Data Plausibility Check for Cycling Traffic Data with Generalized Additive Models

Juntae Kwon, The Anh Vu
theanh_v99@yahoo.de
junjun.kwon@gmail.com

Statistical Consulting
for the Institute of Statistics at LMU Munich
in cooperation with
Dr. Christine Reintinger
from the Mobility Department in Munich
supervised by
Prof. Dr. Helmut Küchenhoff

Munich, May 10, 2024

Contents

Abstract

In Collaboration with the Mobility Department in Munich, we used cycling count data from six counting stations around Munich to find implausible data points. To this end, we additionally used weather information to model the data with a generalized additive model. By introducing new compound variables, we mitigated the problem with time-varying variables and their autocorrelated nature. We then found outliers by constructing a prediction interval around each data point by simulation, and tried to circumvent a problem that the number of detected outliers is large due to the vast amount of data. Based on solid statistical theories, our method can yield reliable results in a simple and effective way.

Introduction

The cycling traffic counts of recent years have shown a significant increase in bicycle traffic in Munich. Since the data was mostly collected only on individual days, random influences such as different weather conditions can not be accounted for in comparisons. For this reason bicycle counting stations were first established in Munich in 2008. Although these counting stations enable the continuous monitoring of the development of bicycle traffic volume, external factors such as failure of the sensors or construction works can influence the data. Together with the Mobility Department in Munich, we used the cycling traffic data available from the Munich Open Data Portal to build a model to find implausible data points.

The model we employed is a generalized additive model consisting of features from the bicycle count data and available weather data with smoothing splines. With the fitted model we construct a prediction interval using simulation to capture the model and data variability. This was done for each data point, so that we were able to find outliers based on the interval. We then filtered it further by marking strong outliers, which was achieved by calculating a range distribution of all outliers based on the interquartile range.

The report provides an analysis of the data structure of the observations, followed by a detailed explanation of the model. In particular, it addresses the challenge of deriving prediction intervals and the pitfall of too many detected outliers. The code can be found in this *github repository*.

Dataset

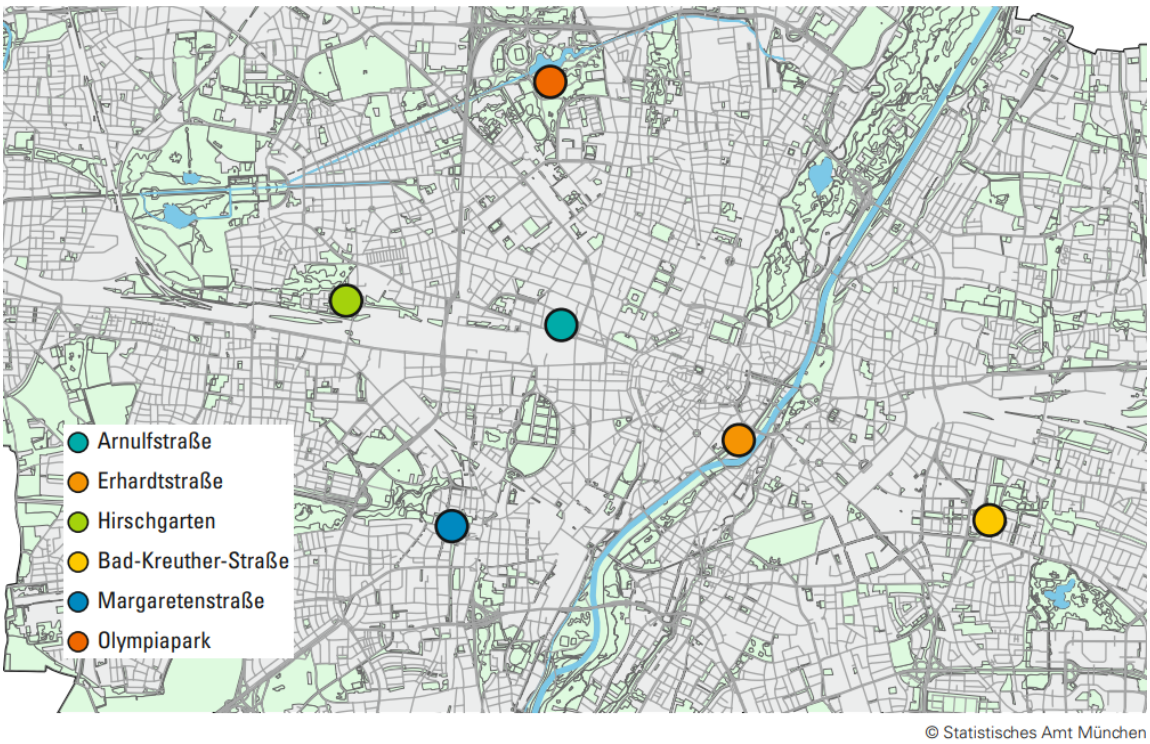


Figure 3.1: Counting Station Map

An important component for implementing new projects is the statistically collected data of cyclists in Munich. Since the summer of 2008, bicycle usage has been measured at various locations in Munich. By now, there are six permanent measuring points in Munich that record passing bicycles. The locations are near the Munich Central Bus Station (ZOB) on Arnulfstraße, on the Isar Cycle Path on Erhardtstraße, on Rudolf-Harbig-Weg in the Olympia Park, and above the Hirschgarten S-Bahn sta-

tion on Birketweg. Additional measuring points exist at Harras on Margaretenstraße, east of the Innsbrucker Ring on Bad-Kreuther-Straße or Joseph-Hörwick-Weg close to Berg am Laim. These permanent measuring points, primarily established for planning purposes, capture all passing bicycles through sensors embedded in the ground. The number of cyclists is counted 24 hours a day, 365 days a year in 15 minute intervals. The data is publicly available on the Open Data Portal from the Mobility Department in Munich. Additionally to this data we utilized the public weather data from the German Weather Service (DWD) to add more information about the cycling behavior for the model.

In the following sections, we will explain the variables we used for the model and examine their structures, which are necessary to be captured by the model in order to show a good fit.

3.1 Time variables

For the time variables, we have unevenly spaced data in time to model the number of cyclists. This includes:

- year
- month
- day
- hour
- day of week (Monday - Sunday)

The problem with using time variables is the interaction between them, as they are usually correlated with each other, making modeling difficult when used separately. For this reason, we have introduced two composite variables with the data:

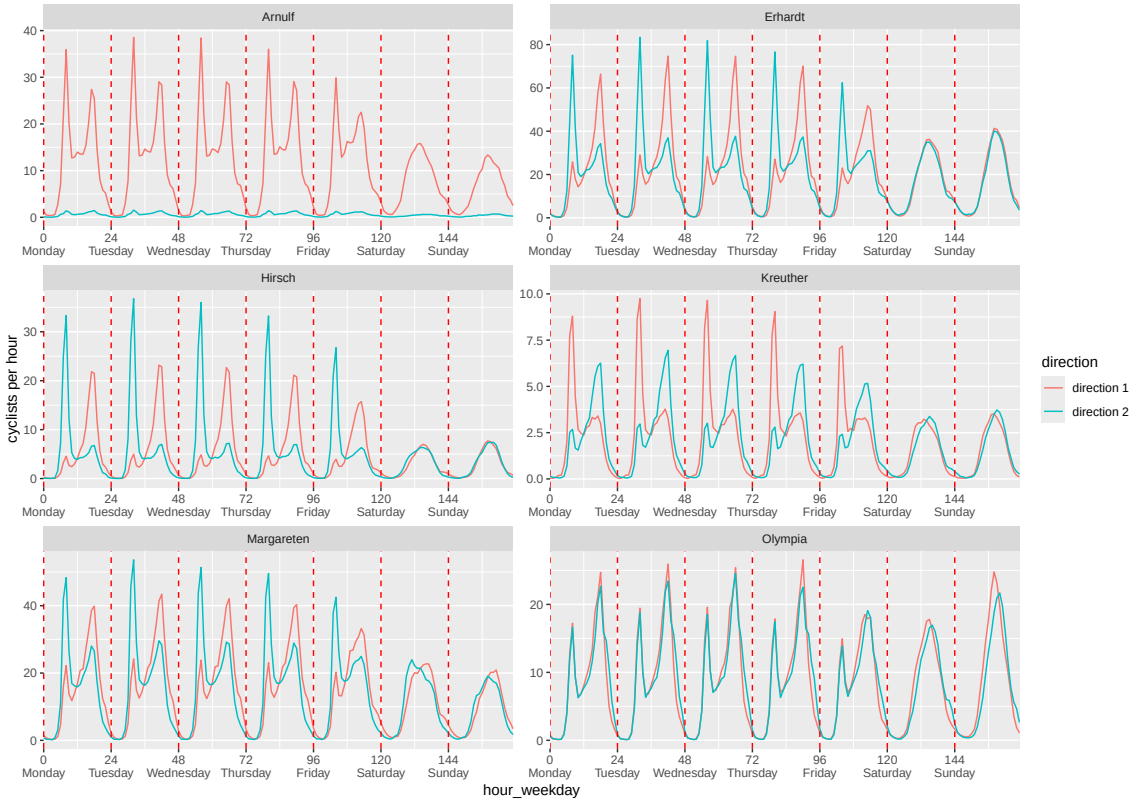


Figure 3.2: cycling volume aggregated by hour_weekday

- hour_weekday := (day of week) * 24 + hour
- month_year := (year – year of first observation) * 12 + (month – month of first observation)

hour_ weekday is the compound of hour and day of week. It encapsulates the cycling behaviour of each day of the week. Figure 3.2 shows the average cycling traffic volume of the observed values over the week separated by directions. From Monday to Friday, peaks occur during the main commuting hours between 06:00 and 09:00 in the morning and between 16:00 and 19:00 in the evening. In the morning, the majority of commuters travel in direction 2 (city inwards), while in the afternoon they travel in direction 1 (city outwards).

The bike counting station at Olympia park is crossed the same amount in both directions. Bad-Kreuther-Straße is the only station where the directions are reversed. For Arnulfstraße there is only a one way bike lane for direction 1. The other direction is for people going the wrong way, thus resulting in close to zero cyclists. On weekends, cycling traffic tends to concentrate more in the afternoon to early evening and is less than during the week. There are no differences regarding directions on weekends. Generally we can say that the bicycle traffic volume is the highest on Tuesday and Wednesday [4].

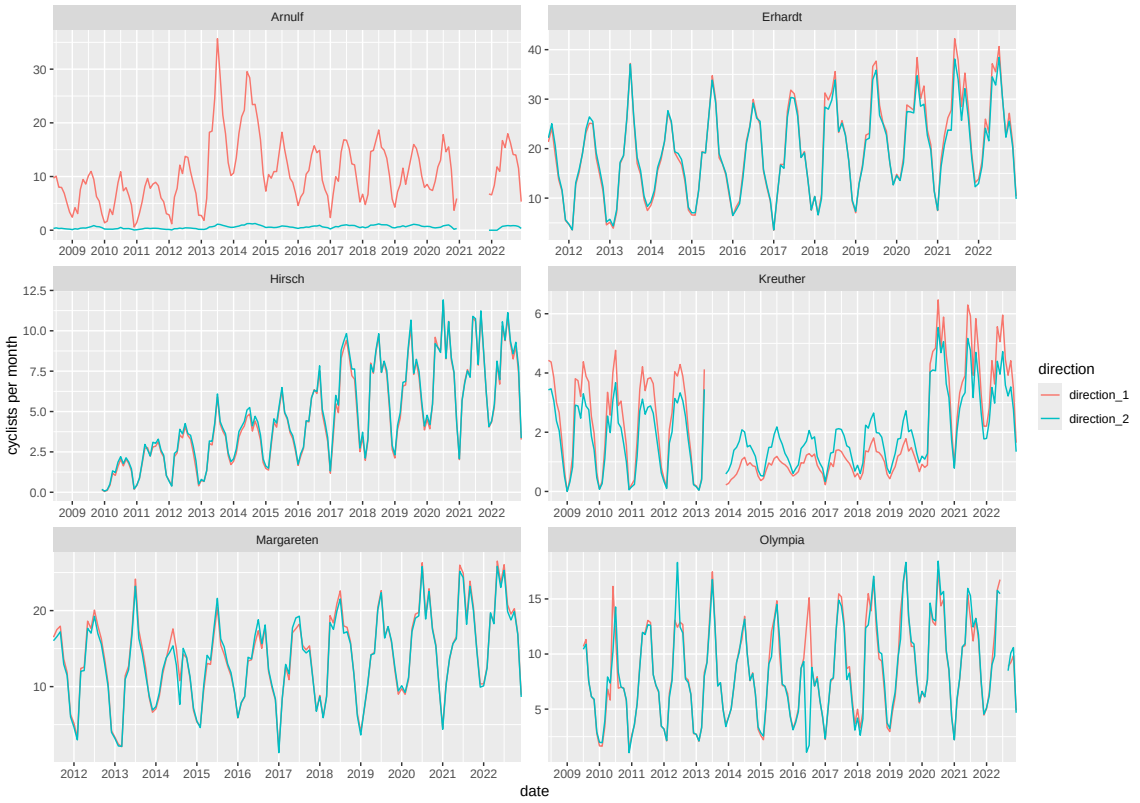


Figure 3.3: cycling volume aggregated by month_year

The idea behind *month_year* is the same. It combines the interaction between month and year into one variable. This allows us to include the yearly trend in our month variable. All in all, we can say that the number of cyclists in Munich has

increased over the years, but very irregularly, as can be seen in Figure 3.3. There are several reasons for such irregularities or missing data. In the data, the NA values are implausible data points from previous validations by the mobility department. For example, they come from construction or damage to the stations.

The irregularities are largely unexplained for us, except for the start of 2020 until around July 2021. We can attribute those to the COVID-19 pandemic. Interestingly there is not really a decline of cyclists even during the official lockdown periods, but multiple smaller declines due to social and work restrictions and fear of infection.

3.2 Weather variables

The public weather data from the German Weather Service provides us with a lot of new variables. To avoid a overly complex model, we restricted the variables to the most common ones:

- precipitation (mm)
- air temperature (°C)
- wind speed (m/s)
- sun time per day (h)

We noticed that wind speed and sun time per day have plenty of missing data resulting in the loss of information when merging with the cycling data. Due to that, we decided to only use air temperature and precipitation for our model. The only change we made is to modify precipitation to a categorical variable with three levels:

- no rain (0 mm)
- drizzle(> 0 & < 0.5 mm)
- rain ($>= 0.5$ mm)

In Figure 3.4 we can see most data points gathering around no precipitation and traffic volume rapidly falling until around 10mm precipitation. We can also notice a lot of single outliers, therefore simplifying with a categorical variable makes sense. The categories are split such that they are around the same size. An alternative approach would be to implement more variables with the official intensity definitions [5]:

- Light rain: trace and 2.5 mm per hour
- Moderate rain: 2,5 mm - 10 mm per hour.
- Heavy rain: $>$ 10 mm per hour
- violent rain: 50 mm per hour

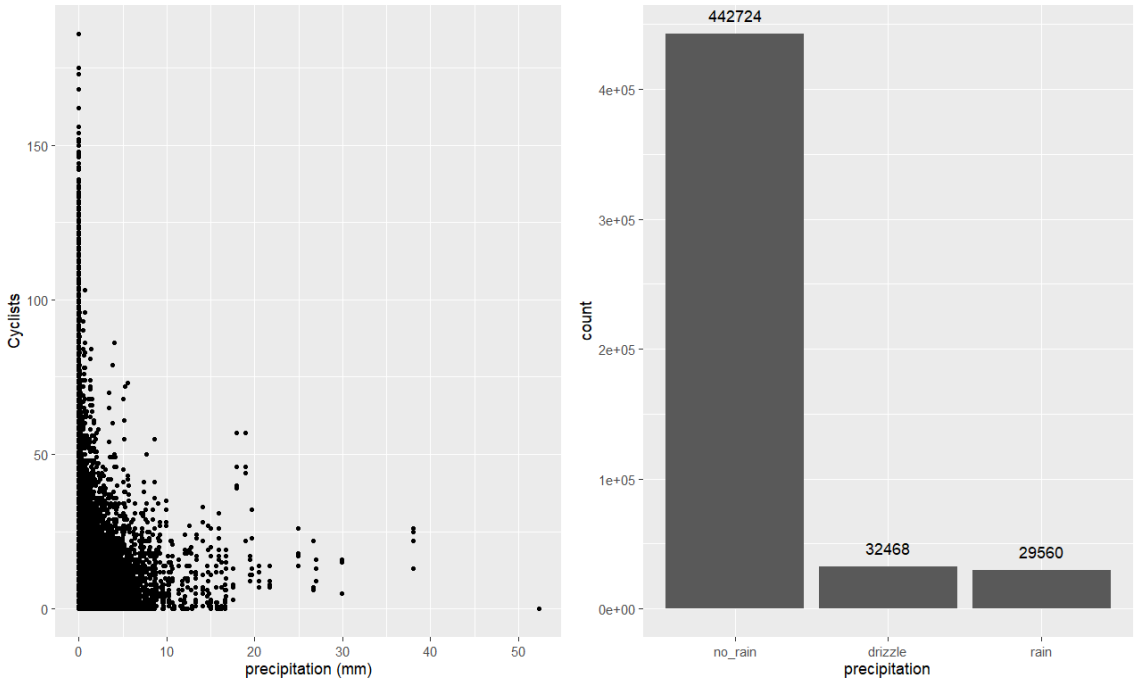


Figure 3.4: distribution of precipitation continuous and categorical

The air temperature in Munich is between a humid oceanic climate and dry continental climate (Cfb/Dfb). It is characterized by a relatively cold winter and a warm

summer with average 20 degrees Celsius. The temperature is gradually rising over the year with July being the hottest month of the year and then falling until January with an average temperature of 1 degree Celsius. The monthly seasonality can mostly be attributed to the air temperature (see Figure 3.5). As expected, the cycling traffic increases with warmer temperatures, except for August, where the summer holiday starts for all students.

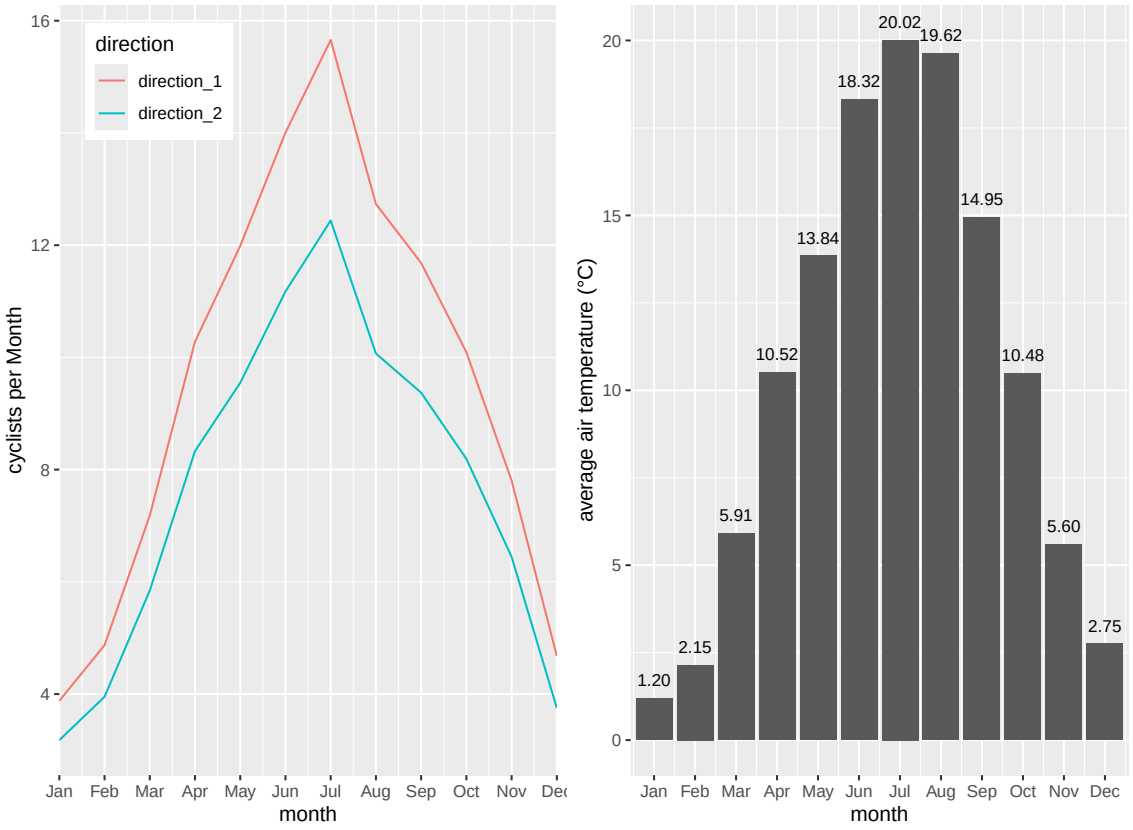


Figure 3.5: average air temperature per month

3.3 Holiday

In addition to the many public holidays, there are 6 holiday periods in Munich that take up 75 working days per year, even more if weekends are included. In our analysis,

we did not distinguish between school and public holidays. Instead, we created a binary holiday variable for both types of holidays.

Figure 3.6: holiday vs non holiday (smoothing by LOESS)

Another observation is the irregularity of the holidays throughout the year. Especially before and after holiday periods, we can see the traffic volume peaking in both directions. This is potentially an artifact of the aggregation over the years and the yearly shifts of the holiday seasons. The smoothing lines in Figure 3.6 show a very rough trend of the cyclists separated by holiday and non-holiday. The noticeable difference are the less cyclists during holidays. For the sake of simplicity, we used a binary variable to model the holiday effects. Further steps to expand on could be:

- separate public and school holidays
- individual effect for each holiday
- include interaction with time variables

Model

To comprehend the relationship between a response variable and predictors, various models, such as a standard linear regression or a deep learning model, can be utilized. While the deep learning model is more complex and often more adept at capturing intricate relationships, it is often more challenging to interpret due to its structure. In the context of the given tasks, which range from data plausibility checks to the impacts of weather on cycling traffic volume, a statistical model, which is one of the most effective options for such tasks, was employed.

4.1 Generalized Linear Model

Let $\mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$ be a predictor vector and a response variable, respectively. Since we focus on modeling the expected value of the response variable, i.e., mean regression, the model can be written as follows:

$$\mu =: \mathbb{E}[y] = h(\eta)$$

where h is a response function and η is a linear predictor that is a linear combination of predictors with coefficients. Since the target variable is count data, implying a non-integer value, it requires to adequately adjust the scale of η by choosing a proper response function. To this end, we chose an exponential.

In addition to the type of the response function, we have to make a suitable assumption on a statistical distribution over the response variable. For modeling count data, a Poisson distribution is often chosen because of its familiarity. Nevertheless, it has

a crucial problem with respect to dispersion: the mean and variance of a Poisson distribution are theoretically equal to each other. In practice, it is a challenge to satisfy this condition and the variance is normally greater than the mean, which is called overdispersion. This phenomenon can be assessed by using the function `dispersiontest` from AER package in R.

By scaling out the variance, other distributions are also available to address this problem. For example, a quasi-poisson distribution is one of the options, with the variance directly proportional to the mean; $Var[y] = \theta\mu$, where θ is a dispersion parameter. Another well-known choice is a negative-binomial distribution, where its variance is quadratic in its mean; $Var[y] = \mu + \frac{\mu^2}{\theta}$. Clearly, the variance of a negative-binomial distribution (NB2) is more flexible, which can lead to better performance. This is also useful in our case, because it is a well-defined probability distribution and makes it easier to construct prediction intervals. For these reasons, we decide to assume a (conditional) negative-binomial distribution over the response variable. Hence, the probability density is defined as

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

where $\mu = \exp(\eta)$.

Now, let us briefly step into how to find the best parameters. Let our samples $y_i|\mathbf{x}_i \overset{\text{ind}}{\sim} NB(\mu_i, \theta)$ for all $i = 1, \ldots, n$. Then, the likelihood function is,

$$L(\mu_1, \ldots, \mu_n, \theta) = \prod_{i=1}^{n} \frac{\Gamma(\theta + y_i)}{\Gamma(\theta) y_i!} \left(\frac{\mu_i}{\theta + \mu_i}\right)^{y_i} \left(\frac{\theta}{\theta + \mu_i}\right)^\theta.$$

For simplicity, the log likelihood function is then given as

$$l(\mu_1, \ldots, \mu_n, \theta) = \sum_{i=1}^{n} \log \frac{\Gamma(\theta + y_i)}{\Gamma(\theta) y_i!} \left(\frac{\mu_i}{\theta + \mu_i}\right)^{y_i} \left(\frac{\theta}{\theta + \mu_i}\right)^\theta.$$

This can be rewritten in terms of parameters we estimate:

$$l(\boldsymbol{\beta}, \theta) = \sum_{i=1}^{n} \log \frac{\Gamma(\theta + y_i)}{\Gamma(\theta) y_i!} \left(\frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{\theta + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})} \right)^{y_i} \left(\frac{\theta}{\theta + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})} \right)^{\theta},$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^\top$. Obtaining the estimates of the maximum likelihood estimator of $\boldsymbol{\beta}$ and θ involves an alternating iteration process. For a given θ, the (log) likelihood function can be optimized via some numerical methods such as Fisher scoring or Newton-Raphson with respect to $\boldsymbol{\beta}$. Then, for fixed $\boldsymbol{\beta}$, the dispersion parameter can be adjusted using score and information iterations. This alternated process is repeated until convergence of both.

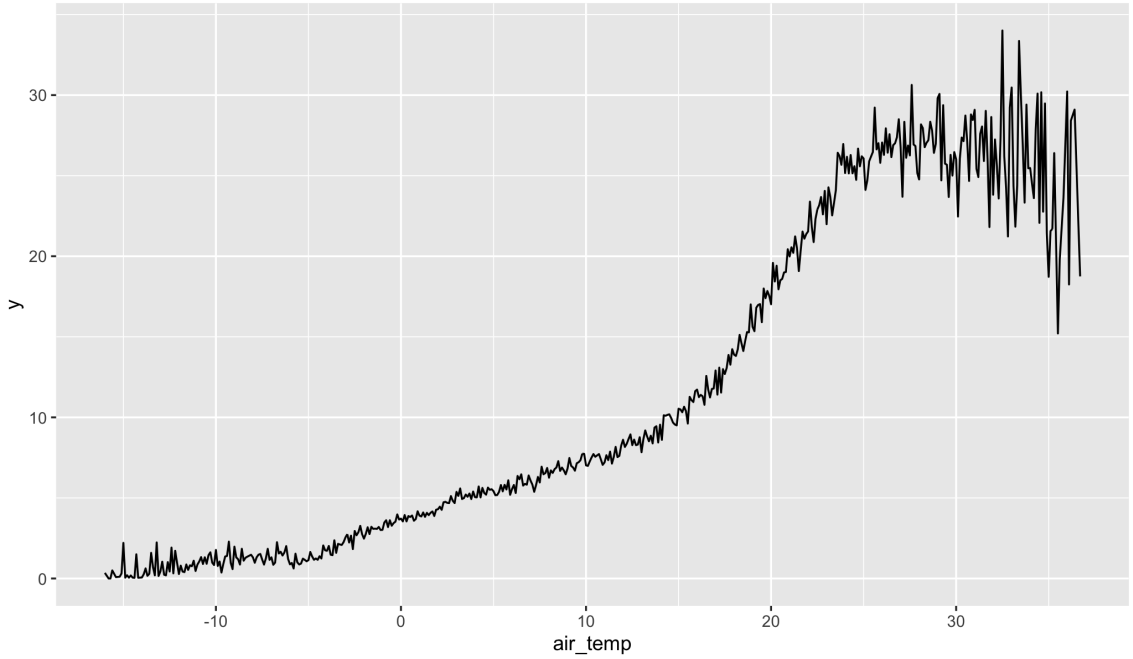


Figure 4.1: the empirical relationship between air temperature and the number of cyclists.

However, GLMs are not that flexible since they only account for linear effects of predictors on the response variable. Some predictors might not be linearly related with the response variable at all. For instance, the number of cyclists tends to

increase with rising air temperatures until it becomes too hot, at which point it starts to drop as illustrated in Figure 4.1. In the next section, we will go into detail on how to overcome this limitation.

4.2 Generalized Additive Model

While GLMs are powerful, they are limited to linear relationships between the transformed response and predictors. This assumption can be overly restrictive when dealing with real-world data that often exhibits non-linear patterns. This is where Generalized Additive Models (GAMs) come into play. Unlike GLMs, GAMs incorporate non-linear effects through the use of smooth functions, such as splines. This allows each predictor to have a customized non-linear relationship with the response variable, enhancing the model's flexibility and capability to capture complex patterns in data.

4.2.1 Splines

To model non-linear effects of continuous predictors on the response variable, we now assume a more complex relationship between them. What we have covered so far is called a parametric approach, meaning a simple structure to estimate the relationship. In other words, there is a finite number of parameters to be estimated. By contrast, a nonparametric approach involves flexible modeling of the effect of the continuous predictors by assuming a large number of (infinitely many) parameters. Rather than choosing the functional form beforehand, we let the data itself inform the structure of the function. Let us step into how splines works in more detail.

Assume a function $f : [a, b] \to \mathbb{R}$ associating the relationship $\mathbb{E}[y] = f(x)$. This f is called a polynomial spline of degree $l \geq 0$ with knots $a = k_1 < k_2 < \cdots < k_m = b$, if

(i) $f(x)$ is $(l-1)$-times continuously differentiable. For $l = 1$, $f(x)$ is continuous but not differentiable. For $l = 0$, $f(x)$ is a step function, that is, no constraint of smoothness.

(ii) $f(x)$ is a polynomial of degree l on the intervals $[k_j, k_{j+1})$ defined by the knots.

This continuity condition guarantees that the function values of each polynomial segment match at the knots. In addition, the derivatives up to the $(l-1)$th order are continuous, enabling alignment at the knots.

Next, it is important to figure out the number of dimensions of the vector space of $(l-1)$-times continuously differentiable functions in order to represent the set of polynomial splines for a given degree l and knots configuration. The degrees of freedom can be computed as $m-1+l$ because of the constraints above: $(l+1)(m-1)$ free parameters over the given domain and $l(m-2)$ constraints for the up to lth derivatives, and hence $(l+1)(m-1)-l(m-2) = m-1+l$. Therefore, the smooth function can be expressed as follows:

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

with $d = m-1+l$, basis functions B_j and their coefficients γ_j for $j = 1, \ldots, d$. With this, we can now introduce basis functions we used for our analysis.

4.2.1.1 B-Splines

A B-spline basis function is formed by $(l+1)$ polynomial pieces of degree l and they are $(l-1)$-times continuously differentiable. More specifically, in a recursive manner,

$$B_j^l(x) = \frac{x - k_{j-l}}{k_j - k_{j-l}} B_{j-1}^{l-1}(x) + \frac{k_{j+1} - x}{k_{j+1} - k_{j+1-l}} B_j^{l-1}(x).$$

B-splines are renowned for their ability to closely approximate any smooth function with high accuracy. They are also recognized for computationally efficient, particularly for large datasets, since they rely on polynomial functions that are easy to evaluate.

4.2.1.2 Cyclic Cubic Regression Splines

Cubic splines are formed by connecting several cubic polynomial segments smoothly at knots. Cyclic cubic splines extend the concept of cubic splines by enforcing an additional condition: the function must also be smooth and continuous at the endpoints of the domain. This is particularly useful in cases like seasonal effects modeling in time series analysis, where the pattern repeats at a regular interval, or in any case where the end of the data range connects back to the beginning in a cyclic manner. In our case, for instance, it is appropriate that every start point and end point for the weekly seasonality remain equally.

4.2.2 Fitting GAMs

Taking all of the above into account, our first trial is defined as follows:

$$\eta = \beta_0 + \mathbb{1}^{holiday}\beta_1 + \mathbb{1}^{rain}\beta_2 + f_1(x^{temperature}) + f_2(x^{month\_year}) + f_3(x^{hour\_weekday}),$$

where f_1 and f_2 are constructed by b-spline basis functions and f_3 by cyclic cubic spline functions due to the nature of *hour_weekday*. Choosing the right number of basis dimensions for each smooth function is critical for optimal performance. Several methods, such as grid search or random search, can help optimize this hyperparameter. A relevant metric is also necessary to serve as a benchmark for optimization. The Akaike Information Criterion (AIC), for instance, is a useful metric for this purpose. Another way to evaluate a model's performance involves using the plot of effects of each component via `plot.gam` from `mgcv` package, checking if the estimated effect appears smooth or wiggly.

Using grid search and the AIC, we determined the most appropriate basis dimension, d. Starting with the default value of 10, we fit the model and examined the shape of the fitted effects for each variable. By increasing it by 5 and comparing the model's AICs, we identified the optimal configuration, as shown in Figure 4.2. As expected, the model manifested a good job of capturing the trend, particularly the two peaks during weekdays in *hour_weekday*, which represent the morning rush hour and the

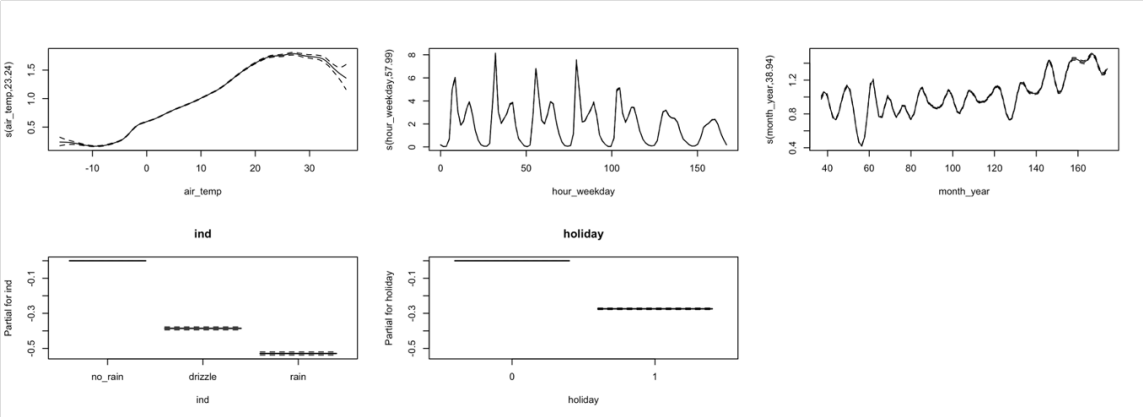
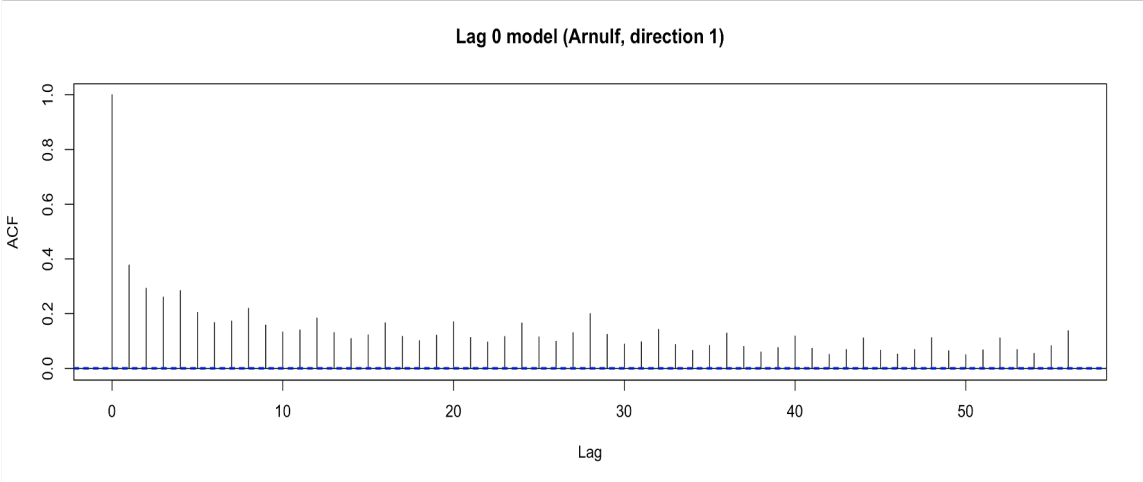


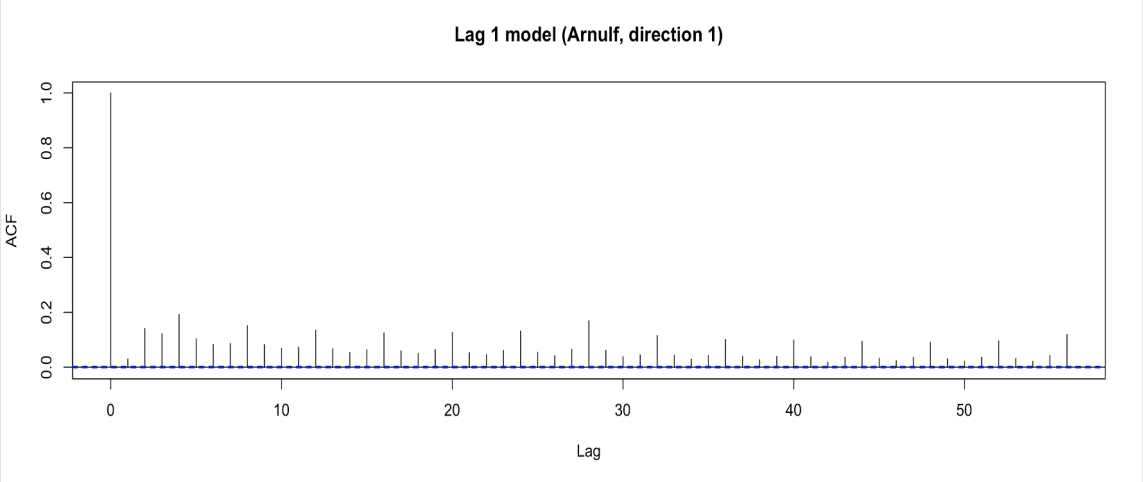
Figure 4.2: The estimated partial effects of the model for Margareten station in direction 2 without lag: $d = 30$ for *air temperature*, $d = 40$ for *month_year* and $d = 60$ for *hour_weekday*.

end of the workday. Regarding the effect of *rain*, we can see the highest number of cyclists when it is not raining, a decrease when there is a drizzle, and the lowest number during rain. People also tend to bike less on *holidays* compared to *non-holidays*, which could be due to fewer people commuting.

Nonetheless, this model also has another significant limitation when it comes to modeling time series data: it does not consider temporal dependency due to the independence assumption. This assumption is problematic given that the data points are spaced only 15 minutes apart, and it is generally understood that such closely spaced data points are likely to influence each other. Among various ways to diagnose this problem, the autocorrelation plot of residuals from the model can be used: it says that whether the residuals exhibit some form of dependency in terms of time. In Figure 4.3a, it is clear that there is significant residual autocorrelation in the data that the wiggly trend components could not account for. In this sense, we fit a new type of model to circumvent this issue; an autoregressive model (AR)[1].



(a)



(b)

Figure 4.3: For Arnulf station and for direction 1, (a) is residual autocorrelation plot of the vanilla model and (b) is of the AR(1) model. It is quite clear that the latter model has an effect of relaxing the higher residual autocorrelation present in the former one.

4.2.3 Autoregressive Model

The autoregressive model incorporates not only predictors but also lagged response variables in the regression function. In this case, it is indeed important to determine the number of lagged responses to include. Considering too many lags can lead to overfitting or may be nonsensical, depending on the nature of data. For these reasons, we fit a new AR(1) model, called "lag 1 model", to the data. That is,

$$y_t|y_{t-1}, \mathbf{x}_t \sim NB(\tilde{\mu}_t, \theta),$$

and therefore

$$\tilde{\mu}_t = \exp(\tilde{\eta}_t), \quad \text{with } \tilde{\eta}_t = \eta_t + f(y_{t-1}).$$

For the smooth function on the lagged response, we considered b-spline basis functions with $d = 20$. This model can mitigate the higher autocorrelation in residuals computed from the plain model as shown in Figure 4.3b.

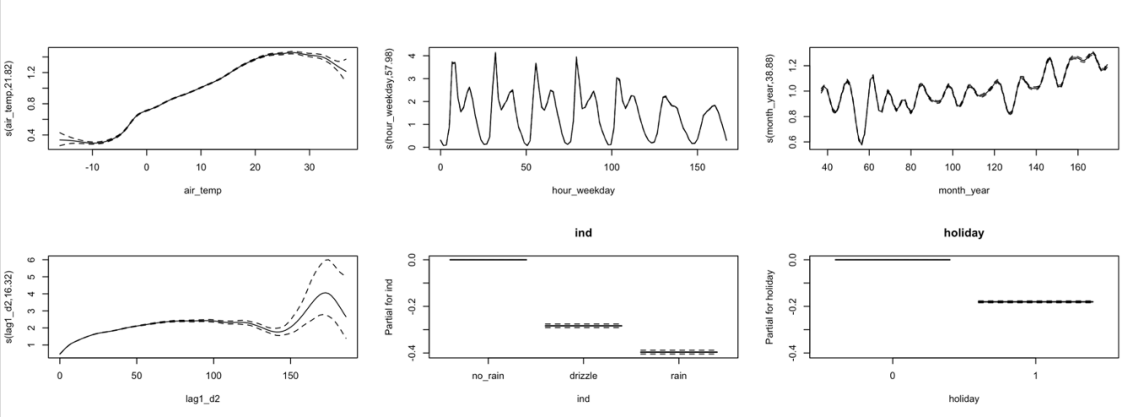


Figure 4.4: The estimated partial effects of the lag 1 model for Margareten station in direction 2.

However, a nuisance still remains for both models. In a classical linear regression model, the raw residual is expected to be symmetrically distributed around zero with constant variance. For count data, as we used a negative binomial family, the desirable properties are no longer accomplished. To deal with this pitfall, the

Pearson residual, weighted by the expected dispersion, or the deviance residual is suggested. It often does not completely lead to a clear answer for residual analysis in the non-normal case, though such residuals are theoretically designed to follow the fundamental properties. One of the methods for residual diagnostics in GLMs/GAMs is a simulation-based approach[3]; by computing simulated response variables and their empirical cumulative densities for each observation. The ideal result is that they look equally likely. Apart from this, there are a lot of discussions about how to sensibly treat the residual from generalized linear models, and hence it would require us to closely investigate the residuals.

Prediction Intervals

Our next work is to construct prediction intervals for each observation in order to detect candidates of outliers, which are considered culprits to make data implausible.

In statistical modeling, understanding the uncertainty in predictions is critical. This is often addressed through the use of intervals - specifically, confidence intervals and prediction intervals. While confidence intervals are used to estimate the uncertainty around model parameters or mean predictions, prediction intervals provide a range within which we expect future observations to fall. The distinction is crucial: confidence intervals focus on the uncertainty in the model's parameter estimates, reflecting the error in estimating the mean response given fixed values of the predictors. In contrast, prediction intervals take into account both the uncertainty in the model parameters and the inherent variability in the data itself. Thus, prediction intervals are generally wider than confidence intervals, because they include more sources of variability.

Data variability represents the randomness inherent in the data that cannot be explained by the model. This variability is important in constructing prediction intervals, because it directly affects the expected spread of observations around the predicted values.

Model variability refers to the uncertainty in the estimates of the model parameters. In generalized additive models (GAMs), this includes not only the linear coefficients, but also the smooth components used to capture non/linear relationships. This variability affects how well the model can predict new data points.

Constructing prediction intervals for GAMs with a negative binomial family presents

unique challenges compared to models that assume normal distributed errors. The negative binomial distribution is inherently asymmetric, especially with a mean-variance relationship that is not constant. This asymmetry means that standard approaches to constructing prediction intervals based on symmetric distributions (such as the normal distribution) are not directly applicable. As a result, the calculation of prediction intervals in this setting often requires numerical methods or simulation-based approaches, such as bootstrapping or Monte Carlo simulations, to account for both the skewed nature of the distribution and the complexity of the model.

With this understanding, our approach is based on simulation. As detailed in Algorithm 1, the process works in a simple manner. The most important thing is that it uses, by means of the Central Limit Theorem, a property of the limiting distribution of the estimators for the coefficients in the model normality. That means that we can obtain simulated coefficients and therefore the estimated means for each observation by multiplying the model matrix we already have. Then, we can finally generate samples of the response variable from the assumed distribution to calculate the lower bound and the upper bound of the prediction interval for each observation.

Algorithm 1 Simulation-based $100(1-\alpha)\%$ prediction interval

Given: model matrix Z including basis function evaluations, the number of simulations n_{sim}, $\hat{\theta}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\Sigma} = \widehat{Cov}(\hat{\boldsymbol{\beta}}) \sim$ model.

1. Draw $\boldsymbol{\beta_j} \sim N(\hat{\boldsymbol{\beta}}, \hat{\Sigma}), \quad j = 1, \ldots, n_{sim}$

2. Compute $\hat{\boldsymbol{\mu}}_{\boldsymbol{j}} = \exp(Z\boldsymbol{\beta_j}), \quad j = 1, \ldots, n_{sim}$

3. Draw $\boldsymbol{Y} \sim NB(\hat{M}, \hat{\theta}), \quad$ where $\hat{M} = [\hat{\boldsymbol{\mu}}_{\mathbf{1}}, \ldots, \hat{\boldsymbol{\mu}}_{n_{sim}}]$

4. Compute $\frac{\alpha}{2}$, $1 - \frac{\alpha}{2}$ quantiles of $\boldsymbol{Y}$ row-wise.

The table 5.1 shows a summary of the outliers detected by both models, using the simulated prediction intervals with 99.5% confidence in direction 1 for each station. It can be seen that both produce a benign coverage, and in particular, zero value outliers, whose original values are zero, are less identified in the lag 1 model. An example of the outliers can be checked in Figure 5.1.

	Lag 0			Lag 1		
Station	Data	Outliers (cov.)	Zero vals	Data	Outliers (cov.)	Zero vals
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

Table 5.1: A summary table including detected outliers in direction 1 by the simulated 99.5% prediction intervals.

5.1 Strong Outliers

However, it is too challenging for humans to manually check the individuals. To overcome this difficulty, we come up with a new idea, employing a classic statistical method. It is very simple and effective in that Algorithm 2 is just based on Tukey's rule: by flagging values that fall below Q_1(first quartile) minus 1.5 times the interquartile range or above Q_3(third quartile) plus 1.5 times the interquartile range. This allows us to identify which of the outliers are the strongest, so we can focus more on them.

Algorithm 2 Strong outliers

Given: true value of a detected outlier y_o and its corresponding fitted value $\hat{y}_o$.
1. Compute the distance between two points for each outlier,
i.e., $d_i \coloneqq |y_{o,i} - \hat{y}_{o,i}| \quad \forall i = 1, \ldots, n_{outliers}$.
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.

	station	direction_1	ds	air_temp	ind	holiday
1	Hirsch	8	2015-06-12 20:00:00	22.7	no_rain	0
2	Hirsch	3	2015-06-12 20:15:00	22.7	no_rain	0
3	Hirsch	3	2015-06-12 20:30:00	22.7	no_rain	0
4	Hirsch	2	2015-06-12 20:45:00	22.7	no_rain	0
5	Hirsch	15	2015-06-12 21:00:00	22.6	no_rain	0
6	Hirsch	7	2015-06-12 21:15:00	22.6	no_rain	0
7	Hirsch	2	2015-06-12 21:30:00	22.6	no_rain	0
8	Hirsch	1	2015-06-12 21:45:00	22.6	no_rain	0
9	Hirsch	6	2015-06-12 22:00:00	20.6	no_rain	0
10	Hirsch	2	2015-06-12 22:15:00	20.6	no_rain	0
11	Hirsch	3	2015-06-12 22:30:00	20.6	no_rain	0
12	Hirsch	7	2015-06-12 22:45:00	20.6	no_rain	0
13	Hirsch	4	2015-06-12 23:00:00	20.4	no_rain	0
14	Hirsch	4	2015-06-12 23:15:00	20.4	no_rain	0
15	Hirsch	7	2015-06-12 23:30:00	20.4	no_rain	0
16	Hirsch	8	2015-06-12 23:45:00	20.4	no_rain	0

Figure 5.1: An example of an outlier from the lag 1 model at Hirsch station.

As demonstrated in Table 5.2, we can see that the method considerably decreased the number of outliers that require attention, and that Arnulf and Olympia station exhibit many strong outliers in comparison to other stations. This observation suggests the need for enhanced monitoring of the measurement devices at these two stations.

Figure 5.2 illustrates two instances of strong outliers. First, the observation of 67 cyclists at approximately midnight at Hirsch station is notably high relative to other data points. Our model predicted this observation as 3. Additionally, the consecutive zero values recorded between 9:45 and 14:00 at Olympia station are similarly anomalous. These results suggest the possibility of an error in the device during the designated time period.

Station	**PIs**		**PIs + Tukey's rule**	
	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

Table 5.2: A comparison of outliers and zero values detected by the lag 1 model's PIs and PIs with Tukey's rule across different stations in direction 1.

	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

(a) Hirsch station

	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

(b) Arnulf station

Figure 5.2: Examples of strong outliers by the lag 1 model in direction 1. In particular, the red boxes in (b) have zero values in a row, which appears peculiar given the time period and other considerations.

Limitations

The first limitation of our model is the model itself. As there are numerous other models to choose from, there are advantages and disadvantages for every one of those. Exploring those may be feasible. Generalized additive models have a few well-known limitations.

Generalized additive models can be prone to overfitting, especially when the number of basis functions or degrees of freedom for the smoothing splines is not properly chosen. This can cause poor generalization. Other design choices for the feature variables together with more complex smoothing functions can enhance this issue. Discerning the needed feature variables from the available data and the number of basis dimensions is key to a good fit. This leads us to hyperparameter optimization of the smoothing splines. In our analysis we only did a rough grid search for the number of basis dimensions and compare them by the Akaike Information Criterion and graphical information by observing the partial effects of the estimated model. A more detailed optimization for each station and direction would result in a better performance.

While GAMs relax the assumption of linearity in the predictors, they still assume additivity of the smooth functions. In real world scenarios, the relationships between predictors and the response variable may not always be purely additive. There could be interactions or non-linear relationships between predictors that violate the additivity assumption. If this assumption is violated, the model may not accurately capture the underlying relationships, leading to biased estimates, poor performance and bad interpretability. Using time series models or Machine Learning models may alleviate this problem, but may come with other problems.

Furthermore, as mentioned in Model, there is a critical point that needs to be carefully examined. Each of the approaches that we have tried only make sense, if the assumptions that underlie the model are met. In contrast to a standard case of regression, where a response variable is treated as being normally distributed, it is the case that the (raw) residuals computed from our model are not symmetric and are not centered at zero due to the negative binomial family.

This can be observed via the plot of (deviance) residuals of the model against its linear predictor. As illustrated in Figure 6.1, it is clearly identified that there is an undesirable pattern in the residuals plot. The presence of patterns in the plot can indicate that the model is not adequately capturing all the relevant variability in the data. This could be due to several factors, such as omitted variables that are influential in predicting the number of cyclists, or incorrect functional forms used to model the relationships between predictors and the response variable. The observed patterns in the residuals could also be a symptom of unmodeled heteroscedasticity - the condition where the variance of the residuals is not constant across levels of the linear predictor.

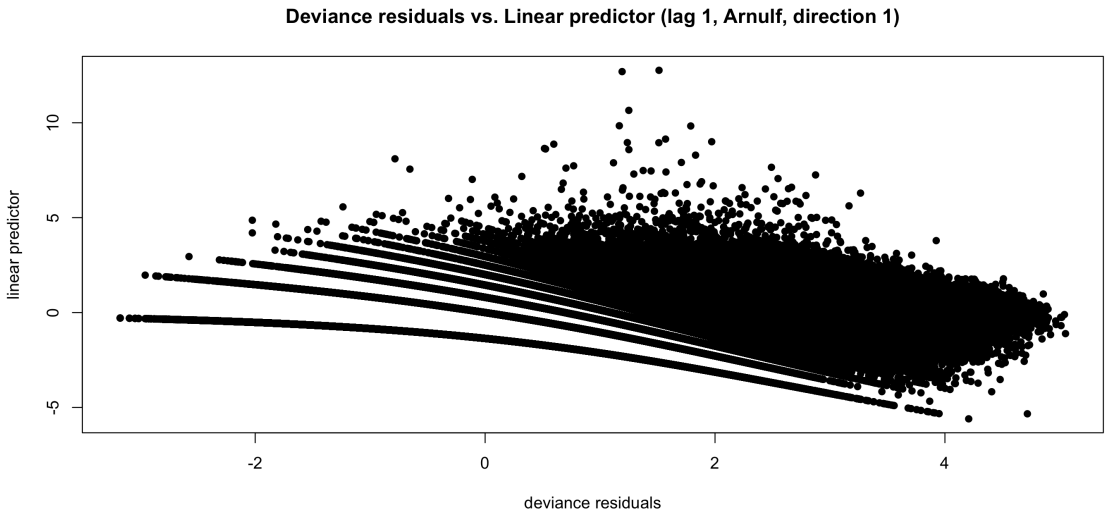


Figure 6.1: A plot of deviance residuals of the lag 1 model for Arnulf station in direction 1 vs. its linear predictor.

It may be necessary to reconsider the model specification to address these limitations. This might include testing for and incorporating interaction effects, exploring the need for considering alternative distributions that might better capture the nature of the data. Advanced diagnostic tools or alternative modeling approaches, such as mixed-effects models or other types of generalized linear models, may also provide insight into improving model performance and better handling of data characteristics. These steps will help ensure that the model not only fits the data better, but also produces more reliable and generalizable results.

Conclusion

To check the Plausibility for Cycling Traffic Count Data we employed a generalized additive model with a negative binomial family and used splines to enable the GAM to capture non-linear patterns. We then cleverly constructed a prediction interval by simulation that captures both model variability and data variability. To ease the load of our collaborators, we calculated a range distribution from all outliers to mark strong outliers, achieved by Tukey's rule. With this approach we were able to find implausible data points. It does need to be double-checked by a human as patterns can not be recognized by the model.

As we have already discussed in Limitations, generalized additive models have their advantages and disadvantages compared to other models and are by no means the definitive best approach to this task, but it is a valid approach.

List of Figures

Bibliography

[1] A. Cameron and Pravin Trivedi. "Regression analysis of count data. 2nd ed". In: vol. 41. Sept. 1999. DOI: 10.1017/CBO9780511814365.

[2] Ludwig Fahrmeir et al. *Regression: Models, Methods and Applications*. Jan. 2013. ISBN: 978-3-642-34332-2. DOI: 10.1007/978-3-642-34333-9.

[3] Florian Hartig. *DHARMa: residual diagnostics for hierarchical (multi-level mixed) regression models*. https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMa.html. [Online; accessed 02/05/2024].

[4] Monika Wreszinski Ulrike Teubner. *„Aufs Radl – Fertig? – Los!" Ergebnisse der Raddauerzählstellen in München 2017 und 2018*. https://stadt.muenchen.de/dam/jcr:9a65625e-952f-470f-b6a1-d4270f4526cb/mb190304.pdf. [Online; accessed 05/05/2024].

[5] Deutscher Wetterdienst. *Wetter- und Klimalexikon*. https://www.dwd.de/DE/service/lexikon/Functions/glossar.html?lv3=101906&lv2=101812. [Online; accessed 07/05/2024].

[6] Simon Wood. *Generalized Additive Models: An Introduction With R*. Vol. 66. Jan. 2006, p. 391. ISBN: 9781315370279. DOI: 10.1201/9781315370279.