Predicting Seoul bike sharing demand with GAMs.

Klaudia Weigel

Dataset

The dataset contains information about shared bikes in Seoul, Korea. The variables in the dataset are:

- `Date` - day indicator. Data has been collected from 2017-12-01 to 2018-11-30.
- `RentedBikeCount` - number of rented bikes, response variable,
- `Hour` - hour of the day,
- `Temperature`, `Humidity`, `WindSpeed`, `Visibility`, `DewPointTemp`, `SolarRadiation`, `Rainfall`, `Snowfall` - variables associated with weather conditions,
- `Seasons` - categorical variable indicating season (winter, spring, summer, autumn)
- `Holiday` - categorical variable indicating whether a particular day is a holiday,
- `FunctioningDay` - functional days of the rental bike system.

```
##         Date RentedBikeCount Hour Temp Humidity WindSpeed Visibility
## 1 01/12/2017             254    0 -5.2       37       2.2       2000
## 2 01/12/2017             204    1 -5.5       38       0.8       2000
## 3 01/12/2017             173    2 -6.0       39       1.0       2000
##   DewPointTemp SolarRadiation Rainfall Snowfall Season    Holiday
## 1        -17.6              0        0        0 Winter No Holiday
## 2        -17.6              0        0        0 Winter No Holiday
## 3        -17.7              0        0        0 Winter No Holiday
##   FunctioningDay
## 1            Yes
## 2            Yes
## 3            Yes
```

Amount of rented bikes by Hour

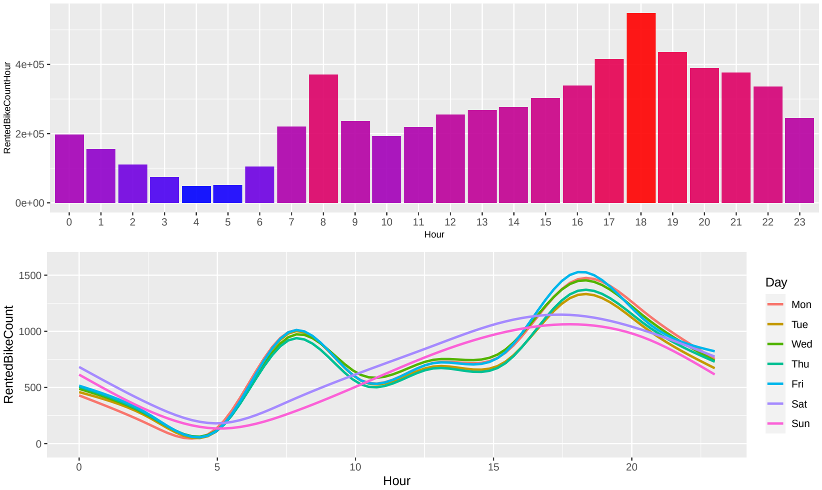


Figure 1: Number of rented bikes with respect to the hour of the rental and factored by day.

Categorical variables

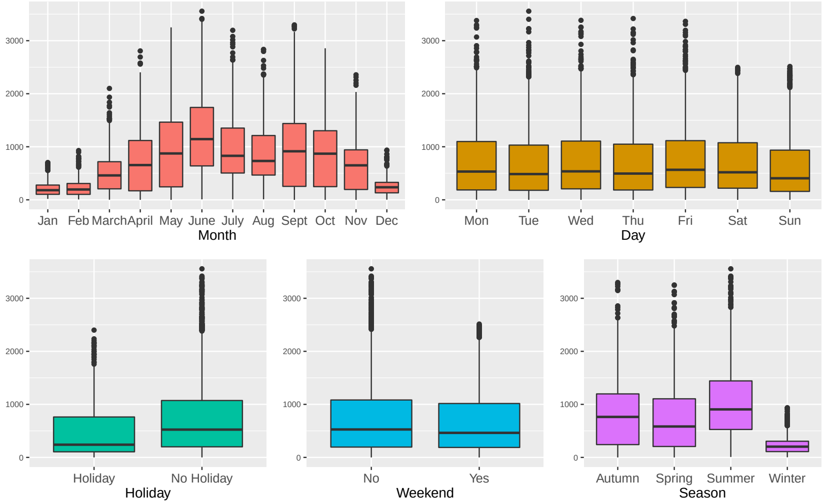
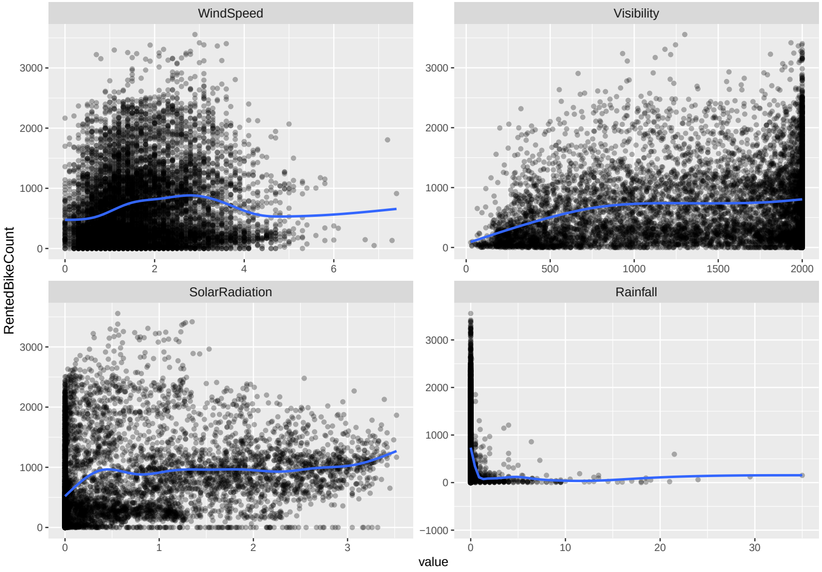
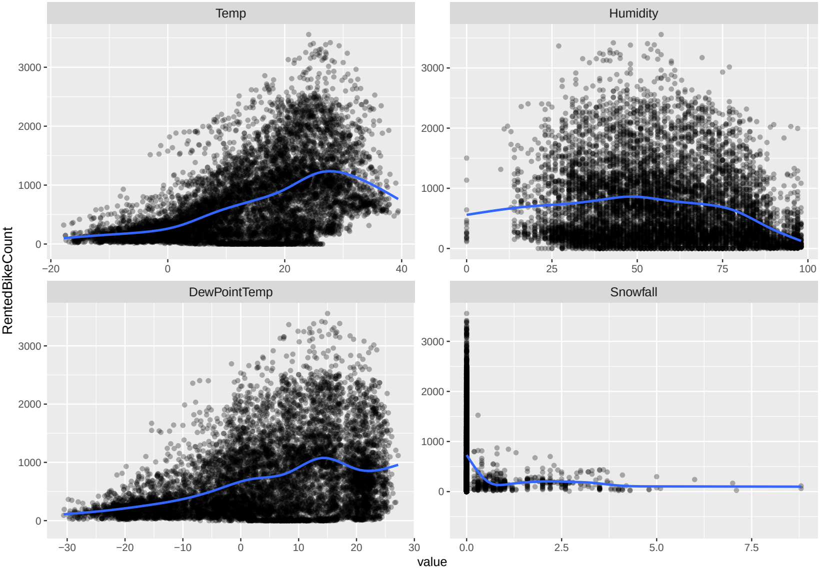


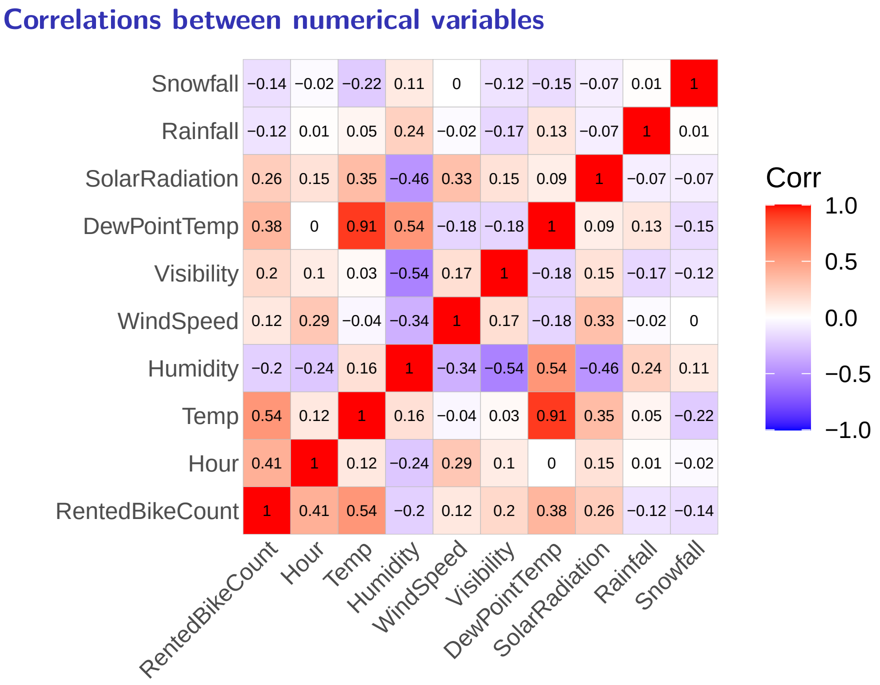
Figure 2: Boxplots of the number of rented bikes with respect to Month, Day, Holiday, Weekend and Season.

Exploring numerical variables



Exploring numerical variables




Correlations between numerical variables
Snowfall −0.14 −0.02 −0.22 0.11 0 −0.12 −0.15 −0.07 0.01 1
Rainfall −0.12 0.01 0.05 0.24 −0.02 −0.17 0.13 −0.07 1 0.01
SolarRadiation 0.26 0.15 0.35 −0.46 0.33 0.15 0.09 1 −0.07 −0.07
DewPointTemp 0.38 0 0.91 0.54 −0.18 −0.18 1 0.09 0.13 −0.15
Visibility 0.2 0.1 0.03 −0.54 0.17 1 −0.18 0.15 −0.17 −0.12
WindSpeed 0.12 0.29 −0.04 −0.34 1 0.17 −0.18 0.33 −0.02 0
Humidity −0.2 −0.24 0.16 1 −0.34 −0.54 0.54 −0.46 0.24 0.11
Temp 0.54 0.12 1 0.16 −0.04 0.03 0.91 0.35 0.05 −0.22
Hour 0.41 1 0.12 −0.24 0.29 0.1 0 0.15 0.01 −0.02
RentedBikeCount 1 0.41 0.54 −0.2 0.12 0.2 0.38 0.26 −0.12 −0.14
RentedBikeCount Hour Temp Humidity WindSpeed Visibility DewPointTemp SolarRadiation Rainfall Snowfall
Corr
1.0
0.5
0.0
−0.5
−1.0

Dealing with overdispersion: quasi-Poisson model

After checking the mean and variance of the response variable

```
mean(seoul_bikes$RentedBikeCount)
```

```
## [1] 704.6021
```

```
var(seoul_bikes$RentedBikeCount)
```

```
## [1] 416021.7
```

For the quasi-Poisson model assumes that the response variable has mean μ and variance $\theta\mu$, where θ is a dispersion parameter. The quasi-Poisson uses the log link function to model the mean

$$\log(\mu_i) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i}.$$

Modelling the data

- Split the dataset into training and validation sets, with 80% split ratio.

	Train	Test
Size	6772	1693

- Features kept in the model: Hour, Temp, Humidity, WindSpeed, Visibility, SolarRadiation, Rainfall, Snowfall, Holiday, Weekend, Month, a total of 11 variables. Also removed observations, where `FunctioningDay == No` (295 obs.).
- Use generalized additive models to model the data

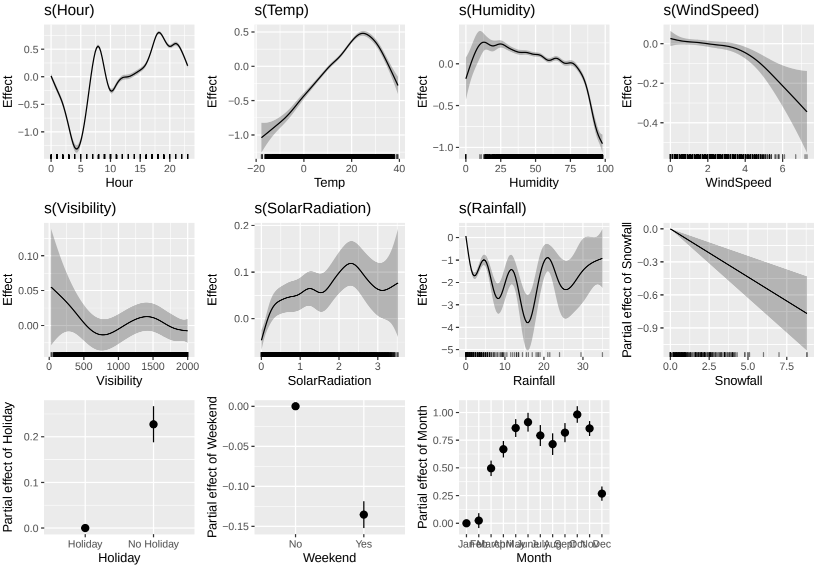
$$g(\mu_i) = \beta_0 + \sum_j f_i(x_{ij}) + \sum_{k \neq j} f_{kj}(x_{ik}, x_{ij}).$$

- Compare models with F-test, RMSE, MAE

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}, \quad MAE = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n}.$$

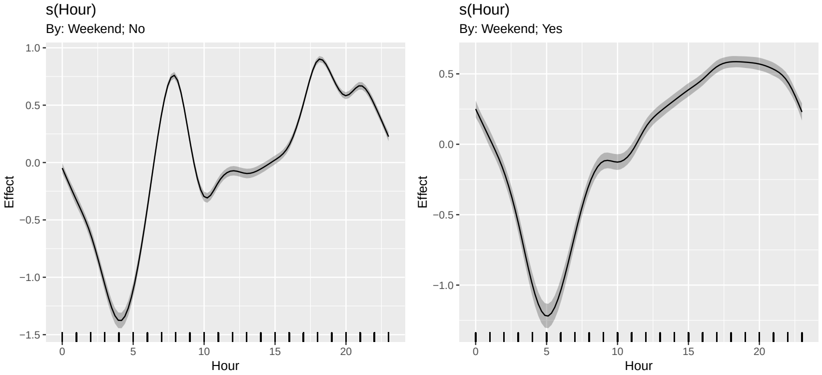
GAM Model

First we will consider a generalized additive model with all numerical variables, except for Snowfall as smooth terms.



Hour-Weekend interaction

First step in improving a basic model is to add an interaction between Hour and Weekend variable. We are fitting separate smooths for each level of Weekend `s(Hour, by = Weekend)`.



Comparing the two models with the F-test, we get p-value equal to

```
anova(gam1, gam2, test = "F")$"Pr(>F)"[2]
```

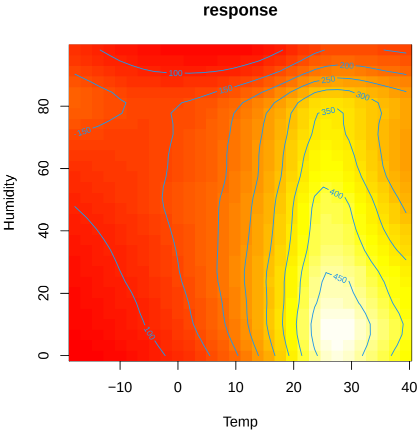
```
## [1] 0
```

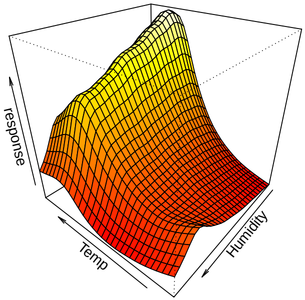
Smooth terms interactions

To introduce interactions we use `ti`, which produces a tensor product interaction, appropriate when the main effects (and any lower interactions) are also present.

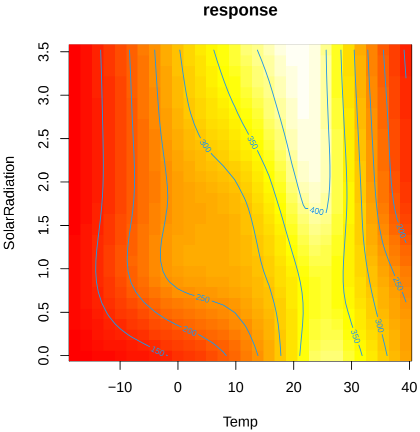
- Temp and Humidity
- Temp and SolarRadiation
- Temp and WindSpeed

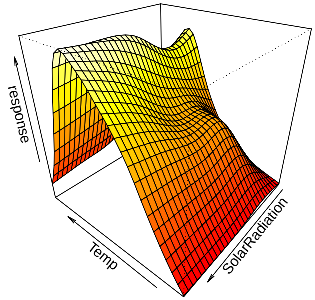
Temp, Humidity



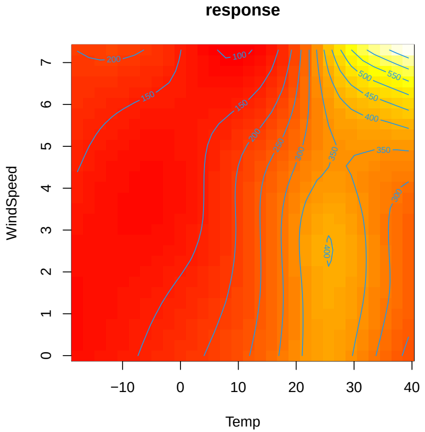


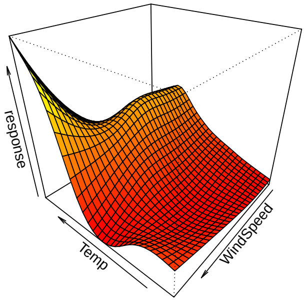
Temp, SolarRadiation





Temp, WindSpeed





Model comparison

- GLM - basic generalized linear model with all predictor variables,

$$
\begin{aligned}
\log(\mu_i) = {} & \beta_0 + \beta_1 * Hour_i + \beta_2 * Temp_i + \beta_3 * Humidity_i + \cdots + \beta_7 * Rainfall_i + \beta_8 * Snowfall_i \\
& + \beta_9 * \mathbb{I}(Holiday_i == "NoHoliday") + \beta_{10} * \mathbb{I}(Weekend_i == "Yes") \\
& + \beta_{11}\ \mathbb{I}(Month == "Feb") + \cdots + \beta_{21}\ \mathbb{I}(Month == "Dec")
\end{aligned}
$$

- GAM1 - generalized additive model with all continuous variables, except Snowfall defined as smooth functions,

$$
\begin{aligned}
\log(\mu_i) = {} & \beta_0 + \beta_1 * Snowfall_1 + \beta_3 * \mathbb{I}(Holiday_i == "NoHoliday") + \beta_4 * \mathbb{I}(Weekend_i == "Yes") \\
& + \beta_4\ \mathbb{I}(Month == "Feb") + \cdots + \beta_{14}\ \mathbb{I}(Month == "Dec") + f_1(Hour_i) + f_2(Temp_i) \\
& + f_3(Humidity_i) + f_4(WindSpeed_i) + f_5(Visibility) + f_6(SolarRadiation) + f_7(Rainfall_i)
\end{aligned}
$$

- GAM2 - GAM1 model with added interaction between Hour and Weekend,
- GAM3 - GAM2 model with tensor interactions.

	RMSE		MAE			
	Train	Test	Train	Test	R-sq.(adj)	Deviance expl.
GLM	369.51	376.43	253.95	259.80	0.66	0.68
GAM1	235.57	236.14	155.24	160.67	0.86	0.87
GAM2	198.01	198.95	125.12	128.59	0.90	0.91
GAM3	187.80	189.50	114.34	119.05	0.91	0.92

The end