Data Plausibility Check for Cycling Traffic Data with Generalized Additive Models

Department of Statistics
Ludwig-Maximilians-Universität München

Juntae Kwon and The Anh Vu
supervised by Küchenhoff Helmut
München, 19th April 2024

Content

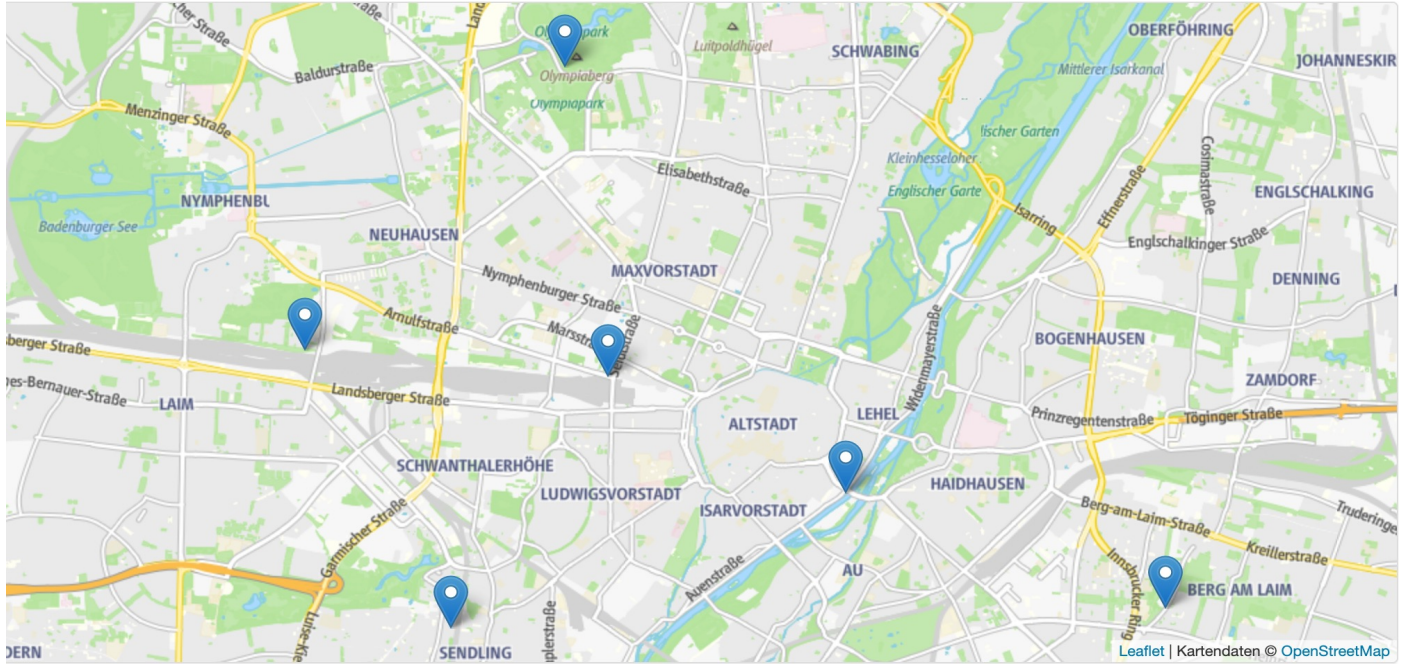
1. Data Introduction
2. Visualization
3. Model
4. Prediction Intervals
5. Limitation

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

1. Data Introduction

Data Introduction

- Counting station at 6 locations:
  - Arnulfstr. (Central Station)
  - Bad-Kreuther-Str. (Innsbrucker Ring)
  - Olympiapark
  - Hirschgarten
  - Margaretenstr. (Harras)
  - Erhardtstr. (Deutsches Museum)



Data Introduction

- 15-minute-interval data between June 2008 to December 2022

Station	Number of observations	Number of NAs
Arnulf	505.536	38.016
Erhardt	403.388	2.309
Hirsch	511.388	53.186
Kreuther	511.008	25.536
Margareten	403.392	672
Olympia	511.388	48.385
Total	2.846.100	168.104

Data Introduction

- The number of cyclists for all directions is NA valued in the following cases:
  - Station not yet in operation
  - Icy bike lane / Not cleared after snowfall / No possible measure
  - Under construction
  - Replacing sensor
  - Failure after damage
  - Some irregularities in the data

Data Introduction

Time variables

- For modeling seasonality, time-related variables are extracted.
  - year
  - month
  - day
  - hour
  - day of week

Data Introduction

Time variables

- For modeling seasonality, time-related variables are extracted.
  - year
  - month
  - day
  - hour
  - day of week
- In addition, two compound variables are created as follows:
  - **month_year := (year – 2008) * 12 + (month – 6)** (count of months starting from the first datapoint)

    e.g., June 2008 = 0, July 2008 = 1, …, December 2022 = 174

Data Introduction

Time variables

- For modeling seasonality, time-related variables are extracted.
  - year
  - month
  - day
  - hour
  - day of week
- In addition, two compound variables are created as follows:
  - month_year := (year – 2008) * 12 + (month – 6) (count of months starting from the first datapoint)

    e.g., June 2008 = 0, July 2008 = 1, …, December 2022 = 174
  - **hour_weekday := (day of week) * 24 + hour** (count of hours in a week)

    e.g., Monday 00:00 = 0, Monday 01:00 = 1, …, Sunday 23:00 = 167

Data Introduction

Weather variables

- Precipitation (mm)
- Air temperature (°C)
- Wind speed (m/s)
- Sun time per day (hour)

Data Introduction

Weather variables

- Precipitation (mm)
- Air temperature (°C)
- ~~Wind speed (m/s)~~
- ~~Sun time per day (hour)~~

Precipitation as categorical variable:

- no rain (0 mm)
- Drizzle (< 0.5 mm)
- Rain (>= 0.5 mm)

Data Introduction

- A variable for holidays including both public and school is also considered, which is encoded as binary.

Public holiday	
Day	**Date**
Mariä Himmelfahrt	2008-08-15
Tag der Deutschen Einheit	2008-10-03
Allerheiligen	2008-11-01
1. Weihnachtsfeiertag	2008-12-25
2. Weihnachtsfeiertag	2008-12-26
Neujahr	2009-01-01
Heilige Drei Könige	2009-01-06
⋮	
Tag der Deutschen Einheit	2022-10-03
Allerheiligen	2022-11-01
1. Weihnachtsfeiertag	2022-12-25
2. Weihnachtsfeiertag	2022-12-26

School holiday	
Day	**Date**
Summer break	2008-08-04
	2008-08-05
	2008-08-06
	2008-08-07
	2008-08-08
	2008-08-09
	2008-08-10
⋮	
Christmas break	2022-12-28
	2022-12-29
	2022-12-30
	2022-12-31

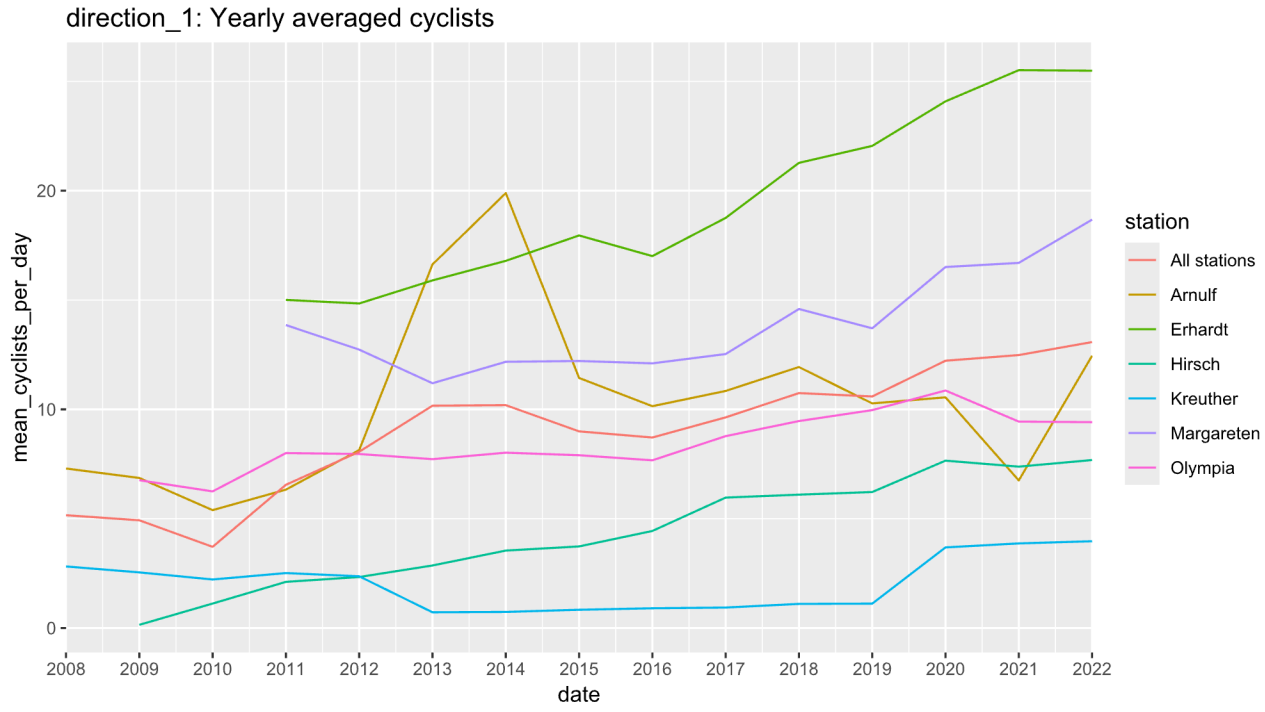
2. Visualization

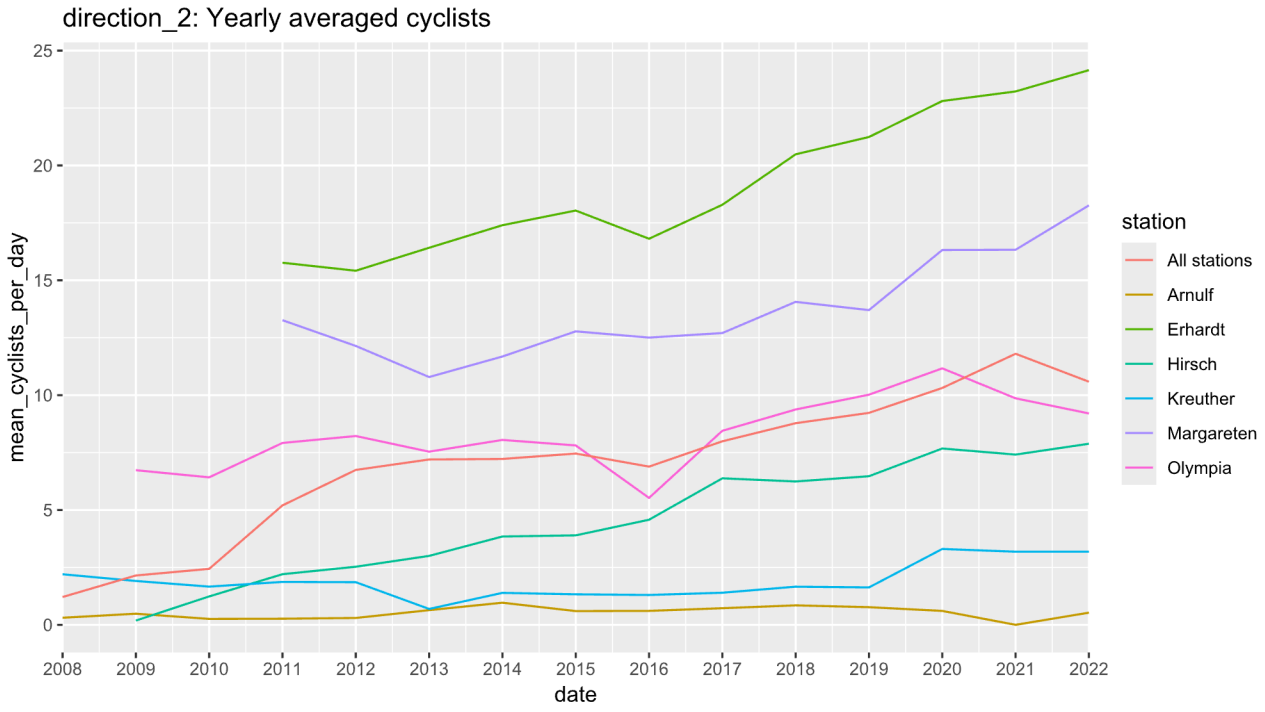
Considerations

- Time
  - Yearly Seasonality
  - Monthly Seasonality
  - Weekly Seasonality
  - Hourly Seasonality
- Holiday
- Direction
- Station

Yearly Seasonality

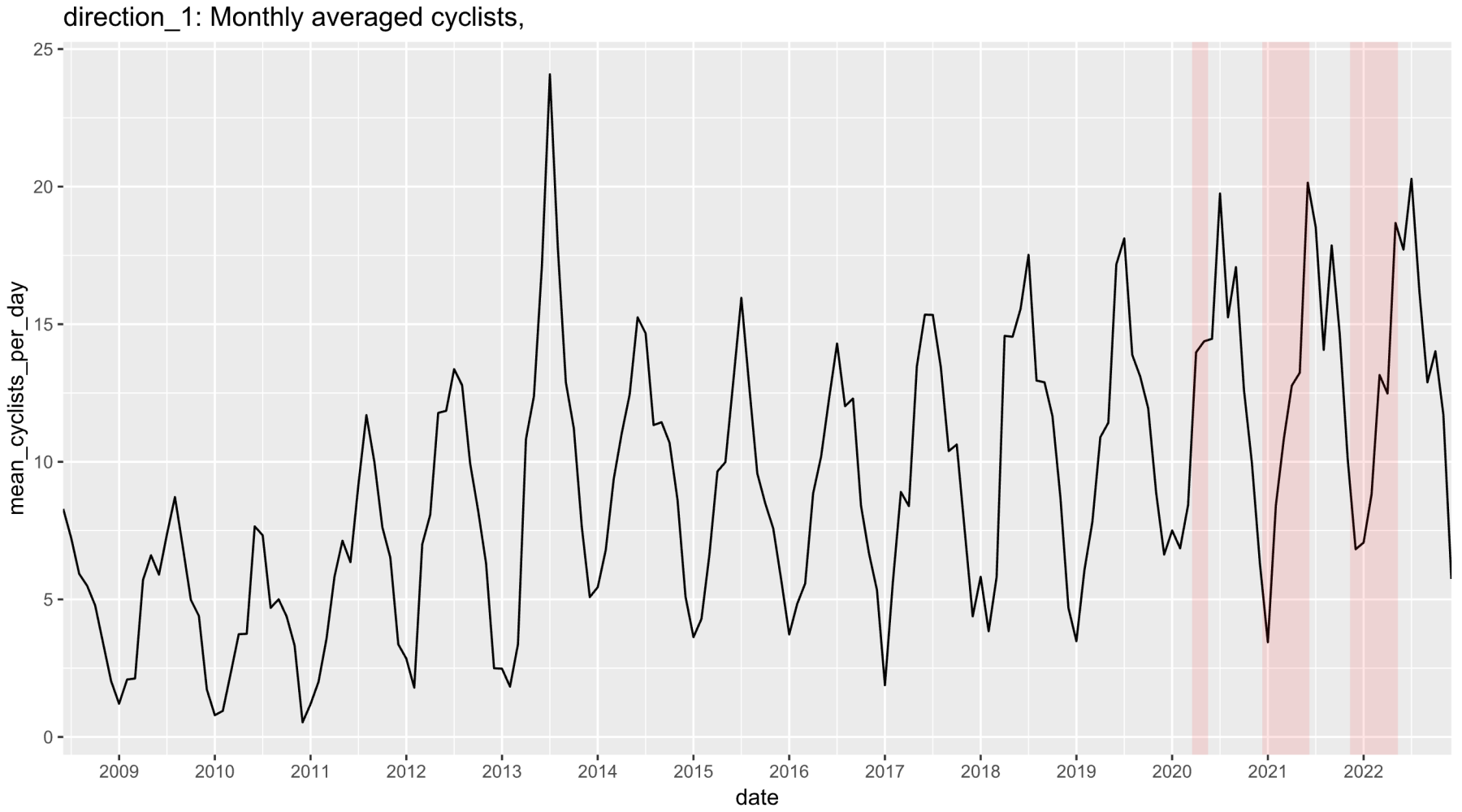
- Steady increase over the years for both directions



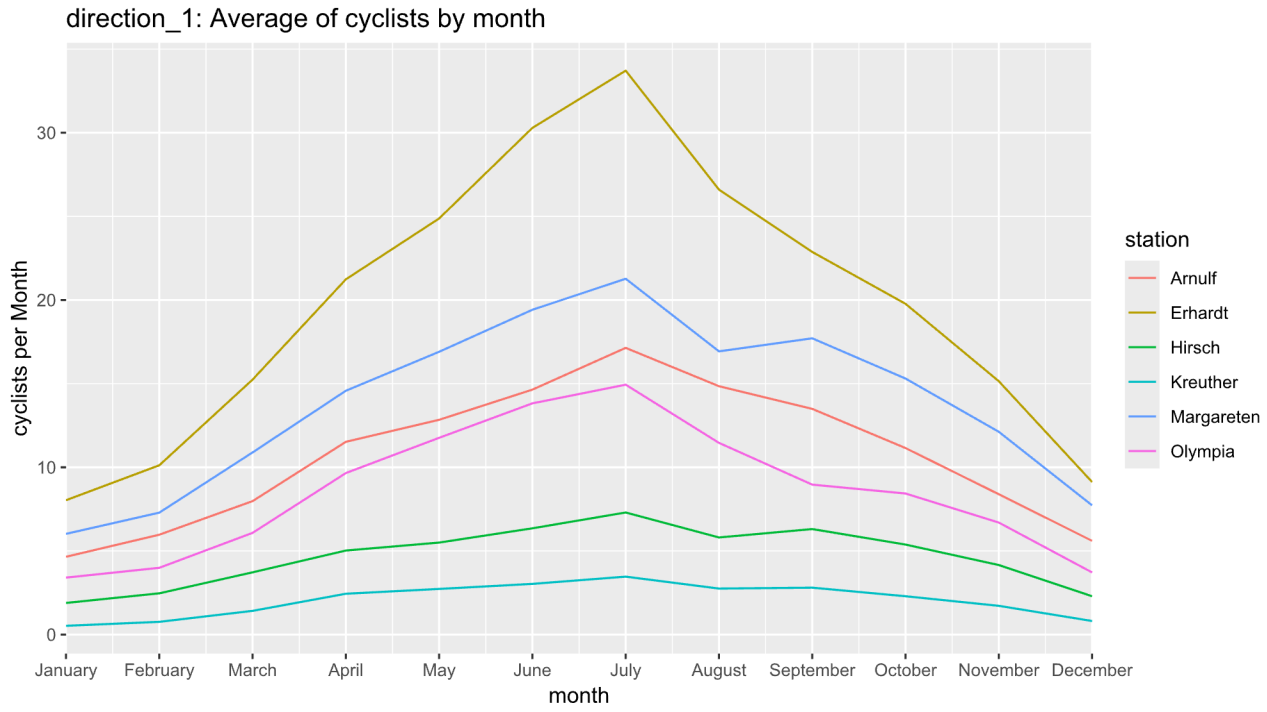


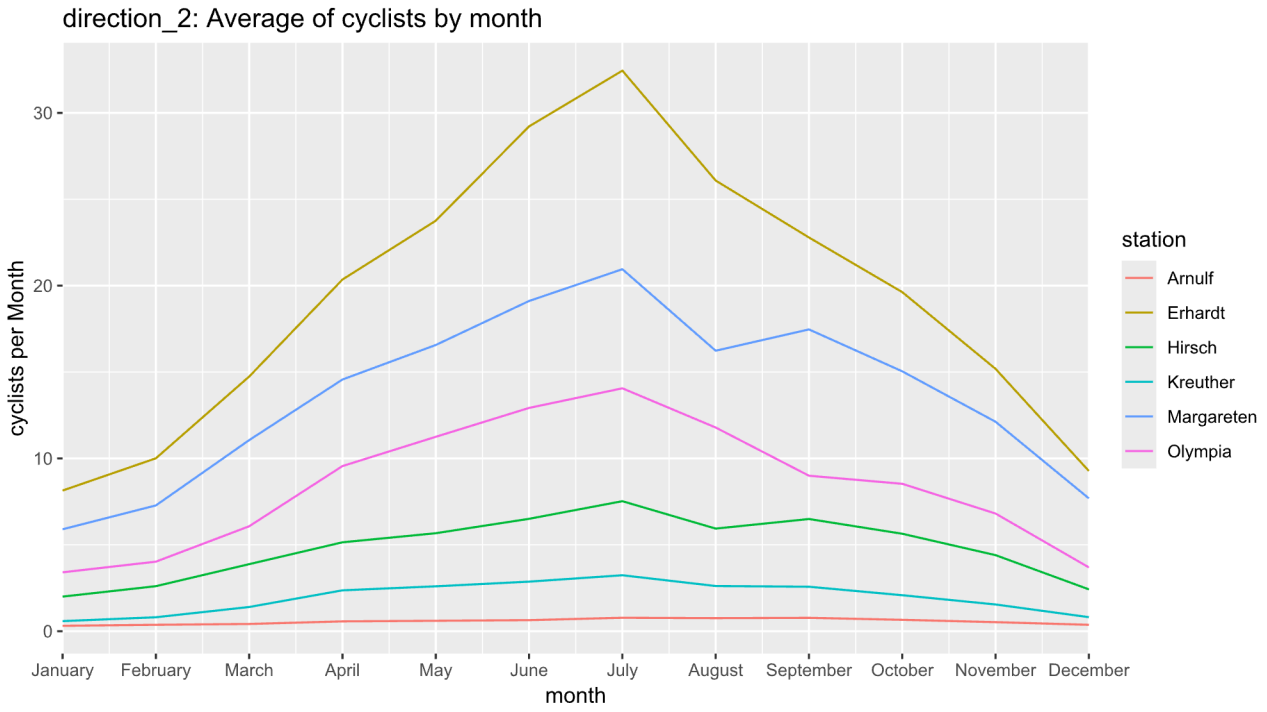
Corona Lockdowns

- Lockdown didn’t effect the quantity of cyclists
- More irregular due to Covid restrictions

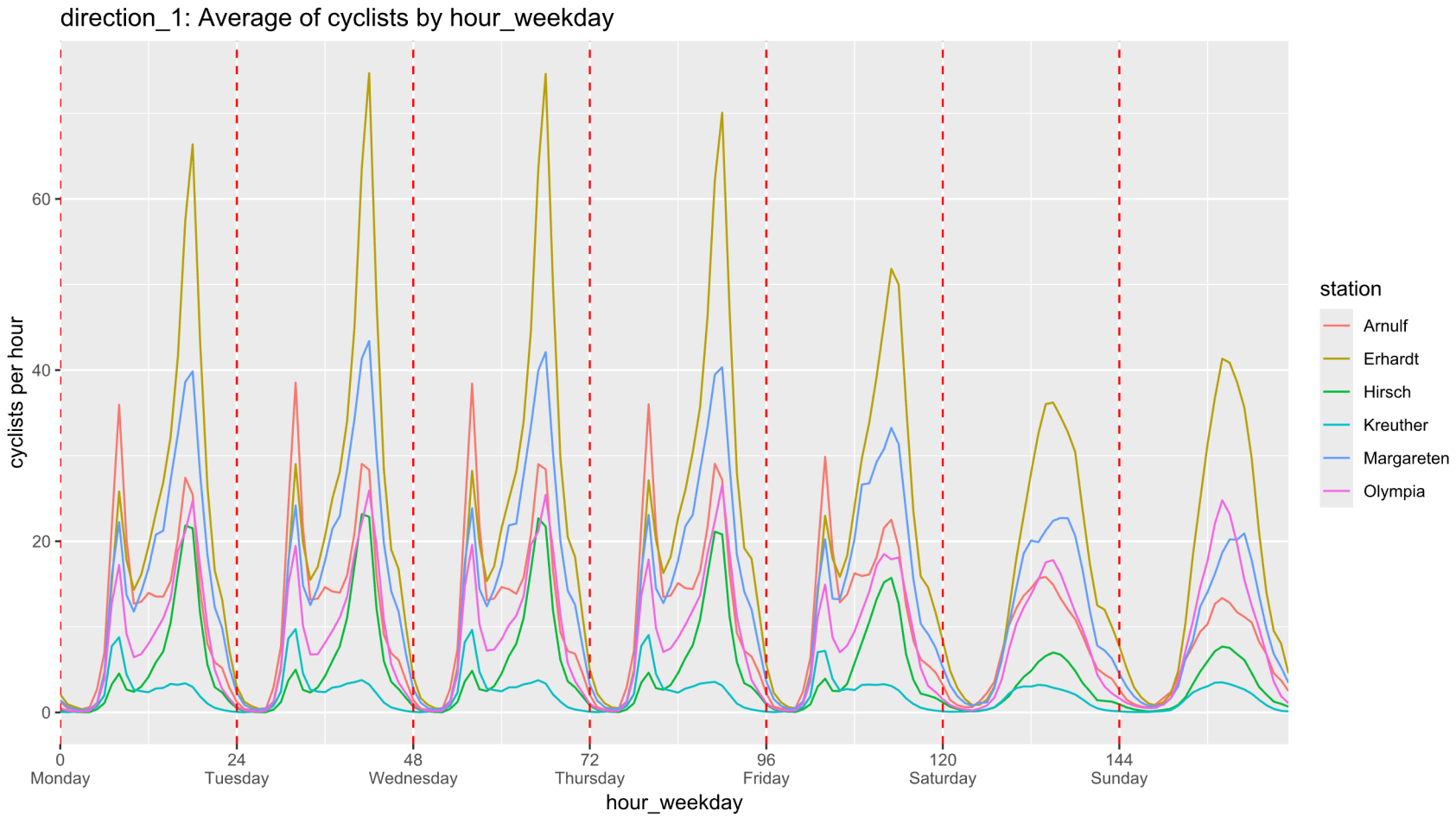


Monthly Seasonality

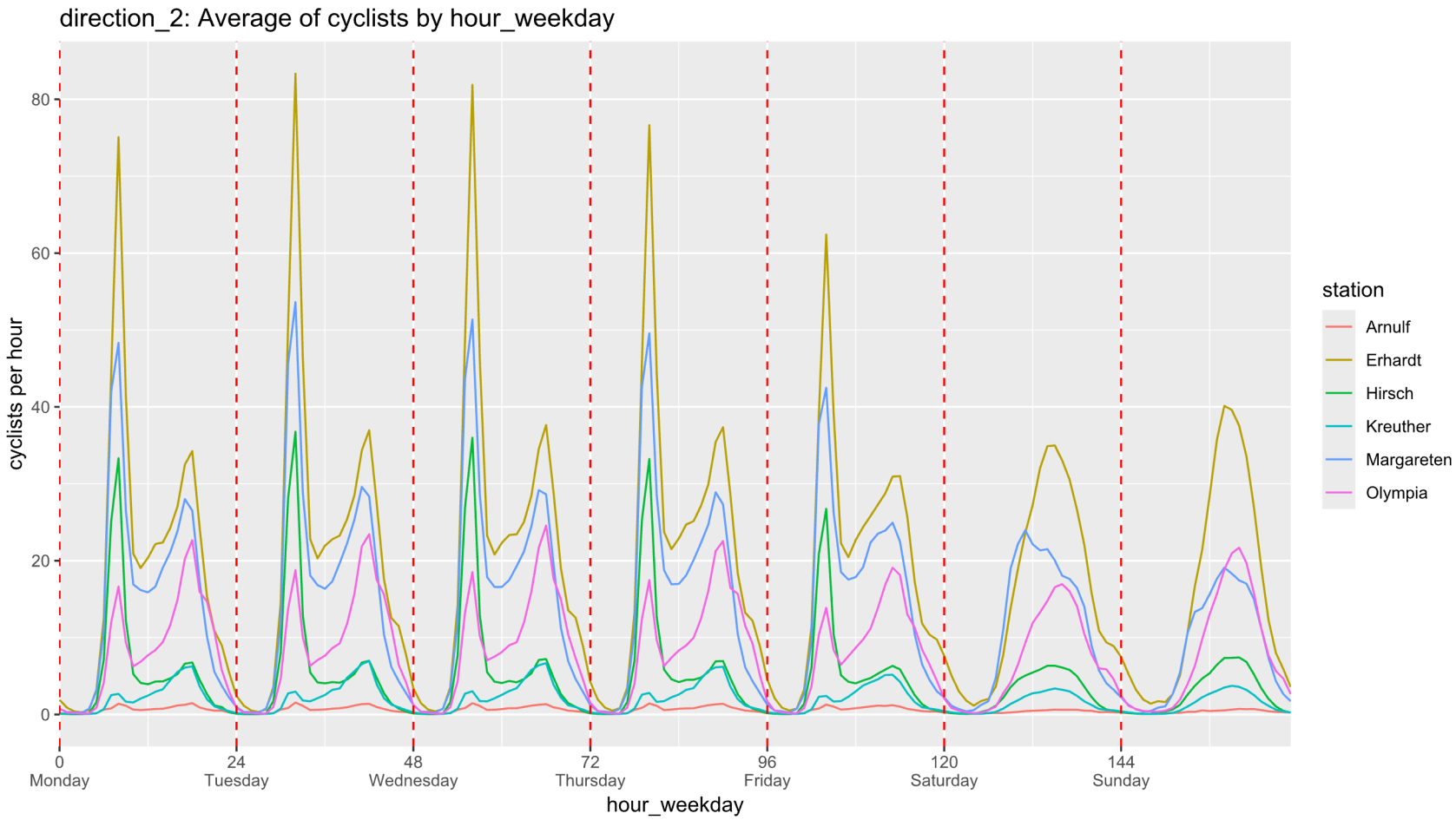




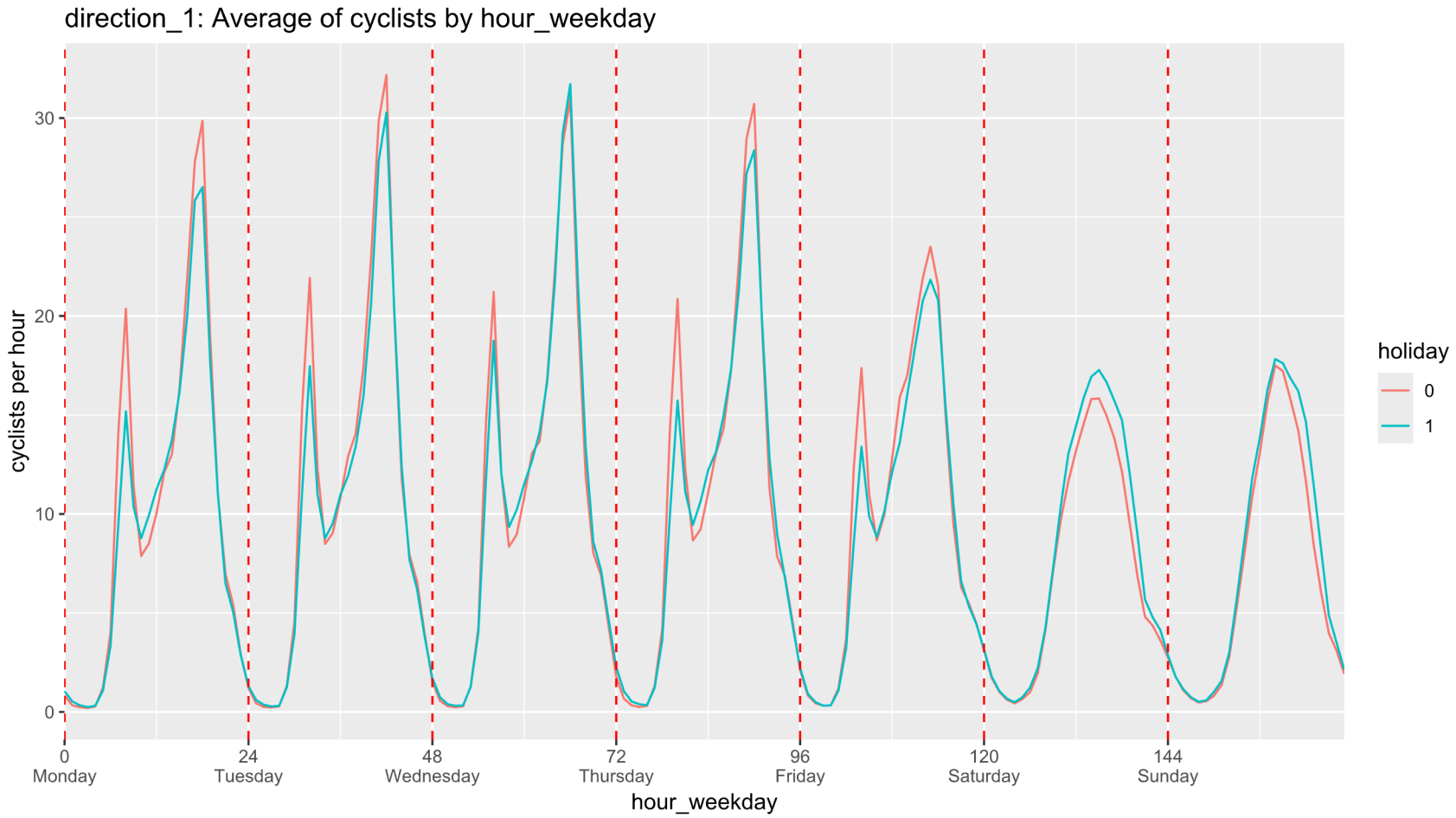
Weekly Seasonality



Weekly Seasonality



Holiday


direction_1: Average of cyclists by hour_weekday
cyclists per hour
30
20
10
0
0
Monday
24
Tuesday
48
Wednesday
72
Thursday
96
Friday
120
Saturday
144
Sunday
hour_weekday
holiday
0
1

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

4. Model

Model

- Why Generalized Additive Model (GAM)?
  - Flexible framework (linear / non-linear effects) through smooth functions.
  - Interpretability (effects of predictors on the response).
  - Can handle unevenly spaced data in time.

$$\mathbb{E}[Y] = \mu = h\left(\beta_0 + \beta_1 X_1 + \cdots \beta_p X_p + f_1(Z_1) + \cdots f_q\left(Z_q\right)\right) =: h(\eta),$$

where h is a response function and η is a linear predictor.

cf. Fitting a gam is done with '`gam()`' function from '`mgcv`' package in R

Model

- Why Negative Binomial (NB) Family?
  - Since our response is count data, we need to choose a family that can handle non-negative integer.
  - Poisson: Not proper in this case due to overdispersed count data (variance exceeds mean), checked by `dispersiontest` from AER package.
  - Quasi-Poisson: Less flexible and robust than NB mainly due to the different assumption on the variance structure. Moreover, NB is a well–defined probability distribution, which makes it straightforward to compute prediction intervals for observations.

$$\text{Quasi-Poisson: } Var[Y] = \theta\mu$$

$$\text{Negative Binomial: } Var[Y] = \mu + \frac{\mu^2}{\theta},$$

where θ is the dispersion parameter.

Model

- Specifically,

$$\log(\mu_t) = \eta_t \,,$$

where $Y_t \sim NB(\mu_t, \theta)$. Therefore, we can define the density as follows:

$$p(y; \mu, \theta) = \frac{\Gamma(\theta + y)}{\Gamma(\theta) y!} \left(\frac{\mu}{\theta + \mu}\right)^y \left(\frac{\theta}{\theta + \mu}\right)^\theta , y \in \mathbb{N}_0$$

Model

- Hence, our model is as follows:

$$\eta_t = \beta_0 + \mathbf{1}_t^{\boldsymbol{holiday}} \cdot \beta_1 + \mathbf{1}_t^{\boldsymbol{rain}} \cdot \beta_2 + f_1\left(x_t^{temperature}\right) + f_2\left(x_t^{month\_year}\right) + f_3\left(x_t^{hour\_weekday}\right)$$

- Components:
  - Y : (non-negative integer) the number of cyclists (direction 1 or 2).
  - $1^{holiday}$: (factor) categorical values of holidays: yes (1) or no (0).
  - 1^{rain} : (factor) categorical values of precipitation: no rain (0), drizzle (<0.5) or rain (≥0.5).

Model

- Hence, our model is as follows:

$$\eta_t = \beta_0 + 1_t^{holiday} \cdot \beta_1 + 1_t^{rain} \cdot \beta_2 + f_1(\boldsymbol{x_t^{temperature}}) + f_2\left(\boldsymbol{x_t^{month\_year}}\right) + f_3(\boldsymbol{x_t^{hour\_weekday}})$$

- Components:
  - $x_{temperature}$: (real) numeric values of air temperature.
  - x_{month_year} : (non-negative integer) seasonality for both monthly and yearly.
  - $x_{hour\_weekday}$: (non-negative integer) seasonality for both hourly and day of week.

Model

- Hence, our model is as follows:

$$\eta_t = \beta_0 + 1_t^{holiday} \cdot \beta_1 + 1_t^{rain} \cdot \beta_2 + \boldsymbol{f_1}\left(x_t^{temperature}\right) + \boldsymbol{f_2}\left(x_t^{month\_year}\right) + \boldsymbol{f_3}\left(x_t^{hour\_weekday}\right)$$

- Components:

$$f(x) = \sum_{j=1}^{k} \gamma_j B_j(x)$$

with basis function $B_j(\cdot)$ and its coefficients γ_j for j $= 1, \ldots, k$ where k is the basis dimension.

Model

- Splines
  - Piecewise polynomial functions used to model non-linear relationships.
  - Dividing data into segments and fitting polynomials to these segments.

Model

- Splines
  - Piecewise polynomial functions used to model non-linear relationships.
  - Dividing data into segments and fitting polynomials to these segments.
- Model Specification
  - B-spline Basis (`'bs', k=30`) for **temperature**: cubic spline order with integrated squared 2nd derivative penalties:

$$B_j^l(x) = \frac{x - \kappa_{j-l}}{\kappa_j - \kappa_{j-l}} B_{j-1}^{l-1}(x) + \frac{\kappa_{j+1} - x}{\kappa_{j+1} - \kappa_{j+1-l}} B_j^{l-1}(x)$$

Model

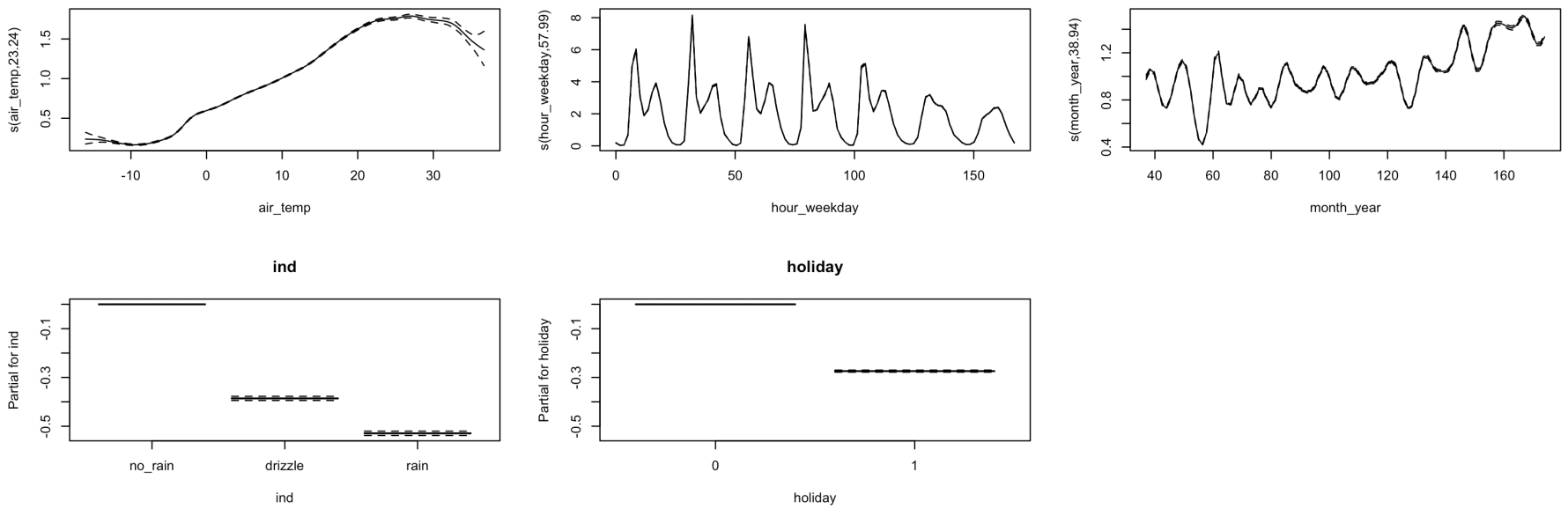
- Splines
  - Piecewise polynomial functions used to model non-linear relationships.
  - Dividing data into segments and fitting polynomials to these segments.
- Model Specification
  - B-spline Basis (`'bs', k=30`) for **temperature**: cubic spline order with integrated squared 2nd derivative penalties.
  - Cyclic cubic regression spline (`'cc', k=60`) for **hour_weekday**: cubic regression splines whose ends match, up to the 2nd derivative, ensuring connections of the start and the end of the cycle.

Model

- Splines
  - Piecewise polynomial functions used to model non-linear relationships.
  - Dividing data into segments and fitting polynomials to these segments.
- Model Specification
  - B-spline Basis (`'bs', k=30`) for **temperature**: cubic spline order with integrated squared 2nd derivative penalties.
  - Cyclic cubic regression spline (`'cc', k=60`) for **hour_weekday**: cubic regression splines whose ends match, up to the 2nd derivative, ensuring connections of the start and the end of the cycle.
  - Thin-Plate regression spline (`'tp', k=40`) for **month_year**: knot free bases for smooths, known as good for handling temporal data.

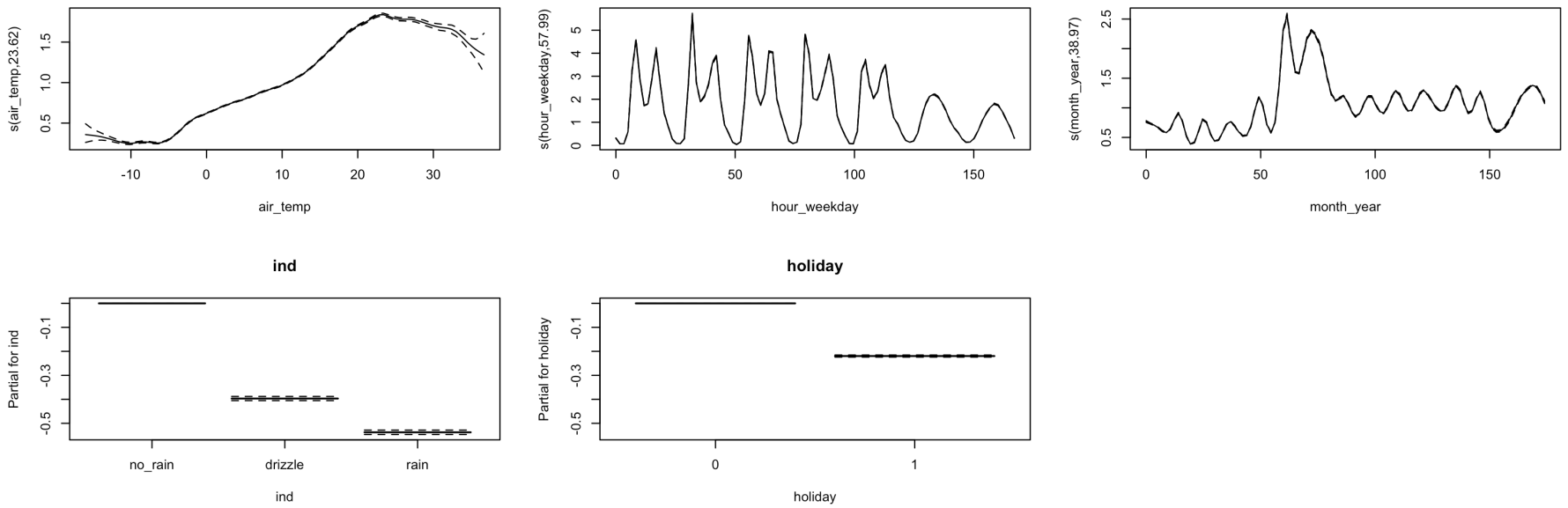
Model

- Partial effects of the estimated model **(Margareten and direction 2)**



Model

- Partial effects of the estimated model **(Arnulf and direction 1)**



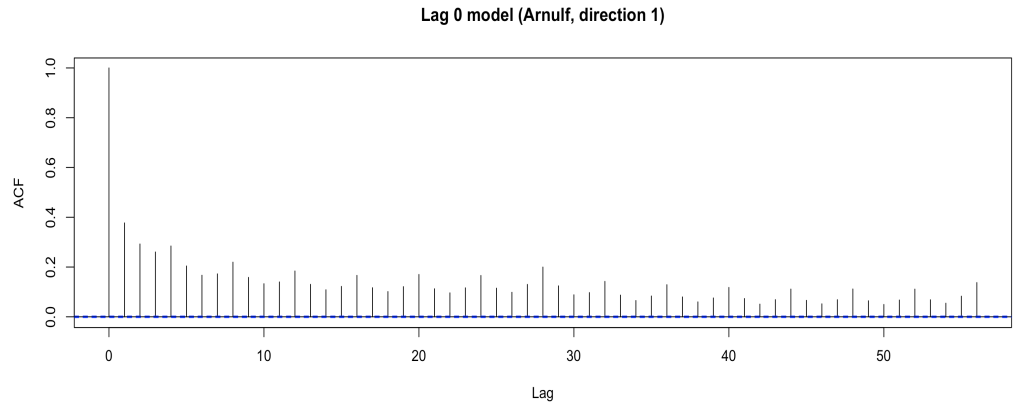
Model

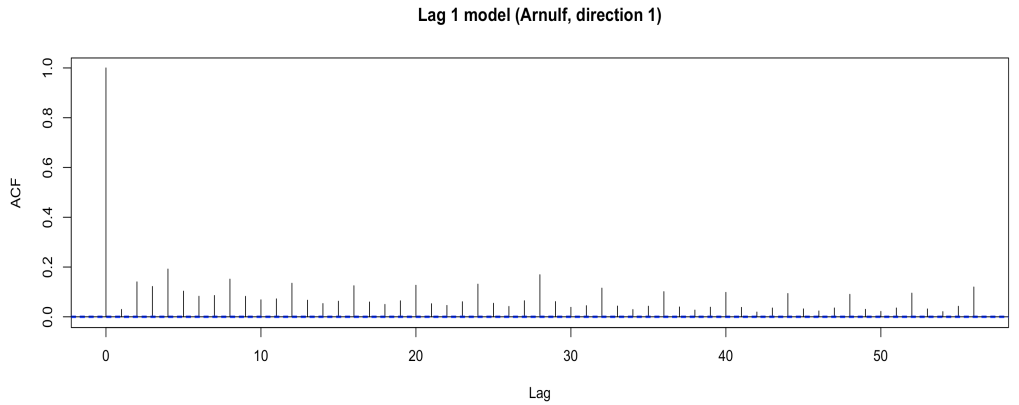
- As this model doesn't capture temporal dependency, we additionally adjust it to include a lagged response variable by one time step (i.e., to account for autoregressive behavior):

$$Y_t|Y_{t-1} \sim NB(\mu_t, \theta), \text{ and therefore}$$

$$\tilde{\eta}_t = \eta_t + f_4(y_{t-1}) \text{ and } \log(\mu_t) = \tilde{\eta}_t$$

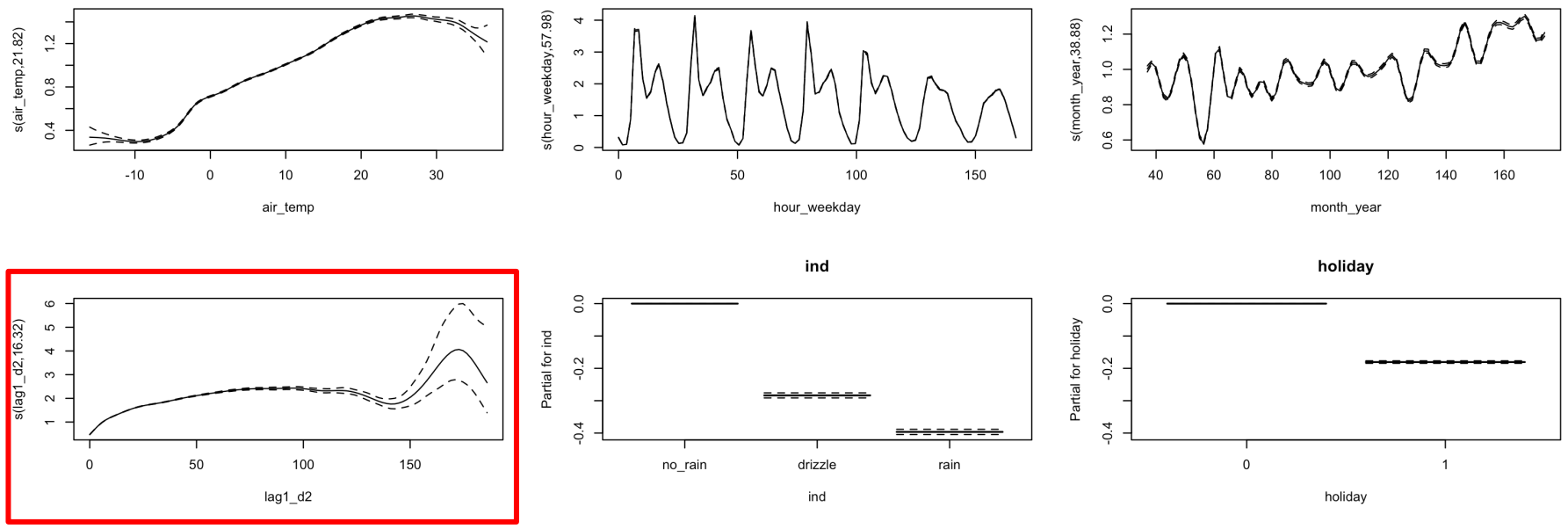
- We call this lag 1 model. This can have an effect of relaxing the higher autocorrelation of the no lag model.





Model

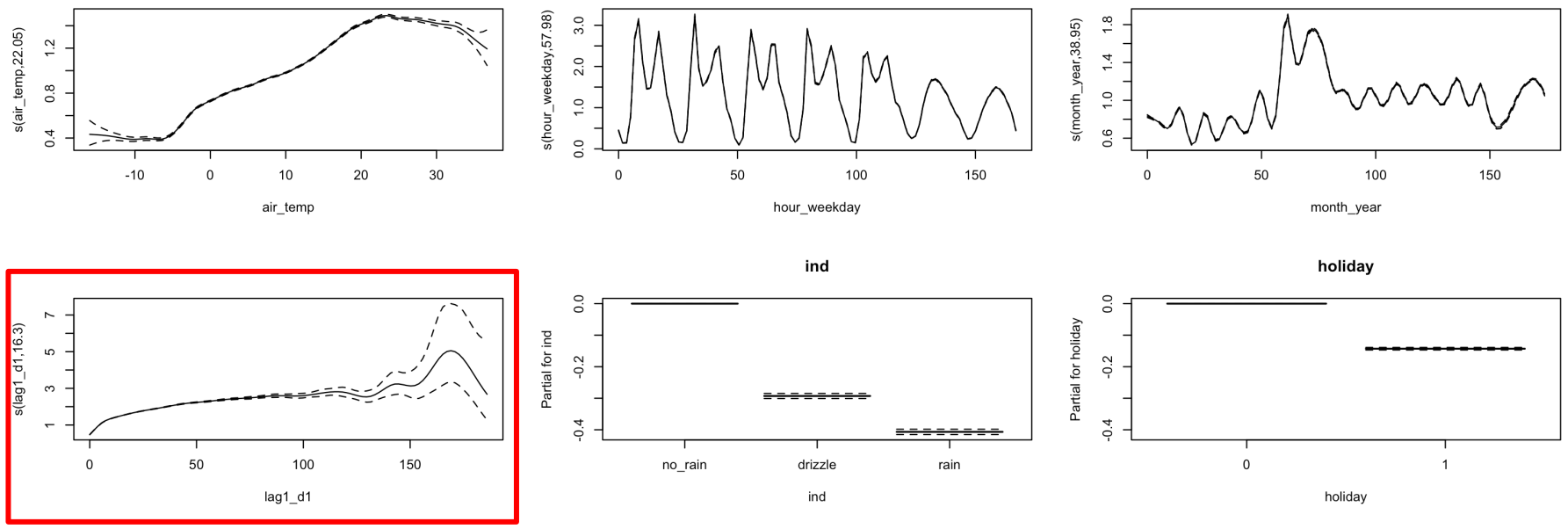
- Partial effects of the estimated lag 1 model **(Margareten and direction 2)**



(B-spline basis with k = 20)

Model

- Partial effects of the estimated lag 1 model **(Arnulf and direction 1)**



(B-spline basis with k = 20)

5. Prediction Intervals

Prediction Intervals

- A range where we expect the true value of the response variable to fall with a certain level of confidence.

Prediction Intervals

- A range where we expect the true value of the response variable to fall with a certain level of confidence.
- This accounts for both model variability and data variability:
  - Model variability - the uncertainty in the model's estimates, including the smooth and linear components; related to how well the model fits the data and the precision of the estimated parameters.

Prediction Intervals

- A range where we expect the true value of the response variable to fall with a certain level of confidence.
- This accounts for both model variability and data variability:
  - Model variability - the uncertainty in the model's estimates, including the smooth and linear components; related to how well the model fits the data and the precision of the estimated parameters.
  - Data variability - the natural variability (dispersion) in the data, which cannot be explained by the model.

Prediction Intervals

- To capture those two variabilities, we conducted the following simulation.
- This is based on the assumption on the data distribution (NB) and the asymptotic normality of MLEs.

1. Construct 100 ·(1- α)% prediction intervals for every observation

Z: design matrix containing basis function evaluations

n_{sim}: the number of samples to be drawn.

Given $\hat{\theta}$, $\widehat{\boldsymbol{\beta}}$, $\hat{\Sigma} = \widehat{Cov(\boldsymbol{\beta})}$ from the trained model:

1) Draw $\boldsymbol{\beta}_j \sim N(\widehat{\boldsymbol{\beta}},\ \hat{\Sigma}),\ \ j = 1, \dots, n_{sim}$
2) Compute $\widehat{\boldsymbol{\mu}_j} = \exp(Z\boldsymbol{\beta}_j),\ \ j = 1, \dots, n_{sim}$
3) Draw $\boldsymbol{Y} \sim NB(\widehat{\mathrm{M}}, \hat{\theta})$, where $\widehat{\mathrm{M}} = [\widehat{\boldsymbol{\mu}}_{\boldsymbol{1}}, \cdots, \widehat{\boldsymbol{\mu}}_{\boldsymbol{n_{sim}}}]$
4) Compute $\frac{\alpha}{2}, 1 - \frac{\alpha}{2}$ quantiles of $\boldsymbol{Y}$ point-wise and set them as lower bound and upper bound, respectively

Prediction Intervals (99.5%)

- A table summarizing detected outliers in direction 1, presented for each station

Direction 1	Lag 0			Lag 1		
Station	data	outliers (coverage)	zero values	data	outliers (coverage)	zero values
Arnulf	466644	4153(99.1%)	1161	466524	4359(99.1%)	576
Kreuther	484712	2777(99.4%)	44	484592	2776(99.4%)	31
Olympia	462255	3697(99.2%)	285	462205	3679(99.2%)	218
Hirsch	457454	2503(99.5%)	184	457451	2635(99.4%)	118
Margareten	401972	3563(99.1%)	984	401971	4019(99.0%)	524
Erhardt	400331	3497(99.1%)	600	400328	4358(98.9%)	397

Prediction Intervals

- To narrow down the number of outliers, we novelly introduce a method to mark strong outliers:

2. Mark strong outliers

Let y_o be a true value of detected outliers and $\widehat{y_o}$ be its corresponding fitted value.

1) Compute the distance between y_o and $\widehat{y_o}$, i.e., $d_i \coloneqq |y_{o,i} - \widehat{y_{o,i}}|$ for all $i = 1, \dots, n_{outliers}$

2) Compute criteria using the distance as follows:

Lower bound: $Q_1 - 1.5 \cdot IQR$

Upper bound: $Q_3 + 1.5 \cdot IQR$

where Q_1, Q_3 and IQR are obtained from the distance distribution.

3) Choose points whose distance falls outside either the lower bound or the upper bound.

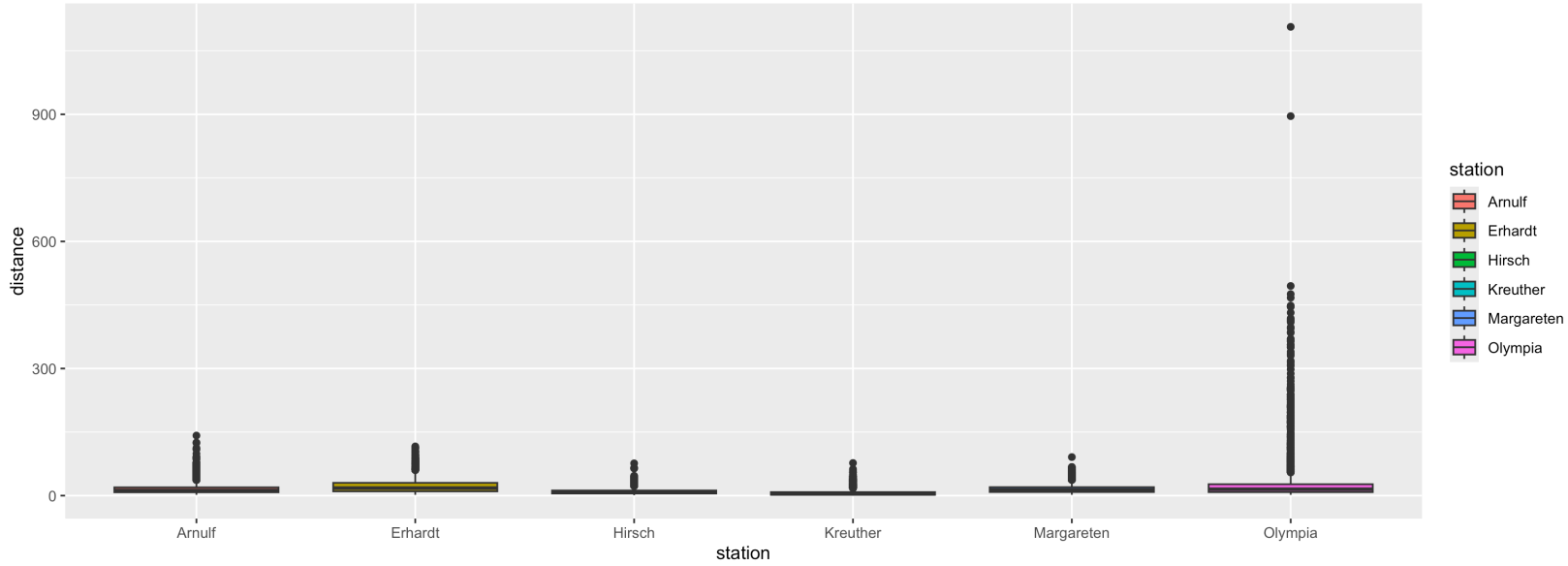
Prediction Intervals (99.5%)

- To narrow down the number of outliers, we novelly introduce a method to mark strong outliers **(lag 1)**:

Direction 1	PIs		PIs + Our method	
Station	outliers	zero values	outliers	zero values
Arnulf	4359	576	255	10
Kreuther	2776	31	192	9
Olympia	3679	218	353	2
Hirsch	2635	118	140	1
Margareten	4019	524	129	7
Erhardt	4358	397	166	6

Prediction Intervals

- Boxplots of distance for each station **(direction 1 / lag 1 model)**



Prediction Intervals

- Example of strong outliers **(lag 1 model)**

	ds	holiday	ind	air_temp	station	direction_1	y_pred
	<char>	<num>	<fctr>	<num>	<char>	<num>	<num>
1:	2018-04-27 20:00:00	0	no_rain	17.0	Hirsch	12	6
2:	2018-04-27 20:15:00	0	no_rain	17.0	Hirsch	10	8
3:	2018-04-27 20:30:00	0	no_rain	17.0	Hirsch	6	8
4:	2018-04-27 20:45:00	0	no_rain	17.0	Hirsch	7	7
5:	2018-04-27 21:00:00	0	no_rain	17.4	Hirsch	2	4
6:	2018-04-27 21:15:00	0	no_rain	17.4	Hirsch	3	4
7:	2018-04-27 21:30:00	0	no_rain	17.4	Hirsch	4	4
8:	2018-04-27 21:45:00	0	no_rain	17.4	Hirsch	3	5
9:	2018-04-27 22:00:00	0	no_rain	16.7	Hirsch	4	3
10:	2018-04-27 22:15:00	0	no_rain	16.7	Hirsch	5	4
11:	2018-04-27 22:30:00	0	no_rain	16.7	Hirsch	6	4
12:	2018-04-27 22:45:00	0	no_rain	16.7	Hirsch	2	4
13:	2018-04-27 23:00:00	0	no_rain	15.4	Hirsch	0	2
14:	2018-04-27 23:15:00	0	no_rain	15.4	Hirsch	2	2
15:	2018-04-27 23:30:00	0	no_rain	15.4	Hirsch	2	3
16:	2018-04-27 23:45:00	0	no_rain	15.4	Hirsch	67	3

Prediction Intervals

- Example of strong outliers **(lag 1 model)**

	ds	holiday	ind	air_temp	station	direction_1	y_pred	outlier
	<char>	<num>	<fctr>	<num>	<char>	<num>	<num>	<char>
1:	2014-07-01 09:00:00	0	no_rain	17.7	Arnulf	82	76	no
2:	2014-07-01 09:15:00	0	no_rain	17.7	Arnulf	72	77	no
3:	2014-07-01 09:30:00	0	no_rain	17.7	Arnulf	48	75	no
4:	2014-07-01 09:45:00	0	no_rain	17.7	Arnulf	0	67	yes
5:	2014-07-01 10:00:00	0	no_rain	18.4	Arnulf	0	50	yes
6:	2014-07-01 10:15:00	0	no_rain	18.4	Arnulf	0	11	yes
7:	2014-07-01 10:30:00	0	no_rain	18.4	Arnulf	0	11	yes
8:	2014-07-01 10:45:00	0	no_rain	18.4	Arnulf	0	11	yes
9:	2014-07-01 11:00:00	0	no_rain	18.7	Arnulf	0	35	yes
10:	2014-07-01 11:15:00	0	no_rain	18.7	Arnulf	0	9	yes
11:	2014-07-01 11:30:00	0	no_rain	18.7	Arnulf	0	9	yes
12:	2014-07-01 11:45:00	0	no_rain	18.7	Arnulf	0	9	yes
13:	2014-07-01 12:00:00	0	no_rain	20.6	Arnulf	0	43	yes
14:	2014-07-01 12:15:00	0	no_rain	20.6	Arnulf	0	10	no
15:	2014-07-01 12:30:00	0	no_rain	20.6	Arnulf	0	10	yes
16:	2014-07-01 12:45:00	0	no_rain	20.6	Arnulf	0	10	no
17:	2014-07-01 13:00:00	0	no_rain	21.3	Arnulf	0	52	yes
18:	2014-07-01 13:15:00	0	no_rain	21.3	Arnulf	0	10	yes
19:	2014-07-01 13:30:00	0	no_rain	21.3	Arnulf	0	10	yes
20:	2014-07-01 13:45:00	0	no_rain	21.3	Arnulf	0	10	yes
21:	2014-07-01 14:00:00	0	no_rain	20.4	Arnulf	0	49	yes
22:	2014-07-01 14:15:00	0	no_rain	20.4	Arnulf	0	11	yes
23:	2014-07-01 14:30:00	0	no_rain	20.4	Arnulf	20	11	no

6. Limitation

Limitation

- Models
  - Machine/Deep Learning Models
  - Classic Time series analysis models with eventually imputations
  - Prophet
- Hyperparameter optimization
  - Number of basis functions
  - Spline type

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

References

[1] Fahrmeir et al. Regression Models, Methods and Applications. Springer, 2013.

[2] Simon Wood. Generalized Additive Models, an Introduction with R (2nd). CRC, 2017.

[3] Colin Cameron and Trivedi. Regression Analysis of Count Data. Cambridge, 1998.

[4] Gavin Simpson's blog posts: https://fromthebottomoftheheap.net/blog/.

[5] Lauer et al. Prospective forecasts of annual dengue hemorrhagic fever incidence in Thailand, 2010–2014. https://doi.org/10.1073/pnas.1714457115