COVID-19: PREDICTION OF THE HOSPITALIZATION RATE

Statistisches Praktikum
WiSe21/22

16.03.2022
Munich

Project partner: Yeganeh Khazaei
Project „Betreuer“: André Klima
Statistisches Beratungslabor StaBLab der LMU
Institut für Statistik

Group:
Alexander Marquard
Phu Nguyen
Qian Feng

AGENDA

BACKGROUND INFORMATION 01

02 DATA PROCESSING

DATA ANALYSIS 03

04 EXCURSION: TIME SERIES

MODEL INTRODUCTION 05

01

BACKGROUND INFORMATION

BACKGROUND INFORMATION

- Background:

  Meaningful evaluations of the data base and determination of measures (such as the reproduction number, incidence, or hospitalization incidence) serve as guiding criteria for measures against the further spread of the virus.

- Task at hand:

  Predict hospitalization rate one to two weeks in the future, taking into account both time and geographical factors.

DEFINITIONS

HOSPITALIZATION (RATE)

The number of COVID-19 patients admitted for treatment (per 100,000 population) in a given time period.

$$= \frac{\text{Number of hospitalizations}}{\text{Respective population}} \cdot 100.000$$

DEFINITIONS

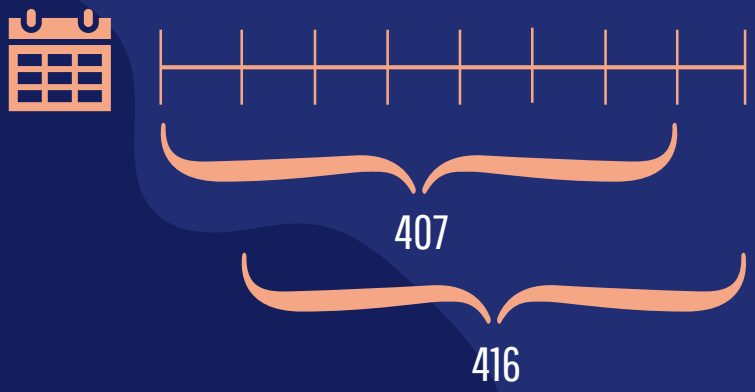
NOWCASTING

Problem: Delays in reporting

- ➔ New hospitalizations being reported daily do not reflect actual numbers
- ➔ **Nowcast**: An estimate is provided at the current point in time by using a statistical procedure

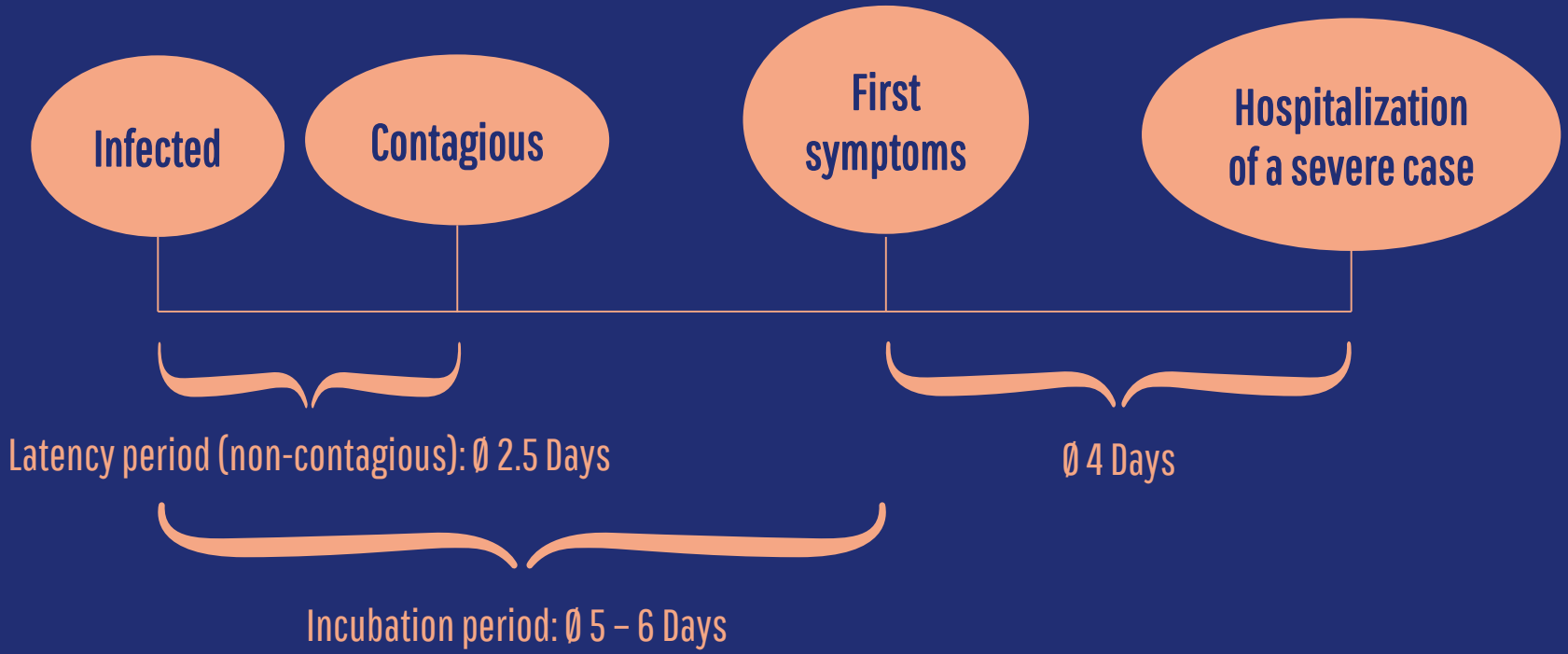
DEFINITIONS

DELAYED REPORTS

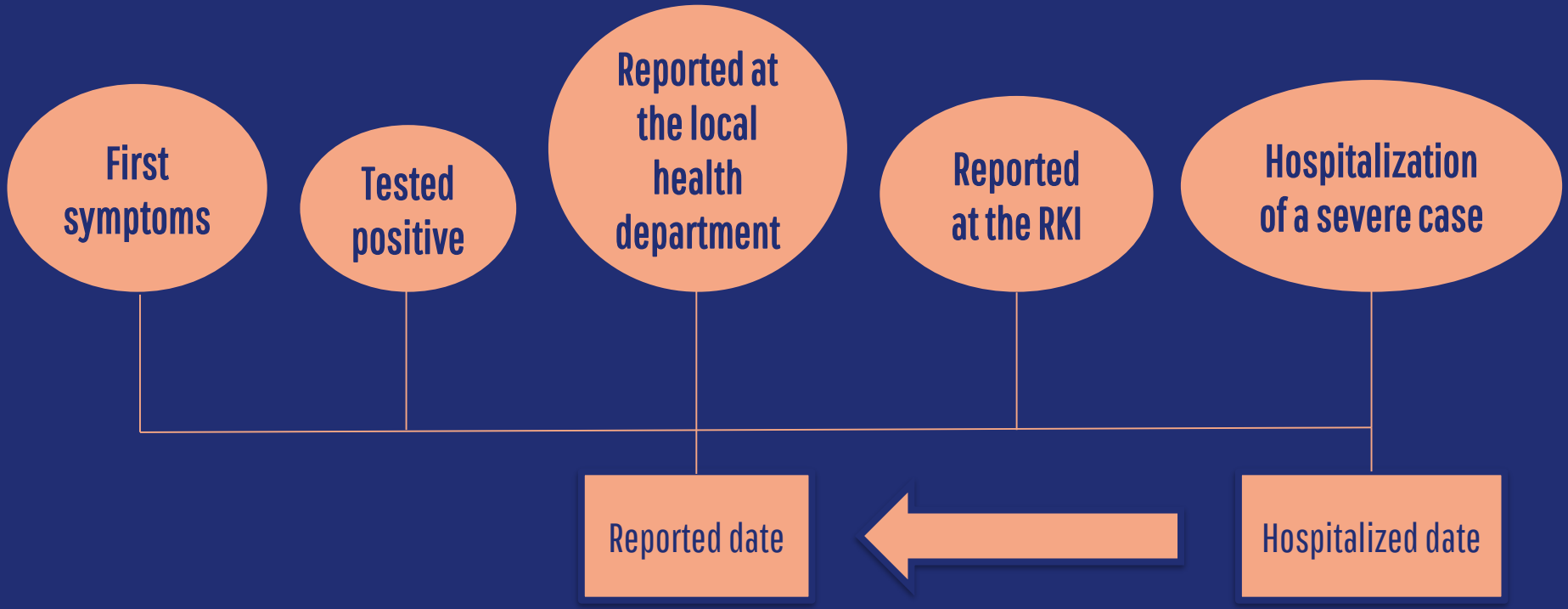


	Reported	Updated
01.08.2021	407	756
02.08.2021	416	786

TYPICAL COVID-19 PROGRESSION



BUREAUCRATIC REPORTING PROCESS OF A COVID-19 CASE



WHAT IS ACTUALLY PREDICTED THEN?

The **updated** hospitalization rate

- on the reported date of infection
- for the next two weeks

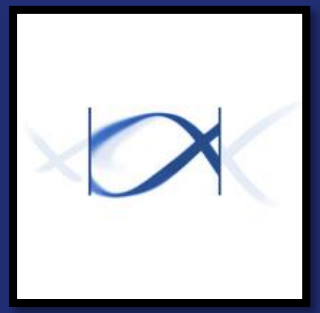
02

DATA PROCESSING

DATA COLLECTION

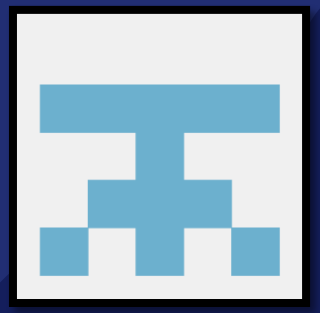
RKI

- New cases
- Hospitalizations
- Vaccination



KITMetricslab

- Population



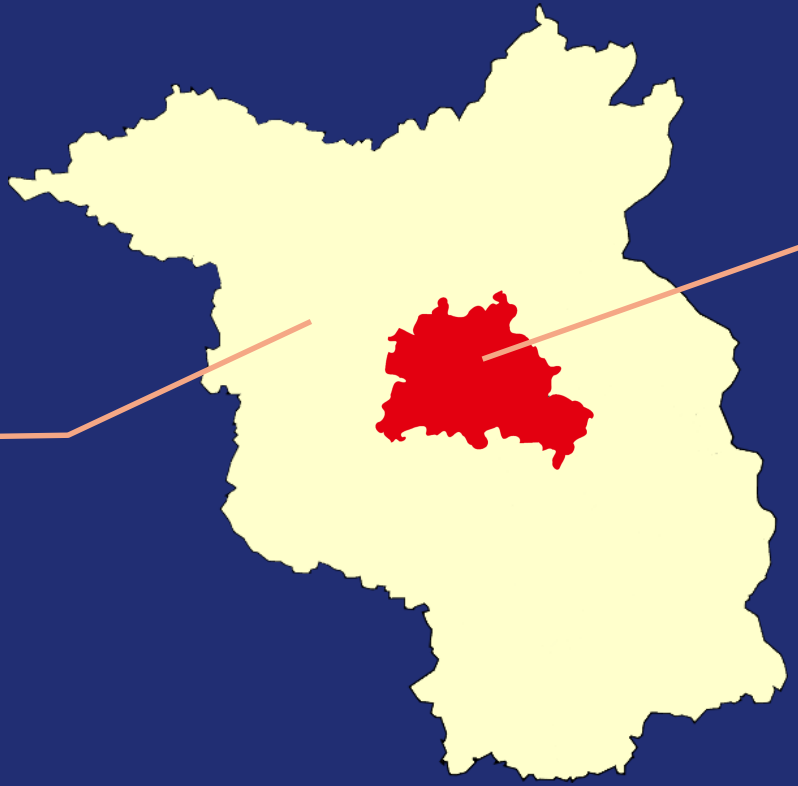
FINALIZED DATA SET

Reported date	Year	Calendar week	Index	State	Age group	Population	New cases	Hospitalizations
2020-03-01	2020	9	9	Bayern	80+	871866	0	0
2020-03-08	2020	10	10	Bayern	80+	871866	3	2
2020-03-15	2020	11	11	Bayern	80+	871866	39	25
2020-03-22	2020	12	12	Bayern	80+	871866	214	132
2020-03-29	2020	13	13	Bayern	80+	871866	759	383

03

DATA ANALYSIS

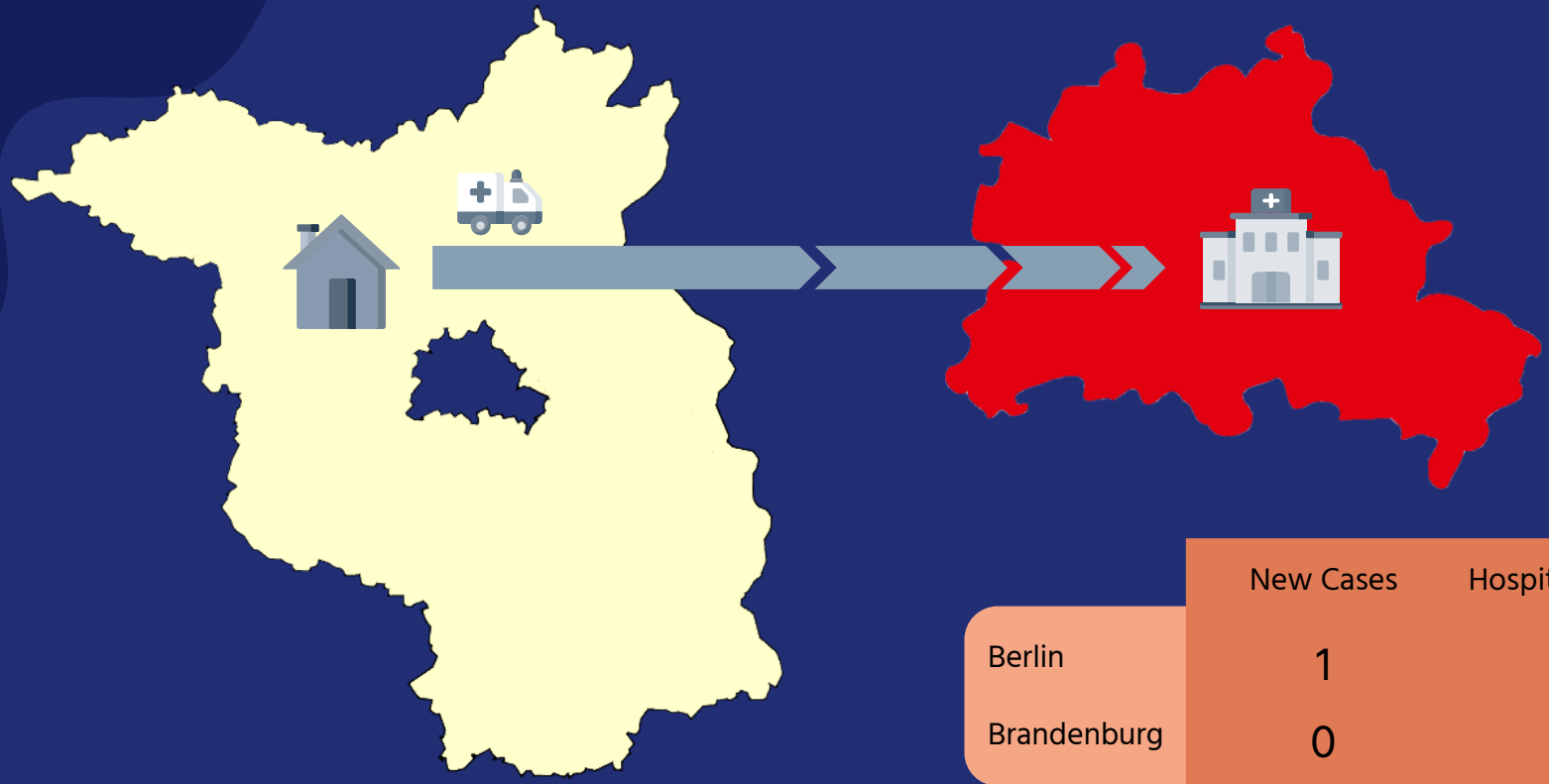
PROBLEMS



Berlin

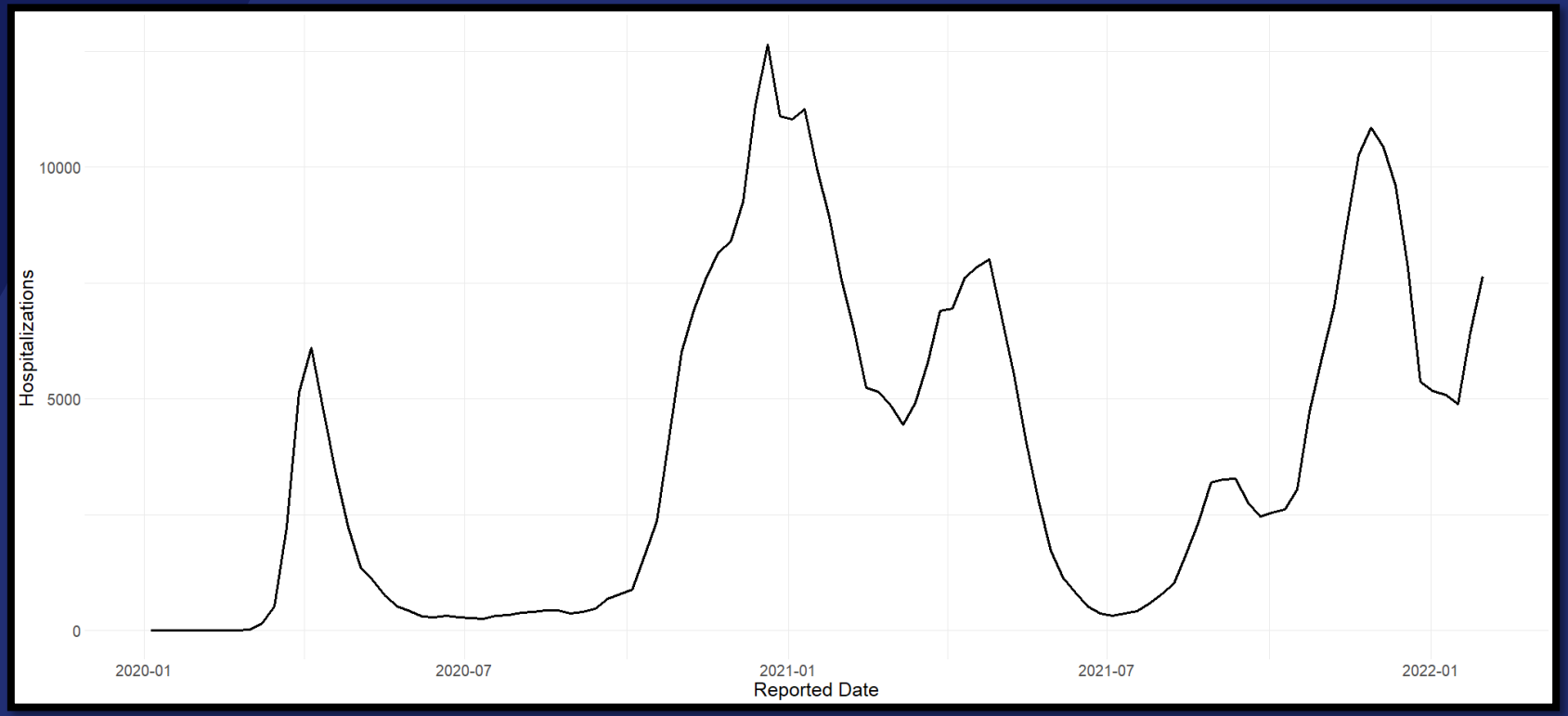
Brandenburg

PROBLEMS

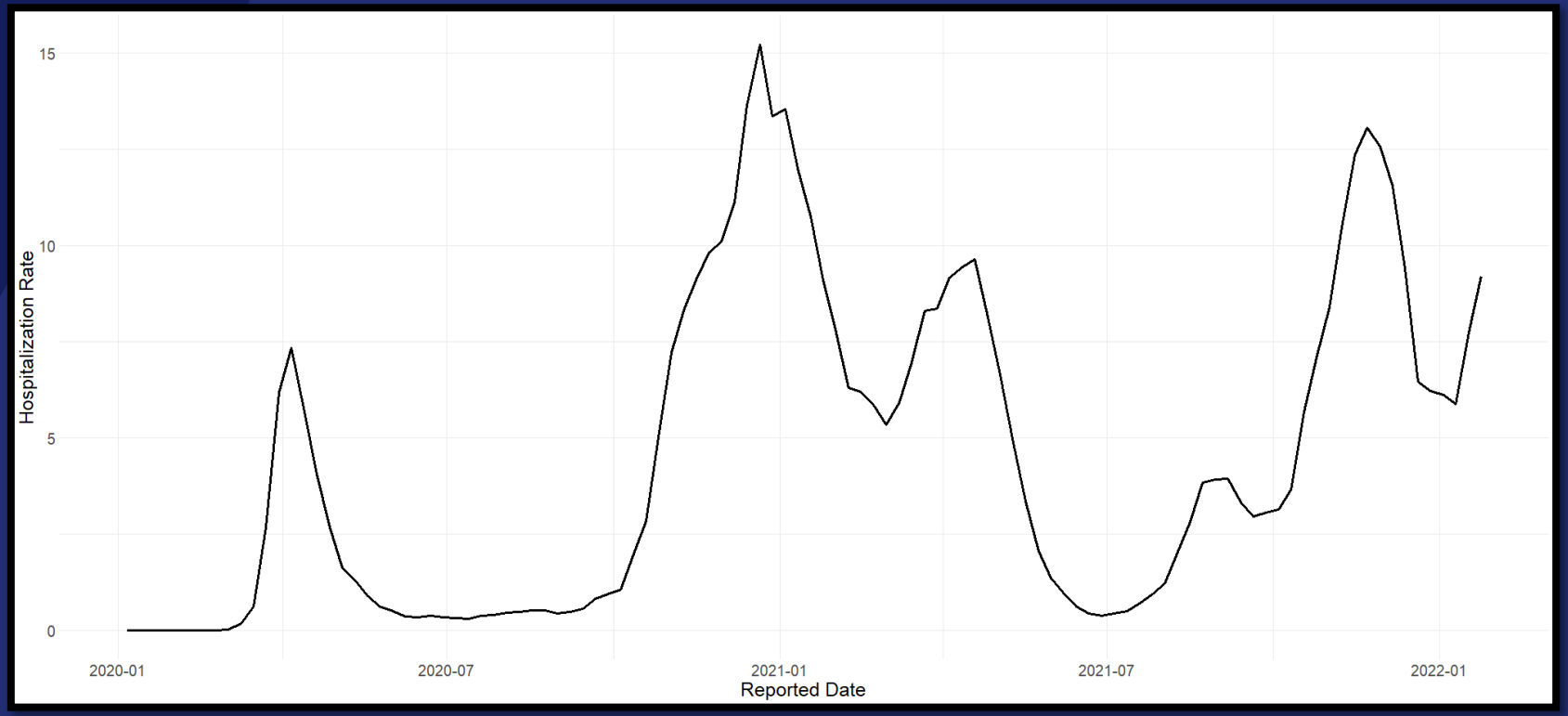


	New Cases	Hospitalizations
Berlin	1	0
Brandenburg	0	1

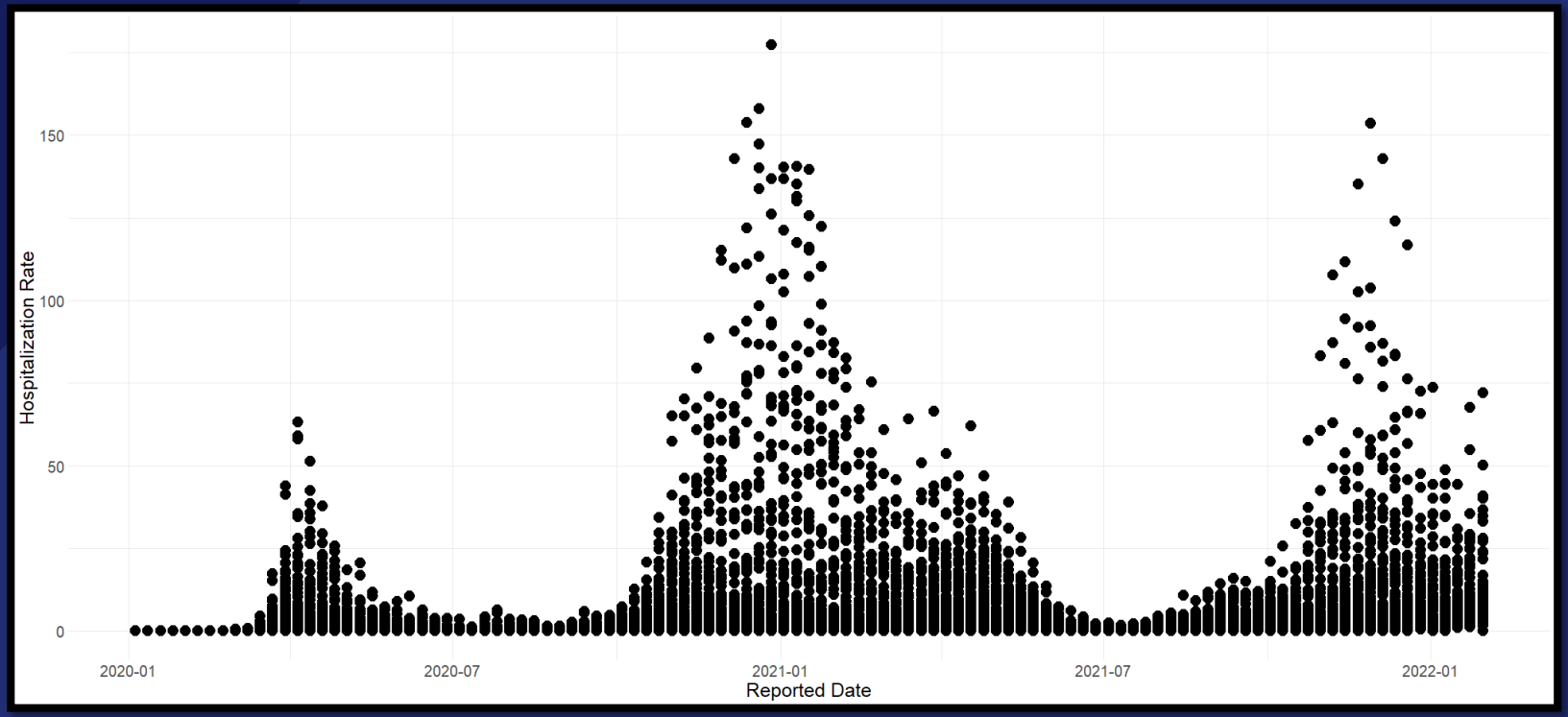
ABSOLUTE NUMBERS OF HOSPITALIZATIONS



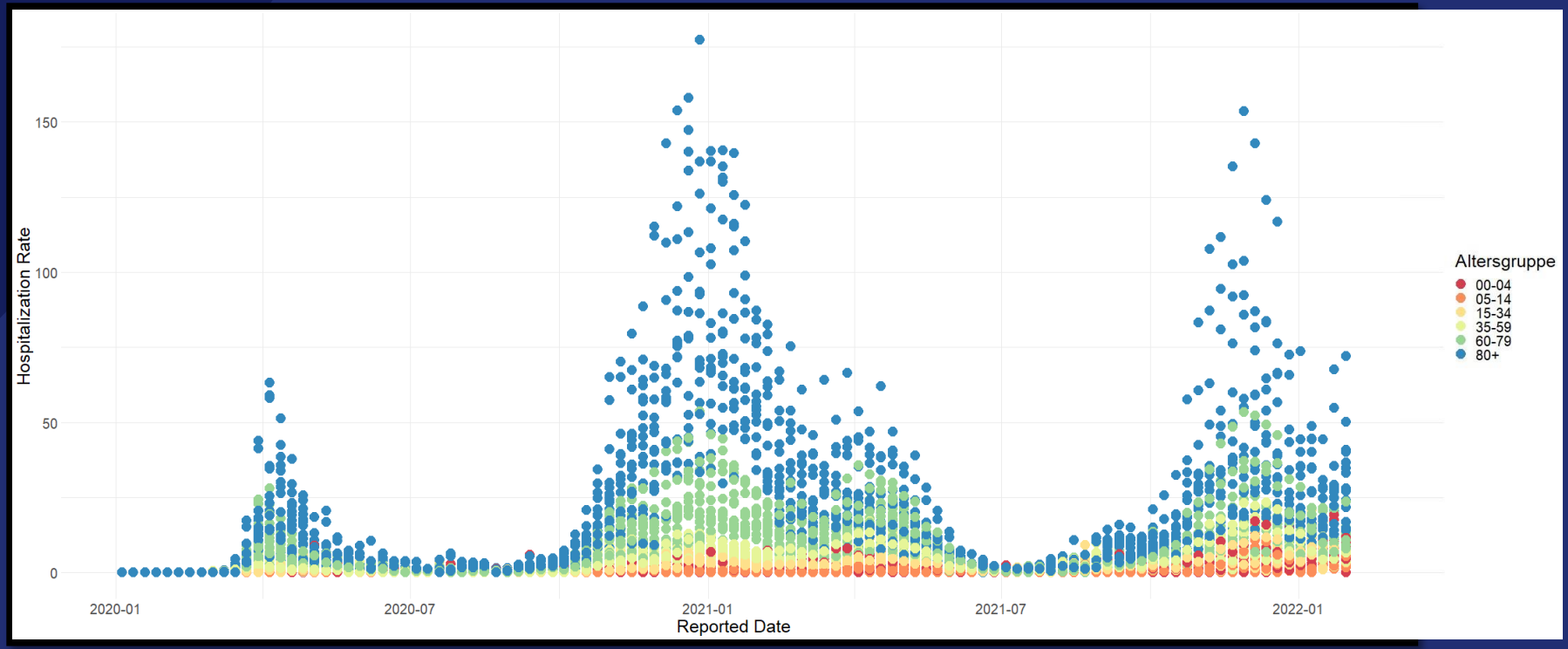
NUMBERS REPORTED BY THE RKI



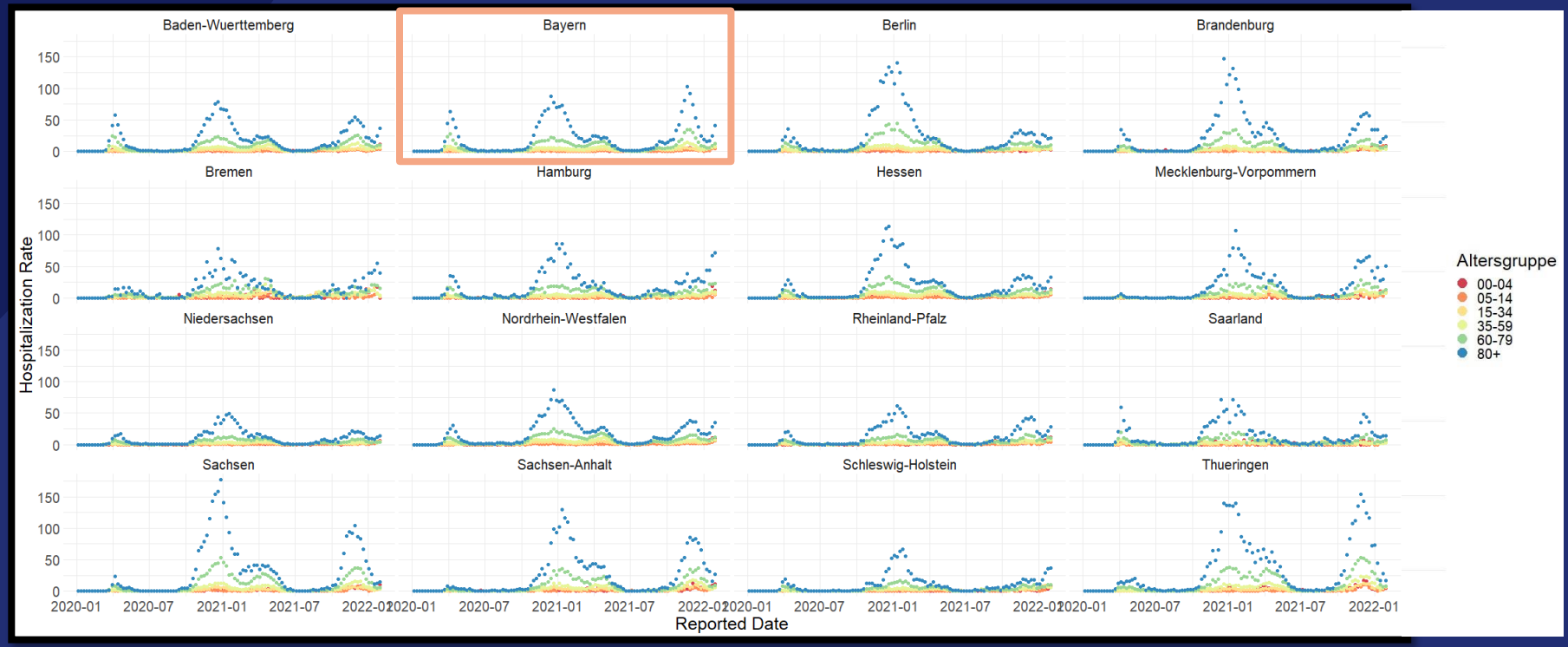
EXPLAINING THE HOSPITALIZATION RATE



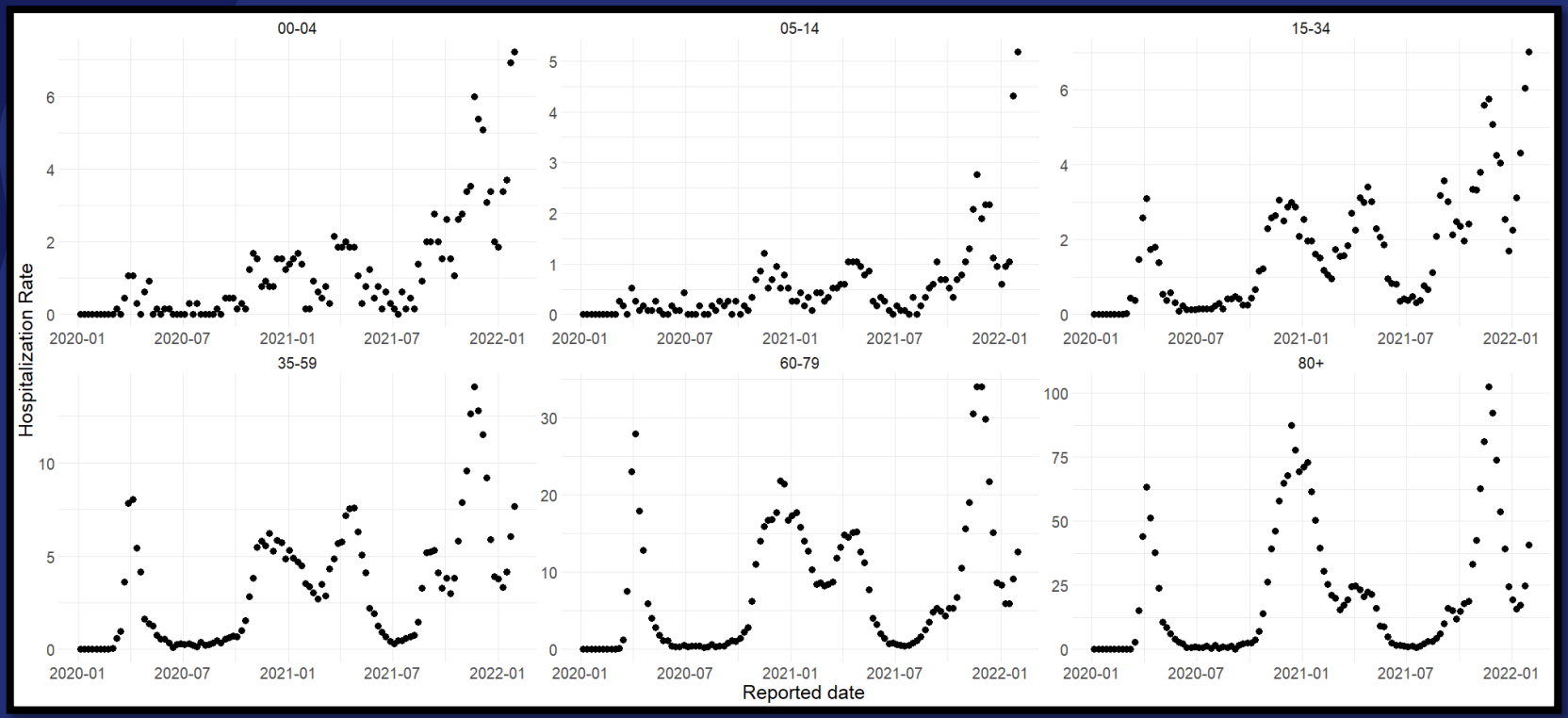
HOSPITALIZATION RATE BROKEN DOWN INTO AGE GROUPS



EXPLAINING THE HOSPITALIZATION RATE



HOSPITALIZATION RATE BY AGE GROUP IN BAYERN



ANALYZING BAYERN BY USING GAM

Distribution assumption: $\text{Hospitalization}_i \mid x_i \sim \text{Poi}(\lambda)$

Link: Log

$$E(\text{Hospitalization}_i) = \exp(\beta_0 + \text{Agegroup}_i + f(\text{index}_i, \text{Agegroup}_i))$$

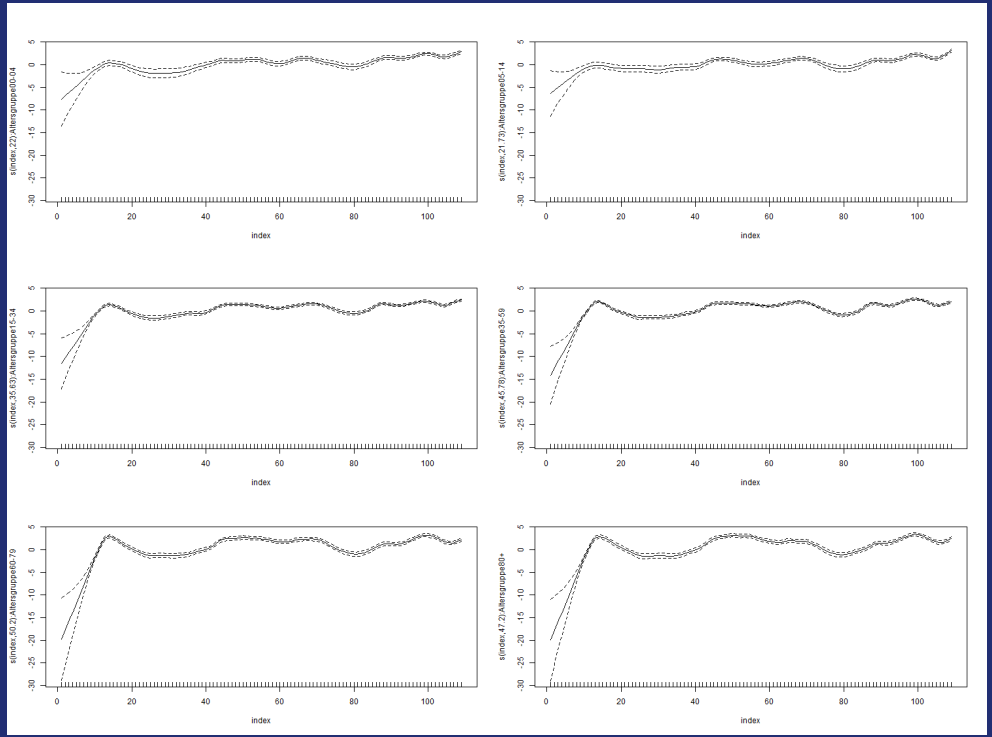
Intercept:	exp(1.06) ~ 3
05-14:	exp(-0.03) ~ 1
15-34:	exp(1.93) ~ 7
35-59:	exp(2.77) ~ 16
60-79:	exp(2.58) ~ 13
80+:	exp(2.3) ~ 10

⇨

00-04 and 05-14 are the least impactful age groups.
This holds true for each individual state.

ANALYZING BAYERN BY USING GAM

The effect of the time varying covariable given the age groups stays nearly the same for each age group. This holds true for each individual state



CONCLUSION

For each age group:

$$E\left(\text{Hospitalization}_{t,j}\right) \propto c_j * \theta_t$$

$$E\left(\text{Hospitalization}_{t+h,j}\right) \propto c_j * \theta_{t+h}$$

$$\Rightarrow E\left(\text{Hospitalization}_{t+h,j}\right) \propto \frac{E(\text{Hospitalization}_{t,j})}{\theta_t} * \theta_{t+h}$$

since $\frac{\theta_{t+h}}{\theta_t}$ is independent from j

$$\Rightarrow X_{t+h,j} \propto \theta * X_{t,j}$$

$$j \in \{1, \dots, 6\}$$

Or more general:

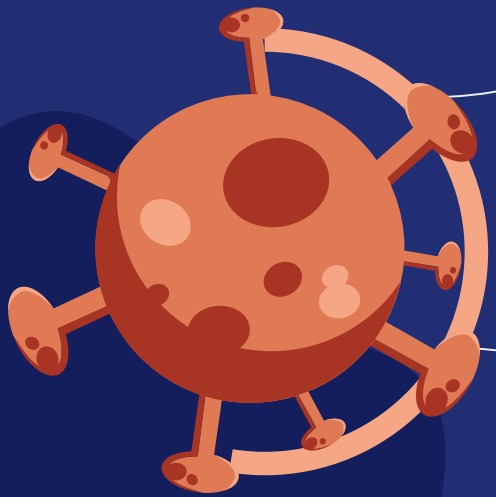
$$X_{t+h} \propto \sum_{j=1}^{6} \theta * X_{t,j}$$

$$\Rightarrow X_{t+h} \propto \theta * X_t$$

04

EXCURSION: TIME SERIES

TIME SERIES



Stationarity

Does the time series show a clear trend or is it stable over the course of time

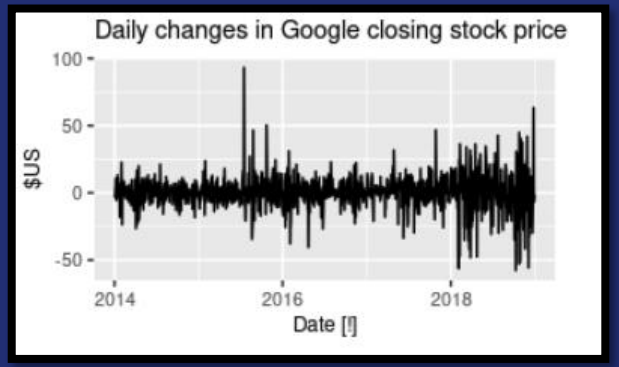
Seasonality

Is there a seasonal pattern?

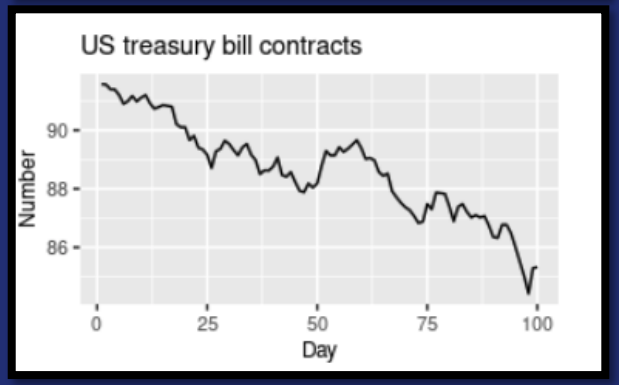
Such as the time of the year or the day of the week?

TIME SERIES (EXAMPLES)

A stationary time series



A nonstationary time series



TIME SERIES

- Every realisation $y_1,\ldots, y_n$ can be seen as a series of random variables $Y_1,\ldots,Y_n$.
- If the (joint) distribution of $Y_1,\ldots, Y_n$ is known, one can predict every realisation of these random variables

3 IMPORTANT ASPECTS OF THE TIME SERIES

Mean

$\mu_t = E(Y_t) = \int_{-\infty}^{\infty} y \cdot f_t(y)\, dy$

With t time and μ_t mean of each random variable $Y_1,\ldots,Y_n$

Covariance

$\gamma(s,t) = Cov(Y_s, Y_t) = E(Y_s - \mu_s)(Y_t - \mu_t)$

Is the covariance function.

It depends on the two timestamps s and t

Correlation

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\cdot\gamma(t,t)}}$$

Is the correlation function

LINEAR TIME SERIES

$$Y_t = \psi_0 \cdot a_t + \psi_1 \cdot a_{t-1} + \ldots$$

$a_t \sim N(0, \sigma^2)$
$t = \ldots, -1, 0, 1, 2, \ldots$
ψ_i weights, $i = 0, 1, 2, \ldots$

$$\sum_{i=0}^{\infty} \psi_i^2 < \infty$$

Backshift-Operator

$Bx_t = x_{t-1}$; $Bx_{t-3} = x_{t-4}$
$B^2x_t = x_{t-2}$; $(1 - B)x_t = x_t - x_{t-1}$
$(1 - B)^2x_t = (1 - 2B + B^2)x_t = x_t - 2x_{t-1} + x_{t-2}$

IMPORTANT ASPECTS OF THE LINEAR TIME SERIES

ACF/Autocorrelation - $\rho(h)$

Measurement of information that Y_t provides to estimate Y_{t+h}

PACF/Partial autocorrelation - $\tau(h)$

Measurement of information that Y_t provides to estimate Y_{t+h} given $Y_{t+1}, \ldots, Y_{t+h-1}$

DIFFERENT TYPES OF TIME SERIES

- Moving average models - MA(q)
- Autoregressive models - AR(p)
- Autoregressive moving average models - ARMA(p,q)

Suitable for stationary time series

- Autoregressive integrated moving average models - ARIMA(p,d,q)
- Autoregressive integrated moving average models with seasonal component - ARIMA$(P,D,Q)_s$

Suitable for non-stationary time series

EXAMPLE: AUTOREGRESSIVE MODEL - AR(p)

AR(p) model is defined by:

$$Y_t = \varphi_1 \cdot Y_{t-1} + \varphi_2 \cdot Y_{t-2} + \dots + a_t \qquad \text{(with } t = 1, 2, \dots\text{)}$$

$$\Leftrightarrow \quad (1 - \varphi_1 \cdot B - \varphi_2 \cdot B^2 \dots) + Y_t = a_t$$

AR(1):

$$Y_t = \varphi \cdot Y_{t-1} + a_t \qquad \text{(with } t = 1, 2, \dots\text{)}$$

$$\Leftrightarrow \quad y_t = b_0 + b_1 \cdot x_t + \varepsilon_t$$

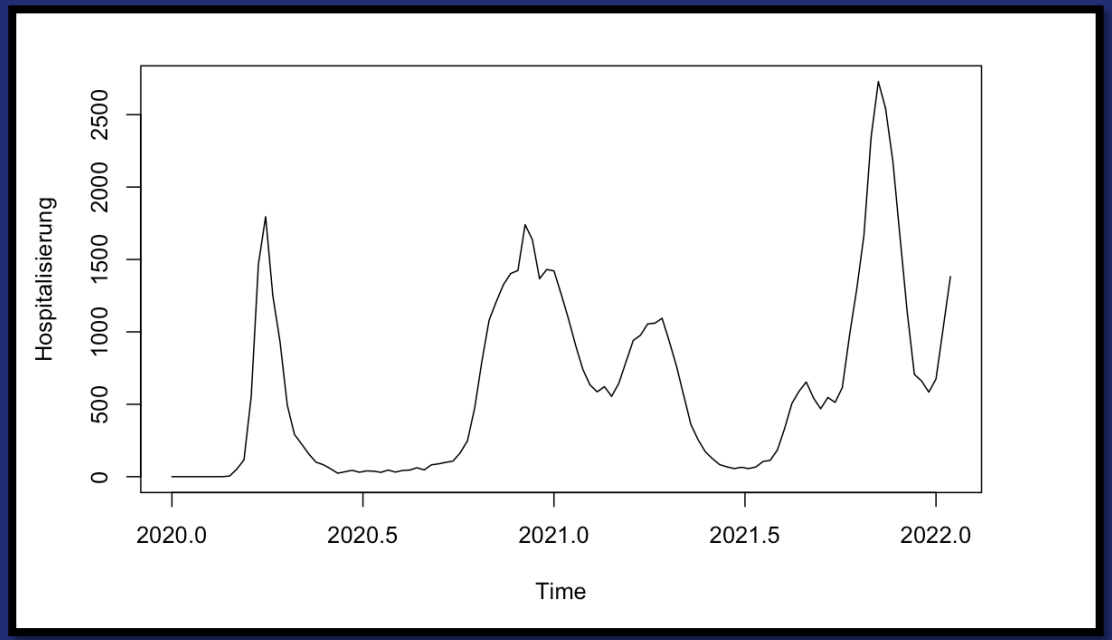
With $b_0 = 0$ and $x_t = y_{t-1}$

05

MODEL INTRODUCTION

ANALYZING THE TIME SERIES (BAYERN)

No clear trend visible
➔ **Stationary time series**



UTILIZING AUTO.ARIMA

- Hyndman-Khandakar algorithm for automatic ARIMA modelling
- auto.arima first checks for stationarity using unit root test
- p and q are chosen by minimising the AIC after differencing the data d times
- Four initial models are fitted:
  - ARIMA(0,d,0)
  - ARIMA(2,d,2)
  - ARIMA(1,d,0)
  - ARIMA(0,d,1)
- Best model with the smallest AIC is then tested on different values for p and q
- Steps are repeated until the lowest AIC is reached

UTILIZING AUTO.ARIMA (BAYERN)

Suggested Model

ARIMA(1,0,2)(0,1,0)[53]

ARIMA(1, 0, 2) for trend

x

ARIMA(0, 1, 0)[53] for season

Differences of 1. Order: $W_t = Y_t - Y_{t-1}$
(The differences are the I in ARIMA)

MODEL EVALUATION (BAYERN)

MAPE (Mean Absolute Percentage Error)

Measure of prediction accuracy of a forecasting method

$$\text{MAPE} = \frac{1}{n} \cdot \sum_{t-1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

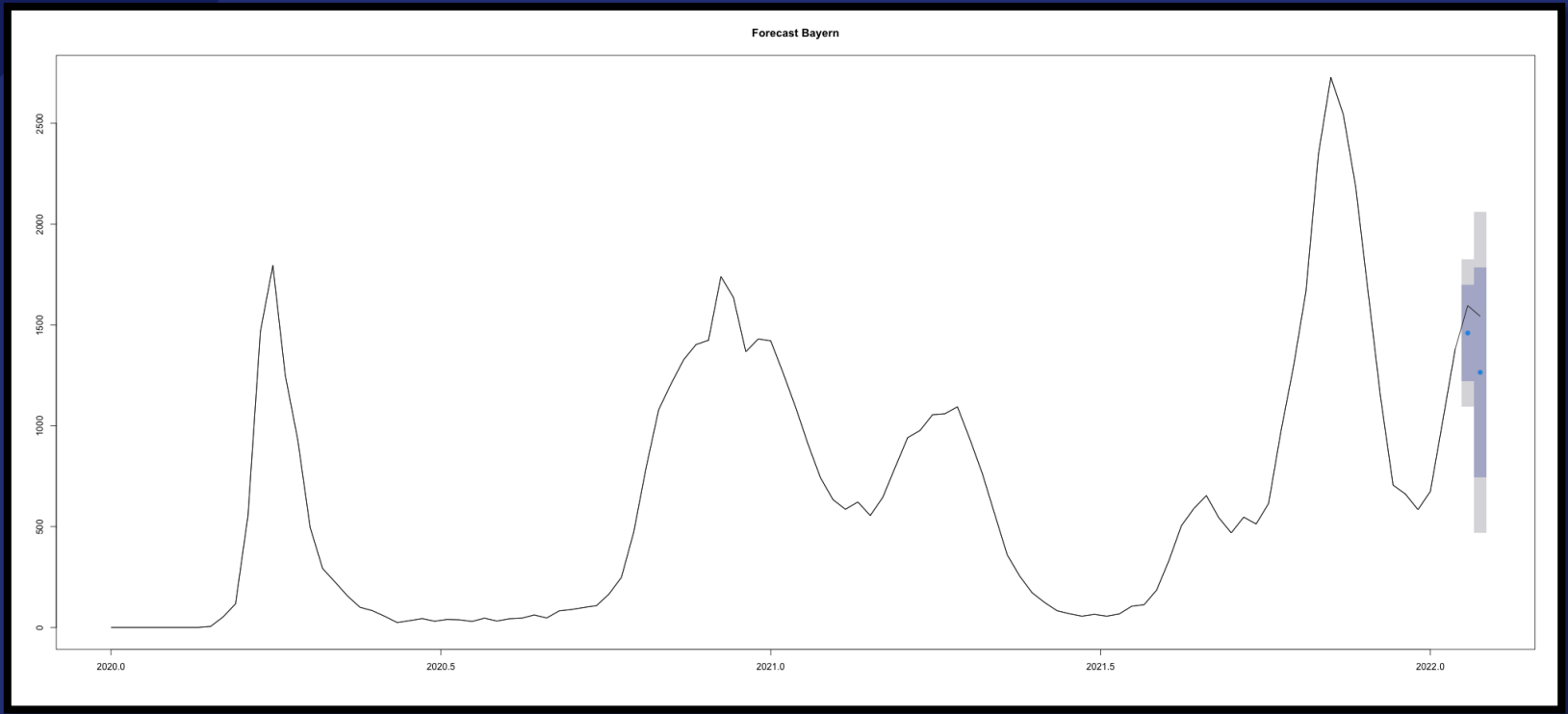
A_t : true value

F_t : forecast value

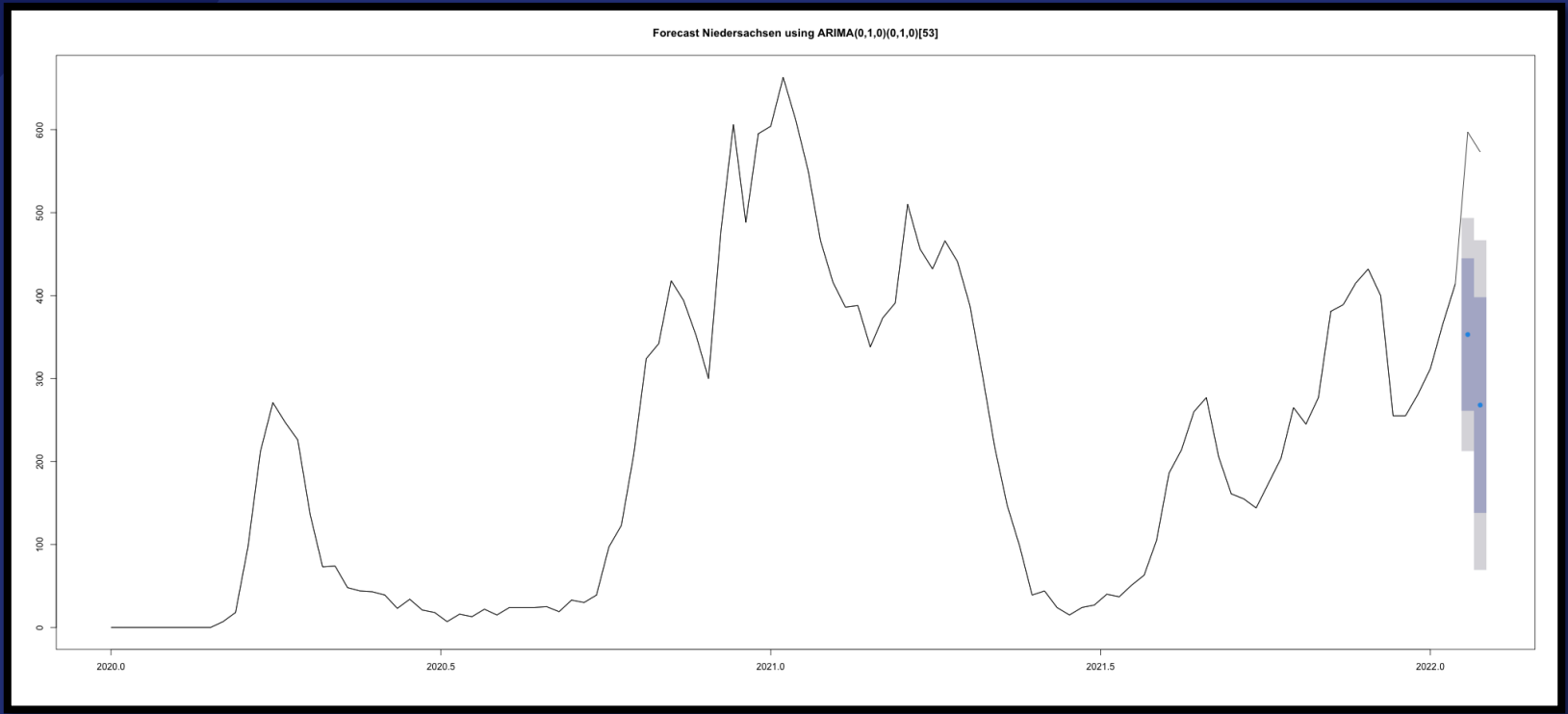
MAPE at one week: 13%

MAPE at two weeks: 25.7%

FORECAST FOR BAYERN



FORECAST FOR NIEDERSACHSEN

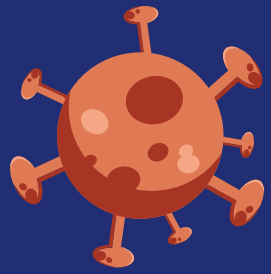


SUMMARY

- The predictions should be used with care
  - No information regarding second infections or hospitalizations due to anonymity
  - Only two years of data
  - A lot of potential information that could be missing in the univariate time series

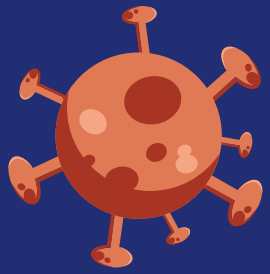
SUMMARY

- Kinds of data we were unable to obtain:
  - The exact number of infections
  - Can people be hospitalized multiple times
  - The exact date of hospitalization
  - Possible indicators (previous illness, smoking etc.)



00

DISCUSSION



ATTACHMENT

Definition ACF

Def. Linear Time Series:

$X_t = \psi_0 * a_t + \ ... = \sum_{i=0}^{\infty} \psi_t * a_{t-i}$ With $\sum_{i=0}^{\infty} \psi_i^2 < \infty$

$$E(X_t) = E\left(\sum_{i=0}^{\infty} \psi_i * a_{t-i}\right) = \sum_{i=0}^{\infty} \psi_i * E(a_{t-i}) = 0$$

$$\mathrm{Cov}(X_t, X_{t+h}) = E\big(X_t - E(X_t)\big)\big(X_{t+h} - E(X_{t+h})\big) = E(X_t X_{t+h}) =$$

$$E\left(\left(\sum_{i=0}^{\infty} \psi_i * a_{t-i}\right)\left(\sum_{i=0}^{\infty} \psi_i * a_{t+h-i}\right)\right) = \cdots = \sum_{i=0}^{\infty} \psi_i \left(\sum_{j=0}^{\infty} \psi_j * E\big(a_{t-i} a_{t+h-j}\big)\right)$$ With $E(a_r a_s) = \begin{cases} 0, if\ r \neq s \\ \sigma^2, else \end{cases}$

$$if\ therefore\colon\ j = i + h \Rightarrow E(X_t X_{t+h}) = \sum_{i=0}^{\infty} \psi_i \psi_{i+h} \sigma^2$$

$$\gamma(h) = \sigma^2 * \sum_{i=0}^{\infty} \psi_i \psi_{i+h}$$ and $$\rho(h) = \frac{\sum_{i=0}^{\infty} \psi_i \psi_{i+h}}{\sum_{i=0}^{\infty} \psi_i^2}\ for\ h = 0,1,2,..$$

Definition MA(q)

$$X_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q \, a_{t-q} \; with \; \theta_q \neq 0$$

Therefore it is a linear time series with:

$$\psi_0 = 1, \psi_1 = -\theta_1, \ldots, \psi_q = -\theta_q \; and \; \psi_k = 0 \; for \; k > q$$

Definition ARMA

Combination of AR(p) and MA(q) models

$$\phi(B)X_t = \theta(B)a_t \leftrightarrow$$
$$\left(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p\right)X_t = \left(1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q\right)a_t$$

Definition ARIMA(p,d,q)

Differences of 1. Order: $W_t = X_t - X_{t-1} = (1 - B)X_t$

Differences of d-th Order: $W_t = (1 - B)^d X_t; t = d + 1, \dots, n$

If W_t is stationary one can try to fit an $ARMA(p, q) - model$

Therefore the new model for X_t is: $\phi(B)(1 - B)^d X_t = \theta(B)a_t$

Definition ARIMA(P,D,Q)_s

Seasonal component s, so $X_t \; and \; X_{t+s} \; are \; correlated.$
Due to this additional relationship we try to fit an ARIMA-model to these values.
This model is called

Where $\Phi(B^s) = (1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps})$
And $\Theta(B^s) = (1 - \Theta_1 B^s - \Theta_2 B^{2s} - \cdots - \Theta_Q B^{Qs})$

The B_t can be described by an ordinary ARIMA(p,d,q)-model

By combining those two models we receive the socalled
$multiplicative \; ARIMA(p,d,q)x(P,D,Q)_s$ model

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D X_t = \theta(B)\Theta(B^s)a_t$$

Definition Unit root Test

Does a time series have a unit root (is it stationary)?

$$I(1)\colon\ x_t = x_{t-1} + \varepsilon_t$$

$$AR(1)\colon\ x_t = \phi x_{t-1} + \varepsilon_t$$

How to decide whether it is an AR(1) time series or an I(1) model?

$$H_0\colon \phi = 1\ (it\ is\ an\ I(1)\ process)$$

$$vs.$$

$$H_1\colon |\phi| < 1\ (it\ is\ an\ AR(1)\ process)$$

The teststatistic is defined by:

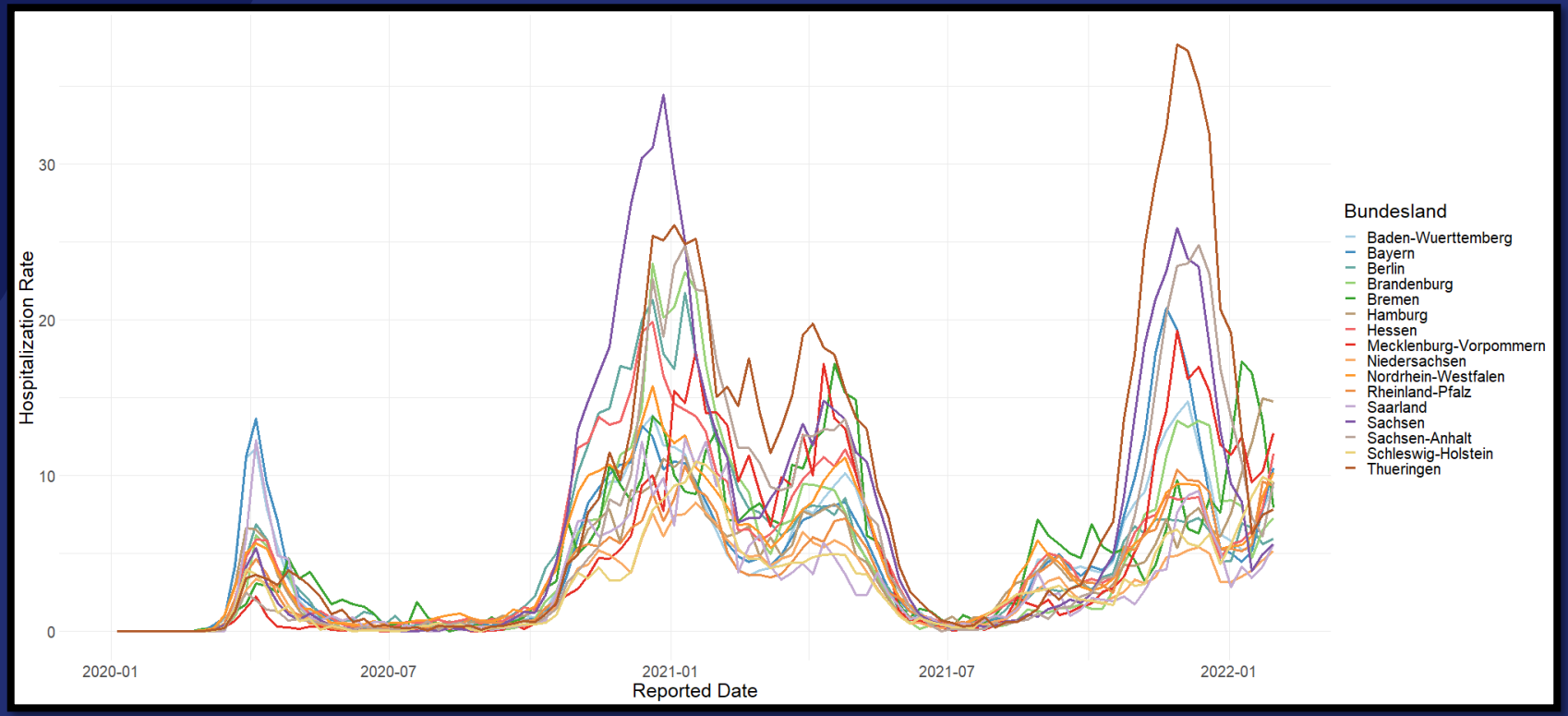
$$T = n(\hat{\phi} - 1)$$

$$n\colon length\ of\ the\ time\ series$$

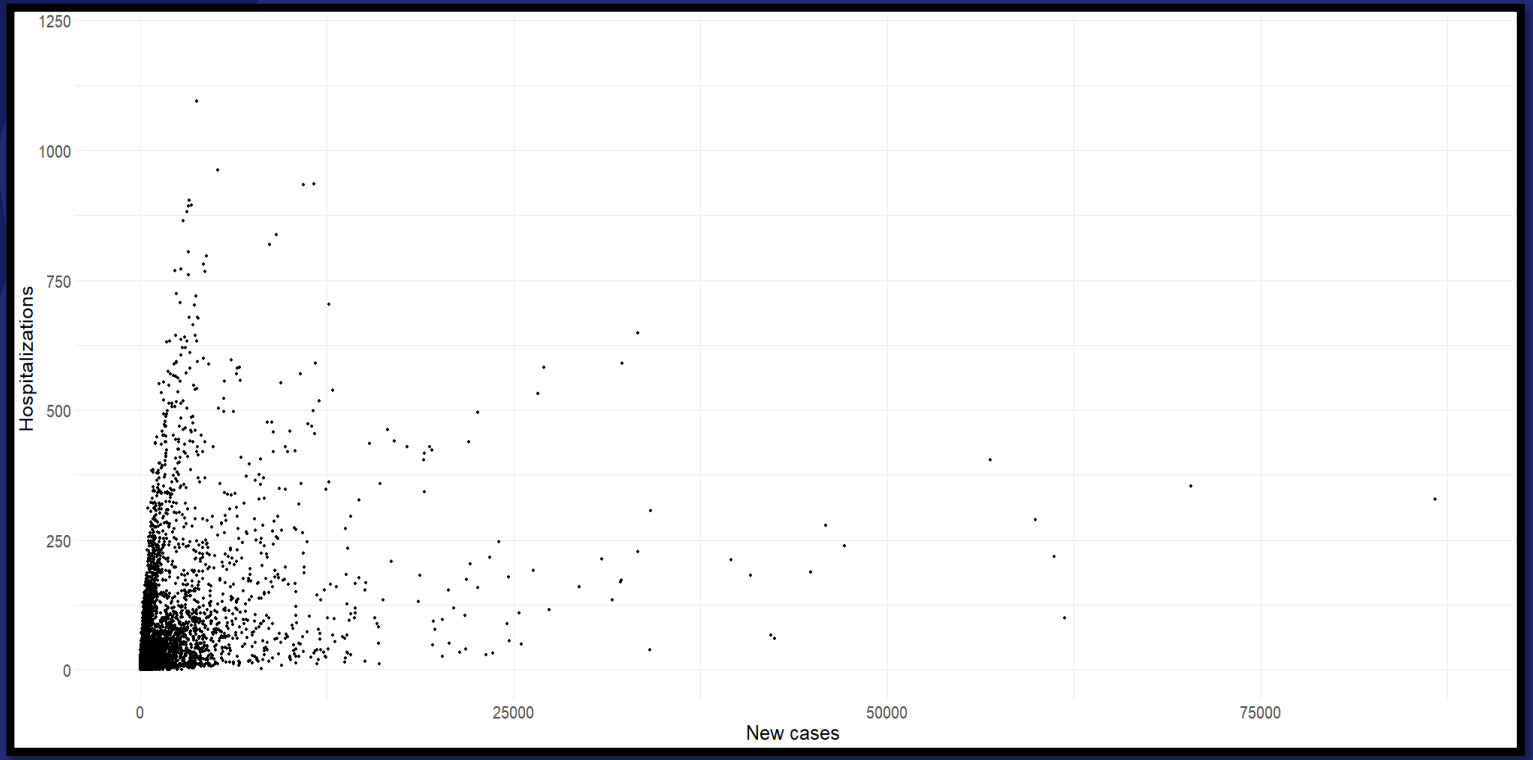
$$\hat{\phi}\colon Estimation\ of\ \phi$$

The critical values can be found in f.e. Halton(1994)

HOSPITALIZATION RATE BY STATE



HOSPITALIZATIONS VS. NEW CASES



HOSPITALIZATIONS VS. NEW CASES