Bias-corrected Bagging in Active Learning with an Actuarial Application

Yangxuan Xu *supervised by* Dr. Hyukjun Gweon and Dr. Shu Li
Department of Statistical and Actuarial Sciences, Western University
yxu827@uwo.ca

Introduction

Variable annuity (VA) is a modern insurance product that offers certain guaranteed protections and tax-deferred treatment. Holding VA contracts could benefit from a separate account which makes outside investment and a general account that guarantees some minimum payoff (Gan, 2013). In Canada this product is called segregated-fund; or unit-linked insurance in the United Kingdom; variable annuities and equity-indexed annuities in the United States (Hardy, 2003).

Because of the inherent complexity of guarantees' payoff, the closed-form solution of fair market values (FMVs) is often not available. Most insurance companies depend on **Monte Carlo (MC) simulation** to price the FMVs of these products, which is an extremely computational intensive and time-consuming approach (Gan, 2013; Gan and Valdez, 2018). However, the **metamodeling approach** can be used to circumvent the heavy computation. The idea of metamodeling approach to predict the FMVs consists of four main steps (Barton, 2015): (i) (sampling stage) acquiring a small number of representative VA contracts; (ii) (labeling stage) labeling these contracts using MC Simulation; (iii) (modeling stage) fitting a machine learning model to the selected representative data; and (iv) (prediction stage) estimating the FMVs of all unlabeled VA contracts in the portfolio through statistical predictive analysis.

In the modeling stage, the bagged tree method has proved to outperform other parametric approaches (Gan et al., 2018). Gweon et al. (2020) also applied a **bias-corrected (BC) bagging model** and showed significant improvement for prediction performance.

When the number of unlabeled data is large, the budget for labeling is limited and labels are expensive to obtain, the **active learning framework** (Cohn et al., 1994) could achieve satisfactory prediction performance with a small amount of labeled data (i.e., representative VA contracts). Two active learning approaches, the weighted random sampling (WRS) and repeated random sampling (RRS) have been tried on selecting representative contracts and showed good performance (Gweon and Li, 2021).

For the summer research program, we investigate whether the BC bagging model is also effective in the active learning framework.

Regression Trees and Bagging

Regression and classification tree (Breiman et al., 1984) is an effective machine learning method, which stratifies or splits the predictor space into different regions. Figure 1 gives a simple regression tree example, which is constructed by using two predictors (x1 and x2). Before reaching the terminal nodes, the algorithm determines the value at each binary split.

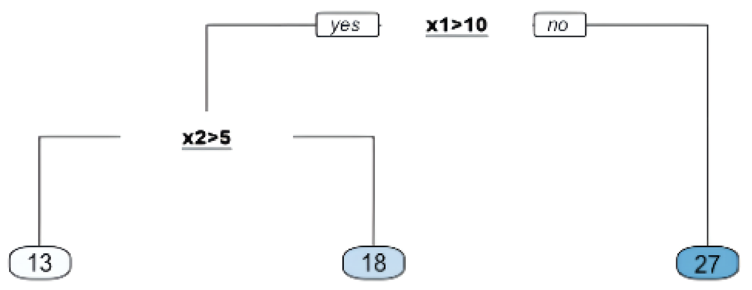

Figure 1. *A simple regression tree example.*

Given the decision tree often suffers from high variance, the **bootstrap aggregation, or bagging** (Breiman, 1996) can help to reduce the variance using bootstrap samples from the original data. The idea of bagging is constructed as follows.

The predictor vector of the portfolio is $X = \{x_1, ..., x_n\}$, where $x_i \in R^p$ and $i = 1, 2, ..., n$. The target, FMV for x_i is $Y = \{y_1, ..., y_n\}$. The relationship between Y and $x \in X$ can be written as

$$Y = f(x) + \epsilon,$$

where $f(x)$ is the approximated functional relationship and ϵ is the error term.

The method of **bootstrap** (Efron, 1979) is used to obtain multiple training sets. We select a small subset Z of X which contains j representative contracts and label them using MC Simulation. The selected training set is $Z = \{z_1, ..., z_j\}$ where $z_j \in X$ with FMVs $\{y_1, ..., y_j\}$. Let $L = \{(z_1, y_1), ..., (z_j, y_j)\}$ be the labeled training dataset. We fit the bagging model for the training set and obtain

$$\hat{f}(z, L) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}(z, L_b), \tag{1}$$

where B is the number of bootstrap samples, $b = 1, ..., B$, L_b is the bth bootstrap sample and $I(*)$ is the indicator function. Then we use this model to predict FMVs for unlabeled contracts.

Bias-corrected Bagging (BC-bagging)

A common measurement for the prediction performance of $\hat{f}(x, L)$ is the mean squared prediction error (MSPE):

$$MSPE(\hat{f}(x, L)) = E((\hat{f}(x, L) - Y)^2).$$

Given $Y = f(x) + \epsilon$, via bias-variance decomposition:

$$MSPE(\hat{f}(x, L)) = E((\hat{f}(x, L) - E(\hat{f}(x, L))^2) + (E(\hat{f}(x, L)) - f(x))^2 + E(\epsilon^2). \tag{2}$$

According to (2), the MSPE is decomposed into variance, bias^2 and noise. It is important to achieve both a low variance and a low bias to obtain a low MSPE. Since bagging effectively reduce the prediction variance, we consider bias as a factor dominating the prediction error (Zhang and Lu, 2012). We use the **BC bagging model** to correct the prediction biases by fitting another bagging model to estimate biases, and subtracting them from the predicted FMVs.

We define the bias $B(x, L)$ as the difference between predicted and true value, then

$$E(\hat{f}(x, L)) - Y = B(x, L) - \epsilon, \tag{3}$$

where $\hat{f}(x, L)$ is the predicted value of bagging, Y is the FMVs computed by MC simulation. We can estimate the prediction bias $B(x, L)$ by fitting another bagging model. The response for the second bagging model is $E(\hat{f}(x, L)) - Y$ instead of Y. Since we cannot predict the FMVs in training stage with untrained model, we consider using the **out-of-bag (OOB) prediction** of the former bagging model. The OOB prediction is defined as

$$\hat{f}^{OOB}(z, L) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}(z, L_b) I((z, y) \notin L_b)),$$

where B is the number of trees that do not use observation (z, y) and $I(*)$ is the indicator function. After we obtain the OOB prediction $\hat{f}^{OOB}(z,L)$ and FMVs computed by MC simulation Y, $\hat{f}^{OOB}(z,L) - Y$ will be the response. Fit a bagging model to the training data to estimate prediction biases $\hat{B}(x, L)$, and the predicted FMV of VA contract x is

$$\hat{Y}^* = E(\hat{f}(x, L)) - \hat{B}(x, L).$$

Active Learning Framework

Active learning approach (Cohn et al., 1994) is considered when the number of unlabeled data is large and labels are expensive to obtain, like in our scenario. We want to reduce the number of representative contracts to save budget and time if we can achieve a good performance of prediction at the same time.

Figure 2 illustrates the active learning framework, check Gweon and Li (2021) for more details:

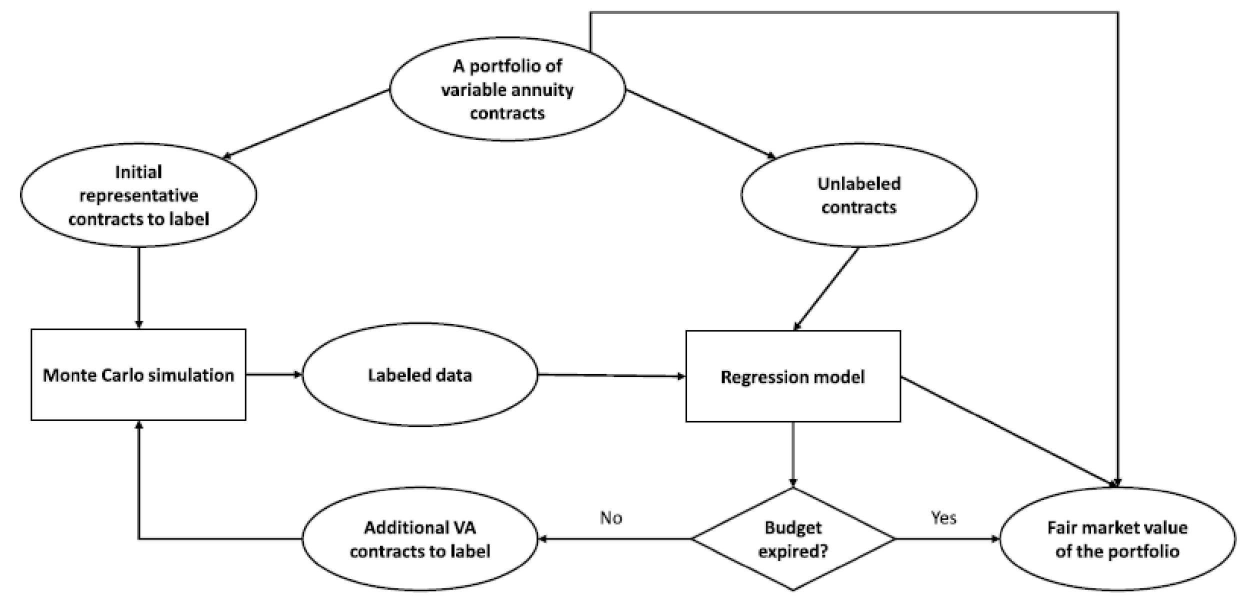

Figure 2. *An illustration of active learning framework (Gweon and Li, 2021).*

Gweon and Li (2021) used **Query-By-Committee (QBC) approach** on the VA portfolios valuation, which can be categorized into uncertainty sampling. Under the QBC-based method, the informativeness of unlabeled data x, the **ambiguity** is defined as

$$V = \frac{1}{B}\sum_{b=1}^{B}(\hat{f}(x, L_b) - \hat{f}(x, L))^2, \tag{4}$$

where B is the number of trees used in bagging, L_b is the bth bootstrap sample and $\hat{f}(x, L)$ is the predicted value for aggregate tree of unlabeled instance x. According to (1), $\hat{f}(x, L) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}(x, L_b)$.

Two uncertainty sampling approaches based on this defined ambiguity are as follows.

- **Weighted Random Sampling (WRS)**

Let U be the current unlabeled data set. For $x_i \in U$, we calculate the sampling weights of prediction ambiguity

$$w_i = \frac{V_i}{\sum_h V_h},$$

where V_i is the ambiguity of the unlabeled contract x_i, $i = 1, ..., h$. Then a subset of unlabeled data can be sampled with the probabilities w and added to the training set.

- **Repeated Random Sampling (RRS)**

The repeated random sampling evaluates the informativeness of a group of contracts, which follows three steps:

1. Draw a sample of size k from unlabeled data U by simple random sampling and denote the batch as S_1.
2. Repeat the random sampling m times and generate groups $S_1, S_2, ..., S_m$.
3. Assess the informativeness of the m groups of contracts and choose the most informative sample.

To measure the informativeness, RRS uses $I_S = A_S + D_S$ where

$$A_S = \frac{1}{k}\sum_{x_i \in S} V_i,$$

and

$$D_S = \frac{1}{k-1}\sum_{x_i \in S}(\hat{f}(x_i, L) - E(\hat{f}(x_i, L))^2.$$

Dataset and Evaluation Measurements

This dataset consists of 190,000 synthetic VA contracts from Gan and Valdez (2017). We use 16 explanatory variables (14 continuous and 2 categorical) and exclude other variables that are identical for all policies. Check Gan and Valdez (2017) for more details. Table 1 summarizes the basic characteristics of the selected predictors in VA dataset.

Table 1. *Summary Statistics of the Predictors in VA Dataset*

Category	Description			Count
Gender	M (Male Policyholder)			113,993
	F (Female policyholder)			76,007
Product type	ABRP (GMAB with return of premium)			10,000
	... (other 18 types)			10,000
Continuous	Description	Minimum	Mean	Maximum
gmwbBalance	GMWB balance	0	35,612	499,709
gbAmt	Guaranteed benefit amount	0	326,835	1,105,732
FundValue1	Account value of the 1st investment fund	0	33,434	1,099,205
FundValue2	Account value of the 2nd investment fund	0	38,543	1,136,896
FundValue3	Account value of the 3rd investment fund	0	26,740	752,945
FundValue4	Account value of the 4th investment fund	0	26,142	610,580
FundValue5	Account value of the 5th investment fund	0	23,027	498,479
FundValue6	Account value of the 6th investment fund	0	35,576	1,091,156
FundValue7	Account value of the 7th investment fund	0	29,973	834,254
FundValue8	Account value of the 8th investment fund	0	30,212	725,745
FundValue9	Account value of the 9th investment fund	0	29,958	927,513
FundValue10	Account value of the 10th investment fund	0	29,862	785,979
age	Age of the policyholder	34.52	49.49	64.46
ttm	Time to maturity in years	0.59	14.54	28.52

We employ three measurements, R^2, mean absolute error (MAE) and percentage error (PE). Let y_i and $\hat{y}_i$ be the true value (computed by MC simulation) and predicted value respectively, for $i = 1, ..., N$ where N is the total number of contracts that needs to be valued, the metrics are defined as

$$R^2 = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \overline{y})^2}, \quad MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|, \quad PE = \frac{\sum(y_i - \hat{y}_i)}{\sum y_i}.$$

Results and Conclusions

Figure 3 shows the comparison of different methods in terms of R^2, MAE and PE. It is obvious that bias-corrected approach (dashed lines) significantly improves the prediction performance based on R^2 and MAE. For percentage error, BC bagging still lowers the absolute values for WRS and RRS.

Figure 4 presents the smoothed curves for prediction errors and density histogram of FMVs. When FMVs are large, the prediction errors tend to be large as well due to the high skewness of the response. However, BC bagging (dashed lines) vastly reduces these errors which could be useful especially for valuing VA contracts with large FMVs.

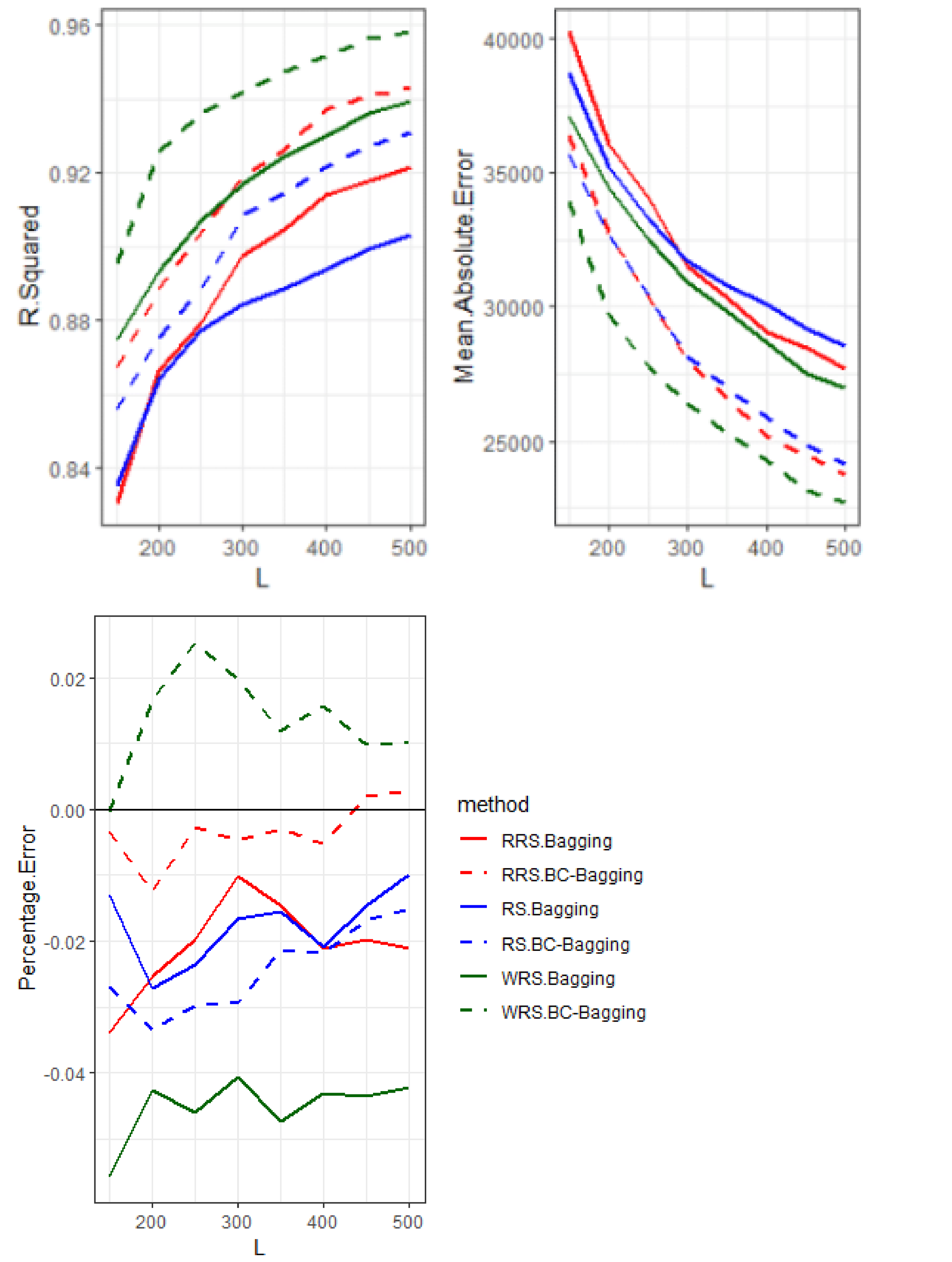

Figure 3. *Comparison of different methods in terms of R^2, mean absolute error and percentage error.*

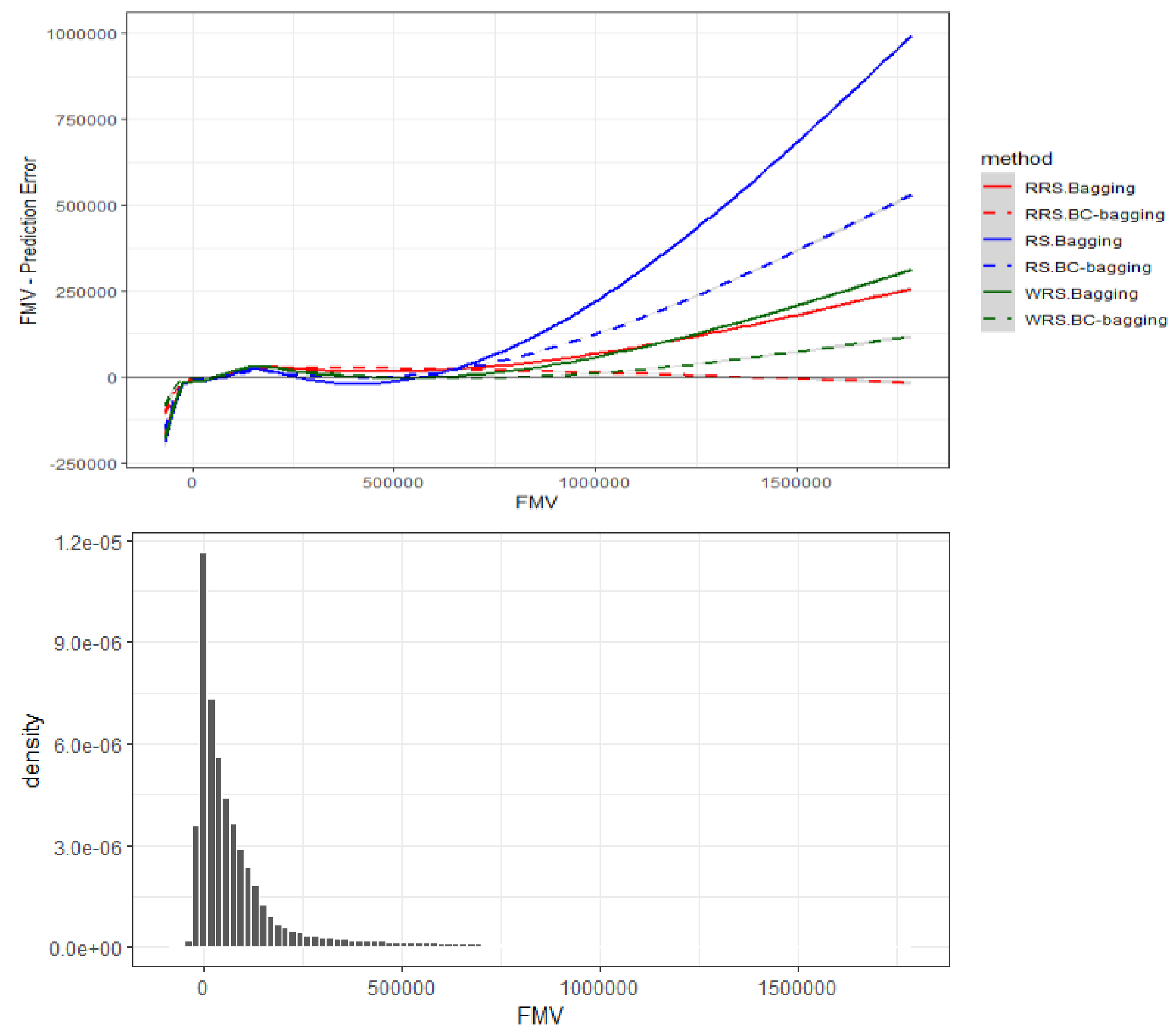

Figure 4. *Upper: difference between the FMVs calculated by MC simulation and the predicted values for different methods. Lower: density histogram of the FMVs calculated by MC simulation.*

In conclusion, the bias-corrected approach vastly improves the prediction performance of bagging under uncertainty sampling methods, WRS and RRS. Especially, for the VA portfolios skewed to the right, BC-bagging performs well in the tail valuation, which make it a promising approach for insurance industry.

Acknowledgment

I would like to express my deepest appreciation to my supervisors Dr. Gweon and Dr. Li for their guidance and support. We also appreciate the funding support from the Multi-hazard Risk and Resilience Interdisciplinary Development Initiative for this USRI project.