Predicting Seoul bike sharing demand with generalized additive models - Semiparametric Regression project write-up.

Klaudia Weigel

Abstract

The objective of this project is to use generalized linear models to predict bike sharing demand in a public bike rental system in Seoul, Korea. First we will introduce generalized linear models and generalized additive models as a more flexible alternative to generalized linear models for modelling data with with non-gaussian response variable. Then we will apply these models to the Seoul bike sharing data. We will compare different models, utilizing the root mean squared error and the mean absolute error metrics.

1 Introduction

Bike sharing systems are bike rental services where the entire process of service registration, bike rental and return has become automated. Through these systems, users are able to easily rent a bike from a particular docking station and return it to another station. First such system was introduced in Amsterdam in 1965 and in recent years various bike sharing systems have been gaining popularity in many countries. Such systems are especially promoted in cities to reduce traffic congestion, reduce exhaust emissions and improve health of the residents. A shared bike also eliminates some of the disadvantages of a private bike like exposure to theft and high purchase costs. [Wikipedia, 2022]. According to bikesharingworldmap.com by August 2021, there were more than 10 million bikes shared in different kinds of schemes. Because of such rapid expansion there is a need for bike sharing programs to effectively understand the factors that influence demand so that they can better maintain inventory, schedule repairs, and manage resources.

In Seoul a bike sharing system called Ddareungi or Ttareungyi (Seoul Bike in English) was first introduced in October 2015 in selected areas by the Han River. Initially with 150 docking stations and 2,000 bicycles across five districts.

[Harezlak et al., 2018]

2 Generalized Linear Models

Generalized linear models were first introduced as an extension of the linear models, and unlike linear models GLMs can accommodate various types of response variables like binary or count variables.

Let us assume that we have a set of observations $\{(y_i, x_{i1}, \dots, x_{ip})\}_{i=1}^n$, where y_i is the response variable and $(x_{i1}, \dots, x_{ip})$ is a set od predictors. In generalized linear models we assume that

1. The response variables $y_1, \dots, y_n$ are instances of independent random variables $Y_1, \dots, Y_n$ from the same distribution that belongs to the family of exponential distributions:

$$f(y_i|\theta_i, \phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right]$$

   The θ is the canonical parameter and represents the location, while ϕ the dispersion parameter and represents the scale.

2. For each i the relation between the mean $\mu_i = E(Y_i)$ and the predictors $(x_{i1}, \dots, x_{ip})$ is assumed to be:

$$\eta_i = g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip},$$

where $(\beta_0, \beta_1, \ldots, \beta_p) \in \mathbb{R}^{p+1}$ is an unknown vector of regression parameters, and is the link function. It can be shown that if Y_i has a distribution in the exponential family then it has mean and variance

$$E(Y_i) = \mu_i = b'(\theta_i),$$

$$\text{var}(Y_i) = \sigma_i^2 = b''(\theta_i)a_i(\phi).$$

The log-likelihood function of β and ϕ is

$$\ell(\beta, \phi) = \sum_{i=1}^{n} \log f(y_i) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi).$$

A closed form solution of log-likelihood maximization can only be found in the case of linear regression. In other cases an algorithm called iteratively re-weighted least squares is most often applied.

An important value in GLMs is the *deviance* of the model, which is defined as

$$D = 2\phi(\ell(\hat{\beta^s}) - \ell(\beta^r)),$$

where $\ell(\hat{\beta^s})$ is the log-likelihood of the saturated model, in which we have one parameter per datapoint. Whereas $\ell(\beta^r)$ is the log-likelihood of the reduced model.

3 Generalized additive models

A generalized additive model (GAM) is an extension of generalized linear model where the transformed mean is defined as a linear combination of smooth functions of covariates. Its main advantage is the flexibility in the specification of the relationship between a dependent variable and its corresponding covariates, contrary to the classical way to model that relationship based on linear associations, which is not always a good assumption in many applications. A very detailed overview of GAMs can be found for example in [Wood, 2017], here we will provide general ideas, based on that source.

A generalized linear has the form

$$g(\mu_i) = A_i\gamma + \sum_j f_j(x_j i), \quad Y_i \sim EF(\mu_i, \phi), \tag{1}$$

where A_i is the ith row of a parametric model matrix, with a vector of parameters γ and f_j is a smooth function of covariate x_j. The response variables are assumed to be instances of random variables from the exponential family with mean μ_i and ϕ is the scale/dispersion parameter.

To estimate each function f_j, using the same methods as for linear models, f_j must be represented in such a way that (1) becomes a linear model. This can be done by choosing a basis, defining the space of functions of which f_j is an element. Choosing a basis, amounts to choosing some basis functions, which will be treated as completely known: if $b_i(x)$ is the ith such basis function, then f is assumed to have a representation

$$f_j(x_j) = \sum_{i=1}^{q_j} b_{ji}(x_j)\beta_{ji},$$

where $\beta_{j1}, \ldots \beta_{jq_j})$ are unknown regression coefficients, different for every f_j. We may define every smooth term with different basis functions and a different number of them. We can therefore write our model as

$$g(\mu) = X\beta,$$

where X is the design matrix and β is the vector of unknown regression coefficients.

3.1 Penalized regression

If the functions $f_1, \dots, f_d$ were chosen to be arbitrarily complex we would most likely have to deal with severe overfitting to the training data. To prevent that, we penalize excessive wobbliness of f_j. We therefore want to maximize the penalized log-likelihood of the regression parameters.

$$\ell_P(\beta) = \ell(\beta) - \frac{1}{2\phi} \sum_{j=1}^{d} \lambda_j \times [\text{roughness measure of } f_j],$$

where λ_j is a smoothing parameter. The larger λ_j is the more we penalize the regression coefficients, so that the function f_j is not too wobbly. Generally the roughness measure may vary depending on the choice of the basis functions. It is commonly chosen to be $\int f_j''(x)dx$, which can be expressed as $\beta^T S_j \beta$, where S_j is a known matrix, known as penalty matrix.

In practice the penalized log-likelihood maximization problem is solved by penalized iteratively re-weighted least squares (P-IRLS), while the smoothing parameters can be estimated using cross validation or related criteria.

The P-IRLS algorithm estimates β, given λ using following steps

1. Initialize $\hat{\mu}_i = y_i + \delta_i$, $\hat{\eta}_i = g(\hat{\mu}_i)$, where δ_i is equal to zero or a small constant, ensuring that $\hat{\eta}_i$ is finite.
2. Compute $z_i = g'(\hat{\mu}_i)(y_i - \hat{\mu}_i)/\alpha(\hat{\mu}_i) + \hat{\eta}_i$ and weights $w_i = \alpha(\hat{\mu}_i)/\{g'(\hat{\mu}_i)^2 V(\hat{\mu}_i)\}$, where $\alpha(\mu_i) = 1 + (y_i - \mu_i)\{V'(\mu_i) + g''(\mu_i)/g'(\mu_i)\}$ and $V(\mu)$ is the variance function determined by the exponential family or by quasi-likelihood.
3. Find $\hat{\beta}$ that minimizes the weighted least squares objective

   $$\|z - X\beta\|_W^2 + \sum_j \lambda_j \beta^T S_j \beta, \quad \text{where} \quad \|a\|_W^2 = a^T W a,$$

   then update $\hat{\eta} = X\hat{\beta}$ and $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$.

We repeat the steps 1 and 2 until convergence.

The vector λ is usually not known and has to be estimated. The estimation is usually done via *generalized cross validation*(GCV) score, defined as

$$\mathcal{V}(\lambda) = \frac{nD(\hat{\beta})}{n - tr(A)}, \quad A = X(X^T X + S_\lambda)^{-1} X^T.$$

The *effective degrees of freedom* for a given smooth term f_j is the sum of diagonal entries of $F = (X^T W X^T + \sum_j \lambda_j S_j)^{-1} X^T W X$ corresponding to the parameters of that f_j.

It can be shown that for large samples [Wood, 2006]

$$\hat{\beta} \sim N(E(\hat{\beta}), V_e), \quad V_e = (X^T W X + S_\lambda)^{-1} X^T W X (X^T W X + S_\lambda)^{-1} \phi.$$

However generally $E(\hat{\beta}) \neq \beta$, so we cannot use this fact for confidence interval prediction.

An alternative is to use a Bayesian approach, which results in a Bayesian posterior covariance matrix for the parameters,

$$V_\beta = (X^T W X + S_\lambda)^{-1} \phi.$$

and a corresponding posterior distribution for those parameters,

$$\beta \sim N(\hat{\beta}, V_\beta).$$

Now we can construct a Wald test for testing significance of parametric terms and smooth terms, details are given in [Wood, 2017].

4 Quasi-Poisson regression

It is often the case that count data are overdispersed, meaning that the variance of the response variable is significantly larger than the mean. In such cases assuming Poisson distribution for the response leads to an inaccurate model. There are however other methods we can apply to model such data like quasi-Poisson regression or negative binomial regression [Ver Hoef and Boveng, 2007].

4.1 Quasi-likelihood

The description quasi-likelihood follows from [Harezlak et al., 2018]. Quasi-likelihood assumes that the variance of the response variable y_i is equal to $\phi V(\mu_i)$. The quasi-likelihood estimator of the regression parameters maximizes $Q = \sum_{i=1}^{n} Q_i$ where

$$Q_i = \int_{y_i}^{\mu_i} \frac{y_i - u}{\phi V(u)} du.$$

We estimate the dispersion parameter from

$$\hat{\phi} = \frac{1}{n - d - 1} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

In case of quasi-Poisson regression we have $\text{var}(y_i) = \phi\mu_i$.

5 Data exploration

Data exploration is an important step of fitting any machine learning model. In order to find significant features, identify relationships between different sets of predictors or eliminate potential collinearity among others we must carefully examine the dataset before actually fitting.

The data was originally introduced in [E and Cho, 2020]. It was collected for a year from 1st December 2017 to 30th November 2018 and provides information about hourly counts of the number of rented bikes. Additional features include weather conditions: temperature (in Celsius degrees), humidity (%), wind speed (m/s), visibility (10m), dew point temperature (in Celsius degrees), solar radiation (MJ/m^2), rainfall (mm), snowfall (cm). There are also variables indicating the season, whether the day was a public holiday and if the service was functional at the time. The dataset contains no missing values.

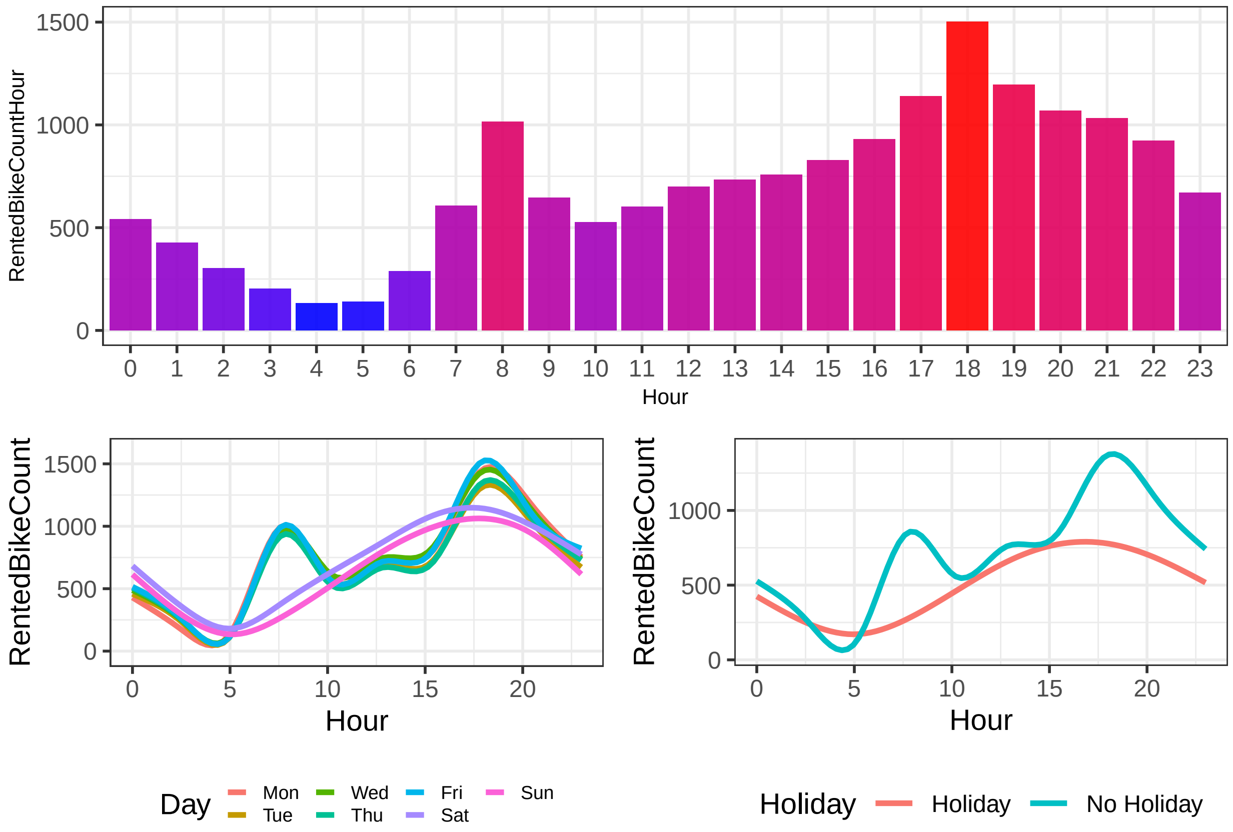

Figure 1: Uppermost plot shows an average amout of bikes being rented out during each hour. Plots in the bottom part show how the distribution varies when grouped by day and by holiday.

From the above plots we see that generally the most bikes are being rented out in the evening. There are also two distinctive peaks at 8 am and 6 pm, which might indicate that a lot of people use bikes to commute to work/school.

The bottom plots show that the distribution of the number of rented bikes with respect to hour is different for when a day is a weekend or a holiday. During those days most bikes are being rented out in the evening hours and the distribution doesn't have as much variation. These are important interactions that should be accounted for in the model.

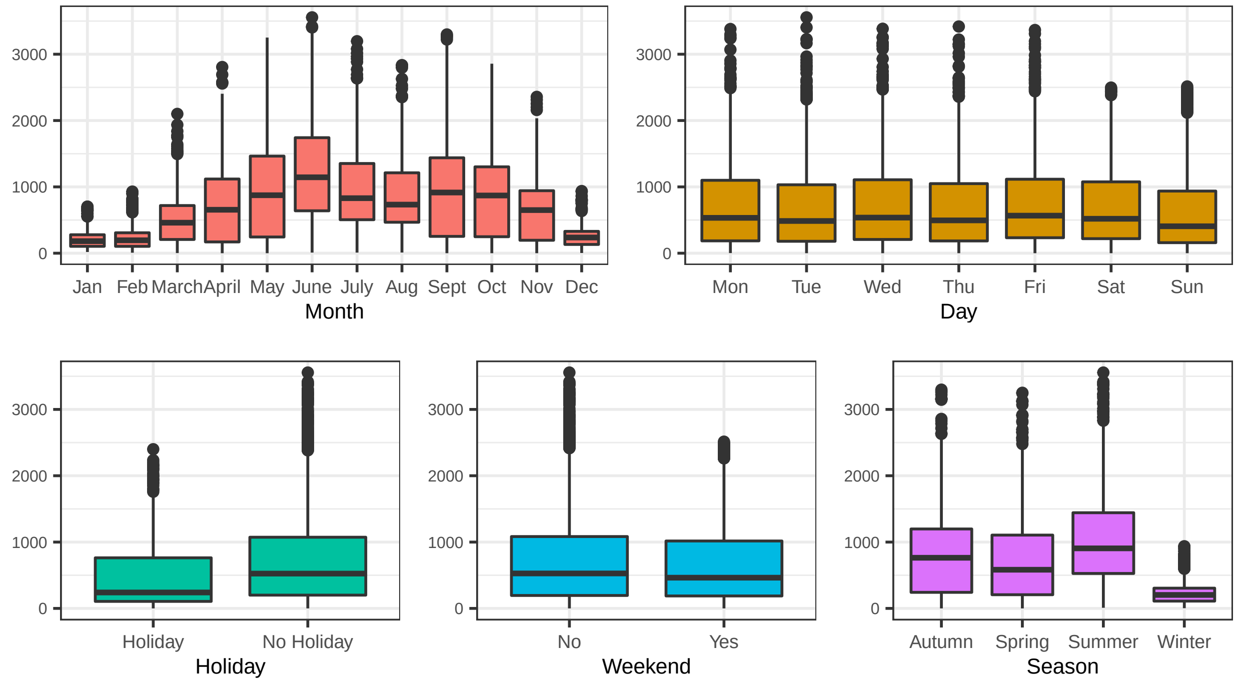

Figure 2: Boxplots of the number of rented bikes with respect to Month, Day, Holiday, Weekend and Season.

From the above boxplots we can conclude that the least number of bikes is being rented out in the winter months, whereas most people ride bikes in late spring, with peak in June and early fall. When it comes to the Day variable we don't see much difference in the amount of rentals between individual weekdays, the distinction however is visible when comparing a weekday to a weekend. Generally there are more rentals during no holiday days and weekdays.

We will now analyze the numerical variables. To analyze the numerical variables we begin with a correlation table.

Table 1: Correlations between numerical variables.

Variable	1	2	3	4	5	6	7	8	9
1. RentedBikeCount	—	—	—	—	—	—	—	—	—
2. Hour	.41	—	—	—	—	—	—	—	—
3. Temp	.54	.12	—	—	—	—	—	—	—
4. Humidity	-.20	-.24	.16	—	—	—	—	—	—
5. WindSpeed	.12	.29	-.04	-.34	—	—	—	—	—
6. Visibility	.20	.10	.03	-.54	.17	—	—	—	—
7. DewPointTemp	.38	.00	.91	.54	-.18	-.18	—	—	—
8. SolarRadiation	.26	.15	.35	-.46	.33	.15	.09	—	—
9. Rainfall	-.12	.01	.05	.24	-.02	-.17	.13	-.07	—
10. Snowfall	-.14	-.02	-.22	.11	-.00	-.12	-.15	-.07	.01

From the correlation matrix we see that the Temp variable and the DewPointTemp variable are very strongly correlated. Since temperature is more strongly correlated with the response variable, DewPointTemp will be removed. When it comes to the response variable, we can see that the amount of rented bikes is most strongly correlated with Hour and Temp.

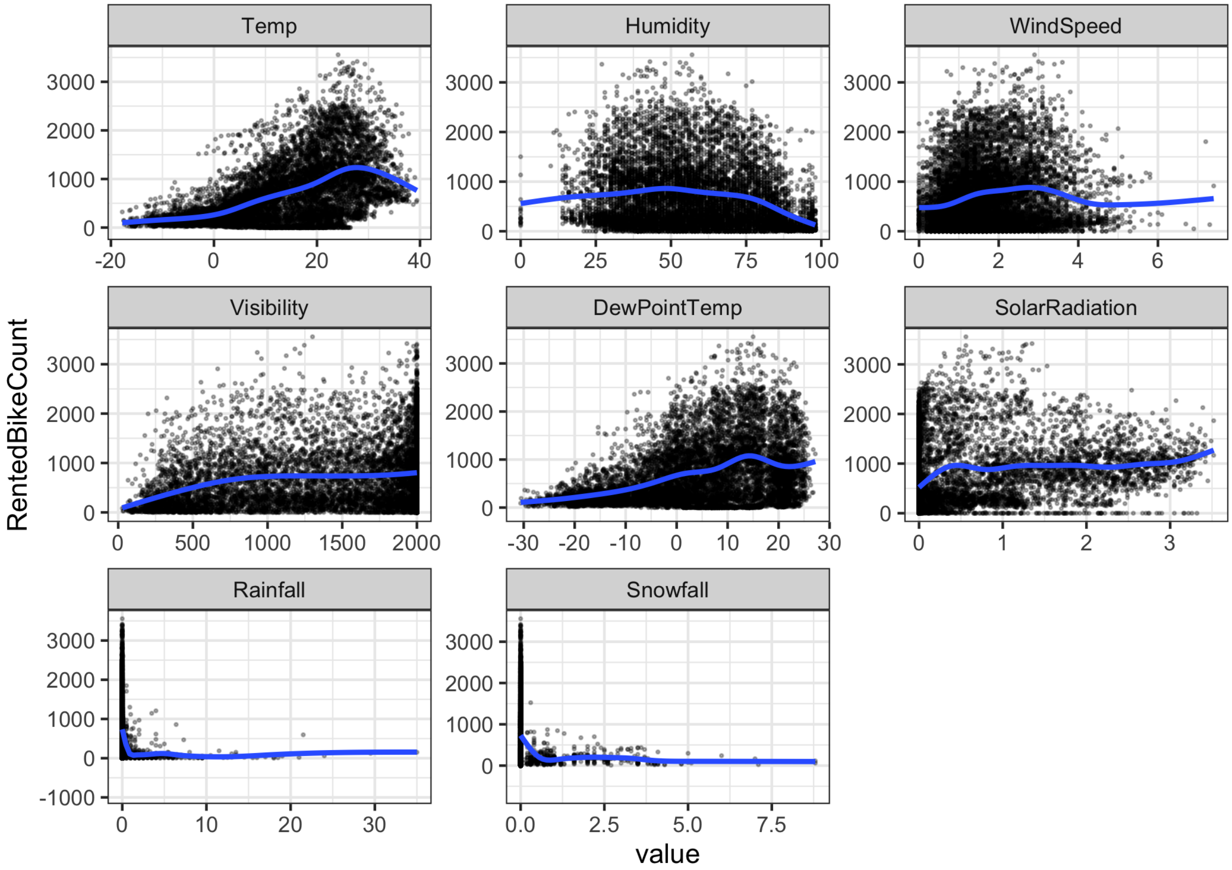

Figure 3: Univariate plots of the numerical variables.

We see that most rentals generally occur within temperatures of around 20-30 degrees. In case of Humidity there is a decreasing trend of the number of rented bikes. For some of the observations the humidity index is equal to zero, which is not physically possible, these observations are a minority however and shouldn't affect the fit. The plots for Rainfall and Snowfall are similar and suggest that a significant number of counts lies along the dates when there was no rain or snow. The count increases with higher visibility and with higher solar radiation. Overall all the plots suggest a non-linear relationship between predictors and the response variable which suggests modeling the data with a generalized linear model.

6 Modeling

Following approach was taken to model the data

1. The observations when the rental service was not functional were removed from the dataset (295 observations).
2. A subset of features was chosen: Hour, Temp, Humidity, WindSpeed, Visibility, SolarRadiation, Rainfall, Snowfall, Holiday, Weekend, Month. Dew point temperature was removed due to high correlation with temperature. Features Month and Weekend were extracted from the Date variable. Season was replaced by more informative Month. A new categorical feature was created as a joint interaction term of Weekend and Holiday (factor with 4 levels). So a total of 12.
3. Modified dataset was split into training and validation sets, with 80% split ratio.

Table 2: Train and test set sizes.

	Train	Test
Size	6772	1693

4. Different generalized additive models were used to model the data. The distribution was specified as quasi-Poisson, since the response variable is overdispersed with mean equal 704.6021 and variance equal to 416021.7.

5. For comparing different models we will evaluate root mean squared error and mean absolute error for the training set and testing set.

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

6.1 Model comparison

All the models were trained in R with *gam* function from the mgcv library. In model definition we define smooth terms with `s()`. We can also specify the number of basis functions `k`, for example we can define `s(Hour, k = 24)`. If not otherwise specified *gam* uses thin plate regression splines. For bivariate smoothing me may use `s(), te(), ti()`. We will compare the following models:

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

$$\begin{aligned}\log(\mu_i) = \beta_0 &+ \beta_1 * Hour_i + \beta_2 * Temp_i + \beta_3 * Humidity_i + \beta_4 * WindSpeed_i \\ &+ \beta_5 * Visibility_i + \beta_6 * SolarRadiation + \beta_7 * Rainfall_i + \beta_8 * Snowfall_i \\ &+ \beta_9 * \mathbb{I}(Holiday_i == "NoHoliday") + \beta_{10} * \mathbb{I}(Weekend_i == "Yes") \\ &+ \beta_{11}\ \mathbb{I}(Month_i == "Feb") + \cdots + \beta_{21}\ \mathbb{I}(Month_i == "Dec")\end{aligned}$$

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

$$\begin{aligned}\log(\mu_i) = \beta_0 &+ \beta_1 * Snowfall_i + \beta_2 * \mathbb{I}(Holiday_i == "NoHoliday") + \beta_3 * \mathbb{I}(Weekend_i == "Yes") \\ &+ \beta_4\ \mathbb{I}(Month_i == "Feb") + \cdots + \beta_{14}\ \mathbb{I}(Month_i == "Dec") + f_1(Hour_i) + f_2(Temp_i) \\ &+ f_3(Humidity_i) + f_4(WindSpeed_i) + f_5(Visibility_i) + f_6(SolarRadiation_i) + f_7(Rainfall_i)\end{aligned}$$

- GAM2 - GAM1 model with added interaction between Hour and Weekend. In the model definition instead of specyfing $s(Hour, k = 24)$ as in GAM1, we now have `s(Hour, by = Weekend, k = 24)`,
- GAM3 - Instead of only grouping the Hour variable by the Weekend attribute, we group it by a joint factor variable WeekendHoliday, `s(Hour, by = WeekendHoliday, k = 24)`.
- GAM4 - GAM3 model with added tensor interaction between Temp and SolarRadiation and a tensor interaction between Temp and Humidity.

Table 3: Comparison of GAM models.

	RMSE		MAE			
	Train	Test	Train	Test	R-sq.(adj)	Deviance expl.
GLM	369.5129	376.4311	253.9520	259.8003	0.6649	0.6839
GAM1	235.5746	236.1368	155.2442	160.6714	0.8626	0.8721
GAM2	198.0070	198.9483	125.1158	128.5908	0.9027	0.9078
GAM3	187.8043	189.5029	114.3393	119.0462	0.9122	0.9160
GAM4	181.1643	182.9959	110.5214	115.3421	0.9181	0.9214

The GAM model that includes interaction between Hour and joint factor WeekendHoliday, tensor interaction between Temp and Humidity and tensor interaction Temp and SolarRadiation achieves the best scores. The RMSE obtained for that model is half of the RMSE for the GLM model. The most significant improvement in the fit between GAM models was achieved when the Hour-Weekend interaction was included. The scores for the training set and for the test set are similar, so there is no evidence of overfitting.

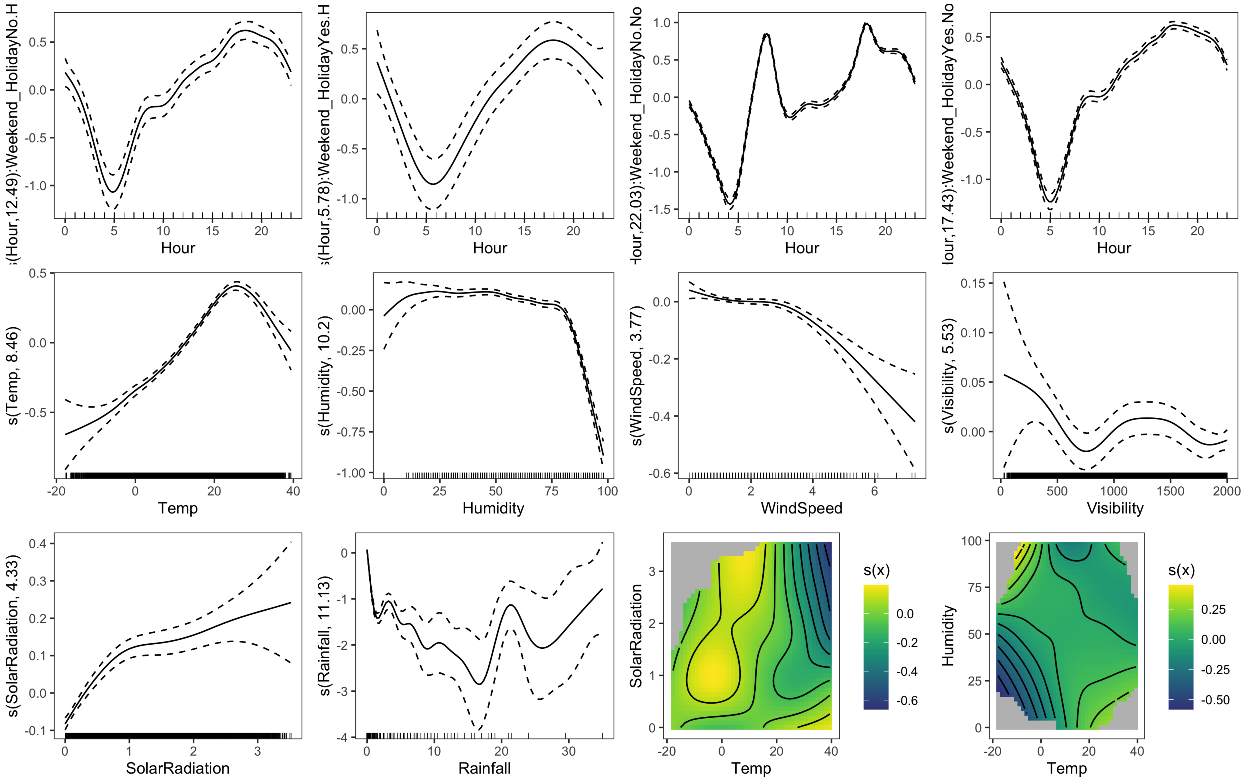

Figure 4: Partial effects of smooth terms in the GAM4 model, together with 95% confidence bands.

The partial effects plot shows that the relation between Hour and whether a day is weekend or holiday has been captured well, with distinctive peaks appearing only when the day is not a holiday and not a weekend. The partial effect of tensor interactions is quite difficult to interpret, as the main effects are also present, we can instead look at the summed effect displayed in Figure 5.

From the summed effects of temp we can conclude that most people rent out bikes in warm temperatures, around 20 degrees, with high solar radiation. When temperatures are on the highest end, we observe a decreasing trend in rentals with an increase in solar radiation. As for the joint effect of humidity of temperature, the highest number of rentals occurs during warm temperatures and when humidity is low.

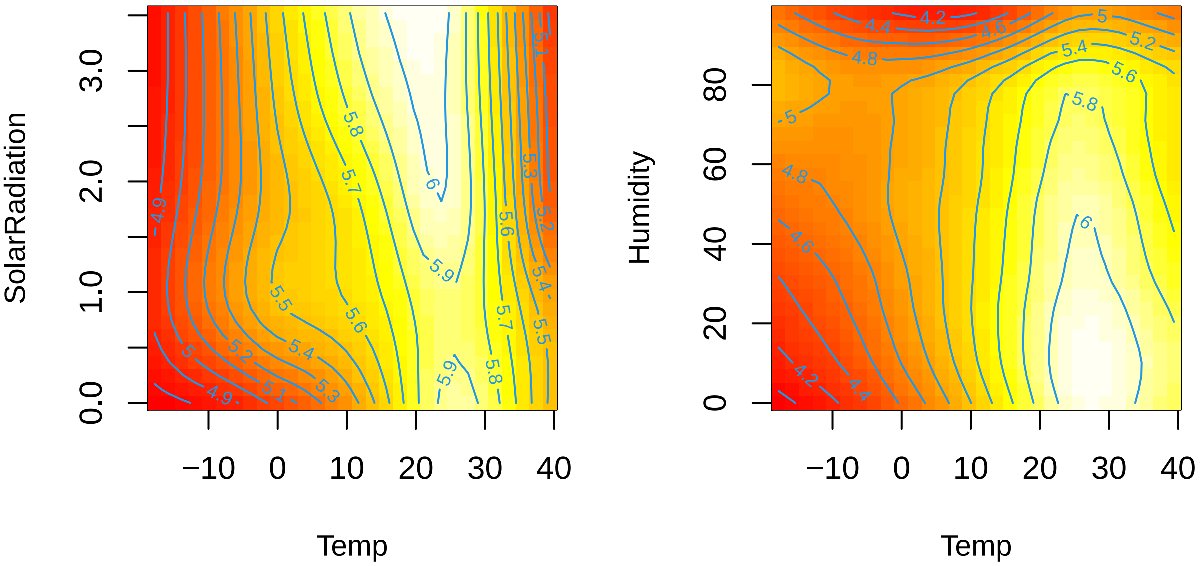

Figure 5: Summed effect of bivariate smooths.

References

Sathishkumar V E and Yongyun Cho. A rule-based model for seoul bike sharing demand prediction using weather data. *European Journal of Remote Sensing*, 53(sup1):166–183, 2020. doi: 10.1080/22797254.2020.1725789. URL https://doi.org/10.1080/22797254.2020.1725789.

Jaroslaw Harezlak, David Ruppert, and M. P. Wand. *Semiparametric regression with R*. Springer, 2018. ISBN 978-1493988518.

Jay M. Ver Hoef and Peter L. Boveng. Quasi-poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology*, 88(11):2766–2772, 2007. doi: https://doi.org/10.1890/07-0043.1. URL https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/07-0043.1.

Wikipedia. Bicycle-sharing system — Wikipedia, the free encyclopedia, 2022. URL https://en.wikipedia.org/w/index.php?title=Bicycle-sharing_system&oldid=1065313848.

Simon Wood. *Generalized Additive Models: An Introduction With R*, volume 66. 01 2006. ISBN 9781315370279. doi: 10.1201/9781315370279.

Simon N Wood. *Generalized Additive Models: An Introduction with R*. CRC Press, United States, 2 edition, 2017. ISBN 978-1498728331.