KAJIAN SIMULASI PERBANDINGAN METODE REGRESI KUADRAT TERKECIL PARSIAL, *SUPPORT VECTOR MACHINE*, DAN *RANDOM FOREST* *

A Andri Fauzi[1], Agus M Soleh[2‡], and Anik Djuraidah[3]

[1]Department of Statistics, IPB University, Indonesia, asepandrif@gmail.com
[2]Department of Statistics, IPB University, Indonesia, agusms@apps.ipb.ac.id
[3]Department of Statistics, IPB University, Indonesia, anikdjuraidah@gmail.com
‡corresponding author

Indonesian Journal of Statistics and Its Applications (eISSN:2599-0802)
Vol 4 No 1 (2020), 203 - 215

Abstract

Highly correlated predictors and nonlinear relationships between response and predictors potentially affected the performance of predictive modeling, especially when using the ordinary least square (OLS) method. The simple technique to solve this problem is by using another method such as Partial Least Square Regression (PLSR), Support Vector Regression with kernel Radial Basis Function (SVR-RBF), and Random Forest Regression (RFR). The purpose of this study is to compare OLS, PLSR, SVR-RBF, and RFR using simulation data. The methods were evaluated by the root mean square error prediction (RMSEP). The result showed that in the linear model, SVR-RBF and RFR have large RMSEP; OLS and PLSR are better than SVR-RBF and RFR, and PLSR provides much more stable prediction than OLS in case of highly correlated predictors and small sample size. In nonlinear data, RFR produced the smallest RMSEP when data contains high correlated predictors.

Keywords: highly correlated predictors, random forest regression, partial least square regression, support vector regression.

1. Pendahuluan

Prediktor berkorelasi tinggi serta hubungan nonlinier antara prediktor dan peubah respon dapat meningkatkan *mean square error* (MSE) yang dihasilkan dari Metode Kuadrat Terkecil (MKT). Hal tersebut banyak dikaji oleh peneliti dalam penelitian

* Received Jan 2020; Accepted Feb 2020; Published online on Feb 2020

empiris. Yeniay & Göktaş (2002) menunjukkan bahwa MKT menghasilkan MSE lebih tinggi daripada metode kuadrat terkecil parsial pada kasus prediktor berkorelasi tinggi. Adamowski *et al*. (2012) menunjukkan bahwa MKT menghasilkan MSE tinggi saat memodelkan peubah respon dan prediktor yang berhubungan nonlinier.

Solusi alternatif dari persoalan tersebut yaitu menggunakan metode pendugaan selain MKT. Penelitian menunjukkan bahwa metode pendugaan Kuadrat Terkecil Parsial (KTP), *Support Vector Machine* (SVM), dan *Random Forest* (RF) lebih baik daripada MKT dalam memodelkan data prediktor yang saling berkorelasi tinggi dan berhubungan nolinier dengan peubah respon. Farahani *et al*. (2010) menunjukkan Regresi Kuadrat Terkecil Parsial (RKTP) lebih baik daripada MKT dalam memodelkan data prediktor berkorelasi tinggi, khususnya jika ukuran contoh kecil. Liu *et al*. (2010) menunjukkan RKTP juga lebih baik daripada MKT dalam memodelkan peubah respon yang berhubungan nonlinier dengan prediktor.

SVM ditemukan oleh Vapnik (Vapnik *et al*., 1997). Pada mulanya SVM digunakan untuk memprediksi kategori atau kelas suatu amatan. Namun, SVM dikembangkan sehingga dapat digunakan untuk memprediksi kuantitas dari suatu amatan atau dikenal dengan pemodelan regresi (Vapnik, 2000). *Support Vector Rergression* dengan kernel *Radial Basis Function* (SVR-RBF) merupakan metode yang baik digunakan dalam pemodelan regresi nonlinier (Ma *et al*., 2018; Pour *et al*., 2016; Roy *et al*., 2019). Karimi *et al*. (2008) menunjukkan bahwa SVR-RBF dapat mengatasi masalah prediktor berkorelasi tinggi yang ada dalam suatu gugus data riil atau empiris. Namun, Dormann *et al*. (2013) menunjukkan bahwa SVR-RBF tidak dapat mengatasi masalah prediktor berkorelasi tinggi dengan menggunakan data simulasi.

RF yang digunakan untuk pemodelan regresi disebut juga dengan *Random Forest Regression* (RFR). Metode ini banyak digunakan dalam pemodelan prediktif (Steinwart & Christmann, 2008). RFR dapat mengatasi masalah prediktor berkorelasi tinggi karena dibuat berdasarkan *regression trees* dan teknik seleksi peubah. RFR juga dapat memodelkan prediktor yang berhubungan nonlinier dengan peubah respon karena RFR menggabungkan (*ensemble*) banyak *regression trees*. James *et al*. (2013) menunjukkan bahwa RFR baik digunakan dalam memodelkan data prediktor berkorelasi tinggi dan berhubungan nonlinier dengan respon.

Perbandingan MKT, RKTP, SVR-RBF, dan RFR telah banyak dikaji dalam berbagai penelitian (Farahani *et al*., 2010; Jing *et al*., 2016; Smith *et al*., 2013; Xu *et al*., 2019). Perbandingan tersebut dilakukan dengan menggunakan data empiris maupun simulasi. Hasil penelitan dengan data empiris hanya berlaku untuk kasus riil tertentu sedangkan data simulasi dapat digunakan untuk generalisasi.

Penelitian ini bertujuan membandingkan MKT, RKTP, SVR-RBF, dan RFR dalam pemodelan prediktif dengan menggunakan data simulasi. Kombinasi antara tingkat keeratan korelasi, jumlah prediktor, ukuran contoh, dan jenis model yang berbeda digunakan dalam pembangkitan data. Perbandingan metode menggunakan *root mean square error prediction* (RMSEP) yang diperoleh dari proses validasi silang *leave-one-out* (LOO).

2. Metodologi

2.1 Pembangkitan Data Simulasi

Data peubah prediktor dibangkitkan dari sebaran normal ganda dengan fungsi

kepekatan peluang:

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2\right),$$

dalam hal ini, $-\infty < x_i < \infty, i = 1,2,\dots,p$; $\boldsymbol{\mu} = \mathbf{0}_{1\times p}$; dan $\boldsymbol{\Sigma}$ adalah matriks peragam sedemikian sehingga setiap pasang prediktor berkorelasi ρ. Korelasi, jumlah prediktor (p),dan ukuran contoh (n) yang digunakan dalam pembangkitan data prediktor yaitu: $\rho = \{0.2, 0.5, 0.8\}, p = \{9, 16, 36\}$, dan $n = \{p+5, 2p+5, 5p+5\}$.

Data peubah respon dibangkitkan dari model linier dan nonlinier. Respon dari model linier dibangkitkan dengan persamaan:

$$\boldsymbol{y} = \beta_0\boldsymbol{x}_0 + \beta_1\boldsymbol{x}_1 + \cdots \beta_p\boldsymbol{x}_p + \boldsymbol{\varepsilon}, \quad (1)$$

sedangkan respon model nonlinier dibangkitkan dengan persamaan

$$\boldsymbol{y} = \log\left(\beta_0 \exp\left(\beta_1\boldsymbol{x}_1 + \cdots + \beta_p\boldsymbol{x}_p\right) + \boldsymbol{\varepsilon}\right). \quad (2)$$

dengan $\boldsymbol{y}$ adalah vektor respon, $\boldsymbol{x}_0$ adalah vektor berisi nilai 1 ($\mathbf{1}_{n\times 1}$), β_0 sampai β_p adalah parameter model regresi, dan $\boldsymbol{\varepsilon}$ adalah vektor galat acak yang menyebar normal. Nilai $\boldsymbol{\varepsilon}$ dibangkitkan dari sebaran $N(0,1)$, dan $\boldsymbol{\beta}$ diperoleh melalui percontohan acak dengan pemulihan dari $\beta = \{0.7, 1.0, 3\}$. Pembangkitan $\boldsymbol{y}$ pada model nonlinier efektif menghasilkan pola nonlinier antara prediktor dan respon tetapi ada potensi menghasilkan nilai tidak terdefinisi akibat log bilangan negatif. Jika respon yang dibangkitkan tidak terdefinisi maka diubah menjadi nol untuk mempertahankan ukuran contoh. Pembangkitan data simulasi menggunakan software *R* (R Core Team, 2019) dengan fungsi mvrnorm() dari paket MASS (Venables & Ripley, 2002). Skenario pembangkitan data simulasi dan penamaan gugus data disajikan pada Tabel 1.

Tabel 1: Informasi gugus data bangkitan.

Nama Gugus Data	Model	ρ	p	n
Dataset 1	Linier	0.2	9	14
…	…	…	…	…
Dataset 27	Linier	0.8	36	185
Dataset 1	Nonlinier	0.2	9	14
…	…	…	…	…
Dataset 27	Nonlinier	0.2	36	185

2.2 Metode Penelitian

a. Metode Kuadrat Terkecil

Model regresi linier dari suatu gugus data berukuran contoh n dapat ditulis sebagai berikut:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \quad (3)$$

$\boldsymbol{y}' = [y_1, y_2, \dots, y_n]$ adalah vektor peubah respon; $\boldsymbol{X} = [\mathbf{1}_{n\times 1}, \boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_p]$ adalah matriks yang berisi prediktor; $\boldsymbol{\beta}' = [\beta_0, \beta_1, \dots, \beta_p]$ adalah vektor parameter model; dan $\boldsymbol{\epsilon}' = [\epsilon_1, \epsilon_2, \dots, \epsilon_n]$ adalah galat acak (*random error*).

Dugaan parameter model, $\widehat{\boldsymbol{\beta}}$, diperoleh dengan cara meminimumkan jumlah kuadrat galat, *L*, yang dedefinisikan sebagai

$$L = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \quad (4)$$

Persamaan (4) dapat diuraikan menjadi

$$L = \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \quad (5)$$

karena $\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y}$ menghasilkan skalar dan $(\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta})' = \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y}$. Penduga MKT diperoleh dengan meminimumkan jumlah kuadrat galat sehingga harus memenuhi persamaan sebagai berikut:

$$\left.\frac{\partial L}{\partial \boldsymbol{\beta}}\right|_{\widehat{\boldsymbol{\beta}}} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{0}. \quad (6)$$

Persamaan (6) dapat disederhanakan menjadi

$$\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y} \quad (7)$$

Persamaan (7) merupakan persamaan normal dari MKT. Kedua ruas pada Persamaan (7) dikali $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ menjadi

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \quad (8)$$

karena $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}$. Dugaan peubah respon, $\widehat{\boldsymbol{y}}$, diperoleh dengan

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}. \quad (9)$$

b. RKTP

Gagasan utama dari RKTP yaitu mendekomposisi prediktor sedemikian sehingga antarkomponen tidak berkorelasi tetapi antara komponen dan peubah respon berkorelasi. Salah satu algoritme yang banyak digunakan adalah *Nonlinier Iterative Partial Least Square* (NIPALS). Algoritme ini mengasumsikan prediktor ($\boldsymbol{X}$) dan peubah respon ($\boldsymbol{Y}$) telah ditransformasi sehingga rata-ratanya nol. Penelitian ini menggunakan satu peubah respon ($\boldsymbol{y}$) dalam pemodelan sehingga proses lebih sederhana. Algortime NIPALS untuk RKTP dengan satu respon adalah sebagai berikut (Geladi & Kowalski, 1986).

1) $\boldsymbol{u} = \boldsymbol{y}$
2) $\boldsymbol{w}' = \boldsymbol{u}'\boldsymbol{X}/\boldsymbol{u}'\boldsymbol{u}$
3) $\boldsymbol{w}' = \boldsymbol{w}'/||\boldsymbol{w}'||$
4) $\boldsymbol{t} = \boldsymbol{X}\boldsymbol{w}/\boldsymbol{w}'\boldsymbol{w}$
5) $\boldsymbol{p}' = \boldsymbol{t}'\boldsymbol{X}/\boldsymbol{t}'\boldsymbol{t}$
6) $\boldsymbol{p}' = \boldsymbol{p}'/||\boldsymbol{p}'||$
7) $\boldsymbol{t} = \boldsymbol{t}\big|\big|\boldsymbol{p}'\big|\big|$
8) $\boldsymbol{w}' = \boldsymbol{w}'\big|\big|\boldsymbol{p}'\big|\big|$
9) $b = \boldsymbol{u}'\boldsymbol{t}/\boldsymbol{t}'\boldsymbol{t}$
10) $\boldsymbol{E}_h = \boldsymbol{E}_{h-1} - \boldsymbol{t}_h\boldsymbol{p}_h'; \boldsymbol{X} = \boldsymbol{E}_0$

$\boldsymbol{E}_h, \boldsymbol{t}_h$, dan $\boldsymbol{p}_h$, secara berurutan, merupakan residu, skor, dan *loading* untuk komponen ke-h. Setelah mencapai langkah (10) proses kembali lagi ke langkah (1) untuk menghitung komponen berikutnya, dengan $\boldsymbol{X} = \boldsymbol{E}_h$. Prediksi peubah respon menggunakan $\hat{\boldsymbol{t}}$ yang diperoleh dari

$$\hat{\boldsymbol{t}}_h = \boldsymbol{E}_{h-1}\boldsymbol{w}_h, \quad (10)$$

dengan $\boldsymbol{E}_h = \boldsymbol{E}_{h-1} - \hat{\boldsymbol{t}}_h\boldsymbol{p}_h'$. RKTP dihitung dengan menggunakan fungsi plsr() dari paket pls (Mevik *et al*., 2019). Penentuan jumlah komponen berdasarkan RMSEP.

c. SVR-RBF

Misalkan data pemodelan $[(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_l, y_l)]$ dengan $\boldsymbol{x}_i$ dan y_i adalah prediktor dan respon amatan ke-*i*. SVR memanfaatkan algortime SVM untuk

mendapatkan fungsi regresi $f(\boldsymbol{x}) = \boldsymbol{w}'.\Phi(\boldsymbol{x}) + b$ terbaik dalam memprediksi respon dengan toleransi galat sebesar ε, yang mana $\boldsymbol{w}$ dan b adalah parameter sedangkan $\Phi(\boldsymbol{x})$ adalah fungsi nonlinier. Permasalahan SVR diformulasikan menjadi masalah optimasi sebagai berikut:

$$\min_{\boldsymbol{w}, b, \xi, \xi^*} \ \frac{1}{2}\boldsymbol{w}'.\boldsymbol{w} + C\sum_{i=1}^{l}(\xi_i + \xi_i^*), \tag{11}$$

dengan kendala
$$y_i - [\boldsymbol{w}'.\Phi(\boldsymbol{x}_i) + b] \le \varepsilon + \xi_i$$
$$[\boldsymbol{w}'.\Phi(\boldsymbol{x}_i) + b] - y_i \le \varepsilon + \xi_i^*$$
$$\xi_i, \xi_i^* \ge 0, i = 1,2,\ldots,l$$

ξ_i dan ξ_i^* merupakan *slack variables* yang menyatakan batas atas dan batas bawah galat dengan batas toleransi ε, dan C adalah konstanta positif sebagai penentu derajat *penalized loss* saat ada galat. SVM menggunakan *penalized loss* Vapnik's ε-*insensitive loss function* (Xu *et al*., 2019).

Permasalahan optimasi tersebut dapat diselesaikan dengan dua gugus Lagrange *multipliers*, α_i dan α_i^*. Fungsi pendekatannya dapat diformulasikan:

$$f(\boldsymbol{x}) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\,\Phi(\boldsymbol{x}_i)'.\Phi(\boldsymbol{x}) + b. \tag{12}$$

Amatan yang memiliki Lagrange *multipliers* $(\alpha_i - \alpha_i^*)$ bernilai tidak nol digunakan sebagai *support vector* dalam fungsi regresi.

Penelitian ini menggunakan SVR dengan fungsi kernel RBF. Fungsi kernel merubah Persamaan 12 menjadi persamaan sebagai berikut:

$$f(\boldsymbol{x}) = \sum_{k=1}^{m}(\alpha_k - \alpha_k^*)\,\mathrm{K}(\boldsymbol{x}_\mathrm{k}, \boldsymbol{x}) + b, \tag{13}$$

dengan $\boldsymbol{x}_\mathrm{k}$ adalah *support vector* ke-*k*, m adalah banyaknya *support vectors*, dan $\mathrm{K}(\boldsymbol{x}_\mathrm{k}, \boldsymbol{x})$ adalah fungsi kernel. Fungsi kernel RBF diformulasikan sebagai berikut: $\mathrm{K}(\boldsymbol{x}_\mathrm{k}, \boldsymbol{x}) = \exp(-\gamma|\boldsymbol{x} - \boldsymbol{x}_i|^2)$. SVR dihitung dengan fungsi ksvm() dari paket kernlab (Karatzoglou *et al*., 2004). Pemilihan parameter C dilakukan dengan metode *grid search* sedangkan parameter lain menggunakan nilai *default* dari fungsi ksvm().

d. RFR

Metode ini memadukan percontohan ulang *bootstrap* dan seleksi peubah untuk mengurangi ragam galat prediksi akibat masalah prediktor berkorelasi tinggi dan meningkatkan ketepatan prediksi peubah respon dari metode regresi pohon. RFR membangun banyak pohon regresi lalu menghitung rata-rata hasil prediksi peubah respon dari semua pohon regresi tersebut.

Gugus data asli yang terdiri dari *p* vektor prediktor $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_p)$ dan satu vektor peubah respon $(\boldsymbol{y})$ dinotasikan dengan $\boldsymbol{Z}$, gugus data baru yang dibuat dengan menggunakan percontohan ulang *bootstrap* dari $\boldsymbol{Z}$ dinotasikan dengan $\boldsymbol{Z}^*$, dan fungsi pohon regresi pada gugus data hasil *bootraping* ke-*b* dinotasikan dengan $T_b(\boldsymbol{x}^*)$. Algoritme RFR, untuk memprediksi peubah respon dari percontohan ulang yang dilakukan sebanyak B adalah sebagai berikut (Hastie *et al*., 2009; Liaw *et al*., 2002):

1) Pada percontohan ulang *boostrap* ke b, dengan $b = 1,2,\ldots,B$, lakukan:
   (a) buat gugus data $\boldsymbol{Z}^*$ dari gugus asli menggunakan metode percontohan

ulang *bootstrap*, dengan $\boldsymbol{Z}^*$ adalah gugus data yang terdiri dari $p/3$ prediktor ($\boldsymbol{x}^*$) dan *n* amatan (jika $p/3$ menghasilkan nilai desimal maka dibulatkan ke atas);

(b) buat pohon *random forest*, $T_b(\boldsymbol{x}^*)$, dari $\boldsymbol{Z}^*$ menggunakan metode pohon regresi.

Prediksi peubah respon dari metode RFR dengan B kali percontohan ulang *bootstrap*, $\hat{f}_{rfr}^{B}(x)$, diperoleh dengan $\hat{f}_{rfr}^{B}(x) = (1/B)\sum_{b=1}^{B} T_b(\boldsymbol{x}^*)$

e. Evaluasi Metode

Evaluasi metode menggunakan ukuran *Root Mean Square Error Prediction* (RMSEP) yang diperoleh dengan teknik validasi silang *Leave-One-Out* (LOO). Proses validasi silang diulang sebanyak 500 kali sehingga diperoleh 500 nilai RMSEP dari setiap metode pemodelan. Validasi silang LOO menjadikan setiap amatan sebagai data uji secara bergantian sehingga dari n amatan menghasilkan n prediksi $\hat{f}_i(x_i)$ yang diperoleh dari n persamaan $\hat{f}_i$. $RMSEP$ dihitung dengan

$$RMSEP = \sqrt{(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2)/n}\,;\; i = 1,2,\ldots,n \tag{14}$$

dengan y_i dan $\hat{y}_i$ adalah nilai aktual dan nilai prediksi peubah respon amatan ke-i.

3. Hasil dan Pembahasan

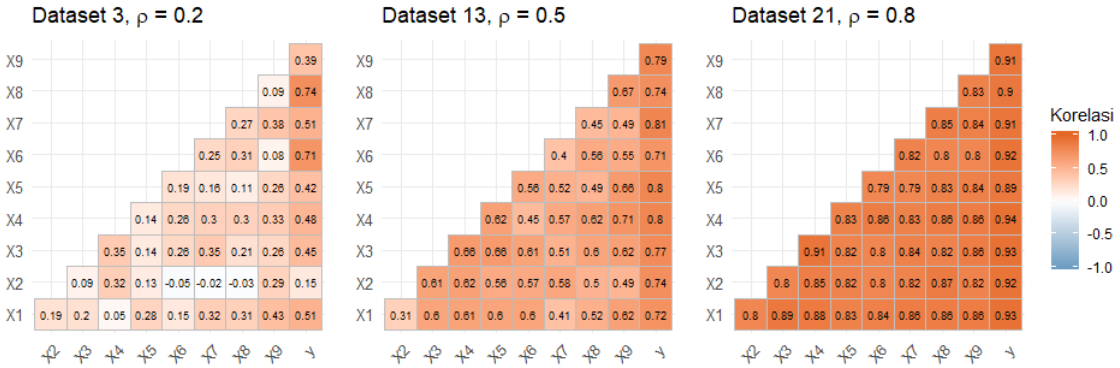
3.1 Deskripsi Data

Gambar 1 menyajikan struktur korelasi Pearson pada sebagian gugus data yang dibangkitkan. Deskripsi ini bertujuan memberikan gambaran korelasi antarprediktor dan korelasi antara prediktor dengan peubah respon. Sebagian gugus data ini telah dapat mewakili gugus data lain.

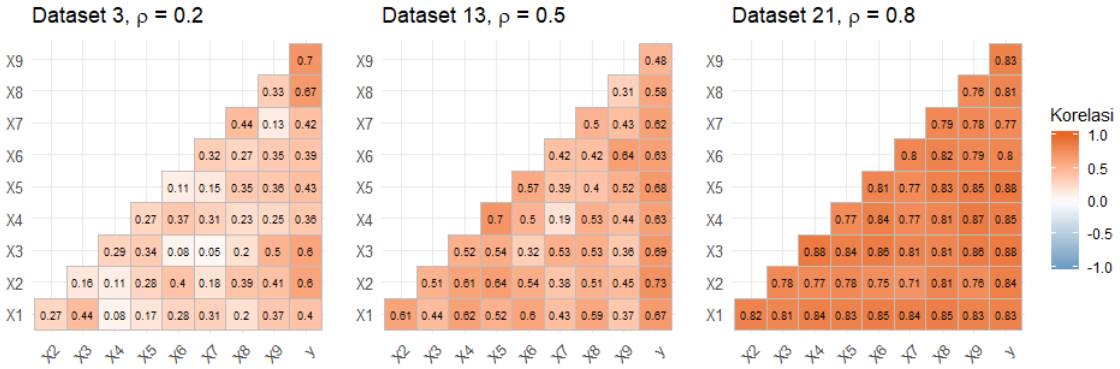
Dari Gambar 1 diketahui korelasi antarprediktor mendekati nilai korelasi (ρ) yang ditentukan. Korelasi antara prediktor dan respon data yang dibangkitkan dari model nonlinier mendekati model linier. Hal ini mengindikasikan hubungan antara prediktor dan respon masih mendekati linier. Meskipun demikian, korelasi antara prediktor dan respon pada model linier mayoritas lebih tinggi daripada model nonlinier.

3.2 Perbandingan Kinerja Metode dalam Pemodelan Prediktif Linier

Pemodelan prediktif linier dalam penelitian ini berarti pemodelan menggunakan prediktor yang berhubungan linier dengan peubah respon. Perbedaan keragaman RMSEP antara satu metode dan yang lain terlalu besar sehingga perbandingan metode menggunakan log(RMSEP). Gambar 2 menyajikan log(RMSEP) untuk data yang dibangkitkan dengan korelasi 0.2 atau korelasi rendah.

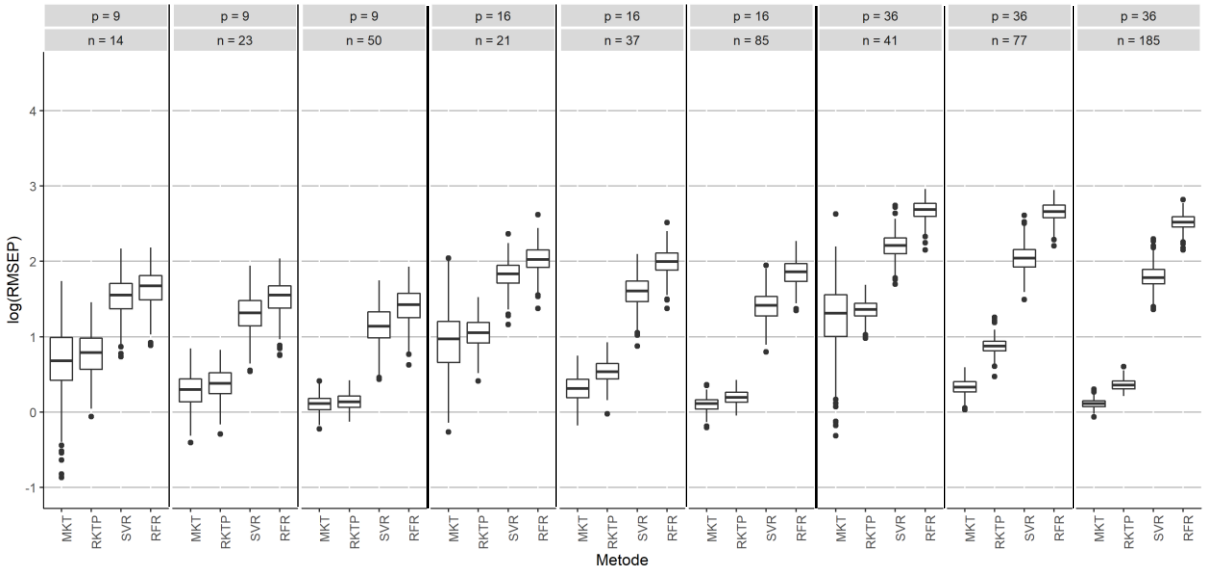


(a)



(b)

Gambar 1: Struktur korelasi dari (a) model linier, (b) model nonlinier.

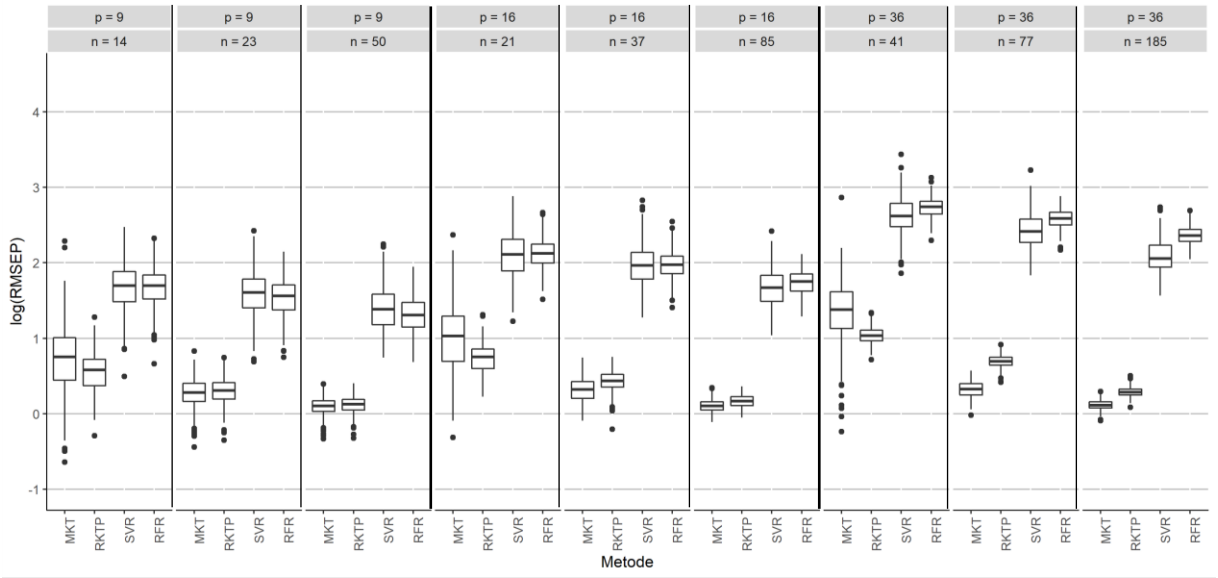


Gambar 2: Perbandingan metode saat model linier dan korelasi = 0.2.

Dari Gambar 2 diketahui median log(RMSEP) SVR-RBF dan RFR selalu lebih tinggi daripada metode lain. Keragaman log(RMSEP) dari dua metode tersebut juga lebih tinggi daripada MKT dan RKTP pada saat ukuran contoh $2p + 5$ dan $5p + 5$. Hal ini menunjukkan SVR-RBF dan RFR tidak cocok untuk memodelkan peubah respon yang berhubungan linier dengan prediktor. Median log(RMSEP) SVR-RBF selalu lebih kecil daripada RFR tetapi keragamannya relatif sama. Median log(RMSEP) MKT pada

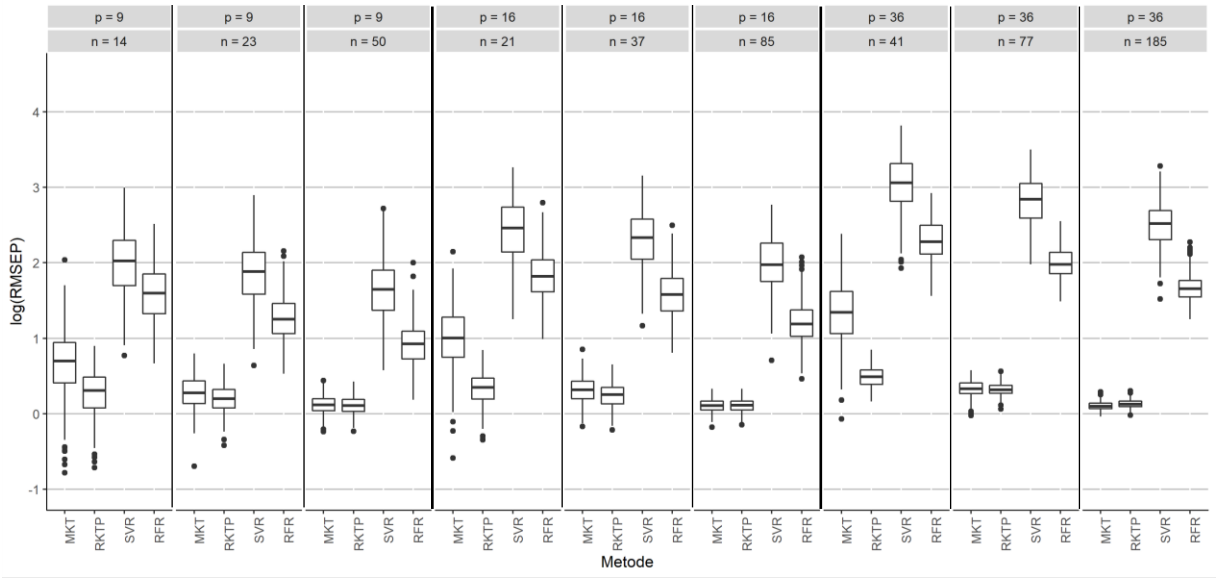
ukuran contoh $p + 5$ lebih rendah tetapi ragamnya lebih tinggi dari RKTP. Seiring bertambahnya ukuran contoh, median log(RMSEP) MKT dan RKTP semakin kecil. Keragaman log(RMSEP) MKT semakin kecil saat ukuran contoh semakin besar.

Gambar 3 menyajikan perbandingan metode saat korelasi 0.5 atau korelasi sedang. Karakteristik perbedaan SVR-RBF dan RFR dengan MKT dan RKTP saat korelasi 0.5 mirip dengan karakteristik perbedaan SVR-RBF dan RFR dengan MKT dan RKTP saat korelasi saat korelasi 0.2. Hal ini menunjukkan bahwa metode SVR-RBF dan RFR tidak cocok untuk memodelkan peubah respon yang berhubungan linier dengan prediktor. Namun, perbedaan antara SVR-RBF dan RFR semakin kecil. Keragaman log(RMSEP) SVR-RBF lebih besar daripada RFR. Ini mengindikasikan bahwa korelasi sedang antarprediktor mempengaruhi SVR-RBF sedangkan RFR tidak terpengaruh. RKTP menghasilkan log(RMSEP) yang lebih kecil dan lebih homogen daripada MKT saat ukuran contoh kecil. Pada ukuran contoh besar MKT lebih baik daripada RKTP.



Gambar 3: Perbandingan metode saat model linier dan korelasi = 0.5.

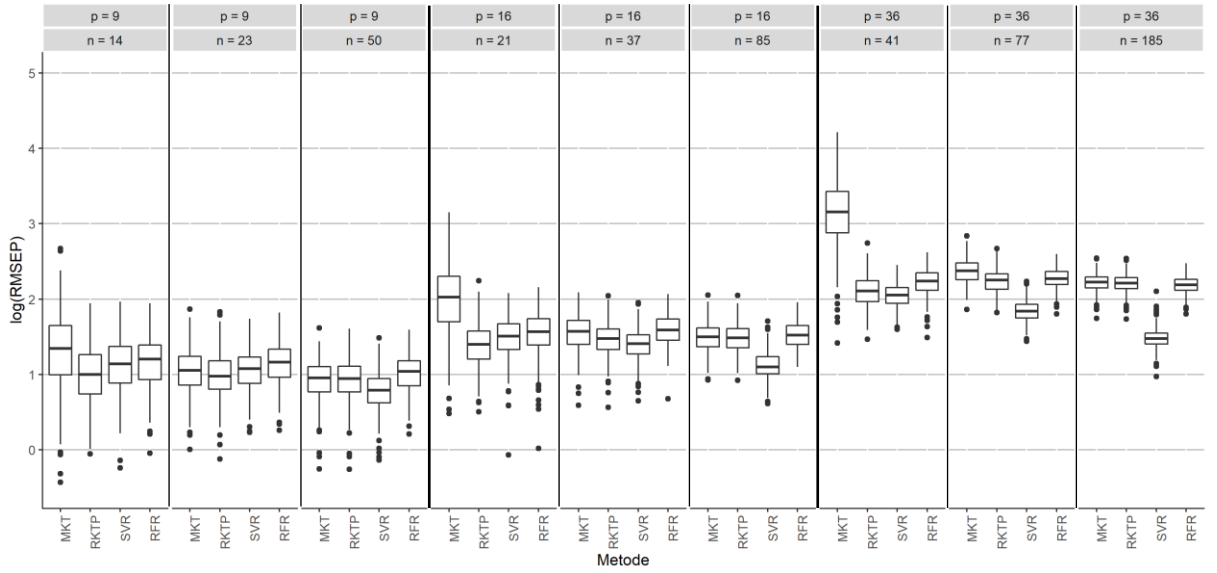
Gambar 4 menyajikan perbandingan metode saat korelasi 0.8 atau korelasi tinggi. SVR-RBF menghasilkan median log(RSEMP) lebih tinggi daripada metode lain. Keragaman log(RMSEP) yang dihasilkan RFR relatif lebih kecil daripada SVR-RBF. Hal ini menunjukkan bahwa SVR-RBF tidak dapat mengatasi masalah prediktor berkorelasi tinggi sedangkan RFR dapat mengatasinya. Keragaman log(RMSEP) MKT saat prediktor berkorelasi tinggi lebih besar daripada korelasi sedang dan rendah. Median dan keragaman log(RMSEP) RKTP pada setiap ukuran contoh tidak berubah secara signifikan sedangkan MKT menghasilkan median dan ragam log(RMSEP) yang semakin tinggi seiiring dengan berkurangnya ukuran contoh. Hal ini menunjukkan bahwa metode RKTP lebih stabil daripada MKT saat prediktor berkorelasi tinggi pada setiap ukuran contoh. Berdasarkan uraian tersebut dapat dikatakan bahwa metode RKTP merupakan metode paling baik untuk memodelkan prediktor berkorelasi tinggi dan berhubungan linier dengan peubah respon.



Gambar 4: Perbandingan metode saat model linier dan korelasi = 0.8.

3.3 Perbandingan Kinerja Metode dalam Pemodelan Prediktif Nonlinier

Pemodelan prediktif nonlinier dalam penelitian ini berarti pemodelan menggunakan prediktor yang berhubungan nonlinier dengan respon. RMSEP pada pemodelan ini juga sangat beragam sehingga perbandingan dilakukan dengan menggunakan log(RMSEP). Gambar 5 menyajikan log(RMSEP) untuk data yang dibangkitkan dengan korelasi 0.2 atau korelasi rendah.

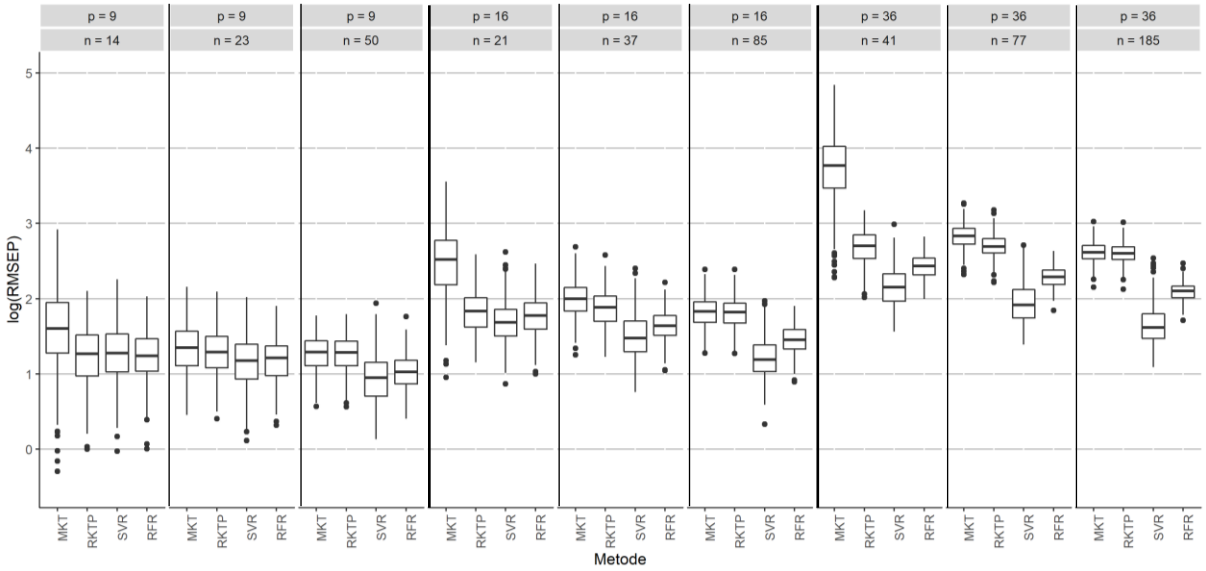


Gambar 5: Perbandingan metode saat model nonlinier dan korelasi = 0.2.

Gambar 5 menunjukkan pada saat jumlah prediktor 9 kinerja metode tidak terlalu jauh berbeda. MKT menghasilkan keragaman log(RMSEP) tertinggi pada ukuran contoh kecil. RKTP, SVR-RBF, dan RFR memiliki kinerja yang tidak terlalu jauh berbeda pada ukuran contoh kecil. SVR-RBF mayoritas menghasilkan log(RMSEP) terkecil. Keragaman yang dihasilkannya tidak terlalu berbeda dengan metode lain. SVR-RBF yang digunakan pada data dengan korelasi 0.2, 36 prediktor dan 185 amatan jauh lebih baik daripada metode lain. Hal ini mengindikasikan bawah metode

SVR-RBF baik digunakan untuk memodelkan peubah respon yang berhubungan nonlinier dengan prediktor yang tidak berkorelasi tinggi. Perbedaan hasil tidak terlalu berbeda antara satu metode dengan metode lain disebabkan pembangkitan data. Seperti telah dijelaskan pada bagian deskripsi data, pola hubungan nonlinier antara peubah respon dan prediktor masih mendekati linier.

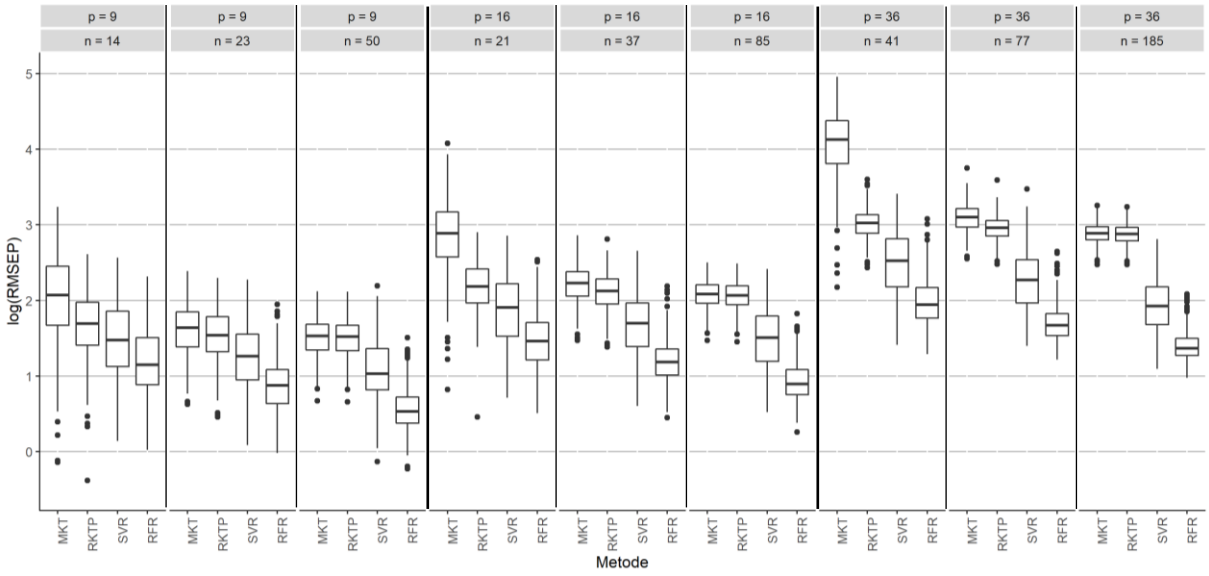
Gambar 6 menyajikan log(RMSEP) untuk data yang dibangkitkan dengan korelasi 0.5 atau korelasi sedang. Berdasarkan Gambar 7 diketahui bahwa median dan keragaman log(RMSEP) dari MKT lebih besar daripada metode lain saat ukuran contoh kecil. SVR-RBF menghasilkan median log(RMSEP) terkecil saat jumlah prediktor 16 dan 36 tetapi keragamannya relatif lebih besar daripada metode lain. Keragaman log(RMSEP) SVR-RBF pada gugus data dengan korelasi 0.5, jumlah prediktor 36, dan ukuran contoh 185 lebih besar tetapi mediannya lebih kecil daripada metode lain. Pada mayoritas gugus data, SVR-RBF dan RFR menghasilkan median dan keragaman log(RMSEP) lebih kecil daripada daripada MKT dan RKTP. Meskipun metode SVR-RBF menghasilkan median log(RMSEP) lebih kecil daripada RFR tetapi keragamannya relatif lebih besar. Hal ini mengindikasikan bahwa metode SVR-RBF terpengaruh oleh korelasi sedang sedangkan metode RFR tidak terpengaruh.



Gambar 6: Perbandingan metode saat model nonlinier dan korelasi = 0.5.

Gambar 6 menyajikan log(RMSEP) untuk data yang dibangkitkan dengan korelasi 0.8 atau korelasi tinggi. Pola pada Gambar 7 berbeda signifikan dari Gambar 5 dan 6. Pada kasus peubah respon berhubungan nonlinier dengan prediktor yang saling berkorelasi tinggi, RKTP menghasilkan median dan keragaman log(RMSEP) lebih kecil daripada MKT. Hal ini menunjukkan bahwa RKPT lebih baik daripada MKT dalam pemodelan nonlinier dan kasus prediktor berkorelasi tinggi.

Berdasarkan Gambar 7 juga diketahui RFR menghasilkan median log(RMSEP) terkecil pada semua gugus data. Keragamannya pun semakin kecil seiring dengan bertambahnya ukuran contoh. SVR-RBF menghasilkan median log(RMSEP) lebih rendah daripada MKT dan RKTP tetapi keragamannya relatif tinggi. Hal ini menunjukkan bahwa SVR-RBF dan RFR lebih baik daripada MKT dan RKTP dalam memodelkan peubah respon yang berhubungan nonlinier dengan prediktor. Selain itu, Gambar 7 juga menunjukkan bahwa metode RFR dapat mengatasi masalah prediktor berkorelasi tinggi sedangkan SVR-RBF tidak dapat mengatasinya.



Gambar 7: Perbandingan metode saat model nonlinier dan korelasi = 0.8.

4. Simpulan

Pada kasus prediktor berkorelasi tinggi dan peubah respon berhubungan linier dengan prediktor, RKTP lebih baik daripada MKT terutama pada saat ukuran contoh kecil. Pada kasus prediktor berkorelasi tinggi dan peubah respon berhubungan nonlinier dengan prediktor, RKTP lebih baik daripada MKT serta RFR lebih baik daripada metode MKT, RKTP, dan SVR-RBF.

Daftar Pustaka

Adamowski, J., Fung Chan, H., Prasher, S. O., Ozga-Zielinski, B., & Sliusarieva, A. (2012). Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. *Water Resources Research*, *48*(1).

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., … Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1): 27–46.

Farahani, H. A., Rahiminezhad, A., Same, L., & others. (2010). A Comparison of Partial Least Squares (PLS) and Ordinary Least Squares (OLS) regressions in predicting of couples mental health based on their communicational patterns. *Procedia-Social and Behavioral Sciences*, *5*: 1459–1463.

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, *185*: 1–17.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. New York (US): Springer Science & Business Media.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York (US): Springer.

Jing, W., Yang, Y., Yue, X., & Zhao, X. (2016). A spatial downscaling algorithm for satellite-based precipitation over the Tibetan plateau based on NDVI, DEM, and land surface temperature. *Remote Sensing*, *8*(8): 1–19.

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab-an S4 package for kernel methods in R. *Journal of Statistical Software*, *11*(9): 1–20.

Karimi, Y., Prasher, S., Madani, A., Kim, S., & others. (2008). Application of support vector machine technology for the estimation of crop biophysical parameters using aerial hyperspectral observations. *Canadian Biosystems Engineering*, *50*(7): 13–20.

Liaw, A., Wiener, M., & others. (2002). Classification and regression by randomForest. *R News*, *2*(3): 18–22.

Liu, Y., Sun, X., Zhou, J., Zhang, H., & Yang, C. (2010). Linear and nonlinear multivariate regressions for determination sugar content of intact Gannan navel orange by Vis–NIR diffuse reflectance spectroscopy. *Mathematical and Computer Modelling*, *51*(11–12): 1438–1443.

Ma, X., Zhang, Y., Cao, H., Zhang, S., & Zhou, Y. (2018). Nonlinear regression with high-dimensional space mapping for blood component spectral quantitative analysis. *Journal of Spectroscopy*, *2018*.

Mevik, B.-H., Wehrens, R., & Liland, K. H. (2019). pls: Partial least squares and principal component regression. *R Package Version*.

Pour, S. H., Shahid, S., & Chung, E.-S. (2016). A hybrid model for statistical downscaling of daily rainfall. *Procedia Engineering*, *154*: 1424–1430.

[R Core Team]. (2019). *R: A language and environment for statistical computing*. Vienna (AT): R Foundation for Statistical Computing.

Roy, A., Manna, R., & Chakraborty, S. (2019). Support vector regression based metamodeling for structural reliability analysis. *Probabilistic Engineering Mechanics*, *55*: 78–89.

Smith, P. F., Ganesh, S., & Liu, P. (2013). A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *Journal of Neuroscience Methods*, *220*(1): 85–91.

Steinwart, I., & Christmann, A. (2008). *Support Vector Machines*. New York (US): Springer Science & Business Media.

Vapnik, V. (2000). *The Nature of Statistical Learning Theory*. New York (US): Springer Science & Business Media.

Vapnik, V., Golowich, S. E., & Smola, A. J. (1997). Support vector method for function approximation, regression estimation and signal processing. *Advances in Neural Information Processing Systems*, 281–287.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. New York (US): Springer.

Xu, R., Chen, Y., & Chen, Z. (2019). Future Changes of Precipitation over the Han River Basin Using NEX-GDDP Dataset and the SVR_QM Method. *Atmosphere*, *10*(11): 688.

Yeniay, Ö., & Göktaş, A. (2002). A comparison of partial least squares regression with other prediction methods. *Hacettepe Journal of Mathematics and Statistics*, *31*: 99–111.