Texas Arbovirus Risk

Lauren Castro[a]
Xi Chen[b]
Nedialko B. Dimitrov[b]
Lauren Ancel Meyers[a]

[a]Section of Integrative Biology
The University of Texas at Austin

[b]Graduate Program in Operations Research
The University of Texas at Austin

Prepared by the University of Texas at Austin
in response to DSHS Contract 2015-047259-001

THE UNIVERSITY OF
TEXAS
AT AUSTIN

June 30, 2015

Executive Summary

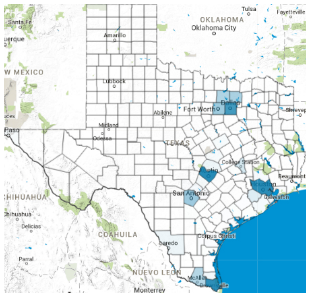
Arboviruses are a major public health concern in Texas. Two viruses that have not yet established local transmission but may pose a threat are chikungunya virus and dengue virus. Chikungunya is a disease that has for decades been endemic in Asia and Africa, but recently has caused large outbreaks in Central America and the Caribbean. Dengue is a concern in tropical and subtropical regions of the world, infecting millions of people every year. Both of these viruses are consistently imported into Texas. Increases in travel and virus outbreaks around the world have lead to an increase in the number of imported cases over the past few years. Although there are significant biological and epidemiological differences between chikungunya and dengue, both viruses share the same primary mosquito vectors, *Aedes aegypti* and *Aedes albopictus*.

In this project, we seek to answer two important problems for chikungunya and dengue surveillance, control, and prevention in Texas:

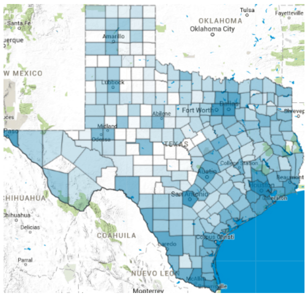
- Where do the mosquito vectors live in Texas?
- Where are the geographic risk zones in Texas?

To assist the Department of State Health Services in its mission, and to answer these questions, this project builds vector habitat suitability maps, import risk maps, and sustained transmission risk maps for Texas. In addition, the project produced a website, arbovirusrisk.org, that can be used to integrate the risk results in the Texas education and surveillance effort for arbovirus. The website offers visualizations of chikungunya and risk, *Ae. spp.* suitability maps, and the ability for DSHS officers to upload and visualize timely state-wide data. The website automates and improves arbovirus surveillance reporting efforts, generating PDF reports and online visualizations specific to the state as a whole, each Health Service Region, as well as each Texas county.

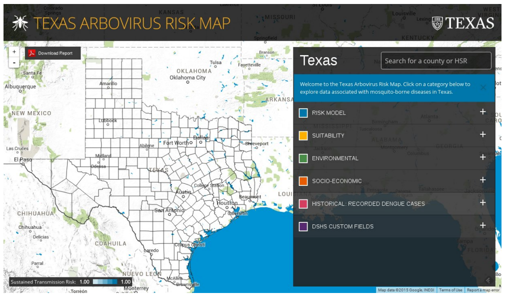
The figures below show just a few example outputs from the project. The first image is a validated predictive model for importing new disease cases, mapped over Texas counties. The second image is a relative risk map for sustained transmission through an *Aedes aegypti* vector. The third image shows the arbovirusrisk.org website. The box on the right of the website user interface allows users to navigate thorough risk maps, high resolution suitability maps, high resolution environmental maps, socio-economic maps, historical import data, and custom DSHS data uploaded by DSHS officers. A user can search for a Health Service Region or a county in the box above, or click on a county, to focus on a new geographic region. A PDF report button on the top left allows users to download an arbovirus surveillance report for their region of interest, along with visualizations of the most recent case reports.



(a) *Import risk*



(b) *Ae. aegypti* transmit risk



Welcome to the Texas Arbovirus Risk Map. Click on a category below to explore data associated with mosquito-borne diseases in Texas.

(c) arbovirusrisk.org

The species habitat models and disease risk models are created using the best available historical disease and vector occurrence data, gathered from multiple sources both within Texas and outside the state. Each model is constructed using state-of-the-art modeling and validation techniques. The model results provide insight into the key geographic areas and counties at risk for arbovirus import, and sustained transmission after import. DSHS officers can use these insights to communicate with primary healthcare providers and patients, to ensure timely response to imports and prevent these diseases from establishing themselves in Texas.

Vector Species Distribution. Based on the environmental characteristics of vector species occurrence for both *Ae. aegypti* and *Ae. albopictus*, we produce maps of the suitable locations for these species across Texas. The output is an estimate of the relative occurrence of the vector species, allowing comparisons of the occurrence of mosquitoes in different counties. Highly suitable areas for these vector species include highly populated areas, specially Houston, Austin, San Antonio, and in the case of *Ae. aegypti*, El Paso. A second tier of risk level is in the counties along the gulf coast. High-resolution maps of vector habitat, accessible from the website, may be useful for future surveillance planning of mosquito trap locations.

Risk Distribution. We construct two key risk maps: an import risk map and a sustained transmission map. Together these maps highlight areas in Texas that are at elevated levels of risk for seeing an imported case and areas that, once a case is imported, could experience secondary local transmission. Most of the risk maps available prior to this project were incidence maps that document the number of cases in each region.

Import Risk Map. We base our import risk map for both chikungunya and dengue on historical dengue import data starting from 2002. Chikungunya has only been imported into Texas in large numbers recently. Model includes 76 potential risk factors in categories of travel, environment, socio-economic and demographic data, as well as vector species distributions. Using a mathematical procedure to validate and compare models, we select the best risk model from the 2^{76} possible models. The predictive ability of the resulting import risk model significantly outperforms an empirical incidence map and other intuitive risk models.

Sustained Transmission Risk Map. There is a risk that chikungunya and dengue may become endemic in some areas of Texas. That risk depends on the pretense of a sufficient mosquito and human population to maintain a transmission cycle after an initial import. Using the output of the vector species distributions, we calculate a sustained transmission risk metric proportional to the reproductive number of the disease in each county. In other words, the metric is proportional to the expected number of secondary cases a single imported generates, for each county. The sustained transmission risk map indicates that counties of high risk are those with a high mosquito population to human population ratio. These are not necessarily the same counties as those with high human populations.

Tools to Facilitate DSHS Mission. The arbovirusrisk.org website provides an effective and direct visualization tool for arbovirus surveillance throughout the state. Users can navigate several categories of data. Each category presents several variables, and each variable can be displayed on a state level, an HSR level, or a county level. Some data, such as vector species distributions and environmental variables can be displayed in high resolution. Other data, such as incidence counts, can be displayed in a time-specific manner using time-selections sliders. More importantly, the website provides tools for DSHS to actively update and edit the data displays, resulting in a valuable tool for future arbovirus surveillance, communication, and control efforts.

Table of Contents

1 Introduction

Arboviruses are a major public health concern in Texas, with emerging viruses such as chikungunya virus (CHIKV), historically imported viruses such as dengue virus (DENV), and viruses endemic to the United States such as west Nile virus (WNV), St. Louis encephalitis (SLEV), and eastern equine encephalitis virus (EEEV). Collectively, these viruses cause hundreds of cases in both humans and domestic animals through the course of the year, leading to a significant economic and health burden. Of special concern are viruses that have the potential to become endemic in Texas.

CHIKV is a re-emerging mosquito-borne infectious disease that has caused large outbreaks around the world, including in non-endemic places such as Italy in 2007 (Thiboutot et al., 2010) and most recently Puerto Rico (Fischer et al., 2014). The United States has had imported cases in almost every state, and there have been an increasing number of locally-acquired cases in Florida and Texas. Within Texas, CHIKV cases have been reported since 2007 when two cases were first imported, originating from India. There was a drastic increase to 114 imported cases in 2014, tracing to travelers coming from countries in the Americas (Florin and Robinson, 2015). The changing origin of imported CHIKV cases in Texas is an indication of the global spread of this disease.

In addition to CHIKV, DENV is another important arbovirus of great concern in tropical and subtropical regions, infecting millions of people every year. Dengue has been present in Texas since 2002, with 33 cases reported in 2014. Although there are important biological and epidemiological differences between CHIKV and DENV, both viruses share the same primary mosquito vectors, *Aedes aegypti* and *Aedes albopictus*. As a result of increased human travel, both of these vectors have spread to new geographic regions and have been documented in Texas for several decades. *Ae. aegypti* has historically been in the gulf region (Micks and Moon, 1980) and *Ae. albopictus* was first documented in Harris County in 1987 (Sprenger and Wuithiranyagool, 1986). Until recently, *Ae. aegypti* was the traditionally associated primary vector for CHIKV, but a mutation documented by Dubrulle et al. (2009) has led to a new strain that is efficiently transmitted by *Ae. albopictus*. The presence of both of these mosquito vectors and the repeated introduction of imported cases call attention to CHIKV and DENV as growing threats in Texas and highlights the need for local and state level preparedness through increased understanding of the ecology and risk factors of these viruses.

1.1 CHIKV and DENV Surveillance

In the face of a growing threat of CHIKV and DENV, optimizing mosquito and arbovirus surveillance within individual counties and throughout the state of Texas would increase the chances of early detection of circulating virus. Targeted surveillance allows for a cost effective allocation of resources before and during an outbreak. Effective surveillance also enables early detection and allows public health officials to control secondary infections. One must define the question that the surveillance activity addresses in order to make it effective. Two potential questions that surveillance may address include:

(a) Where do the vectors live in Texas?

(b) Where are the risk zones in Texas for imported or sustained transmission of CHIK or DENV?

There are ongoing efforts to answer these two questions within Texas. Regarding the first, arbovirus surveillance is routinely conducted throughout Texas in an effort to detect the presence of pathogens in mosquitoes before the development of human cases. Specific surveillance activities include:

- monitoring mosquito populations
- identification of mosquito species
- laboratory testing of mosquito samples for circulating viruses
- recording human infections
- educating and training the community on matters related to arbovirus control

Out of these activities, trapping and mosquito population monitoring is largely coordinated at the county level, while laboratory testing may occur within individual county mosquito control divisions or samples may be sent to the Department of State Health Services (DSHS). Because of the county-based mosquito surveillance programs, data on presence and abundance of the different *Ae.* is irregular and scattered throughout disparate resources in the state. The lack of a centralized database of data poses a significant challenge towards gaining an acute situational awareness of risk throughout the state. Nevertheless, understanding which geographic areas are at risk for CHIKV and DENV is a key goal for the surveillance system.

Risk maps can be developed to address the second question of identifying risk zones for imported and sustained transmission of arbovirus. Risk maps show different areas of high and low risk and can be important tools for making decisions about how to best allocate resources and conduct surveillance. Current risk maps used by authorities for arbovirus disease surveillance are often incidence maps that document the number of cases that have occurred in a given area within a specific time period. Although this type of risk map can indicate the risk level to some extent, it may be possible to construct more effective and accurate risk maps that predict potential for new cases in Texas.

In this project we aim to improve upon existing arbovirus surveillance practices by producing three measures of risk for arboviruses transmitted by *Ae. aegypti* and *Ae. albopictus*, specifically for CHIKV and DENV. First, we model the vector distribution of each mosquito species throughout Texas, producing a 1-km grid of relative abundances of mosquitoes. While this output is used as input for risk models, it can be considered a measure of ecological risk on its own. Second, we develop a data-driven risk map of import risk, highlighting counties where we are likely to see the next imported case based on historical data. Third, we use the output of the vector species distribution to help quantify sustained transmission risk, identifying counties where once an importation has occurred, one is likely to see a sustained chain of secondary cases.

As a result of the project, the risk maps and vector distribution maps are integrated into a larger web app tool, the Texas Arbovirus Risk Map, that can be used to easily visualize historical data, compare relative levels of importation and sustained transmission risk among counties, and generate important educational and arbovirus activity reports. The web app also allows DSHS to visualize novel data fields that epidemiologists believe may be relevant to arbovirus risk, further contributing to surveillance practices.

2 Vector Species Distribution

One type of risk modeling previously used for arbovirus surveillance is species distribution modeling (SDM). SDM's goal is to identify geographic areas where vector species live. This is a key aspect of arbovirus risk because local transmission can only occur in the presence the vector. Moreover, the likelihood of local transmission can be linked to the relative abundance of the vector in a given geographic location.

SDM estimates the relationship between known presence sites of a species and the environmental characteristics of those sites. SDM is a key tool in assessing the impact of the environment on species distributions. For infectious diseases, SDM has been used for risk mapping of DENV in Colombia (Arboleda et al., 2009) and Mexico (Machado-Machado, 2012), and in Texas to assess risk of Chagas, another arthopod transmitted disease, based on areas of suitable habitat for its triatomine vector species (Sarkar et al., 2010). Strengths of SDM in describing geographic vector distribution include:

- ability to incorporate data at a high resolution spatial scale
- ability to extrapolate from sparse data sets
- ability to identify important predictors
- ability to output habitat suitability maps over the geographic area of interest.

SDM's output maps can be especially useful for understanding potential transmission areas, and areas with relatively high vector presence.

The best vector distributions prior to this project are provided by the AgriLife Extension Service Project as seen in the *Ae. aegypti* and *Ae. albopictus* figures found on the Agricultural and Environmental Safety, Texas A&M AgriLife Extension webpage ((Johnsen) http://agrilife.org/aes/public-health-vector-and-mosquito-control/mosquitoes-of-texas/). The distributions are generated based on literature and mosquito trapping conducted by the Agricultural and Environmental Safety Unit personnel. These maps provide a baseline knowledge of where vector species have historically been found. The maps present presence and absence data on a county level.

SDM contributes to this mapping effort by producing higher-fidelity maps that provide relative abundance counts at a finer geographic resolution. Relative abundances are useful to derive relative risks of sustained transmission in different counties. Furthermore, finer geographic resolution may identify that a mosquito population is contained within one focal population within the county, and not uniformly spread throughout. Finally, SDM contributes to our understanding of what environmental factors about these counties make them suitable for *Ae. aegypti* and *Ae. albopictus*.

We use a popular SDM software package, MaxEnt((Phillips et al., 2006) https://www.cs.princeton.edu/~schapire/maxent/), version 3.3.3k, to separately model habitat suitability for *Ae. aegypti* and *Ae. albopictus* across locations in Texas. MaxEnt is a common method for SDM because of its predictive performance and ability to work with sparse presence-only data sets (Elith et al., 2011).

MaxEnt uses maximum entropy methods to estimate relationships between environmental variables and the habitat suitability using species presence-only data. The output is distribution of vector occurrence over a specified geographic space (Phillips et al., 2006). Intuitively, maximum entropy chooses the distribution that is closest to a uniform across the geographic space, while fulfilling environmental constraints specified by the presence-only records. The constraints enforce environmental characteristics, such as the mean temperature, over the predicted species distribution to match the empirical characteristics derived from the input presence-only data. The output of MaxEnt can be interpreted as a relative species occurrence rate across the region of interest.

2.1 Data

The two inputs of MaxEnt are environmental variables and occurrence data. Environmental variables are the key constraints placed on the species distribution. For our study, the input environmental variables consist of bioclimatic, topographic, and socioeconomic data at a resolution of 30 arc-seconds and are listed in Table 1. These consist of a standard set of 19 bioclimatic variables derived from the WorldClim database (last accessed 3-April-2015). The data layers for aspect and slope are derived from the elevation layer using the SDMTools package in R Statistical Software. Three additional data layers are also included based on the disease ecology of the vector species. *Ae. spp.* are urban dwelling mosquitoes that breed in artificial containers and feed almost exclusively on humans (Harrington et al., 2001): population count from the Socioeconomic Data and Applications Center (http://sedac.ciesin.columbia.edu/data/set/grump-v1-population-density), maximum vegetation index from the USGS Land Cover Institute (http://landcover.usgs.gov/green_veg.php), and measurement of artificial surface from the FAO GeoNetwork (http://sedac.ciesin.columbia.edu/data/set/grump-v1-population-density) as variables that relate to the food source, and therefore predict survival.

Occurrence data refers to geo-referenced species presence-only locations found within the geographic study area. In order to find enough occurrence points for of both *Ae. spp.* for MaxEnt to reliably estimate each vector species' distribution, the geographic study area includes the states of Arizona, New Mexico, and Oklahoma, as well as the country of Mexico. Documented observations of *Ae. aegypti* and *Ae. albopictus* were collected from a thorough literature search and mosquito control units in the city of Lubbock and El Paso. A presence location is collected if the observation is accompanied by a geo-referenced latitude and longitude coordinate. Geo-referencing errors are calculated using the MaNIS geo-referencing protocol (Wieczorek, 2012). To remain consistent with the spatial resolution of the environmental layers, presence points with estimated errors of greater than 30 arc-seconds are removed from the analysis. The final set of data points used to run the models contained 188 presence locations for *Ae. aegypti* and 76 locations *Ae. albopictus*. References for these points can be found in Table 2 and Table 3.

Temperature	Precipitation	Topographic	Urban
Annual Mean Temperature	Annual Precipitation	Slope	Population Count
Mean Diurnal Range	Precipitation of Wettest Month	Aspect	Artificial Surface Cover
Isothermality	Precipitation of Driest Month	Elevation	Maximum Green Vegetation Cover
Temperature Seasonality	Precipitation Seasonality		
Max Temperature of Warmest Month	Precipitation of Wettest Quarter		
Min Temperature of Coldest Month	Precipitation of Driest Quarter		
Temperature Annual Range	Precipitation of Warmest Quarter		
Mean Temperature of Wettest Quarter	Precipitation of Coldest Quarter		
Mean Temperature of Driest Quarter			
Mean Temperature of Warmest Quarter			
Mean Temperature of Coldest Quarter			

Table 1: Complete Set of Environmental Predictors

Country	State	Num. of Points	reference
United States	Texas	42	(Soto, May 4, 2015)
United States	Arizona	37	(Merrill et al., 2005)
United States	Texas	31	(Barney, 2008)
United States	Texas	16	(Merrill et al., 2005)
United States	Texas	4	(Vitek et al., 2014)
United States	Texas	2	(Kavanaugh, 2008)
United States	Texas	2	(Cano et al., 2015)
United States	Texas	1	(McPhatter et al., 2012)
Mexico		36	(Gorrochotegui-Escalante et al., 2002)
Mexico	Chihuhua	3	(de la Mora-Covarrubias et al., 2010)
Mexico	Nuevo Leon	1	(Moffett et al., 2009)

Table 2: Presence Points Used for MaxEnt Input, *Ae. aegypti*

Country	State	Num. of Points	reference
United States	Texas	10	(White, 2008)
United States	Texas	6	(Segura, May 1,2015)
United States	Texas	5	(McPhatter et al., 2012)
United States	Texas	3	(PHCR-West, 2002)
United States	Texas	3	(Kavanaugh, 2008)
United States	Texas	2	(Soto, May 4, 2015)
United States	New Mexico	1	(Powers et al., 2006)
Mexico		8	(Reyes-Villanueva et al., 2013)
Mexico	Nuevo Leon	6	(Pesina et al., 2001)
Mexico		2	(Marina et al., 2011)
Mexico	Coahuila	2	(Ibáñez Bernal and Martínez-Campos, 1994)
Mexico	Coahuila	1	(Sanchez-Rodríguez et al., 2014)

Table 3: Presence Points Used for MaxEnt Input, *Ae. albopictus*

2.2 Model Construction

The model construction process includes determining which environmental variables to include and how the variables interact with each other. From the twenty-five variables, we test numerous combinations to identify the combination that is most predictive of species occurrence. For each combination tested, 70% of the occurrence points are used as input training data and 30% of the occurrence points are withheld to be used as independent test data.

Each combination of environmental variables is tested 100 times, using a different randomization of training and test set each time. With the exception of reducing complexity, discussed further in Section 3.2.4, and sub-sampling the data into test and training sets, all MaxEnt parameters are left on default settings. Although there is the possibility to further customize models when estimating the distribution of a single species, Phillips et al. (2006) suggests the default parameters when there is the possibility of bias in the presence points, and especially with a small number of input presence points.

2.3 Assessment of Model of Performance

The MaxEnt software package uses the standard interpretative metric of area under the receiver operating characteristic curve (AUC) measure model performance. The AUC is a comparison of the model's specificity and sensitivity, with an optimal model having an AUC close to 1 and a model that predicts species occurrences at random having an AUC closer to 0.5. A natural statistical interpretation of this metric is model's ability to correctly identify presence and absence locations over a geographic space.

We use the average test AUC over the 100 runs for evaluating model performance. We choose to use the test AUC instead of the train AUC because the train AUC may favor models that over fit (Warren and Seifert, 2011). While the AUC may appear to be an intuitive metric for assessing, its implementation in MaxEnt must be taken with caution. In order to calculate true negative and false positive rates, MaxEnt generates a set of pseudo-absences from the background points that are not indicated as presence locations in the input. However, these pseudo-absences are not necessary true absences of the species in that location.

Presence of a species may not be detected in an area for a number of reasons. One important reason is bias sampling, where some areas receive higher sampling effort than others. To qualify a location as absent of a species requires persistent sampling over time. Despite these uncertainties in the MaxEnt implementation, and the tendency to have an unjustified level of confidence in the AUC, we use the test AUC as a measure of relative model performance because of its intuitive meaning, its ease of implementation in the MaxEnt software package, and its previous use in SDM literature.

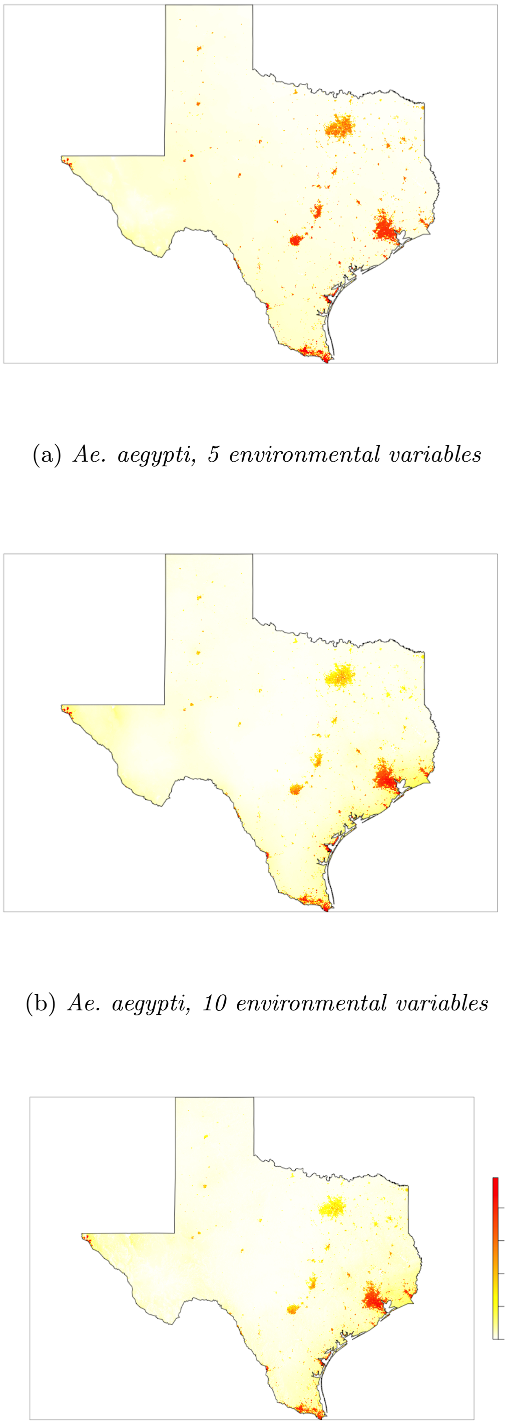
2.4 Model Selection

From the twenty-five environmental variables used to model *Ae. aegypti* and *Ae. albopictus* species distribution, we select ten and nine variables, respectively, for the final models. Four environmental variables (mean temperature of coldest quarter, mean temperature of the driest quarter, mean temperature of the warmest quarter, and mean temperature of the wettest quarter) have discontinuities where the data was not interpolated correctly at the 30 arc-second resolution, and therefore can be eliminated. With the remaining twenty-one variables, an analysis of the importance of each variable allows us to down-select to the most important environmental predictors.

The goal of down-selection is to generate a simple yet robust model, in order to avoid over-fitting and construct interpretable results. Starting with the twenty-one variables, each run eliminates the environmental variable that contributes the least to the model. This backwards elimination process continues until we reach a model with the fewest number of environmental variables that can predict presence and pseudo-absence locations as well as the model with all twenty-one environmental variables.

We determine the stopping point for backwards elimination using a two-sample t-test of the test AUC from the 100 runs of each model combination. We further check for functional changes in the map to see if fewer environmental variables led to significantly different suitability maps. As seen in Figure 1, only including the top five most predictive environmental variables produces a different suitability map for *Ae. aegypti* than the model with the top ten variables. However, the suitability map produced by the model with the top ten predictive variables and that from all twenty-one are functionally as well as statistically the same.

Based on the top models, Table 4 lists the environmental variables in their order of importance. For both vector species, the most influential variables are those associated with urban environments, such as population count and artificial surfaces. variables in the other three categories of



(a) *Ae. aegypti, 5 environmental variables*

(b) *Ae. aegypti, 10 environmental variables*

(c) *Ae. aegypti, 21 environmental variables*

Figure 1: Comparison of suitability maps produced by models with 5, 10, and 21 environmental variables.1a AUC: 0.937 (0.020), 1b AUC: 0.951 (0.015), 1c AUC: 0.958(0.013). The scaling of the figures is logarithmic to highlight differences in smaller values of relative occurrence rate. The raw format has only a small number of sites with relatively large values and is not visually informative. Throughout the rest of the report, all visuals will be displayed on a logarithmic scale, but all analysis was completed using the raw output.

temperature, precipitation, and topography are also included in the best models for both vector species. The environmental predictors chosen and their order is not identical between the two *Ae. spp.* The model for *Ae. aegypti* includes all three urban focused variables whereas *Ae. albopictus* only includes *artificial surfaces.*

In addition to testing the combination of environmental variables to include, we also explore the effects of limiting the complexity of the nonlinear response curves MaxEnt can build into model. MaxEnt offers five feature classes for environmental variables: the raw environmental variables (linear), features derived from squares of the variable that constrain the variance (quadratic), features of the product of pairs of environmental variables (products), binary features based on environmental threshold for the variable, and hinge features. By default MaxEnt uses the number of presence locations to determine which feature classes to explore. MaxEnt explores all features if the number of input presence locations is greater than 80.

As *Ae. albopictus* had fewer than 80 points, we limit models for *Ae. albopictus* to only linear, product, and quadratic feature classes. We tested the effects of incorporating higher or lower levels of complexity for *Ae. aegypti.* Through those tests, it is clear that the more feature classes the model includes, the more constrained the model becomes, with a smaller geographic spread of the species distribution. This is a sign of over-fitting. For *Ae. aegypti*, we choose to also restrict the possible feature classes to linear, quadratic, and product functions. We choose to stay with models that include product and quadratic features in addition to linear features because their AUC was significantly better than the model with only linear features.

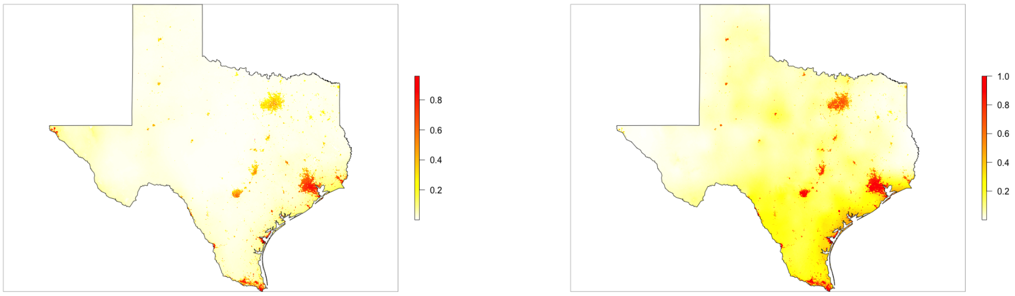
2.5 Habitat Suitability of Vector *Ae. spp.* in Texas

Figure 2 presents high resolution maps for the final habitat suitability models for both vector species. The locations that have a high relative occurrence of each vector species are consistent with the Texas A&M AgriLife Extension distributions, further validating the final models. The sum over all 30 arc-second cells within a county generates a county-level relative suitability index, presented in Figure 3. The top counties are listed in Table 5.

Comparing the results of Figures 2 and 3, it is clear that the high resolution map contains more information on suitable locations. Areas, particularly along the coast, have regions of high relative occurrence rates in the high resolution map, but in the county map the entire county shows up as being at a higher ecological risk for habitat suitability. These fine resolution geographic differences

Ae. aegypti	*Ae. albopictus*
population count	artificial surfaces
artificial surfaces	isothermality
elevation	temperature seasonality
temperature seasonality	mean diurnal range
annual mean temperature	elevation
maximum green vegetation fraction	minimum temperature of coldest month
precipitation of driest quarter	precipitation of coldest quarter
precipitation seasonality	precipitation seasonality
minimum temperature of coldest month	precipitation of warmest quarter
mean diurnal range	

Table 4: Environmental predictors in order of importance. The importance of the variables is determined by its percent contribution to the gain of the model.

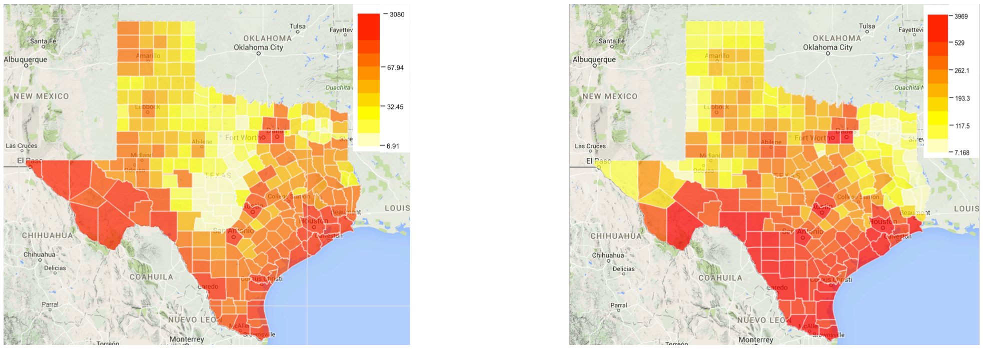


(a) *Ae. aegypti*, High Resolution

(b) *Ae. albopictus*, High Resolution

Figure 2: High Resolution relative habitat suitability distributions of *Ae. aegypti* and *Ae. albopictus* in Texas. Scores indicate relative occurrence rates of the vector species, with the the location of the highest relative occurrence rate receiving a score of 1. The *Ae. aegypti* model had an average AUC of 0.965 with a standard deviation of 0.013. The *Ae. albopictus* model has an average AUC 0.939 with a standard deviation of 0.022. Both highlight urban areas as having higher relative occurrence rates, although *Ae. albopictus* estimates the relative occurrence to be slightly more uniform throughout eastern and central Texas. *Ae. aegypti* has a high relative occurrence in the area around El Paso, which in contrast is low for *Ae. albopictus*

.



(a) *Ae. aegypti*, County Level

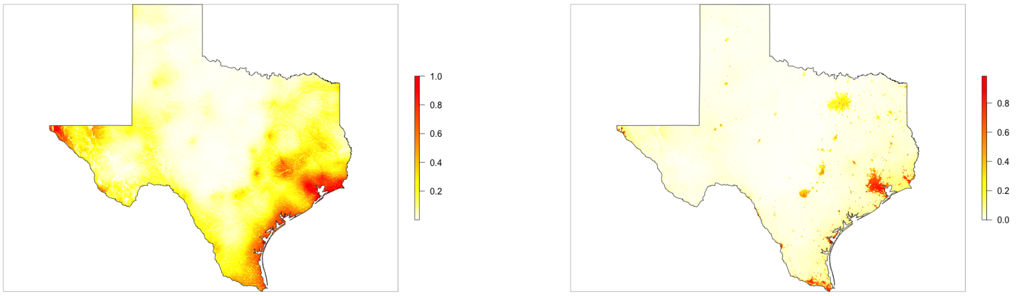
(b) *Ae. albopictus*, County Level

Figure 3: County level relative habitat suitability distribution of *Ae. aegypti* and *Ae. albopictus* in Texas. The relative occurrence rates of each 30 arc-second location was aggregated over the county to produce the relative occurrence rates for each county. The suitability of both vector species have higher ecological habitat suitability in counties along the eastern gulf and in the eastern interior of the state. High population counties such as El Paso have a high suitability score for *Ae. aegypti* but not for *Ae. albopictus*.

Ae. aegypti	*Ae. albopictus*
Harris	Harris
Hidalgo	Cameron
Bexar	Hidalgo
Montgomery	Bexar
El Paso	Nueces
Cameron	Webb
Travis	Dallas
Brazoria	Tarrant
Dallas	Galveston
Fort Bend	Brazoria

Table 5: Top ten counties for habitat suitability for *Ae. aegypti* and *Ae. albopictus*.

may be important in targeting communication and surveillance efforts.



(a) *Ae. aegypti*: bioclimatic data layers

(b) *Ae. aegypti*: bioclimatic and urban data layers

Figure 4: Comparison of suitability maps produced by models excluding (4a) and including (4b) urban focused data layers. The relative occurrence of the vector species is more uniformly distributed across Texas when based only on the precipitation and temperature environmental variables in 4a. When urban focused data layers are added, the relative occurrence becomes more constricted to areas of high population. The AUC for 4a is 0.921 (0.017) and for 4b is 0.958(0.013), indicating a better predictive performance when the data layers of population, artificial surface coverage, and maximum green vegetation fraction are included.

High-suitability locations occur in urban areas. This concurs with the biological understanding of the vector species, specifically container breeding and primarily feeding on humans. Figure 4 compares models that only include bioclimatic variables (temperature, precipitation, and topographic) and those models that also included the urban data layers. Models with urban layers have significantly better predictive performance. Bioclimatic-data-only models estimate more uniform rates of occurrence over a broader geographic region. Consistent trapping data could be an independent evaluation of relative occurrence rates and will be discussed further in the report.

3 Geographic Risk Distribution

It is possible to define three types of risk for a non-endemic disease: the risk of importing a case, the risk of an imported case subsequently leading to a chain of autochthonous transmission, and risk for endemic establishment. In this project, we consider how to assess risk based on the first and second categories, which we refer to as ”Import Risk” and ”Sustained Transmission Risk.” The following sections detail models of import risk and sustained transmission risk for IKV and DENV in Texas.

In other countries, risk maps for DENV and CHIKV have used a wide variety of modeling approaches and predictor variables. Effective predictor variables include demographic, socio-economic, environmental, and historical case data. However, depending on the location, different categories of predictor variables may be more important than others. Most of the risk maps have been descriptive and based on the historical data. Analysis approaches in these modeling approaches include logistic regression models, multinomial models, generalized linear models, general additive models,generalized linear mixed models, environmental niche, maximum entropy approach of species distribution modeling, Kernel estimations, geographically weighted regression, kriging and co-kriging, Knox test concept and etc. (Louis, 2014).

3.1 Import Risk Map

DENV has been imported into Texas regularly since 2002, while CHIKV was first reported only recently. The lack of direct CHIKV data limits the creation of a data-driven CHIKV model directly. However, both DENV and CHIKV are transmitted by the same *Ae.* mosquitoes. As such, we use DENV import cases as a proxy for CHIKV risk. In other words, we construct a data driven DENV import risk map that serves as a proxy to indicate CHIKV import risk level. We use a Maximum entropy approach, as again as have presence-only DENV data and the historical DENV data is sparse.

3.1.1 Data

Human DENV occurrence data from 2002 to 2012 is available to DSHS for modeling. DENV is not endemic in Texas and only sparse DENV occurrence data is available for each county over the past ten years. We use the historic DENV imports to construct the import risk model. In addition to this import data, the model also takes as input environmental, socio-economic and demographic data. The model output is the probability that the next imported case happens in each county.

The input variables to the model include environmental, socio-economic, and demographic data. Environmental data, such as temperature, is known to be important in adult vector survival, viral replication, and infection periods (Murray NEA, 2013). As the proliferation of *Ae.* mosquitoes depends on climate, temperature, humidity, and precipitation we include related variables as potential influential factors for DENV risk map construction. Socio-economic and demographic data, such as age, gender, education, and level of income are commonly used demographic and socioeconomic variables (Louis, 2014), and are also included as input variables. We include additional socio-economic and demographic data such as county population size, employment status, population below poverty, ways of communicating to work, and health insurance coverage. In addition, we include travel data for each county, adding an additional variable in comparison to published risk model studies. Counties with more travelers or to DENV and CHIKV endemic countries should exhibit higher import risk. While direct travel data to endemic areas would be ideal, such data was not available. The best available travel data described the size of tourism industry for each county.

Since DENV and CHIKV are primarily transmitted by *Ae. aegypti* and *Ae. albopictus*, the results from our vector species distribution were also included. Our full data set contained 76 factors from four categories which are shown in Table 6.

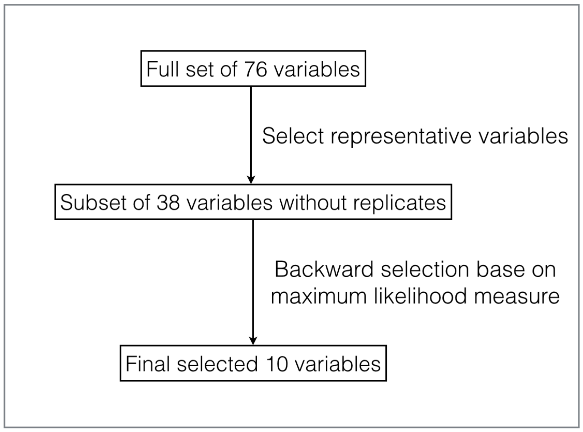
Environmental	Socio-economic	Demographic, Travel and Vector Suitability
Annual Mean Temperature	Employed Population	Male Population
Annual Precipitation	Unemployed Population	Female Population
Slope	Employed Population in Percentage	Male Population in Percentage
Population Count	Unemployed Population in Percentage	Female Population in Percentage
Isothermality	Population below Poverty Level in Percentage	Local(dollars)
Precipitation of Driest Month	Families below Poverty Level in Percentage	State(dollars)
Elevation	Population with Health Insurance	Total Direct Spending(dollars)
Maximum Green Vegetation Cover	Percentage with Health Insurance	Visitor Spending
Temperature Seasonality	Population without Health Insurance	Earnings(dollars)
Precipitation Seasonality	Percentage without Health Insurance	Travel Employment
Max Temperature of Warmest Month	Mean Travel Time to Work(Minutes)	Albopictus Abundance(Total)
Precipitation of Wettest Quarter	Population Walk to Work	Albopictus Abundance(Average)
Min Temperature of Coldest Month	Population Walk to Work in Percentage	Average MGV (percentage per km)
Precipitation of Driest Quarter	Percentage Commuting to Work with Taxi	Total Approximate MGV Cover (km)
Temperature Annual Range	Commuting to Work with Taxi	Aegypti Abundance (Total)
Precipitation of Warmest Quarter	Percentage Commuting to Work with Public Transportation	Aegypti Abundance(average per km)
Mean Temperature of Wettest Quarter	Commuting to Work with Public Transportation	
Precipitation of Coldest Quarter	Commuting to Work with Car, Truck or Van (Carpooled)	
Mean Temperature of Driest Quarter	Commuting to Work with Car, Truck or Van(Alone)	
Mean Temperature of Warmest Quarter	Percentage Commuting to Work with Car, Truck or Van(Carpooled)	
Mean Temperature of Coldest Quarter	Percentage Commuting to Work with Car, Truck or Van(Alone)	
Mean Diurnal Range	Commuting to Work with Other Means	
Precipitation of Wettest Month	Percentage Commuting to Work with Other Means	
Aspect	Education Attainment below 9th grade	
Artificial Surface Cover(Percentage)	Education Attainment below 9th grade in Percentage	
Total Artificial Surface Cover (km)	Education Attainment between 9th and 12th grade	
	Percentage Education Attainment between 9th and 12th grade	
	High School Graduates	
	High School Graduates in Percentage	
	College without diploma	
	College without diploma in Percentage	
	Associates degree	
	Associates degree in Percentage	
	Bachelor's degree	
	Bachelor's degree in Percentage	
	Graduate or professional degree	
	Graduate or professional degree in Percentage	

Table 6: Complete Set of Variables for Import Risk Map Modeling

3.1.2 Modeling Approach

To model the import risk, we use the maximum entropy method introduced in Section 2 to estimate the relative probability for the next DENV importation case to happen in each Texas county. The input historical DENV cases represent an empirical distribution of import. That empirical distribution produces estimated expectations on each of the input variables, for example expected travel industry size, or expected poverty level. The maximum entropy method produces an output probability distribution that agrees with those empirical, observed expectations. From all the distributions that agree with those observed expectations, the method selects the distribution with maximum entropy—in other words, the one closest to uniform.

We use an out-of-sample likelihood calculation to evaluate potential models. Specifically, ten years DENV import data from 2002 to 2012 is divided into two data sets, train years and test years. The test years consist of three years of DENV imports. The train years and a combination of input variables generate a probability distribution over Texas counties using the maximum entropy method. To evaluate the resulting model, we compute the likelihood of observing the test years' data under the model's output distribution. This allows us to compare different models, with better models producing higher out-of-sample likelihoods. Practically, we compare log-likelihoods instead


Full set of 76 variables
Select representative variables
Subset of 38 variables without replicates
Backward selection base on maximum likelihood measure
Final selected 10 variables

(a) *Flow Chart of Variable Selection Procedure*

Figure 5: We start with full set of 76 variables. After removing variables based on our replicated variables elimination procedure we have 38 variables remaining. We sub-setting these remaining 38 variables through our backward selection process backward selection. In each step, we drop the variable that contributes the least to the model performance. Backward selection continues until all the variables are eliminated. In each step, model performance is measured using the out-of-sample log-likelihood measure.

of likelihoods.

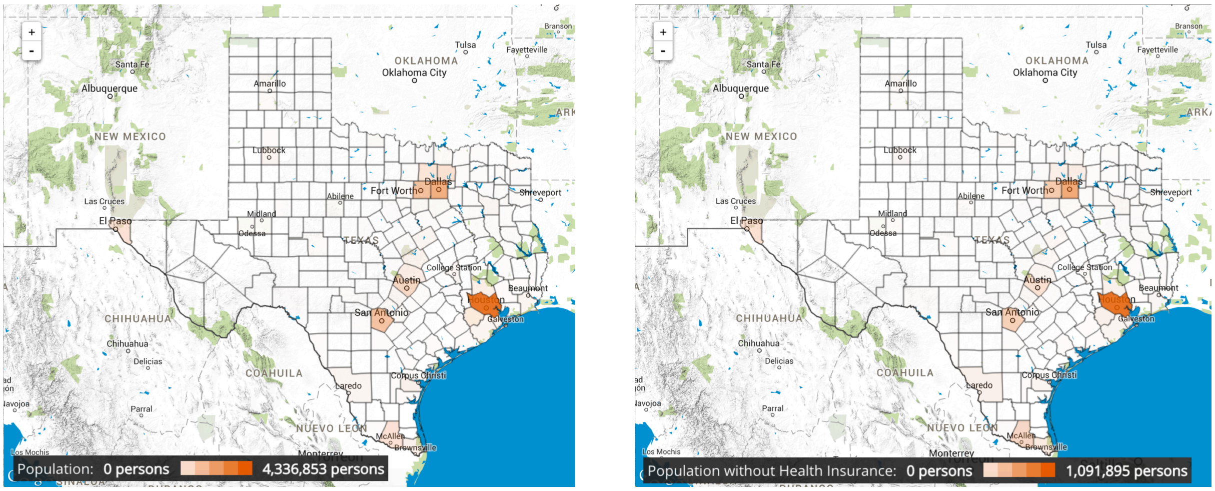
This out-of-sample log-likelihood comparison allows us to select the best model out of combinatorially many competing models. Each combination of the 76 factors yields a potential import risk model, for a total of at least $2^{76} \approx 10^{23}$ models. In addition, other models can be created, such as simple empirical models. In the following sections we describe methods to select the best model out of this combinatorially large set of possibilities.

3.1.3 Variable Selection

Selecting the best model consists of several steps, depicted in Figure 5. As an overview, the first step is to reduce the 76 available variables to a smaller set of 38 representative variables. This step removes variables that introduce duplicate information. The second step is to perform backward selection on the 38 representative variables to reduce the model size to about 10 input variables. We describe each of these steps in turn.

The first step in selecting a model is to remove duplicate variables—variables that essentially bring the same information to the model. We call this step selecting representative variables. Selecting representative variables has nothing to do with the DENV import data, and only has to do with the information contained in the input variables. As an example, Figure 6 depicts county population without health insurance is total county total population. These two input variables are essentially constant multiples of each other. As such, they provide essentially the same constraints in the maximum entropy model, and only one of them is necessary regardless of the DENV import data.

Selecting the representative variables done with a variant of the facility location problem. The $\ell{-}\infty$ norm of the difference between two unit-norm variables is assigned as the distance between the two variables. This distance measure is derived from the maximum difference in expectations



(a) *Population without health insurance*

(b) *Population estimated in July, 2013*

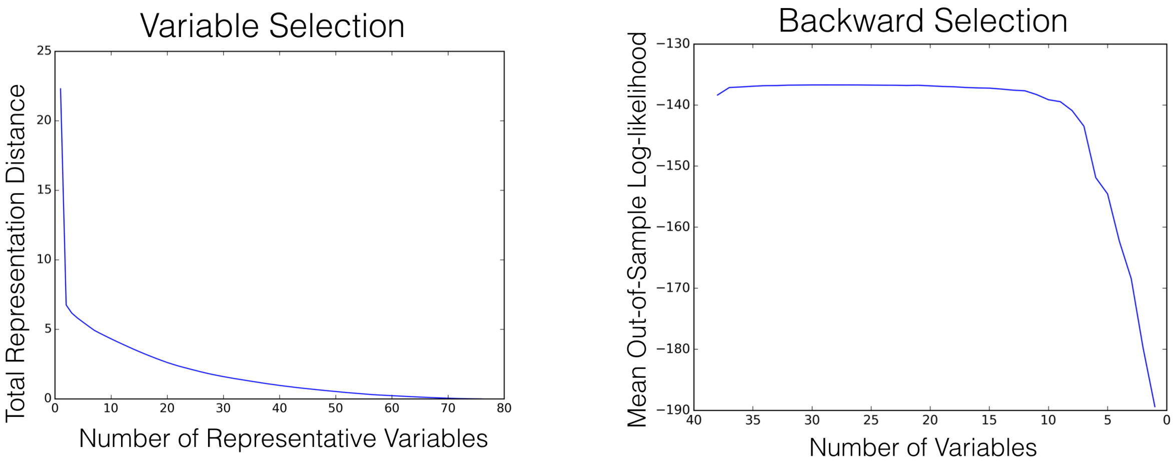
Figure 6: County population without health insurance and county population are essentially constant multiples of each other. As such, they provide essentially the same constraints in the maximum entropy model, and only one of them is necessary regardless of the DENV import data.

that the two variables can produce, under any probability distribution. The facility location model allows us to select the k variables that best represent others, subject to the computed distances. The objective function for selecting representative variables is to minimize the distance between the k representative variables and the remaining variables. Each variable is represented by exactly one of the k representatives. As more representative variables are selected, k is closer to 76, the objective value decreases since more variables represent themselves. Figure 7a below shows the objective function as the number of representatives k changes. Based on this output, we select 38 representative variables.

The second step of selecting a good import risk model is backward selection. The backward selection procedure starts with all 38 variables of interest. In each step, the variable that contributes the least to the model performance is dropped. Backward selection continues until all the variables are eliminated. In each step, model performance is measured using the out-of-sample log-likelihood measure.

Selecting representative variables should be done prior to backward selection, as we have done, to avoid spurious results. With duplicate variables, backward selection can eliminate the first duplicate arbitrarily as contributing nothing to the model. If we re-run backward selection, it may eliminate the second duplicate. In the presence of duplicates, backward selection cannot robustly identify which input variables contribute most to model performance.

To ensure a robust backward selection procedure on the 38 representative variables, we implemented cross-validation to evaluate out-of-sample log-likelihoods. Specifically, we do not measure out-of-sample log-likelihood on just one set of 3 years. Instead, we sample 7 combinations of test and train years, and the model is fit 7 times. The mean of the out-of-sample log-likelihoods for these 7 runs of the model is used as the selection measure. Figure 7b depicts the change in mean out-of-sample log-likelihood as variables are eliminated during backward selection. Model performance increases at first, as some of the 38 variables are removed, but decreases after reaching the maximum at 29 variables. The figure also shows that the drop in model performance is negligible



(a) *objective values in variable selection*

(b) *objective values in backward selection*

Figure 7: (a) The error of using 38 variables as representatives is small, below 5% of the error of using just 1 representative variable. (b) The maximum log-likelihood model contains 29 variables, but a model with just 10 performs comparably well.

until less than 10 variables remain.

3.1.4 Results Analysis

To summarize, half of the variables (38 out of 76 variables) are selected as representative variables. Figure 7a shows that the error of using these 38 variables as representatives is small, below 5% of the error of using just 1 representative variable. Through backward selection, we produce a model with just 10 input variables. Figure 7b shows that the maximum log-likelihood model contains 29 variables, but a model with just 10 performs about as well. The 10 variables selected in the final model are shown in Table 7.

The maximum entropy method uses the 10 input variables as constraints for generating the output probability distribution. As such, maximum entropy does not naturally compute correlations between the output variable and input variables. To gain intuitive understanding between these variables and the probability of import, in Table 7 we compute correlation coefficients at the end of model generation. Intuitively, as imported DENV cases are brought by the travelers from DENV endemic areas and countries, variables which indicate international travel should play an important role in generating the import risk probabilities. Even though the suitability of the mosquitoes plays an important role in sustained transmission risk, the suitability of the *Ae. spp.* should not be considered important for import risk. Our model selection procedures confirm this intuition as none of the variables indicating the suitability of the *Ae. spp.* are selected. The majority of the variables selected have to do with education, and indicators of being in a city.

We compare our model with three intuitive models: an empirical distribution often used for plotting an incidence map and a maximum entropy model with 10 and 5 intuitively selected variables, with variables shown in Table 8 and Table 9. The empirical probability for each county is calculated by dividing the cases occurring in the train years in that county with the total cases occurring in Texas in the same period. To compare the models, we employ a cross-validation procedure. Every combination of three years from 2002 to 2012 is treated as test years, and the remaining seven

Variables in Order of Importance	*Correlation Coefficient*
Educational Attainment with Bachelor's degree	0.8267
Minimum Temperature of Coldest Month	0.2265
Percentage of Using Public Transportation to Work	0.7766
Educational Attainment in some college with no degree	0.7909
Walked to Work	0.8155
Commuting to Work with Other Means	0.8387
Educational Attainment less than 9th degree	0.7856
Percentage of Educational Attainment with Graduate or professional degree	0.3833
Percentage of Walked to Work	-0.0907
Average Artificial Surface (Percentage)	0.8271

Table 7: Top 10 variables after the variable and backward selection procedures. Correlation coefficients between variables and relative probability distribution reveal more direct relationships.

years are treated as train years. The average out-of-sample log-likelihood over all combinations is calculated as the model performance for each model. Table 10 shows that the model resulting from our method produces significantly better performance than any of the intuitive models.

10 Variables in Intuitive Model	
Total Population	Percentage of people under poverty level
Earnings(dollars)	Local(dollars)
Travel Employment	Percentage of people being unemployed
Number of people without health insurance	Average Artificial Surface (percentage)
Number of people with Bachelor's degree	Percentage of people with graduate or professional degree

Table 8: 10 Variables Selected for Intuitive model.

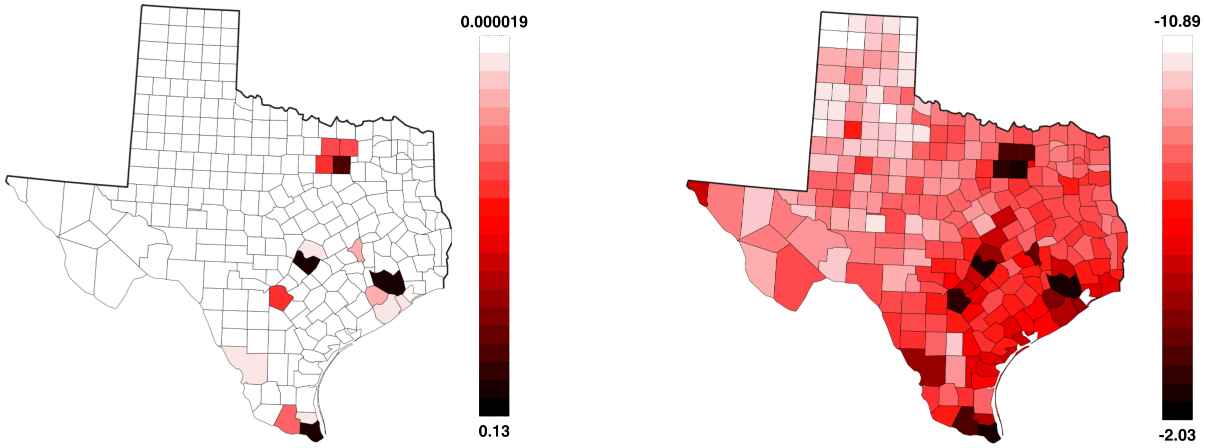
5 Variables in Intuitive Model	
Travel Employment	Percentage of people being unemployed
Number of people without health insurance	Average Artificial Surface (percentage)
Number of people with Bachelor's degree	

Table 9: 5 Variables selected for Intuitive model.

Two types of risk maps are generated based on the import probabilities, shown in Figure 8a and Figure 8b The first depicts the probability that the next import occurs in the specified county, and the second depicts the logarithm of the import probability. Harris county, Travis county and Cameron county are the counties in highest import risk for the next DENV importation case to occur. Dallas county is one level lower than above three counties but still in a much higher risk range than the remaining counties. Bexar county and Tarrant county are in the same risk level and Denton county and Collin county are in a one level lower risk range. All the counties in white color are in a much lower risk level than the others. The logarithm plot provides easier visualization of the relative risk levels of neighboring counties.

Models Ranked in Order of Goodness	*Mean out-of-sample log-likelihood*
Our Model (10 variables)	-146.3940
Intuitive Model (10 variables)	-158.1337
Empirical Distribution	-176.8625
Intuitive Model (5 variables)	-179.4058

Table 10: Model comparison: Cross-validation with all combinations of a three-year test period is used for model evaluation. Mean out-of-sample log-likelihoods across all combinations function as model performance measures. An empirical probability model is calculated by dividing the cases occurring in the train years in each county with the total cases occurring in Texas in the same years. Relative probabilities among counties from our model are calculated by fitting 10 selected variables from variable selection and backward selection procedures into the maximum entropy model. The intuitive models are generated by employing maximum entropy with 10 and 5 intuitively selected variables, described in Tables 8 and 9. The model resulting from our method performs significantly better than the competing intuitive models.



(a) *Risk map based on probability of next import.*

(b) *Risk map based on log-probability of next import.*

Figure 8: Two types of risk maps are generated based on the import probabilities. Figure (a) depicts the probability that the next import occurs in the specified county, and figure (b) depicts the logarithm of the import probability.

3.2 Sustained Transmission Risk

Once an infection is imported into the state, there is a risk that the infection will spread further through the local mosquito population. An imported human case of CHIKV or DENV can cause a chain of secondary cases if a susceptible *Ae.* mosquito bites the infected human and subsequently bites and infects another humans in quick succession. A basic model of the spread of infection among humans via the vectored transmission of mosquitoes can be is described in Keeling and Rohani (2008):

$$\begin{aligned}
\frac{dX_H}{dt} &= \upsilon_H - rT_{HM}Y_MX_H - \mu_HX_H, \\
\frac{dY_H}{dt} &= rT_{HM}Y_MX_H - \mu_HY_H - \gamma_HY_H, \\
\frac{dX_M}{dt} &= \upsilon_M - rT_{MH}Y_HX_M - \mu_MX_M, \\
\frac{dY_M}{dt} &= rT_{MH}Y_HX_M - \mu_MY_M.
\end{aligned}$$

From the above set of equations, we can derive a reproductive number R_0 for the infection. The reproductive number is defined as the average number of secondary cases produced by an infectious individual in a totally susceptible population. In vectored transmission, this definition encompasses two components. Starting with an infectious individual, the number of secondary human cases depends on 1) the expected number of mosquitoes the infectious individual transmits the virus to and 2) the number of infected humans from an infected mosquito. The R_0 is then given by the product of the average number of mosquitoes infected by one infectious human,

$$\frac{bT_{MH}N_M}{(\gamma_H + \mu_M)N_H}$$

and the number of humans infected from the primary mosquito,

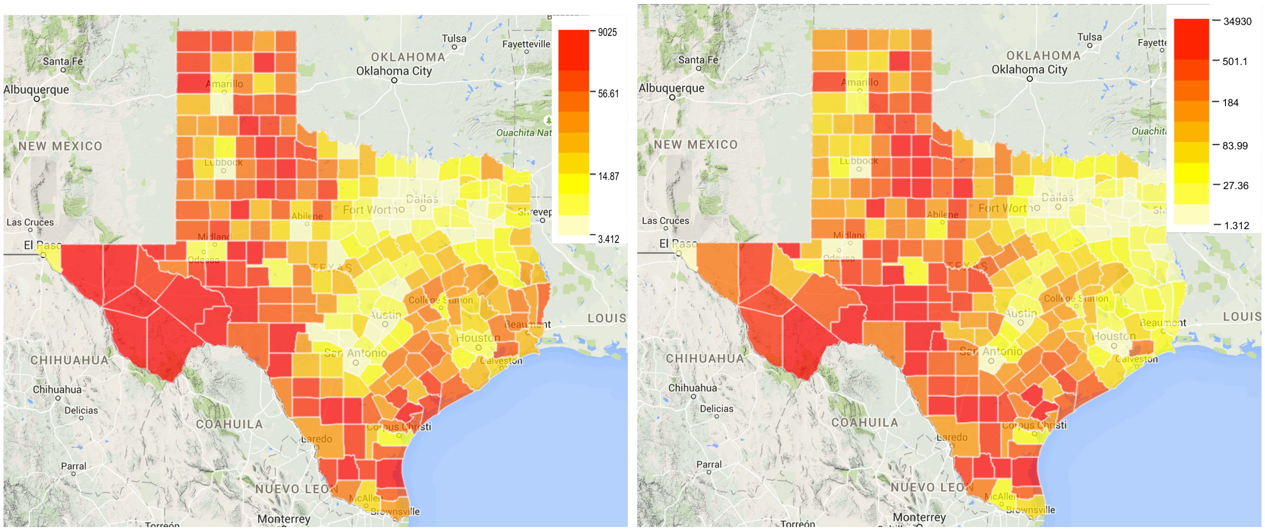
$$\frac{bT_{HM}}{\mu_M}.$$

The R_0 is:

$$R_0 = \frac{b^2T_{HM}T_{MH}N_M}{\mu_M(\gamma_H + \mu_H)N_H}. \tag{1}$$

Symbol	Parameter
υ_i	birth rate of species i
μ_i	per capita death rate for host species i
r	rate at which humans are bitten
γ	recovery rate for humans
T_{ij}	transmission probability following a bite to species i to species j
X_i	number of individuals of species i that are susceptible
Y_i	number of individuals of species i that are infectious

Table 11: Parameters in the transmission differential equation model



(a) *Ae. aegypti*

(b) *Ae. albopictus*

Figure 9: Relative sustained transmission risk based on the MaxEnt output of Figure 2 and the population of each county. Red indicates areas of higher risk and lighter areas indicate lower areas of sustained transmission risk.

To generate a sustained transmission risk map, we compare the relative values of R_0 between the counties. Because we are computing relative R_0 values, as opposed to exact values as in Equation 1, we can eliminate all parameters in the equation that are the same between counties. Parameters such as the mosquito biting rate b, the probability of transmission of the infection from mosquito to human and vice-versa T_{MH} and T_{HM}, the recovery infection time for humans γ_H, and the death rates for both humans and mosquitoes μ_H and μ_M can be assumed to be the same between counties. Removing these parameters, the key comparison for relative values of R_0 reduce to the ratio of mosquitoes to humans.

The output from MaxEnt SDM produces relative occurrence rates of the mosquitoes. Although this is not a measure of absolute abundance, it gives the relative occurrences of the mosquitoes between the counties. We can take the ratio of the MaxEnt output to the number of humans in each county to give us relative R_0 comparison, a comparison of sustained transmission risk between counties,

$$R_0 \propto \frac{N_M}{N_H}. \tag{2}$$

Sustained transmission risk yields significantly different maps, presented in Figure 9, than the habitat suitability maps in Figure 2. The main population centers have lower sustained transmission risk, R_0, relative to other areas in the state whereas they have higher vector suitabilities. Looking at relative R_0 formula we see that R_0 increase proportionally with the number of mosquitoes but decrease with the number of humans. This is because if the number of mosquitoes per human is low, than the initially infected human may not be bitten by any mosquitoes before recovering. Therefore, the sustained transmission risk map highlights areas in Texas that have a ratio of mosquitoes to humans is large, producing a higher risk for having an arbovirus successfully spread and invade once an initial import is present.

Ae. aegypti	*Ae. albopictus*
Loving	Kenedy
Kenedy	Loving
Hudspeth	Martin
Culberson	Edwards
Terrell	Terrell
King	King
McMullen	Borden
Jeff Davis	Kinney
Presidio	Kent

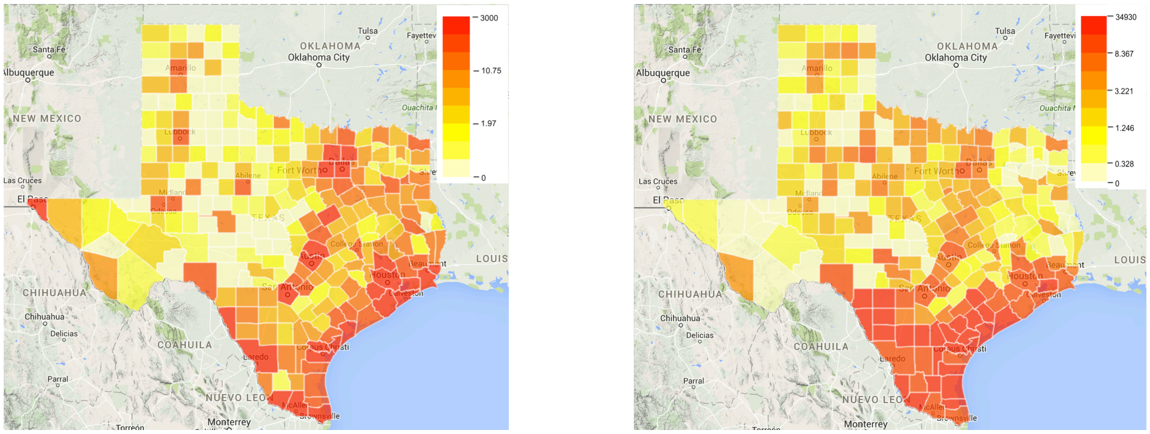
Table 12: Top ten counties for highest risk of sustained transmission

Many of the counties that are indicated as having the highest risk of sustained transmission also have the lowest human population sizes. There is the possibility that these small human populations could not sustain a sufficient mosquito population for transmission to occur by not providing sufficient breeding container opportunities or food source. One of the properties of maximum entropy is that it chooses the species distribution that is closest to uniform across geographic space while satisfying the environmental constraints. In this sense, counties that may not actually have any mosquito populations will still be assigned some small level of relative occurrence. For counties with small population sizes, in the R_0 calculation, the relative occurrence N_M is high enough that when divided by a small population N_H, the R_0 is large.

We consider how sustained transmission risk would change across the state by choosing a threshold of relative occurrence rate under which counties are re-assigned as 0 *Ae.* occurrence. The threshold is visually determined by progressively increasing the threshold until known areas of trapped *Ae.* mosquitoes are below the threshold. Figure 10 depicts sustained transmission risk maps based on these thresholds. The counties with the highest sustained transmission risk change, with many of the rural areas in west Texas that had a higher sustained transmission risk in Figure 9 now at lower risk. This map of sustained transmission risk has a greater intuitive agreement with the habitat suitability maps. Areas along the coast and high population density areas of Dallas, Houston, Austin, San Antonio, and El Paso in the case of *Ae. aegypti* are at high sustained transmission risk.

Ae. aegypti	*Ae. albopictus*
Harris	Kenedy
Cameron	Refugio
Dallas	Dimmit
Hidalgo	Kinney
Tarrant	Goliad
Bexar	Brooks
Jefferson	Jackson
Montgomery	Lampasas
Galveston	Zavala
Brazoria	Calhoun

Table 13: Top ten counties for highest risk of sustained transmission when thresholds of 0.1 and 0.20 are applied to the MaxEnt ouput



(a) *Ae. aegypti*; Threshold 0.1

(b) *Ae. albopictus*; Threshold 0.2

Figure 10: Sustained Transmission Risk Maps with thresholds applied to the MaxEnt output. The values of risk were calculated from the high resolution relative occurrence rate (Figure 2) and the high resolution population data according to equation (2). For each county the proportional R_0 was calculated as the sum of all 30 arc-second cells.

4 Tools to Help DSHS Mission

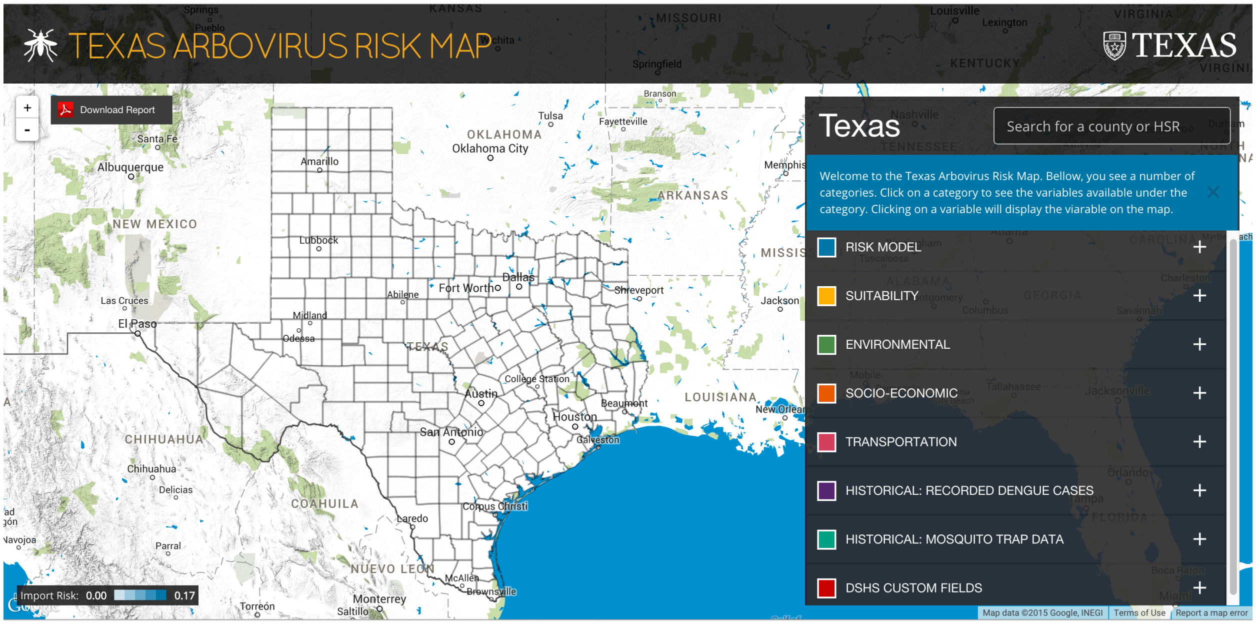
A state-of-the-art website was developed to facilitate DSHS in arbovirus surveillance, control and prevention. We aimed to integrate our results in the Texas education and surveillance effort for arbovirus with visualized dengue and chikungunya risk maps and *Ae. spp.* suitability maps.

Screenshots of the website are shown in Figures 11 through 14. After landing on the Texas Arbovirus Risk Map home, Figure 11, a user can navigate several categories of data. Each category presents several variables, and each variable can be displayed on a state level, an HSR level, or a county level. Some data, such as the environmental variables used for for constructing *Ae. spp.* suitability maps, can be displayed in high resolution. For example, as in Figure 12a, users can view the high resolution data of annual mean temperature by clicking the high resolution selector. Similarly, a user can view time-specific incidence counts of arbovirus disease through a time-selection slider. For example, historical dengue cases from year 2009 to 2012 are shown in Figure 12b.

Users can choose to view specific county data by clicking the county in the map or typing the county name in the search box as shown in Figure 13a. Users can also view a specific Health Service Region (HSR) by typing the HSR name in the search box. Whether at a county, HSR, or state zoom level, a user can click the PDF report button on the top left to generate an arbovirus report for their geographic region of interest.

A user can select variables of interest by navigating the categories of variables on the right. When selecting a variable of interest, a legend on the bottom left of the website describes the connection between visualization color and variable value. For example, in Figure 13b, the population of Travis county shown under the socio-economic category. The legend also shows the minimum and maximum value for the variable across all counties in Texas.

The import risk map, logarithm import risk map and sustained transmission risk map discussed in Section 3 are available under the risk model category. The *Ae. spp.* suitability maps are avail-



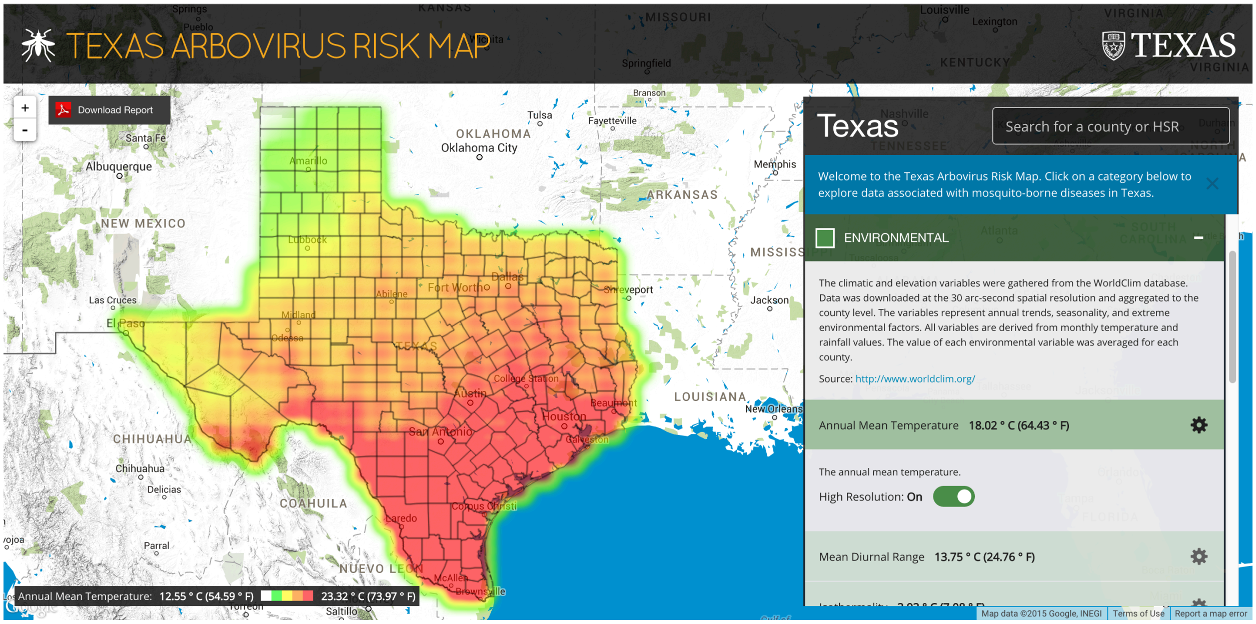
Welcome to the Texas Arbovirus Risk Map. Bellow, you see a number of categories. Click on a category to see the variables available under the category. Clicking on a variable will display the viarable on the map.

Figure 11: Homepage of the Texas Arbovirus Risk Map. A user can use the website to visualize risk maps, and important indicator variables. In addition, the website can be used to visualize timely arbovirus incidence data and generate weekly arbovirus risk reports. After landing on the website, a user can use the right-hand-side search box to search for an Health Service Region or a County. In addition, a user can select a variable of interest by navigating the categories on the right. A report is generated for the geographic region of interest by clicking the PDF report button on the top left.

able under the suitability category. All the environmental data, socio-economic data, travel data, historical reported dengue cases and historical mosquito trap data we used are also available.

The website also allows DSHS to add custom fields for display on the website, and to upload the most recent Arbovirus surveillance data through an administrative interface, displayed in Figure 14. Through this interface, DSHS officers can upload data for arbovirus surveillance reports, download the auto-generated reports, and download all website data in a zip file. DSHS can update the data displayed on the website by first editing appropriate spreadsheets in Google Drive, then clicking the load data button on the administrative interface.

This website provides an effective and direct visualization tool for the centralized surveillance repository in state level. In addition, each county and HSR can focus on its specific data. The website, as a publicly accessible and user friendly tool, can help authorities communicate arbovirus risk and help coordinate privatization the limited resource in arbovirus surveillance. In addition, the tools provided for DSHS to actively update and edit the website data, should make it a valuable tool for future arbovirus control efforts.



Welcome to the Texas Arbovirus Risk Map. Click on a category below to explore data associated with mosquito-borne diseases in Texas.

ENVIRONMENTAL

The climatic and elevation variables were gathered from the WorldClim database. Data was downloaded at the 30 arc-second spatial resolution and aggregated to the county level. The variables represent annual trends, seasonality, and extreme environmental factors. All variables are derived from monthly temperature and rainfall values. The value of each environmental variable was averaged for each county.

Source: http://www.worldclim.org/

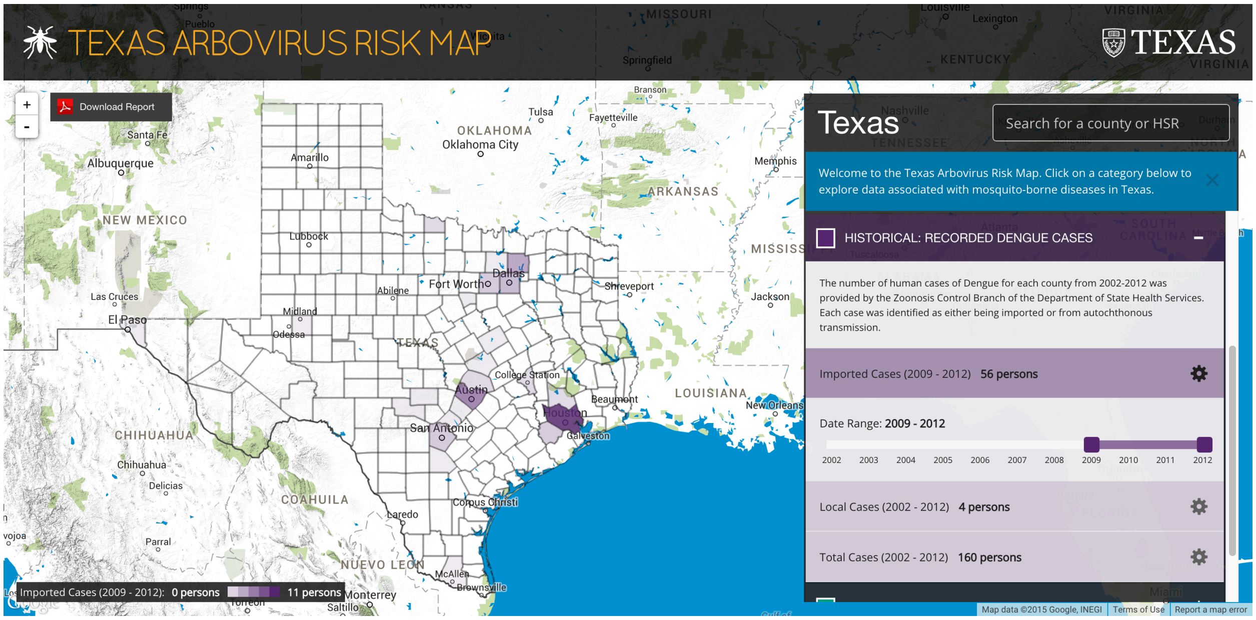
Annual Mean Temperature 18.02 ° C (64.43 ° F)

The annual mean temperature.

High Resolution: On

Mean Diurnal Range 13.75 ° C (24.76 ° F)

(a) *Average Mean Temperature in High Resolution*



Welcome to the Texas Arbovirus Risk Map. Click on a category below to explore data associated with mosquito-borne diseases in Texas.

HISTORICAL: RECORDED DENGUE CASES

The number of human cases of Dengue for each county from 2002-2012 was provided by the Zoonosis Control Branch of the Department of State Health Services. Each case was identified as either being imported or from autochthonous transmission.

Imported Cases (2009 - 2012) 56 persons

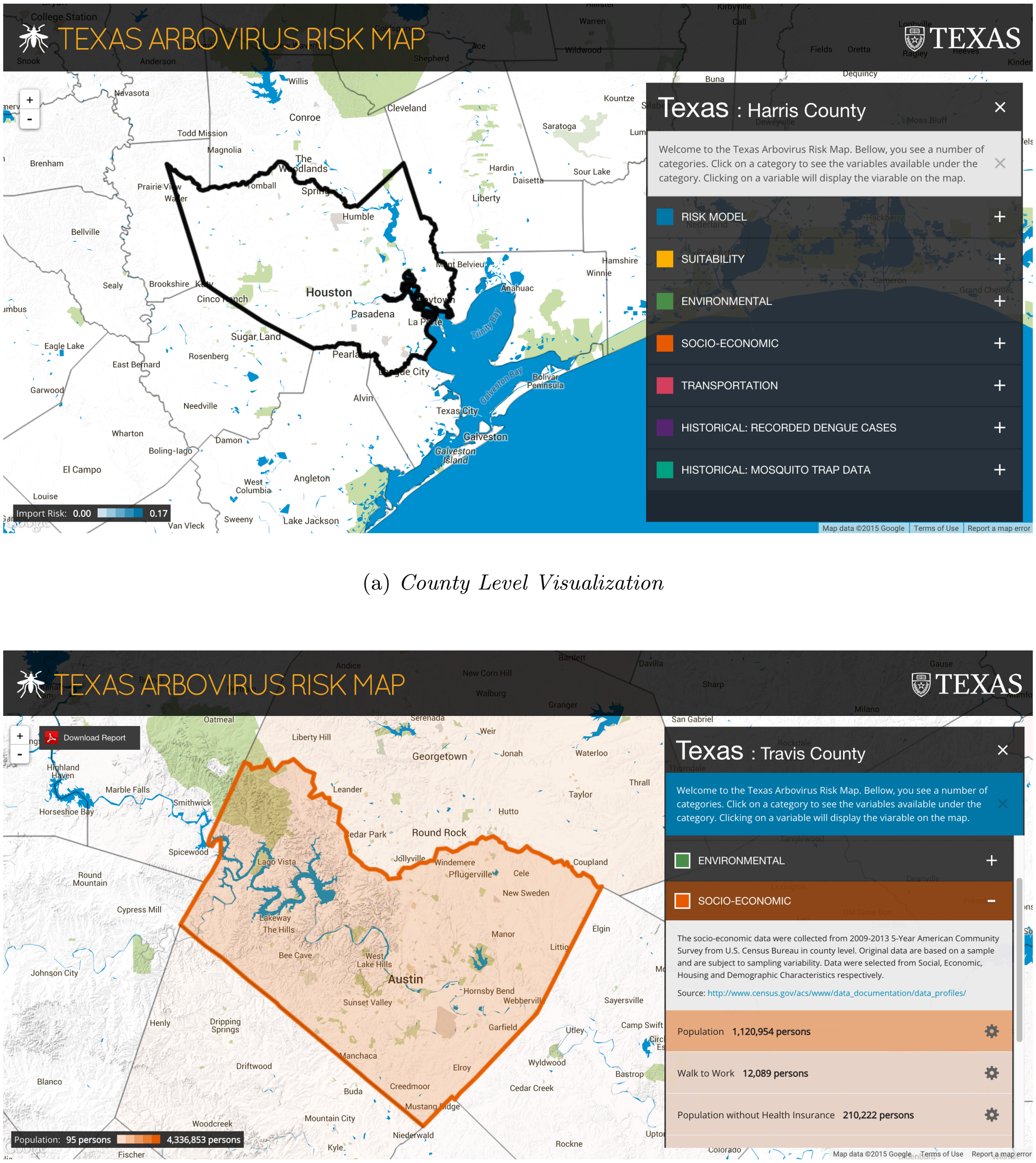
Date Range: 2009 - 2012

Local Cases (2002 - 2012) 4 persons

Total Cases (2002 - 2012) 160 persons

(b) *Dengue Cases from 2009 to 2012*

Figure 12: The Texas Arbovirus Risk Map can visualize several types of data. In addition to county-level data, the map can visualize high resolution data as in Figure 12a. The website can also be used to visualize time-sensitive incidence data through a time-selection slider as in Figure 12b.



Texas : Harris County

Welcome to the Texas Arbovirus Risk Map. Bellow, you see a number of categories. Click on a category to see the variables available under the category. Clicking on a variable will display the viarable on the map.

- 🟦 RISK MODEL +
- 🟨 SUITABILITY +
- 🟩 ENVIRONMENTAL +
- 🟧 SOCIO-ECONOMIC +
- 🟥 TRANSPORTATION +
- 🟪 HISTORICAL: RECORDED DENGUE CASES +
- 🟩 HISTORICAL: MOSQUITO TRAP DATA +

(a) *County Level Visualization*

Texas : Travis County

Welcome to the Texas Arbovirus Risk Map. Bellow, you see a number of categories. Click on a category to see the variables available under the category. Clicking on a variable will display the viarable on the map.

- 🟩 ENVIRONMENTAL +
- 🟧 SOCIO-ECONOMIC −

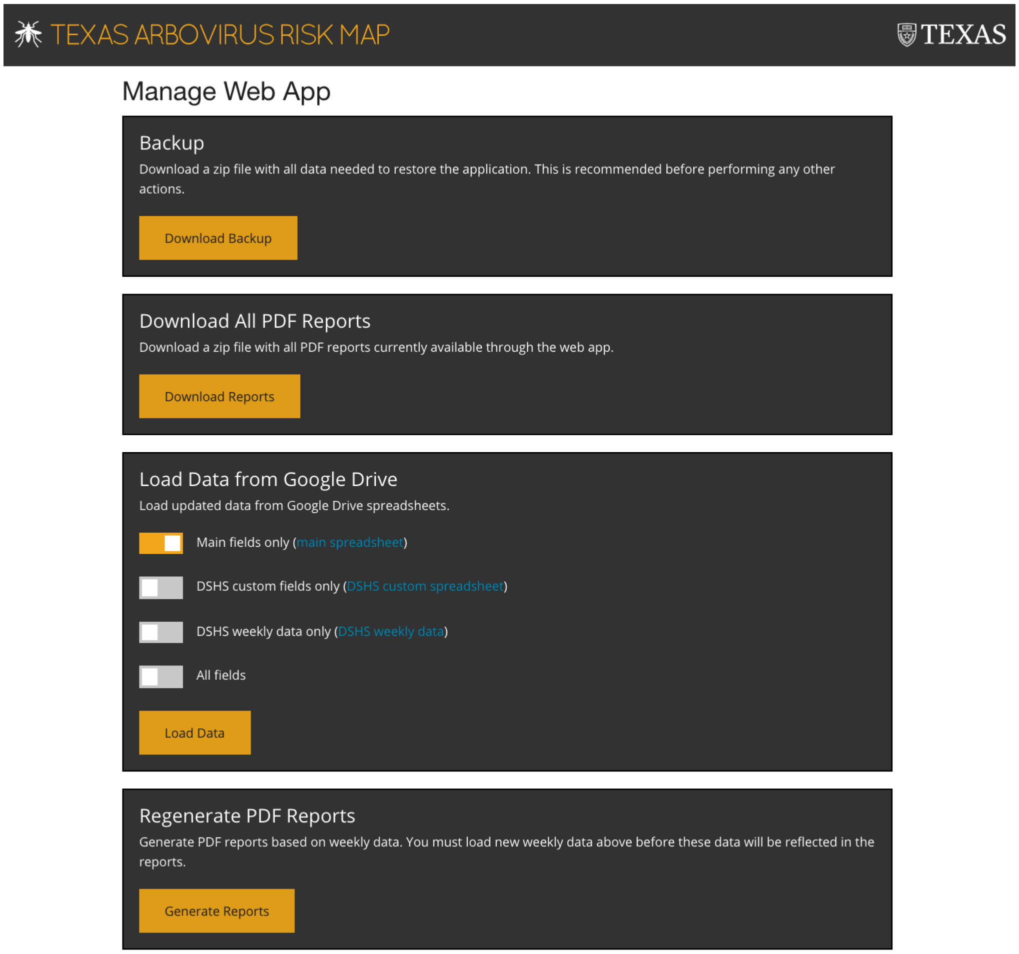
The socio-economic data were collected from 2009-2013 5-Year American Community Survey from U.S. Census Bureau in county level. Original data are based on a sample and are subject to sampling variability. Data were selected from Social, Economic, Housing and Demographic Characteristics respectively.

Source: http://www.census.gov/acs/www/data_documentation/data_profiles/

Population	1,120,954 persons
Walk to Work	12,089 persons
Population without Health Insurance	210,222 persons

(b) *Travis County Population*

Figure 13: A user can select a geographic area and variable of interest. The geographic area of interest can be either the entire state, an HSR, or a county. The user can either click on a county, or type a county name or HSR name in the search box to zoom to that region. Using the categories on the right, the user can select a variable of interest. A legend on the bottom left connects the visualization color to the variable value.



Manage Web App

Backup

Download a zip file with all data needed to restore the application. This is recommended before performing any other actions.

Download All PDF Reports

Download a zip file with all PDF reports currently available through the web app.

Load Data from Google Drive

Load updated data from Google Drive spreadsheets.

- ☑ Main fields only (main spreadsheet)
- □ DSHS custom fields only (DSHS custom spreadsheet)
- □ DSHS weekly data only (DSHS weekly data)
- □ All fields

Regenerate PDF Reports

Generate PDF reports based on weekly data. You must load new weekly data above before these data will be reflected in the reports.

Figure 14: DSHS administrative interface or the Texas Arbovirus Risk Map. Through this interface, DSHS can upload the latest arbovirus surveillance data, download auto-generated reports, update DSHS custom fields, and backup all website data.

5 Conclusions and Discussions

Through a careful assessment of the ecological and socioeconomic factors that influence the feasibility of transmission of DENV and CHKIV, we have highlighted areas of Texas that are at relatively higher risk.

For CHIKV and DENV, our models predict higher suitable habitat and a higher relative probability for the imported cases in urban areas in the eastern half of Texas and along the gulf coast. These results are consistent with known distributions of both vector species and where historical imported cases have occurred. The species distributions are an improvement on past mapping efforts because they provide higher resolution on areas that are more highly suitable. In general the pattern of relative occurrence rate results are consistent with the ecology of both vector species in that the higher rates are concentrated in the urban areas. The majority of these major urban areas are in the eastern half of the state, but *Ae. aegypti* has high levels of relative occurrence also in the western corner of El Paso.

In constructing the importation model, originally, there were millions of models to choose.

We developed an innovative way to eliminate duplicate variables. Half of the variables could be well represented by the rest. We use backward selection to effectively select the most significant variables. The resulting model, based on just 10 variables, performs significantly better than competing models. The population with Bachelor's degree was the most important variable and had a correlation coefficient of 0.83 with the relative probability for the next DENV cases to occur. An explanation was that the people with Bachelor's degree are more likely to have a job, such as consulting, which requires more international travel.

The import risk model indicates Harris county, Travis county and Cameron county as the counties in highest risks for the next dengue importation case to occur, with a relative probability of at least as 20 times that of most other counties. Dallas county is one level lower but still in a much higher risk range than other counties. If we draw the risk map with the logarithm of relative probability for next DENV case to occur, counties in southeast areas suffered a higher risk than the counties in northwest areas.

The sustained transmission risk map offers a useful insight: it is not necessarily the counties that have the highest relative abundance for *Ae.* that are at risk for establishing these arboviruses, but those counties that have a lower human population. Counties will smaller urban areas may still provide adequate habitat opportunities for *Ae. spp.* and with fewer persons available to bite. Such counties have a higher probability that an infected individual would be bitten and spread the infection. Both geographic risk distribution maps demonstrate that there is a legitimate extent of risk of DENV and CHIKV in Texas, but that across and within counties that are great differences in the risk.

Maximum entropy was an appropriate method to use in this analysis because of the scarcity of historical data. However there are limitations with maximum entropy method and the MaxEnt software that should be acknowledged when interpreting the results:

- In constructing the suitability maps, two caveats are the lack of data points within Texas and the uncertainty associated with the sampling effort across the states. Although MaxEnt has reportedly had good performance with interpolating across geographic region, more Texas specific data points would allow for the model to be fit just for Texas.

- Without knowing the effort level put into surveying other areas of the state it is not possible to know for certain where these species are currently absent.

- The modeling method produces a static output. With the available data, the suitability models cannot be more temporally resolute. Mosquito abundances and therefore arboviruses have a seasonal component in Texas, historically seen with West Nile Activity. If time-stamped data points were available it would be possible to generate seasonal maps. This could be especially useful if different areas of the state experienced different peak seasons in mosquito abundances.

- Maximum entropy outputs are as close as possible to a uniform distribution. The results then suggest that each county has some level of occurrence. When the model outputs a level of occurrence close to zero, the reality may be that vectors are not at all present in that geographic region.

- The analysis does not consider vector population dynamics in modeling relative occurrence rates. In more temporally resolute models that predict abundance and the course of an outbreak, vector population dynamics can be important in the increase and decrease in number of cases.

- Even though we included the travel data into our model, in contrast to previous studies, the availability of a more detailed country-to-Texas-county travel data could facilitate our ability to produce a more accurate import risk model.

- Availability of the data resources, quality of the data, and feasibility of the data acquisition always plays an important role in the accuracy and prediction ability of the risk maps. Absence of the CHIV data also brought limitations for the import risk maps. We use probability of DENV import, based on historical DENV data, as a proxy for CHIKV import in the absence of the CHIKV data. The practicability of cross disease risk representation was based on the same transmission vectors for both diseases. However, the import arbovirus disease was brought by the travelers from arbovirus endemic countries and areas. And the DENV endemic countries are not exactly the same as the CHIKV endemic countries and areas. This is especially true with the recent high levels of CHIKV activity in Central America and the Caribbean.

6 Future Work

We put forth several contributions in this paper. First, we include the travel data in risk model construction. Second, we construct an innovative way of eliminating the replicated variables in model construction. Third, we use the maximum likelihood method in model valuation and selection instead of the original methods used in MaxEnt software. The vectors suitability model, import risk model, and sustained transmission risk map offer useful insights that the relative abundance for *Ae.* doesn't contribute much in the relative probability for the importation and sustained transmission risks compare to some socio-economic and demographic factors. Based on the results of the risk maps, we have potential avenues of future work to improve mosquito surveillance in counties across Texas:

- In order to test the vector species distribution hypothesis put forward by the maps and to gain a quantitative absolute risk assessment, time series data on vector abundances from traps need to be collected. Time-series would not only confirm the relative occurrence rates generated from the MaxEnt output, it would also allow real-time modeling of actual abundances and more temporal fluctuations in population sizes. Similar forecasting of *Ae.* mosquito populations has been investigated in Florida. Validation and forecasting in this form may best be first approached on a county by county basis, as to complete this evaluation on a state-wide level would require a highly coordinated data intensive effort and mosquito surveillance is presently set up by county.

- In addition to collecting detailed trap data for individual counties, future work could include optimizing the locations of the traps in order to improve early detection and situational awareness of arbovirus outbreaks. Based on an evaluation of how well the current trap locations contribute to timeliness, accuracy, and spatio-temporal representativeness of the arbovirus surveillance system, maximally informative locations and monitoring schedule could be determined. Optimization processes could be completed for surveillance of other important arboviruses such as West Nile Virus in addition to DENV and CHIKV. Key counties in Texas, including Harris and Travis Counties have approached the UT team for further discussion of optimization possibilities.

- A more detailed county-to-county travel data collection is considered as an important future work to reflect the actual population mobility to improve the prediction ability and early warning ability of the importation and sustained transmission models.

- Lastly, future work could include developing methods for integrating the two types of risk, into a "spark and fire" risk model. This type of risk would evaluate the probability that there will be an imported case and that the imported case will lead to secondary cases. This risk quantification takes into account both stages of transmission, as opposed to the sustained transmission model which considers the possibility of secondary cases, given that there is an imported case.

Acknowledgments

Many people helped by providing useful discussions, guidance, and data to produce the results presented here. The authors would like to thank the DSHS officers who helped facilitate this project especially Bruce Clements, Nicole Evert, David Florin, Laura Robinson, Tom Sidwa, Bethany Bolling, Martha Thompson and the others who have helped guide the project. The authors would also like to thank Samuel V. Scarpino, Michael Johansson, Sahotra Sarkar, Timothy Segura, Danny Soto, and Christopher Vitek for data, help, and guidance throughout the project.

References

Sair Arboleda, Nicolas Jaramillo-O, and A Townsend Peterson. Mapping environmental dimensions of dengue fever transmission risk in the Aburrá Valley, Colombia. *International journal of environmental research and public health*, 6(12):3040–55, December 2009. ISSN 1660-4601. doi: 10.3390/ijerph6123040. URL http://www.mdpi.com/1660-4601/6/12/3040/htm.

Carolyn Elizabeth Barney. *Dengue Risk Factor Distribution in Harris County, Texas*. PhD thesis, The University of Texas School of Public Health, 2008.

Frida Cano, Jaycob Gorski, and Jessica Eastridge. Mosquito Surveillance in the Brazos County (Diptera : Culicidae), 2015.

Antonio de la Mora-Covarrubias, Florinda Jiménez-Vega, and Sandra Maritza Treviño Aguilar. Distribución geoespacial y detección del virus del dengue en mosquitos Aedes (Stegomyia) aegypti de Ciudad Juárez, Chihuahua, México. *Salud Publica de Mexico*, 52(2):127–133, 2010. ISSN 00363634. doi: 10.1590/S0036-36342010000200004.

Mathieu Dubrulle, Laurence Mousson, Sara Moutailler, Marie Vazeille, and Anna-Bella Failloux. Chikungunya virus and Aedes mosquitoes: saliva is infectious as soon as two days after oral infection. *PloS one*, 4(6):e5895, January 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0005895. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0005895.

Jane Elith, Steven J. Phillips, Trevor Hastie, Miroslav Dudík, Yung En Chee, and Colin J. Yates. A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, 17(1):43–57, January 2011. ISSN 13669516. doi: 10.1111/j.1472-4642.2010.00725.x. URL http://doi.wiley.com/10.1111/j.1472-4642.2010.00725.x.

Marc Fischer, J Erin Staples, et al. Notes from the field: chikungunya virus spreads in the americas-caribbean and south america, 2013-2014. *MMWR Morb Mortal Wkly Rep*, 63(22):500–501, 2014.

David Florin and Laura E. Robinson. Chikungunya. pdf document, private communication, 2015.

Norma Gorrochotegui-Escalante, Consuelo Gomez-Machorro, Saul Lozano-Fuentes, Ildefonso Fernandez-Salas, Maria De Lourdes Munoz, Jose a. Farfan-Ale, Julian Garcia-Rejon, Barry J. Beaty, and William C. Black IV. Breeding structure of aedes aegypti populations in Mexico varies by region. *American Journal of Tropical Medicine and Hygiene*, 66(2):213–222, 2002. ISSN 00029637.

Laura C Harrington, John D Edman, and Thomas W Scott. Why do female aedes aegypti (diptera: Culicidae) feed preferentially and frequently on human blood? *Journal of Medical Entomology*, 38(3):411–422, 2001.

Sergio Ibáñez Bernal and Carmen Martínez-Campos. Aedes albopictus in Mexico. *Journal of the American Mosquito Control Association*, 10(2):231–232, 1994.

M.M. Johnsen. Mosquitoes of texas. http://agrilife.org/aes/public-health-vector-and-mosquito-control/mosquitoes-of-texas/.

Michael David Kavanaugh. *Influence of stormwater drainage facilities on mosquito*. PhD thesis, University of North Texas, 2008.

Matt J Keeling and Pejman Rohani. *Modeling infectious diseases in humans and animals*. Princeton University Press, 2008.

Valérie R. Louis. Modeling tools for dengue risk mapping-a systematic review. *International journal of health geographics*, 2014. 13.1 (2014): 50.

Elia Axinia Machado-Machado. Empirical mapping of suitability to dengue fever in Mexico using species distribution modeling. *Applied Geography*, 33:82–93, April 2012. ISSN 01436228. doi: 10.1016/j.apgeog.2011.06.011. URL http://www.sciencedirect.com/science/article/pii/S0143622811001275.

Carlos F. Marina, J. Guillermo Bond, Mauricio Casas, José Muñoz, Arnoldo Orozco, Javier Valle, and Trevor Williams. Spinosad as an effective larvicide for control of Aedes albopictus and Aedes aegypti, vectors of dengue in southern Mexico. *Pest Management Science*, 67(1):114–121, 2011. ISSN 1526498X. doi: 10.1002/ps.2043.

Lee P. McPhatter, Farida Mahmood, and Mustapha Debboun. Survey of Mosquito Fauna in San Antonio, Texas. *Journal of the American Mosquito Control Association*, 28(3):240–247, September 2012. ISSN 8756-971X. doi: 10.2987/12-6230R.1. URL http://dx.doi.org/10.2987/12-6230R.1.

Samuel A. Merrill, Frank B. Ramberg, and Henry H. Hagedorn. Phylogeography and population structure of Aedes aegypti in Arizona. *Am J Trop Med Hyg*, 72(3):304–310, March 2005. URL http://www.ajtmh.org/content/72/3/304.full.

Don W Micks and William B Moon. Aedes aegypti in a texas coastal county as an index of dengue fever receptivity and control. *The American journal of tropical medicine and hygiene*, 29(6):1382–1388, 1980.

Alexander Moffett, Stavana Strutz, Nelson Guda, Camila González, Maria Cristina Ferro, Víctor Sánchez-Cordero, Sahotra Sarkar, et al. A global public database of disease vector and reservoir distributions, 2009.

Wilder-Smith A Murray NEA, Quam MB. Epidemiology of dengue: past, present and future prospects. *Clinical Epidemiology*, 2013.

Hector Orta Pesina, Roberto Mercado Hernandez, and Martha a. Valdez Rodriguez. Aedes albopictus in Allende City, Nuevo Leon, Mexico. *Journal of the American Mosquito Control Association*, 17(4):260–261, 2001.

U.S. Army Public Health Command Region North PHCR-West. Vectormap data portal, 2002. URL http://www.vectormap.org. Accessed on 11 March, 2015.

Steven J. Phillips, Robert P. Anderson, and Robert E. Schapire. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3-4):231–259, January 2006. ISSN 03043800. doi: 10.1016/j.ecolmodel.2005.03.026. URL http://www.sciencedirect.com/science/article/pii/S030438000500267X.

N.R. Powers, K. Cox, R. Romero, and M.A. DiMenna. The Reintroduction and Possible Establishment of Aedes. *Journal of the American Mosquito Control Association*, 22(4):756–757, 2006.

Filiberto Reyes-Villanueva, Javier A. Garza-Hernandez, Alberto M. Garcia-Munguia, Annabel F.V. Howard, Aldo I. Ortega-Morales, Monsuru A. Adeleke, and Mario A. Rodriguez-Perez. Aedes albopictus in northeast Mexico: An update on adult distribution and first report of parasitism by Ascogregarina taiwanensis, 2013. URL http://www.researchgate.net/profile/Javier_Garza-Hernandez/publication/258501379_Aedes_albopictus_in_northeast_Mexico_An_update_on_adult_distribution_and_first_report_of_parasitism_by_Ascogregarina_taiwanensis/links/00b4952956e266a7d6000000.pdf.

Olga S. Sanchez-Rodríguez, Rosa M. Sanchez-Casas, Maricela Laguna-Aguilar, Marcela S. Alvarado-Moreno, Ewry A. Zarate-Nahon, Rocio Gamirez-Jimenez, Carlos E. Mdeina de la Garza, Raul Torres-Zapata, Marco Dominguez-Galera, Pedro Mis-Avila, and Ildefonso Fernandez-Salas. Natural transmission of dengue virus by Aedes albopictus at Monterrey, Northeaster Mexico. *Southwestern Entomologist*, 39(3):459–468, 2014.

Sahotra Sarkar, Stavana E Strutz, David M Frank, Chissa-Louise Rivaldi, Blake Sissel, and Victor Sánchez-Cordero. Chagas disease risk in Texas. *PLoS neglected tropical diseases*, 4(10):e836, January 2010. ISSN 1935-2735. doi: 10.1371/journal.pntd.0000836. URL http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0000836.

Tim Segura. private communication, May 1,2015. Vector Control Coordinator, Lubbock.

Danny Soto. private communication, May 4, 2015. Environmental Services Department, El Paso.

D. Sprenger and T Wuithiranyagool. The discovery and distribution of Aedes albopictus in Harris County, Texas. *Journal of the American Mosquito Control Association*, 2(2):217–219, 1986.

Michelle M Thiboutot, Senthil Kannan, Omkar U Kawalekar, Devon J Shedlock, Amir S Khan, Gopalsamy Sarangan, Padma Srikanth, David B Weiner, and Karuppiah Muthumani. Chikungunya: a potentially emerging epidemic? *PLoS neglected tropical diseases*, 4(4):e623, January 2010. ISSN 1935-2735. doi: 10.1371/journal.pntd.0000623. URL http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0000623.

Christopher J Vitek, Joann A Gutierrez, and Frank J Dirrigl. Dengue vectors, human activity, and dengue virus transmission potential in the lower rio grande valley, texas, united states. *Journal of medical entomology*, 51(5):1019–1028, 2014.

Dan L. Warren and Stephanie N. Seifert. Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. *Ecological Applications*, 21(2): 335–342, March 2011. ISSN 1051-0761. doi: 10.1890/10-1171.1. URL http://www.esajournals.org/doi/abs/10.1890/10-1171.1.

Stephanie L Y N White. *No Title*. PhD thesis, Texas A&M University, 2008.

Bloom David Constable Heather Fang Janet-Koo Michelle Spencer Carol Yamamoto Kristina Wieczorek, John. Georeferencing quick reference guide. pdf document, 2012. Retrieved Feb 25, 2015.