Combining traditional and online-media information for forecasting emerging climate sensitive mosquito-borne diseases

L. Leticia Ramírez Ramírez, CIMAT, MEXICO

Yulia Gel, UTDALLAS, USA
Slava Lyubchich, UMCES, USA

Oct, 2017

What is Chikungunya?

Chikungunya (pronunciation: *c̃hik-en-gun-ye*) virus is transmitted to people by mosquitoes.

Symptoms The most common symptoms of chikungunya are fever and joint pain.

Progress The disease has the sudden onset of fever 2–3 days after exposure that last 2–7 days.

Treatment and Prevention Only to relieve sypmptoms.

Chikungunya Transmission

Chikungunya is transmitted via bites infected mosquito, mainly the *Aedes aegypti* and *Aedes albopictus* species.

These are the same mosquitoes that transmit **dengue** virus.

Chikungunya Outbreaks

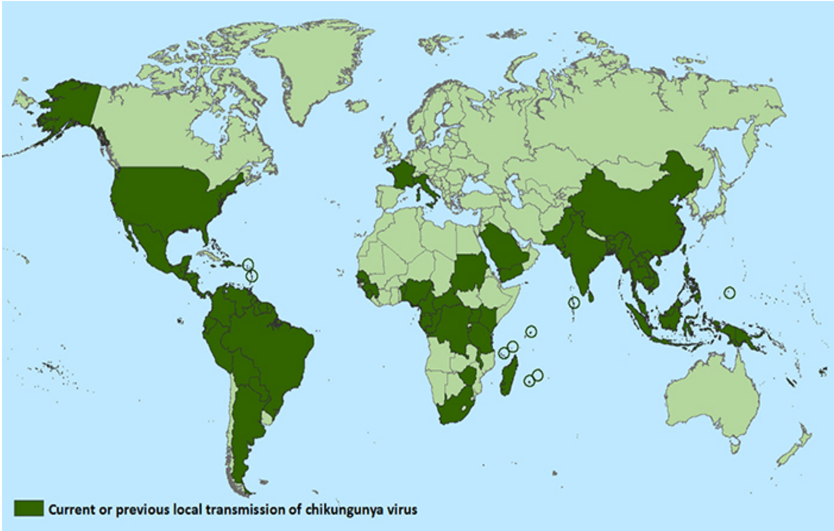
Chikungunya was first identified in 1952 in West Africa. Outbreaks have occurred later in 60 countries in Africa, Asia, Europe, and the Indian and Pacific Oceans.

In late 2013, chikungunya virus was found for the first time in the Americas on islands in the Caribbean.

Since 2013 and up to April 2015 in the Americas, there were 46 countries with confirmed cases and reported more than 1,300,000 confirmed and suspected cases (PAHO).

DARPA of USA launched a chikungunya forecasting challenge on August 15, 2014 with a goal to develop new data analytics tools for predicting chikungunya dynamics.

Chikungunya Geographics (as of April 22, 2016; CDC)



Data Challenges

The key problem is that

- Chikungunya has been recorded for the 1st in Americas only late 2013; some Central/Latin American countries report cases only from September 2014
- but **no prior data** in the region to mitigate and predict virus dynamics!

What can be done? Or the Big Data Opportunities for Global Infectious Disease Surveillance...

Search for Alternative Data

As an alternative, we can try to use other less traditional data sources such as Google Dengue Trend (GDT).

How does this work? Google has found that certain search terms are good indicators of dengue activity. Similarly to Google Flu Trend (GFT), GDT uses aggregated Google search data to estimate dengue activity around the world in near real-time.

According to El-Metwally (2015) and Gluskin et al.(2014), the dengue season is variable but it tends to coincide with the rainy season. For the countries where have favorable climate for vector, GDT is accurate.

Goal and Model

We focus on forecasting chikungunya in the Dominican Republic (DR) in 2014.

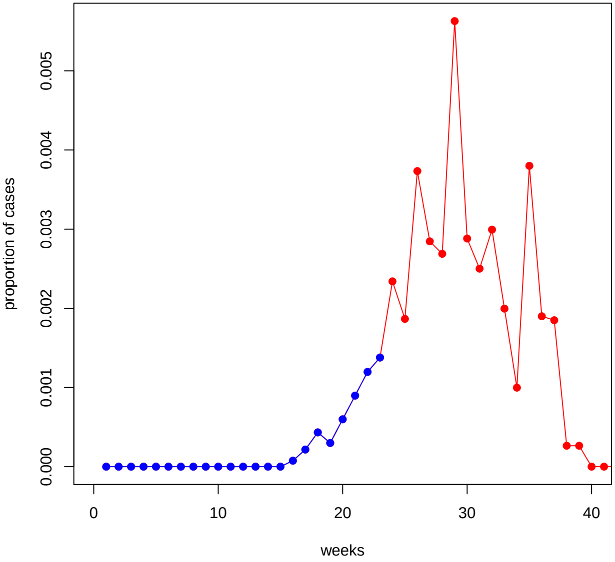
However, we have the first recorded case only in mid April 2014 and the spike has occurred already in mid July 2014, only 14 weeks after!

Our Goal: Predicting spike location and severity of chikungunya in DR, using the first 8 weeks of reported cases.

The model: A Box-Jenkings ARIMAX model with exogenonous regressors from two different sources:

- Non-Traditional information source (GDT)
- Epidemic deterministic model and prior information on its parameters

Chikungunya in DR during 2014



Harnessing Non-Traditional Sources

- We use Google Dengue Trend (GDT) as a proxy to unobserved mosquito density and as an evaluator of social activity.
- However, GDT is not available in the Dominican Republic!
- Instead, we use GDT in Mexico on a state-wide level. There are 17 states in Mexico with GDT starting 2003. Not all states are available over 2003-2014.

Non-Traditional Sources

1. We use the *TRUST* algorithm (Ciampi et al., 2010) to find time-space clusters and identify the GDT states in Mexico with similar information to chikungunya occurences in DF for the first 8 weeks.

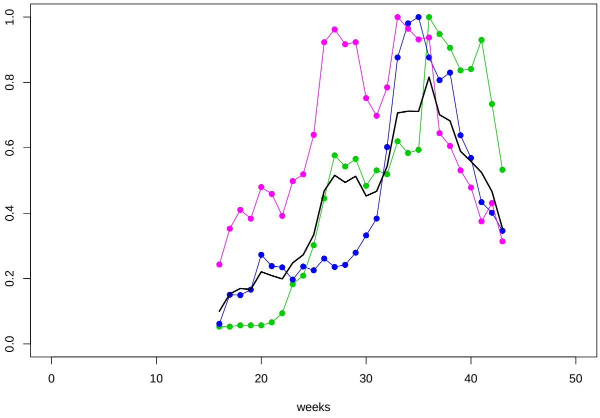
2. We select Mexican states that exhibit correlation with chikungunya in DR during the 3 previous years.

Map of Mexico

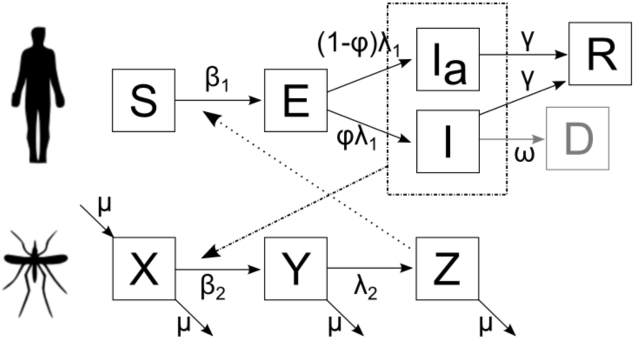
Morelos, Yucatan, and **Nayarit**.

Note that it is not the geographical proximity that drives resemblance of chikungunya and dengue!

Google Dengue Trend

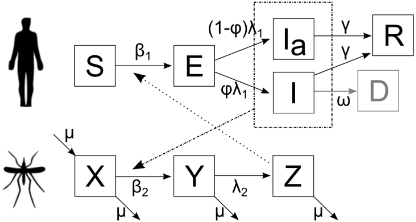


Deterministic Epidemic Model



Yakob and Clements (2013)

Deterministic Epidemic Model



$$
\begin{aligned}
dS &= -\beta_1 SZ \\
dE &= \beta_1 SZ - \lambda_1 E \\
dI &= \phi\lambda_1 E - \gamma I \\
dD &= \omega I \\
dI_a &= (1-\phi)\lambda_1 E - \gamma Ia \\
dR &= \gamma(I + Ia) \\
dX &= \mu - \beta_2 X(I + Ia) - \mu X \\
dY &= \beta_2 X(I + Ia) - \lambda_2 Y - \mu Y \\
dZ &= \lambda_2 Y - \mu Z
\end{aligned}
$$

Yakob and Clements (2013)

Model (Forward map, parameters)

Based on the weekly incidence information from the Réunion Island in 2005, Yakob and Clements fitted the β_1 and β_2 parameters (LS method).

$$\begin{aligned} \beta_1 &= 0.14, \\ \beta_2 &= 0.40, \\ \gamma &= 0.25, \\ \lambda_1 &= 0.50, \\ \lambda_2 &= 0.50, \\ \phi &= 0.97, \\ \omega &= 0.25, \\ \mu &= 0.05. \end{aligned}$$

Uncertainty Quantification (Bayesian Inference)

- **Prior** The prior is consider to be $\pi_\theta = \prod_{i=1}^{8} \pi_{\theta_i}$, where

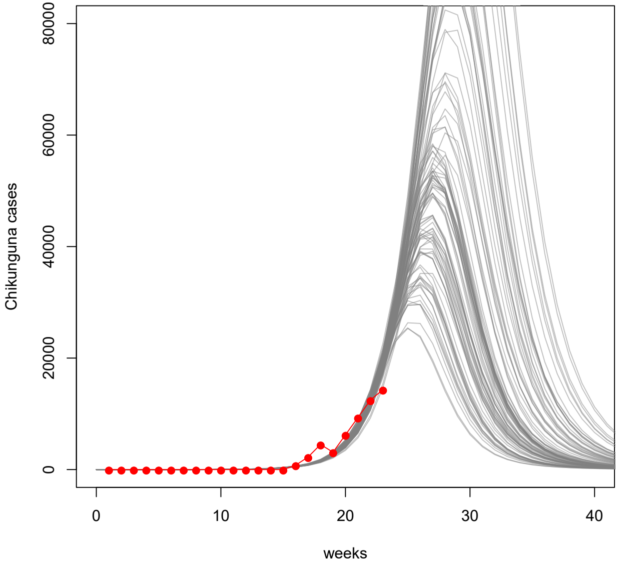
$$\pi_{\theta_i} \sim U(\theta_{0,i}(1-0.3), \theta_{0,i}(1+0.3)),$$

  where $\theta_{0,i}$ is the point estimates proposed by Yakobs and Clements.
- **Likelihood** We assume the observations are Negative Binomial with mean the FM ($G(t;)$) and variance $aG(i;)$. We consider a fixed ($a =$50, 100, 200 render similar results)
- **The posterior estimation: t-walk** (Christen and Fox, 2010) is a "A General Purpose Sampling Algorithm for Continuous Distributions" to sample from many objective functions.

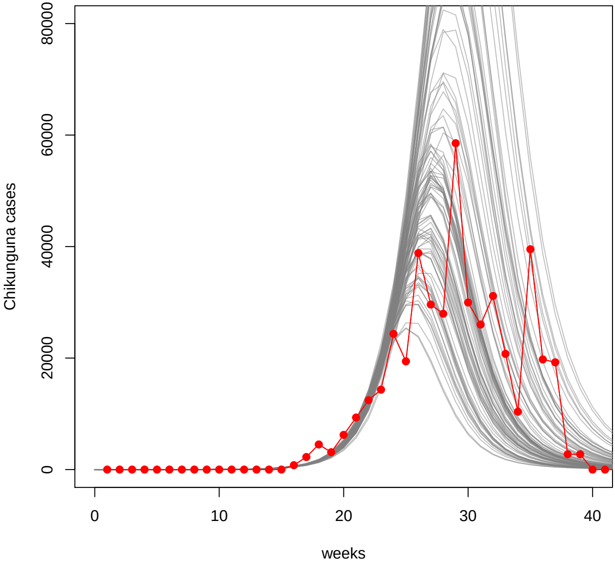
Uncertainty Quantification (Bayesian Inference)

In the case of the parameter ϕ we establish a distribution that can take into consideration the subreporting in the surveillance system. This problem is common for countries with a system of reporting and diagnosis that are not yet consolidated. For Dominican Republic we consider that this under reporting can go from 0 to obtain one reported case for every 40 real cases. Then we consider $\phi \sim$ Unif(0.0323,0.97).

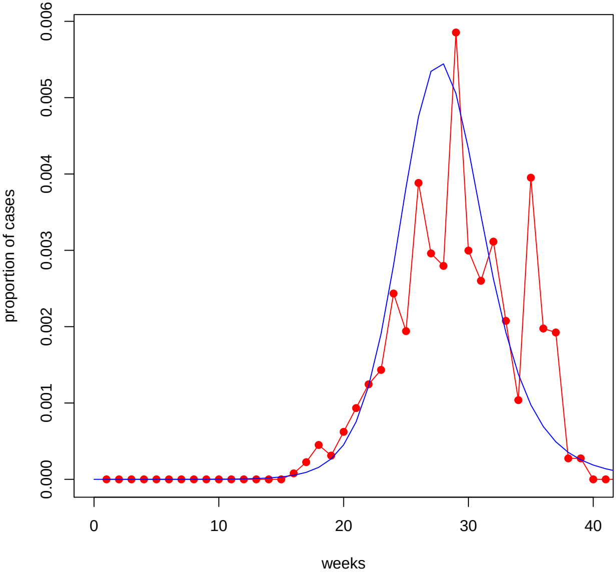
Sample of accepted curves



Sample of accepted curves



Averaged curve



The Model

Let $y_1, \ldots, y_t$ be observed weekly chikungunya counts and $\mathbf{X}_t = D_t | Z_t$ where D_t are averaged weekly Google Dengue reports in the past year for the selected states in Mexico, and Z_t the number of cases predicted from the ODE epidemic model.

We then use the Box-Jenkins ARIMAX:

$$y_t = \mathbf{X}_t' \boldsymbol{\beta} + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \psi_1 v_t + \ldots + \psi_q v_q,$$

where

- $\boldsymbol{\beta}$ is a vector of regression coefficients;
- v_t is $\mathrm{WN}(0, \sigma^2)$
- $\phi(\lambda) = 1 + \phi_1 \lambda + \ldots + \phi_p \lambda^p = (1 - \lambda)^d \tilde{\phi}(\lambda), \quad d \in \mathbf{Z}^+$;
- $\tilde{\phi}(\lambda) \neq 0, \quad \forall |\lambda| \leq 1$;
- $\psi(\lambda) = 1 + \psi_1 \lambda + \ldots + \psi_q \lambda^q$ and $\psi(\lambda) \neq 0, \quad \forall |\lambda| \leq 1$

Forecast 20-weeks ahead

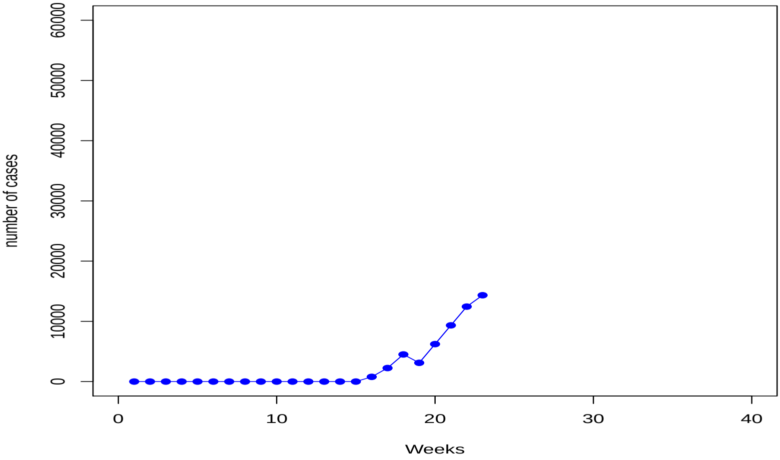
We estimate p and q with AIC and ARIMAX parameters with LS.

Now, we forecast chikungunya 20-weeks ahead, i.e., over the period of 24–44 weeks in 2014.

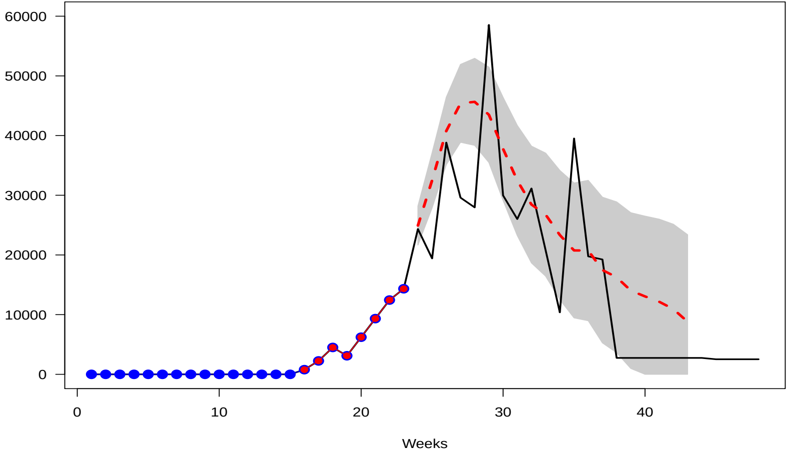
About the regressors:

- Since we do not know out-of-sample values of GDT in Morelos, Yucatan, and Nayarit for weeks 24–40 in 2014, we use the respective GDT values from 2013.
- We use the weekly averaged number of cases predicted by the deterministic model with accepted parameter values.

Step 3: Forecast 20-weeks ahead.



Step 3: Forecast 20-weeks ahead.



Discussion and Future Work

- With a very limit history of chikungunya in the Americas, the proposed method is able to estimate the spike location.
- Unfortunately GDT is not longer working but the results encourages to use web crawlers and text mining to estimate the infectious agent activity.
- Some open questions are:
  - How to automatically find optimal regional web-activity predictors?
  - The Other variables to add? Meteorological? Socio-demographics? All subject to data availability...

References

1. Ciampi, A; Appice, A and Malerba, D (2010) Discovering trend- based clusters in spatially distributed data streams, in *ECML/PKDD Workshop on Mining Ubiquitous and Social Environments*.
2. Yakob, L and Clements, ACA (2013) A Mathematical Model of Chikungunya Dynamics and Control: The Major Epidemic on Reunion Island. *PLoS ONE* 8(3):e57448.
3. Gluskin RT, Johansson MA, Santillana M and Brownstein JS(2014) Evaluation of Internet-Based Dengue Query Data: Google Dengue Trends. *PLoS neglected tropical diseases*. 8(2):e27
4. Christen, JA and Fox, C (2010) A General Purpose Sampling Algorithm for Continuous Distributions (the t-walk), *Bayesian Analysis*, 5(2), 263-282.
5. El-Metwally (2015) Google Search Trend of Dengue fever in developing Countries in 2013-2014: An Internet-Based Analysis. *J. Health Informatics in Developing Countries* 9(1).