# Aedes Arbovirus Risk in Texas

Xi Chen[a], Nedialko B. Dimitrov[b]

a: Ph.D, carol.chen@utexas.edu; b: Faculty, ned@austin.utexas.edu; OR/IE Group, Department of Mechanical Engineering, The University of Texas at Austin

## INTRODUCTION

- Arboviruses are a major public health concern in Texas, with endemic and non-endemic diseases collectively causing hundreds of cases in humans and domestic animals each year
- There are three mosquito-borne viruses spreading widely in tropical and subtropical regions of the Americas---dengue virus (DENV), chikungunya virus (CHIKV), and Zika virus (ZIKV)---that have the potential to emerge in Texas but have not yet established local transmission.
- Risk maps are visual tools that can help decision-makers to identify geographic areas of high or low risk for disease activity. These maps are often based on statistical models that have been fit to historical data reflecting risk factors.
- Here, we present a framework for mapping the risks of arbovirus introductions and transmission, to support surveillance efforts for DENV, CHIKV, and ZIKV in Texas.
- To build a predictive model for county-level import risk, we combined a maximum entropy method with a novel model selection procedure to systematically identify the ten most informative predictor variables from among 76 candidates.

## METHODOLOGY

### Maximum Entropy Method

$$\max_{p_i} \; -\sum_{i=1}^{n} p_i \log p_i$$

$$\text{s.t.} \quad \sum_{i=1}^{n} p_i f_j(x_i) = E(f_j(X)) \quad \forall j$$

$$\sum_{i=1}^{n} p_i = 1$$

$$p_i \geq 0 \quad \forall i$$

| Symbol | Definition |
|---|---|
| $x_i$ | representing the counties of Texas (i.e. $x_1$ represents the county, Dallas) |
| $p_i$ | the probability for $x_i$ to have an imported DENV case |
| $f_j(X)$ | functions of socio-economic, travel, mosquito and environmental variables |

### Representative Variable Selection

$$\min_{x_{ij}, y_j} \; \sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij} x_{ij}$$

$$\text{s.t.} \quad \sum_{j=1}^{n} x_{ij} = 1 \quad \forall i$$

$$\sum_{j=1}^{n} y_j = k$$

$$x_{ij} \leq y_j \quad \forall i, j$$

$$x_{ij} \in \{0,1\} \quad \forall i, j$$

$$y_j \in \{0,1\} \quad \forall j$$

| Symbol | Definition |
|---|---|
| $f_j$ | 76 variables represented by vectors $f_j, j = 1, 2, ...76$ |
| $d_{ij}$ | distance between two variables, measured as $d_{ij} = \| \frac{f_i}{\|f_i\|_2} - \frac{f_j}{\|f_j\|_2} \|_\infty$ |
| $x_{ij}$ | $x_{ij} = 1$ if vector i is represented by vector j; $x_{ij} = 0$, otherwise; |
| $y_j$ | $y_j = 1$, if vector j is selected as representative; $y_j = 0$, otherwise |

## METHODOLOGY

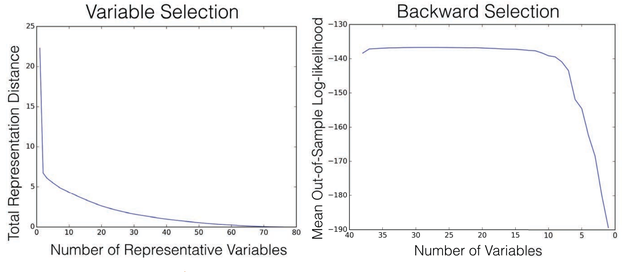
Our full data set contained 76 variables. First, we removed duplicate variables through representative variable selection. Then, we selected the 10 most predictive variable through backward selection.

### Predictive Variable Selection

| | Backward Selection |
|---|---|
| 1 | **function** BACKWARD SELECTION $(N)$ |
| 2 | Set $V = N$ |
| 3 | **While** $\|V\| > 1$ **do** |
| 4 | Set $e = argmax_{e \in V} C(S(V - e))$ |
| 5 | Set $V = V - \{e\}$ |
| 6 | Record $V$ and $C(S(V - e))$ |

| | |
|---|---|
| N | The complete set of variables |
| C | Return the out-of-sample log-likelihood, averaged over of seven randomly sampled cross validation folds |
| S | Fit a maximum entropy model given a set of variables $f_j$ |

### Results of Variable Selection

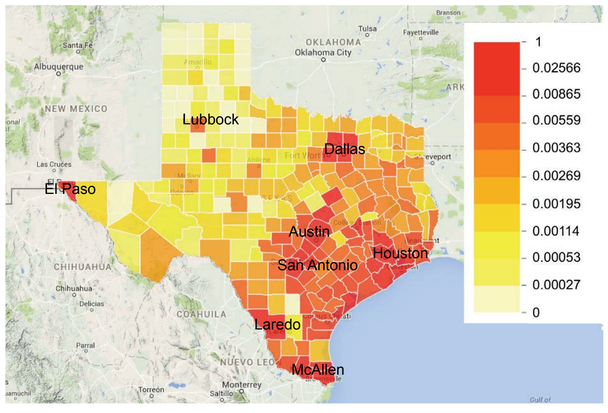



Methods for down selecting predictor variables from 76. Representative Variable Selection: Representation distance - a measure of how much information is dropped, with higher distance dropping more information - versus number of representative variables. Backwards Selection: Mean out-of-sample log-likelihood versus number of variables used in the model.

## RESULTS

### Relative Importation Risk

Relative risk of arbovirus importation for each Texas county, based on ten selected predictor variables. The six largest of Texas' 25 metropolitan areas are labeled, as well as Lubbock (11th) and Laredo (12th)




Number of reported DENV importations (2002-2012) and CHIKV importations (2014-2015) versus county population size for each of Texas' 254 counties.