Dengue Forecasting Project

In areas where dengue is endemic, incidence follows seasonal transmission patterns punctuated every few years by much larger epidemics. Because these epidemics are currently unpredictable and of major consequence to the affected populations, the initial focus of the pilot project will be forecasting key metrics for historical dengue seasons using only data from time periods prior to those seasons. The targets, data, forecasts, and evaluation are described below.

Overall Timeline

- Training data release for model development: June 5, 2015
- Model description and training forecasts due: August 12, 2015
- Testing data release for model evaluation: August 19, 2015
- Testing forecasts due: September 2, 2015
- Model and forecast evaluation workshop: date to be decided, mid to late September, 2015

Forecast targets

For a given transmission season (a 12-month period commencing with the location-specific, historical lowest incidence week), the following targets will be forecasted:

A. Timing of peak incidence, the week when the highest incidence of dengue occurs during the transmission season,
B. Maximum weekly incidence, the number of dengue cases reported during the week when incidence peaks, and
C. Total number of cases in a transmission season, the number of dengue cases (confirmed or suspected, depending on the location) reported in the transmission season.

Data

Dengue data. Historical surveillance data is provided for two endemic locations: Iquitos, Peru and San Juan, Puerto Rico. The data include weekly laboratory-confirmed and serotype-specific cases for each location. All data are "final", i.e. the case counts reflect all cases for each week regardless of whether those reports were actually available at that time. Metadata and references are provided for each dataset.

Other data. Environmental data from weather stations, satellites, and climate models are also provided along with relevant metadata. These data are on a daily time scale. Participants may also use other data sources such as social media or demographic data, but should NOT use other data on dengue in the study locations or nearby locations unless they make those data available to all participants through the project organizers. Dengue-related data on the mosquito vectors or disease in humans from other surveillance systems or research studies should not be used even if published and publicly available without explicit approval by the project organizers. Other types of data such as search queries or social media including the word "dengue" may be used.

Temporal range. Data will be provided for two time periods at each location, a training period and a testing time period. The initial data only covers the training period, 1990-2009 for San Juan and 2000-2009 for Iquitos. Testing data for 2009-2013 will be provided later, as described in the "Forecasts" section below. Note that no forecast should use data that comes from a later time period. For example, a forecast made at Week 28 of the 2010/2011 season should only use data (dengue data or other data) from prior weeks and years.

Forecasts

Forecasts will be made in two stages. First, the training data should be used by each team to develop and select the optimal model for each location and prediction target. Once this has been accomplished, the team should write a brief description of the model and data used (Appendix 1). If different models are used for different targets or locations, each model should be described. The team should also prepare forecasts for the years 2005-2009 using the selected model(s). For each of these four transmission seasons, forecasts should be made every 4 weeks (weeks 0, 4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44, 48) for each target. Each forecast should include a point estimate and a probability distribution. Note that forecasts should be made for peak incidence even after peak incidence has occurred. These forecasts reflect the probability that peak incidence has already occurred (e.g. late season forecasts should be non-zero if there is some chance of a second, higher peak).

One “csv” file should be prepared for each location and each target using the supplied templates (Appendix 2). The initial model description and forecasts should be submitted to predict@cdc.gov by August 12, 2015. These forecasts will be used to verify that the format is correct and to provide metrics on fit to the training data.

All teams with verified submissions by August 12, 2015, will receive the testing data by email in the same format as the training data on August 19, 2015. They will have two weeks to submit forecasts for the 2009-2013 testing period using the already selected model. These forecasts should use exactly the same model and same format as the first submission and must be submitted to predict@cdc.gov by September 2, 2015.

IMPORTANT NOTE: Much of the data for 2009-2013 is currently accessible to researchers; it is therefore contingent upon the researcher to NOT use these data for model development or evaluation. The data are supplied only for “future” forecasts within the testing period. For example, forecasts made for the 2011/2012 season at Week 4 may use data from any date up to Week 4 of the 2011/2012 season, but no data of any type from later dates. The historical data may be used to dynamically update coefficients or covariates, but there should be no structural changes to the model and no consideration of data from Week 5 of that season or later.

Evaluation

Forecasts will be quantitatively evaluated for each target using two metrics. Point forecasts will be evaluated using relative Mean Absolute Error (relative MAE, Appendix 3) to assess performance relative to a seasonal autoregressive model and to other forecasts. The probability distributions will be evaluated using the logarithmic scoring rule (Appendix 3). For each target, relative MAE and the logarithmic score will be calculated across all seasons and forecast times (week of the season) as well as for specific seasons and forecast time to identify potential differences in model strengths. The primary comparisons will be made for the testing period (2009-2013), however forecasts will also be compared between the training and testing periods to assess how forecast accuracy changes when predicting on data that was excluded from the model development process. IMPORTANT NOTE: A different model may be employed for each target and location.

Of all teams completing the submission process within the specified time frame, up to six will be selected to give in depth presentations on their forecasting approaches at the September workshop. The selected teams will be those achieving the highest logarithmic score over forecasts at weeks 0 through 24 across all four testing seasons for each of the three targets in each of the two cities, i.e. there will be one model selected for each target and each city. An interagency panel will evaluate the submissions and select the presenters.

Following the workshop, the organizers will put together a manuscript to publish the evaluation results. All submitted forecasts will be included in the manuscript. All team leaders will be invited as authors. The metrics for all forecasts will be shared with all participants at the workshop for independent analysis. However, these data (data related to forecasts from other participants, not the participant’s own forecasts) should not be used in publications until one year after the workshop or until there is a joint publication related to the workshop, whichever comes first. After this embargo, any participant can use these data as long as no identifiers are used. To include identifiers, permission must be obtained from each team leader.

Appendix 1 – Model description

Once model development has been finished, each team should select their best model for future forecasts. Note again that there may be different models for different targets and locations, but only one for each target and location (though that may be an ensemble model). If different models are selected for different targets/locations, the description should include each of those models. The description should include the following components:

1. **Team name:** This should match the name used in the submission file names.
2. **Team members:** List every person involved with the forecasting effort and their institution. Include the email address of the team leader.
3. **Agreement:** Include the following statement: "By submitting these forecasts, I (we) indicate my (our) full and unconditional agreement to abide by the project's official rules and data use agreements."
4. **Model description:** Is the model mechanistic, statistical? Is it an instance of a known class of models? The description should include sufficient detail for another modeler to understand the approach being applied. It may include equations, but that is not necessary. If multiple models are used, describe each model and which target each model was used to predict.
5. **Variables:** What data is used in the model? Historical dengue data? Weather data? Other data? List every variable used and its temporal relationship to the forecast (e.g. lag or order of autocorrelation). If multiple models are used specify which variables enter into each model.
6. **Computational resources:** What programming languages/software tools were used to write and execute the forecasts?
7. **Publications:** Does the model derive directly from previously published work? If so please include references.

Appendix 2 – Forecast format

Forecasts should be submitted as separate csv files for each location and each forecasting target using the provided templates. The name convention for the files is [teamname]_[target]_[location]_[dataset].csv. For example, Team A's forecasts for the week of peak incidence for San Juan in the training seasons should be submitted as "teama_ peakweek_sanjuan_train.csv".

Templates are provided for each location and forecasting target. The header row in each of these indicates the transmission season for which the forecast should be made and the most recent data that may be used for the forecast. The format of the column names is "2009/2010_wk4" where "2009/2010" indicates the dengue season over which the forecast should occur and "wk4" is the most recent data that may be used in that forecast. For example, a peak incidence forecast for "2009/2010_wk4" would be the forecast for peak weekly incidence in the 2009/2010 season using only data available through Week 4 of the 2009/2010 season. Note that predictions at Week 0, cannot use any data from the target transmission season, only previous seasons.

The row names in each template indicate the prediction that should be made. "point" refers to a point prediction, i.e. the most likely outcome for the specific target. The units should be weeks for the week of peak incidence and cases for the peak weekly incidence and total incidence for the season. The following rows indicate a binned probability distribution for the target which may differ according to the target and the location. For peak week, these rows are "p(peak_week=1)" through "p(peak_week=52)". "p(peak_week=1)" is the probability (between 0 and 1) that the peak occurs in the first week of the season.

For the peak weekly incidence and total incidence for the season, these probabilities are based on the distributions of cases observed in the each location and represent a range of numbers of cases. As an example, for peak weekly incidence in San Juan, "p(0<=peak_incidence<50)" is the probability that maximum weekly number of cases observed during the season is less than 50 cases and "p(150<=peak_incidence<200)" is the probability that the maximum is between 150 and 200 cases. For total incidence in the season, the probabilities are expressed as "p(5000<=season_incidence<6000)". For the specific bins for each location, please see the templates themselves.

EXAMPLE: Each forecast is a column in a csv file that uses a specific set of data and makes a point and probability distribution forecast for a specific target in a specific season. For total seasonal incidence in San Juan in 2010/2011, a prediction may be made at Week 8 of the season. It may use any data from Week 8 of the 2010/2011 and data from any earlier week. An example is shown below. The forecast consists of a point forecast (6,291 cases) and a probability distribution which sums to 1.0. Not that there is a probability assigned to each outcome, even if that probability is zero.

	…	2010/2011_wk8	…
point		6291	
p(0<=season_incidence<1000)		0.0	
p(1000<=season_incidence<2000)		0.0	
p(2000<=season_incidence<3000)		0.0	
p(3000<=season_incidence<4000)		0.0	
p(4000<=season_incidence<5000)		0.007	
p(5000<=season_incidence<6000)		0.27	
p(6000<=season_incidence<7000)		0.638	
p(7000<=season_incidence<8000)		0.085	
p(8000<=season_incidence<9000)		0.0	
p(9000<=season_incidence<10000)		0.0	
p(10000<=season_incidence)		0.0	

Appendix 3 – Evaluation metrics

Relative Mean Absolute Error (relative MAE)

Mean absolute error (MAE) is the mean absolute difference between predictions $\hat{y}$ and observations y over n data points:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|.$$

Relative MAE for models A and B is:

$$relMAE_{A,B} = \frac{MAE_A}{MAE_B}.$$

An important feature of this metric is that it can be interpreted directly in terms of accuracy of predictions. For example, a relative MAE(A,B) = 0.8 indicates that, on average, predictions from model A were 20% closer to the observed values than those from model B. Additionally, comparing multiple candidate models to a common baseline model with relative MAE allows for the assessment of the relative accuracy of the candidate models. For example, the relative MAE for model A versus the baseline model can be divided by the relative MAE for model B versus the baseline, resulting in the relative MAE for model A versus model B.

References:

Hyndman RJ and AB Koehler. (2006) Another look at measures of forecast accuracy. International Journal of Forecasting. 22(4):679-688. Available at: http://www.buseco.monash.edu.au/ebs/pubs/wpapers/2005/wp13-05.pdf

Reich NG, J Lessler, K Sakrejda, SA Lauer, S Iamsirithaworn, and DAT Cummings. (2015) Case studies in evaluating time series prediction models using the relative mean absolute error. Available at: http://works.bepress.com/nicholas_reich/11

Logarithmic Score

The logarithmic score is a proper scoring rule based on a binned probability distribution of the prediction, $\mathbf{p}$. The score is the log of the probability assigned to the observed outcome, i:

$$S(\mathbf{p}, i) = \log(p_i).$$

For example, a single prediction for the peak week includes probabilities for each week (1-52) in the season. The probability assigned to a week when the peak is unlikely would be low, near to zero, while the probability assigned to the forecast peak week is the highest. The total of these probabilities across all weeks must equal 1, i.e.:

$$\sum_{i=1}^{52} p_i = 1$$

If the observed peak week is week 30 and $p_{30} = 0.15$, the score for this prediction is $\log(0.15)$, or approximately -1.9.

Note that the logarithmic score is based on a *probabilistic estimate* – a complete probability distribution over all possible outcomes. This is desirable because it requires the forecaster to consider the entire range of possible outcomes, and to estimate the likelihood if each one of them. Two forecasters may agree on which outcome is the most likely, and

therefore submit the same point estimate, which would be scored identically by MAE. However, their predictions may differ substantially on *how likely* this outcome is, and how likely other outcomes are. By considering this probability, the logarithmic score enables scoring of the confidence in the prediction, not just the value of the point prediction.

Another advantage of logarithmic scores is that they can be summed across different time periods, targets, and locations to provide both specific and generalized measures of model accuracy. The bins that will be used for forecasts of the peak week, maximum weekly incidence, and total cases in the season will be specified in the forecast templates (see Appendix 2).

References:

Gneiting T and AE Raftery. (2007) Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association. 102(477):359-378. Available at https://www.stat.washington.edu/raftery/Research/PDF/Gneiting2007jasa.pdf.

Rosenfeld R, J Grefenstette, and D Burke. (2012) A Proposal for Standardized Evaluation of Epidemiological Models. Available at: http://delphi.midas.cs.cmu.edu/files/StandardizedEvaluation_Revised_12-11-09.pdf