Supplement: Real-time Zika risk assessment in the United States

Lauren A Castro BA [1*], Spencer J Fox BS[1*@], Xi Chen MS[2], Kai Liu MS[3], Steve Bellan PhD[4], Nedialko B Dimitrov PhD[2], Alison P Galvani PhD[5,6], Lauren Ancel Meyers[1,7],

1 Section of Integrative Biology, The University of Texas at Austin, Austin, USA
2 Graduate Program in Operations Research Industrial Engineering, The University of Texas at Austin, Austin, TX, USA
3 Institute for Cellular and Molecular Biology, The University of Texas at Austin, Austin,TX, USA
4 Center for Computational Biology and Bioinformatics, The University of Texas at Austin, Austin, TX, USA
5 Center for Infectious Disease Modeling and Analysis, Yale School of Public Health, New Haven, CT, USA
6 Department of Ecology and Evolution, Yale University, New Haven, CT, USA
7 The Santa Fe Institute, Santa Fe, NM, USA

*contributed equally to this manuscript
@ corresponding author: spncrfx@gmail.com

1 Importation Risk Analysis

Maximum Entropy Here we provide an overview of the maximum entropy method used to estimate Texas importation risk. Suppose we have set $X = \{x_1, x_2, ..., x_n\}$ representing the counties of Texas (i.e. x_1 represents the county, Dallas). Let the probability for x_i to have an imported DENV, CHIK, and ZIKA case be π_i. We construct an estimate of this unknown probability distribution using the historical import data. Call the estimated probability for county x_i, p_i. The vector of p_i sums to one over all counties. The relative probabilities $p_1, p_2, ...p_n$ can be constrained with known mean, variance, or other moments of some known $f_j(X)$. The functions $f_j(X)$ can be functions of socio-economic, environmental, and travel variables in our case (Table 4). Mathematically, we want to:

$$\max_{p_i} \quad -\sum_{i=1}^{n} p_i \log p_i \tag{1a}$$

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

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

$$p_i \geq 0 \qquad \forall i \tag{1d}$$

When we use Shannon's measure of entropy as the objective (1a), the constraints (1d) are automatically satisfied. The right-hand-side of (1b), $E(f_j(X))$, is estimated by the weighted arithmetic mean of $f_j(x_1), f_j(x_2), ..., f_j(x_n)$ based on the n counties of Texas [1].

Representative Variable Selection In the first step, we removed duplicate variables—variables that essentially bring the same information to the model. We call this step *representative variable selection*. Selecting representative variables was independent of the DENV, CHIK, and ZIKA import data, and only deals with the information contained in the variables themselves.

Selecting the representative variables was done with a variation of the facility location problem [2]. The goal was to select k variables to represent the entire variable set. k selected factors would represent themselves and the remained $72-k$ variables would be represented by exactly one variable from k selected variables. The $\ell-\infty$ norm of the difference between two unit-norm variables, denoted by f_i, f_j in Table 1, was assigned as the distance between the two variables. This distance measure was derived from the maximum difference in expectations that the two variables can produce, under any probability distribution. The facility location model allowed us to select the k variables that best represent others as represented by (2c). The objective function 2a for selecting representative variables was to minimize the distance between the k representative variables and all the variables in the entire variable set. Each variable was represented by exactly one of the k representatives, as represented by 2b.

$$\min_{x_{ij}, y_j} \quad \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} x_{ij} \tag{2a}$$

$$\text{s.t.} \quad \sum_{j=1}^{n} x_{ij} = 1 \qquad \forall i \tag{2b}$$

$$\sum_{j=1}^{n} y_j = k \tag{2c}$$

$$x_{ij} \leq y_j \quad \forall i, j \tag{2d}$$

$$x_{ij} \in \{0, 1\} \quad \forall i, j \tag{2e}$$

$$y_j \in \{0, 1\} \qquad \forall j \tag{2f}$$

Symbol	Definition
f_j	72 variables represented by vectors $f_j, j = 1, 2, ...72$
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 vector; $y_j = 0$, otherwise;

Table 1: **Parameters in representative variable selection method used to down select from 72 variables to 20.**

Predictive Variable Selection After selecting the k most representative variables, we chose the most predictive variables within k representative variables. One existing method of selecting predictive variables, once we have created a representative variable set, is to use hypothesis testing to choose between nested models [3]. We propose a different method, outlined in Table 2. Using a backward selection approach, in each iteration, the variable that contributed the least to model performance was dropped. Backward selection continued until all the variables were eliminated.

Model performance Model performance was measured base on out-of-sample data and cross validation was incorporated to strengthen the robustness of the model performance results. For each iteration, ten years DENV importation cases were divided in to two subsets: train data and test data. The model was fit using 7 years of train data and model performance was measured using 3 years of out-of-sample test data. To improve the robustness of the variable selection procedure and as cross-validation, we ran each set of variables on 6 randomly selected partitions of the 10 years of available data. From the 6 runs, we calculated the average of the out-of-sample log-likelihood of the model and eliminated the variable that resulting the largest mean out-of-sample log-likelihood with its elimination. A summary of the algorithm for Backward Selection is showed in Table 2.

Algorithm	Backward Selection
1	**function** BACKWARD SELECTION (N)
2	Set $V = N$
3	**While** $\lvert V \rvert > 1$ **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 representative 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

Table 2: **Algorithm for the backward variable selection of the 38 representative variables to 10 that was included in the final maximum entropy model**

.

Variables ordered by importance
Total Direct Spending(dollars)
Graduate or professional degree in Percentage
Local (dollars)
Male Population
Commuting to Work with Other Means
Max Temperature of Warmest Month
Population below Poverty Level in Percentage
Precipitation of Wettest Quarter
Population without Health Insurance
Graduate or professional degree population

Table 3: **Import risk model variables**. These 10 variables were selected from 72 variables using a combination of representative variables selection and backwards selection. The importance of each variable (from top to bottom) is determined by order of exclusion in backwards selection, with the most important variables remaining in the model the longest.

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
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)
Max Temperature of Warmest Month	Mean Travel Time to Work(Minutes)	
Precipitation of Wettest Quarter	Population Walk to Work	
Temperature Annual Range	Commuting to Work with Taxi	
Precipitation of Warmest Quarter	Percentage Commuting to Work with Public Transportation	
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 4: **Complete Set of Variables for Import Risk Map Modeling**

2 Transmission Risk Analysis

Estimating R_0 in Texas We estimated reproduction numbers (R_0) for Texas counties following the methodology in [4].

We estimate R_0 according to the Ross-Macdonald formulation, given by,

$$R_0 = \frac{mbc\alpha^2 e^{\mu n}}{\mu\gamma}, \tag{3}$$

where m,b, c, α, n, and μ denote the mosquito to human ratio, the mosquito-to-human transmission probability, the human-to-mosquito transmission probability, the mosquito biting rate, the extrinsic incubation period, and the average mosquito lifespan respectively (Table 2).

Of these, we assumed that n and μ varied with temperature. To calibrate our model for August temperatures, we collected average temperature estimates of each Texas county from a period of 1980 to 2010 [5]. The average temperature of Texas ranged from 24 to 31 °C. To estimate temperature-dependent extrinsic incubation periods, we used the log-normal distribution model estimated in [6] for DENV viruses in *Ae. aegypti*. Although μ does vary with temperature, a field mark-release-recapture experiment of *Ae. aegypti* in Puerto Rico estimated that adult longevity stays roughly the same over the range of temperatures that Texas may experience in August (for 50% of the population) and therefore we only used one estimate (14 days).

We used recent data on the susceptibility of Brazil populations of *Ae. aegypti* to the currently circulating Asian genotype of ZIKV to get an estimate of the human-to-mosquito transmission probability [7]. To estimate the mosquito-to-human transmission probability, we used estimates from published fitted parameters of a ZIKV β, which encompasses the mosquito-to-human transmission probability, from the 2013-2014 French Polynesia outbreak in [8] and our estimate of biting rate from [9] to derive an estimate of mosquito-to-human transmission probability. Finally, we also allowed m to vary among Texas counties. We used estimates of occurrence probabilities of *Ae. aegypti* for each Texas county obtained from a predicted global distribution of *Ae. aegypti* in [10] and estimated mosquito abundance assuming mosquito abundance follows a Poisson distribution [11]. We then multiplied mosquito abundances by a log linear function of the 2014 gross domestic product economic index for each Texas county extended from the fitted function derived in [4][12], as described in to incorporate economic effects on mosquito-human contacts. We present a sensitivity analysis of this function below.

We provide a sensitivity analysis of the function derived to estimate m used to relate GDP to decreases in mosquito-human contact ratios below.

Parameter	Description	Value	Reference
α	Mosquito biting rate: the expected number of bites per day.	0.63	[9]
n	Extrinsic incubation: The expected days between initial infection and infectiousness in *Ae. aegypti*	6-18	[6]
μ	Average lifespan of female *Ae. aegypti* mosquito (days)	14	[13]
b	Mosquito-to-human probability of transmission per bite	0.634	[8]
c	Human-to-mosquito probability of transmission per bite	0.77	[7]

Table 5: **Parameters for estimating ZIKV (R_0 in Texas counties**

.

Scenario	Function
Medium	$ln(MF) = -1.79 - .14 * ln(GDP)$
Weak	$ln(MF) = -2.6 - .14 * ln(GDP)$
Strong	$ln(MF) = -0.9 - .14 * ln(GDP)$
Mixed	$ln(MF) = -1.35 - 1.8 * ln(GDP)$

Table 6: **Sensitivity Analysis of R_0**

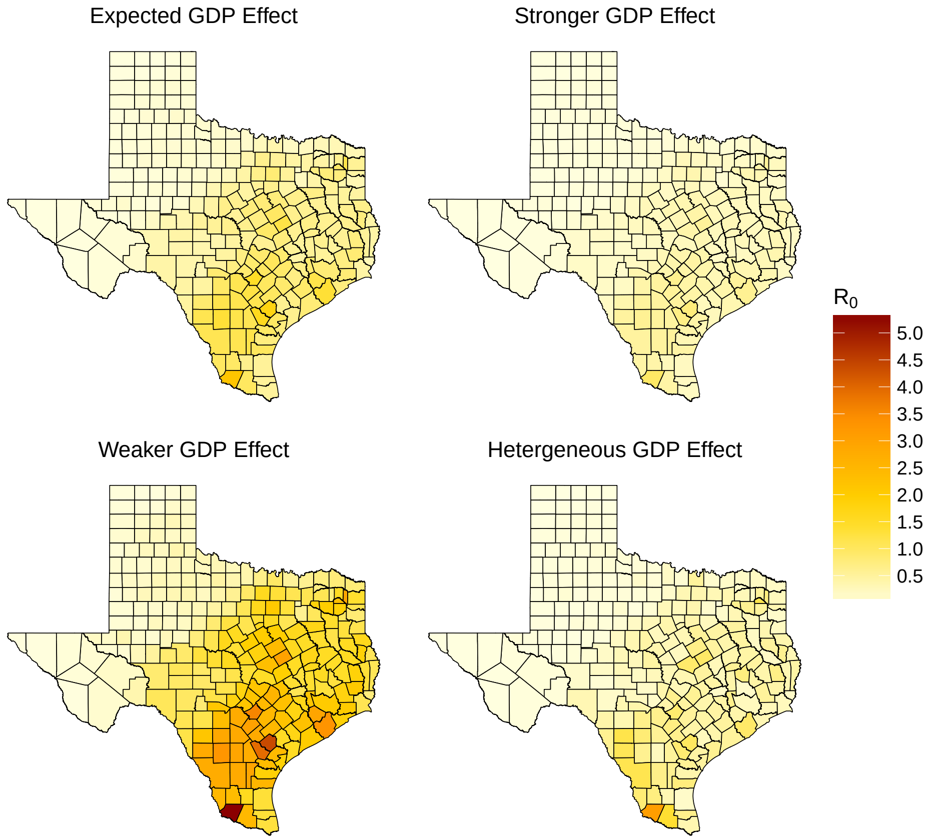

Figure 1: **Sensitivity Analysis of Estimated $\mathbf{R_0}$'s by the Effect of GDP on Mosquito-Human Contact.** We explore the uncertainty in our R_0's estimates resulting from the relationship between GDP and mosquito-human contact, which we estimated as an extension of the fitted function derived in [4]. In *Expected* we show the R_0 estimates used in the main analysis.*Stronger* shows estimated R_0's if we consider that the effect of GDP on mosquito-human contact is greater (reducing contact) than in *Expected*. This results in fewer counties having R_0's >1, or fewer counties can sustain ZIKV transmission. In *Weaker* we show estimated R_0's if the effect of GDP on the relationship is minimal, meaning mosquito-human contact levels are similar to ratios of mosquito abundance and population sizes in each county. Across the state, county GDP levels do not reduce the mosquito-human contact as strongly as in *Expected* and *Stronger*, resulting in higher R_0 estimates and more counties capable of sustaining ZIKV transmission. In this scenario, R_0 estimates are approximately two-fold higher than in our *Expected* estimates and the majority of Eastern and Southern Texas is at risk for sustained ZIKV transmission. The effect of increasing GDP is held constant in these first three panels. In *Heterogeneous* we estimate R_0's if increases in GDP have a greater effect on reducing mosquito-to-human contact than that in the first three scenarios. In this *Heterogeneous* case, counties with lower GDP would have higher levels of mosquito-human contact than in *Expected*, while counties with higher GDP would have lower mosquito-human contact levels than in *Expected*. This results in higher heterogeneity in R_0's overall, with more at risk counties having higher $R_0's$ than in *Expected* and less at risk counties having lower R_0's.

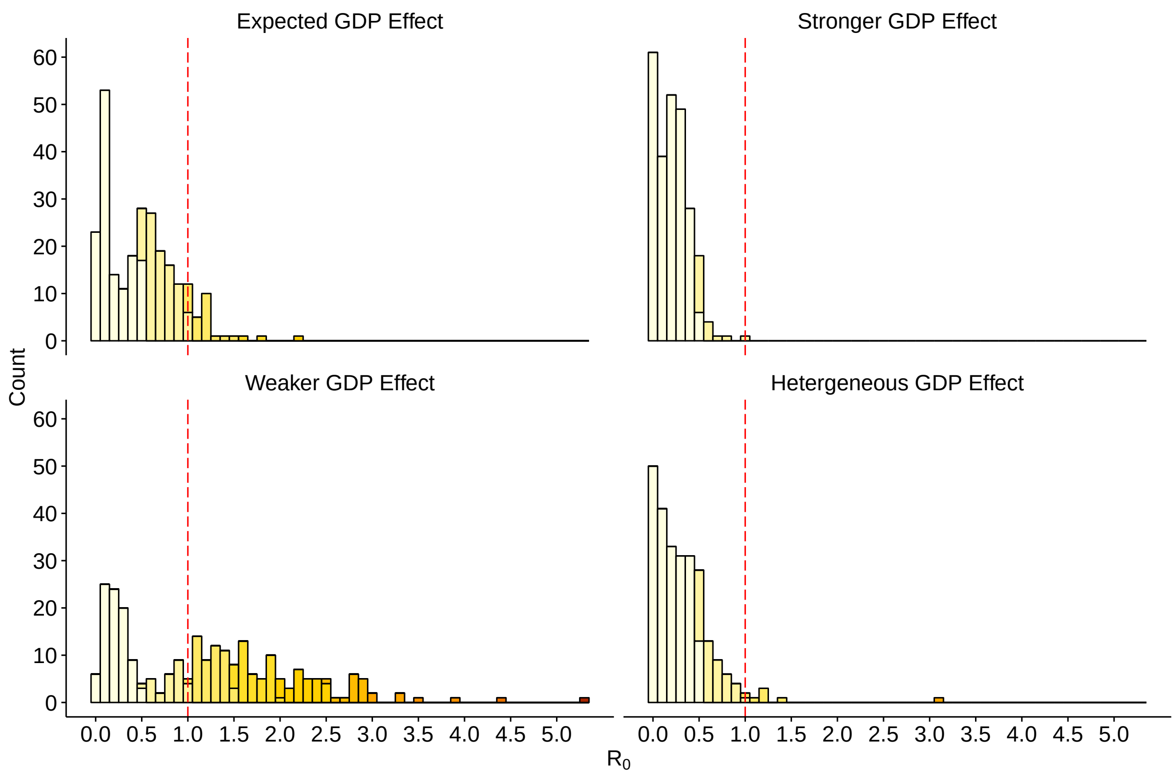

Figure 2: **Distributions of Estimated R_0's**. We show the distribution of estimated R_0's for each scenario from Fig.1. The spatial observations in Fig.1 are reflected in the number of counties above and below the threshold of $R_0 = 1$. In the *Expected* scenario, which we used as our expected R_0 estimates in the manuscript, there are 33 counties at high risk (above the threshold of $R_0 \geq 1$. As the effect of GDP on mitigating mosquito-human contact is increased, effectively reducing contact and risk of exposure, only one county remains at high risk for sustained transmission (upper right). On the other hand, as the effect of GDP on mosquito-human contact is reduced, R_0's are increased, with over 50% of the counties being at high risk for sustained transmission (lower left).

3 Supporting Figures and Tables

Model parameters We estimate some model parameters directly from epidemiological data and base others on published studies of ZIKV during the current and previous outbreaks (Table S7).

Parameter	Description	Range of Values (or median 95%)	Source
Transmission Rate (β)	The expected number of secondary infections per infectious person per day.	0.14-0.21	[10]
Infectious Period (γ)	The average length of the infectious period. Achieved with number of compartments, $n_{\text{Infectious}} = 3$, and daily recovery probability, 0.304.	9(3-22) days	[14]
Meta-Latent Period (α)	Average latent period before becoming infectious (see model assumptions). Achieved with number of compartments, $n_{\text{Incubation}} = 6$, and daily recovery probability, 0.584	10(6-17) days	[15],[14]
Reproduction Number (R_0)	The expected total number of secondary infections from one infectious individual ($\beta * \gamma$)	0.1-1.9	[10]
Serial Interval (SI)	The average length of time between consecutive exposures. $SI = \alpha + \frac{1}{2\gamma}$	15 (9.5-23.5) days	[15]
Reporting Rate (η)	The daily probability of an infectious individual being reported.	Daily: $1\% - 5\%$ Overall: $5\% - 40\%$	[16]
Importation Rate (μ)	The expected number of infectious ZIKV importations per day. (Statewide)	(0.3, 0.8, 4.5)	[17]

Table 7: **Branching Process Model Parameters**

.

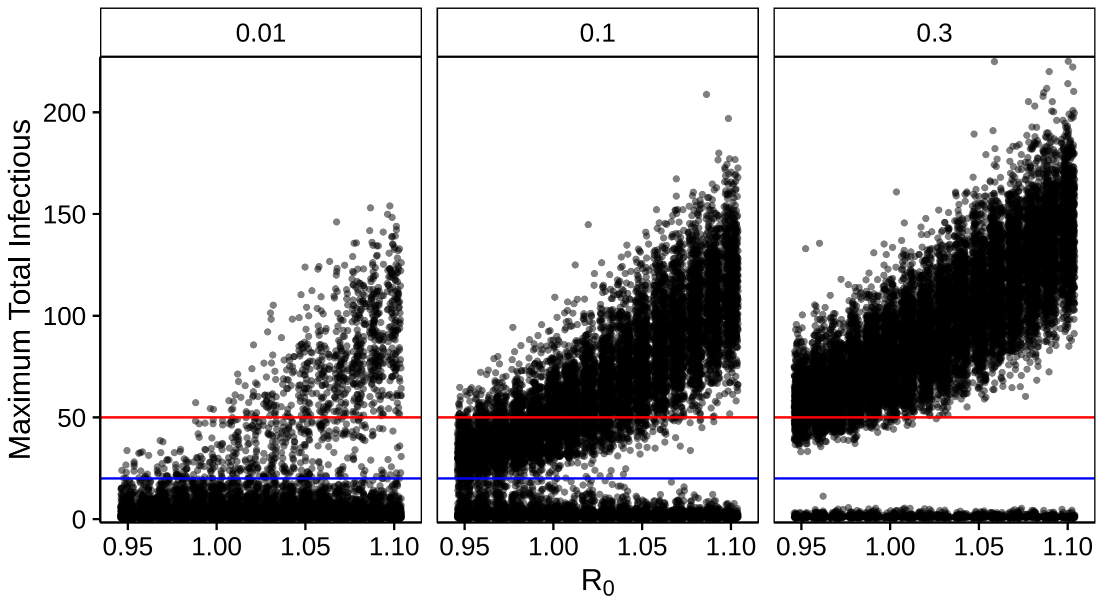

Figure 3: **Determination of threshold for surveillance triggers.** For each R_0 value we plot the maximum daily total infectious individuals for 1,000 of our 10,000 trials (black dots). Blue line indicate the prevalence threshold cutoff signifying extensive transmission determined to be 20. Red line indicates the epidemic threshold cutoff value (50), chosen to differentiate epidemics with $R_0 > 1$ from outbreaks with $R_0 < 1$. Panels differ by the daily importation rate for the simulations. Larger importation rates lead to larger maximum prevalences.

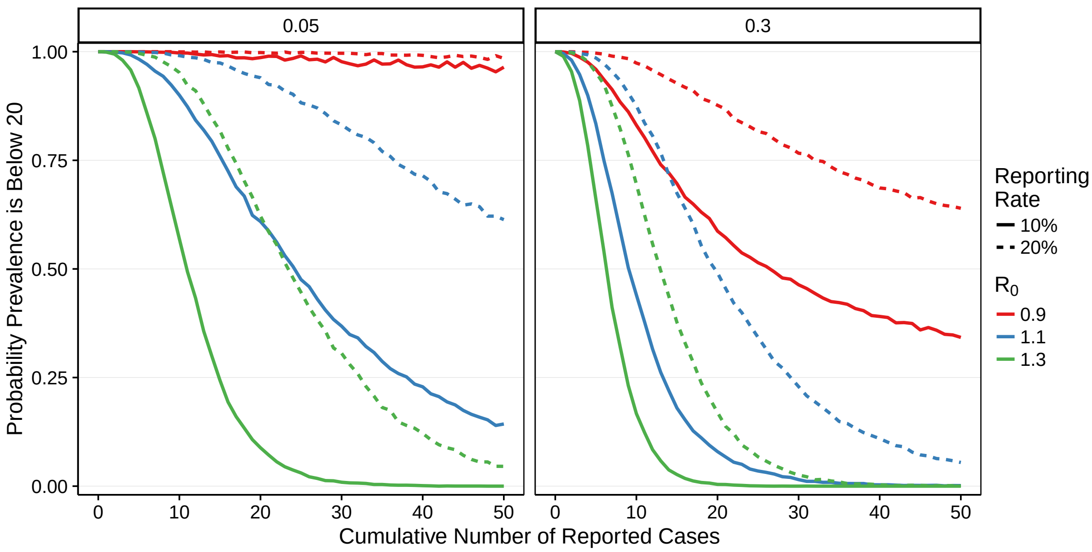

Figure 4: **Probability of exceeding prevalence threshold based on reported cases**. Lines indicate the probability that current cases for various R_0 values (colors) fall below a prevalence threshold, under low and high importation rates (panels). Line-type corresponds to either a high (20%, dashed) or low (10%, solid) reporting rate of ZIKV cases. Intuitively, the probability that current cases are below a threshold (e.g. 20 cases) for high R_0 and low reporting rate decreases rapidly, as fewer cases are reported while the outbreak is growing. When the importation rate is low, there is a high certainty that low R_0 outbreaks are below threshold concern. However, when there are high levels of importation, a low reporting rate can cause an outbreak with a low R_0 outbreak to be more of a concern than an high R_0 outbreak with a higher detection probability.

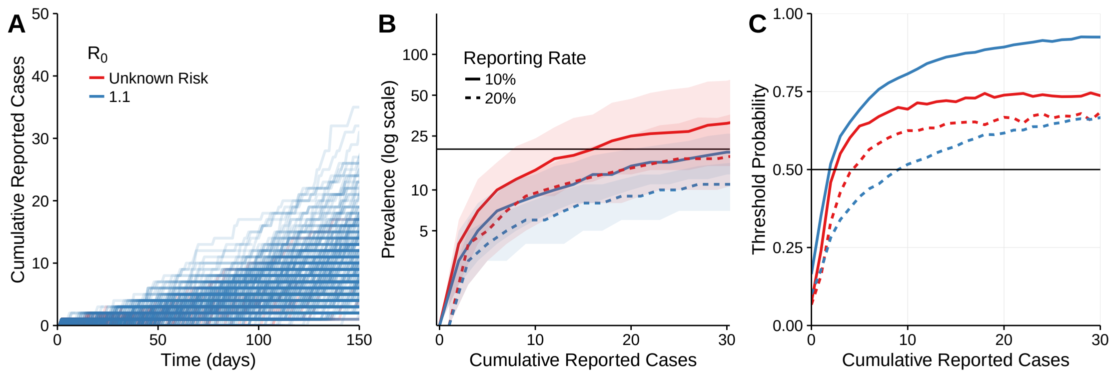

Figure 5: **Surveillance triggers for detecting and forecasting ZIKV transmission.** (A) Simulated outbreaks, assuming an importation rate of 0.1 case per day, for a known (moderate risk) R_0 (blue) or an unknown risk R_0 (red). 2,000 randomly sampled simulations are shown for each scenario. (B) Current prevalence as a function of the cumulative detected cases, assuming an importation rate of 0.1 case per day, for a known R_0 (blue) or an unknown risk R_0 (red), and a relatively high (dashed) or low (solid) reporting rate. Ribbons indicate 50% quantiles. (C) The increasing probability of imminent epidemic expansion across a range of reported cases, compared across the unknown risk (red) and known moderate risk (blue) for a low (solid) and high (dashed) reporting rate. Suppose cases arise in an unknown risk county and a policymaker wishes to trigger a response as soon as the chance of sustained transmission reaches 50% (horizontal line). Then, if the reporting rate is 20%, he or she should trigger the response as soon as the 4th case is reported.

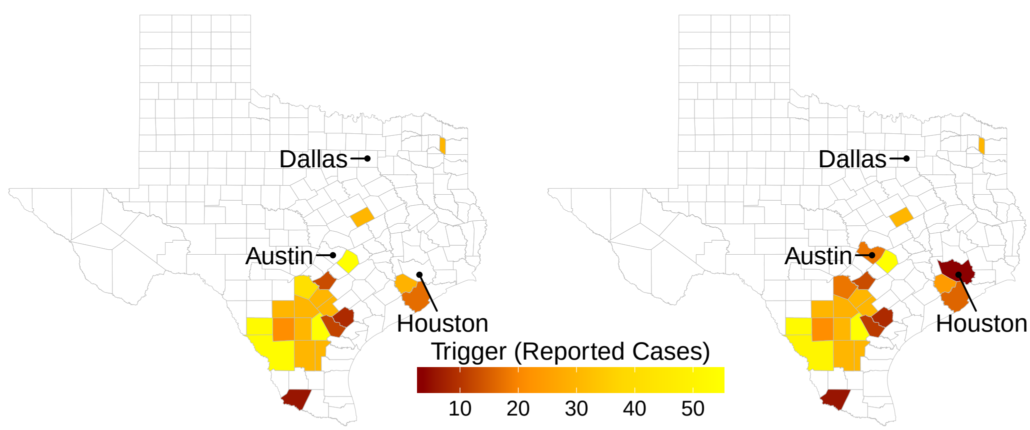

Figure 6: **ZIKV surveillance triggers across Texas.** Recommended county-level surveillance triggers for detecting that the probability of current prevalence has T= 20, with p_T=0.70, assuming a reporting rate of 20%. These reflect (A) the baseline importation scenario for August 2016 (81 cases statewide per 90 days) projected from historical arbovirus data, and (B) the elevated importation scenario (405 cases statewide per 90 days) that assumes recent ZIKV importations represent only 20% of all importations. White counties indicate that less than 1% of the 10,000 simulated outbreaks resulted in sustained transmission.

References

[1] Kapur JN, Kesavan HK. Entropy optimization principles with applications. Academic Pr; 1992.

[2] Wolsey LA. Integer programming. vol. 42. Wiley New York; 1998.

[3] Halvorsen R, Mazzoni S, Bryn A, Bakkestuen V. Opportunities for improved distribution modelling practice via a strict maximum likelihood interpretation of MaxEnt. Ecography. 2015;38(2):172–183.

[4] Perkins A, Siraj A, Warren Ruktanonchai C, Kraemer M, Tatem A. Model-based projections of Zika virus infections in childbearing women in the Americas. bioRxiv [Internet]. 2016 Feb [cited 2016 Apr 6];Available from: http://biorxiv.org/lookup/doi/10.1101/039610.

[5] USA.com: Texas[Internet]. World Media Group, LCC; 2016 [cited 2016 May 25]. Available from: http://www.usa.com/texas-state.htm.

[6] Chan M, Johansson MA. The incubation periods of dengue viruses. PloS one. 2012;7(11):e50972.

[7] Chouin-Carneiro T, Vega-Rua A, Vazeille M, Yebakima A, Girod R, Goindin D, et al. Differential Susceptibilities of Aedes aegypti and Aedes albopictus from the Americas to Zika Virus. PLoS Negl Trop Dis. 2016;10(3):e0004543.

[8] Kucharski AJ, Funk S, Eggo RM, Mallet HP, Edmunds WJ, Nilles EJ. Transmission dynamics of Zika virus in island populations: a modelling analysis of the 2013–14 French Polynesia outbreak. PLOS Negl Trop Dis. 2016;10(5):e0004726.

[9] Scott TW, Amerasinghe PH, Morrison AC, Lorenz LH, Clark GG, Strickman D, et al. Longitudinal studies of Aedes aegypti (Diptera: Culicidae) in Thailand and Puerto Rico: blood feeding frequency. Journal of medical entomology. 2000;37(1):89–101.

[10] Kraemer MUG, Sinka ME, Duda KA, Mylne A, Shearer FM, Barker CM, et al. The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus. eLife [Internet]. 2015 Jun [cited 2016 Apr 19];4:e08347. Available from: http://elifesciences.org/content/4/e08347v3.

[11] Wright DH. Correlations Between Incidence and Abundance are Expected by Chance. J Biogeogr [Internet]. 1991 [cited 2016 Apr 19];18(4):463–466. Available from: http://www.jstor.org/stable/2845487.

[12] U S Bureau of Economic Analysis. Texas Counties: Per Capita Income [Internet]; 2014 [cited 2016 Apr 19]. Available from: http://www.txcip.org/tac/census/morecountyinfo.php?MORE=1011.

[13] Brady OJ, Johansson MA, Guerra CA, Bhatt S, Golding N, Pigott DM, et al. Modelling adult Aedes aegypti and Aedes albopictus survival at different temperatures in laboratory and field settings. Parasites & Vectors. 2013;6(1):1–12. Available from: http://dx.doi.org/10.1186/1756-3305-6-351.

[14] Lessler J, Ott CT, Carcelen AC, Konikoff JM, Williamson J, Bi Q, et al. Times to Key Events in the Course of Zika Infection and their Implications for Surveillance: A Systematic Review and Pooled Analysis. bioRxiv [Internet]. 2016 Mar [cited 2016 Apr 26];Available from: http://biorxiv.org/content/early/2016/03/02/041913.abstract.

[15] Majumder MS, Cohn E, Fish D, Brownstein JS. Estimating a feasible serial interval range for Zika fever. Bulletin of the World Health Organization [Internet]. 2016 [cited 2016 May 10];Available from: http://www.who.int/bulletin/online_first/16-171009.pdf.

[16] Duffy MR, Chen TH, Hancock WT, Powers AM, Kool JL, Lanciotti RS, et al. Zika virus outbreak on Yap Island, Federated States of Micronesia. New Eng J Med [Internet]. 2009 Jun [cited 2016 Apr 4];360(24):2536–2543. Available from: http://www.nejm.org/doi/abs/10.1056/NEJMoa0805715.

[17] Texas Department of State Health and Human Services. Zika in Texas [Internet]; 2016 [cited 2016 Apr 19]. Available from: http://www.texaszika.org/.