



COVID-19: PREDICTION OF THE HOSPITALIZATION RATE

Statistisches Praktikum
WiSe21/22

16.03.2022
Munich

Project partner:
Yeganeh Khazaei
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

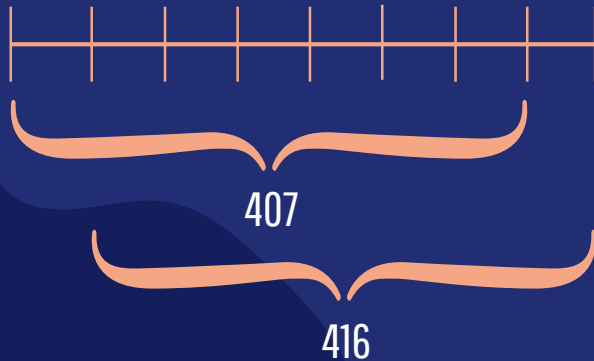
NOWCASTING

Problem: Delays in reporting (period between onset of illness and 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



01.08.2021

Reported

407

Updated

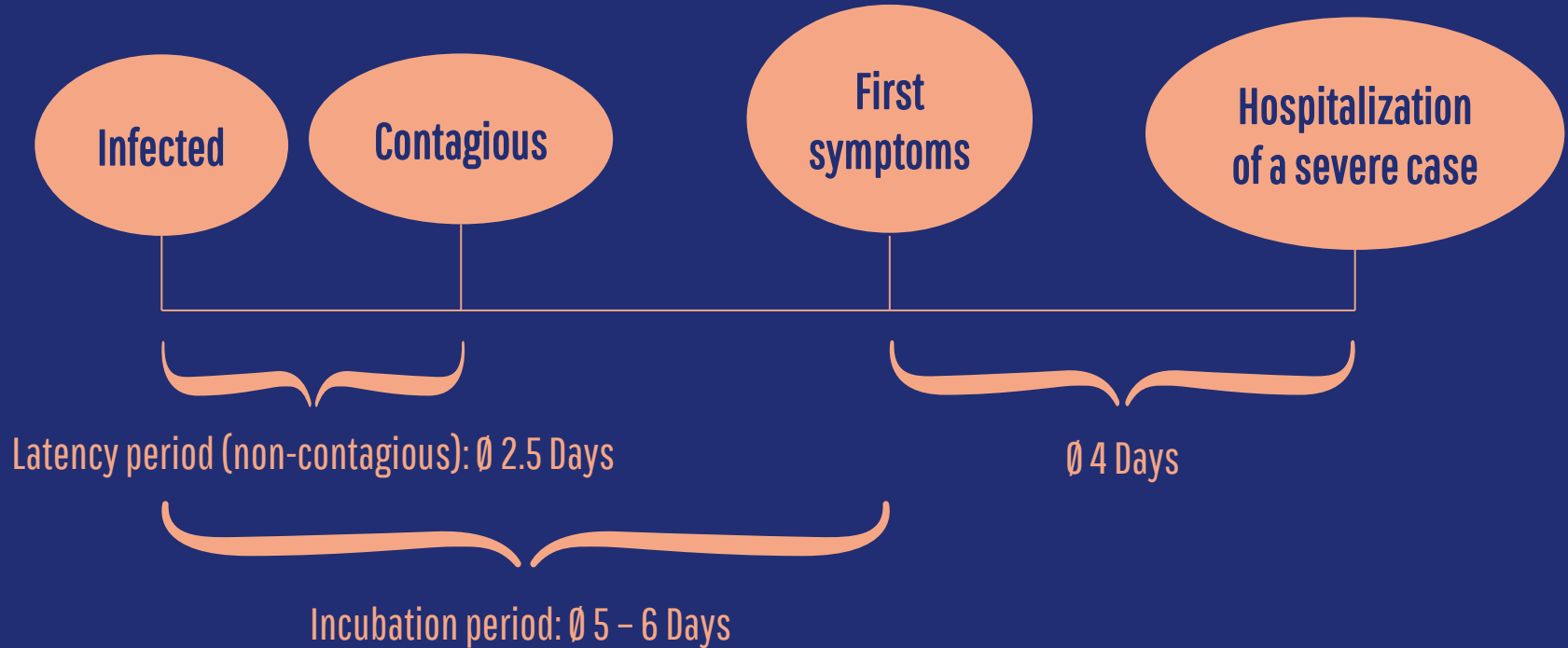
756

02.08.2021

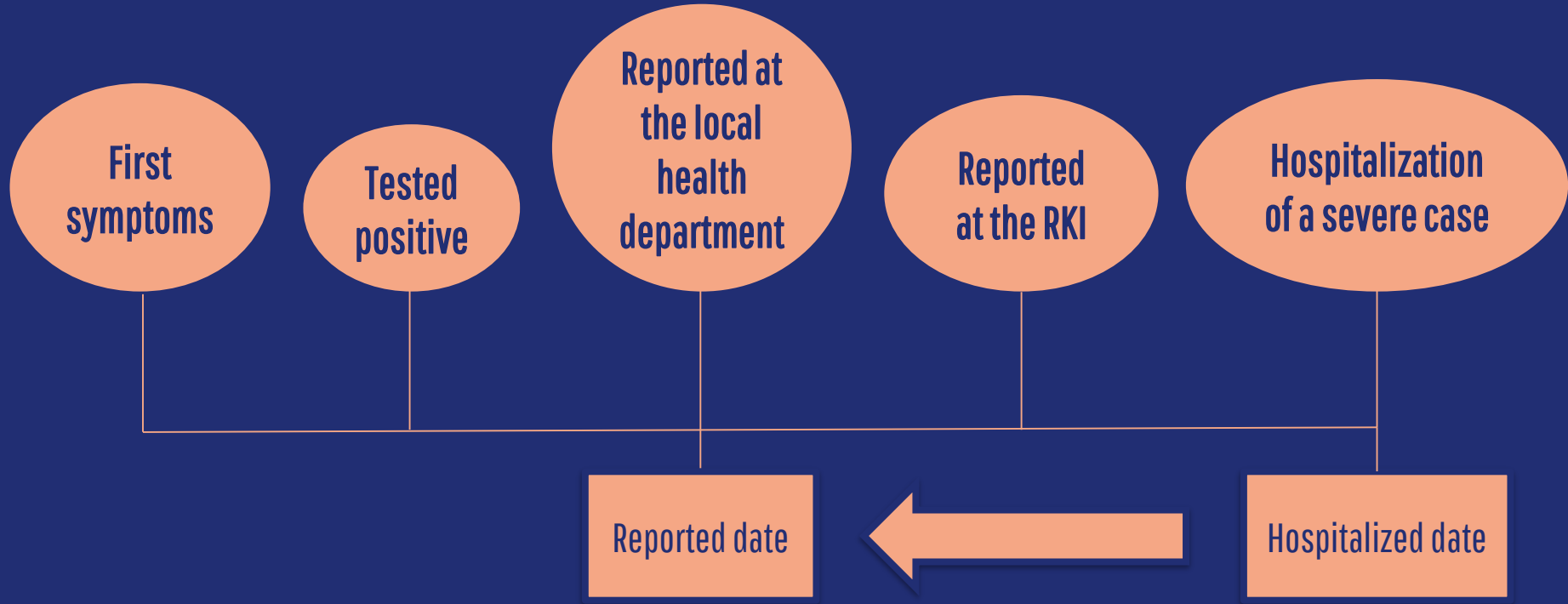
416

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



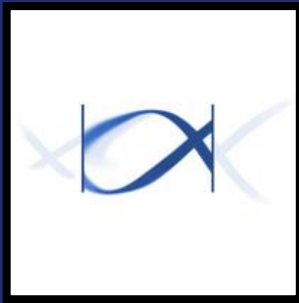
02

DATA PROCESSING

DATA COLLECTION

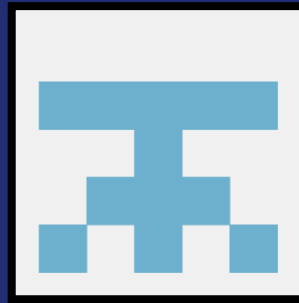
RKI

- New cases
- Hospitalizations



KITMetricslab

- Population



LMU

- Vaccination



FINALIZED DATA SET

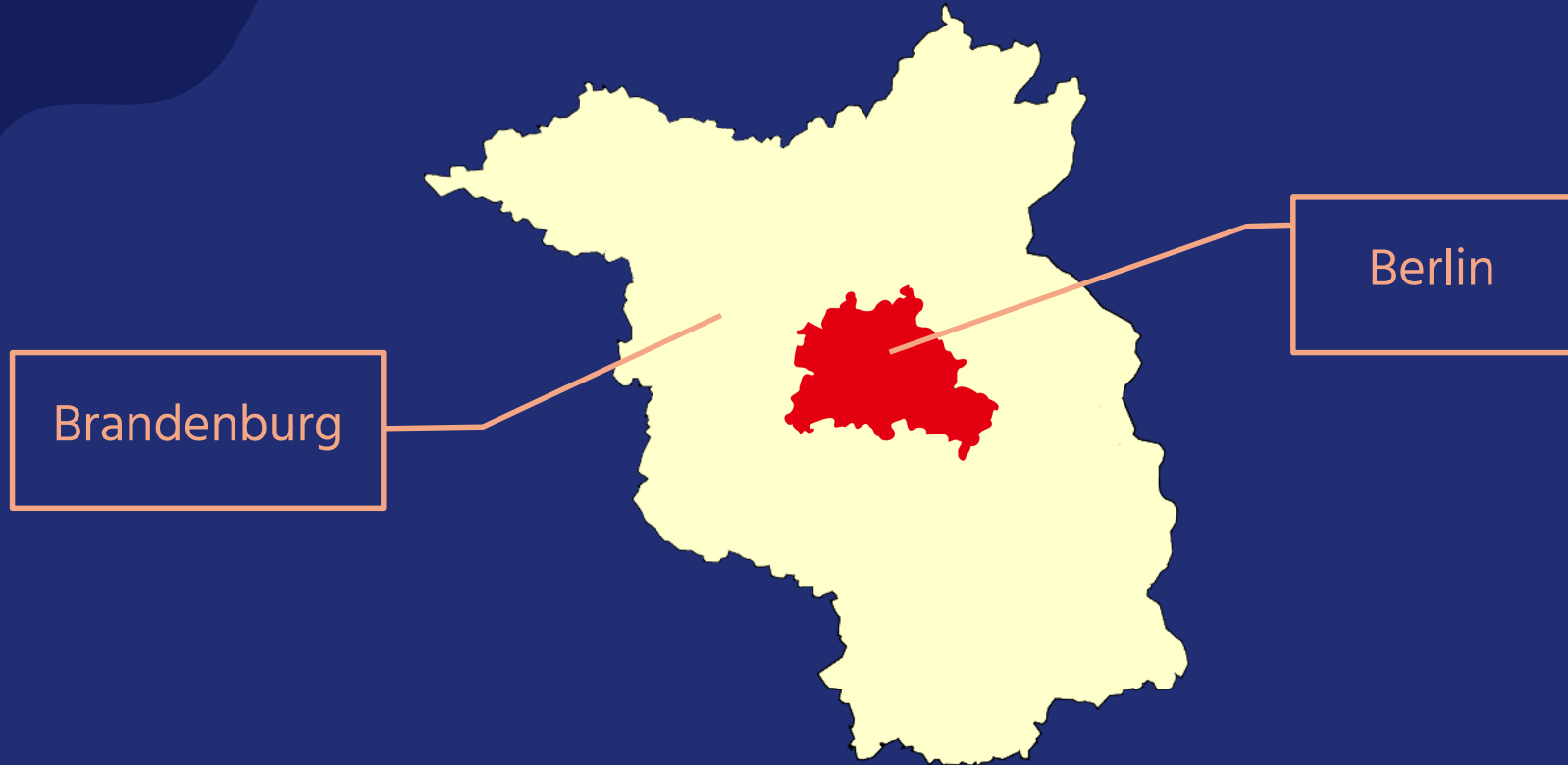
Reported date	Year	Calendar week	Index	State	Age group	Population	New cases	Hospitalizations
2020-03-01	2020	9	9	Bayern	80+	871866	0	0
2020-03-08	2020	10	10	Bayern	80+	871866	3	2
2020-03-15	2020	11	11	Bayern	80+	871866	39	25
2020-03-22	2020	12	12	Bayern	80+	871866	214	132
2020-03-29	2020	13	13	Bayern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



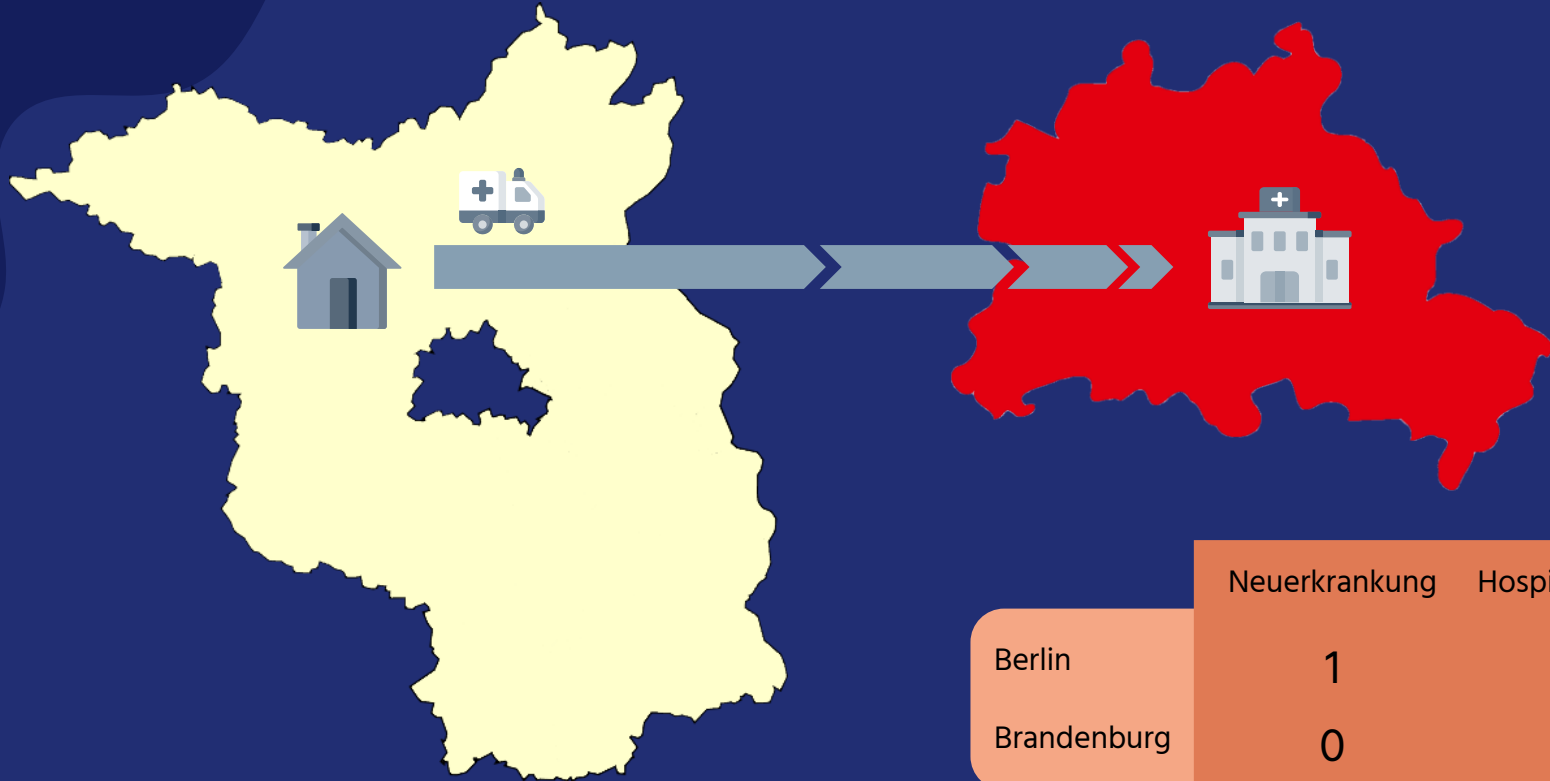
03

DATA ANALYSIS

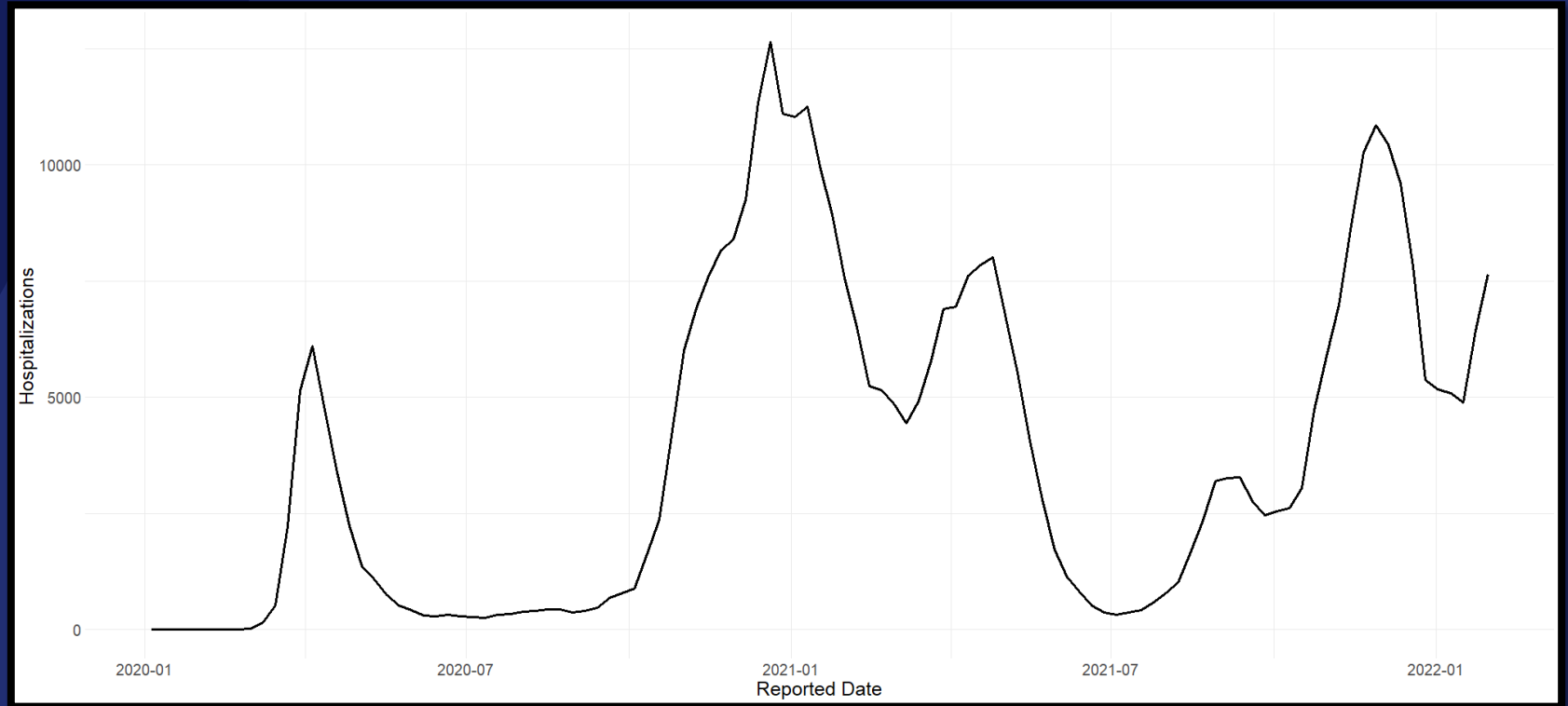
PROBLEMS



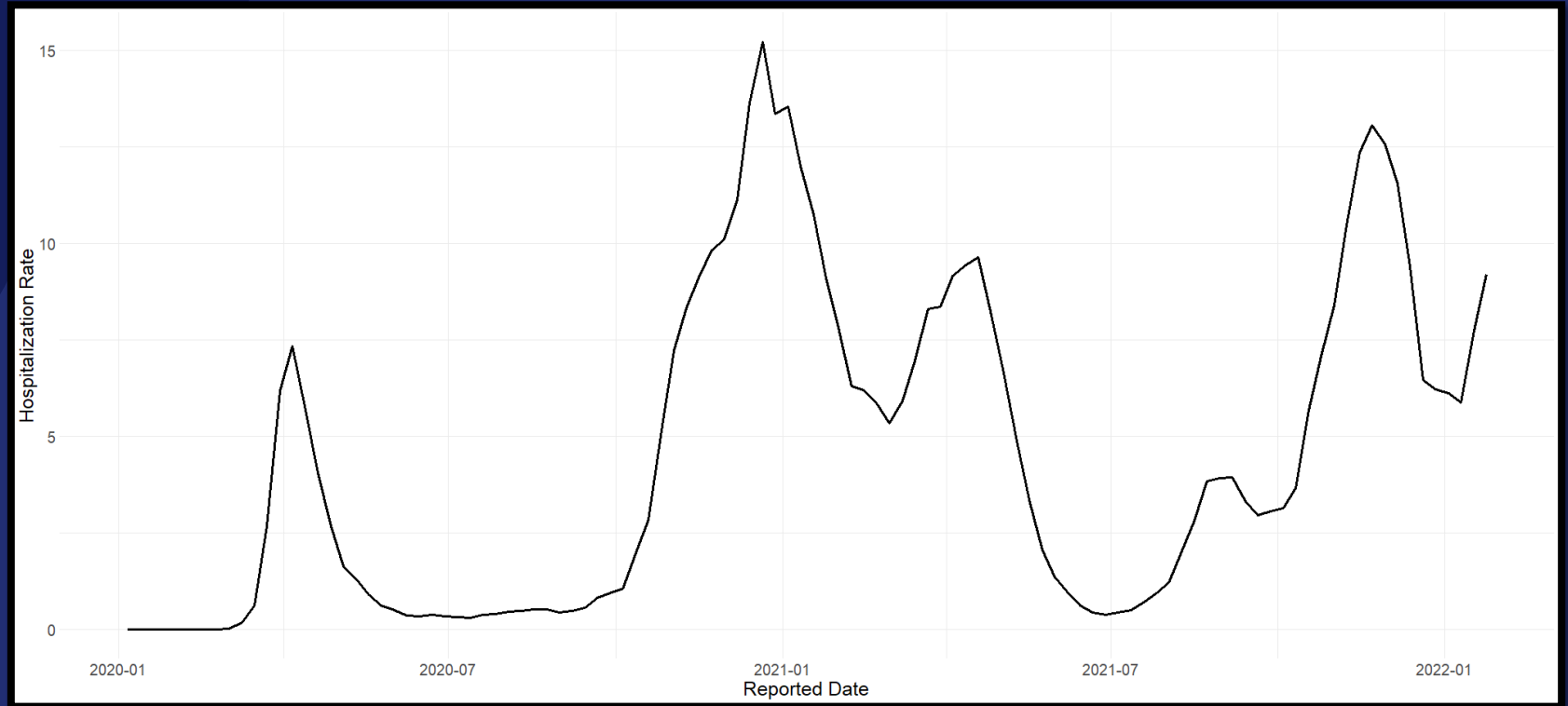
PROBLEMS



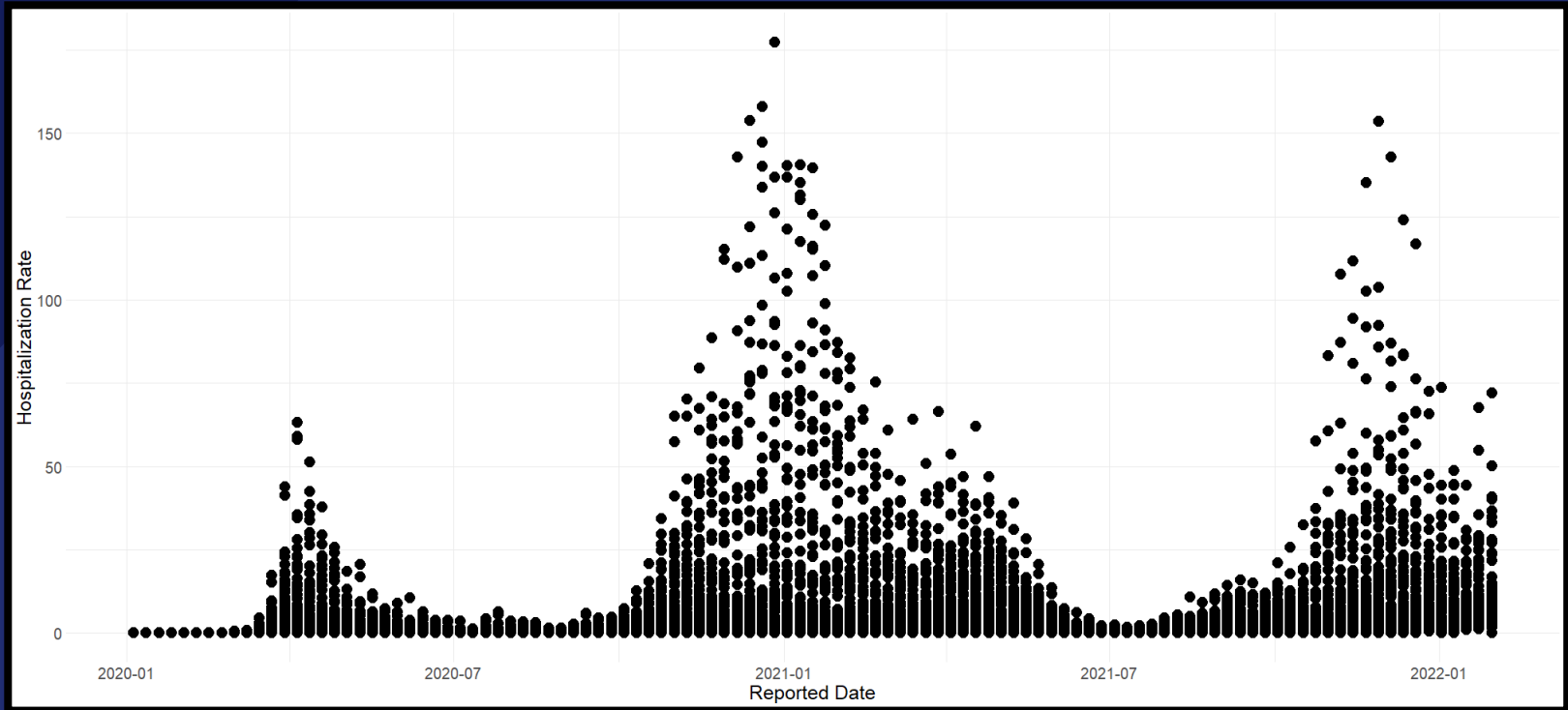
ABSOLUTE NUMBERS OF HOSPITALIZATIONS



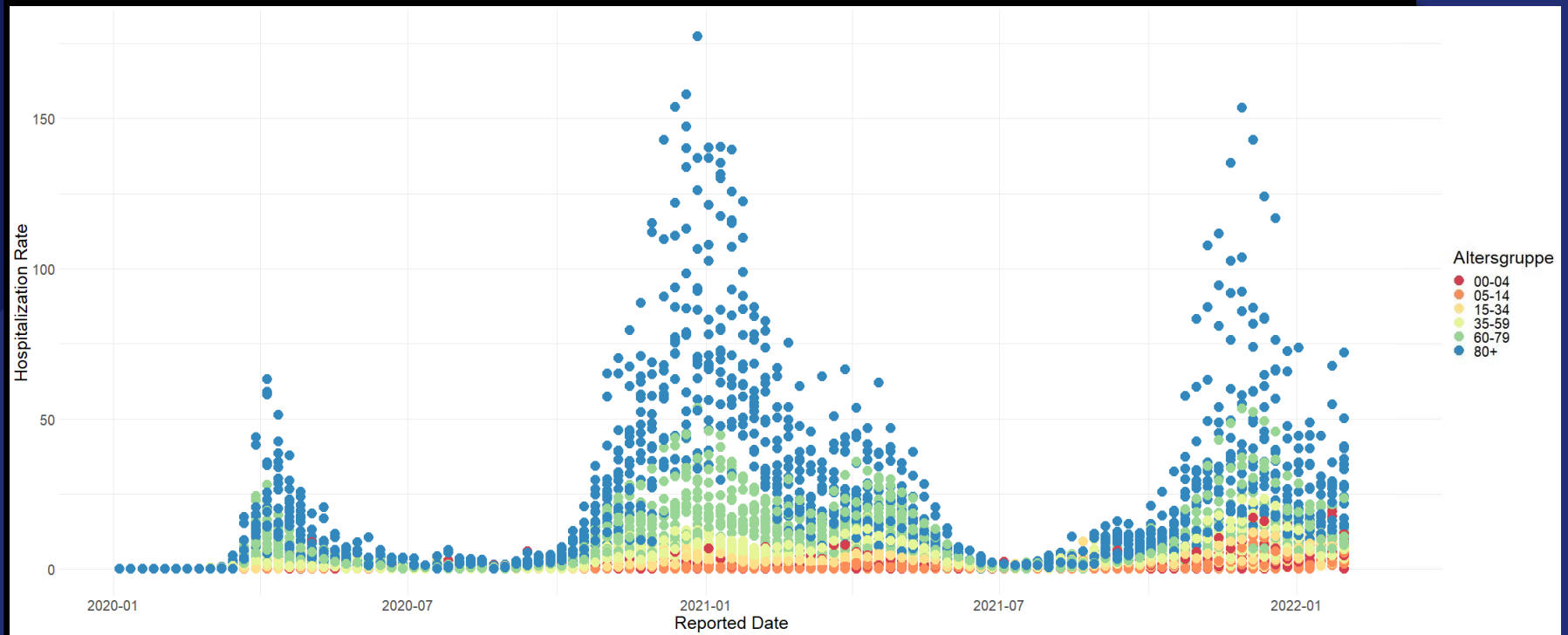
NUMBERS REPORTED BY THE RKI



EXPLAINING THE HOSPITALIZATION RATE



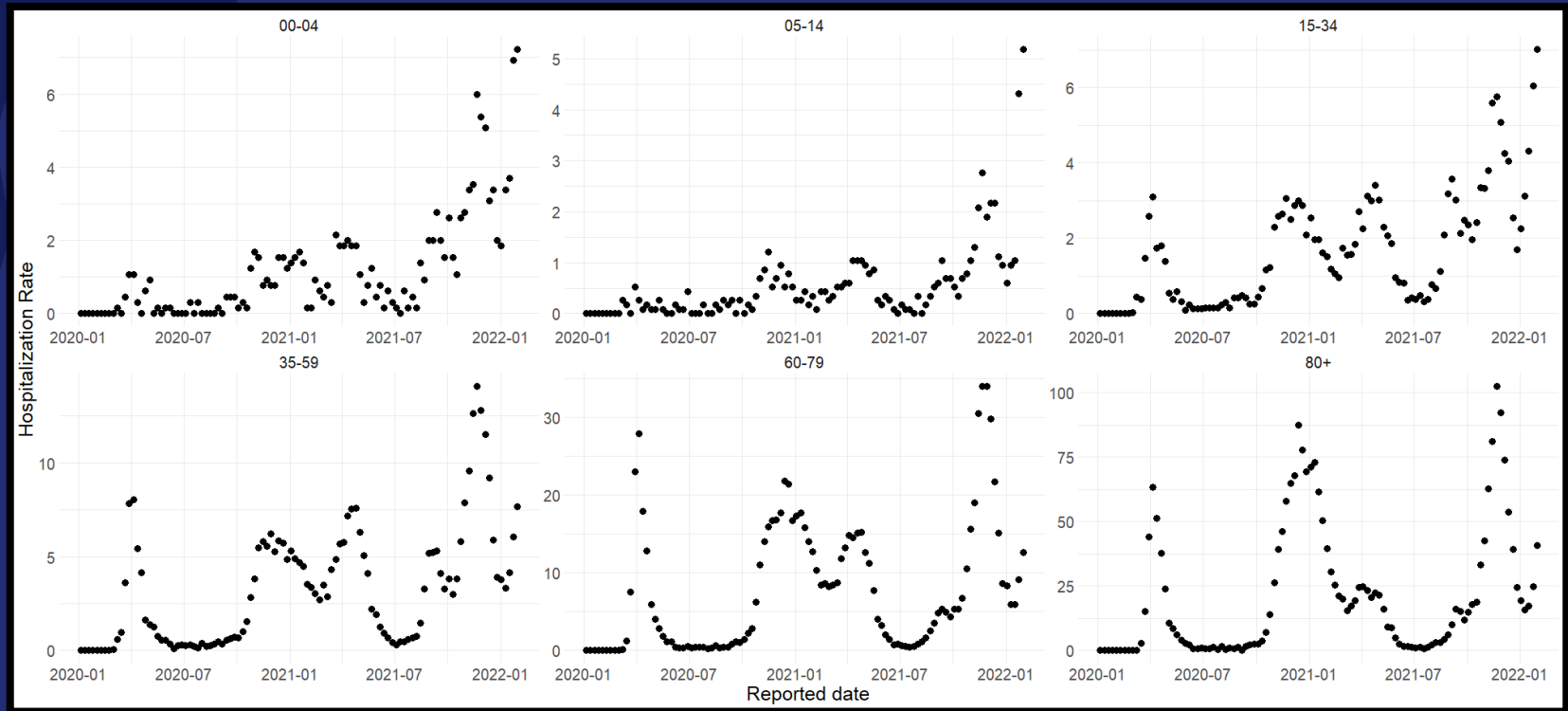
HOSPITALIZATION RATE BROKEN DOWN INTO AGE GROUPS



EXPLAINING THE HOSPITALIZATION RATE



HOSPITALIZATION RATE BY AGE GROUP IN BAYERN



ANALYZING BAYERN BY USING GAM

Distribution assumption:

Hospitalization_i | x_i ~ Poi(λ)

Link:

Logit

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

Effect of the individual coefficients:

Intercept:

$\exp(1.06) \sim 3$

05-14:

$\exp(-0.03) \sim 1$

15-34:

$\exp(1.93) \sim 7$

35-59:

$\exp(2.77) \sim 16$

60-79:

$\exp(2.58) \sim 13$

80+:

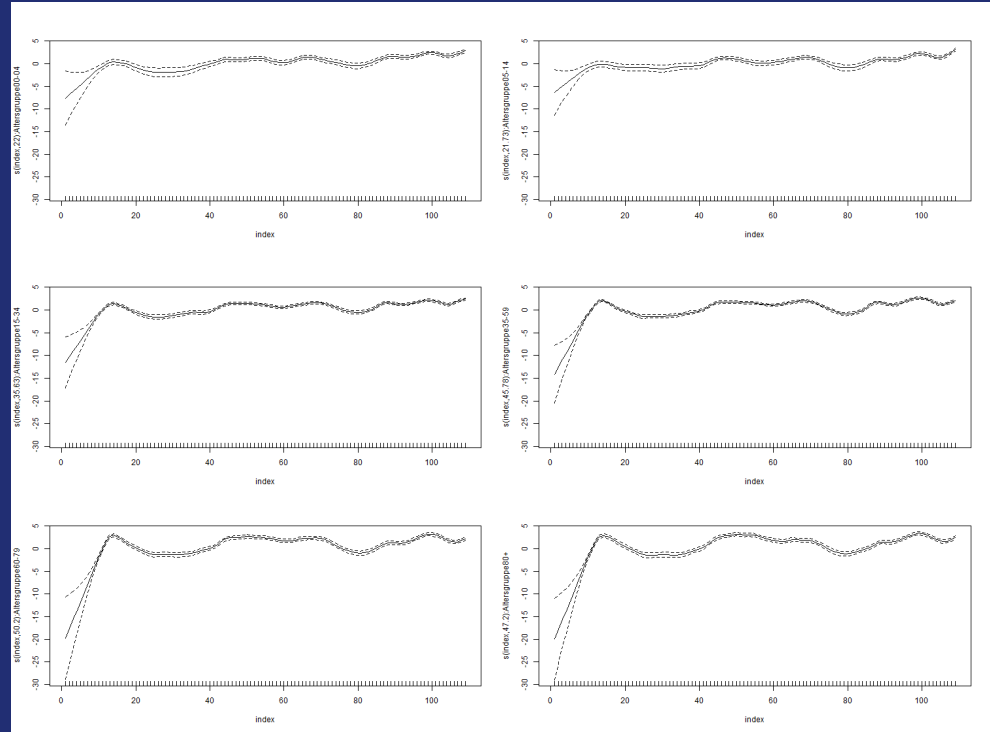
$\exp(2.3) \sim 10$



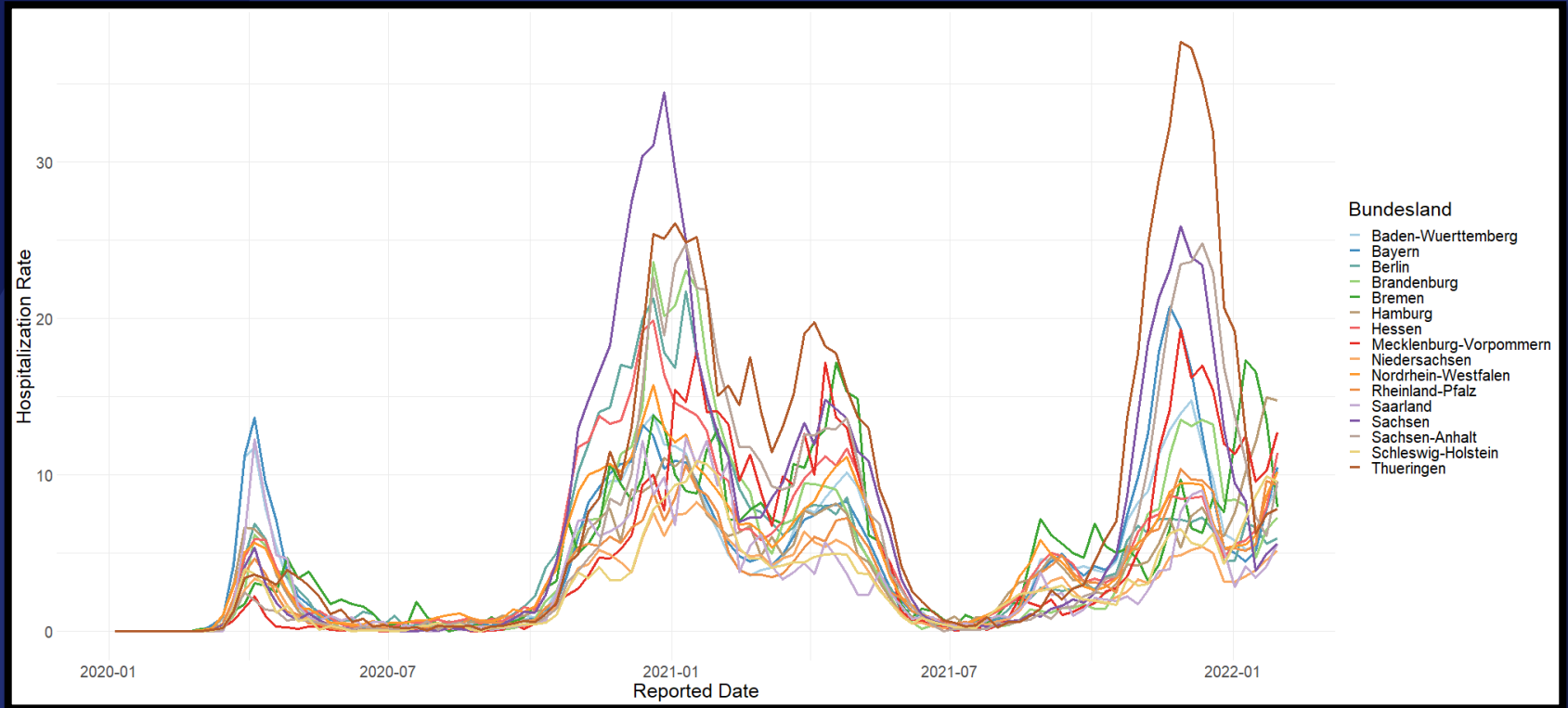
00-04 and 05-14 are the least impactful age groups.
This holds true for each individual state.

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



HOSPITALIZATION RATE BY STATE



SUB-CONCLUSION

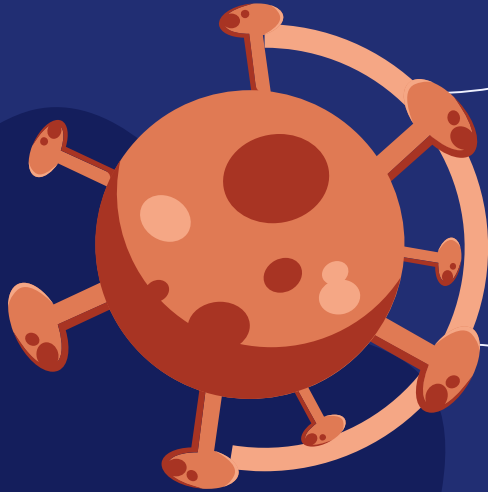
- Differentiating between states seems to be beneficial in improving the prediction, due to the differences in trend for each state
- The trend for each age group regarding the hospitalization looks however quite similar



04

EXCURSION:
TIME SERIES

TIME SERIES



Stationarity

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

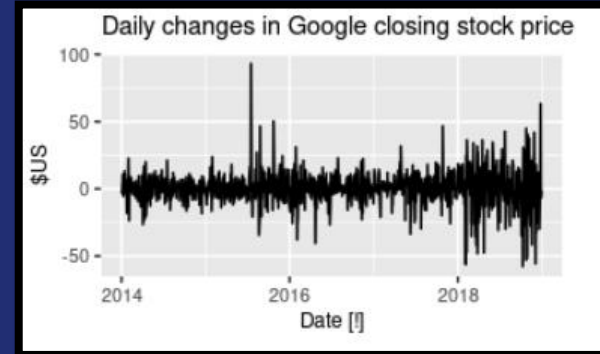
Seasonality

Is there a seasonal pattern?

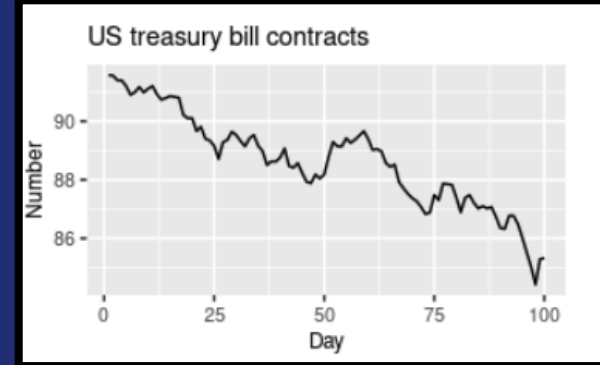
Such as the time of the year or the day of the week?

TIME SERIES (EXAMPLES)

A stationary time series



A nonstationary time series



TIME SERIES

- Every realisation y_1, \dots, y_n can be seen as a series of random variables Y_1, \dots, Y_n .
- If the (joint) distribution of Y_1, \dots, 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, \dots, Y_n

Covariance

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

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

$$Y_t = \chi_0 \cdot a_t + \chi_1 \cdot a_{t-1} + \dots$$

$$a_t \sim N(0, \sigma^2)$$

$$t = \dots, -1, 0, 1, 2, \dots$$

χ_i weights, $i = 0, 1, 2, \dots$

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 $Y_{t+1}, \dots, Y_{t+h-1}$

DIFFERENT TYPES OF TIME SERIES

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

EXAMPLE OF AUTOREGRESSIVE MODELS - AR(p)

AR(p) model is defined by:

$$\begin{aligned} Y_t &= \varphi_1 \cdot Y_{t-1} + \varphi_2 \cdot Y_{t-2} + \dots + a_t & (\text{with } t = 1, 2, \dots) \\ \Leftrightarrow (1 - \varphi_1 \cdot B - \varphi_2 \cdot B^2 \dots) \cdot Y_t &= a_t \end{aligned}$$



AR(1):

$$\begin{aligned} Y_t &= \varphi \cdot Y_{t-1} + a_t & (\text{with } t = 1, 2, \dots) \\ \Leftrightarrow y_t &= b_0 + b_1 \cdot x_t + \varepsilon_t \\ \text{Where } b_0 &= 0 \text{ and } x_t = y_{t-1} \end{aligned}$$

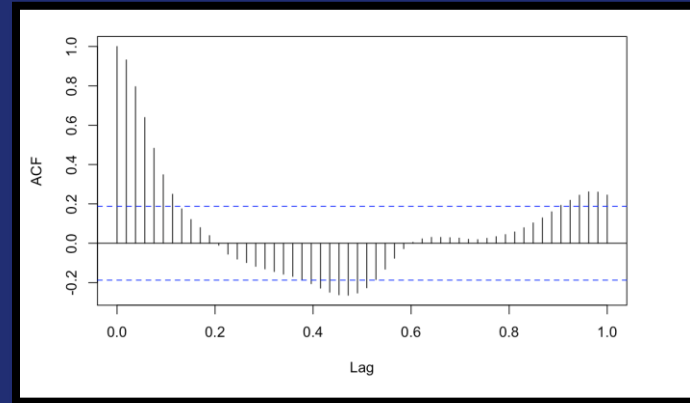


05

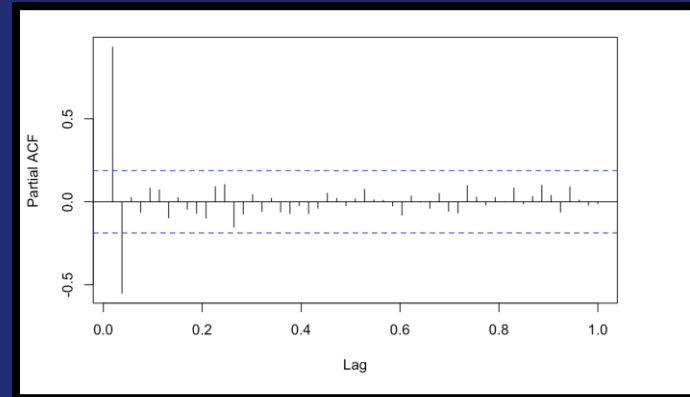
MODEL INTRODUCTION

ANALYZING ACF AND PACF (BAYERN)

ACF
Autocorrelation
 $\rho(h)$

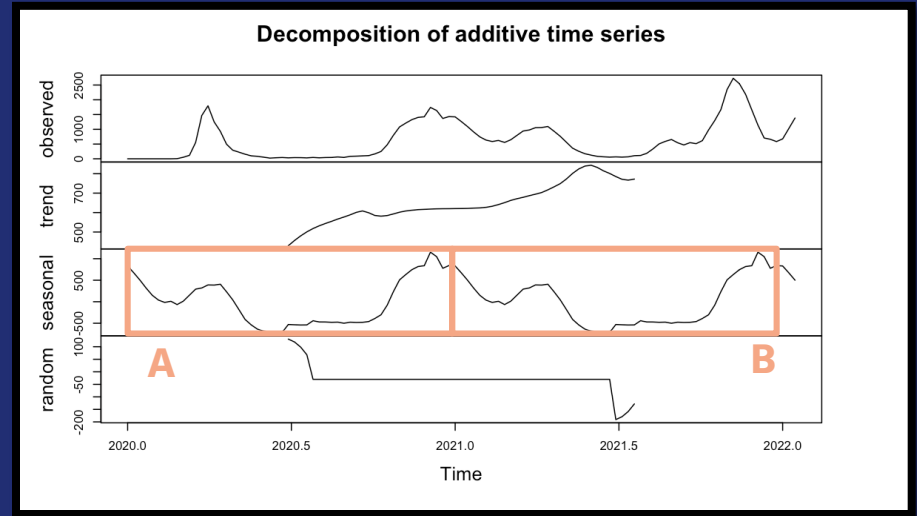


PACF
Partial autocorrelation
 $\tau(h)$



MODEL DECISION

- The time series is not stationary
- The time series is seasonal
→ $\text{ARIMA}(P, D, Q)_s$
- Using model selection we receive
→ $\text{ARIMA}(0, 1, 1)_{53}$



MODEL EVALUATION (BAYERN)

MAPE (Mean Absolute Percentage Error)

Measure of prediction accuracy of a forecasting method

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

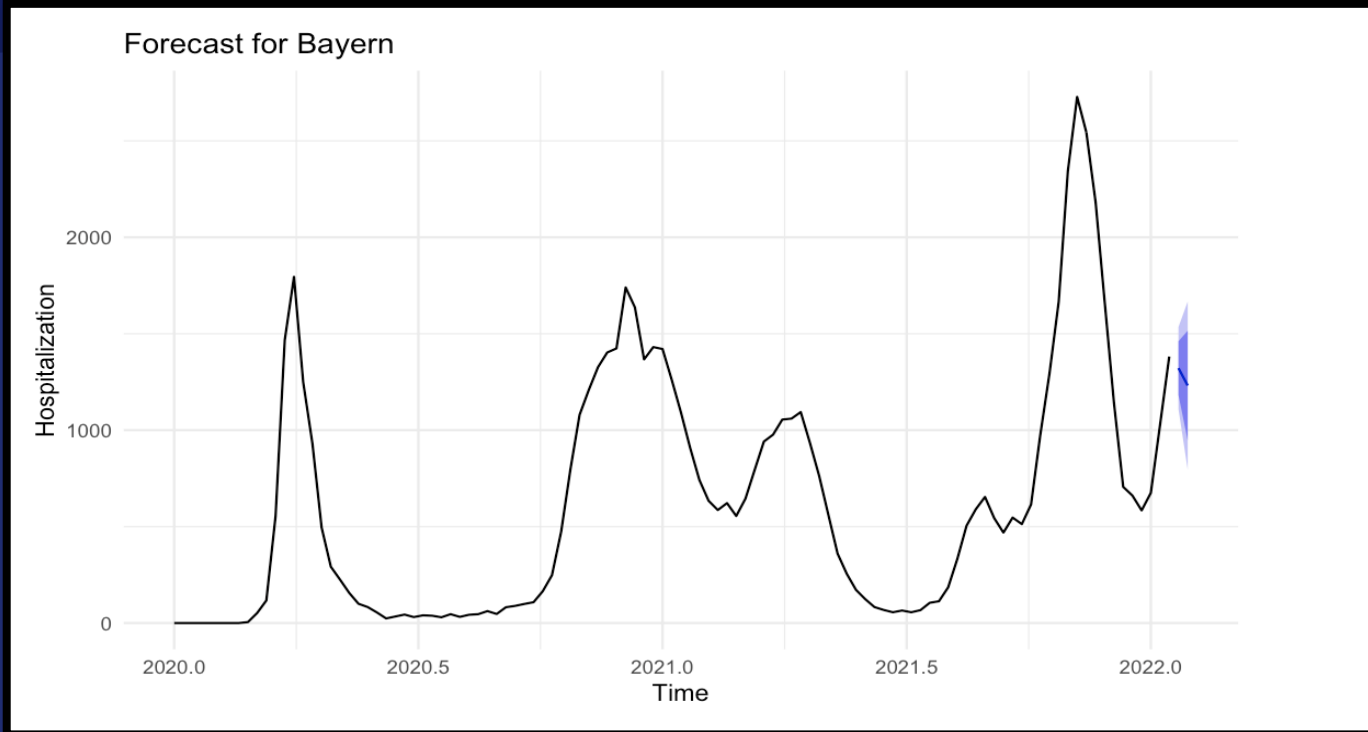
A_t : true value

F_t : forecast value

MAPE at one week:
4%

MAPE at two weeks:
19.6%

FORECAST AT TWO WEEKS FOR BAYERN



FORECASTING RESULTS

SUMMARY

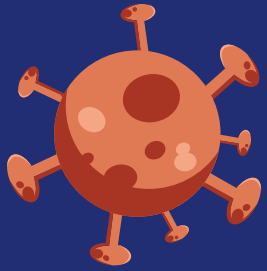
- Difficult to forecast
- Infections cannot be accounted for hospitalization (?)
- Right now (as of February) there is a clear positive trend

SUMMARY

- The predictions should be used with care
 - Unlikely that the hospitalization will keep following the current course
 - No information regarding second infections or hospitalizations due to anonymity
 - A lot of potential information that could be missing in the univariate time series which could then lead to better predictions

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 (illness, smoking etc.)

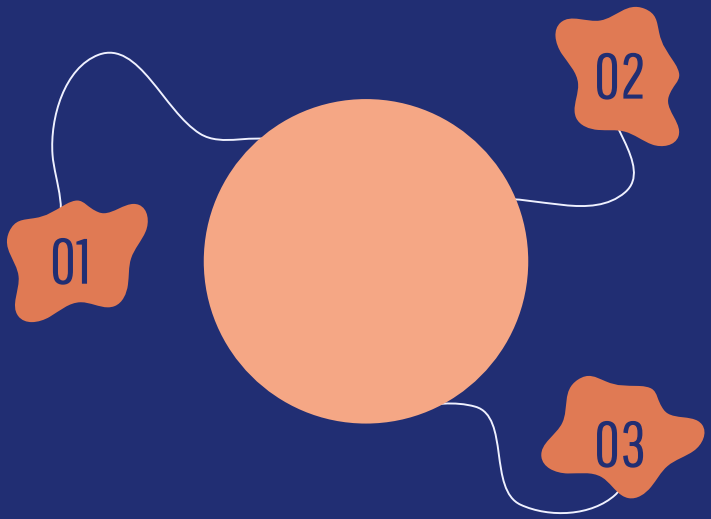


05

Diskussionsrunde



TIME SERIES



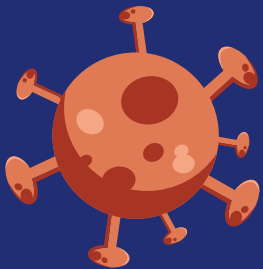
Saturn
Saturn is composed
of hydrogen and
helium

Mercury

Mercury is the
closest planet to
the Sun

Jupiter

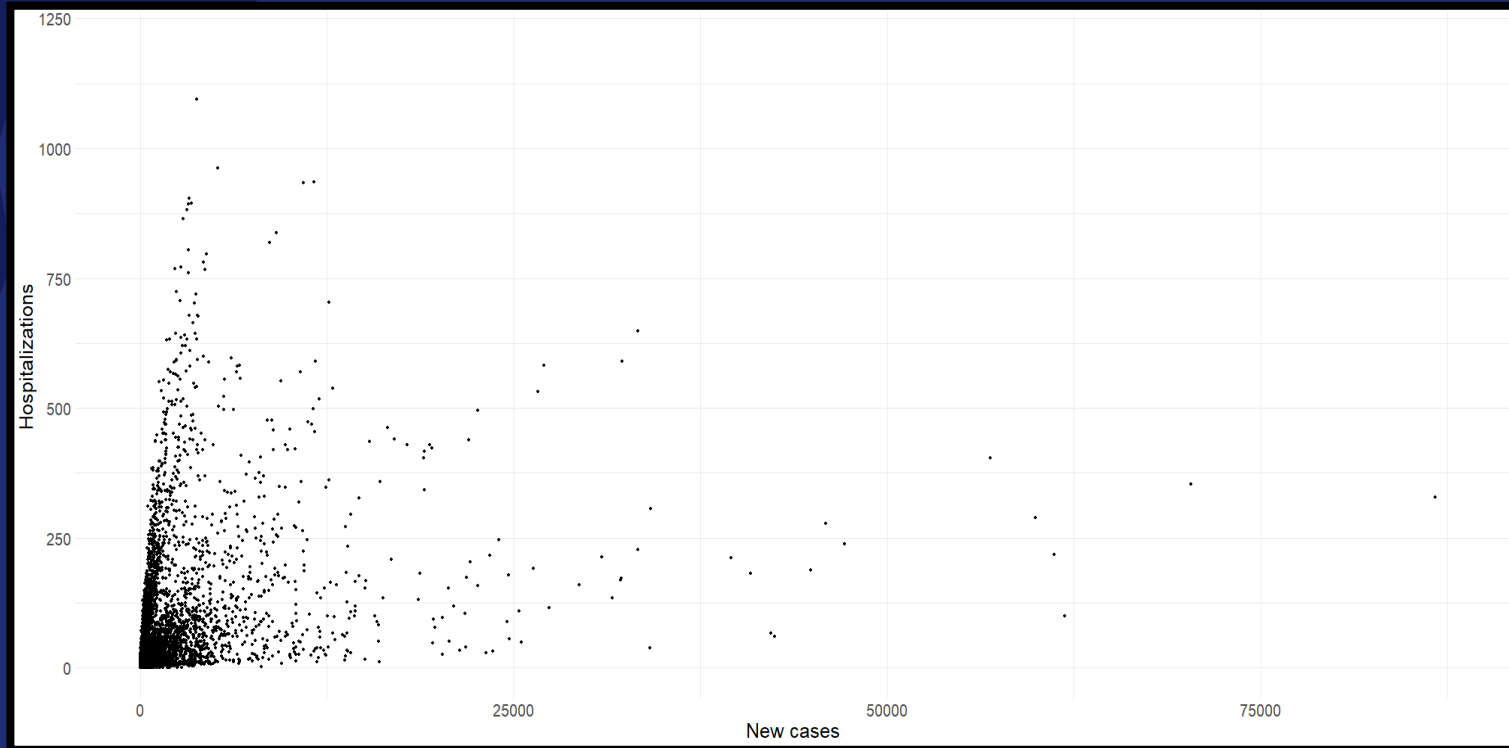
It's the biggest
planet in the Solar
System



ANHANG



HOSPITALIZATIONS VS. NEW CASES



HOSPITALIZATIONS VS. NEW CASES

