

CSIS Discussion Paper No. 141

A Method for Comparing Multiple Regression Models

Yuki Hiruta Yasushi Asami

Department of Urban Engineering, the University of Tokyo
e-mail: hiruta@ua.t.u-tokyo.ac.jp asami@csis.u-tokyo.ac.jp

January 2016

Abstract

In recent years, multiple regression models have been developed and are becoming broadly applicable for us. However, there are not many options for comparing the model qualities based on the same standard. This paper suggests a simple way for evaluating the different types of regression models from two points of view: the ‘data fitting’ and the ‘model stability’.

1. Introduction

In recent years, multiple regression models have been developed and are becoming broadly applicable for us. However, it is not a simple issue comparing the model qualities to the same standard. Because of the diversity of the models that are available, it is becoming inappropriate to apply just a single criterion such as Akaike information criterion (AIC). There are models that estimate a number of parameters; in some cases, the number of parameters exceeds the number of independent variables. Criteria such as GCV, namely generalized cross-validation (Craven & Wahba, 1978), are used to compare the model quality. However, in practical use, we sometimes need to select models by understanding the characteristic of the models from two points of view. First, how much the model fits the observation. Secondly, how stable the model can make the estimation. For standardizing the balance between these two points of view, this paper suggests a simple way for evaluating the different types of models based on the relative positions among models that are coordinated by two criteria. We call the first criterion the ‘data fitting criterion’ and the second the ‘model stability criterion’ hereafter.

2. Methods

2-1 Data set

Table 1 shows the descriptive statistics for the data set we apply, and Figure 1 displays a scatter diagram of them.

For obtaining two variables that have a non-linear relation in an unintended manner, we applied a sequence of x from -100–100 by 1 as an independent variable, while the dependent variable y was provided by Equation (1):

$$y = (x^5 + x^4 + x^3 + x^2 + x^1)/10000 + u \quad \cdots(1)$$

u is random error term generated from the uniform distribution with a lower limit of 0 and an upper limit of 1,010,101 which is the value of the first term $(x^5 + x^4 + x^3 + x^2 + x^1)/10000$ when x equals 100.

Table 1. Descriptive statistics for the employed data set

	y	x
Min.	-590811	-100
1st quantile	179671	-50
Median	505674	0
Mean	489201	0
3rd quantile	749369	50
Max.	1665768	100
Standard deviation	407724	58

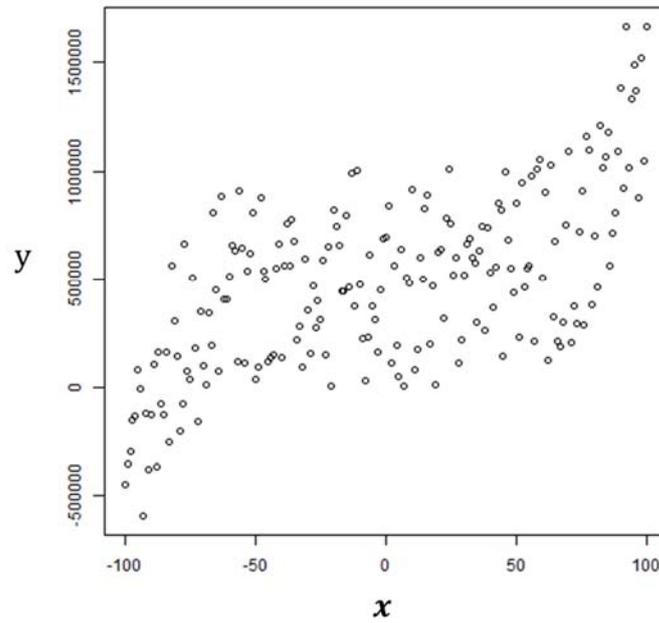


Figure 1. Scatter diagram of the employed data set

2-2 'Data fitting criterion' and 'model stability criterion'

We applied the method based on bootstrap sampling. First, permitting replacement, we randomly sampled the data from the original data set (called the 'Original Sample' hereafter), and obtained 100 sets of the data (called 'Bootstrap Samples' hereafter). Each 'Bootstrap Sample' has a sample size 201, the same as the 'Original Sample'. Secondly, we built the models (called 'Bootstrap Models' hereafter) using each of all the 100 'Bootstrap Samples', while the models (called 'Original Models' hereafter) were also built using the 'Original Sample'. Thirdly, two criteria for all models were acquired. One is called the 'data fitting criterion,' which is the total of the squared residuals of all 'Bootstrap Models'. The other is called the 'model stability criterion', which is the total value of the difference between the fitted value of the 'Bootstrap Models' and that of the 'Original Models'. Finally, relative positions among models that are coordinated by these two criteria are represented in a diagram.

2-3 Models

Some of the models have a smoothing or complexity parameter. In such models we can control the complexity of the model fitness by controlling the value of the smoothing or complexity parameter. It is known that there is a tradeoff; the complex models tend to fit the observations well but the estimation is not stable; the smooth models tend to be stable but do not fit the observations well. By controlling the smoothing or complexity parameter, we can also control such a tradeoff (Hastie et al., 2001).

To figure out whether the relative positions among models coordinated by two criteria ('data fitting criterion' and 'model stability criterion') are reasonable enough or not, we used the models that have a smoothing or complexity parameter, and confirmed whether the two criteria are able to represent such tradeoffs.

We employed four types of models that have a smoothing or complexity parameter: namely, GAM, SVM, Regression Tree and MARS. We used a sequence of a smoothing or complexity parameter to build the models that have different complexities. The settings and explanations about the models are shown in Table 2.

Table 2. Multiple Regression models applied

	Name of the Models (abbr.)	Explanations	Parameters applied
1	Generalized Additive Models (GAM)	Described in Wood, (2004); and Wood (2006)	The degree of model smoothness is controlled by the smoothing parameter 'sp', larger sp leads to a straight line estimate, while sp=0 results in an un-penalized regression spline estimate. A sequence of the 'sp' from 2^{20} to 2^{-20} by 2^1 was applied.
2	Support vector machine (SVM)	Described in Meyer, et al. (2015)	The degree of model complexity is controlled by the combination of the parameters 'C' and ' γ '. 'C' is the constant of the regularization term in the Lagrange formulation. ' γ ' is a parameter needed for all kernels. A sequence of C from 2^{-17} to 2^{17} by 2^1 was applied while ' γ ' is fixed to 1.
3	Multivariate adaptive regression splines (MARS)	Described in Friedman, (1991); and Milborrow, (2015)	The degree of model complexity is controlled by the maximum number of model terms 'nk'. A sequence of 'nk' from 1 to 60 by 2 was applied.
4	Regression Tree	Described in Therneau et al. (2015)	The size of regression trees is controlled by the parameter 'cp'. Larger 'cp' indicate the complex model with many branches, and smaller 'cp' indicate simple model with less branches. A sequence of 'cp' from 2^{-10} to 2^{-1} by $2^{0.5}$ was applied.

3. Results

3-1 Simulations

We observed the trade-off: the flexible models fit the data well but their estimations are not stable, and smooth models do not fit the data well but their estimations are stable, by the test under taken by GAM and SVM. MARS and Regression Tree showed a volatile result but it was assumed reasonable.

(1) GAM

In Figure 2, GAMs with different parameters were plotted based on the coordination by ‘data fitting criterion’ and ‘model stability criterion’. The horizontal axis shows the degree of the ‘data fitting criterion’, while the vertical axis shows the degree of the ‘model stability criterion’. The smaller ‘data fitting criterion’ indicates that the estimation by the model fits the data well. The smaller ‘model stability criterion’ indicates that the estimation by the model is stable.

Corresponding to the increase of the smoothing parameter ‘sp’, the degree of ‘data fitting criterion’ continuously increased with the decrease of the degree of ‘model stability criterion’. This result corresponds to the trade-off.

(2) SVM

We implemented the same test for SVM as represented in Figure 3. SVM has two parameters that relate to complexity: ‘C’ and ‘ γ ’. In advance, we tuned the model to acquire the parameters for the best fit using ‘Original Sample’. We implemented the same test by shifting the parameter ‘C’ while ‘ γ ’ was fixed as 1. As same as the results of GAMs, the results shown in Figure 3 corresponded to the trade-off.

(3) MARS

The same test was implemented for MARS as shown in Figure 4. Corresponding to the decrease of the parameter 'nk', the degree of 'data fitting criterion' increased continuously with the decrease of the degree of 'model stability criterion' as whole. However, we can recognize that only the model with 3 in 'nk' shows a higher 'model stability criterion' against the trade-off among other models. The observations were scattered with almost rotational symmetry with the three parts. It looks reasonable that the estimation cannot be stable if the model tries to fit such a rotational symmetry by two pieces of a piecewise function. It is also reasonable that the model with 1 in 'nk', which is the model of a single horizontal line, shows the highest degree of 'data fitting criterion' (i.e. it does not fit the data well) but the lowest 'model stability criterion' (i.e. stable); other estimations are unlikely to be made by a single horizontal line based on similar observations.

(4) Regression Tree

The same test was implemented for Regression Tree as shown in Figure 5. Although the degree of 'data fitting criterion' increased continuously corresponding to the increase of the parameter 'cp', 'model stability criterion' shows a volatile trace. In the range 'cp' translated from 2^{-6} to 2^{-2} , 'model stability criterion' sharply increased against the trade-off among the models. Similar to MARS, Regression Tree is a model of piecewise functions and is a floor function. It seems reasonable that the estimation becomes unstable, because many patterns are possible if we try to fit the observations by a small number of 'floors'. In addition, as the same as MARS, it is reasonable that the model with 2^{-1} in 'cp', which is the model of a single horizontal line, shows the highest degree of 'data fitting criterion' (i.e. it does not fit data well) but the lowest 'model stability criterion' (i.e. stable).

Summarizing the results, models that smoothly change their flexibility by transition of the smoothing or complexity parameter such as GAMs and SVMs, showed the trade-off relation well.

On the other hand, models with piecewise functions such as MARS and Regression Tree, although the degree of ‘data fitting criterion’ decreased (i.e. it does fit data well) continuously corresponding to the increase of the flexibility of the models, ‘model stability criterion’ showed volatile but reasonable responses. To the extent of these results acquired, ‘data fitting criterion’ and ‘model stability criterion’ reasonably reflect the model characteristics.

Additionally, the tendencies of the obtained results were similar in the cases where we reduced the number of the samples in a phased manner down to 10.

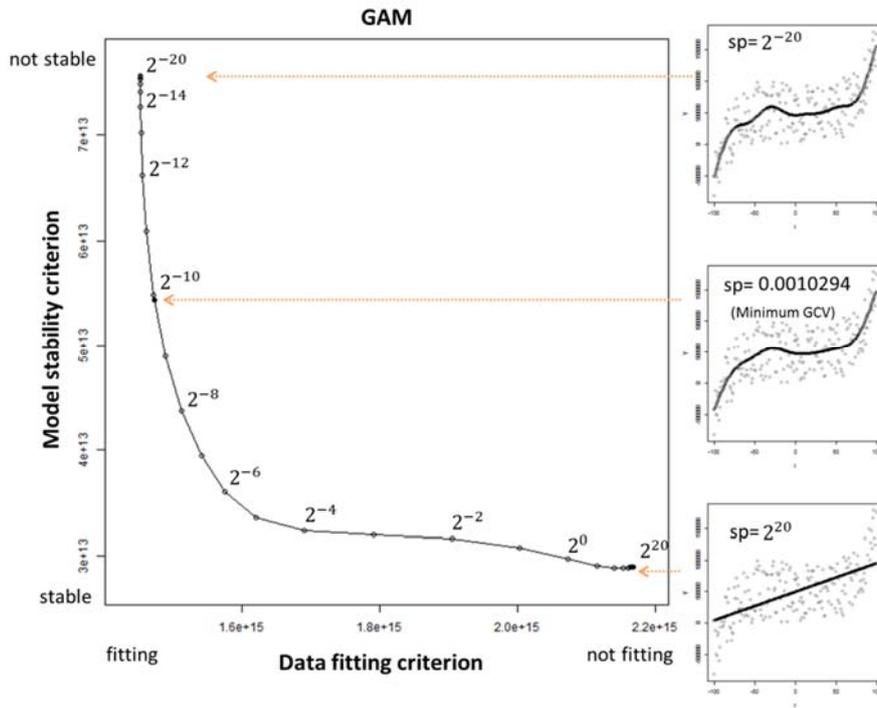


Figure 2. Evaluation on GAM

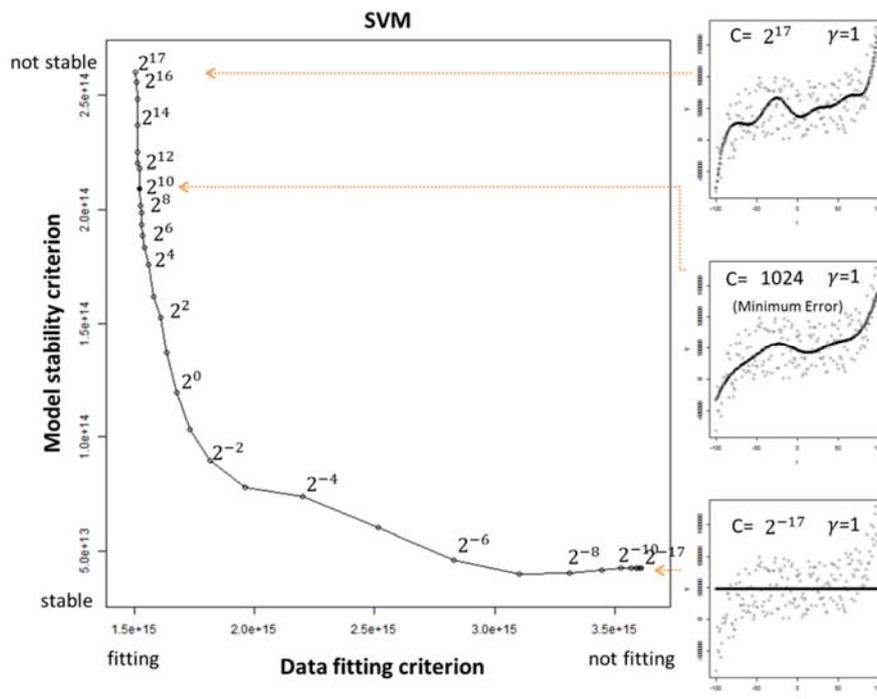


Figure 3. Evaluation on SVM

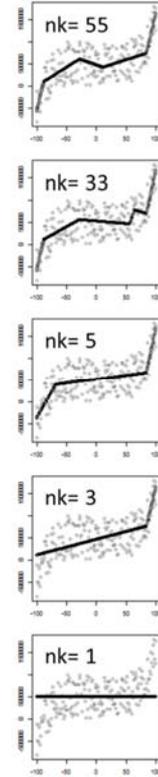
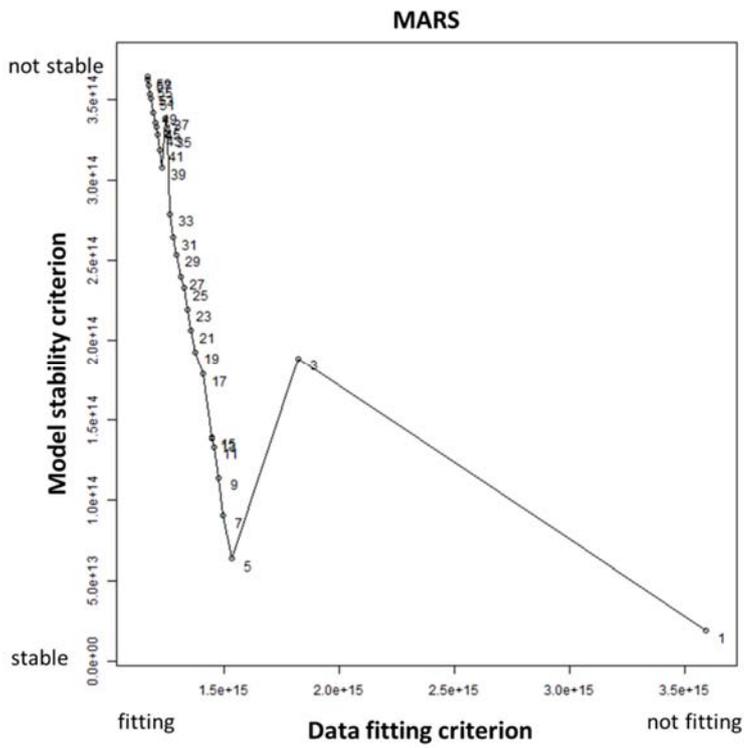


Figure 4. Evaluation on MARS

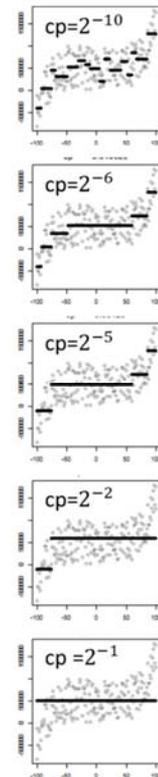
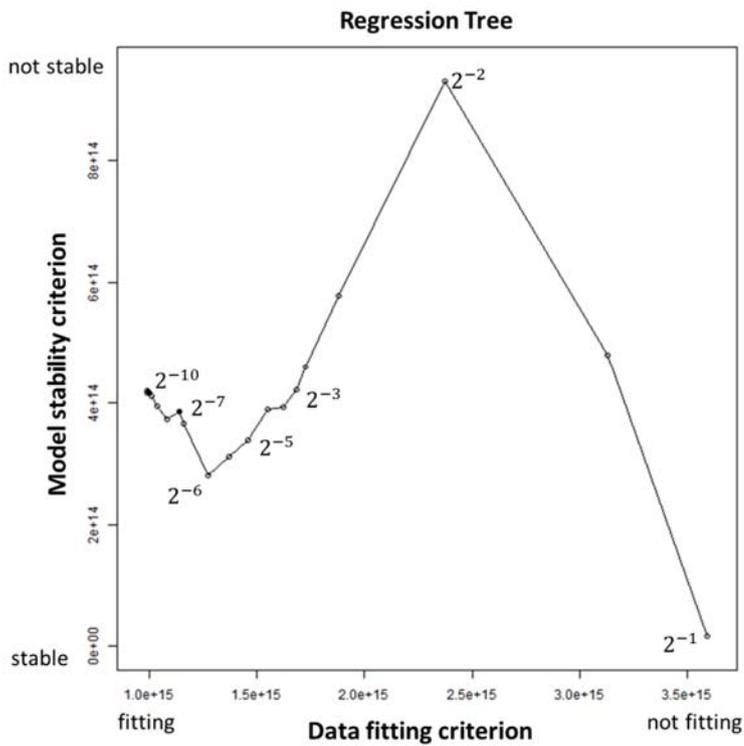


Figure 5. Evaluation on Regression Tree

3-2 Model comparison

Figure 6 shows the relative positions of all models based on the coordination by two criteria. The figure reasonably reflects the characteristics of the models.

In Figure 6, the smoothest models among MARSs, Regression Trees, and SVMs are placed very close together, and \hat{y} predicted by them are uniformly distributed. The smoothest model among GAMs shows the same prediction as \hat{y} predicted by OLS (ordinary least squares). Comparing the relative positions between the models that predict uniformly distributed \hat{y} and OLS, OLS is as stable as the models that predict uniformly distributed \hat{y} , but fits better the observations.

In contrast, the best-fit model built by Regression Tree is unstable, but fits data the best. \hat{y} predicted by the fittest model shows a fluctuating pattern, as if the estimation traces the observations.

GAM with a small smoothing parameter seems to be a good model for this sample dataset in terms of both model stability and data fitting.

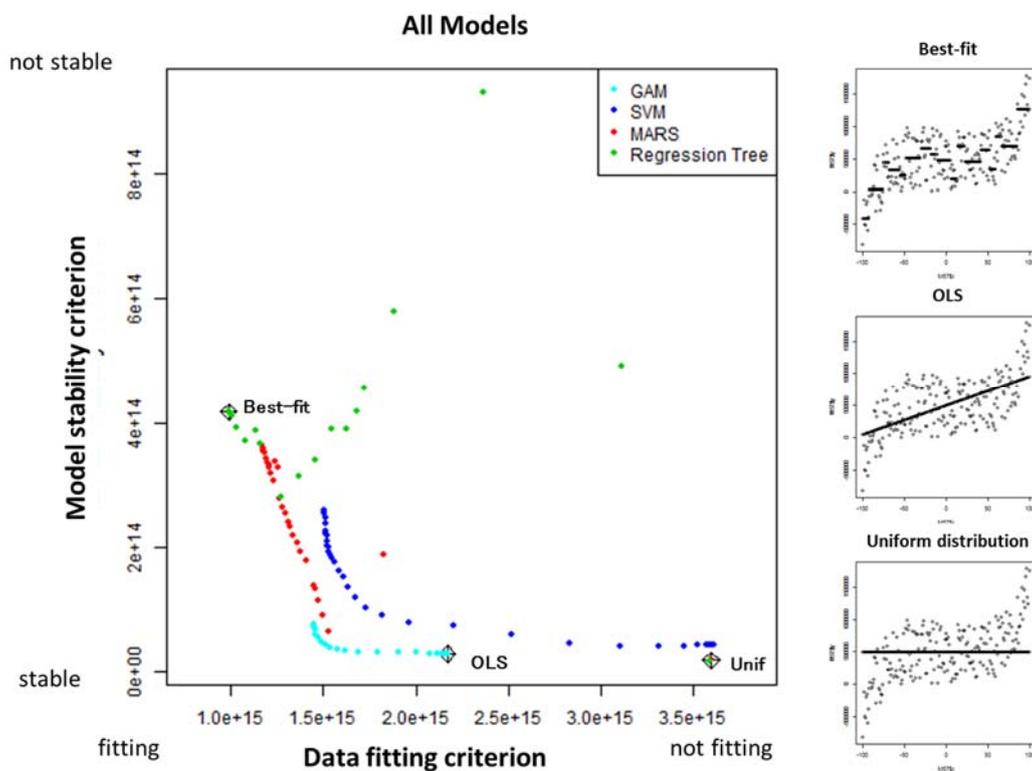


Figure 6. Model comparison

4. Conclusions

This paper has suggested a way for evaluating the different types of regression models based on the relative positions among models, which are coordinated by two criteria: ‘data fitting criterion’; ‘model stability criterion’.

It is known that complex models tend to fit observations well, but the estimation is not stable; smooth models tend to be stable but do not fit the observations well. To understand whether the criteria can properly represent the tradeoff, we used the models that have a parameter through which we can control model smoothness.

As model complexities were controlled by the parameters, the models that smoothly change their flexibility according to the transition of the smoothing or complexity parameter such as GAMs and SVMs clearly demonstrated the trade-off. On the other hand, models with piecewise functions such as MARSs and Regression Trees showed volatile but reasonable responses.

It is considered that two criteria represent the model characteristic property in terms of data fitting and model stability. Since only the observed values and the predicted values of each model are necessary for these criteria, the suggested criteria are very simple. We are able to calculate these criteria for multiple types of models, from the simplest OLS to machine learning algorithms, and to compare them to the same standard. By visualizing the positions coordinated by two criteria, we can understand the characteristics of multiple models as considering the balance between data fitting and model stability. It is considered this method can be one of the reliable options for selecting the appropriate model that meets the purpose of the modeling.

As a next step, by applying other data sets with known expected values, we will examine the substitutability between ‘data fitting criterion’ and the bias of the models, as well as ‘model stability criterion’ and the variance of the models. We will also examine applicability of the method by comparing the two indicators we suggested to the existing criteria such as GCV.

References

- Craven, P., & Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik*, **31**(4), 377–403. Available from: doi:10.1007/BF01404567
- Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, **19**(1), 1–67. Available from: doi:10.1214/aos/1176347963
- Hastie, T. J., Tibshirani, R. J., & Friedman, J. H. (2001). *The elements of statistical learning: data mining, inference, and prediction*. New York ; Tokyo: Springer.
- Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2015). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. Retrieved from <http://cran.r-project.org/package=e1071>
- Milborrow, S. (2015). *earth: Multivariate Adaptive Regression Splines*. Retrieved from <http://cran.r-project.org/package=earth>
- Therneau, T., Atkinson, B., & Ripley, B. (2015). *rpart: Recursive Partitioning and Regression Trees*. Retrieved from <http://cran.r-project.org/package=rpart>
- Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, **99**(467), 673–686.
- Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, Boca Raton.