

Hotspots 2 Brief Proposal & Data Overview

Proposed Cost: \$515,571.14

Proposed Deliverables: Hotspots 2 and 3 data, models, and methods.

Proposed timeline: Hotspots 2 data delivered immediately, and enhanced Hotspots 3 data, models, and methods delivered at the end of Year 1 and Year 2.

Hotspots 2 Data & Model Overview

The Hotspots II use boosted regression trees (BRTs, a type of machine learning model) to predict the distribution of zoonotic disease emergence events based on a set of environmental predictors. Maps of two variables are provided: predict.observed; predict.actual

These maps display the model's per-grid-cell relative risk of observed emerging infectious disease (EID) events, and relative risk of EID events adjusted for bias caused by the uneven distribution of disease reporting effort across the population. There are two versions provided in the original email: low-resolution; high-resolution

The low-resolution version of the model has a few technically superior aspects that are not possible in the high-resolution version. This is because time slicing of predictor data, where EID events were matched with population size, cropland, and pasture data from relevant decades, was only possible in the low-resolution model. The data used in the low-resolution model was not temporal, and methods used in the high-resolution model require temporal data.

Furthermore, the in house generated reporting effort dataset was not available aggregated to high resolution at time of map generation, and so was resampled to the higher resolution with bilinear filtering. Maps are provided in .png image files, and Esri asc II (.asc) and GeoTIFF (.tif) formats. High-resolution .pngs are provided alongside Winsorized versions, which truncate extreme high and low values to better display the variation of mid-range values.

Hotspots 3 - Future Work

Hotspots2 represents a more advanced understanding of EID risk, but also highlight our currently limited understanding of disease emergence. Most glaring is the use of only one EID event per disease. This decision was made to capture the drivers of disease emergence (i.e., the areas where novel diseases are most likely to emerge into the human population). However, different diseases emerge due to multiple different factors, and emergence is a complex process that can take varying amounts of time.

To properly assess the risk of disease emergence and accurately make future predictions, it is necessary to build separate models for individual diseases, and groups of biologically related diseases, to gain a better understanding of the causal factors and processes governing their emergence. Actual disease emergence risk is the aggregate of the risk of multiple diseases emerging or potentially the interaction of simultaneous emergence events.

Data & Methods Overview

Explanatory variables

We compiled spatial data layers for predictors in four broad categories to decompose which factors increased ‘landscape suitability’ for disease emergence. These reflected the most frequently hypothesized drivers of zoonotic disease emergence and included (Table 1): human presence/activity, animals/hosts, the environment, and observation bias. Predictors came from a variety of data sources, and all were rescaled or transformed to a spatial grid of 1° resolution (WGS84, c. 110 km at the equator) prior to their use in models.

“Human Activity” data were compiled and six predictors derived based on the following rationale:

1) Population density likely influences EID risk in multiple ways. Firstly, EID events are defined as diseases emerging in the human population so that their frequency may be proportional to population density. To represent this, we treated human population conceptually as a multiplicative factor in our models. Secondly, population density may affect transmission dynamics such that introduction of a novel disease into a denser population may be more likely to produce outbreaks large enough to be detected as EID events. Human population density was represented in our models by a human population dataset, which provides gridded estimates of human population every five years for 1970–2000.

2) Population change acts as a proxy for changing demands on ecosystems leading to environmental perturbation that has been hypothesized to lead to disease emergence. We created a measure for population change by calculating the inter-decadal difference in values of human population density. We analyzed these data against the EID event database, matching them both decade-by-decade;

3) Land-use type represents largely anthropogenic influence on the landscape (as opposed to ‘land cover’ below) and has been hypothesized to play a role in disease emergence and spatial distribution. We used the percentage of land-use types in each grid cell of a global dataset every ten years to derive predictors representing percentage of land used for cropland and percentage used for pasture. We used dates present in both human population and land-use datasets;

4) Land-use change has been hypothesized as a key mechanism for disease emergence in that it brings humans into close proximity to wildlife. To test for an effect of land-use change on EID risk, we created metrics of change for pasture and cropland by calculating the between-decade difference in values for each grid cell for cropland and pasture.

Andrew Huff, Ph.D., M.S.
huff@ecohealthalliance.org

“Animal/host” data were represented by two predictors: **1) Mammalian species richness**. The diversity and prevalence in a host population of potentially zoonotic pathogens in an area is likely to be a key factor in the risk of novel pathogen emergence. However, spatial data on global pathogen diversity do not currently exist, and it is estimated that we have identified less than 1% of mammalian viral diversity. Consistent with previous studies, we therefore assume that the number of available pathogens in an area is proportional to the diversity of wildlife species. Since the overwhelming majority of emerging zoonoses have mammalian hosts, we used mammal species richness as a proxy for pathogen species richness. To do this, we used the most up to date mammal species distribution maps available, derived from species-specific distribution models accounting for habitat suitability. In each analysis, we scaled mammal richness to the study grid; **2) Domestic animal density**. A number of past EID events with wildlife origin have emerged through farmed or domestic animal intermediate or amplifier hosts (e.g. Hendra and Nipah virus, SARS). Additionally, there is growing evidence that the global trend of intensification of livestock production has an inherent risk of the emergence of novel wildlife-origin zoonoses, e.g. Nipah virus in Malaysia, influenza viruses and others. We used the a dataset that contains data for poultry, goat, buffalo, cattle, sheep and pig headcounts. We summed mammals to a single predictor (domestic mammal headcount) and retained poultry as a discrete predictor.

We analyzed thirteen predictors from two datasets representing “Environmental” variables: **1) Climate**. Climatic factors have been repeatedly hypothesized as important in the global biogeography of human infectious diseases, including EIDs. Climate may influence disease distribution through enhanced suitability for vectors of wildlife origin zoonoses (e.g., West Nile virus), more rapid vector reproduction rates and biting rates, changes in the efficiency or rates of pathogen transmission among hosts and vectors, and changes in the ability of pathogens to persist in the environment, among other factors. Climate was represented by a single layer in our study, the Global Environmental Stratification, which uses a quantitative model to stratify the Earth's surface into zones of similar climate on a single scalar measure; **2) Land cover type**: Land cover type is associated with the distribution of terrestrial mammals and other taxa, potentially exposing humans present to different assemblages of viral species. It is also likely that different land use types favor different types of contact between wildlife and people. For land cover, we used twelve classes (different types of vegetation, urban land, and barren areas).

Observation bias

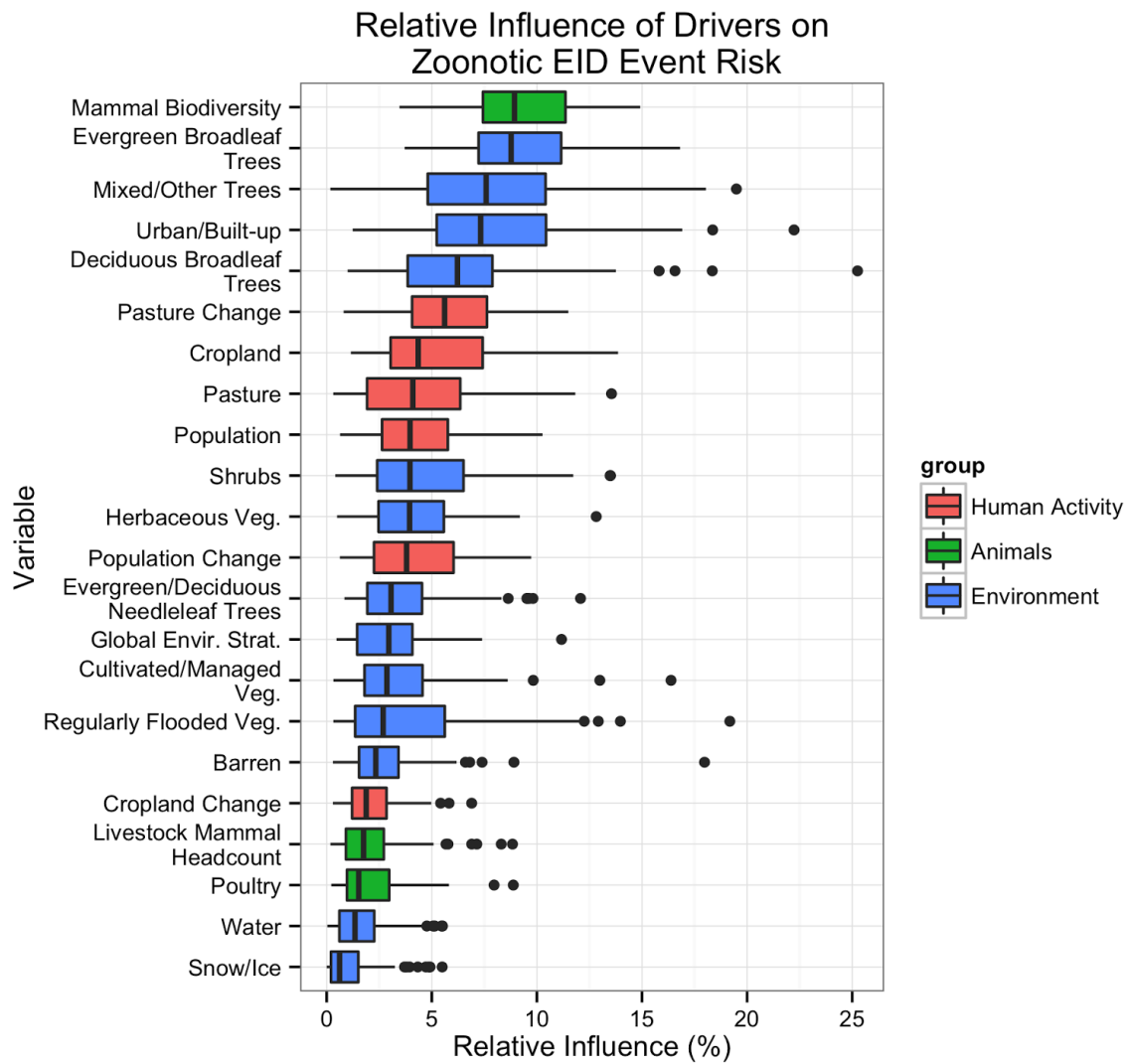
The distribution of reported EID events is likely strongly influenced by unequal detection and reporting disease outbreaks spatially and temporally. Previous studies have used proxies of observation effort such as the interpolated locations of known sampling sites (‘sampling effort’); frequency of countries of residence for all authors of all articles in the Journal of Infectious Disease (‘reporting effort’); and PubMed searches for “‘infectious disease’ + ‘country name’” for each country (‘reporting bias’). Other studies have used occurrence records for a similar class of

Andrew Huff, Ph.D., M.S.
huff@ecohealthalliance.org

observations as a surrogate for background sampling effort, e.g. in ecology, modeling the distribution of a particular species and utilizing occurrence records from multiple other species as a background sample. We adapted these approaches by deriving an index for observation effort (“Force of Observation”) based on the spatial and temporal distribution of publications in peer-reviewed biomedical literature. We searched articles in the PubMed Central Open-Access Subset for place names to generate a spatial layer representing relative reporting or observation effort globally. For each of PMCOAS’s approximately 760,000 articles, we used an automated machine-learning tool which we designed (“Pubcrawler”) to extract body text and search it for country names and names of cities with populations over 1000 from the GeoNames database which includes data on population, country, and geographical coordinates for each city. We weighted each citation of a place for which the country was also cited as 1. Where multiple places matched the same name (e.g. Birmingham, Alabama and Birmingham, UK), for each article, each place name’s weight of 1 was divided among its matches proportional to their relative populations. Weighted place name citations were then summed to the study grid. Rather than using the raw publication effort data in model fitting, we smoothed this layer by substituting it for the results of a model used to predict publication effort. The model used the predictors, and explained xx% of variation in publication effort globally. Using the model output rather than the raw data eliminated the large number of grid cells with zero-value reporting effort in the raw count and made our analysis more robust to outlier grid cells resulting from false positives (areas where place names were incorrectly identified). This approach, of using boosted regression trees (BRT) to fit the observational bias allowed us to normalize imperfections in our underlying data of deriving place names from the published literature and in the GeoNames database.



Variable(s)	Type	Spatial Resolution	Temporal Resolution
Human population	Human activity	0°5'	5 years
Cropland	Human activity	0°5'	10 years
Pasture	Human activity	0°5'	10 years
Mammal species richness	Animals/hosts	300m (Mollweide projection)	N/A
Domestic livestock headcount	Animals/hosts	0.05°	N/A
Poultry headcount	Animals/hosts	0.05°	N/A
Global environmental stratification	Environment	0°0'30"	N/A
Land cover	Environment	0°0'30"	N/A
Reporting effort	Reporting effort	N/A	N/A



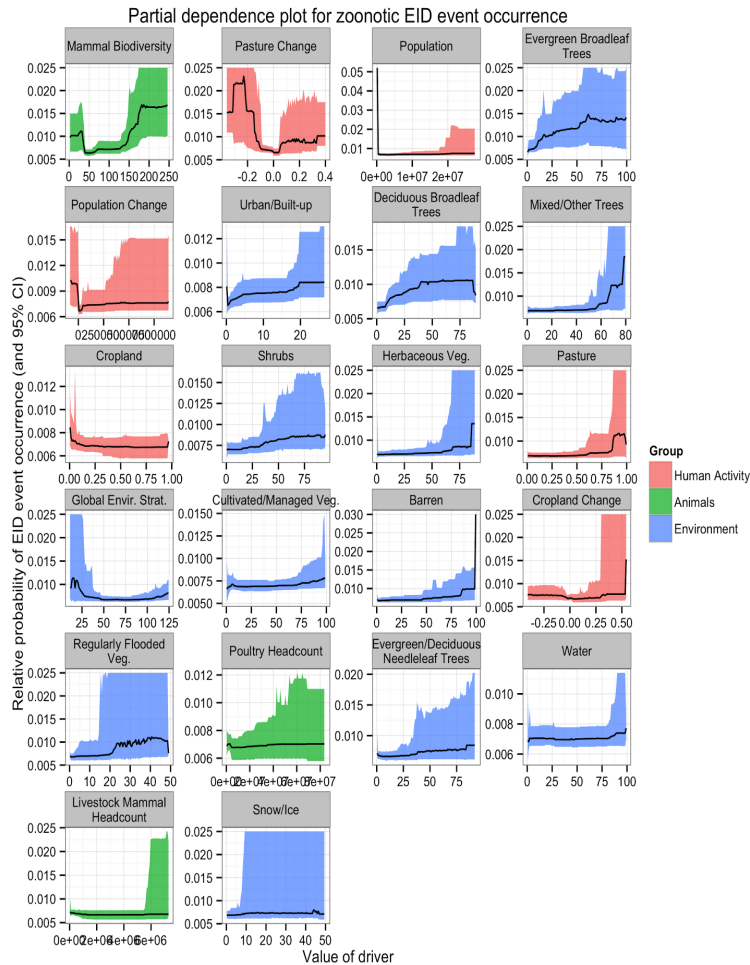


Fig. 2: Partial dependence plots for the zoonotic EID events for all variables in the boosted regression tree (BRT) model. X axes show the full range of observations (e.g. biodiversity or degree of pasture change per grid square). Y axes show the relative risk of an EID event in grid squares over the full range of observed drivers. Thin lines show individual model runs and dark lines show a smoother applied to all model runs. Partial dependence plots are a common method of visualizing BRTs, and display the response for an individual variable in the model while holding all other variables constant (De'ath 2007, Elith et al. 2008).