# ARIMA Modelling

To predict the development of the dengue cases with an Auto-Regressive Integrated Moving Average model (ARIMA), 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. The null hypothesis assumes the presence of a unit root, implying non-stationarity, while the alternative hypothesis suggests stationarity. On the other hand, 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, 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:

yt=c+ϕ1yt−1+ϕ2yt−2+⋯+ϕpyt−p+εt

εt = white noise  
 ϕ1,…,ϕp = parameters  
yt-1,…, yt-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:

yt=c+εt+θ1εt−1+θ2εt−2+⋯+θqεt−q

θ1,…, θq = parameters

I(d) 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, Athanasopoulos, 2018).

Quellen:

Selva Prabhakaran (2017), Time Series Analysis, r-statistics.co

Aishwarya Venkat (2018), Time Series Analysis for Epidemiological Data, <https://rstudio-pubs-static.s3.amazonaws.com/354672_b7cb732a2b61469390d6fc72621bc9c4.html>

Hyndman, R.J., & Athanasopoulos, G. (2018) *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2. Accessed on 12.07.2023

# Results – Decomposition of the time series

The time series of total dengue cases in Thailand was additively decomposed in the components trend, seasonal and random, as shown in Figure x . The time series shows a seasonal pattern at regular intervals of 12 months. The trend shows fluctuation over the time period with two periods of comparatively high dengue cases.

*Abb. der decomposition  
Bildunterschrift: Figure X: Decomposition of the time series of the total dengue cases in Thailand.*

The composition 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 y.

*Abb. der trends decomposition  
Bildunterschrift: Figure Y: Trend of the total dengue cases in Thailand compared to the trend of the average temperature in Thailand for the years 2006 till 2020.*

# Results – ARIMA

ARIMA modelling was performed to predict the development of the total dengue cases in Thailand. The ADF test results in a p-value of 0.01. The KPSS test results in a p-value of 0.1. Therefore both test indicate stationarity of the time series and differencing isn’t necessary. Observation of the ACF of the time series 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 AR and MA 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 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 couldn’t 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 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 residual 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 extend beyond the data. The forecast is shown in figure Z.

*Abb. Des Auto Arima forecast  
Bildunterschrift: Figure Z: ARIMA forecast of the dengue cases in Thailand for 2021-2024.*

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 A.

*Abb. Des ARIMA 2016*

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

<http://r-statistics.co/Time-Series-Analysis-With-R.html?utm_content=cmp-true>

2017 Selva Prabhakaran, Time Series Analysis, r-statistics.co

* Stationary
* acf pacf

<https://www.analyticsvidhya.com/blog/2021/06/statistical-tests-to-check-stationarity-in-time-series-part-1/#How_to_Check_Stationarity>?

* ADF und KPPS

<https://rstudio-pubs-static.s3.amazonaws.com/354672_b7cb732a2b61469390d6fc72621bc9c4.html>

Time Series Analysis for Epidemiological Data

Aishwarya Venkat

2018-02-01

AR und MA

Hyndman, R.J., & Athanasopoulos, G. (2018) *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2. Accessed on 12.07.2023

für später:

When data have a trend, the autocorrelations for small lags tend to be large and positive because observations nearby in time are also nearby in size. So the ACF of trended time series tend to have positive values that slowly decrease as the lags increase.

When data are seasonal, the autocorrelations will be larger for the seasonal lags (at multiples of the seasonal frequency) than for other lags.

a significant decline in the ACF and a small or negligible partial autocorrelation beyond a few lags, it suggests the possibility of an MA (Moving Average) component in your time series.

The presence of a sine pattern in the ACF suggests the presence of seasonality in the data. at regular intervals. This pattern may be observed as oscillations in the ACF, where the autocorrelation values show a repeating up-and-down pattern. The increase in ACF values at lag 1 suggests a strong positive autocorrelation between adjacent observations. This indicates that the value at time t is highly influenced by the value at time t-1.

2 It suggests that there is a direct relationship between observations at those lags, even after accounting for the influence of other intermediate lags. The smaller pACF values for the remaining lags indicate a weaker or negligible direct relationship between observations at those lags.

The ACF plot is used to identify the potential MA order by examining the decay pattern. If the ACF shows a significant decline and cuts off after a few lags, it indicates that an MA term may be appropriate.

the presence of significant high values in the initial lags of the pACF indicates the need to consider autoregressive terms in the model. The specific order or combination of autoregressive and moving average terms would require further analysis, such as model diagnostics and statistical tests.

The presence of both positive and negative significant values in the initial lags of the pACF suggests that both autoregressive and moving average components might be necessary to adequately model the time series. The positive autocorrelation at the first lag indicates the need for an autoregressive term to capture the dependence on previous observations, while the negative autocorrelation at the second lag suggests the inclusion of a moving average term to capture the influence of the residual errors from those observations.

Thaddl

To generate a GAM the temperature and incidence of all provinces of Thailand and all months of the years 2006 until 2020 was used vllt a Gam was modeld / deveoped based on. 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. cool

To figure out the optimal model the degrees of freedom of the smoothing function was alternated. The effective degrees of freedom, which was found to be 13.82, resembles the complexity of the smooth. In this case it is a relative high value, 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. In this case, the model explains approximately 4,7% of 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). -> to show: Mehrwert

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, as well as the evaluating graphs. Nice

Subsequently the model was used to predict the average dengue case incidences over the period of 2021 until 2040, based on forecasted average temperature. Voll gut

Noch checken: komma nach to…