Computational Statistics and Data Analysis 54 (2010) 1197–1205

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

The Bayesian Additive Classification Tree applied to credit risk modelling

Junni L. Zhang[a,*], Wolfgang K. Härdle[b]

[a] *Department of Business Statistics and Econometrics, Guanghua School of Management, Peking University, Beijing 100871, PR China*

[b] *Center for Applied Statistics and Economics, Wirtschaftswissenschaftliche Fakultät, Humboldt-Universität zu Berlin, Spandauer Straße 1, 10178, Berlin, Germany*

RESEARCH ARTICLE

WILEY

Forecasting with many predictors using Bayesian additive regression trees

Jan Prüser[1,2]

[1]Faculty of Economics and Business Administration, University of Duisburg-Essen, Essen, Germany

[2]Ruhr Graduate School in Economics, RWI—Leibniz Institute for Economic Research, Essen, Germany

Correspondence
Jan Prüser, Ruhr Graduate School in Economics, RWI—Leibniz Institute for Economic Research, Hohenzollernstrasse 1–3, D-45128 Essen, Germany.
Email: jan.prueser@rgs-econ.de

Abstract

Forecasting with many predictors provides the opportunity to exploit a much richer base of information. However, macroeconomic time series are typically rather short, raising problems for conventional econometric models. This paper explores the use of Bayesian additive regression trees (Bart) from the machine learning literature to forecast macroeconomic time series in a predictor-rich environment. The interest lies in forecasting nine key macroeconomic variables of interest for government budget planning, central bank policy making and business decisions. It turns out that Bart is a valuable addition to existing methods for handling high dimensional data sets in a macroeconomic context.

KEYWORDS

fat data, forecasting, nonlinearity, variable selection

Accelerated Bayesian additive regression trees for fast multi-class classification

Meijia Wang, Jingyu He, Saar Yalov, Jared Murray, P. Richard H...

March 26, 2021

Focus Article

Bayesian treed response surface models

Hugh Chipman,[1] Edward I. George,[2] Robert B. Gramacy[3] and Robert McCulloch[3]*

Tree-based regression and classification, popularized in the 1980s with the advent of the classification and regression trees (CART) has seen a recent resurgence in popularity alongside a boom in modern computing power. The new methodologies take advantage of simulation-based inference, and ensemble methods, to produce higher fidelity response surfaces with competitive out-of-sample predictive performance while retaining many of the attractive features of classic trees: thrifty divide-and-conquer nonparametric inference, variable selection and sensitivity analysis, and nonstationary modeling features. In this paper, we review recent advances in Bayesian modeling for trees, from simple Bayesian CART models, treed Gaussian process, sequential inference via dynamic trees, to ensemble modeling via Bayesian additive regression trees (BART). We outline open source R packages supporting these methods and illustrate their use. © 2013 Wiley Periodicals, Inc.

How to cite this article:
WIREs Data Mining Knowl Discov 2013, 3: 298–305 doi: 10.1002/widm.1094

ISSN 1980-5411 (On-line version)
Production, 28, e20170110, 2018
DOI: 10.1590/0103-6513.20170110

PRODUCTION

Application of bayesian additive regression trees in the development of credit scoring models in Brazil

Daniel Alves de Brito Filho[a], Rinaldo Artes[a*]
[a]Insper, São Paulo, SP, Brasil
*rinaldoa@insper.edu.br

Abstract

Paper aims: This paper presents a comparison of the performances of the Bayesian additive regression trees (BART), Random Forest (RF) and the logistic regression model (LRM) for the development of credit scoring models.

Originality: It is not usual the use of BART methodology for the analysis of credit scoring data. The database was provided by Serasa-Experian with information regarding direct retail consumer credit operations. The use of credit bureau variables is not usual in academic papers.

Research method: Several models were adjusted and their performances were compared by using regular methods.

Main findings: The analysis confirms the superiority of the BART model over the LRM for the analyzed data. RF was superior to LRM only for the balanced sample. The best-adjusted BART model was superior to RF.

Implications for theory and practice: The paper suggests that the use of BART or RF may bring better results for credit scoring modelling.

Keywords
Credit. Machine learning. Logistic regression. BART. Random Forest.

Communications for Statistical Applications and Methods
2017, Vol. 24, No. 6, 543–559

https://doi.org/10.29220/CSAM.2017.24.6.543
Print ISSN 2287-7843 / Online ISSN 2383-4757

A review of tree-based Bayesian methods

Antonio R. Linero[1,a]

[a]Department of Statistics, Florida State University, USA

Abstract

Tree-based regression and classification ensembles form a standard part of the data-science toolkit. Many commonly used methods take an algorithmic view, proposing greedy methods for constructing decision trees; examples include the classification and regression trees algorithm, boosted decision trees, and random forests. Recent history has seen a surge of interest in Bayesian techniques for constructing decision tree ensembles, with these methods frequently outperforming their algorithmic counterparts. The goal of this article is to survey the landscape surrounding Bayesian decision tree methods, and to discuss recent modeling and computational developments. We provide connections between Bayesian tree-based methods and existing machine learning techniques, and outline several recent theoretical developments establishing frequentist consistency and rates of convergence for the posterior distribution. The methodology we present is applicable for a wide variety of statistical tasks including regression, classification, modeling of count data, and many others. We illustrate the methodology on both simulated and real datasets.

Keywords: Bayesian additive regression trees, boosting, random forests, semiparametric Bayes

Computational Statistics and Data Analysis 54 (2010) 1197–1205

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

The Bayesian Additive Classification Tree applied to credit risk modelling

Junni L. Zhang[a,*], Wolfgang K. Härdle[b]

[a] *Department of Business Statistics and Econometrics, Guanghua School of Management, Peking University, Beijing 100871, PR China*

[b] *Center for Applied Statistics and Economics, Wirtschaftswissenschaftliche Fakultät, Humboldt-Universität zu Berlin, Spandauer Straβe 1, 10178, Berlin, Germany*

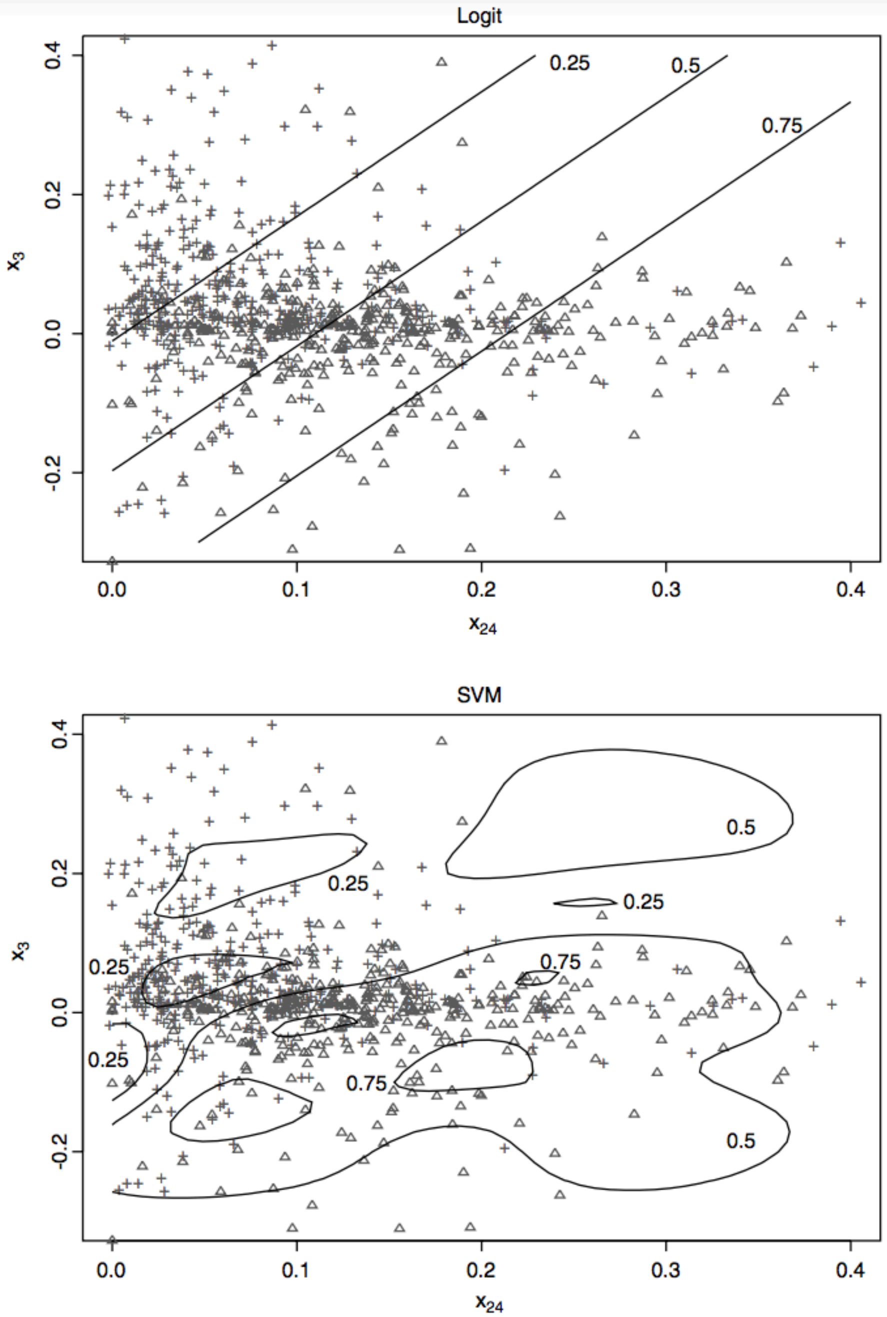

Fig. 2. The contour plots for the logit model and SVM. The triangles and pluses represent insolvent firms and solvent firms respectively. The numbers by the contours indicate the probabilities of insolvency.

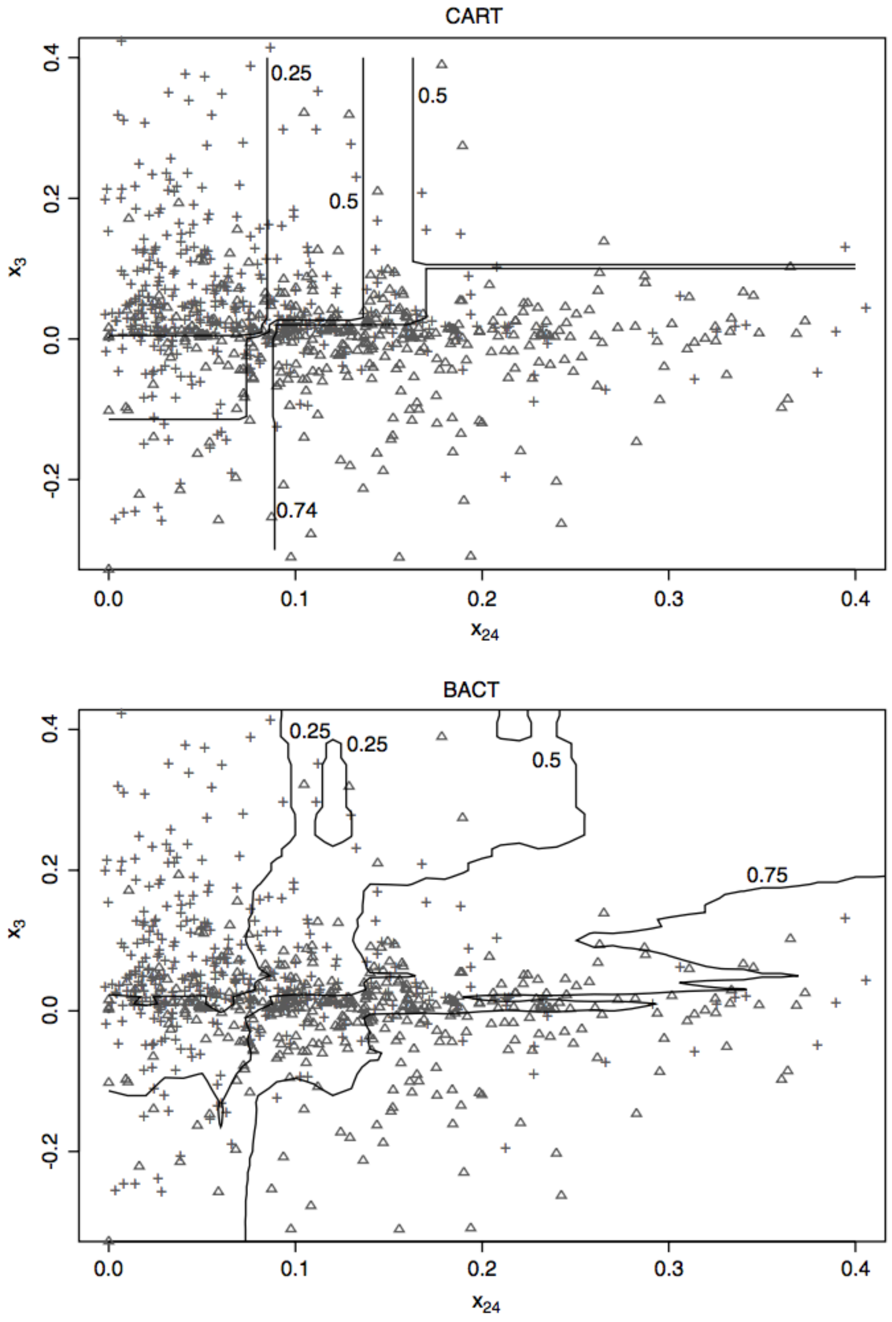

Fig. 3. The contour plots for CART and BACT. The triangles and pluses represent insolvent firms and solvent firms respectively. The numbers by the contours indicate the probabilities of insolvency.

Table 3
Definition of financial variables to be used for classification for the Creditreform data.

Var.	Definition
$x1$	Net income/total assets
$x2$	Net income/total sales
$x3$	Operating income/total assets
$x4$	Operating income/total sales
$x5$	Earnings before interest and tax/total assets
$x6$	Earnings before interest, Tax, Depreciation and amortization/total assets
$x7$	Earnings before interest and tax/total sales
$x8$	Own funds/total assets
$x9$	(Own funds − intangible assets) /(total assets − intangible assets − cash and cash equivalents − lands and buildings)
$x10$	Current liabilities/total assets
$x11$	(Current liabilities − cash and cash equivalents)/total assets
$x12$	Total liabilities/total assets
$x13$	Debt/total assets
$x14$	Earnings before interest and tax/interest expense
$x15$	Cash and cash equivalents/total assets
$x16$	Cash and cash equivalents/current liabilities
$x17$	(Cash and cash equivalents − inventories)/current liabilities
$x18$	Current assets/current liabilities
$x19$	(Current assets − current liabilities)/total assets
$x20$	Current liabilities/total liabilities
$x21$	Total assets/total sales
$x22$	Inventories/total sales
$x23$	Accounts receivable/total sales
$x24$	Accounts payable/total sales
$x25$	log(total assets)
$x26$	Increase (decrease) in inventories/inventories
$x27$	Increase (decrease) in liabilities/total Liabilities
$x28$	Increase (decrease) in cash flow/cash and cash equivalents

Table 5
The average values of AR and the three types of misclassification rates for the Logit model, CART, random forest, gradient boosting and BACT.

Performance measure	Logit (%)	CART (%)	Random forest (%)	Gradient boosting (%)	BACT (%)
AR	52.1	58.7	58.6	61.0	60.4
Overall misclassification rate	30.2	33.8	27.4	26.7	26.6
Type I Misclassification rate	28.3	27.2	26.9	26.8	27.6
Type II Misclassification rate	30.3	34.3	27.5	26.7	26.5

Accelerated Bayesian additive regression trees for fast multi-class classification

Meijia Wang, Jingyu He, Saar Yalov, Jared Murray, P. Richard Hahn

March 26, 2021

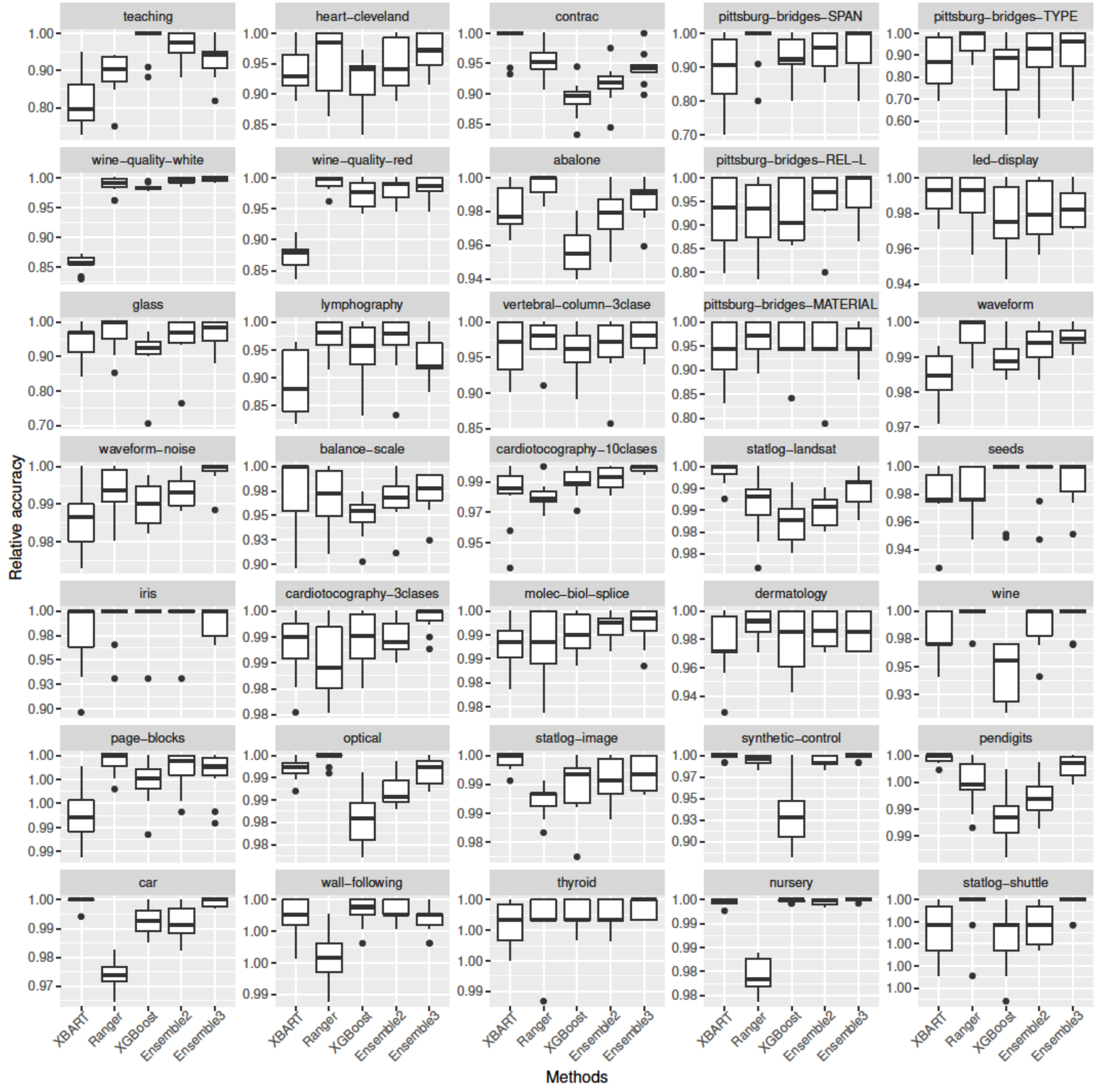

Figure 4: Relative accuracy of XBART, Ranger, XGBoost, and ensemble models on 35 UCI classification datasets. Ensemble2 combines Ranger and XGBoost; Ensemble3 combines all three methods.

RESEARCH ARTICLE

WILEY

Forecasting with many predictors using Bayesian additive regression trees

Jan Prüser[1,2]

[1]Faculty of Economics and Business Administration, University of Duisburg-Essen, Essen, Germany

[2]Ruhr Graduate School in Economics, RWI—Leibniz Institute for Economic Research, Essen, Germany

Correspondence
Jan Prüser, Ruhr Graduate School in Economics, RWI—Leibniz Institute for Economic Research, Hohenzollernstrasse 1–3, D-45128 Essen, Germany.
Email: jan.prueser@rgs-econ.de

Abstract

Forecasting with many predictors provides the opportunity to exploit a much richer base of information. However, macroeconomic time series are typically rather short, raising problems for conventional econometric models. This paper explores the use of Bayesian additive regression trees (Bart) from the machine learning literature to forecast macroeconomic time series in a predictor-rich environment. The interest lies in forecasting nine key macroeconomic variables of interest for government budget planning, central bank policy making and business decisions. It turns out that Bart is a valuable addition to existing methods for handling high dimensional data sets in a macroeconomic context.

KEYWORDS

fat data, forecasting, nonlinearity, variable selection

TABLE B1 Out-of-sample results: MAE relative to AR(1)

Variable	BartMSE	BartMAE	BartPL	BartBM	Fac1	Fac2	Fac3	Lasso1	Lasso2
One quarter ahead									
rGDP	0.94	0.97	0.94	0.94	**0.92**	1.03	1.01	0.98	0.98
rPCE	**0.95**	0.99	0.97	0.95	0.98	1.04	1.01	1.06	1.06
IND	**1.01**	1.03	1.08	1.05	1.09	1.15	1.10	1.20	1.20
UNEM	0.93	**0.92**	0.93	0.93	1.04	1.04	0.95	1.10	1.11
GDPdef	1.18	1.02	1.10	**0.99**	2.78	1.94	1.86	2.12	2.10
PCEdef	**0.91**	1.33	0.95	0.91	1.86	1.34	1.38	1.02	1.02
CPI	0.64	0.66	0.64	**0.61**	0.95	0.91	0.85	0.64	0.64
FED	1.33	**0.89**	0.90	1.03	1.07	1.11	1.15	1.28	1.27
GS10	0.99	**0.98**	1.02	0.99	1.02	1.01	1.05	1.02	1.02
One year ahead									
rGDP	1.05	1.04	**1.02**	1.02	1.03	1.15	1.16	1.08	1.08
rPCE	1.06	1.09	1.06	**1.04**	1.09	1.16	1.19	1.25	1.26
IND	**1.02**	1.03	1.06	1.02	1.03	1.18	1.14	1.12	1.13
UNEM	0.98	0.95	0.95	**0.92**	1.11	1.14	1.00	1.03	1.03
GDPdef	0.90	**0.83**	0.84	0.87	2.90	1.97	2.03	2.22	2.23
PCEdef	0.79	**0.75**	0.83	0.76	1.77	1.33	1.41	0.80	0.81
CPI	0.50	0.48	0.50	0.49	0.92	0.92	0.80	**0.45**	0.46
FED	0.97	0.93	**0.90**	0.95	0.93	0.97	1.11	1.02	1.04
GS10	0.96	0.98	0.95	**0.90**	0.98	1.00	0.99	0.96	0.95

Note. The table shows the forecasting performance of the Bart model with four different specifications, the factor model with one to three factors, and the Lasso approach with two different hierarchical priors. The forecasting performance is measured by the mean absolute forecasting error (MAE) and values below indicate that the model outperforms the AR(1) model.

TABLE B3 Out-of-sample results: PL – PL of AR(1)

Variable	BartMSE	BartMAF	BartPL	BartBM	Fac1	Fac2	Fac3	Lasso1	Lasso2
One quarter ahead									
rGDP	15.67	**18.31**	5.73	16.10	5.62	6.49	9.19	11.49	12.24
rPCE	8.49	12.77	**13.04**	11.36	−1.49	2.44	7.45	−1.63	−1.79
IND	**17.57**	15.63	3.69	15.15	−3.62	0.15	10.79	−8.67	−9.89
UNEM	13.46	13.25	13.97	**16.25**	−3.72	2.50	11.93	−11.69	11.83
GDPdef	−9.22	6.92	3.95	**12.22**	−103.43	−70.71	−62.21	−69.11	−69.14
PCEdef	15.57	−1.77	**16.39**	15.61	−47.26	−20.61	−20.78	13.60	13.00
CPI	0.54	40.30	40.44	41.86	5.11	8.09	15.88	41.13	**43.41**
FED	29.79	16.99	15.11	**30.81**	6.52	5.63	4.32	6.05	6.08
GS10	**1.14**	0.57	−1.8	0.58	−3.90	−3.65	−5.99	−1.85	−1.26
One year ahead									
rGDP	9.01	−0.87	**15.79**	12.26	−3.02	−5.01	−3.39	−1.15	−1.03
rPCE	−6.17	−12.11	−6.32	**−0.55**	−11.46	−12.41	−12.41	−21.80	−22.27
IND	12.13	3.52	**15.01**	13.31	−0.76	−10.34	3.81	−9.74	−9.66
UNEM	1.23	**15.38**	10.75	15.03	−4.62	−3.14	10.77	−3.32	−4.55
GDPdef	29.05	30.42	**30.43**	29.16	−3.02	−5.01	−3.39	−72.14	−71.26
PCEdef	32.04	**33.64**	17.84	33.21	−48.75	−20.53	−28.90	32.76	32.84
CPI	68.23	**72.13**	69.70	70.81	6.02	7.92	19.96	73.71	73.77
FED	9.94	14.46	**21.37**	21.52	9.02	7.42	1.94	8.81	7.69
GS10	1.72	0.46	2.41	3.29	3.67	−1.77	−3.00	3.89	**4.03**

Note. The table shows the forecasting performance of the Bart model with four different specifications, the factor model with one to three factors and the Lasso approach with two different hierarchical priors. The forecasting performance is measured by the sum of log-predictive likelihoods (PL) and positive values indicate that the model outperforms the AR(1) model.

ISSN 1980-5411 (On-line version)
Production, 28, e20170110, 2018
DOI: 10.1590/0103-6513.20170110

PRODUCTION
PRODUÇÃO

Application of bayesian additive regression trees in the development of credit scoring models in Brazil

Daniel Alves de Brito Filho[a], Rinaldo Artes[a]*
[a]Insper, São Paulo, SP, Brasil
*rinaldoa@insper.edu.br

Abstract

Paper aims: This paper presents a comparison of the performances of the Bayesian additive regression trees (BART), Random Forest (RF) and the logistic regression model (LRM) for the development of credit scoring models.

Originality: It is not usual the use of BART methodology for the analysis of credit scoring data. The database was provided by Serasa-Experian with information regarding direct retail consumer credit operations. The use of credit bureau variables is not usual in academic papers.

Research method: Several models were adjusted and their performances were compared by using regular methods.

Main findings: The analysis confirms the superiority of the BART model over the LRM for the analyzed data. RF was superior to LRM only for the balanced sample. The best-adjusted BART model was superior to RF.

Implications for theory and practice: The paper suggests that the use of BART or RF may bring better results for credit scoring modelling.

Keywords
Credit. Machine learning. Logistic regression. BART. Random Forest.

4. Database

The database used in this paper was provided by Serasa Experian and contains data regarding customers of direct retail consumer credit operations. The database provided by the credit bureau has 10,356 customer observations of direct retail consumer credit operations and 198 variables for the year 2014. Although random trees and BART were designed for larger datasets it is not unusual to find papers that aim to compare estimation methods designed for big datasets with sample sizes equivalent to ours, see, for instance, Chipman et al. (2010), Yeh et al. (2012), Leong (2016), Abellán & Castellano (2017), Bequé & Lessmann (2017), and several papers analysed in Lessmann et al. (2015) review.

The first group of predictor variables includes the amount of demand for credit of a specific borrower, in several different segments and in different periods of time. The segments are checks, real estate, banks, financial agencies, industries, insurance, services, telephony, retailer, utilities and others. The credit demand periods are up to 30 days, from 31 to 60 days, from 61 to 90 days, from 91 to 180 days and from 181 to 360 days, totaling 68 independent variables.

The second group of predictor variables is related to the first group. These variables measure the time in days since the first demand and since the last credit demand of a specific borrower by several segments. The segments are checks, banks, financial agencies, insurance, telecommunication and retail. This group has a total of 12 independent variables.

The third group is related to the number of events of the borrower registered in the credit bureau during certain periods of time. Events recorded at the bureau are active or settled debts, protests, bounced checks, active or resolved refusals by bank or financial agency, active or resolved refusal by companies that are not banks or financial agencies and active creditors. The time periods are 1 month, 2 months, 3 months, 6 months, 12 months, 2 years and 5 years. This group has a total of 60 independent variables.

Finally, the fourth group of predictors is related to the third group and measures the financial value registered in the credit bureau related to the described events. This group has a total of 40 independent variables.

Thirteen variables were excluded due the large number of missing values.

In addition to the described variables, whether the borrower was a “good” or “bad” payer was also indicated; this variable was used as a dependent variable in the calibration of the credit scoring model based on past data. However, the credit bureau did not report the criteria used to qualify borrowers as “good” or “bad” payers.

Table 6. Comparison of the AUC of the different models.

Sample	Hypotheses	Test			
		Delong		Bootstrap	
		z	p	z	p
Balanced	H_0: Log. Reg. = R. Forest	-1.958	0.050	-1.934	0.053
	H_0: Default BART = R. Forests	-1.246	0.213	-1.246	0.213
	H_0: Default BART = Log. Reg.	-3.322	0.001	-3.332	0.001
Unbalanced	H_0: Logistic Reg. = R. Forest	-0.922	0.356	-0.967	0.334
	H_0: BART = R. Forests	-2.028	0.043	-2.004	0.045
	H_0: BART = Logistic Regression	-2.869	0.004	-2.788	0.005
	H_0: Default BART = R. Forest	-0.884	0.376	-0.907	0.365
	H_0: Default BART = Log. Reg.	-1.977	0.048	-1.958	0.050

Communications for Statistical Applications and Methods
2017, Vol. 24, No. 6, 543–559

https://doi.org/10.29220/CSAM.2017.24.6.543
Print ISSN 2287-7843 / Online ISSN 2383-4757

A review of tree-based Bayesian methods

Antonio R. Linero[1,a]

[a]Department of Statistics, Florida State University, USA

Abstract

Tree-based regression and classification ensembles form a standard part of the data-science toolkit. Many commonly used methods take an algorithmic view, proposing greedy methods for constructing decision trees; examples include the classification and regression trees algorithm, boosted decision trees, and random forests. Recent history has seen a surge of interest in Bayesian techniques for constructing decision tree ensembles, with these methods frequently outperforming their algorithmic counterparts. The goal of this article is to survey the landscape surrounding Bayesian decision tree methods, and to discuss recent modeling and computational developments. We provide connections between Bayesian tree-based methods and existing machine learning techniques, and outline several recent theoretical developments establishing frequentist consistency and rates of convergence for the posterior distribution. The methodology we present is applicable for a wide variety of statistical tasks including regression, classification, modeling of count data, and many others. We illustrate the methodology on both simulated and real datasets.

Keywords: Bayesian additive regression trees, boosting, random forests, semiparametric Bayes

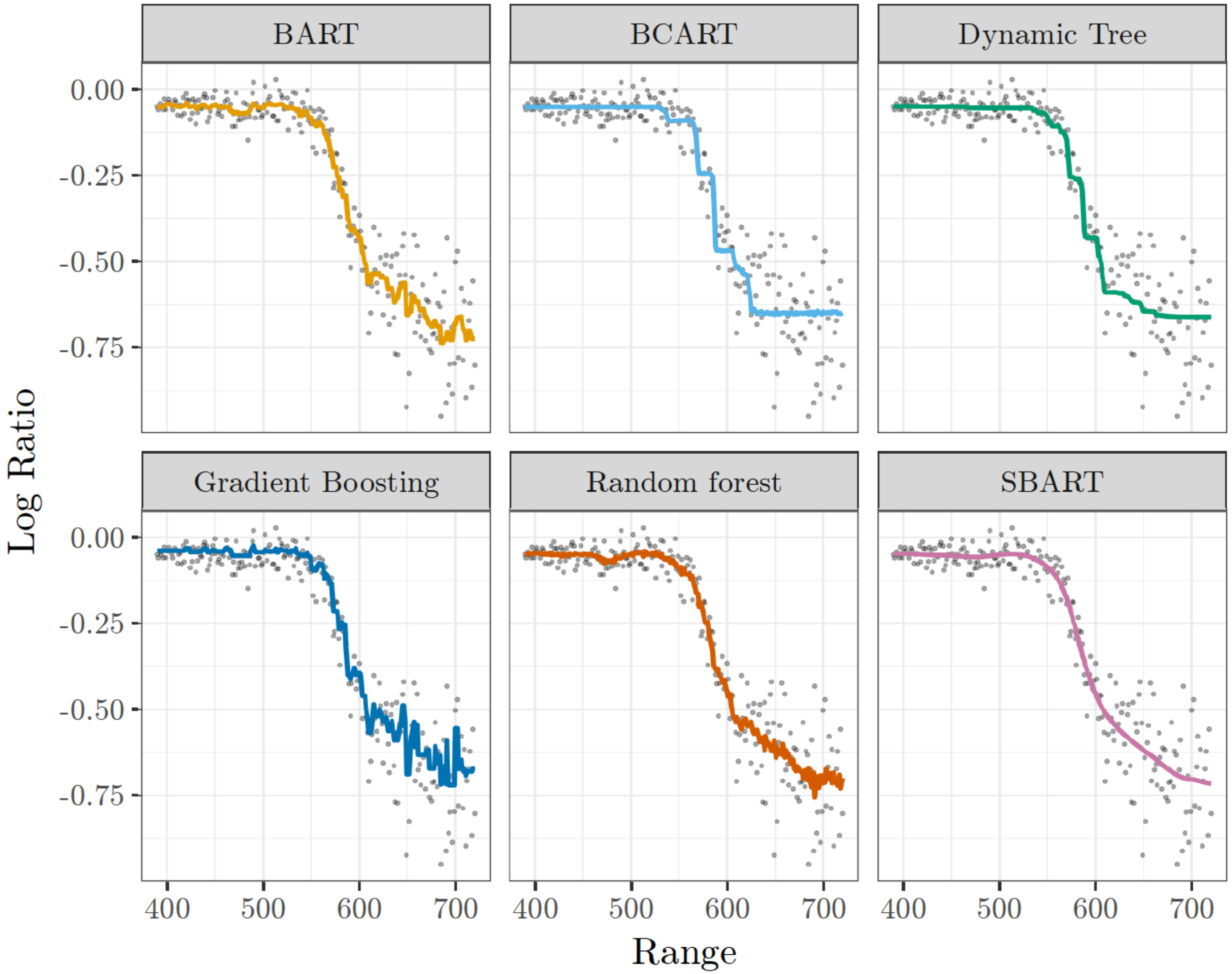

Figure 3: *Fits of various methods to the* `lidar` *dataset. BART = Bayesian additive regression trees; BCART = Bayesian classification and regression trees; SBART = smoothed Bayesian additive regression trees.*

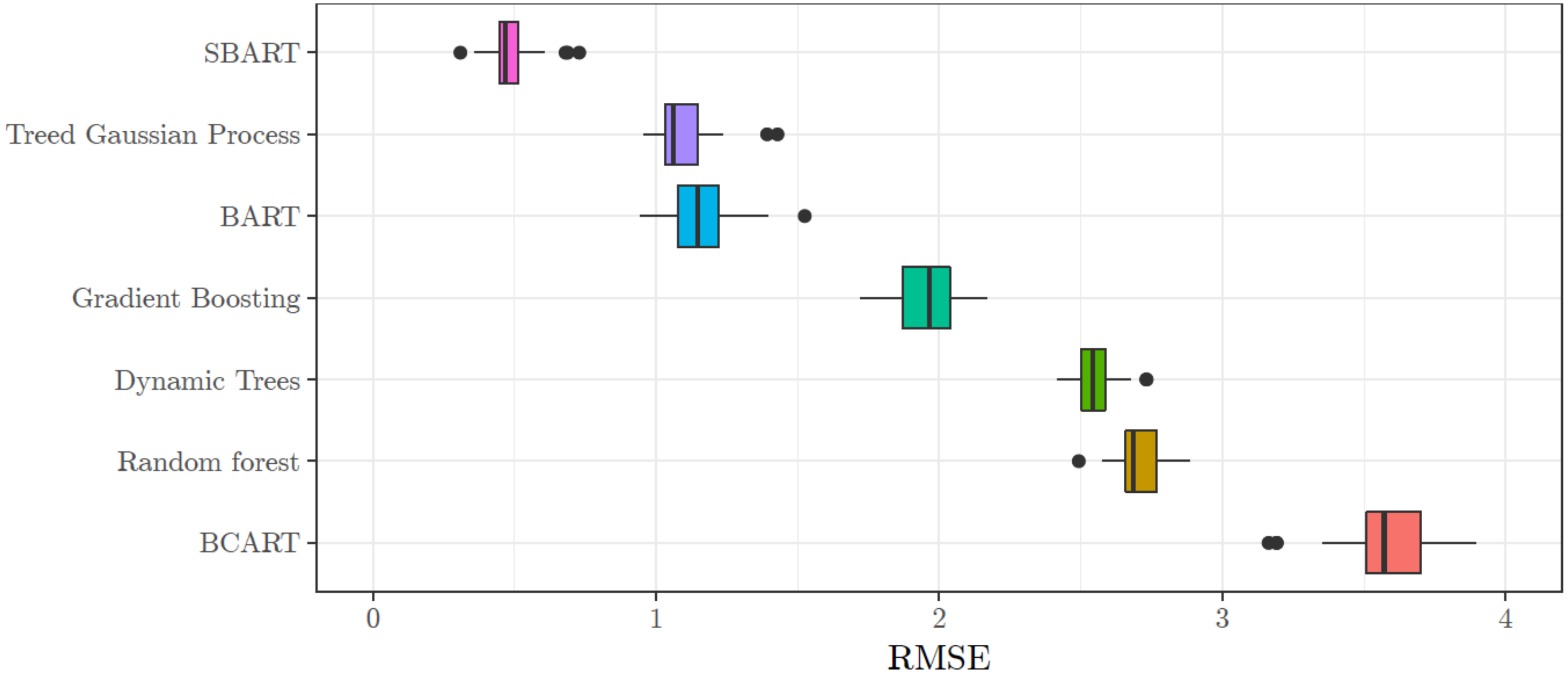

Figure 4: *Integrated RMSE under* (4.1) *across 30 independent replications, with* $n = 250, \sigma^2 = 1$. *SBART = smoothed Bayesian additive regression trees; BART = Bayesian additive regression trees; BCART = Bayesian classification and regression trees; RMSE = root mean squared error.*

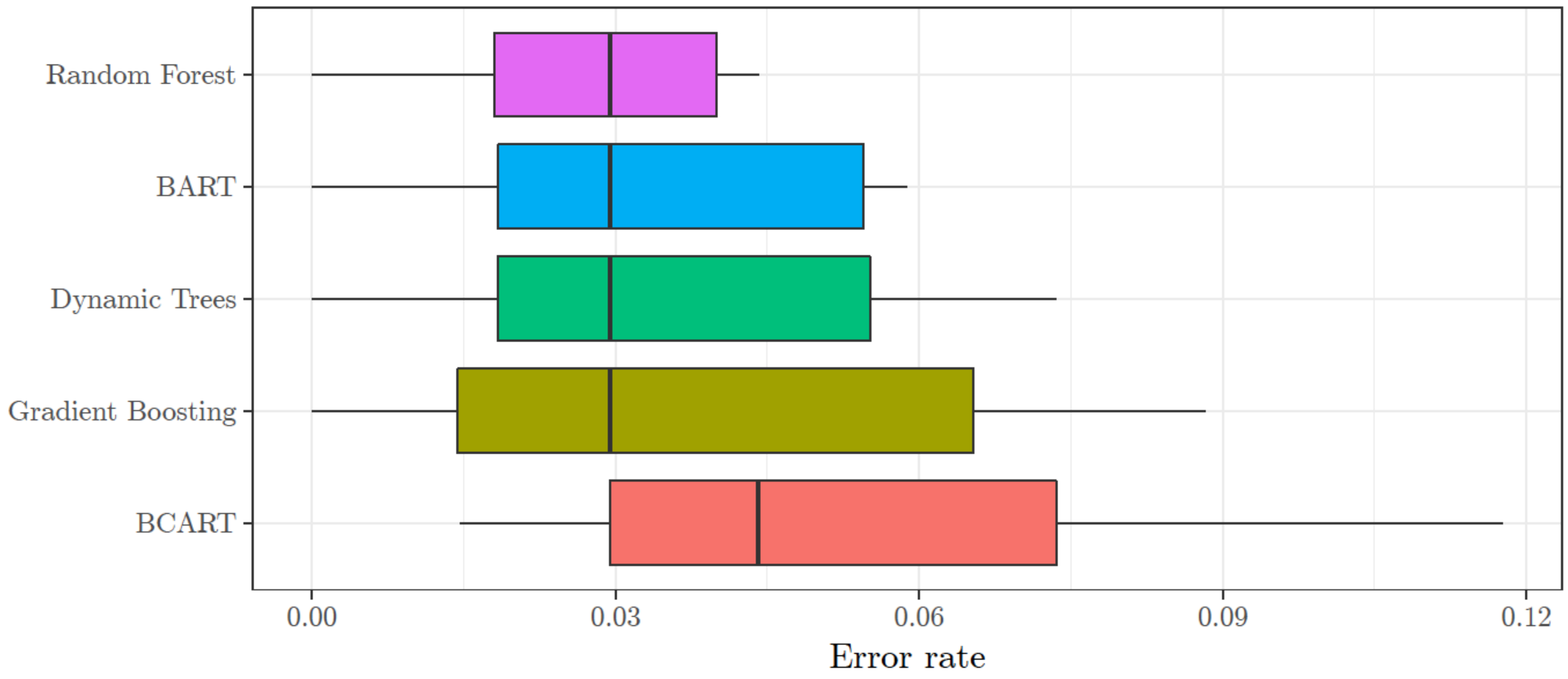

Figure 5: *Error rates for the competing methods on the Wisconsin breast cancer dataset for each fold in the cross-validation experiment. BART = Bayesian additive regression trees; BCART = Bayesian classification and regression trees.*

Focus Article

Bayesian treed response surface models

Hugh Chipman,[1] Edward I. George,[2] Robert B. Gramacy[3] and Robert McCulloch[3]*

Tree-based regression and classification, popularized in the 1980s with the advent of the classification and regression trees (CART) has seen a recent resurgence in popularity alongside a boom in modern computing power. The new methodologies take advantage of simulation-based inference, and ensemble methods, to produce higher fidelity response surfaces with competitive out-of-sample predictive performance while retaining many of the attractive features of classic trees: thrifty divide-and-conquer nonparametric inference, variable selection and sensitivity analysis, and nonstationary modeling features. In this paper, we review recent advances in Bayesian modeling for trees, from simple Bayesian CART models, treed Gaussian process, sequential inference via dynamic trees, to ensemble modeling via Bayesian additive regression trees (BART). We outline open source R packages supporting these methods and illustrate their use. © 2013 Wiley Periodicals, Inc.

How to cite this article:
WIREs Data Mining Knowl Discov 2013, 3: 298–305 doi: 10.1002/widm.1094

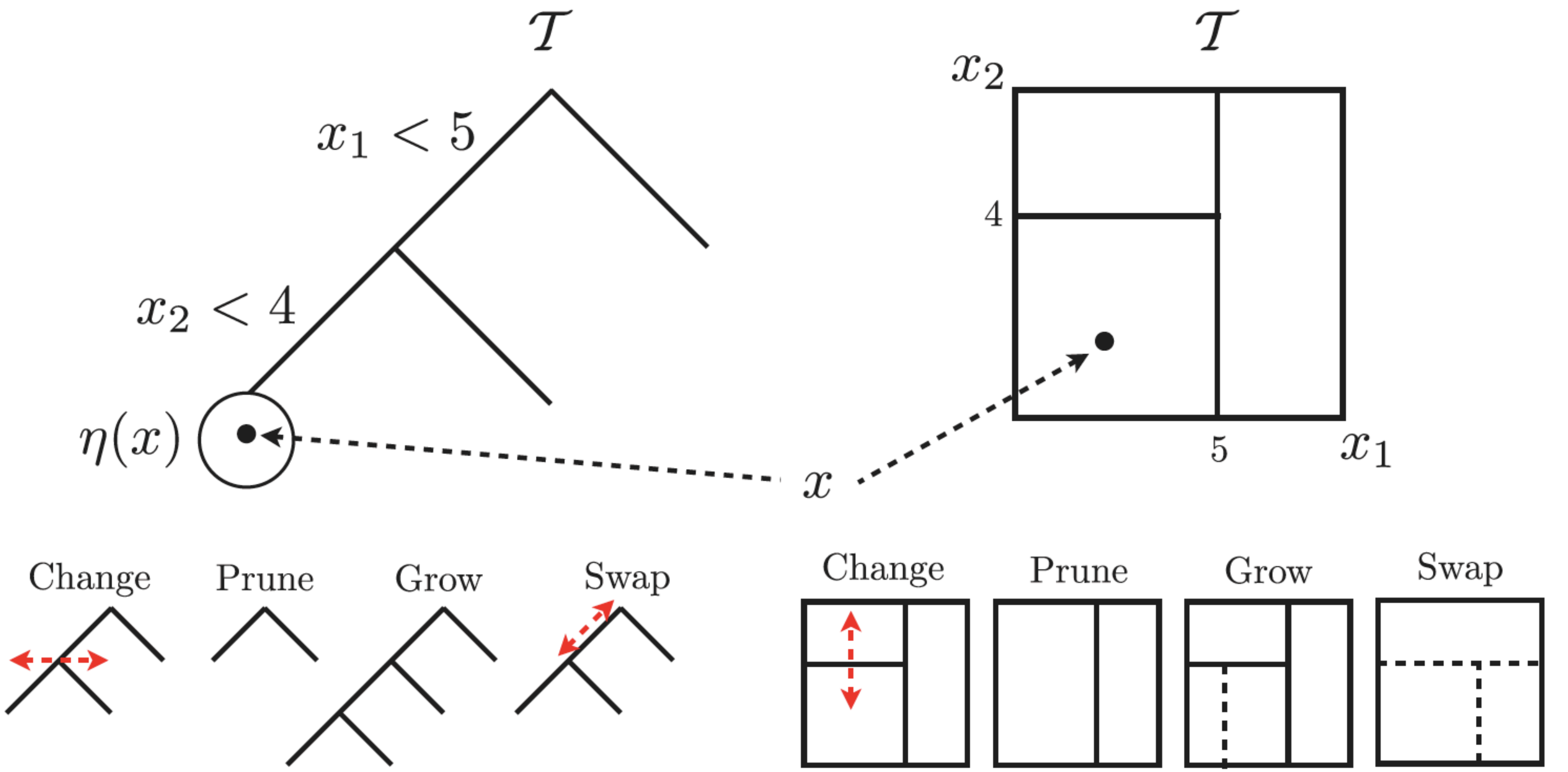

FIGURE 1 | Illustrating trees (top row) diagrammatically (left) and geographically (right). Each predictive location x falls in a leaf node $\eta(x)$. Tree operations (bottom row) show possible perturbations of trees that are possible steps in a stochastic search MCMC algorithm. x–y data.

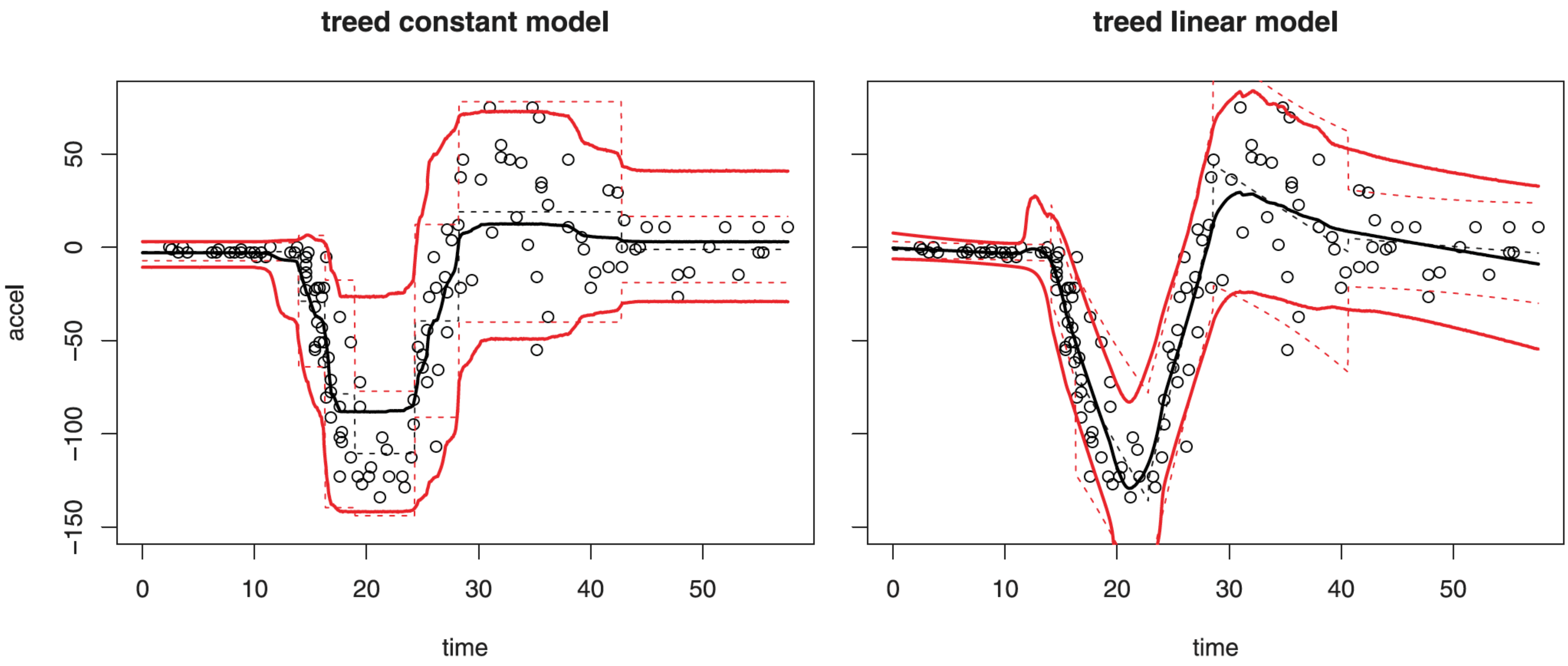

FIGURE 2 | Predictive surfaces for the treed constant model (*left*) and treed linear model (*right*); posterior mean in bold, and mode dashed.

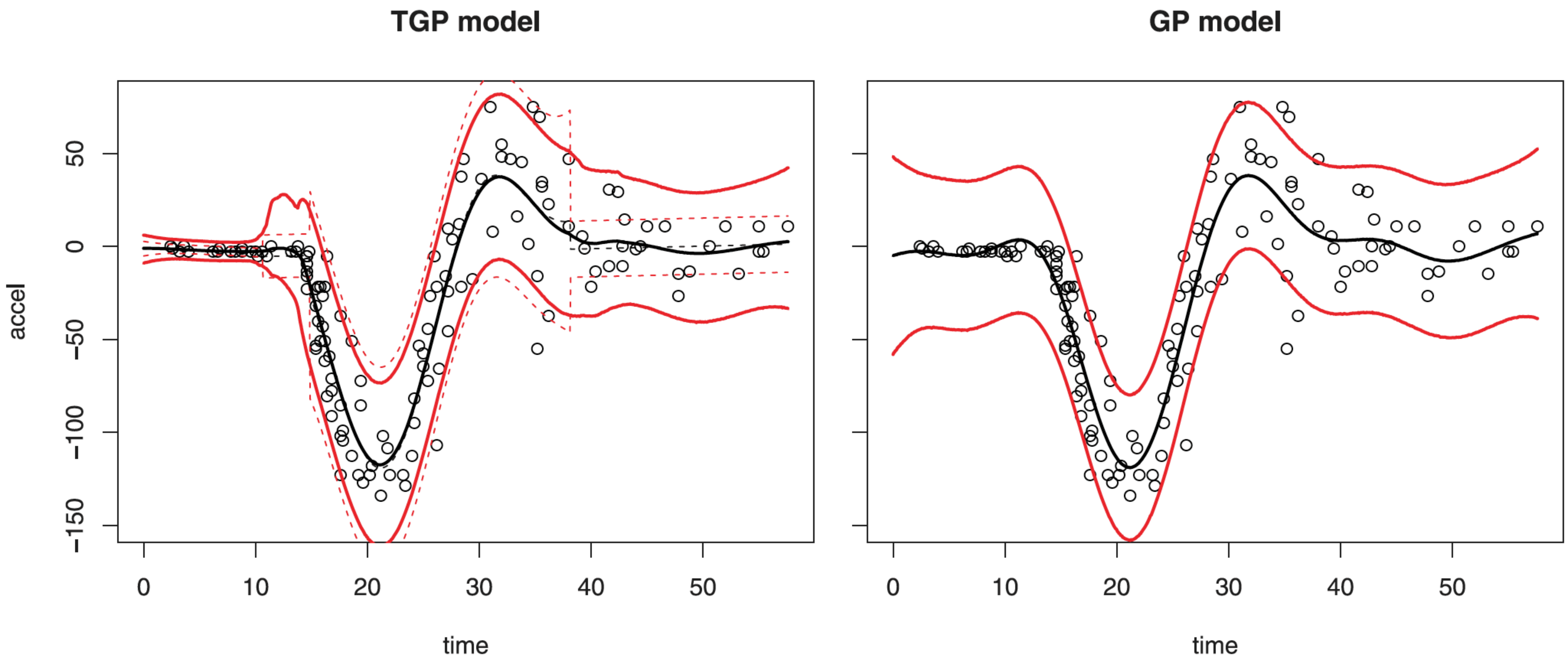

FIGURE 3 | Predictive surfaces for the treed GP model (left: posterior mean in bold, and mode dashed), and for the nontreed GP (right).

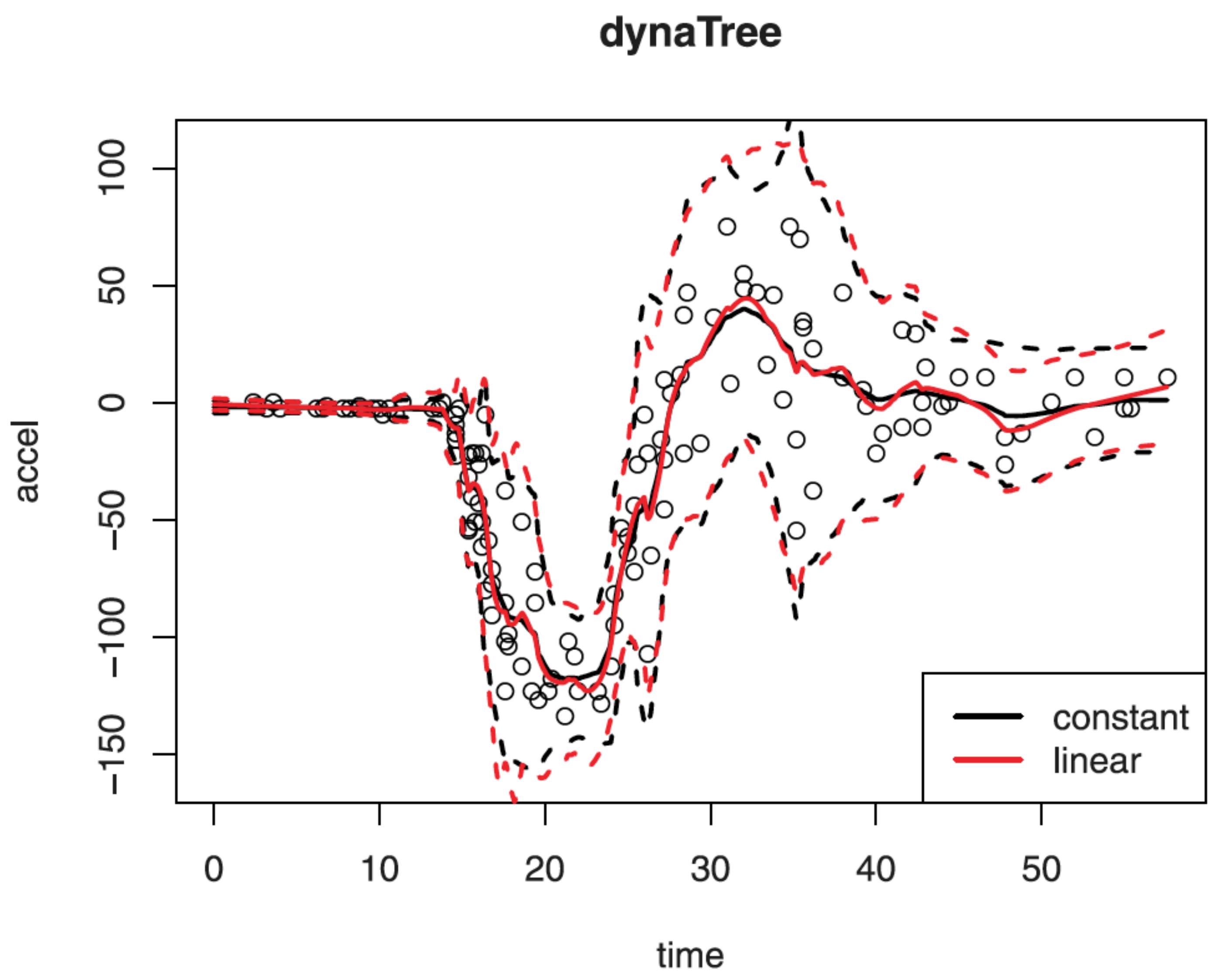

FIGURE 4 | Predictive surfaces for the DT models.

- tgp: for BCART and BTLM models, as well as BTGP and limiting linear model setups. Provides for sequential design via ALC and EI, and sensitivity analysis via Sobol indices. Multithreaded compilation is possible. Only regression is supported.

```
R> library(tgp)
R> library(MASS)
R> XX < − seq(0,max(mcycle[,1]), length=
  1000)
R> out.bcart < − bcart(X=mcycle[,1], Z=
  mcycle[,2], XX=XX)
R> out.btlm < − btlm(X=mcycle[,1], Z=
  mcycle[,2], XX=XX)
R> out.bgp < − bgp(X=mcycle[,1], Z=
  mcycle[,2], XX=XX)
R> out.btgp < − btgp(X=mcycle[,1], Z=
  mcycle[,2], XX=XX, bprior=”b0”)
```

- dynaTree: for dynamic treed regression and classification. Supports variable selection by relevance, and Sobol indices for sensitivity. Online inference is possible via datapoint retirement (e.g., with ALC) and forgetting factors for drifting concepts.

```
R> library(dynaTree)
R> out.dtc < − dynaTrees(X=mcycle[,1],
  y=mcycle[,2], XX=XX)
R> out.dtl < − dynaTrees(X=mcycle[,1],
  y=mcycle[,2], XX=XX, model=”linear”)
```

- BayesTree: for BART modeling of sum of trees regression and classification. Relevance indices also provided.

```
R> library(BayesTree)
R> bartfit < − bart(mcycle[,1],mcycle[,2])
```

NOTES

[a]... and when the practitioner is accustomed to the smooth fits GPs provide.