Exploring Signals of Chikungunya in Puerto Rico

Wesley Chen and Mauricio Santillana

(File to be Read with “Interpretation.xlsx”)

General Discussion of Curve and Certain Features

*(reference sheet named “PAHO Cases Lagged 1 Week”)*

* The PAHO data does not start until late June, a couple months after the first confirmed cases of Chikungunya in PR.
* We see 3 spikes – one initial outbreak in mid-August followed by a larger one towards October and the largest spike predicted in mid-November
* Interestingly, the confirmed cases and suspected cases correlation from week to week is only 0.54
* We see a very low suspected number for Week of December 21st – which is strange and may be an artifact of poor data towards the end of the year (including irregular data between weeks)

Decisions Made

* Our golden standard for number of cases is PAHO’s suspected numbers to see if we can match PAHO’s model predictions (and not just confirmed numbers since Twitter can hopefully capture more than just the cases that are confirmed)
* We lag our Tweets by one week since we believe that people will first complain and self-report on Twitter before
* We decided to test LASSO with running windows as we believe that “trending” topics will artificially inflate Twitter numbers the first time something is mentioned and then will die out over time and lead to self-reporting only later on as the initial “crisis” response subsides

2 Static Training Stories

* 2 Divisions of the data into phases to try to understand which variables have the most weight at different phases of the trending nature, in-sample predictions only

3 Dynamic Stories

* There are 3 versions of LASSO regression with pretty different weightings of categories looking to predict numbers given a certain frame of training prior to that week

Static Divisions Per Outbreak (First, Second and Third Outbreaks)

*(reference sheet named “Static By Outbreak”)*

* We divide by the 3 outbreaks (including one week after each outbreak)
* There is a very high correlation (though the predicted values themselves aren’t as good as the division by time method
* The only method that really gets the magnitude of the last prediction well
* Large mix in the beginning – Chikungunya song, news and jokes
* Second phase is self reporting, educational and some jokes still
* Last category is a “unsure” category hard to explain, but general noise due to the unsure levels of the last peak even by PAHO (one week has high number of confirmed and PAHO may have over-predicted themselves)

Static Divisions In Time (Pre During and Post Trends)

*(reference sheet named “Static By Time”)*

* We divide by time, before the outbreaks, during the first two outbreaks and then post-trending outbreaks
* We see that we are able to track the onset and the first couple of outbreaks very well (while the matter is hot and tweeted a lot)
* The last peak is not tracked as well, though we still see a relative peak
* Pretty varied categories of just mentioning Chikungunya at start
* During the peaks phase is dominated by Chikungunya Bracelet (a preventive thing) and the Chikungunya song (perhaps cultural awareness)
* Last phase is also the “unclear” phase – perhaps due to the noise of data set again

Dynamic All Data

*(reference sheet named “All Dynamic”)*

* We see the best shape detection with dynamic data – the only method of the 3 to correlate to the 3rd (but largest) peak
* However, the magnitude is not close at all since despite seeing a peak, the volume of tweets is scaled down since the news is “old” now
* To support this, we see over time that self-reporting has an increasingly large weight
* The joke tweets are weights more in between and the “other category” becomes fairly even and slightly diminished (still a fragment of the beginning since we continue to train on all data)
* Since the model takes time to train, the early stages are not predicted well and though this is expected, there just simply wasn’t enough gold-standard data to train on since the suspected cases came so late

Dynamic 3 Week Moving Window

*(reference sheet named “3 Week Moving”)*

* When we narrow our window, we see that it picks up on week to week Twitter “trends” better and is better able to match the “crisis” response time for the first onset
* However later on, when there are fewer changes from week to week, a 3 week window does not pick up on the new crises – but is good at picking up levels between peaks (see Mid October to Early November between peaks)
* We see that the weights heavily swing – and very early on is an emphasis in self-reporting, then a quick change to news
* We see 3 peaks of self-reporting weights – which correspond to the 3 peaks of outbreak
* But the window is so small that the signal (potentially noisy) from the “other” categories obscure this – different “other” groups see different peaks at different times when we investigate into the components of the “other” groups

Dynamic 8 Week Moving Window

*(reference sheet named “8 Week Moving”)*

* We look to expand the rolling window to see if we can get the best of both worlds, but unfortunately there does not seem to be enough data to allow for an 8 week window to predict well post-trending on Twitter (after November when news is not hot)
* As with the 3 week rolling window method, we get the onset much better (when different categories interact very differently
* We also see a transition towards self-reporting weight right about when the outbreak happens BUT ONE WEEK EARLY (maybe after the news subsides, the 1 week lag may not be useful!)