2025 SPRINT: 2ND INFODENGUE-MOSQLIMATE DENGUE CHALLENGE (IMDC)

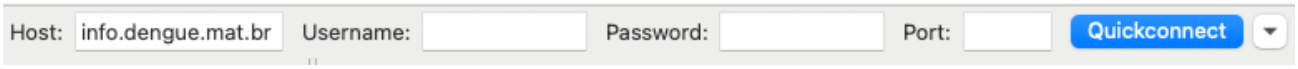
Data documentation

This section provides detailed documentation of the data available for the Sprint, including table descriptions and other essential information for data analysis and modelling.

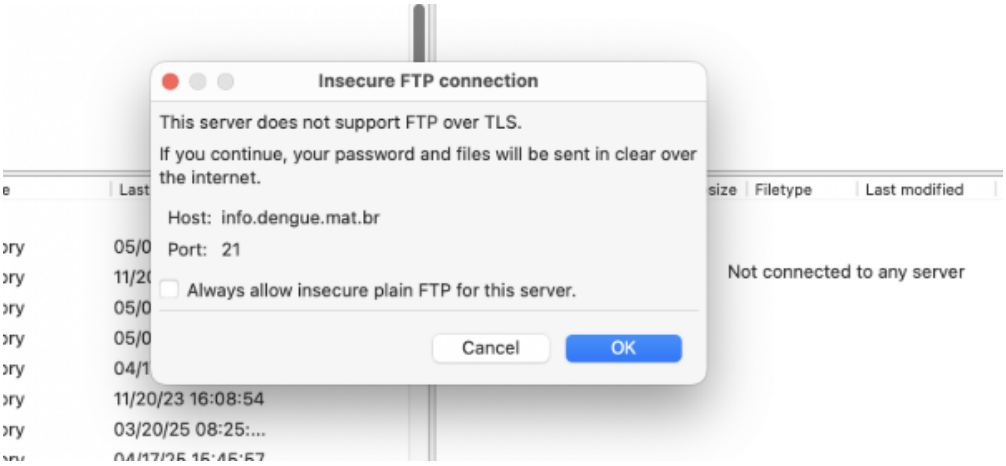
The data were uploaded to an FTP server. There are several ways to access the data on an FTP server. We will propose some here:

1- Using FileZilla

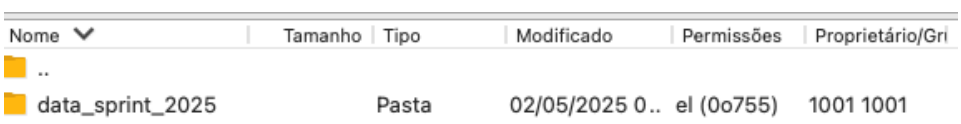
1. Download the FileZilla app from the official website: https://filezilla-project.org/.
2. Open the application, enter info.dengue.mat.br in the Host field, and click the button to connect.



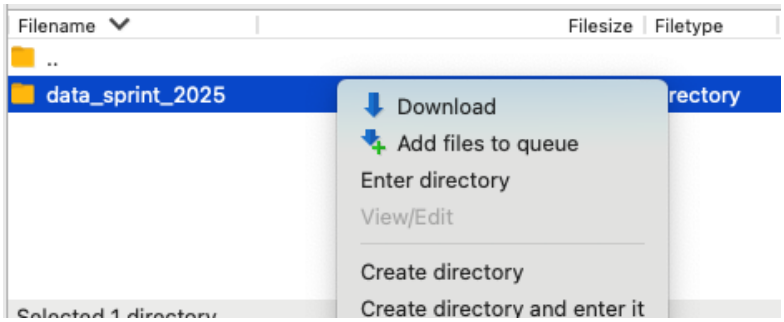
If the message below appears to you, just click ok:



3. Open the data_sprint_2025 folder to visualize the available datasets.

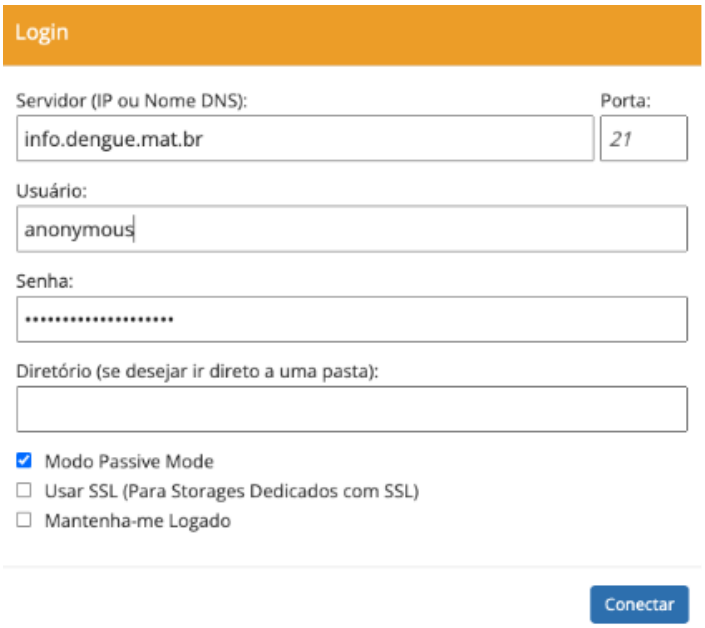


4. To download all the datasets in the folder (described in detail in this document) right-click on the folder and click on the Download option:



2 - Using FTPWeb

1. Open the link: https://www.ftpweb.com.br/index2.php and fill server= info.dengue.mat.br, user = anonymous and password= anonymous@domain.com.



2. Open the data_sprint_2025 folder to visualize the available datasets and download them.

Inside the data_sprint_2025 folder, there are the following files:

- Population: *datasus_population_2001_2024.csv.gz*,
- Environmental: *environ_vars.csv.gz*,
- Ocean temperature indicators: *enso.csv.gz, iod.csv.gz, pdo.csv.gz*,
- Shapefile of the cities: *shape_muni.gpkg*,
- Shapefile of the regional health divisions: *shape_regional_health.gpkg*,
- Shapefile of the macroregional health divisions: *shape_macroregional_health.gpkg*,
- Link between each city and its regional and health region and macroregion: *map_regional_health.csv*
- Weekly time series of dengue cases: *dengue.csv.gz*
- Weekly time series of climatic variables: *climate.csv.gz*
- Monthly time series of climate variable forecasts: *climate_forecast.csv.gz*

Each of these datasets is described in detail below.

Disease data

Period: epiweek 201001 to epiweek 202517[1].

Aggregation: cases aggregated by the epidemiological week of dengue symptom onset and by municipality.

File: *dengue.csv.gz*.

Sources: from SINAN and IBGE, organized by Infodengue.

Note: Data for the state of ES are not available due to reporting issues.

Table 1. Description of the columns in *dengue.csv.gz*

Column name	Type	Description
date	YYYY-MM-DD	First day of the epiweek (Sunday).
epiweek	int (YYYYWW)	Epidemiological week is defined by the date of symptom onset.
geocode	int	IBGE's municipality code.
casos	int	Number of cases per week, classified as probable dengue cases[2]. This column is equivalent to the column `casprov` in the infodengue table available in the mosqlimate API.
regional_geocode[3]	int	Health district code.
macroregional_geocode[3]	int	Health macroregion code.
uf	str	Federative Unit (state).
train_1	bool	Data for the first training (pre-season 22/23).
train_2	bool	Data for the second training (pre-season 23/24).
train_3	bool	Data for the third training (pre-season 24/25).
target_1	bool	Data for the first validation (season 22/23).
target_2	bool	data for the second validation (season 23/24).

Column name	Type	Description
target_3	bool	data for the third validation (season 24/25). This column does not go up to epiweek 202540 as this data has not yet been reported. However, please send the forecasts for the whole period ([EW 41 2024- EW40 2025]), since by the end of the challenge, the data will be reported, and the forecasts can be evaluated.

Climate — reanalysis

Reanalysis of hourly data from ERA5, summarized by week by the Mosqlimate project.

Period: epiweek 201001 to epiweek 202517[4].

Aggregation: temperature, humidity, and precipitation, originally by hour, were first aggregated by day (min, max, mean), and these daily measures were aggregated by epidemiological week (mean).

File: climate.csv.gz.

Sources: Copernicus ERA5, organized by Mosqlimate.

Table 2. Description of the columns of climate.csv.gz. The daily values of these variables are available in the mosqlimate API. *Atmospheric pressures are given as if the place were at sea level.

Column name	Type	Description
date	YYYY-MM-DD	First day of the epiweek (Sunday).
epiweek	int (YYYYWW)	Epidemiological week.
geocode	int	IBGE's municipality code.
temp_min	float (°C)	Minimum temperature.
temp_med	float (°C)	Mean temperature.
temp_max	float (°C)	Maximum temperature.
precip_min	float (mm/h)	Minimum precipitation rate.
precip_med	float (mm/h)	Average precipitation rate.

Column name	Type	Description
precip_max	float (mm/h)	Maximum precipitation rate.
precip_tot	float (mm)	Total precipitation.
pressure_min	float (atm)	Minimum daily sea level atmospheric pressure*.
pressure_med	float (atm)	Average atmospheric pressure*.
pressure_max	float (atm)	Maximum atmospheric pressure*.
rel_humid_min	float (%)	Minimum relative humidity.
rel_humid_med	float (%)	Average relative humidity.
rel_humid_max	float (%)	Maximum relative humidity.
thermal_range	float (°C)	Difference between the daily maximum and minimum temperature averaged by week
rainy_days	int	Number of days in the week for which `precip\_tot > 0.03`.

Climate Forecast

Seasonal forecasts (up to six months ahead) of climate variables from Copernicus, generated using System 51 by the ECMWF center.

Period: January 2010–April 2025.

File: *climate_forecast.csv.gz*.

Sources: Copernicus.

Table 3. Description of the columns of *climate_forecast.csv.gz*.

Column name	Type	Description
geocode	int	IBGE's municipality code.
reference_month	YYYY-MM-DD	Reference month.
forecast_months_ahead	int	The number of months into the future relative to the reference month for which the forecast is made.
temp_med	float (°C)	Mean temperature.

Column name	Type	Description
precip_tot	float (mm)	Total precipitation.
rel_humid_med	float (%)	Average relative humidity.

Ocean temperature and level oscillations

Period: 1993-01-04 — 2025-03-03 (weekly).

File: *ocean_climate_oscillations.csv.gz*.

Sources: https://sealevel.jpl.nasa.gov/.

Table 4. Description of the columns of *ocean_climate_oscillations.csv.gz*.

Column name	Type	Description
date	YYYY-MM-DD	Week (starting on Monday).
enso	float	**El Niño-Southern Oscillation** is a climate pattern in the Pacific Ocean that has two phases: El Niño and La Niña. In a normal year, in the Pacific Ocean, the trade winds blow westward along the Equator and push warm surface waters near Australia and Indonesia. On the other side of the Pacific Ocean, nutrient-rich cold waters come up off the coast of Central and South America, creating favorable conditions for fishing. During an El Niño event, the trade winds weaken, and warm, nutrient-poor waters are not pushed anymore by the winds, and sea level rises in the eastern tropical Pacific and falls in the western tropical Pacific. La Niña is the opposite phase of El Niño, with warm water piling up in the western Pacific and colder water in the eastern Pacific. This causes a higher sea level in the western tropical Pacific and a lower sea level in the eastern tropical Pacific.
iod	float	**The Indian Ocean Dipole.** Is a climate pattern affecting the Indian Ocean. During a positive phase, warm waters are pushed to the Western part of the Indian Ocean, while cold deep waters are brought up to the surface in the Eastern Indian Ocean. This pattern is reversed during the negative phase of the IOD.

Column name	Type	Description
pdo	float	**The Pacific Decadal Oscillation PDO.** It is a long-term (10-20 year) oscillation of the Pacific Ocean in response to the changes in the atmosphere. During a warm (positive) phase, the response of the ocean to low atmospheric pressure over the Aleutian Islands causes ocean currents to bring warm waters in the Eastern Pacific Ocean and along the coast of North America, and cool nutrient-rich waters in the western Pacific Ocean. This leads to higher sea levels along the coastlines of the Northeast Pacific. During a cool (negative) phase, the Eastern Pacific Ocean becomes cooler and the Western Pacific Ocean becomes warmer. This leads to lower sea levels along the coastlines of the Northeast Pacific.

Environmental data

Environmental characteristics of the municipalities. Other variables can be aggregated as necessary.

Period: 2010 (koppen) and 2024 (biome).

File: *environ_vars.csv.gz*.

Sources: IBGE, Embrapa.

Table 5. Description of the columns of *environ_vars.csv.gz*.

Column name	Type	Description
geocode	int	IBGE's municipality code.
uf_code	int	IBGE's state code.
koppen	str	main climate type
biome	str	main biome type .

Demographic data

Table 6. Geometry of cities in *shape_muni.gpkg* (source = IBGE).

Column name	Type	Description
geocode	int	IBGE's municipality code.
geocode_name	str	Municipality name.
uf	str	Two-letter state name.
uf_code	int	IBGE's state code.
geometry	geometry	municipality geometry.

Table 7. Geometry of the regional health divisions in *shape_regional_health.gpkg* (source = DATASUS).

Column name	Type	Description
regional_geocode	int	Regional health code.
regional_name	str	Regional health name.
uf_code	int	IBGE's state code.
geometry	geometry	Regional health geometry.

Table 8. Geometry of the macroregional health divisions in *shape_macroregional_health.gpkg* (source = DATASUS).

Column name	Type	Description
macroregional_geocode	int	Macrorregional health code.
macroregional_name	str	Macrorregional health name.
uf	str	Two-letter state name.
uf_code	int	IBGE's state code.
geometry	geometry	Macroregional health geometry.

Table 9. Link between each city and its regional and macroregional health center in *map_regional_health.csv* (source = IBGE).

Column name	Type	Description
macroregion_code	int	Macroregion code (1- Norte, 2- Nordeste, 3- Sudeste, 4 - Sul, 5 - Centro-Oeste).

Column name	Type	Description
macroregion_name	str	Macroregion name.
uf_code	int	IBGE's state code.
uf	str	Two-letter state name.
uf_name	str	State name.
macroregional_geocode	int	Macrorregional health code.
macroregional_name	str	Macrorregional health name.
regional_geocode	int	Regional health code.
regional_name	str	Regional health name.
geocode	int	IBGE's municipality code.
geocode_name	str	Municipality name.

Table 10. Population data (source: SVS). Files with population by city and year (2001 - 2024) in *datasus_population_2001_2024.csv.gz*

Column name	Type	Description
geocode	int	IBGE's municipality code.
year	int	Year (YYYY)
population	int	Population of the city.

Additional datasets

- **REGIC:** The publication Regiões de Influência das Cidades 2018 presents the research conducted to identify the hierarchy and areas of influence of Brazilian cities, describes the general characteristics of the detected urban network, and includes thematic analyses to highlight specific features of cities within this network. Available here: https://www.ibge.gov.br/geociencias/cartas-e-mapas/redes-geograficas/15798-regioes-de-influencia-das-cidades.html.

- **EPISCANNER:** This dataset contains estimates of the epidemiological parameters of reproduction number, peak week, epidemic duration, and size of the epidemic for all the Brazilian cities between 2010 and 2025. The methodology to obtain these estimates is available here: https://arxiv.org/abs/2407.21286. There is a dashboard showing the estimates: https://info.dengue.mat.br/epi-scanner/, and this data can be downloaded using the mosqlimate API: https://api.mosqlimate.org/docs/datastore/GET/episcanner/.

- **NVDI (Normalized Difference Vegetation Index) from BDC (Brazil Data Cube):** The BDC is a research, development, and technological innovation project of the National Institute for Space Research (INPE), Brazil. It is producing data sets from big volumes of medium-resolution remote sensing images for the entire national territory and developing a computational platform to process and analyze these data sets using artificial intelligence, machine learning, and image time series analysis. It has your library to get the images. A tutorial to get an image in Python and process it (based on a bbox region) is available here: https://github.com/brazil-data-cube/code-gallery/blob/master/jupyter/Python/stac/stac-image-processing.ipynb. There is also this tutorial: https://github.com/brazil-data-cube/code-gallery/blob/master/jupyter/Python/wtss/wtss-introduction.ipynb explaining how to retrieve a time series of these indicators for a specific latitude and longitude coordinates.

---

1. Note that the last weeks are subject to update as cases are still being reported. This data will be updated before the submission of the 2026 forecasts. ↩
2. Case definition: Probable cases = Suspected cases - discarded cases. ↩
3. Regional and Macroregional are the subdivisions used by the Ministry of Health. ↩ ↩
4. This data will be updated before the submission of the 2026 forecasts. ↩