A dynamic, ensemble learning approach to forecast dengue fever epidemic years in Brazil

Using weather and population susceptibility cycles

Aindrila Garai
M.Sc Statistics, IIT Kanpur

- under the guidance of **Dr. Tanujit Chakrabarty**
Assistant Professor of Mathematics (Statistics AI)
Sorbonne University and Sorbonne Center for AI

08/03/2023

Citation

Cite this article:

McGough SF, Clemente L, Kutz JN, Santillana M. 2021 A dynamic, ensemble learning approach to forecast dengue fever epidemic years in Brazil using weather and population susceptibility cycles. J. R. Soc. Interface 18: 20201006. https://doi.org/10.1098/rsif.2020.1006

Severity of Dengue

- Transmission of dengue fever depends on a complex interplay of human, climate and mosquito dynamics and most importantly **immunity** of human being.

Severity of Dengue

- Transmission of dengue fever depends on a complex interplay of human, climate and mosquito dynamics and most importantly **immunity** of human being.
- Aedes aegypti and Aedes albopictus mosquitoes infect an estimated 390 million people per year.

Severity of Dengue

- Transmission of dengue fever depends on a complex interplay of human, climate and mosquito dynamics and most importantly **immunity** of human being.
- Aedes aegypti and Aedes albopictus mosquitoes infect an estimated 390 million people per year.
- The global burden of dengue has doubled every 10 years over the last three decades.

Material and methods

Our workflow combines elements from

Material and methods

Our workflow combines elements from

1. Signal prepossessing
2. Time-series feature extraction
3. Independent model training and prediction
4. Model selection
5. Ensemble prediction
6. Dengue cycles

Signal pre-processing

Windowing technique

Windowing is typically used to improve signal clarity to incorporate the information in the days both within and around each time interval.

Signal pre-processing

Windowing technique

Windowing is typically used to improve signal clarity to incorporate the information in the days both within and around each time interval.

- Each time interval is defined by a start date between early June and late September and a period between 10 and 95 days.

Signal pre-processing

Windowing technique

Windowing is typically used to improve signal clarity to incorporate the information in the days both within and around each time interval.

- Each time interval is defined by a start date between early June and late September and a period between 10 and 95 days.
- Windowing approach:
  Define a rectangle of 5 $\times$ 6 - rows are five consecutive start dates and columns are six consecutive spanning period lengths.

Signal pre-processing

Windowing technique

Windowing is typically used to improve signal clarity to incorporate the information in the days both within and around each time interval.

- Each time interval is defined by a start date between early June and late September and a period between 10 and 95 days.
- Windowing approach:
  Define a rectangle of 5×6 - rows are five consecutive start dates and columns are six consecutive spanning period lengths.
- Each time interval and weather variable is summarized by 30 data points.

Signal pre-processing

Windowing technique

Windowing is typically used to improve signal clarity to incorporate the information in the days both within and around each time interval.

- Each time interval is defined by a start date between early June and late September and a period between 10 and 95 days.
- Windowing approach:
  Define a rectangle of 5×6 - rows are five consecutive start dates and columns are six consecutive spanning period lengths.
- Each time interval and weather variable is summarized by 30 data points.
- Build hundreds of models on different intervals to select those with the strongest signals.

Time-series feature extraction

Time-series feature extraction

It involves computing summary features of the time series, which can range from simple means to complex wavelet transforms.

Time-series feature extraction

Time-series feature extraction

It involves computing summary features of the time series, which can range from simple means to complex wavelet transforms.

To test the feasibility we extracted the following features within each time interval -

Time-series feature extraction

Time-series feature extraction

It involves computing summary features of the time series, which can range from simple means to complex wavelet transforms.

To test the feasibility we extracted the following features within each time interval -

1. The arithmetic means of daily temperature.
2. Mean precipitation frequency defined as the time interval between peaks of daily precipitation.

Independent model training and prediction

Goal: To identify the periods of the year that are most predictive of annual dengue outbreaks.

Independent model training and prediction

Goal: To identify the periods of the year that are most predictive of annual dengue outbreaks.

- Train a collection of independent SVM classifiers on an initial 7-year training period resulting 432 independent models trained per year.

Independent model training and prediction

Goal: To identify the periods of the year that are most predictive of annual dengue outbreaks.

- Train a collection of independent SVM classifiers on an initial 7-year training period resulting 432 independent models trained per year.
- Predictions are made by classifying the 30 out-of-sample data points corresponding to the weather information preceding the target year and taking a majority vote.

Independent model training and prediction

Goal: To identify the periods of the year that are most predictive of annual dengue outbreaks.

- Train a collection of independent SVM classifiers on an initial 7-year training period resulting 432 independent models trained per year.
- Predictions are made by classifying the 30 out-of-sample data points corresponding to the weather information preceding the target year and taking a majority vote.

Remark

Radial basis function and sigmoid kernels are used and model parameters (gamma, soft margin cost function and coefficient) using 10-fold cross-validation are tuned.

Model selection

The best-performing models(11) are selected each year based on -

- Historical out-of-sample prediction accuracy.

Model selection

The best-performing models(11) are selected each year based on -

- Historical out-of-sample prediction accuracy.
- Out-of-sample prediction accuracy of neighbouring models.

Model selection

The best-performing models(11) are selected each year based on -

- Historical out-of-sample prediction accuracy.
- Out-of-sample prediction accuracy of neighbouring models.

These models represent strongly predictive periods of the year preceding outbreaks as we allow all models to generate **4 years of out-of-sample predictions**.

Model selection

The best-performing models(11) are selected each year based on -

- Historical out-of-sample prediction accuracy.
- Out-of-sample prediction accuracy of neighbouring models.

These models represent strongly predictive periods of the year preceding outbreaks as we allow all models to generate **4 years of out-of-sample predictions**.

Note

The selection of the 11 models changes from year to year as the model-building process is dynamic.

Ensemble prediction

Ensemble learning combines the results of multiple trained predictors in order to generate a single robust prediction.

Ensemble prediction

Ensemble learning combines the results of multiple trained predictors in order to generate a single robust prediction.

- Use a simple majority vote of the 11 models to decide a single forecast which are produced for the last 6 years of the 17-year data-set.

Ensemble prediction

Ensemble learning combines the results of multiple trained predictors in order to generate a single robust prediction.

- Use a simple majority vote of the 11 models to decide a single forecast which are produced for the last 6 years of the 17-year data-set.
- These represent the culmination of a prediction process that involves:
  7-year initial training period,
  4-year out-of-sample model calibration period,
  6-year out-of-sample ensemble prediction period.

Dengue cycles

Endemic transmission of dengue fever is typically distinguished by periodic outbreak cycles of around 3–4 years because of

1. An exhaustion of susceptible after an outbreak.
2. Short-term cross-immunity to other circulating DENV serotypes after infection.

Dengue cycles

Title of the paper

Aindrila Garai

Citation

Severity of Dengue

Material and methods

Results

Conclusion

Acknowledgement

Endemic transmission of dengue fever is typically distinguished by periodic outbreak cycles of around 3–4 years because of

1. An exhaustion of susceptible after an outbreak.
2. Short-term cross-immunity to other circulating DENV serotypes after infection.

- So, implement a decision rule governed by the second- and third-order Markov transition probabilities, reflecting the transition between consecutive sequences of epidemic and non-epidemic states.

Dengue cycles

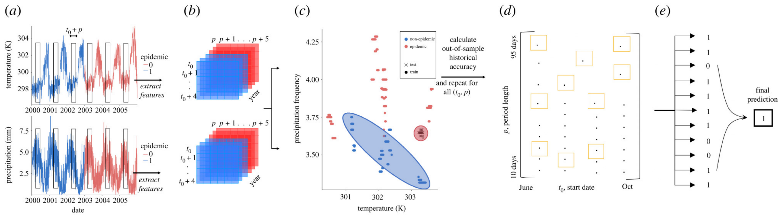
Endemic transmission of dengue fever is typically distinguished by periodic outbreak cycles of around 3–4 years because of

1. An exhaustion of susceptible after an outbreak.
2. Short-term cross-immunity to other circulating DENV serotypes after infection.

- So, implement a decision rule governed by the second- and third-order Markov transition probabilities, reflecting the transition between consecutive sequences of epidemic and non-epidemic states.
- The transition probabilities corresponding to the following 3- and 4-year cycles: 001, 110, 0001 and 1110.

Results

Figure: Ensemble forecast workflow



1. Extract features of temperature and precipitation.
2. An array of features corresponding to the mean value.
3. Train an SVM to classify next year's epidemic status.
4. Repeat for all 432 intervals and top 11 models are selected.
5. Contribute to a majority voting system.

Exploiting weather signals to create a data-driven forecast system

- Weather patterns are extracted and analysed across hundreds of partially overlapping time intervals spanning the last seven months of a year.

Exploiting weather signals to create a data-driven forecast system

- Weather patterns are extracted and analysed across hundreds of partially overlapping time intervals spanning the last seven months of a year.
- Each of these patterns is then assessed for its ability to predict an outbreak year for the subsequent year.

Exploiting weather signals to create a data-driven forecast system

- Weather patterns are extracted and analysed across hundreds of partially overlapping time intervals spanning the last seven months of a year.
- Each of these patterns is then assessed for its ability to predict an outbreak year for the subsequent year.
- Out-of-sample forecasts trained on a yearly expanding window are produced for 10 years and for each time interval using support vector machines.

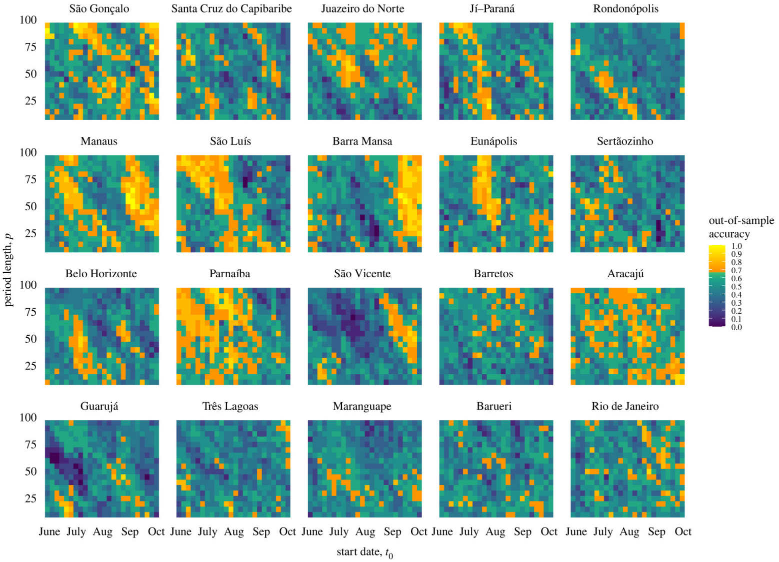
Exploiting weather signals to create a data-driven forecast system

- Weather patterns are extracted and analysed across hundreds of partially overlapping time intervals spanning the last seven months of a year.
- Each of these patterns is then assessed for its ability to predict an outbreak year for the subsequent year.
- Out-of-sample forecasts trained on a yearly expanding window are produced for 10 years and for each time interval using support vector machines.
- Every year, the time intervals with high historical predictive power are automatically selected and evaluated in the upcoming year to produce out-of-sample predictions for the subsequent dengue season.

Exploiting weather signals to create a data-driven forecast system

- Weather patterns are extracted and analysed across hundreds of partially overlapping time intervals spanning the last seven months of a year.
- Each of these patterns is then assessed for its ability to predict an outbreak year for the subsequent year.
- Out-of-sample forecasts trained on a yearly expanding window are produced for 10 years and for each time interval using support vector machines.
- Every year, the time intervals with high historical predictive power are automatically selected and evaluated in the upcoming year to produce out-of-sample predictions for the subsequent dengue season.
- An ensemble approach determines the system's final prediction: whether a year would be epidemic or not.

Out-of-sample forecast accuracy for 2008-2017



Weather-based forecasting performance

- Correctly forecast 81% of all epidemic years across 20 municipalities in Brazil between 2012 and 2017.

Weather-based forecasting performance

- Correctly forecast 81% of all epidemic years across 20 municipalities in Brazil between 2012 and 2017.
- Identifies 58% of non-epidemic years correctly.

Weather-based forecasting performance

- Correctly forecast 81% of all epidemic years across 20 municipalities in Brazil between 2012 and 2017.
- Identifies 58% of non-epidemic years correctly.
- This results in an overall accuracy of 72%.

Weather-based forecasting performance

- Correctly forecast 81% of all epidemic years across 20 municipalities in Brazil between 2012 and 2017.
- Identifies 58% of non-epidemic years correctly.
- This results in an overall accuracy of 72%.
- **Cities with better prediction accuracy have stronger weather signatures** such that
    1. South-eastern municipality of Barra Mansa,
    2. Manaus

  Five out of six ensemble years predict correctly for both of these two.

Weather-based forecasting performance

- Correctly forecast 81% of all epidemic years across 20 municipalities in Brazil between 2012 and 2017.
- Identifies 58% of non-epidemic years correctly.
- This results in an overall accuracy of 72%.
- **Cities with better prediction accuracy have stronger weather signatures** such that
  1. South-eastern municipality of Barra Mansa,
  2. Manaus

  Five out of six ensemble years predict correctly for both of these two.
- Weather-based predictions are less successful due to no clear temporal trend like Rio de Janeiro.

Incorporating dengue susceptibility cycles

- Given the previously observed sequence of consecutive outbreak and non-outbreak years (dengue fever cycles), the Markov model computes the probability of the next year being an outbreak or a non-outbreak year.

Incorporating dengue susceptibility cycles

- Given the previously observed sequence of consecutive outbreak and non-outbreak years (dengue fever cycles), the Markov model computes the probability of the next year being an outbreak or a non-outbreak year.
- While the weather conditions are identified to be an outbreak, there is stronger evidence that the population may have low susceptibility to infection based on multiple consecutive preceding years.

Results

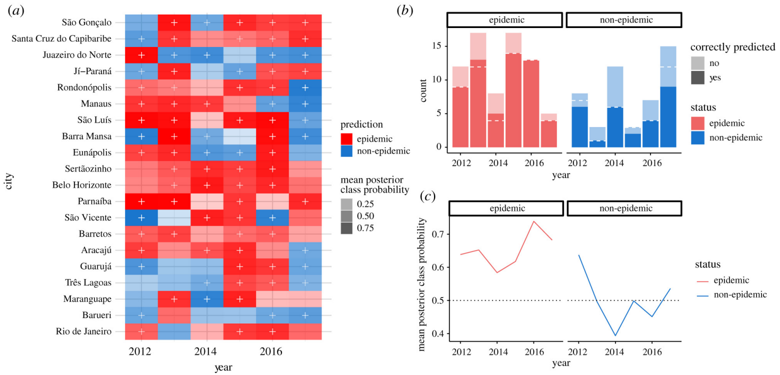


Figure: Weather-based prediction results for 120 municipality years

1. Annual out-of-sample forecasts of outbreak status.
2. Number of correctly forecasting year.
3. The mean posterior class probability across municipalities,

Model performance by year

- 2015–2017 epidemics are predicted by the weather-only models with at least 80% accuracy, with 100% of the 13 outbreaks in 2016 correctly.

Model performance by year

- 2015–2017 epidemics are predicted by the weather-only models with at least 80% accuracy, with 100% of the 13 outbreaks in 2016 correctly.
- Non-epidemic years (2013–2014) are with only one-third and one-half of cities correctly forecasting non-epidemics for these years.

Model performance by year

- 2015–2017 epidemics are predicted by the weather-only models with at least 80% accuracy, with 100% of the 13 outbreaks in 2016 correctly.
- Non-epidemic years (2013–2014) are with only one-third and one-half of cities correctly forecasting non-epidemics for these years.
- In 2012, the most successful non-epidemic predictions are occurred with 75% accuracy.

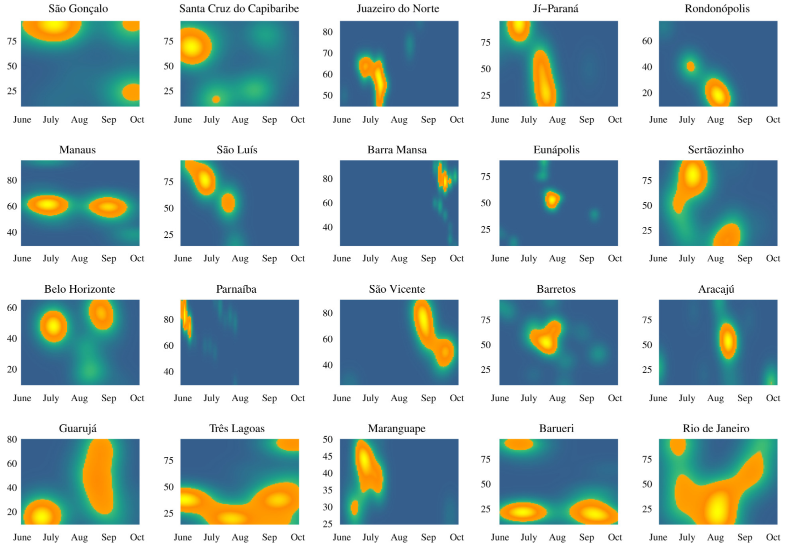
Model performance by year

- 2015–2017 epidemics are predicted by the weather-only models with at least 80% accuracy, with 100% of the 13 outbreaks in 2016 correctly.
- Non-epidemic years (2013–2014) are with only one-third and one-half of cities correctly forecasting non-epidemics for these years.
- In 2012, the most successful non-epidemic predictions are occurred with 75% accuracy.
- 80% and 85% of municipalities are correctly classified as epidemics or non-epidemics in 2015 and 2016.

Model performance by year

- 2015–2017 epidemics are predicted by the weather-only models with at least 80% accuracy, with 100% of the 13 outbreaks in 2016 correctly.
- Non-epidemic years (2013–2014) are with only one-third and one-half of cities correctly forecasting non-epidemics for these years.
- In 2012, the most successful non-epidemic predictions are occurred with 75% accuracy.
- 80% and 85% of municipalities are correctly classified as epidemics or non-epidemics in 2015 and 2016.
- In 2014 and 2017, 45% and 35% municipalities are misclassified respectively.

Periods of the year selected into the ensemble forecast model for 2012–2017



Quantifying the strength of predictions

- To assess the strength of each individual classifier, we estimate whether the separability between the two classes is well captured by the classifier by extracting posterior probabilities of each SVM model.

Quantifying the strength of predictions

- To assess the strength of each individual classifier, we estimate whether the separability between the two classes is well captured by the classifier by extracting posterior probabilities of each SVM model.

Note

The posterior probability reflects the distance to the separation boundary distinguishing epidemic and non-epidemic years on the basis of weather.

Quantifying the strength of predictions

- To assess the strength of each individual classifier, we estimate whether the separability between the two classes is well captured by the classifier by extracting posterior probabilities of each SVM model.

Note

The posterior probability reflects the distance to the separation boundary distinguishing epidemic and non-epidemic years on the basis of weather.

- A higher probability represents how strongly the weather patterns of the prediction year aligned with predictions.

Quantifying the strength of predictions

- To assess the strength of each individual classifier, we estimate whether the separability between the two classes is well captured by the classifier by extracting posterior probabilities of each SVM model.

Note

The posterior probability reflects the distance to the separation boundary distinguishing epidemic and non-epidemic years on the basis of weather.

- A higher probability represents how strongly the weather patterns of the prediction year aligned with predictions.
- But only this measure of separability is not a particularly good indicator of accuracy.

Conclusion

Discussions:

- Complementing our weather-based approach with observed 3- to 4-year outbreak cycles is key to achieve higher accuracy and improve in predicting non-epidemic years.

Conclusion

Discussions:

- Complementing our weather-based approach with observed 3- to 4-year outbreak cycles is key to achieve higher accuracy and improve in predicting non-epidemic years.
- Provide timely information on dengue fever activity to policymakers **months ahead of outbreak seasons**.

Conclusion

Discussions:

- Complementing our weather-based approach with observed 3- to 4-year outbreak cycles is key to achieve higher accuracy and improve in predicting non-epidemic years.
- Provide timely information on dengue fever activity to policymakers **months ahead of outbreak seasons**.
- These models can be extended to other locations requiring no location-specific manipulation.

Conclusion

Discussions:

- Complementing our weather-based approach with observed 3- to 4-year outbreak cycles is key to achieve higher accuracy and improve in predicting non-epidemic years.
- Provide timely information on dengue fever activity to policymakers **months ahead of outbreak seasons**.
- These models can be extended to other locations requiring no location-specific manipulation.
- Our models miss half of non-epidemics in 2014 because of the immunity provided by a large outbreak in 2013.

Conclusion

Discussions:

- Complementing our weather-based approach with observed 3- to 4-year outbreak cycles is key to achieve higher accuracy and improve in predicting non-epidemic years.
- Provide timely information on dengue fever activity to policymakers **months ahead of outbreak seasons**.
- These models can be extended to other locations requiring no location-specific manipulation.
- Our models miss half of non-epidemics in 2014 because of the immunity provided by a large outbreak in 2013.
- Our approach achieves an overall accuracy of 75%.

Acknowledgement

Thank you,