HEALTH

DengAI: Predicting Disease Spread

Using environmental data collected by U.S. Federal Government agencies, can you predict the number of dengue fever cases reported each week in San Juan, Puerto Rico and Iquitos, Peru?

Intermediate practice **9 months** left **14,149** joined

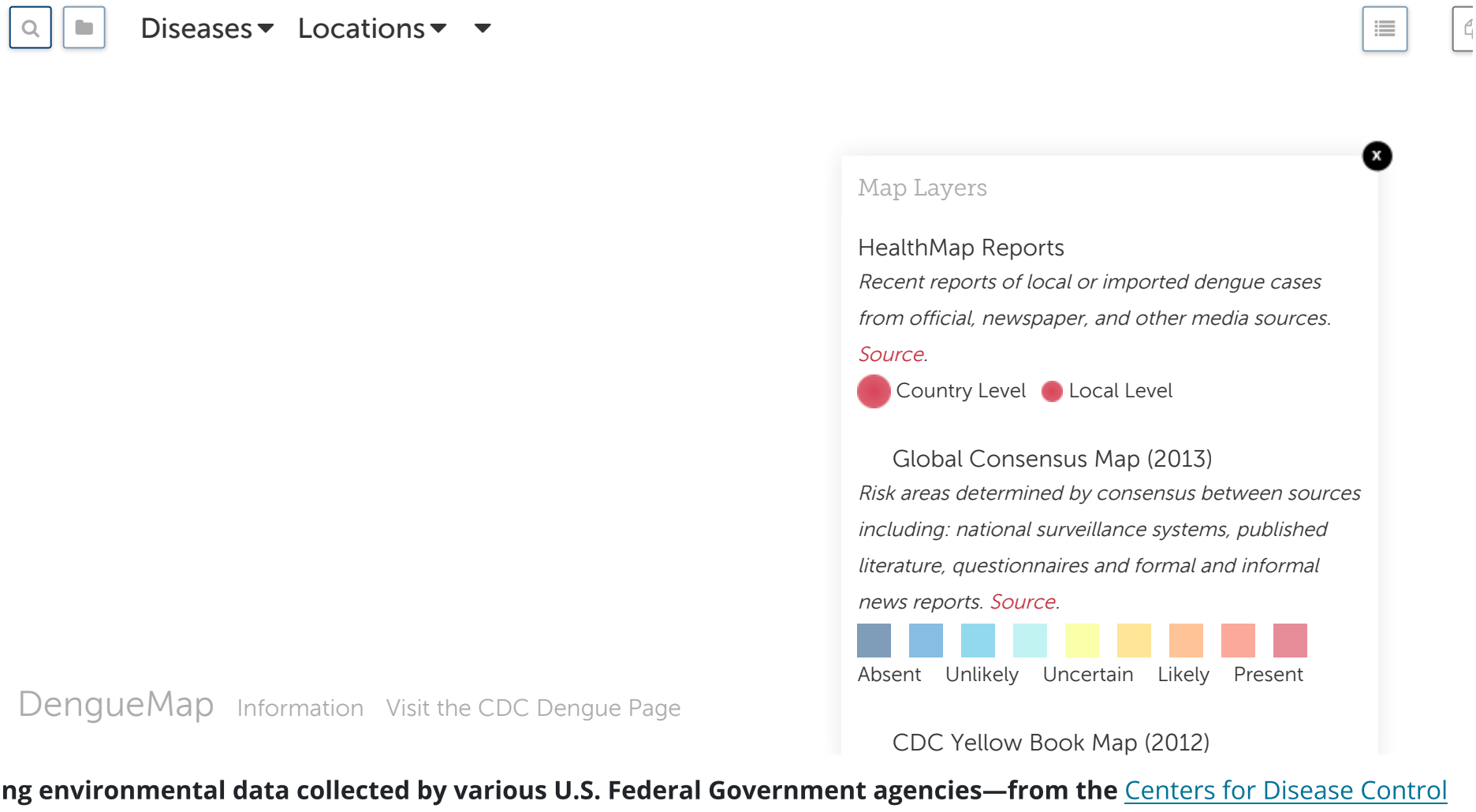
Challenge Summary

Can you predict local epidemics of dengue fever?

Dengue fever is a mosquito-borne disease that occurs in tropical and sub-tropical parts of the world. In mild cases, symptoms are similar to the flu: fever, rash, and muscle and joint pain. In severe cases, dengue fever can cause severe bleeding, low blood pressure, and even death.

Because it is carried by mosquitoes, the transmission dynamics of dengue are related to climate variables such as temperature and precipitation. Although the relationship to climate is complex, a growing number of scientists argue that climate change is likely to produce distributional shifts that will have significant public health implications worldwide.

In recent years dengue fever has been spreading. Historically, the disease has been most prevalent in Southeast Asia and the Pacific islands. These days many of the nearly half billion cases per year are occurring in Latin America:



Using environmental data collected by various U.S. Federal Government agencies—from the Centers for Disease Control and Prevention to the National Oceanic and Atmospheric Administration in the U.S. Department of Commerce—can you predict the number of dengue fever cases reported each week in San Juan, Puerto Rico and Iquitos, Peru?

This is an intermediate-level practice competition. Your task is to predict the number of dengue cases each week (in each location) based on environmental variables describing changes in temperature, precipitation, vegetation, and more.

An understanding of the relationship between climate and dengue dynamics can improve research initiatives and resource allocation to help fight life-threatening pandemics.

Competition End Date:

Oct. 5, 2024, 7:38 p.m.

This competition is for learning and exploring, so the deadline may be extended in the future.

How to compete

1. Click the "Join the competition" button on the sidebar to enroll in the competition.
2. Get familiar with the problem on the Problem Description and About page.
3. Download the data from the Data tab.
4. Create and train your own model. The benchmark blog post is a great place to start!
5. Use your model to generate predictions that match the submission format. Click “Submit” in the sidebar, and then “Make new submission”. You’re in!
6. Bonus: share your work! Click the "+" icon on the Submissions page and add a link to your approach.

Mosquito image courtesy of flickr user sanofi-pasteur

DRIVENDATA

HEALTH

DengAI: Predicting Disease Spread

Using environmental data collected by U.S. Federal Government agencies, can you predict the number of dengue fever cases reported each week in San Juan, Puerto Rico and Iquitos, Peru?

Intermediate practice **9 months** left **14,149** joined

Predict the Next Pandemic Initiative

The data for this competition comes from multiple sources aimed at supporting the Predict the Next Pandemic Initiative. Dengue surveillance data is provided by the U.S. Centers for Disease Control and prevention, as well as the Department of Defense's Naval Medical Research Unit 6 and the Armed Forces Health Surveillance Center, in collaboration with the Peruvian government and U.S. universities. Environmental and climate data is provided by the National Oceanic and Atmospheric Administration (NOAA), an agency of the U.S. Department of Commerce.

In their own words:

> Accurate dengue predictions would help public health workers … and people around the world take steps to reduce the impact of these epidemics. But predicting dengue is a hefty task that calls for the consolidation of different data sets on disease incidence, weather, and the environment.

This is a complicated and messy problem, to be sure. But real data is often complicated and messy. Study up using the resources below—your insights could save lives!

You can learn more here:

- Dengue Forecasting Homepage
- CDC Dengue Overview
- NOAA Wiki

NOAA logo

HEALTH

DengAI: Predicting Disease Spread

Using environmental data collected by U.S. Federal Government agencies, can you predict the number of dengue fever cases reported each week in San Juan, Puerto Rico and Iquitos, Peru?

Intermediate practice **9 months** left **14,149** joined

Problem description

Your goal is to predict the `total_cases` label for each `(city, year, weekofyear)` in the test set. There are two cities, San Juan and Iquitos, with test data for each city spanning 5 and 3 years respectively. You will make one submission that contains predictions for both cities. The data for each city have been concatenated along with a `city` column indicating the source: `sj` for San Juan and `iq` for Iquitos. The test set is a pure future hold-out, meaning the test data are sequential and non-overlapping with any of the training data. Throughout, missing values have been filled as `NaN`s.

Features

List of features

Example of features

Performance metric

Mean absolute error

Submission Format

Format example

The features in this dataset

You are provided the following set of information on a `(year, weekofyear)` timescale:

(Where appropriate, units are provided as a `_unit` suffix on the feature name.)

City and date indicators

- `city` – City abbreviations: `sj` for San Juan and `iq` for Iquitos
- `week_start_date` – Date given in yyyy-mm-dd format

NOAA's GHCN daily climate data weather station measurements

- `station_max_temp_c` – Maximum temperature
- `station_min_temp_c` – Minimum temperature
- `station_avg_temp_c` – Average temperature
- `station_precip_mm` – Total precipitation
- `station_diur_temp_rng_c` – Diurnal temperature range

PERSIANN satellite precipitation measurements (0.25x0.25 degree scale)

- `precipitation_amt_mm` – Total precipitation

NOAA's NCEP Climate Forecast System Reanalysis measurements (0.5x0.5 degree scale)

- `reanalysis_sat_precip_amt_mm` – Total precipitation
- `reanalysis_dew_point_temp_k` – Mean dew point temperature
- `reanalysis_air_temp_k` – Mean air temperature
- `reanalysis_relative_humidity_percent` – Mean relative humidity
- `reanalysis_specific_humidity_g_per_kg` – Mean specific humidity
- `reanalysis_precip_amt_kg_per_m2` – Total precipitation
- `reanalysis_max_air_temp_k` – Maximum air temperature
- `reanalysis_min_air_temp_k` – Minimum air temperature
- `reanalysis_avg_temp_k` – Average air temperature
- `reanalysis_tdtr_k` – Diurnal temperature range

Satellite vegetation - Normalized difference vegetation index (NDVI) - NOAA's CDR Normalized Difference Vegetation Index (0.5x0.5 degree scale) measurements

- `ndvi_se` – Pixel southeast of city centroid
- `ndvi_sw` – Pixel southwest of city centroid
- `ndvi_ne` – Pixel northeast of city centroid
- `ndvi_nw` – Pixel northwest of city centroid

Feature data example

For example, a single row in the dataset, indexed by (city, year, weekofyear): (sj, 1994, 18), has these values:

week_start_date	1994-05-07
total_cases	22
station_max_temp_c	33.3
station_avg_temp_c	27.7571428571

station_precip_mm	10.5
station_min_temp_c	22.8
station_diur_temp_rng_c	7.7
precipitation_amt_mm	68.0
reanalysis_sat_precip_amt_mm	68.0
reanalysis_dew_point_temp_k	295.235714286
reanalysis_air_temp_k	298.927142857
reanalysis_relative_humidity_percent	80.3528571429
reanalysis_specific_humidity_g_per_kg	16.6214285714
reanalysis_precip_amt_kg_per_m2	14.1
reanalysis_max_air_temp_k	301.1
reanalysis_min_air_temp_k	297.0
reanalysis_avg_temp_k	299.092857143
reanalysis_tdtr_k	2.67142857143
ndvi_location_1	0.1644143
ndvi_location_2	0.0652
ndvi_location_3	0.1321429
ndvi_location_4	0.08175

Performance metric

Performance is evaluated according to the mean absolute error.

Submission format

The format for the submission file is simply the `(city, year, weekofyear)` and the predicted `total_cases` for San Juan or Iquitos (for an example, see `SubmissionFormat.csv` on the data download page). The `total_cases` should be represented as integer values.

For example, if you just predicted that there were 5 cases each week for 5 weeks in San Juan and 3 cases each week for 5 weeks in Iquitos, for a total of 10 weeks, you would have the following predictions:

city	year	weekofyear	total_cases
sj	2008	18	5
sj	2008	19	5
sj	2008	20	5
sj	2008	21	5
sj	2008	22	5
...			
iq	2013	22	3
iq	2013	23	3
iq	2013	24	3
iq	2013	25	3
iq	2013	26	3

Your `.csv` file that you submit would look like:

```
city,year,weekofyear,total_cases
sj,2008,18,5
sj,2008,19,5
sj,2008,20,5
sj,2008,21,5
sj,2008,22,5
...
iq,2013,22,3
iq,2013,23,3
iq,2013,24,3
iq,2013,25,3
iq,2013,26,3
```

Keep in mind that you need to submit one csv with predictions for both cities! Hence the requirement of the `city` column. Results will be parsed on our end and MAE scores will be given for each city's predictions.

Good luck!

Looking for a great tutorial to get you started? Check out the benchmark walkthrough created for this challenge.

Good luck and enjoy this problem! If you have any questions you can always visit the user forum!