**Table 1** Summary of variable selection methods for random forest regression. Methods are presented in chronological order by year of publication

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Abbreviation in Paper** | **Publication** | **R package/**  **Implementation** | **Approach** | **Type of forest method** | **Summary** | **Parameter Settings** |
| None | [Breiman 2001 [1]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7508310/#R2) | *ranger* | N/A | Axis-based | No variable selection | Default |
| Svetnik | [Svetnik 2004 [16]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7508310/#R23) | Uses *party*, code from Hapfelmeier [15] | Performance Based | Conditional Inference | Uses backward elimination based on importance measures and k-fold validation | # trees=100, # folds=5, # repetitions=20 |
| Jiang | [Jiang 2004 [17]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7508310/#R17) | Uses *party*, code from Hapfelmeier [15] | Performance Based | Conditional Inference | Similar to Svetnik but provides mechanism to prevent overfitting | # trees=500, # Standard Error Rule = minimum |
| Caret | [Kuhn 2008 [8]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7508310/#R18) | *caret* | Performance Based | Axis-based | Uses recursive feature elimination,1 criteria to remove variables based on maintaining similar error rate to full model | Default |
| Altmann | [Altmann 2010 [18]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7508310/#R1) | *vita* | Test Based | Axis-based | Based on a parametric test of repeated permutations of importance measures | Default |
| Boruta | [Kursa 2010 [5]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7508310/#R19) | *Boruta* | Test Based | Axis-based | Based on a permutation test using a hold out approach for importance measures | Default |
| Menze | [Menze 2011](file:///C:\Users\jspeiser\AppData\Local\Microsoft\Windows\INetCache\Content.Outlook\538G6FYJ\Menze,%20Bjoern%20H.,%20et%20al.%20%22On%20oblique%20random%20forests.%22%20Machine%20Learning%20and%20Knowledge%20Discovery%20in%20Databases:%20European%20Conference,%20ECML%20PKDD%202011,%20Athens,%20Greece,%20September%205-9,%202011,%20Proceedings,%20P) | *aorsf* | Performance Based | Oblique | p-values are calculated for predictors at non-leaf nodes. Variable importance is based on the proportion of time a p-value for a predictor is <0.01. Recursive feature elimination is applied based on this importance metric.1 | Default  n\_predictor\_min = 2 |
| RRF | [Deng 2013 [11]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7508310/#R9) | *RRF* | Performance Based | Axis-based | Based on a regularized random forest, which uses forward selection to sequentially add variables until there is no further information gain | Default |
| SRC | [Ishwaran 2014 [10]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7508310/#R15) | *randomForestSRC* | Performance Based | Axis-based | Calculates the minimal depth of the maximal subtree (i.e., the largest subtree whose root node splits on the predictor) for each predictor. | Default |
| VSURF | [Genuer 2015 [7]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7508310/#R12) | *VSURF* | Performance Based | Axis-based | Three step variable selection procedure. The "thresholding step" removes irrelevant variables. The "interpretation step" aims to select all variables related to the response for interpretation purpose. The "prediction step" refines the selection by eliminating redundancy in the set of variables selected by the second step. | Default |
| Negation | [Jaeger 2023](https://www.tandfonline.com/doi/full/10.1080/10618600.2023.2231048) | *aorsf* | Performance Based | Oblique | Variable importance is based on prediction accuracy after negating the signs of all coefficients linked to a variable. Recursive feature elimination is applied based on this importance metric.1 | Default |
| Permutation Oblique | [Jaeger 2023](https://www.tandfonline.com/doi/full/10.1080/10618600.2023.2231048) | *aorsf* | Test Based | Oblique | Variable importance is based on prediction accuracy after permuting values for a given variable. Recursive feature elimination is applied based on this importance metric.1 | Default |
| Permutation Axis |  | *ranger* | Test Based | Axis-based | Selects all variables with variable importance > 0 based on permutation. | Default |
| Hapfelmeier | [Hapfelmeier 2023 [15]](https://www.sciencedirect.com/science/article/pii/S0167947322002699) | *rfvimptest* | Test Based | Conditional Inference | Similar to Altmann, but uses unbiased importance measures | # type = SAPT |

1Recursive feature elimination is a stepwise procedure where a single predictor is dropped at each step until a stopping criterion is met. At each step, out-of-bag prediction accuracy is computed. The set of variables that maximize out-of-bag prediction accuracy is determined as the final selection.

**Figure 1:** Characteristics of datasets used for the study



**Table 2**: Distribution of R-Squared, Computation Time, and Percent Variable Reduction for variable selection procedures

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **R-Squared (Axis)** | **R-Squared (Oblique)** | **Variable Percent Reduced** | **Time (seconds)** |
| **Altman** | | | | |
| Mean (SD) | 0.41 ( 0.017 ) | 0.43 (0.018) | 0.75 ( 0.012 ) | 338.02 ( 37.636 ) |
| Median [IQR] | 0.39 [ 0.183 , 0.626 ] | 0.44 [0.223, 0.651] | 0.82 [ 0.667 , 0.9 ] | 52.63 [ 19.956 , 379.291 ] |
| **Menze** | | | | |
| Mean (SD) | 0.42 ( 0.018 ) | 0.45 (0.018) | 0.61 ( 0.015 ) | 17.8 ( 2.427 ) |
| Median [IQR] | 0.42 [ 0.212 , 0.693 ] | 0.47 [0.233, 0.724] | 0.67 [ 0.439 , 0.812 ] | 3.57 [ 1.198 , 12.561 ] |
| **Permute -** | | | | |
| **Oblique** | | | | |
| Mean (SD) | 0.42 ( 0.019 ) | 0.45 (0.019) | 0.56 ( 0.016 ) | 42.67 ( 6.464 ) |
| Median [IQR] | 0.4 [ 0.22 , 0.694 ] | 0.48 [0.236, 0.724] | 0.6 [ 0.364 , 0.8 ] | 4.91 [ 1.145 , 24.119 ] |
| **Boruta** | | | | |
| Mean (SD) | 0.41 ( 0.019 ) | 0.43 (0.018) | 0.46 ( 0.019 ) | 40.58 ( 4.363 ) |
| Median [IQR] | 0.42 [ 0.208 , 0.677 ] | 0.44 [0.225, 0.673] | 0.45 [ 0.133 , 0.774 ] | 8.35 [ 3.519 , 47.411 ] |
| **CARET** | | | | |
| Mean (SD) | 0.43 ( 0.019 ) | 0.44 (0.019) | 0.48 ( 0.02 ) | 3544.69 ( 580.392 ) |
| Median [IQR] | 0.42 [ 0.23 , 0.694 ] | 0.44 [0.218, 0.681] | 0.5 [ 0.111 , 0.826 ] | 171.95 [ 52.846 , 1170.799 ] |
| **Hapfelmeier** | | | | |
| Mean (SD) | 0.18 ( 0.017 ) | 0.21 (0.017) | 0.92 ( 0.002 ) | 1068.56 ( 151.485 ) |
| Median [IQR] | 0.08 [ -0.005 , 0.32 ] | 0.1 [0, 0.401] | 0.93 [ 0.9 , 0.95 ] | 143.83 [ 35.449 , 563.047 ] |
| **Jiang** | | | | |
| Mean (SD) | 0.42 ( 0.019 ) | 0.44 (0.019) | 0.65 ( 0.015 ) | 455.36 ( 63.503 ) |
| Median [IQR] | 0.42 [ 0.22 , 0.693 ] | 0.44 [0.232, 0.72] | 0.7 [ 0.474 , 0.872 ] | 41.57 [ 14.373 , 260.241 ] |
| **Min Depth** | | | | |
| **Medium** | | | | |
| Mean (SD) | 0.41 ( 0.018 ) | 0.41 (0.018) | 0.35 ( 0.019 ) | 30.08 ( 1.16 ) |
| Median [IQR] | 0.39 [ 0.199 , 0.666 ] | 0.42 [0.143, 0.634] | 0.27 [ 0 , 0.667 ] | 20.87 [ 13.91 , 47.381 ] |
| **Negation** | | | | |
| Mean (SD) | 0.41 ( 0.019 ) | 0.44 (0.019) | 0.53 ( 0.016 ) | 44.32 ( 7.204 ) |
| Median [IQR] | 0.4 [ 0.182 , 0.671 ] | 0.41 [0.188, 0.726] | 0.58 [ 0.325 , 0.776 ] | 5.06 [ 1.166 , 20.333 ] |
| **None** | | | | |
| Mean (SD) | 0.4 ( 0.017 ) | 0.4 (0.017) | 0 ( 0 ) | 0 ( 0 ) |
| Median [IQR] | 0.4 [ 0.182 , 0.621 ] | 0.4 [0.137, 0.635] | 0 [ 0 , 0 ] | 0 [ 0 , 0 ] |
| **Permute -** | | | | |
| **Axis** | | | | |
| Mean (SD) | 0.4 ( 0.018 ) | 0.4 (0.017) | 0.13 ( 0.01 ) | 0.37 ( 0.032 ) |
| Median [IQR] | 0.39 [ 0.196 , 0.624 ] | 0.4 [0.161, 0.633] | 0.05 [ 0 , 0.243 ] | 0.08 [ 0.041 , 0.373 ] |
| **RRF** | | | | |
| Mean (SD) | 0.4 ( 0.017 ) | 0.4 (0.017) | 0.02 ( 0.004 ) | 2.43 ( 0.256 ) |
| Median [IQR] | 0.4 [ 0.17 , 0.623 ] | 0.39 [0.137, 0.633] | 0 [ 0 , 0 ] | 0.6 [ 0.152 , 2.735 ] |
| **Svetnik** | | | | |
| Mean (SD) | 0.41 ( 0.018 ) | 0.43 (0.018) | 0.69 ( 0.016 ) | 1197.21 ( 134.788 ) |
| Median [IQR] | 0.37 [ 0.181 , 0.642 ] | 0.4 [0.204, 0.68] | 0.76 [ 0.577 , 0.904 ] | 245.47 [ 76.233 , 1317.819 ] |
| **VSURF** | | | | |
| Mean (SD) | 0.43 ( 0.018 ) | 0.43 (0.019) | 0.76 ( 0.014 ) | 256.19 ( 41.346 ) |
| Median [IQR] | 0.41 [ 0.205 , 0.69 ] | 0.43 [0.216, 0.714] | 0.84 [ 0.707 , 0.918 ] | 23.71 [ 8.851 , 125.959 ] |

**Figure 2: Distributions of Log computation time and Standardized Log-computation Time**

****

**Figure 3: Distributions of Percent Reduction (Absolute and Standardized)**

****

**Figure 4: R-Squared for Axis Forests Variable Selection Method**



**Figure 5:** **R-Squared for Oblique Forests Variable Selection Method**



**Figure 6: Comparison of Mean and Median R-Squared by Forest type and Method of Variable Selection**

![A screenshot of a graph

Description automatically generated]()

**Figure 7: Comparison of Accuracy by Time and Percent Reduction**

****

****

**Table 3: Instances when no variables were selected**

|  | | |
| --- | --- | --- |
| Method | # of Time Occurred | # of Datasets Occurred In |
| Altman | 13 | 5 |
| Hapfelmeier | 94 | 42 |
| VSURF | 9 | 4 |

**Figure 8: Comparison of Accuracy by Time by Percent Reduction (Complete Case)**





**Figure 9: Median R-Squared by Variable Selection type based on High vs Low P:N**

**P:N >0.10**

A screen shot of a graph

Description automatically generated

**P:N <0.10**

A graph with numbers and lines

Description automatically generated