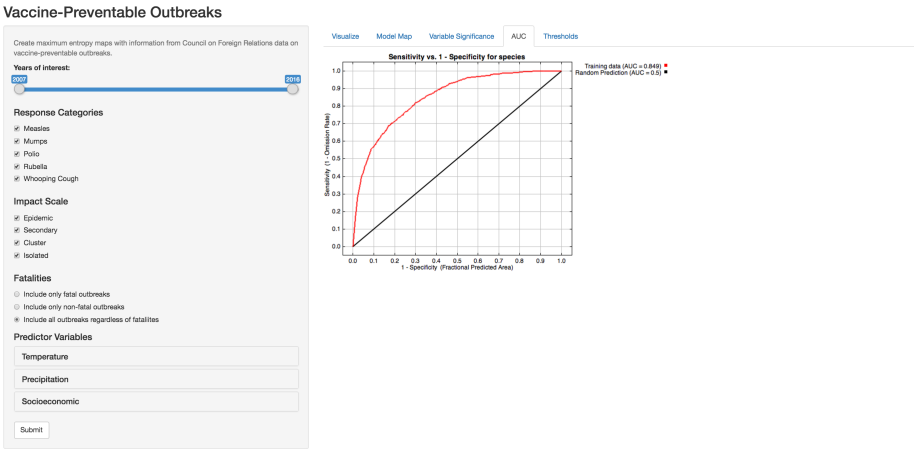

Figure 4: Area under curve (AUC) plot showing true positive vs false positive rate, in this instance a value of 0.849 is calculated.

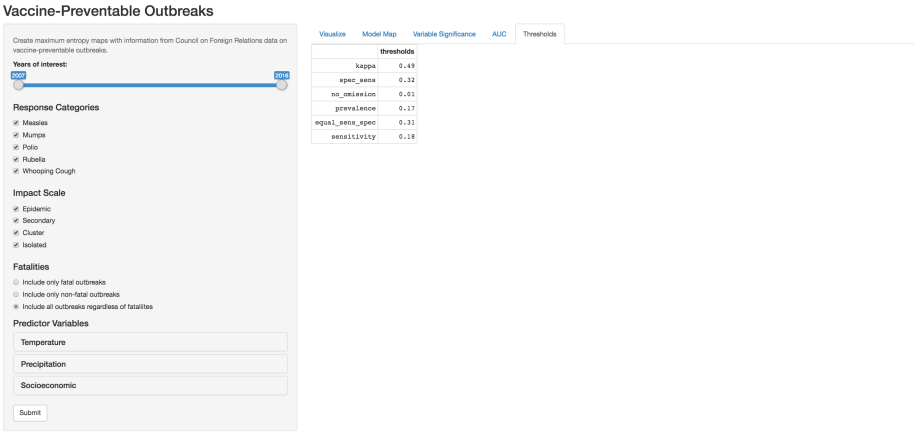

Figure 5: Thresholds for binary classification are given which are used to maximize a given set of measurements (kappa, no omission, and TPR+TNR)

Interpretation

The model in no way is able to predict when and where outbreaks will occur due to the nature of the data. What it does provide though is an accurate image of where outbreaks are likely to occur based on the selected features along with the importance of each feature. The limitation in expanding the scope of the model is the availability of data. It is simple to merge and drop layers of raster data as long as they are scaled to fit the same dimensions at the same resolution. Hopefully the model can be expanded to include predictors with regards to availability and access to vaccines, but no data could be found at a fine enough spatial resolution to justify it: data exists on a national level which is not feasible to use.

References

Steven J. Phillips, Miroslav Dudik, Robert E. Schapire (2004). A maximum entropy approach to species distribution modeling. *Proceedings of the Twenty-First International Conference on Machine Learning* (pp. 655-662).

Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones and A. Jarvis, 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* 25: 1965-1978.

Center for International Earth Science Information Network - CIESIN - Columbia University, United Nations Food and Agriculture Programme - FAO, and Centro Internacional de Agricultura Tropical - CIAT. 2005. Gridded Population of the World, Version 3 (GPWv3): Population Count Grid. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). http://dx.doi.org/10.7927/H4639MPP. Accessed 1 April 2016.