Hindawi
Complexity
Volume 2020, Article ID 3567894, 13 pages
https://doi.org/10.1155/2020/3567894

Research Article

Prediction Error and Forecasting Interval Analysis of Decision Trees with an Application in Renewable Energy Supply Forecasting

Xin Zhao [1] **and Xiaokai Nie** [2]

[1]*School of Mathematics, Southeast University, Nanjing 211189, China*
[2]*School of Automation, Southeast University, Nanjing 210096, China*

Correspondence should be addressed to Xiaokai Nie; xiaokai.nie@hotmail.com

Received 10 July 2020; Revised 25 August 2020; Accepted 13 September 2020; Published 26 October 2020

Academic Editor: Shubo Wang

Renewable energy has become popular compared with traditional energy like coal. The relative demand for renewable energy compared to traditional energy is an important index to determine the energy supply structure. Forecasting the relative demand index has become quite essential. Data mining methods like decision trees are quite effective in such time series forecasting, but theory behind them is rarely discussed in research. In this paper, some theories are explored about decision trees including the behavior of bias, variance, and squared prediction error using trees and the prediction interval analysis. After that, real UK grid data are used in interval forecasting application. In the renewable energy ratio forecasting application, the ratio of renewable energy supply over that of traditional energy can be dynamically forecasted with an interval coverage accuracy higher than 80% and a small width around 22, which is similar to its standard deviation.

1. Introduction

Renewable energy such as solar and wind has been playing an integral role in sustaining power supply and relieving the environment pollution and global warming crisis. With the increasing penetration of renewable energy, determining the amounts of renewable energy generation is critical to maintain the energy balance and the stability and reliability of power networks. Forecasting the mixing shares of the energy generation offers the guidance of setting up the power generation for each energy source and ensures the load demand of power networks to be satisfied [1, 2]. Data-based prediction methods, in particular machine learning methods, provide a promising solution to infer the required ratios of energy generation, among which decision tree is a well-recognized approach due to its satisfactory accuracy and interpretation [3–6].

Although decision tree provides an effective method in forecasting, the theory explaining when and how it performs well is rarely discussed. The required ratios of renewable energy generation can be seen as a linear time series. In this context, we explore how the tree model performs in terms of the bias, variance, and prediction error. In addition, point prediction is not sufficient in time series prediction, so we also provide prediction interval choices like Gaussian and quantile intervals in theories with the application in renewable energy ratio forecasting.

Decision tree [7] is a nonparametric supervised learning method used for discovery and prediction-oriented classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Decision tree, compared to other data mining methods, has its own advantages. (1) For casual relationship, it can deal with nonlinear models. In most cases, economics pay more attention to linear models, while if it is a nonlinear model, it will be transferred to be a linear model. In problems like consumer behavior analysis, the number of variables exceeds the normal extent to be tens or even hundreds, which will definitely lead to high correlation among variables. In that case, coefficients may have the wrong meaning in reality. Decision tree, however, provides variable importance

ranking criteria, which helps a lot. (2) In terms of comprehensibility, it tends to be relatively better than "black-box" models like neural network, which means it can interpret data structure more clearly and help readers understand the information involved. These undoubtedly bring convenience to decision making in medical treatment [8–10], e-commerce, [11–13] and so on.

We now explore the performance of trees when fitted to data generated from a linear model. The corresponding bias, variance, and prediction error between the fitted simplified tree and the true simple linear model will be calculated. Then, how those errors vary will be explored when the linear data distribution changes. The motivation is to explore how the trees perform under different distributions. Afterwards, prediction interval is proposed using Gaussian and quantile intervals, which explains why quantile interval is chosen in the study by Zhao et al. [14]. The simple linear model in use is

$$Y = \alpha + \beta X + \varepsilon, \tag{1}$$

where $f(X) = \alpha + \beta X$ is the true model. It is supposed that, throughout this paper, $X \sim U(a, b)$ independently and $\varepsilon \sim N(0, \sigma^2)$. Uniform distribution guarantees that if the tree has k terminal nodes, the sample size in each node will be equal which is convenient in theory and simulation analysis. Decision tree analysis under the uniform distribution assumption includes the work by Hancock [15], Jackson and Servedio [16], and White and Liu [17]. Other distributions can also be considered but the analysis will be much more complex as the sample size of each terminal node depends on many parameters.

The expected squared prediction error (SPE) is one of the important metrics to measure how well the trained model is applied to further unseen data. As shown in Hastie et al. [18], SPE of a regression fit $\widehat{f}(X)$ at an input point $X = x_0$ is

$$\begin{aligned}
\mathrm{SPE}(x_0) &= E\left[\left(Y - \widehat{f}(x_0)\right)^2 \mid X = x_0\right] \\
&= \sigma^2 + \left[E\widehat{f}(x_0) - f(x_0)\right]^2 + E\left[\widehat{f}(x_0) - E\widehat{f}(x_0)\right]^2 \\
&= \sigma^2 + \mathrm{Bias}^2\left(\widehat{f}(x_0)\right) + \mathrm{Var}\left(\widehat{f}(x_0)\right) \\
&= \text{irreducible error} + \text{bias}^2 + \text{variance}.
\end{aligned} \tag{2}$$

In (2), the first term is the variance of the target around its true mean $f(x_0)$ and cannot be avoided no matter how well the $f(x_0)$ is estimated, unless $\sigma^2 = 0$. The second term is the squared bias, the amount by which the average of the estimate differs from the true mean; the last term is the variance, the expected squared deviation of $\widehat{f}(x_0)$ around its mean. Typically the more complex the model $\widehat{f}$ is, the lower the (squared) bias but the higher the variance [18] will be.

In Section 2, the performance of regression trees is analyzed when fitted to data which simply follow a uniform distribution, with additive Gaussian noise. When we predict this time series using simplified trees, the prediction error is calculated and decomposed into variance and other errors. When Gaussian or uniform effect is strong, those errors have different kinds of behavior. Other exploration is conducted in Section 3 including the best tree depth with minimum prediction error and the performance of Gaussian and quantile prediction intervals under different conditions. A real interval forecasting application is conducted in Section 4. Conclusions are drawn in Section 5. All calculations were done using *R* [19]; 'waveslim' [20] was used for wavelet decomposition and 'ctree' [21] for CTree.

2. Bias-Variance Exploration

2.1. Decomposition Background. For the i^{th} observation Y_i, the (unconditional) expectation is

$$\begin{aligned}
E(Y_i) &= E(\alpha + \beta X_i + \varepsilon_i) \\
&= \alpha + \beta E(X_i) + E(\varepsilon_i) \\
&= \alpha + \beta \cdot \frac{a+b}{2},
\end{aligned} \tag{3}$$

and the variance is

$$\begin{aligned}
\mathrm{Var}(Y_i) &= \mathrm{Var}(\alpha + \beta X_i + \varepsilon_i) \\
&= \beta^2 \mathrm{Var}(X_i) + \mathrm{Var}(\varepsilon_i) \\
&= \beta^2 \cdot \frac{(b-a)^2}{12} + \sigma^2.
\end{aligned} \tag{4}$$

They both have no relationship to i. In that case, $E(Y) = E(Y_i)$ and $\mathrm{Var}(Y) = \mathrm{Var}(Y_i)$. Accordingly, for N observations, the expectation and variance for the average $\overline{Y}$ are shown in

$$\begin{aligned}
E(\overline{Y}) &= \frac{1}{N}\left(E(Y_1) + E(Y_2) + \cdots + E(Y_N)\right), \\
E(Y) &= \alpha + \beta \cdot \frac{a+b}{2},
\end{aligned} \tag{5}$$

$$\begin{aligned}
\mathrm{Var}(\overline{Y}) &= \frac{1}{N^2}\left(\mathrm{Var}(Y_1) + \mathrm{Var}(Y_2) + \cdots + \mathrm{Var}(Y_N)\right) \\
&= \frac{1}{N}\mathrm{Var}(Y) \\
&= \frac{1}{N}\left\{\beta^2 \cdot \frac{(b-a)^2}{12} + \sigma^2\right\}.
\end{aligned} \tag{6}$$

2.2. Decomposition in the Context of Decision Trees. In the context of decision trees, the fitted model is $\widehat{f}(X)$ in a simplified form is

$$\widehat{f}(X) = \overline{Y}^i, \quad i = 1, 2, \ldots, k, \tag{7}$$

where k is the number of terminal nodes in the tree $\widehat{f}(X)$ and $\overline{Y}^i$ is the mean of y in terminal node i. In a tree with only the root node, $k = 1$, and the fitted model is $\widehat{f}(X) = \overline{Y}$. Then, for point x_0,

$$\begin{aligned}\text{Bias}^2\left(\widehat{f}(x_0)\right) &= \left[E\left\{\widehat{f}(x_0)\right\} - f(x_0)\right]^2 \\ &= \left[\alpha + \beta \cdot \frac{a+b}{2} - \alpha - \beta \cdot x_0\right]^2 \\ &= \beta^2\left(\frac{a+b}{2} - x_0\right)^2,\end{aligned} \tag{8}$$

and the variance is

$$\begin{aligned}\text{Var}\left(\widehat{f}(x_0)\right) &= E\left[\widehat{f}(x_0) - E\widehat{f}(x_0)\right]^2 \\ &= E\left[\overline{Y} - \left(\alpha + \beta \cdot \frac{a+b}{2}\right)\right]^2 \\ &= E\left(\overline{Y}^2\right) + \left(\alpha + \beta \cdot \frac{a+b}{2}\right)^2 - 2E(\overline{Y})\left(\alpha + \beta \cdot \frac{a+b}{2}\right) \\ &= E^2(\overline{Y}) + \text{Var}(\overline{Y}) - \left(\alpha + \beta \cdot \frac{a+b}{2}\right)^2 \\ &= \frac{\sigma^2}{N} + \beta^2\frac{(b-a)^2}{12N}.\end{aligned} \tag{9}$$

Thus, the SPE at point x_0 is

$$\begin{aligned}\text{SPE}\left(\widehat{f}(x_0)\right) &= \sigma^2 + \text{bias}^2\left(\widehat{f}(x_0)\right) + \text{Var}\left(\widehat{f}(x_0)\right) \\ &= \sigma^2 + \beta^2\left(\frac{a+b}{2} - x_0\right)^2 + \frac{\sigma^2}{N} + \beta^2\frac{(b-a)^2}{12N} \\ &= \left(1 + \frac{1}{N}\right)\sigma^2 + \beta^2\left(\frac{a+b}{2} - x_0\right)^2 + \beta^2\frac{(b-a)^2}{12N}.\end{aligned} \tag{10}$$

Then, the mean squared prediction error (MSPE) is

$$\begin{aligned}\text{MSPE} &= \int_a^b \text{SPE}(\widehat{f}(x))P(X = x)\text{d}x \\ &= \int_a^b \text{SPE}(\widehat{f}(x))\frac{1}{b-a}\text{d}x \\ &= \left(1 + \frac{1}{N}\right)\sigma^2 + \beta^2\frac{(b-a)^2}{12N} + \beta^2\int_a^b\left(\frac{a+b}{2} - x\right)^2\frac{1}{b-a}\text{d}x \\ &= \left(1 + \frac{1}{N}\right)\sigma^2 + \beta^2\frac{(b-a)^2}{12N} + \beta^2\frac{1}{12}(b-a)^2 \\ &= \left(1 + \frac{1}{N}\right)\left(\sigma^2 + \beta^2\frac{(b-a)^2}{12}\right).\end{aligned} \tag{11}$$

Comprising variance is

$$E(\text{Var}) = \frac{1}{N}\left(\sigma^2 + \beta^2\frac{(b-a)^2}{12}\right), \tag{12}$$

and squared bias is

$$E\left(\text{Bias}^2\right) = \beta^2\frac{(b-a)^2}{12}. \tag{13}$$

Now the number of terminal nodes in the decision tree is extended from $k = 1$ to a general k; then, the MSPE, bias^2, and variance for $x \in [a, b]$ are equal to those for $x \in [a, (a + ((b-a)/k))]$ since the decision tree is assumed to make k equal terminal nodes with the same number of observations in each terminal node. In that case, for $x \in [a, b]$ for a general k, the MSPE is

$$\text{MSPE} = \left(1 + \frac{k}{N}\right)\left(\sigma^2 + \beta^2\frac{(b-a)^2}{12k^2}\right), \tag{14}$$

with variance as

$$E(\text{Var}) = \left(\frac{k}{N}\right)\left(\sigma^2 + \beta^2\frac{(b-a)^2}{12k^2}\right) \tag{15}$$

and squared bias as

$$E\left(\text{Bias}^2\right) = \beta^2\frac{(b-a)^2}{12k^2}. \tag{16}$$

It is easy to see that with a lower $|\beta|$, $b - a$, and σ^2 and higher N, variance, squared bias, and MSPE will all decrease.

2.3. Optimal k to Minimize MSPE. The ideal number of terminal nodes can be found by minimizing the MSPE with respect to k. Here k is a discrete integer, so the target k will be the nearest integer from the differentiate result. Calculating the first derivative of MSPE, we get

$$\frac{\text{dMSPE}(k)}{\text{d}(k)} = \frac{\sigma^2}{N} - \frac{\beta^2(b-a)^2}{6k^3} - \frac{\beta^2(b-a)^2}{12\text{Nk}^2}, \tag{17}$$

and the second derivative of MSPE is always positive. Therefore, we only need to solve

$$\frac{\text{dMSPE}(k)}{\text{d}(k)} = 0, \quad k \in [1, N]. \tag{18}$$

The real root of (18) is

$$\begin{aligned}k_{\min} = \sqrt[3]{\frac{\beta^2(b-a)^2}{12\sigma^2}} \cdot \Bigg(&\sqrt[3]{N + \sqrt{N^2 - \frac{\beta^2(b-a)^2}{324\sigma^2}}} \\ &+ \sqrt[3]{N - \sqrt{N^2 - \frac{\beta^2(b-a)^2}{324\sigma^2}}}\Bigg).\end{aligned} \tag{19}$$

Having

$$N \gg \frac{\beta(b-a)}{18\sigma}, \tag{20}$$

we can approximate $k_{\min}$ by

$$k_{\min} \approx \sqrt[3]{\frac{\beta^2(b-a)^2N}{6\sigma^2}}. \tag{21}$$

In addition, the constraint for root k is also $k_{\min} \in [1, N]$. If $k_{\min}$ is not in $[1, N]$, MSPE might always decrease.

By substituting $k_{\min}$ in (19) back into (16), we will get

$$E\left(\text{Bias}^2\right) = 48\left(\frac{12\sigma^2\beta(b-a)}{N}\right)^{(2/3)}, \tag{22}$$

and it is easy to see that, with the increase of σ and $\beta(b-a)$, when N is fixed, $E(\text{Bias}^2)$ will increase. The others will be shown as figures.

Accordingly, how will the ratios $(E(\text{Var})/\text{MSPE})$, $(E(\text{Bias}^2)/\text{MSPE})$, (σ^2/MSPE) vary when parameters change? Since a, b, and β appear together, they are regarded as one parameter. For b and a, the thing that matters is their difference, so we use $a = 0$ and only change b. Here, k is set to be $k_{\min}$ calculated using given parameters for (19), and if $k_{\min}$ does not exist, the results will not be shown. The results in Figure 1 (changing $\beta^2(b-a)^2$) and Figure 2 (changing σ^2) show that, under both circumstances, MSPE, $E(\text{Var})$, and $E(\text{Bias}^2)$ all increase.

In Figure 1, when $\beta^2(b-a)^2$ gets bigger, X is more likely to be uniformly distributed and $k_{\min}$ increases as y is more accurately described with a uniform distribution; besides, the ratio of Var and Bias^2 over MSPE gets larger while σ^2 increases. In Figure 2, when σ^2 gets bigger, the Gaussian distribution will play a bigger role in data generation and $k_{\min}$ decreases. That is why (σ^2/MSPE) increases. $(E(\text{Var})/\text{MSPE})$ and $(E(\text{Bias}^2)/\text{MSPE})$ generally decrease. The decrease speed slows with bigger b and β as expected.

2.4. Simulation. In this simulation, a simplified tree model will be designed to confirm the theory results using simulated data. That is, when parameters of the simulated data change, the distribution of X and y will also change. The question is, how will the statistics of Var, Bias^2, MSE, and $k_{\min}$ change accordingly?

In the simplified tree, X is evenly split into k intervals, $i = 1, 2, \ldots, k$. For specific k, a, b, N, α, and β, we are going to calculate the statistics of MSPE, Var, and Bias^2 for the i^{th} interval in k from simulated data. Thus, for the i^{th} interval, the x range is

$$R_i = \left[a + \frac{(i-1)(b-a)}{K}, a + \frac{(i)(b-a)}{K}\right]. \tag{23}$$

The number of observations in interval i $(i = 1, 2, \ldots k-1)$ is n_i:

$$n_i = \left\lfloor \frac{\left(N - \sum_{j=0}^{i-1} n_j\right)}{(k-i)} \right\rfloor, \tag{24}$$

defining $n_0 = 0$ and $n_k = N - \sum_{j=0}^{k-1} n_j$.

(i) Step 1: for the data (x, y) in R_i, we train a model from them as

$$\widehat{f}_i(x) = \overline{y}, \tag{25}$$

for simulated y, and $\overline{y}$ is the averaged value of y in R_i.

(ii) Step 2: repeat Step 1 s times. Then, we have s trained models $\left\{\widehat{f}_j(x)\right\}$, $j = 1, 2, \ldots, s$.

(iii) Step 3: simulate one x_0 uniformly from the x range R_i. We are going to calculate the $\text{SPE}(x_0)$, $\text{Var}(x_0)$, and $\text{Bias}^2(x_0)$ for this specific x_0.

(iv) Step 4: simulate s values of y_j using x_0.

(v) Step 5: calculate the statistics of $\text{SPE}(x_0)$, $\text{Var}(x_0)$, and $\text{Bias}^2(x_0)$ for this specific x_0 as

$$\text{SPE}(x_0) = \frac{1}{s}\sum_{j=1}^{s}\left(\widehat{f}_i(x_0) - y_j\right)^2, \text{Bias}^2(x_0) = \left\{\frac{1}{s}\sum_{j=1}^{s}\widehat{f}_i(x_0) - f(x_0)\right\}^2, \text{Var}(x_0) = \text{variance}\left(\widehat{f}_i(x_0)\right). \tag{26}$$

(vi) Step 6: repeat Step 3 to Step 5 for n.repeat times and calculate the mean of $\text{SPE}(x_0)$, $\text{Var}(x_0)$, and $\text{Bias}^2(x_0)$ as MSPE_i, Bias_i^2, and Var_i.

Follow Step 1 to Step 6 for all i, $i = 1, 2, \ldots, k$, and calculate the mean as MSPE, Bias^2, and Var.

The results of simulations with 200 trials are shown in Figures 3 and 4. For Figure 3, $k_{\min} \in [1, N]$, we have a minimum MSPE. However, when $k_{\min} \notin [1, N]$ as in Figure 4, MSPE keeps decreasing.

3. Prediction Interval

Instead of point prediction, a prediction interval is also desirable especially for time series with high variance. If both the point prediction and the prediction interval can be provided, we will be more confident for the prediction. This study also helps us decide the proper prediction interval method for decision-tree-based regression problems. Gaussian-based prediction interval and quantile interval are compared under different parameters distributions.

3.1. Probability Function of Y. Since our linear model,

$$Y = \alpha + \beta X + \varepsilon, \tag{27}$$

is the sum of uniform and Gaussian distributions, the probability function for Y is

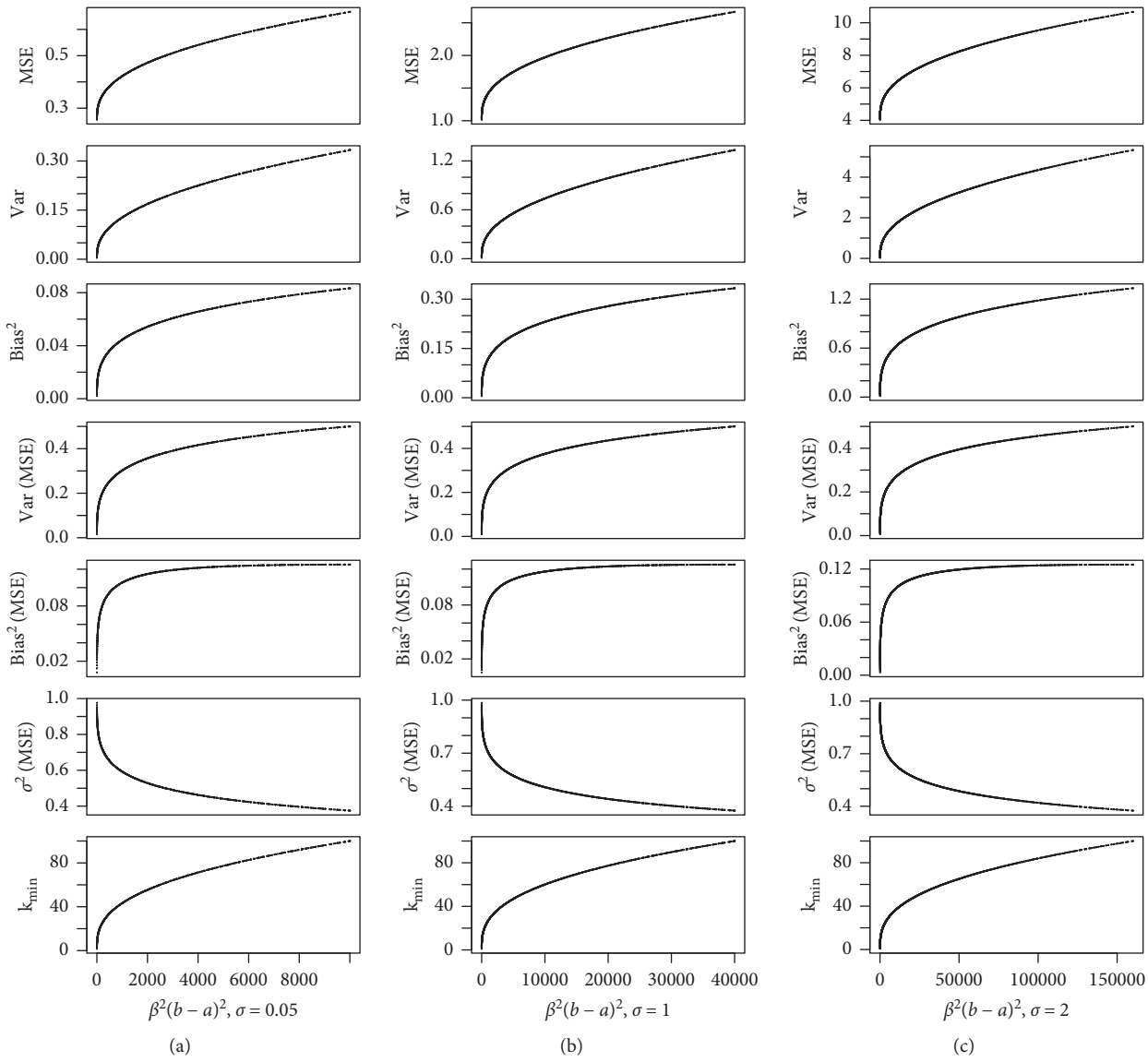

Figure 1: Ratios $(E(\text{Var})/\text{MSPE})$, $(E(\text{Bias}^2)/\text{MSPE})$, (σ^2/MSPE) with different $\beta^2(b-a)^2$. $N = 100$ and $\alpha = 0$.

$$P_Y(y) = \int_a^b P(Y = y|X = x)P_X(x)\mathrm{d}x = \int_a^b \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(y-(\alpha+\beta x))^2}{2\sigma^2}\right\}\frac{1}{b-a}\mathrm{d}x. \tag{28}$$

By letting $t = ((\alpha + \beta x - y)/\sigma)$, we obtain

$$P_Y(y) = \frac{1}{\beta(b-a)}\int_{(\alpha+\beta a-y/\sigma)}^{(\alpha+\beta b-y/\sigma)}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{t^2}{2}\right\}\mathrm{d}t = \frac{1}{\beta(b-a)}\left[\Phi\left(\frac{\alpha+\beta b-y}{\sigma}\right)-\Phi\left(\frac{\alpha+\beta a-y}{\sigma}\right)\right]. \tag{29}$$

Now we get the probability of Y as (29). However, $P_Y(y)$ is in a complex form meaning that the parameters are not easily solvable in theory by a given value for $P_Y(y)$.

3.2. Prediction Interval as a Gaussian Distribution. If we want to get the prediction interval, say $[y_1, y_2]$ for Y at $(1-p)$ level, the theoretical way is to obtain y_1 and y_2 from the equations

$$\int_{-\infty}^{y_1} P_Y(y)\mathrm{d}y = \frac{p}{2}, \qquad \int_{y_2}^{\infty} P_Y(y)\mathrm{d}y = 1-\frac{p}{2}. \tag{30}$$

However, the integral of Φ is not analytically solvable without approximating Φ with other suitable expressions. The results will also be quite complex. If we know the parameters values, then y_1 and y_2 can easily be found numerically.

From Figure 5, if the uniform (Gaussian) distribution plays a main role, then Y can be approximately described by a uniform (Gaussian) distribution. Under the conditions

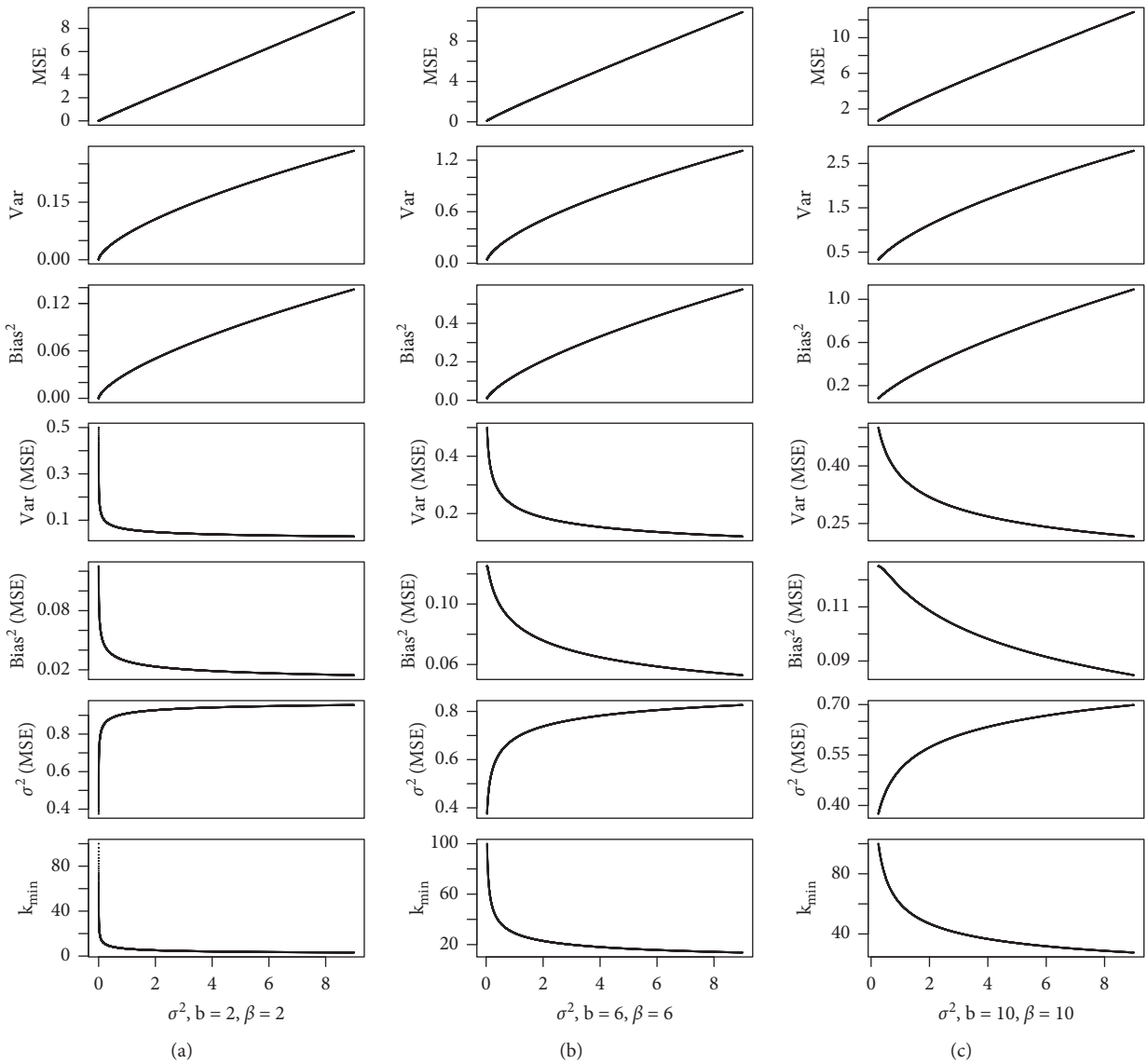

Figure 2: Ratios $(E(\text{Var})/\text{MSPE})$, $(E(\text{Bias}^2)/\text{MSPE})$, (σ^2/MSPE) with different σ^2. $N = 100$ and $\alpha = 0$.

that β is not too large, σ is not too small, and k is 1 (with only one interval), we will approximate the distribution of Y as a Gaussian distribution $N(\mu_Y, \sigma_Y^2)$:

$$Y \sim N\left(\alpha + \beta \cdot \frac{a+b}{2}, \beta^2 \cdot \frac{(b-a)^2}{12} + \sigma^2\right). \tag{31}$$

Then, the prediction interval under 95% criteria for this Gaussian distribution is around

$$\left[\overline{Y} - 1.96\sqrt{\left(1+\frac{1}{N}\right)\left(\beta^2 \cdot \frac{(b-a)^2}{12} + \sigma^2\right)}, \overline{Y} + 1.96\sqrt{\left(1+\frac{1}{N}\right)\left(\beta^2 \cdot \frac{(b-a)^2}{12} + \sigma^2\right)}\right]. \tag{32}$$

Then, for a general k, the prediction interval becomes

$$\left[\widehat{f} - 1.96 \times \sqrt{\left(1+\frac{k}{N}\right)\left(\beta^2 \cdot \frac{(b-a)^2}{12k^2} + \sigma^2\right)}, \widehat{f} + 1.96 \times \sqrt{\left(1+\frac{k}{N}\right)\left(\beta^2 \cdot \frac{(b-a)^2}{12k^2} + \sigma^2\right)}\right], \tag{33}$$

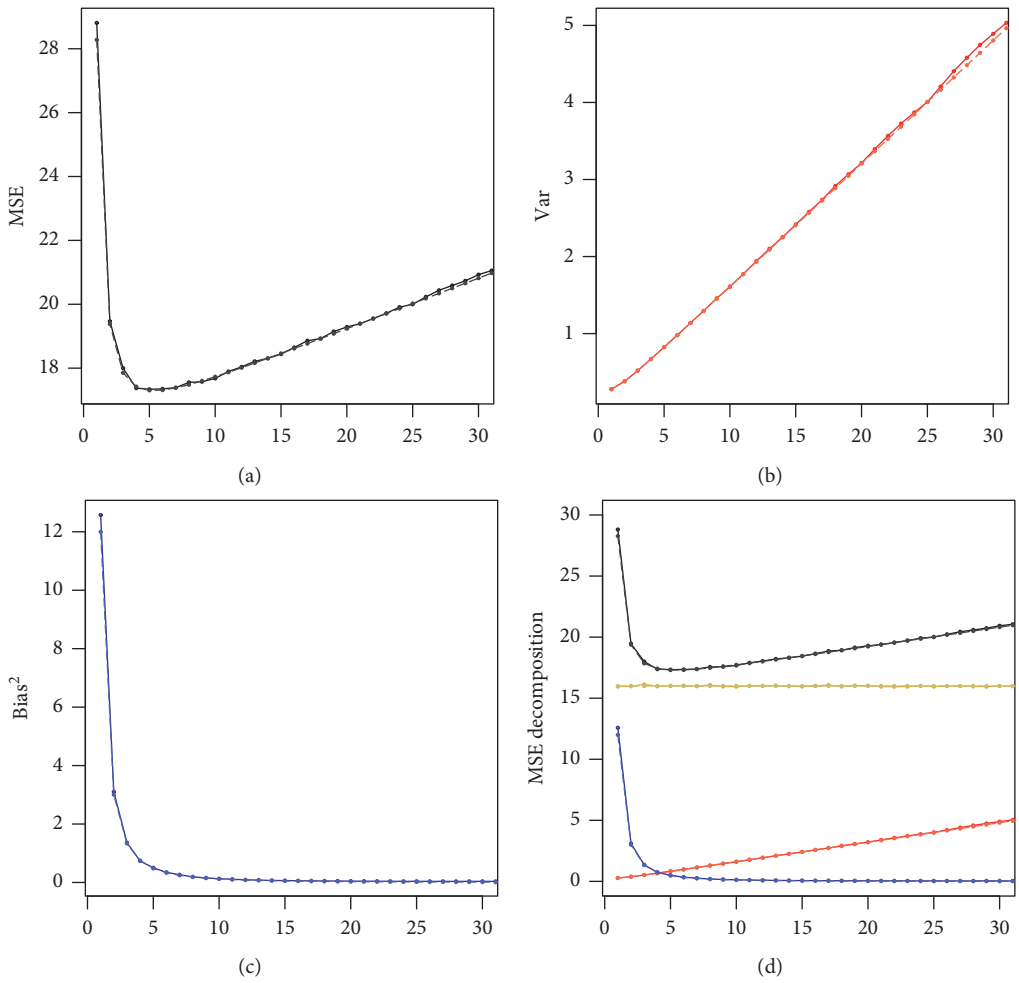

FIGURE 3: An example when $k_{\min}$ exists. The x axis label is k: the number of trees splits. MSPE, Bias^2, and Var are averaged calculations from 200 simulation trials. The black line is the MSPE, the blue line is the Bias^2, the orange line is σ^2, and the red line is the Var. The solid lines are from simulated data, and the dashed lines are theoretical calculations. The parameters values are given as follows: $\alpha = 2$, $\beta = 4$, $N = 100$, $a = 0$, $b = 3$, and $\sigma = 4$.

which is the form

$$[\widehat{f} - 1.96 \times \text{RMSPE}, \widehat{f} + 1.96 \times \text{RMSPE}], \tag{34}$$

a typical Gaussian prediction interval.

3.3. Prediction Simulation Using Gaussian Prediction Interval and Quantile Interval. In this simulation, we explore the performance of Gaussian prediction intervals and quantile intervals under different parameter combinations. The parameters include σ, $b - a$, β, and k. When the other parameters are fixed, a higher σ means a stronger Gaussian distribution effect, in which case, Gaussian prediction interval may work well. When $\beta^2 (b - a)^2$ is large, the uniform distribution plays a bigger role. Then, Gaussian prediction interval may not work so well. Both Gaussian prediction interval and quantile interval are influenced by the observation size of the terminal node. When the sample size is large, they can have stable performance, but when sample size is small, performance differs.

The Gaussian prediction interval in use is

$$[\widehat{f} - c\text{RMSPE}, \widehat{f} + c\text{RMSPE}], \tag{35}$$

where c is 1.96 and RMSPE is the root mean squared error estimated from the training data in each terminal node.

The quantile interval $[L, U]$ comes from the 0.025 and 0.975 quantiles of each terminal node from the training data.

(i) Step 1: training data generation.

Using given parameters α, β, a, b, σ, N, data are generated according to the model

$$Y = \alpha + \beta X + \varepsilon. \tag{36}$$

Therefore, we get the true fitted values for Y.

(ii) Step 2: RMSPE and quantiles from training data.

From this training data, the trained model, RMSPE, and quantiles are calculated as in the following steps.

(i) Step 2.1: model training.

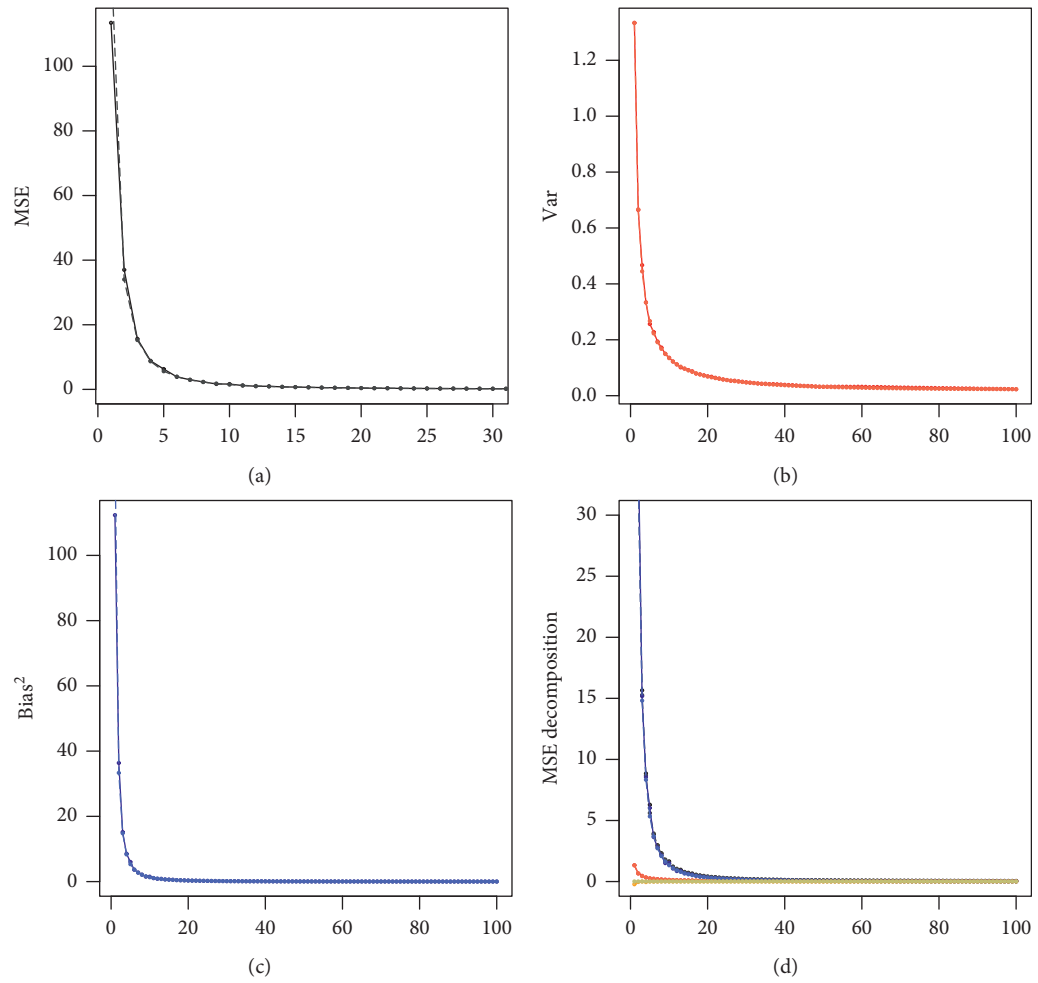

Figure 4: An example when $k_{\min}$ does not exist. The x axis label is k: the number of trees splits. MSPE, Bias^2, and Var are averaged calculations from 200 simulation trials. The black line is the MSPE, the blue line is the Bias^2, the orange line is σ^2, and the red line is the Var. The solid lines are from simulated data, and the dashed lines are theoretical calculations. The parameters values are given as follows: $\alpha = 2$, $\beta = 12$, $N = 100$, $a = 0$, $b = 4$, and $\sigma = 0.2$.

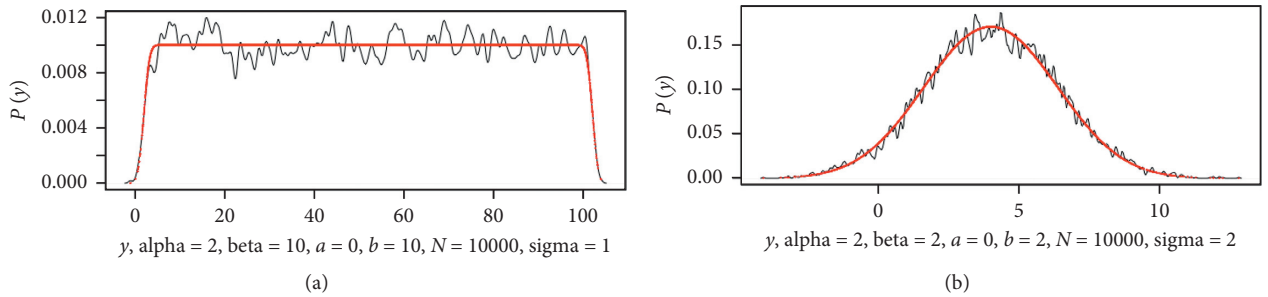

Figure 5: Probability density function of Y with different parameter values. The red line is the theoretical probability, and the black line is the simulated probability.

For training data A (the rest of data B is the test data), we sort the data x in an ascending order, so y will also be rearranged following x, and then A_{training} is divided into k roughly successive equal folds, making a total of N observations. The number of observations in fold i $(i = 1, 2, \ldots k)$ is n_i:

$$n_i = \left\lfloor \frac{\left(N - \sum_{j=0}^{i-1} n_j\right)}{(k-i)} \right\rfloor, \tag{37}$$

defining $n_0 = 0$.

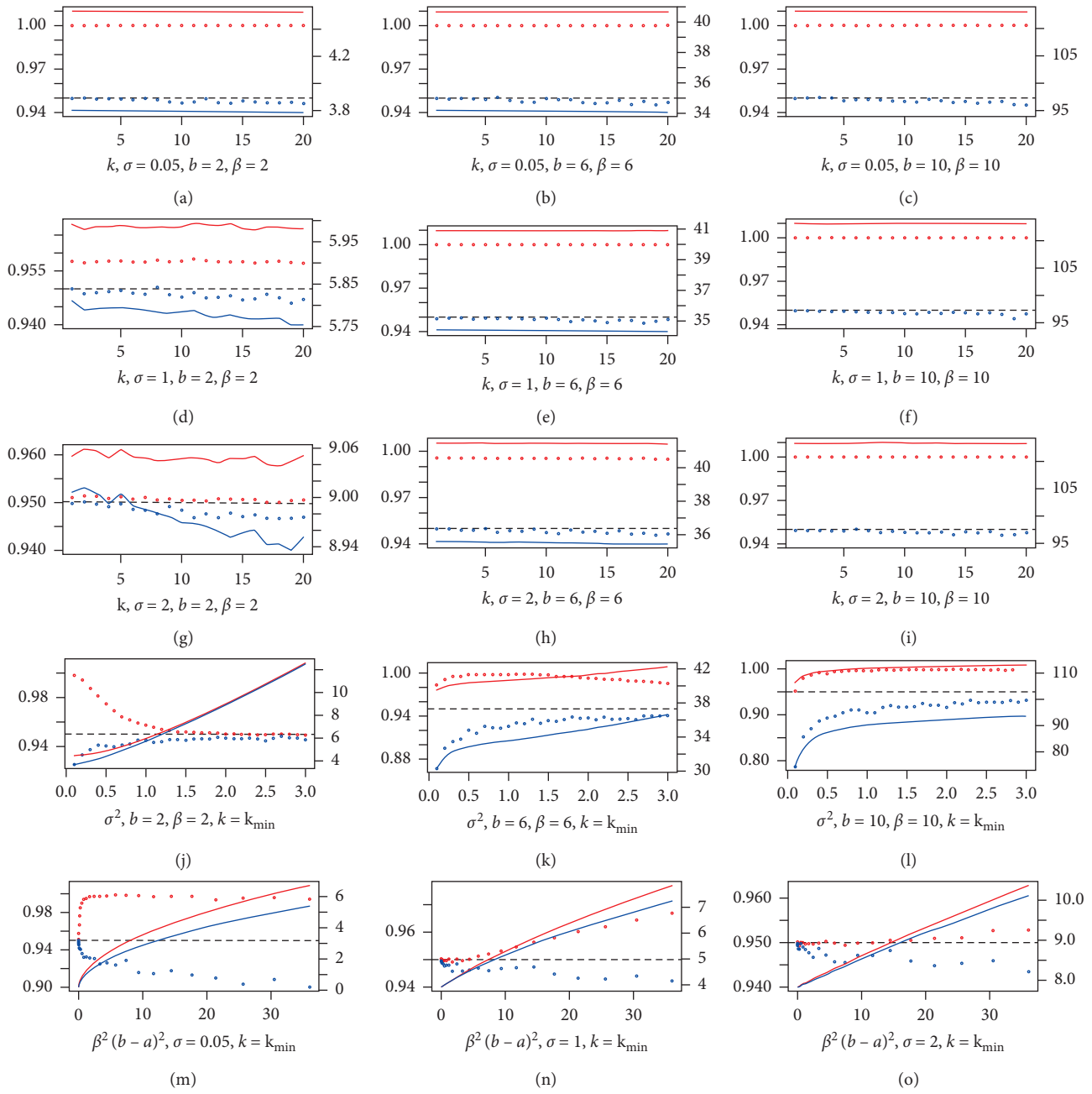

FIGURE 6: Prediction coverage using RMSPE. The black dash line is 0.95. The dash lines are coverage. The solid lines are width. The red lines represent Gaussian prediction interval, and the blue lines represent quantile prediction interval. The right-hand y axis (purple) is the axis for width. $N = 10000$.

For the i^{th} fold in A, giving x_i and y_i, the predicted value will be

$$\widehat{y} = \overline{y}_i, \quad \$x \in [\min(x_i), \max(x_i)]\$, \tag{38}$$

in the trees context. The predicted value of a tree model is the averaged response values of each terminal node. Samples being split into those terminal nodes will have the corresponding averaged value as the predicted value.

(ii) Step 2.2: RMSPE and quantile calculation.

When the model for each i is trained as model_{A_i}, the predicted values for y in A will be $\widehat{y}$. Then, the RMSPE for the training data is

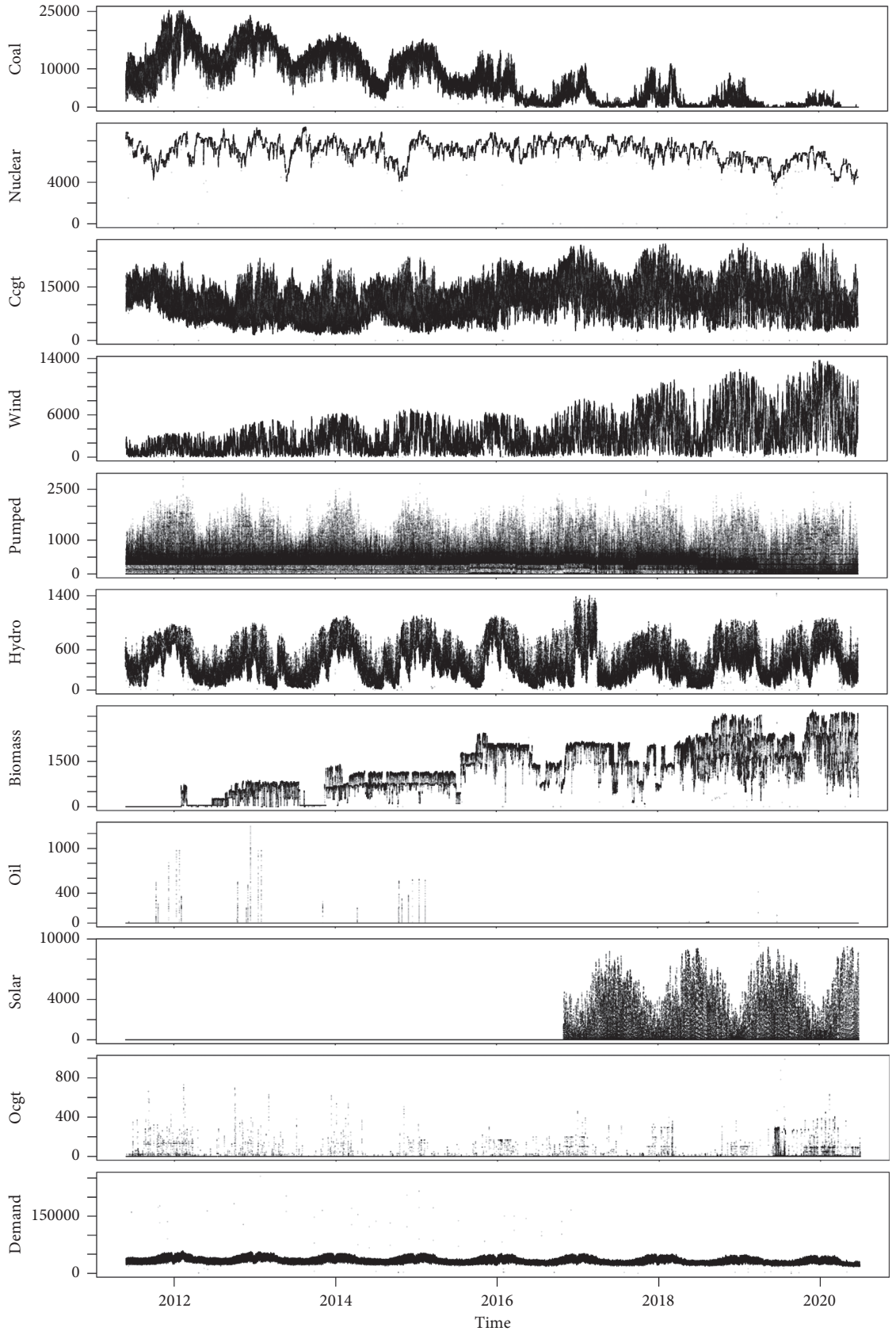

Figure 7: The supply and demand of energy in UK from year 2011 to year 2020. The time gap is 5 minutes.

$$\text{RMSPE} = \sqrt{\frac{\sum_{r=1}^{N} (y - \hat{y})^2}{N}}. \quad (39)$$

The quantile intervals L and U are the 0.025 and 0.975 quantiles of the i_{th} training data y.

(iii) Step 3: test data generation and model testing.

Using the same parameters α, β, a, b, and σN as in Step 1, data are generated according to

$$Y = \alpha + \beta X + \varepsilon. \quad (40)$$

Then, the test data B are put into model$_A$ and the coverage is computed as

$$\frac{1}{N} \sum_{r=1}^{N} I(\hat{y} - c\,\text{RMSPE} < y < \hat{y} + c\,\text{RMSPE}). \quad (41)$$

(iv) Step 4: repeating Steps 1 to 3.

Repeat Steps 1 to 3 s times to get an averaged coverage. Using parameters $a = 0$, $\alpha = 2$, and $s = 200$, the results are shown in Figures 6.

The results show that quantile interval coverages are closer to the 0.95 reference line for fixed σ, b, and β. Gaussian prediction interval is only closer to the 0.95 coverage when σ is large; otherwise, it is larger than 0.95 at the cost of wider width. When k is chosen as the best k_{min}, the coverages get closer to the 0.95 reference line as σ increases for both quantile and Gaussian prediction intervals. However, when the uniform distribution effect gets stronger, the coverages all go far away from 0.95. Accordingly, when the number of observations for each terminal node is large and the data distribution is not obviously Gaussian, quantile intervals are suggested. When the data follows obvious Gaussian distribution, Gaussian prediction intervals are recommended.

4. Real Application

We have explored the performance of decision trees under different circumstances. A real application is conducted in this section. The data come from UK Gridwatch (http://www.gridwatch.templar.co.uk/), which are the demand data of grid and the supply data of each energy source. The time series begin from year 2011 to year 2020, making a total of 953824 observations with a record every 5 minutes. The details are shown in Figure 7.

From the figure, we can see that the demand of grid changes in period as expected since there are peak and valley values daily and seasonally. The general trend of grid demand changes a little. Some kinds of energy like wind and biomass increase a lot in supply these years; they will be more frequently used in the future than traditional energy like coal as they are more environmentally friendly. We construct a metric ratio to measure the ratio of other energy supply over that of coal. By deleting observations which have none or zero values of coal, we have 847922 observations left, as shown in Figure 8.

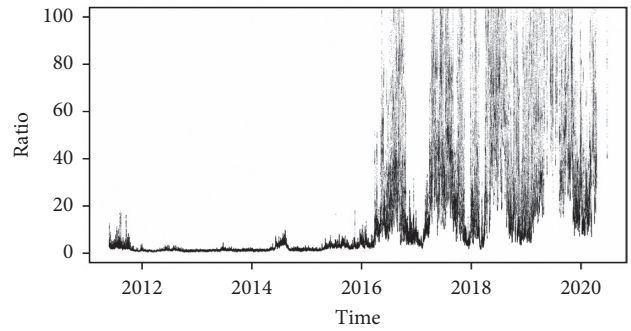

Figure 8: The ratio of other energy supply over that of coal.

Table 1: Results of interval forecasting on the time series ratio.

Coverage (%)	Averaged width	Computational time (s)
80.31	22.95	11.86

The experiment is run on a 3.2 GHz 8-Core Intel Xeon W processor.

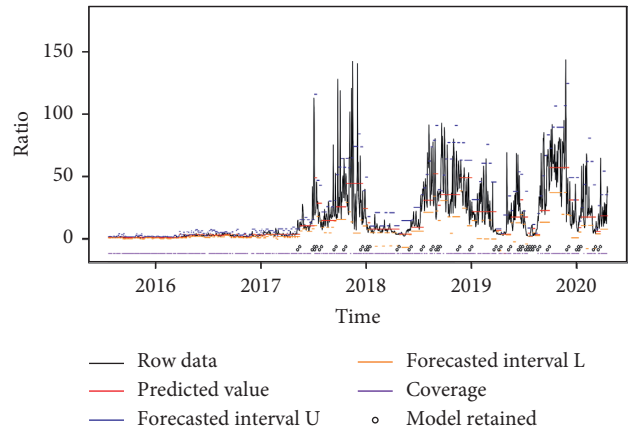

Figure 9: The dynamic interval forecasting result. The black lines are the raw daily ratio time series; the red lines are the predicted values; the blue line is the upper forecasted interval; and the yellow line is the lower forecasted interval. The purple line describes whether the data are coved by the interval or not. The circle means the model is retrained on that time point. The upper and lower forecasted interval $[U, L]$ give an interval that in most cases the real future value will be in, which is similar to the 95% confidence interval that the fitted value is most likely to be covered in.

We average the time series ratio from a frequency of 5 minutes to a daily basis, ending with 2954 observations left. A forecasting method is conducted on ratio to help us know how much renewable energy is needed in the near future. The interval forecasting method we use is from our designed method, Zhao et al. [14], called ctreeone, which uses the tree method ctree in a dynamic interval forecasting context. The different parameter we choose is 7 for time gap (weekly dynamic forecasting), leaving the other parameters unchanged.

Interval forecasting provides not only the point forecasting results but also the prediction interval that the predicted point belongs to. Small change of the ratio often happens, which influences a little the the energy supply and demand system, so no action is needed in this circumstance. When the predicted ratio changes a lot, out of a preset limit, an alarm may be raised to help the system accommodate to the new circumstance, for example, by producing more renewable energies in advance to meet the instant demand. The interval forecasting model provides such an alerting system to adjust the energy production.

The results are shown in Table 1 and Figure 9. The coverage and width make a good balance; that is, a higher coverage costs a relatively higher width. We end with a coverage of 80.31% and a suitable width of 22.95 which is similar to the standard deviation of ratio of 19.78.

5. Conclusion

In this paper, the data are constructed using a simple model that includes both Gaussian and uniform distributions. We explore the squared prediction error in the context of trees and decompose that error into bias, variance, and irreducible error. The bias decreases when the tree gets bigger. However, for squared prediction error and variance, the relationship is not monotonic. We also calculate the best tree depth with a minimum mean squared prediction error. When Gaussian effect dominates, the best tree depth density decreases. However, when uniform effect dominates, the best tree depth increases. Under both circumstances, mean squared error, variance, and bias all increase.

After that, two options are given for the prediction interval using Gaussian prediction interval or quantile interval. When Gaussian distribution is obviously dominant, Gaussian prediction intervals are suggested. Otherwise, quantile intervals are suggested, which is also why quantile intervals are chosen as the prediction intervals in our regression application, although they both perform poorly when the uniform distribution is quite strong. When the number of observations is small in the terminal node, both interval constructions perform poorly in terms of coverage.

In the real data application, we applied our method to the UK grid energy supply and demand data to forecast the ratio of renewable energy supply over that of coal. We have good forecasting results as 80.31% in interval coverage and 22.95 in interval width. The method can be extended to other models as well besides decision trees.

We use the model decision tree for interval forecasting. In practice, other models can also be considered. For example, Hall et al. [22] used multiple nonlinear regression to forecast and analyze the changes of climate and weather dynamics and proposed a simple model averaging approach to reduce model and prediction uncertainty. Besides decision tree, other dynamic regression models can also be considered, for example, Gu et al. [23] used dynamic regression model to predict the dynamics of a specific space weather index and proposed a new approach for prediction uncertainty analysis using point-cloud model parameters. Dynamic regression model was also applied to social dynamic behavior modeling and analysis [24].

In the future research, the model can be applied to more kinds of datasets to test its generation ability. In the simulation, instead of linear model, nonlinear model can also be considered to test the tree performance.

Data Availability

The source code and simulation data in the theory exploration are available from the corresponding author upon request. The real data in application can be openly accessed from Elexon Portal (cited June 2020) [25].

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research is improved from the PhD thesis of Xin Zhao. She is grateful for the financial support of the Fundamental Research Funds for the Central Universities (nos. 2242020R40073 and 2242020R10051) and Jiangsu Science Foundation for Youths (no. SBK2020040696) during this research, which was mostly completed during her PhD studies at the University of Leeds. Xiaokai Nie is grateful for the financial support of the Fundamental Research Funds for the Central Universities (no. 2242020R10053), Nanjing Prioritized Fund for Science and Technology Innovation (no. 1108000241), and Essential Science Indicator Improvement Funds of Southeast University (no.4016002011) during this research.

References

[1] F. von Loeper, P. Schaumann, M. de Langlard, R. Hess, R. Bäsmann, and V. Schmidt, "Probabilistic prediction of solar power supply to distribution networks, using forecasts of global horizontal irradiation," *Solar Energy*, vol. 203, pp. 145–156, 2020.

[2] Y. Zhang, J. Xia, X. Zhang et al., "Modeling and prediction of the reliability analysis of an 18-pulse rectifier power supply for aircraft based applications," *IEEE Access*, vol. 8, pp. 47063–47071, 2020.

[3] J. Huan, H. Li, M. Li, and B. Chen, "Prediction of dissolved oxygen in aquaculture based on gradient boosting decision tree and long short-term memory network: a study of chang zhou fishery demonstration base, China," *Computers and Electronics in Agriculture*, vol. 175, 2020.

[4] J. Liu and Y. Li, "Study on environment-concerned short-term load forecasting model for wind power based on feature extraction and tree regression," *Journal of Cleaner Production*, vol. 264, 2020.

[5] D. Raspopov and P. Belousov, "Development of methods and algorithms for identification of a type of electric energy consumers using artificial intelligence and machine learning models for smart grid systems," *Procedia Computer Science*, vol. 169, pp. 597–605, 2020.

[6] C.-s. Shi and D.-p. Meng, "Combined prediction model for supply risk in nuclear power equipment manufacturing

industry based on support vector machine and decision tree," *Nuclear Power Engineering*, vol. 32, no. 5, p. 138, 2011.

[7] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*, Wiley, Hoboken, NJ, USA, 1984.

[8] F. De Felice, D. Crocetti, M. Parisi et al., "Decision tree algorithm in locally advanced rectal cancer: an example of over-interpretation and misuse of a machine learning approach," *Journal of Cancer Research and Clinical Oncology*, vol. 146, no. 3, pp. 761–765, 2020a.

[9] W. Kuang, Y.-L. Chan, S.-H. Tsang, and W.-C. Siu, "Machine learning-based fast intra mode decision for hevc screen content coding via decision trees," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 5, pp. 1481–1496, 2020.

[10] X. Zhao, S. Barber, C. C. Taylor, and Z. Milan, "Classification tree methods for panel data using wavelet-transformed time series," *Computational Statistics & Data Analysis*, vol. 127, pp. 204–216, 2018.

[11] C. Chen, L. Geng, and S. Zhou, "Design and implementation of bank crm system based on decision tree algorithm," *Neural Computing and Applications*, p. 1, 2020.

[12] M. Durica, J. Frnda, and L. Svabova, "Decision tree based model of business failure prediction for polish companies," *Oeconomia Copernicana*, vol. 10, no. 3, p. 453, 2019.

[13] X. Hu, Y. Yang, L. Chen, and S. Zhu, "Research on a customer churn combination prediction model based on decision tree and neural network," in *Proceeding of the 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics*, pp. 129–132, Chengdu, China, April 2020.

[14] X. Zhao, S. Barber, C. C. Taylor, and Z. Milan, "Interval forecasts based on regression trees for streaming data," *Advances in Data Analysis and Classification*, 2019.

[15] T. R. Hancock, "Learning k μ decision trees on the uniform distribution," in *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, pp. 352–360, Santa Cruz, CA, USA, July 1993.

[16] J. C. Jackson and R. A. Servedio, *Learning Random Log-Depth Decision Trees under the Uniform Distribution, Learning Theory and Kernel Machines*, pp. 610–624, Springer, Berlin, Germany, 2003.

[17] A. P. White and W. Z. Liu, "Bias in information-based measures in decision tree induction," *Machine Learning*, vol. 15, no. 3, pp. 321–329, 1994.

[18] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, Berlin, Germany, 2001.

[19] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018, https://www.R-project.org/.

[20] B. Whitcher, *Waveslim: Basic Wavelet Routines for One-, Two- and Three-Dimensional Signal Processing*, 2019, https://CRAN.R-project.org/package=waveslim.

[21] T. Hothorn, K. Hornik, and A. Zeileis, *Ctree: Conditional Inference Trees*, The Comprehensive R Archive Network, 2015.

[22] R. J. Hall, H. L. Wei, and E. Hanna, "Complex systems modelling for statistical forecasting of winter north atlantic atmospheric variability: a new approach to north atlantic seasonal forecasting," *Quarterly Journal of the Royal Meteorological Society*, vol. 145, no. 723, pp. 2568–2585, 2019.

[23] Y. Gu, H. L. Wei, R. J. Boynton, S. N. Walker, and M. A. Balikhin, "System identification and data-driven forecasting of AE index and prediction uncertainty analysis using a new cloud NARX model," *Journal of Geophysical Research: Space Physics*, vol. 124, no. 1, pp. 248–263, 2019.

[24] H. Wei and G. R. Bigg, "The dominance of food supply in changing demographic factors across africa: a model using a systems identification approach," *Social Sciences*, vol. 6, no. 4, p. 122, 2017.

[25] E. portal, *G.b. national grid status*, 2020, http://www.gridwatch.templar.co.uk/.