Temperature sensitivity of dengue in Thailand

Thaddeus Kühn, Dana Levi, Frederik Racky, Fiona Tanner

Bioinformatics Project group 4 team 4 climate sensitive infectious diseases Summer Term 2023

Supervisor: **Dr. Stella Dafka**

The full abstract goes here

Contents

Abbreviations

DHF	Dengue Hemmorhagic Fever
ARIMA	Autoregressive moving average
GAM	General additive model
ADF	Augumented Dickey Fuller test
ACF	Autocorrelation function
pACF	Partial autocorrelation function
AIC	Akaike's information criterion

1 Introduction

The Dengue virus is a vector borne virus, consisting of four serotypes (DENV 1-4). It is transmitted to humans by mosquitoes, more specifically by *Aedes aegypti* and *Aedes albopictus* (Phanitchat *et al.*, 2019). The resulting infection goes asymptomatic in many cases, but can also cause Dengue Hemmorhagic Fever (DHF), characterized by symptoms like fever, headache, joint and muscle pain as well as (internal) bleeding and bruising (Gubler, 1998). Dengue is an emerging public health issue, with half of the worlds population at risk of infection. 390 million cases

per year are estimated worldwide, with most of them not showing symptoms and thus not being reported. Affected areas range from sub-tropical to tropical regions, with south-east Asia, including Thailand, being one of the most seriously affected regions. Newly affected areas also include Europe (WHO, 2023). The disease control of dengue is challenging due to the absence of an effective treatment or vaccine, which leaves only the treatment of symptoms with painkillers like paracetamol (WHO, 2023). Responsible institutions in affected areas currently focus on prevention, vector control, case control and prediction of possible future outbreaks (Phanitchat *et al.*, 2019).

An important factor in predicting the epidemiological dynamics of Dengue is the climate: climate fluctuations due to recurring weather phenomenons and climate change are shown to influence *Aedes* biology and infections (Descloux *et al.*, 2012; Phanitchat *et al.*, 2019). In several studies, maximal temperature has an effect on dengue transmission (Descloux *et al.*, 2012), being associated with higher incidence (Phanitchat *et al.*, 2019) between 27 °C and 29,5 °C. The temperature with the highest epidemic potential was shown to be 29,3 °C with a low temperature range throughout the day (Liu-Helmersson *et al.*, 2014). Extreme global climate events like el Niño have been shown to affect disease outbreaks like Dengue as well (Anyamba *et al.*, 2019).

In Thailand, a significant climate factor is the monsoon, which can be separated into two seasons: The south-west monsoon between may and october is characterized by higher temperature and high precipitation. The north-east monsoon between november and february is characterized by lower temperature and low precipitation (Kripalani *et al.*, 1995). It has been shown that in northern districts of Thailand, there has been a detectable rise in temperature since the mid 20th century (Masud *et al.*, 2016).

In this analysis, we are going to investigate the correlation between temperature and Dengue in Thailand from 2006 to 2020 to assess the significance of climate change, recurring climate fluctuations and extremes and geographical factors on Dengue infections. It is concentrated on three main points: Time periods with DHF incidence will be compared to time points of extreme weather events. It will be examined if provinces with higher temperature also show higher incidences. The observation of trends in temperature and dengue cases over the course of the given time period will be analysed and compared. Based on the findings, two models will be generated to forecast the development of dengue fever cases: Autoregressive moving average (ARIMA) will be used to model the near future. Then, Generalized Additive Model (GAM) will be generated to predict the spatial distribution of incidences over Thailand in the future.

2 Material and Methods

2.1 Material

ERA5 data (climate) The ERA5 database is a global climate database by the European Centre for Medium-Range Weather Forecasts (ECMWF), covering the earth in a 31 km horizontal grid up to 80 km in the atmosphere in the time period from 1950 to present. It was generated from measurements of various climate variables combined with a reanalysis of existing data and past reanalysis data to accurately model and complete the dataset in the given resolution (Hersbach *et al.*, 2020). For this project, monthly temperature data 2m above ground for every province in Thailand in the timeframe of 2006 - 2020 was extracted.

Dengue data The Dengue case numbers are retrieved from annual infectious disease reports published by the Thailand ministry of health. Monthly case numbers of DHF for every province in the timeframe of 2006 - 2020 are used in this project.

The datasets are used with a resolution at province level. As of 2011, Thailand has a total of 77 provinces, but had 76 provinces before 2011, as Bueng Kan was split from Nong Khai in 2011. For better compatibility of the data before and after 2011, the two new provinces are merged into a province equivalent to Nong Khai before 2011. This excludes data used for mapping, there the temperature and incidence values of Nong Khai are copied and used for Bueng Kan for compatibility with the spatial data.

Climate forecast For GAM modeling, temperature data from the CORDEX climate model is used. It includes the temperature 2m above ground for June to August (south west monsoon) of the years 2021 -2040 at a 22 km grid (Copernicus Climate Change Service, 2019).

Population data We used ... population data of every Thailand province from 2006 - 2020 (Quelle?).

Spatial data To associate our data with the different provinces and visualize it, we use spatial data of Thailand's provinces. The two main data types in spatial data are vector-data and raster-data. Vector-data consists of a list of points with their exact location, which can then form lines or polygons. Raster-data assigns a value to every square of a raster. In this case, maps consisting of polygons for each province are used (Pebesma and Bivand, 2023). With the sf package, objects associating our data for each province with its coordinates and polygons are created, describing its shape. The function geom_sf of ggplot2 are used for mapping.

2.2 Methods

Descriptive analysis Linear regression is a method to describe a linear relationship between variables, using the minimum sum of squares between regression line and data points to identify possible trends (Schneider *et al.*, 2010).

ARIMA To predict the development of the dengue cases with an ARIMA model, a time series with the total dengue cases of Thailand was created. A time series can be decomposed in the components trend, seasonality and random. A requirement for fitting an ARIMA model is a stationary time-series. This is obtained, when the mean value doesn't change over time, the variance doesn't increase and the seasonality effect is minimal (Prabhakaran, 2017). Two tests were used to test for stationarity of the data. The Augmented Dickey-Fuller (ADF) test examines whether the time series has a unit root, indicating non-stationarity. In the ADF test, the null hypothesis assumes the presence of a unit root, implying non-stationarity, while the alternative hypothesis suggests stationarity. The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test is also a unit root test but focuses on the presence of a deterministic trend in the series. In contrast to ADF-test, for the KPSS-test the null hypothesis assumes stationarity. If the time series is initially found to be non-stationary, the differences between consecutive observations can be calculated, and the stationarity tests can be applied again (Hyndman and Athanasopoulos, 2018). ARIMA models combine an autoregressive model AR(p) and a moving average model MA(q). The autoregressive model computes the current value from previous values and the error term:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \epsilon_t$$

ϵ_t = white noise
$1, \ldots, \phi_p$ = parameters
$y_{t-1}, \ldots, y_{t-p}$ = lagged values

For the moving average the current value consists of the mean value of the time series and weighted current and past error terms:

$$y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon t - 2 + ... + \theta_q \epsilon_t - q$$

$\theta_1, \ldots, \theta_q$= parameters

I(t) is the number of times differencing was performed to make the time series stationary (Venkat, 2018). To find the optimal values for p and q, the Autocorrelation function (ACF) and partial Autocorrelation function (pACF) were evaluated. The ACF plot shows the correlations of a time-series with lags of itself. The pACF calculates the relationship between a time series and

its lag, excluding the influence of linear dependencies among other lags (Prabhakaran, 2017). A second evaluation tool is the auto.arima function. The function automatically fits the best ARIMA model by minimizing the Akaike's Information Criterion (AIC), which is a criterion for the quality of the model. The auto.arima function can also consider seasonal models. Optimal models will yield uncorrelated residuals with zero mean and constant variance. This can be evaluated by plotting the ACF of the residuals and by performing a portmontreau test, for example the Ljung-Box test (Hyndman and Athanasopoulos, 2018).

GAM GAM provides insight into the shape and direction of the relationship between temperature and dengue cases. Additionally, it was used to forecast the future dengue case development based on the temperature development prediction. A GAM is a flexible extension of Generalized Linear Models (GLMs) that allows for the modelling of non-linear relationships between the response variable and predictor variables. A linear model can be described as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$$

GAMs are now a nonparametric form of regression where the linear predictors ($\beta_i x_i$) of the regression are replaced by smooth functions of the explanatory variables, $f(x_i)$. The model can be defined as: $y_i = f(x_i) + \in_i$ where y_i is the response variable, x_i is the predictor variable, and f is the smooth function (Wood, 2006). It is called an additive model, because all the $f(x_i)$ functions and therefore their predictor variables contribute individually to the response variable and are added up. The advantage is that the different smooth functions can capture complex relationships by flexibly fitting curves to the data. There are two different ways to interpolate the functions of the predictor values. Finding a polynomial with a specific degree that passes through all the data points is a polynomial interpolation approach. Because high-degree polynomials may result into wide oscillation or overfitting, piece-wise interpolation is sometimes more accurate. Here, the data is divided into smaller intervals, each one is described by an individual function. Defining the number of knots determines into how many segments the model is divided. All polynomials together are called splines with a degree of k. They connect the knots one by one. The spline can be differentiated k-1 times. A smaller number of knots makes the response smoother, while a higher k value results in a curve that closely follows the individual data points. Choosing from various types of splines, in the GAM smoothing splines were used, which try to fit the data closely as well as maintaining smoothness (Peri, 2021). The obtained model includes the relationship between incidence of dengue cases and temperature. Such GAMs can be used to forecast the development of the response variable (incidence of dengue fever) based on the predictor variables (temperature). The predicted values can than be plotted on a map of Thailand. The prediction is based on the average temperature for the months June to August for the south west monsoon over the time period of 2021 until 2040 in Thailand.

3 Results

3.1 Descriptive Analysis

3.2 ARIMA model

Time series decomposition The time series of total dengue cases in Thailand was additively decomposed in the components trend, seasonal and random, as shown in Figure 1. The time series shows a seasonal pattern with a frequency of 12 months. The trend shows fluctuation over the time period with two periods of comparatively high dengue cases in 2013/14 and 2015/16.

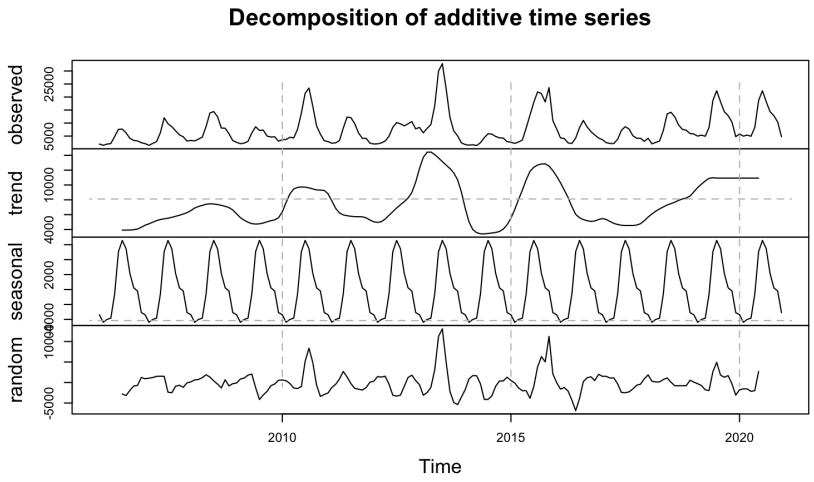

Figure 1: Decomposition of the additive time series of the total dengue cases in Thailand

The decomposition of the time series of the dengue cases was compared to the decomposition of the time series of temperatures. Both time series have an annual seasonal pattern. The trends show in general similar patterns with differences in the amplitudes and slight shifts on the time axis, as shown in Figure 2. In 2010 and 2012/13, there were temperature peaks, followed by peaking dengue cases a few months after. In 2015/16 rising temperatures were accompanied by rising dengue cases.

ARIMA forecast ARIMA modeling was performed to forecast the dengue case development in Thailand. For ADF test results, a p-value of 0.01 was used, for the KPSS test results a p-value of 0.1 was used. Therefore both test indicate stationarity of the time series and differencing isn't necessary. Observation of the time series ACF showed oscillations which suggests the presence of seasonality in the data. The pACF exhibits a notable pattern with an initial positive peak followed by a significant negative peak. This pattern indicates the need to consider both autoregressive and moving average components in the model. To test for the optimal ARIMA model, models with different combinations of p, d and q values were generated and the best model

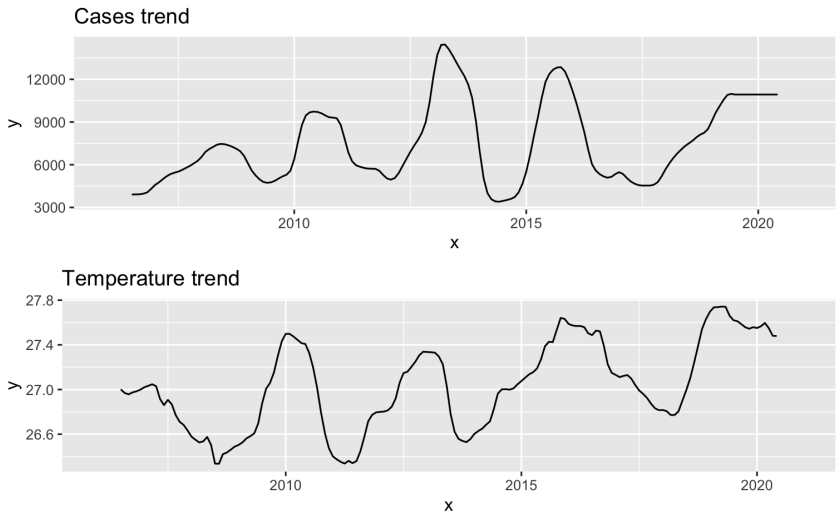

Figure 2: Trend of the total dengue cases in Thailand compared to the trend of the average temperature in Thailand for the years 2006 till 2020.

was selected by comparing the AICs. The best model was found to be (2,2,2) with an AIC of 3319.732. Because this method could not account for seasonality in the data, the auto.arima function was used. "ARIMA(1,0,2)(1,1,0)[12] with drift" was found to be the best model. This model consists of an autoregressive term and two moving average terms. No differencing was performed. The model also has seasonal compounds with a period of 12. "With drift" means that a constant is included in the model. As the AIC of 3108.35 is lower and thus more accurate than the AIC of the former model, the auto.arima model was used for further analysis. The evaluation of the histogram of the residuals revealed that the residuals were distributed similar to a normal distribution. The ACF plot indicates some significant autocorrelation at lag 1, while the remaining lags show no significant autocorrelation. The p-value of the Ljung-Box test was 0.41. Thus, the residuals are not distinguishable from white noise and the model has adequately captured the information in the data. Based on the model, a forecast was made for the next four years, which extends beyond the data (see Figure 3).

To compare the accuracy of the forecast, the time series was cropped after December 2016 and the years 2017 till 2020 were forecasted, as shown in Figure 4.

Although the forecast is more uniform than the actual cases in th years 2017 to 2020, the reported cases are in the range of deviation shown in grey.

3.3 GAM model

A GAM model was generated based on the monthly temperature and incidence of all provinces of Thailand over the years 2006 - 2020. The predictor value (temperature) was passed into a smoothing function to create a smoothed spline curve. For the response variable (incidence) a quasi-Poisson distribution was chosen. The link function is a log transformation. To figure

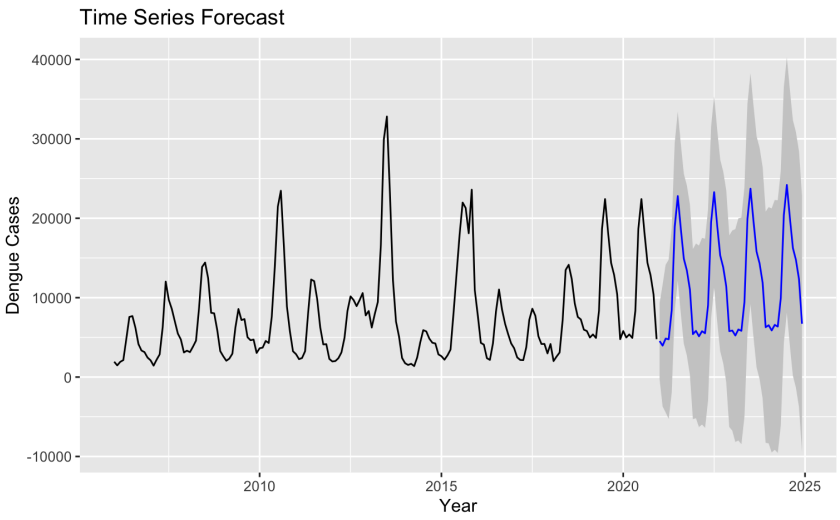

Figure 3: ARIMA forecast of the dengue cases in Thailand for 2021-2024.

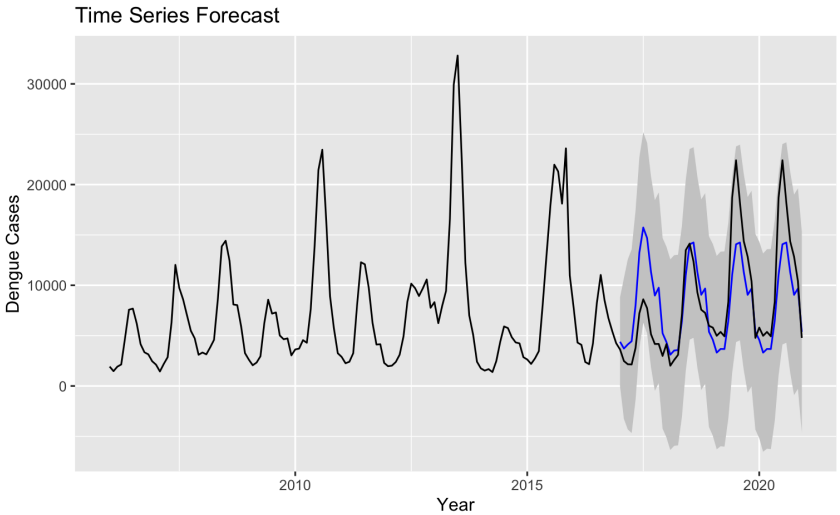

Figure 4: ARIMA forecast of the dengue cases in Thailand for the years 2017-2020 (blue) compared to the actual dengue cases (black).

out the optimal model, the degrees of freedom of the smoothing function were alternated. The effective degrees of freedom, which were found to be 13.82, resemble the complexity of the smooth curve. This value is relatively high, which indicates a "wigglier" spline. F-value and p-value indicate the statistical significance of the smooth function. In this case, the smooth function of temperature is highly significant (p-value $<$ 2e-16). The computed R-squared value of 0.0465 indicates how well the model explains the variance in the dependent variable. Here, the model explains approximately 4,7of the variance in the data. A low GCV (Generalized Cross Validation) value suggests a good fit of the model. The generated model has a GCV of 11.77. Additionally, the AIC value of the smooth model (112947.1) was compared with that of a linear model (113361.2). The smooth model had a lower value, which indicates a better fit. Taking all these parameters into account, the GAM was computed with a k of 16. The model and its relationship of temperature and incidence is shown in the appendix Figure ..., as well as the evaluating graphs. Subsequently the model was used to predict the average dengue case incidences over the period of 2021 until 2040, based on forecasted average temperature. You can see the average incidences of the past and future below.

4 Discussion

Our results show ...

References

Anyamba, A., Chretien, J.-P., Britch, S. C., Soebiyanto, R. P., Small, J. L., Jepsen, R., Forshey, B. M., Sanchez, J. L., Smith, R. D., Harris, R., Tucker, C. J., Karesh, W. B., and Linthicum, K. J. (2019). Global Disease Outbreaks Associated with the 2015–2016 El Niño Event. Scientific Reports *9*, 1930.

Copernicus Climate Change Service (2019). CORDEX regional climate model data on single levels.

Descloux, E., Mangeas, M., Menkes, C. E., Lengaigne, M., Leroy, A., Tehei, T., Guillaumot, L., Teurlai, M., Gourinat, A. C., Benzler, J., Pfannstiel, A., Grangeon, J. P., Degallier, N., and de Lamballerie, X. (2012). Climate-Based Models for Understanding and Forecasting Dengue Epidemics. PLOS Neglected Tropical Diseases *6*, e1470, doi: 10.1371/JOURNAL.PNTD.0001470.

Gubler, D. J. (1998). Dengue and dengue hemorrhagic fever. Clinical Microbiology Reviews *11*, 480–496.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G. D., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J. N. (2020). The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society *146*, 1999–2049, doi: 10.1002/QJ.3803.

Hyndman, R., and Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd ed). OTexts: Melbourne, Australia. https://otexts.com/fpp2/.

Kripalani, R. H., Singh, S. V., Panchawagh, N., and Brikshavana, M. (1995). Variability of the summer monsoon rainfall over Thailand—comparison with features over India. International Journal of Climatology *15*, 657–672.

Liu-Helmersson, J., Stenlund, H., Wilder-Smith, A., and Rocklöv, J. (2014). Vectorial Capacity of Aedes aegypti: Effects of Temperature and Implications for Global Dengue Epidemic Potential. PLoS ONE *9*, doi: 10.1371/JOURNAL.PONE.0089783.

Masud, M. B., Soni, P., Shrestha, S., and Tripathi, N. K. (2016). Changes in climate extremes over North Thailand, 1960–2099. downloads.hindawi.com , 1960–2099.

Pebesma, E., and Bivand, R. (2023). Spatial Data Science: With Applications in R (Chapman and Hall/CRC).

Peri, S. P. (2021). GAMs and Smoothing Splines. Towards AI .

Phanitchat, T., Zhao, B., Haque, U., Pientong, C., Ekalaksananan, T., Aromseree, S., Thaewnongiew, K., Fustec, B., Bangs, M. J., Alexander, N., and Overgaard, H. J. (2019). Spatial and temporal patterns of dengue incidence in northeastern Thailand 2006–2016. BMC Infectious Diseases *19*, 743.

Prabhakaran, S. (2017). Time Series Analysis With R. r-statistics.co .

Schneider, A., Hommel, G., and Blettner, M. (2010). Linear Regression Analysis: Part 14 of a Series on Evaluation of Scientific Publications. Deutsches Ärzteblatt International *107*, 776.

Venkat, A. (2018). Time Series Analysis for Epidemiological Data.

WHO, W. H. O. (2023). Dengue and severe dengue, https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue.

Wood, S. (2006). Generalized Additive Models: An Introduction With R, vol. 66.