Piecewise Regression through the Akaike Information Criterion using Mathematical Programming ⋆

Ioannis Gkioulekas, Lazaros G. Papageorgiou

Centre for Process Systems Engineering, Department of Chemical Engineering, UCL (University College London), Torrington Place, London WC1E 7JE, UK

Corresponding author: l.papageorgiou@ucl.ac.uk

Abstract: In machine learning, regression analysis is a tool for predicting the output variables from a set of known independent variables. Through regression analysis, a function that captures the relationship between the variables is fitted to the data. Many methods from literature tackle this problem with various degrees of difficulty. Some simple methods include linear regression and least squares, while some are more complicated such as support vector regression. Piecewise or segmented regression is a method of analysis that partitions the independent variables into intervals and a function is fitted to each interval. In this work, the *Optimal Piecewise Linear Regression Analysis* (*OPLRA*) model is used from literature to tackle the problem of segmented analysis. This model is a mathematical programming approach that is formulated as a mixed integer linear programming problem that optimally partitions the data into multiple regions and calculates the regression coefficients, while minimising the Mean Absolute Error of the fitting. However, the number of regions is a known priori. For this work, an extension of the model is proposed that can optimally decide on the number of regions using information criteria. Specifically, the Akaike Information Criterion is used and the objective is to minimise its value. By using the criterion, the model no longer needs a heuristic approach to decide on the number of regions and it also deals with the problem of overfitting and model complexity.

Keywords: Mathematical programming, Regression analysis, Optimisation, Information criterion, Machine learning

1. INTRODUCTION

In statistics, regression analysis is a process for estimating the relationship between variables. Specifically, given a dataset with dependent and independent variables, regression aims to understand how the value of the dependent variable changes when the independent variables vary. The goal of the analysis is to retrieve a mathematical function that correlates the predictors (i.e. independent variables) with the response (i.e. dependent variables)

There are many examples in the literature for regression analysis such as linear and least squares regression, Support Vector Regression (SVR) (Smola and Schölkopf, 2004), K-nearest neighbour (Korhonen and Kangas, 1997), Multivariate Adaptive Regression Splines (MARS) (Friedman, 1991) and Random Forest (Breiman, 2001). More recent work includes the automated learning of algebraic models for optimisation (ALAMO) (Cozad et al., 2014) and an optimization based regression approach called Optimal Piecewise Linear Regression Analysis (*OPLRA*) (Yang et al., 2016).

1.1 Over-fitting and Information Criteria

One of the challenges in statistics and machine learning is to develop a model that fits a set of training data but is also able to make predictions in other datasets. During this process, the problem of over-fitting arises. In over fitting, the proposed statistical model describes the set of training data but also its noise. That leads to a model that is tailored to fit the training dataset rather than reflecting the overall population, resulting in poor predictive performance (Hawkins, 2004). In the same sense, there is also the problem of under-fitting, where the constructed model is so simple that it is not able to capture the underlying relationship that might exist in the data.

Choosing the suitable model complexity can be achieved through *model selection and assessment.* In model selection, given a set of candidate models, the objective is to select the best model according to some criterion. Once the final model has been selected, the estimation of its prediction error in new data will assess the quality of the model. Such an assessment method is *cross validation.*

Tackling the model selection problem can be achieved through the use of information criteria. Information criteria are measures of the relative goodness of fit of a statistical model. Akaike proposed that a model should

⋆ This work was supported by The Leverhulme Trust under Grant number RPG-2015-240

be evaluated in terms of the goodness of the results by assessing the closeness between the predictive distribution defined by the model and the true distribution. The core idea is that introducing parameters to the model yields good results, but having a high degree of freedom will lead to an increase in the instability of the model, producing unreliable results (Konishi and Kitagawa, 2008).

1.2 Contribution of this work

The focus of this work is the topic of piecewise linear regression. Such regression methods partition the independent variable into multiple segments, also called regions, and a function is fitted to each one. The boundaries between the different regions are called break points and identifying the position of the break points is a key task for piecewise regression.

In order to perform piecewise linear regression, the *Optimal Piecewise Linear Regression Analysis* (*OPLRA*) model is used from literature (Yang et al., 2016). The main objective of that model is to receive a multivariate dataset as input, identify a single feature and partition the samples into multiple regions based on this feature. The partitioning of the samples into regions is solved as an optimisation problem, minimising the absolute error of the fitting. The number of regions has to be specified, enabling the model to decide on the optimal position of the break points and fit a linear function to each region. Knowing the number of regions in advance however is typically never the case. So an iterative approach was proposed by the authors to choose the optimal number of regions. An additional region is added with each iteration and the mathematical model is solved again. Convergence is achieved by satisfying a heuristic rule.

This work extends the *OPLRA* mathematical model and addresses the issue of selecting the appropriate number of regions. In an attempt to avoid over-fitting the data and generate a large number of regions, the Akaike Information Criterion is used. New binary variables are introduced to the mathematical model that are able to 'activate' regions so that samples can be allocated to them and the AIC is now the objective function of the optimisation problem. Minimising its value will lead to a model that balances complexity with predictive accuracy.

2. MATHEMATICAL FORMULATION

2.1 Akaike Information Criterion

The *Akaike Information criterion* is used in this work for model selection. The objective of the AIC is to estimate the amount of information that is lost when a model is used to approximate the true process that occurs and generates the data, by using the Kullback-Leibler distance. The criterion calculates the maximised log-likelihood, but Akaike showed that this is a biased term. This bias is approximately equal to the number of K parameters of the tested model, hence this term is introduced to the formulation (Burnham and Anderson, 2003).

Since the criterion is based on the 'difference' between the model and the true underlying mechanism, it is important to understand that the AIC is simply a measure of the relative quality of statistical models and it is not useful for assessing the model. As a result, if all of the models are poor, AIC will select the one that is estimated to be the best, even though this model might not perform well. When it comes to regression analysis, if all the candidate models assume normally distributed errors with a constant variance, then AIC can be easily computed as (Burnham and Anderson, 2003):

$$AIC = n \cdot \ln\left(\frac{RSS}{n}\right) + 2K \tag{1}$$

where:

RSS	residual sum of squares
n	number of observations
K	number of regression parameters

With this formulation, K is the total number of estimated regression parameters including the intercept. Since the criterion can only be used as a measure of relative quality, it stands to reason that the absolute value of the AIC is not important. Instead it is the relative value of the model that matters. The criterion chooses as the best model the one that has the lowest AIC value. An easy interpretation of the criterion and equation (1) is to think that the AIC 'rewards' descriptive accuracy via the residuals and 'penalizes' for model complexity.

The AIC has been established as one of the most frequently used information criterion for model selection problems, with a wide variety of applications. In human body composition predictions based on bioelectricity measurements, the AIC was included to reduce redundant influence factors (Chen et al., 2016), cancer research where AIC was used to develop a prognostic model in patients with germ cell tumors who experienced treatment failure with chemotherapy (Group, 2010) and outlier detection (Lehmann and Lösler, 2016). Also, Carvajal et al. (2016) considered an optimisation approach for model selection using the AIC, by incorporating the ℓ_0-(pseudo)norm as a penalty function to the log-likelihood function.

2.2 Mathematical Formulation of OPLRA

In this section the *OPLRA* mathematical programming model is described as found in literature (Yang et al., 2016).

Indices

s	observation, $s = 1, 2, ..., S$
m	independent input variable $m = 1, 2, ...M$
r	region, $r = 1, 2, ..., R$
m^*	the variable where sample partitioning takes place

Parameters

A_s^m	numeric value of observation s on variable m
Y_s	output value of observation s
U_1, U_2	arbitrary large positive numbers
ϵ	very small number

Positive variables

$X^r_{m^*}$	break-point r on partitioning variable m^*
D_s	training error between predicted output and real output for sample s

Variables

W^r_m	regression coefficient for variable m in region r
B^r	intercept of regression function in region r
Pr^r_s	predicted output for observation s in region r

Binary variables

F^r_s	1 if observation s falls into region r; 0 otherwise

Equations and Constraints

If there is total number of R regions, then there are $R-1$ break points. The following equations arranges them in an ordered way:

$$X^{r-1}_m \leq X^r_m \qquad \forall\ m = m^*,\ r = 2,3,...,R-1 \tag{2}$$

In order to assign samples into regions, binary variables are introduced to the model for the formulation of the following *big M* constraints:

$$X^{r-1}_m - U_1 \cdot (1- \ F^r_s) + \epsilon \leq A^m_s \qquad \forall\ s,\ r = 2,3,...,R,\ m = m^* \tag{3}$$

$$A^m_s \leq X^r_m + U_1 \cdot (1- \ F^r_s) - \epsilon \qquad \forall\ s,\ r = 1,2,...,R-1,\ m = m^* \tag{4}$$

The addition of the small number ϵ is done to ensure strict separation of samples into regions.

To enforce the logical constraint that each sample belongs only to one region, we use the following constraint:

$$\sum_r F^r_s = 1 \qquad \forall\ s \tag{5}$$

The prediction of the fitted model, Pr^r_s, is given by the following constraint:

$$Pr^r_s = \sum_m A^m_s \cdot W^r_m + B^r \qquad \forall\ s,r \tag{6}$$

The following two equations are used to formulate the absolute deviation between the real output and the predicted output of the model.

$$D_s \geq Y_s - Pr^r_s - U_2 \cdot (1 - F^r_s) \qquad \forall s,r \tag{7}$$

$$D_s \geq Pr^r_s - Y_s - U_2 \cdot (1 - F^r_s) \qquad \forall s,r \tag{8}$$

The objective function of the model is the minimisation of the absolute deviation error:

$$\min \sum_s D_s \tag{9}$$

The resulting model can be summarised as:
objective function (9)
subject to (2)-(8) constraints

and is formulated as an MILP problem that can be solved to global optimality.

This literature work introduced a heuristic procedure in order to identify the partitioning feature and then find the optimal number of regions. The partitioning feature is identified by solving the optimisation problem defined above for all the variables, while fixing the number of regions to 2, and choosing the one that yields the minimum fitting error.

Another iterative approach was used for selecting the optimal number of regions by introducing a new parameter which was used as a threshold to the reduction percentage of the absolute error. If the reduction percentage of the error is above that parameter, then a new region is added and the model will be solved again to improve the fitting.

2.3 Extended AIC approach

Using the Akaike criterion for model selection involves minimising its value to identify the best model in a set of candidate models. In this proposed approach, an extended mathematical model is constructed that aims to solve the problem of identifying the optimal number of regions by directly minimising the AIC value. So instead of solving multiple MILP models using the heuristic approach mentioned in section (2.2), the final result will be acquired by solving a single model called *AICO* (Akaike Information Criterion Optimisation).

Because of the logarithmic nature of the Akaike criterion, some adjustments have to be made in order to overcome the non-linear problems and formulate the model as an MILP. The first change is the approximation of the logarithmic function. To achieve this we approximate the value of the function with piecewise linear expressions.

The new additions to the model are presented below:

Indices

i	number of breaking points for the approximation , $i = 1,2,...N$

Variables

AIC	*Akaike information criterion* value
λ_i	SOS2 variable
G	The approximation of the logarithm

Binary variables

E_r	1 if region r is selected; 0 otherwise

Parameters

γ_i	The break points for the approximation
β_i	The 'output' of the breaking points (define the equation to be approximated)

Constraints

Some additional constraints have to be introduced in order to create a model that will choose the number of regions without any iterative approaches.

The following constraint ensures that observation s belongs to region r only if that region is selected:

$$F^r_s \leq E_r \qquad \forall\ r,s \tag{10}$$

That means that if a specific region r is selected, then the equivalent binary variable will be set to $E_r = 1$ allowing variable F_s^r to receive an value. Otherwise, if $E_r = 0$ then $F_s^r = 0$ as well.

The following constraint ensures that if region r is not selected, then all of the following regions in the set will not be selected as well:

$$E_{r+1} \leq E_r \qquad \forall\ r = 1, 2, ..., R-1 \tag{11}$$

The next set of equations are responsible for the linear approximation of the logarithm in the *Akaike criterion.* To achieve this, we introduce λ_i variables which are a SOS2 set. That means that at most two variables within this ordered set can take on non-zero values. Those two values have to be for adjacent variables in that set.

The first step is to define the function that needs to be approximated, in this case the $ln(x)$. Parameter γ_i is used to discretise the domain of that function.

$$\beta_i = \ln(\gamma_i) \qquad \forall i$$

The above equation is not part of the optimisation model. It is used to define the function that needs to be approximated by selecting γ_i points and calculating the equivalent β_i ‘output‘ points.

The new constraints that are introduced to the model are presented below. In these constraints another simplification is applied by using absolute error values instead of *RSS* for the AIC:

$$\sum_s D_s = \sum_i \gamma_i \cdot \lambda_i \tag{12}$$

$$G = \sum_i \beta_i \cdot \lambda_i \tag{13}$$

$$\sum_i \lambda_i = 1 \tag{14}$$

Equation (12) is used to describe the independent variable. In this case, the independent variable that we want to approximate is the absolute error $\sum_s D_s$. Equation (13) is used to calculate the value of the response, meaning the final approximated value, that was described in equation (12). Constraint (14) ensures that the sum of all the SOS2 variables will be 1.

The objective function of the optimisation model is the minimisation of the AIC value. So using equation (1) and modifying it to fit the notation of this work, the objective is formulated as follows:

$$\min AIC = S \cdot G - S \cdot \ln(S) + 2(M+1) \cdot \sum_r E_r \tag{15}$$

The last term in equation (15) is the penalty factor of the criterion, based on the number of parameters of the model. M is the total number of variables in the dataset and $\sum_r E_r$ is the total number of regions that will be selected. S is the total number of observations in the dataset and G is the approximation of the logarithm of the error of the fitting, as discussed in equations (12)-(14). The partitioning feature is once again identified by using the same heuristic that was discussed in section (2.2). The next step is to select the maximum number of regions R for the new extended model. The binary variable E_r that was introduced will decide the optimal number of regions that will be selected and constraint (10) will ensure that all of the samples belong to those regions. Overall, the proposed MILP model can be summarised as follows:

minimise objective function (15)

subject to constraints (2) - (8) and (10) - (14)

One key difference of the *AICO* model when compared to the previous approach is the computational time. This approach requires only one MILP model to solved, once the partitioning variable has been identified, instead of solving multiple MILP models iteratively. But one point of attention is the maximum number of allowable regions. This number should be large enough to ensure that the model will capture all of the necessary regions, but at the same time small enough to avoid generating unnecessary binary variables.

3. COMPUTATIONAL PART

3.1 The examined datasets

To test the proposed methods a number of real world datasets have been used.

Table 1. Datasets used in this work

Dataset	**Samples**	**Variables**
Pharmacokinetics	132	4
Bodyfat	252	14
Yacht Hydrodynamics	308	6
Sensory	576	11
Cooling efficiency	768	8
Heating efficiency	768	8
Earthquake	1000	4
Concrete	1030	8
White wine quality	4898	11

The datasets reported in table (1) are derived from different online sources. More specifically the pharmacokinetics and earthquake data are available through a package in R (R Development Core Team, 2016), bodyfat and sensory data are available through StatLib (Pantelis Vlachos, 2005) and the rest from the UCI repository (Lichman, 2013).

The datasets that are taken from the UCI repository are also used in the original work (Yang et al., 2016). The yacht hydrodynamics set predicts the residuary resistance of sailing yachts for estimating the required propulsive power. The energy efficiency dataset (Tsanas and Xifara, 2012) assesses the heating and cooling load requirements of different buildings. The concrete dataset (Yeh, 1998) tries to predict the compressive strength of concrete as a structural material. The wine dataset (Cortez et al., 2009) tries to predict the quality of white wine according to some of it’s properties.

As mentioned in section (2.2), the original *OPLRA* model used a heuristic approach to identify the number of regions

Table 2. Cross validation results

	Original datasets					*New datasets*			
	Yacht	Cooling	Heating	Concrete	Wine	Bodyfat	Sensory	Pharma	Earthquake
OPLRA	0.689	**1.275**	**0.805**	4.845	0.551	1.273	0.632	1.613	**7.238**
AICO	**0.678**	**1.275**	0.806	4.838	0.555	0.631	0.626	**1.288**	**7.238**
KNN	5.788	2.237	2.063	8.924	0.577	2.869	0.642	1.981	8.464
SVR	3.673	1.820	1.456	4.864	0.518	1.391	0.613	1.834	7.250
RandFor	2.454	1.326	0.861	**4.029**	**0.439**	1.532	**0.562**	1.677	7.978
Mars	1.079	1.340	0.826	4.932	0.569	**0.389**	0.616	1.420	7.389

and introduced a parameter as a stopping criterion. This user specified parameter was set at 0.03 after a sensitivty analysis was performed with the same UCI datasets. So in a way, the algorithm was tailored to those datasets. In order to test the accuracy and robustness of the proposed extension, new datasets are introduced in this work.

The pharmacokinetics dataset contains data from a study of the kinetics of the anti-asthmatic drug theophylline (Boeckmann et al., 1994) on twelve subjects, with the aim of predicting the final drug concentration. The earthquake dataset gives the location of seismic events that occurred near Fiji since 1964. The bodyfat dataset uses features such as age, weight and height to try and measure the percentage of bodyfat in a subject. The sensory dataset has data for the evaluation of wine quality by a total of 6 judges.

3.2 Validation of the method

In order to test the accuracy of the proposed method, 5-fold cross validation is performed for each dataset. The method of k-folds cross validation splits the data into k smaller sets of equal size. Then it uses one of these sets as a testing set while the rest are being used to train the model. The method stops when all of the k sets have been used as the testing set.

For each run of the cross validation the mean absolute error is calculated for every dataset. The final score is the average of 10 runs of 5-fold cross validation. Both the proposed method and the original *OPLRA* paper (Yang et al., 2016) are implemented in the General Algebraic Modeling System (GAMS) (GAMS Development Corporation, 2016) and are solved using the CPLEX MILP solver, with optimality gap set at 0 and a time limit of 400s for the proposed approach. The R programming language (R Development Core Team, 2016) is used to create the random partitioning of the data for the k-folds validation.

A number of methods from the literature are also implemented in this work and are compared with the proposed methods on the same datasets. The methods include *KNN* regression (Korhonen and Kangas, 1997), *Random Forest* regression, *MARS* regression (Friedman, 1991) and Support Vector regression (*SVR*) (Smola and Schölkopf, 2004). All of those methods are implemented in the R programming language (R Development Core Team, 2016) using the appropriate packages for each method. Once again, 10 runs of 5-fold cross validation are performed and the final results are compared to the proposed method.

4. RESULTS

Before applying all of the methods, *feature scaling* is performed to the datasets according to the following equation:

$$\frac{A_{s,m} - \min_s A_{s,m}}{\max_s A_{s,m} - \min_s A_{s,m}}$$

That means the predictors of the datasets are now all within the range of [0,1]. The main advantage of scaling is not having predictors with great numeric ranges that can potentially dominate those in smaller ranges. As a result, all of the breaking points that will be determined by the model will also be within that same range.

Table (2) compares the proposed extension in this work with the original model from literature (Yang et al., 2016). For each tested dataset, the lowest mean absolute error (MAE) achieved is marked with bold. For the Yahct Hydrodynamics dataset, the *AICO* provides the lowest MAE value, outperforming the other tested methods. For the Cooling efficiency, the lowest score is achieved by both the *AICO* and the *OPRLA*. On heating efficiency, the *OPLRA* model emerges as the best performer, while the *AICO* is a close second. On the concrete dataset, even though there is an improvement from the *OPRLA* model, the *AICO* approach still doesn't perform better than the competition. Finally, for the wine dataset, the *AICO* method performs almost as well as the *OPLRA* model.

Since the original *OPLRA* model and the value of the parameter mentioned in section (2.2) were constructed using the datasets from UCI, new datasets are introduced to try and eliminate any biased performance scores. From table (2) we can see that the proposed *AICO* approach outperforms *OPLRA* in all of the new datasets. More specifically, there is a noticeable difference in performance for the bodyfat and pharmacokinetics datasets and minor difference in the sensory dataset, while having the same score for the earthquake dataset.

We can see that overall the proposed method has the lowest error in only four of the datasets. However, examining the results closer, it is obvious that the method performs well since the error scores are always very close to the ones that have the best overall performance. To demonstrate this we are going to develop a graph comparing the overall performance of each dataset. This graph is very similar to the one found in the original work (Yang et al., 2016).

In this graph, the method that performed the best gets awarded 10 points, while the one that performed the worst gets 1 point. Everything else is within this range. The final performance score is the average across all of the available datasets.

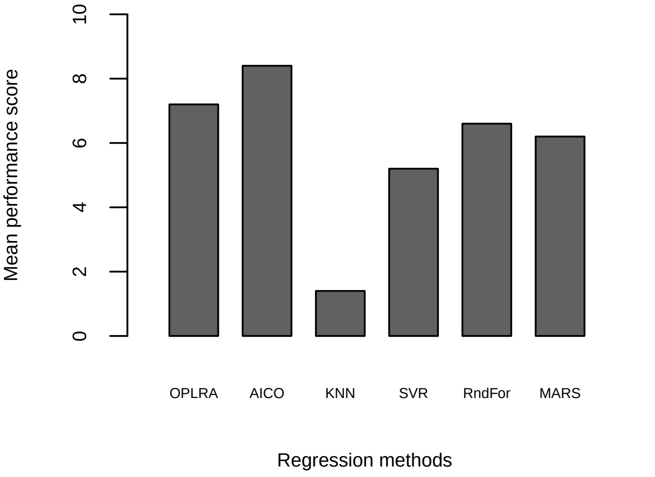

Fig. 1. The overall performance score of each method. The final score of each method is the average performance across all datasets.

5. CONCLUSION

This work addresses the problem of piecewise multivariate regression analysis. An extension of a current method from the literature is proposed that uses the Akaike information criterion in order to select the appropriate number of regions to split the data. The final optimisation model is able to select the number of regions, decide on the position of the break points and calculate the regression coefficients.

To test the method, several real world examples have been employed. The performance of the proposed method is compared to the original *OPLRA* work as well as other established regression methods. Computational experiments indicate that the new proposed approach has consistently better predictive performance than the original *OPLRA* work. That means that by using the AIC, the model is able to identify different number of regions compared to the original *OPLRA* model, leading to an increased performance. Additionally, the proposed method provides very competitive performance when compared to other regression methods. Figure (1) is a comparison of the predictive performance between all of the methods. Overall, the new proposed model is able to outperform the original *OPLRA* work as well as the other established methods that were considered in this work.

REFERENCES

Boeckmann, A., Sheiner, L., and Beal, S. (1994). *NONMEM Users Guide: Part V*. University of California, San Franscisco.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.

Burnham, K.P. and Anderson, D.R. (2003). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.

Carvajal, R., Urrutia, G., and Agero, J.C. (2016). An optimization-based algorithm for model selection using an approximation of akaike's information criterion. In *2016 Australian Control Conference (AuCC)*, 217–220.

Chen, B., Gao, X., Zheng, Q., and Wu, J. (2016). Research on human body composition prediction model based on akaike information criterion and improved entropy method. In *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 1882–1886.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *In Decision Support Systems*, 47, 547–553.

Cozad, A., Sahinidis, N.V., and Miller, D.C. (2014). Learning surrogate models for simulation-based optimization. *AIChE Journal*, 60(6), 2211–2227.

Friedman, J.H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, 1–67.

GAMS Development Corporation (2016). General Algebraic Modeling System (GAMS) Release 24.7.1, Washington, DC, USA.

Group, I.P.F.S. (2010). Prognostic factors in patients with metastatic germ cell tumors who experienced treatment failure with cisplatin-based first-line chemotherapy. *Journal of Clinical Oncology*, 28(33), 4906–4911.

Hawkins, D.M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1), 1–12.

Konishi, S. and Kitagawa, G. (2008). *Information criteria and statistical modeling.* Springer Science & Business Media.

Korhonen, K.T. and Kangas, A. (1997). Application of nearest-neighbour regression for generalizing sample tree information. *Scandinavian Journal of Forest Research*, 12(1), 97–101.

Lehmann, R. and Lösler, M. (2016). Multiple outlier detection: hypothesis tests versus model selection by information criteria. *Journal of surveying engineering*, 142(4).

Lichman, M. (2013). UCI machine learning repository.

Pantelis Vlachos (2005). StatLib-statistical datasets. *Available at* `http://lib.stat.cmu.edu/datasets/`.

R Development Core Team (2016). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Smola, A.J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, 14(3), 199–222.

Tsanas, A. and Xifara, A. (2012). Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, 49, 560 – 567.

Yang, L., Liu, S., Tsoka, S., and Papageorgiou, L. (2016). Mathematical programming for piecewise linear regression analysis. *Expert systems with applications*, 44, 156–167.

Yeh, I. (1998). Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*, 28(12), 1797 – 1808.