Behavior Research Methods (2023) 55:2447–2466
https://doi.org/10.3758/s13428-022-01901-9

Common, uncommon, and novel applications of random forest in psychological research

Dustin A. Fife[1] · **Juliana D'Onofrio**[1]

Accepted: 5 June 2022 / Published online: 1 August 2022

Abstract

Recent reform efforts have pushed toward a better understanding of the distinction between exploratory and confirmatory research, and appropriate use of each. As some utilize more exploratory tools, it may be tempting to employ multiple linear regression models. In this paper, we advocate for the use of random forest (RF) models. RF is able to obtain better predictive performance than traditional regression, while also inherently protecting against overfitting as well as detecting nonlinear effects and interactions among predictors. Given the advantages of RF compared to other statistical procedures, it is a tool commonly used within a plethora of industries, including stock trading, banking, pharmaceuticals, and patient healthcare planning. However, we find RF is used within the field of psychology comparatively less frequently. In the current paper, we advocate for RF as an important statistical tool within the context of behavioral and psychological research. In hopes of increasing the use of RF in the field of psychology, we provide information pertaining to the limitations one might confront in using RF and how to overcome such limitations. Moreover, we discuss various methods for how to optimally utilize RF with psychological data, such as nonparametric modeling, interaction and nonlinearity detection, variable selection, prediction and classification modeling, and assessing parameters of Monte Carlo simulations. Throughout, we illustrate the use of RF with visualization strategies, aimed to make RF models more comprehensible and intuitive.

Keywords Prediction · Classification · Variable importance · Multiple regression

There is little doubt the research landscape is rapidly changing. Many researchers are pushing for greater transparency and open science practices in general (Munafò et al., 2017; Nosek, Ebersole, DeHaven, & Mellor, 2018). There has been a corresponding call for more meaningful statistics (Cumming, 2014; Kruschke & Liddell, 2018), cumulative approaches to research (Cumming, 2014; Schmidt & Oh, 2016), and greater reliance on data visualization (Fife, 2020; Fife & Rodgers, 2021; Tay, Parrigon, Huang, & LeBreton, 2016). While some push for stricter standards of research practices (Nelson, Simmons, & Simonsohn, 2018), others (e.g., Fife & Rodgers, 2021) advocate that we broaden our perspective on what constitutes scientific research to include greater use of exploratory data analysis (EDA).

EDA is a data analytic philosophy that emphasizes "listening" to one's data. (Tukey, 1986), the father and a vocal advocate for EDA, suggested that confirmatory research is akin to a prosecutor that places hypotheses on trial. EDA, on the other hand, is more like a detective, hunting for clues and letting the evidence speak for itself. As such, while confirmatory research is hypothesis-driven, EDA tends to be hypothesis-generating.

There are several reasons to expand the use of EDA. First, few applied researchers are actually ready to conduct confirmatory research, since it requires a detailed analysis plan that fully anticipates all analytic strategies without any deviation. In the words of McArdle, "… it can be said that exploratory analyses predominate our actual research activities. To be more extreme, we can assert there is actually no such thing as a true confirmatory analysis of data, nor should there be" (McArdle, 2012, p. 405).

Second, while most researchers have EDA intentions, some might mistakenly utilize CDA tools (Fife & Rodgers, 2021).[1] Fife & Rodgers, 2021 distinguished between "tools"

✉ Dustin A. Fife
fife@rowan.edu

[1] Rowan University, Glassboro, NJ, United States

[1] We don't mean to imply that multiple regression is the *only* appropriate tool for those with CDA intentions, while those with EDA intentions can *only* use random forests. What we mean is that the

and "intentions." There are many tools that were developed under the EDA paradigm that may have a place in CDA. For example, residual analyses were developed under the EDA paradigm (as a way to look for patterns the researcher failed to model), but can be used with CDA analyses as well (e.g., to verify/demonstrate the researcher met assumptions). On the other hand, CDA tools (e.g., hypothesis tests) are probably not appropriate for analyses with EDA intentions since their probability distributions rely on many assumptions not likely to be met when doing EDA (e.g., the sample size is planned in advance, multiple tests are corrected for multiple comparisons).

For example, one frequently misused and abused CDA tool is multiple regression. Researchers routinely perform multiple tests of significance with dozens of variables to identify a hypothesis that is supported by the data. However, the p values associated with these significance tests have no probabilistic meaning without protections in place (e.g., corrections for multiple comparisons, adherence to distributional assumptions, (Cramer et al., 2016; Fife & Rodgers, 2021)). Even with these protections, multiple regression is a poor tool to use for exploratory research, particularly when multiple variables are involved, mostly due to its tendency to capitalize on chance (see, e.g., McNeish, 2015).

As the research landscape evolves, we hope it becomes more friendly toward EDA analyses, and more receptive to EDA tools as well. One such tool is random forest (RF), a machine learning algorithm well equipped to handle the shortcomings of multiple regression. RF has several advantages over traditional approaches. First, RF models reduce (but do not eliminate, (Gashler, Giraud-Carrier, & Martinez, 2008; Segal, 2004)overfitting (Breiman, 2001). Second, these models and the classification (and decision trees upon which they are based) are nonparametric, providing more flexibility when statistical assumptions are untenable (Malley, Kruppa, Dasgupta, Malley, & Ziegler, 2012; Steinberg & Colla, 1995). Third, RF natively detects interaction and nonlinear effects without requiring the user to explicitly model these relationships (Ryo & Rillig, 2017; Touw et al., 2013). Finally, RF can be used in situations where the number of variables far exceeds the number of subjects (Breiman, 2001; Matsuki, Kuperman, & Van Dyke, 2016)[2]

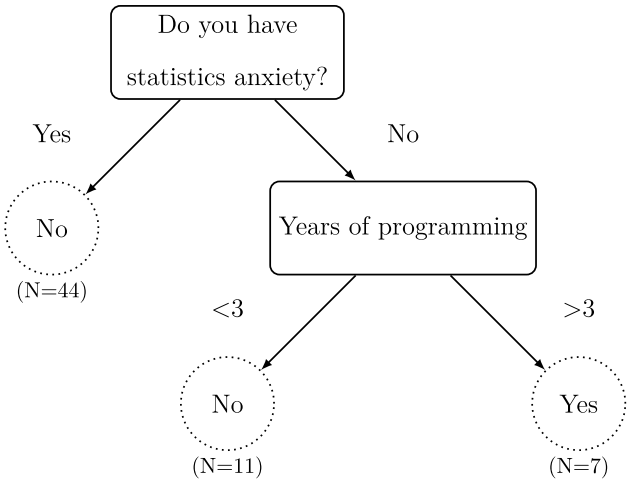

Fig. 1 Example of a decision tree. This fictitious decision tree attempts to predict whether someone will use random forest based on their statistics anxiety and years of programming. *Rectangles* are called "nodes." The *dotted circles* are the predictions of the model with the sample size of those meeting each condition listed underneath the fitted predictions

In this paper, we hope to accomplish multiple goals. First, we describe how RF works and why it offers several advantages over traditional statistical models. Second, we discuss its strengths and limitations as well as address common misconceptions. Finally, we conclude by discussing several novel and/or underutilized applications of RF in psychology, in hopes to guide researchers in how best to utilize this statistical tool.

How random forest algorithms work

In this section, we intentionally provide a nontechnical overview of the RF procedure in an attempt to make RF less mystical and/or daunting. In addition, we have found that one need not understand the technical nuances of RF in order to capitalize on its strengths and use it to make interesting discoveries. For those more interested in a more technical treatment of RF, see (Strobl, Malley, & Tutz, 2009), as well as Chapter 8 of (James, Witten, Hastie, & Tibshirani, 2013).

Decision trees

The basic unit of analysis for RF models is a decision tree. Figure 1 shows an example of a decision tree, which is aimed

Footnote 1 (continued)

arsenal of multiple regression tools (e.g., p values, power calculation formulae, parametric properties) are probably better equipped to handle CDA intentions. Likewise, the tools associated with RF (e.g., bootstrapped sampling of individuals, sampling of variables, internal cross-validation, nonparametric) are better equipped to handle EDA intentions. Presumably, one can choose an a priori hypothesis with RF, but one cannot make probabilistic inferences or plan appropriate samples sizes, which are requirements of "strict CDA" (Fife & Rodgers, 2021). Likewise, one *can* use multiple regression when their intentions are exploratory (for example, by using stepwise regression), but the EDA tools for RF are so much better.

[2] This is often called the "$n < p$" problem, where n is the number of participants and p is the number of variables. If this occurs, traditional statistical models cannot be estimated.

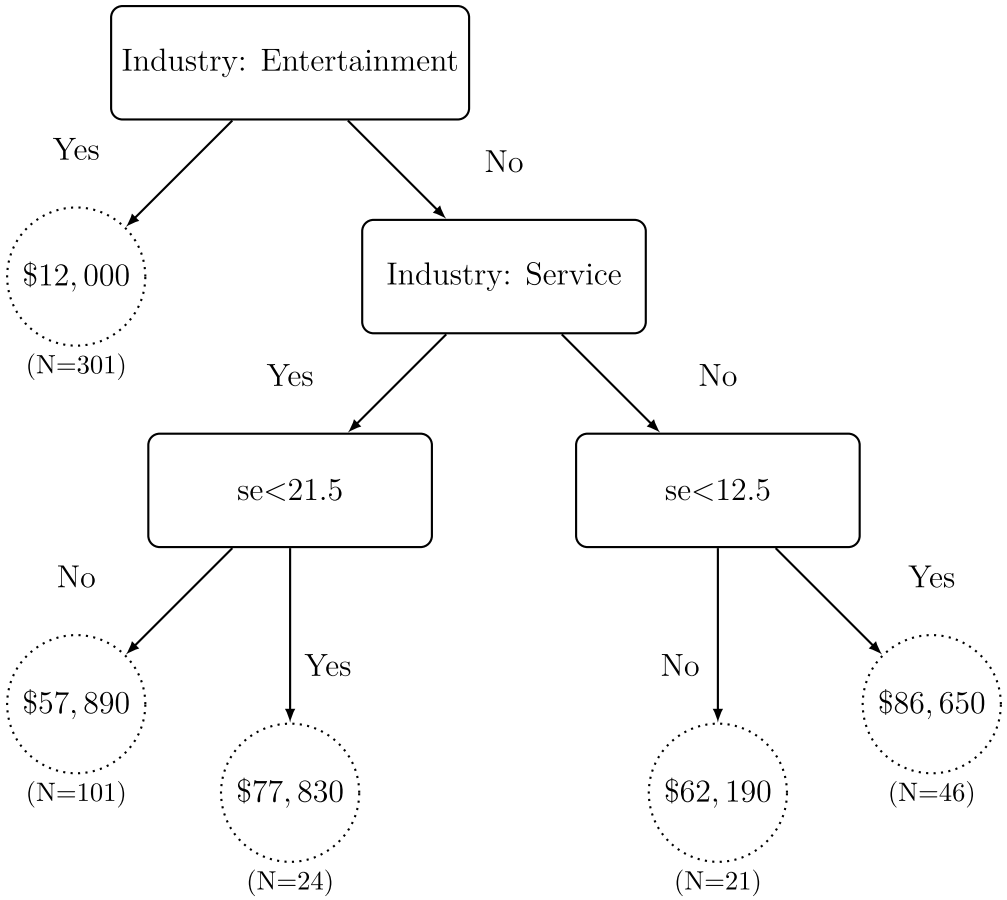

Fig. 2 Another example of a decision tree, though this example includes a continuous outcome variable (income). This fictitious decision tree attempts to predict someone's income from their industry (entertainment, science and technology, and service) and their self-efficacy score. As before, *dotted circles* represent the prediction and the *numbers below the circles* indicate how many individuals matched those conditions

at determining whether an individual will choose to use an RF model for their personal research. The top box, or "node" asks whether the individual has statistics anxiety. For those who do, we predict they will not use RF. Model predictions are indicated by dotted circles, with the number of individuals who meet those conditions written below the dashed circles (e.g., 44 individuals reported having statistics anxiety). The second node asks how many years of experience the individual has in computer programming. For those with more than 3 years of experience, the model predicts they will use RF for their research. For those who do not, the model predicts they will not. Those nodes early in the decision tree (e.g., statistics anxiety) are called "branching nodes," or "internal nodes," while those nodes that do not branch into other nodes (e.g., years of programming) are called "terminal nodes," or "leaf nodes." (Note: the predictions, indicated by circles, are not considered nodes).

Decision trees have been used for decades to model statistical relationships. When entered into a computer algorithm, the computer will decide, algorithmically, where in the tree the nodes fall and also determine the optimal cutoff (e.g., 3 versus 8 years of programming experience). From this single decision tree, the computer generates predictions. In this example, those with more than 3 years of experience in computer programming and who do not report anxiety about statistics will use RF, while those who do not meet both conditions will not. As with any other statistical model, we can identify how well we classify individuals. Specifically, when we speak of classification accuracy at the node level, we call it "node purity."[3]

This example can be extended to continuous outcomes. Suppose one wanted to predict an individual's income based on the following variables: a) their type of industry (science and technology, service, or entertainment), and b) their level of self-efficacy. As before, a computer can optimally compute the cutoffs for each variable and levels within each

[3] Accuracy can be determined both at the node level (which is called "node purity") and at the forest level (which is called "prediction accuracy").

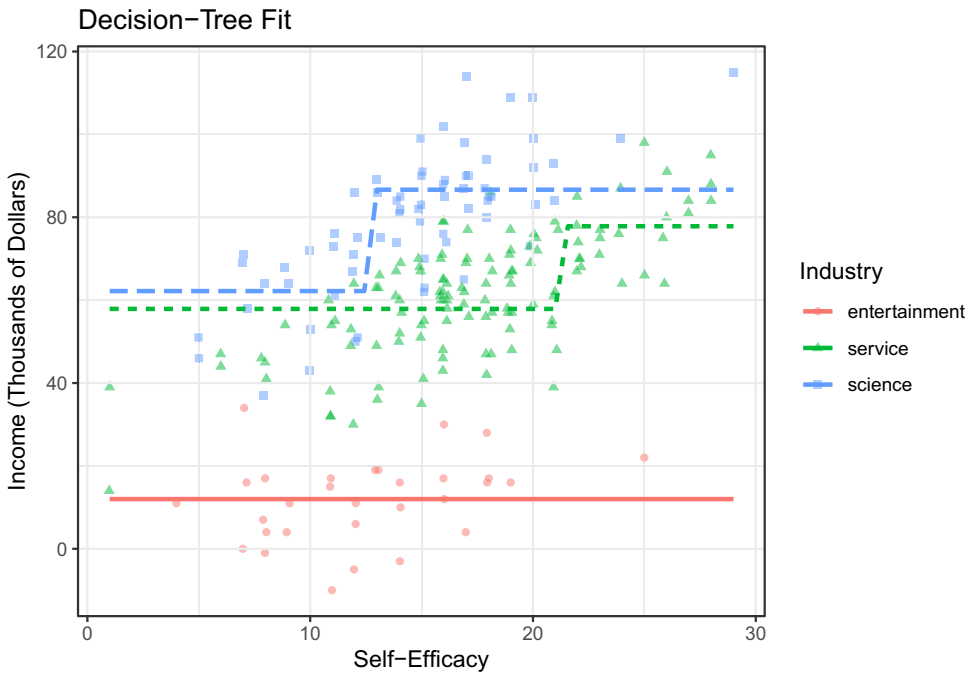

Fig. 3 This plot shows raw data and predictions from the decision tree shown in Fig. 2. The lines visually reflect the predictions from the decision tree. These lines are nonlinear and suggest an interaction, even though the model never explicitly modeled nonlinear or interaction effects

variable.[4] However, with continuous outcomes, the predictions for each individual are no longer binary, but rather individuals are assigned a value as a prediction (e.g., an income of $63,690). Prediction accuracy might then be evaluated using sum of squared residuals as is done in a simple regression model. Also similar to regression, the fit of the model can be visualized with scatterplots. Figure 2 shows the decision tree of this model, while Fig. 3 shows the corresponding fit of the model overlaid on a scatterplot. Notice that the fit of the model is nonlinear, and that the decision-tree allows the fit of the model to "bend" with the data, as appropriate. Likewise, notice how the different prediction lines suggest an interaction effect. Specifically, self-efficacy increases with income for the science and technology, and service industries, but not for entertainment. Although we never explicitly asked the decision-tree to model an interaction, it was detected through analysis. The tree simply and innately generated different predictions for different combinations of the variables.

Random forest

RF utilizes multiple decision trees, often in the hundreds, if not thousands, hence the name random *forest.* Each tree in the forest will utilize a different and randomly selected set of observations and predictor variables, which is where the term "random" comes from. RF will randomly sample participants. By default, it samples with replacement (i.e., the sample is "boostrapped"), though one can easily sample without replacement. Typically, the sample size is set to 67% of the entire sample. This 67% sample is used to calibrate the decision tree. The remaining 33% are reserved for cross-validation in what is called the "out of bag" (OOB) sample (Breiman, 2001), each OOB observation is passed through each tree to generate predictions. The prediction accuracy of the model is then evaluated based on the OOB sample. RF has repeatedly been shown to outperform regression models (both logistic and standard regression, (Couronné, Probst, & Boulesteix, 2018; Kirasich, Smith, & Sadler, 2018; Muchlinski, Siroky, He, & Kocher, 2016)).[5] Although the algorithm

[4] It may seem odd that Fig. 2 shows different cutoff values for se (21.5 vs. 12.5). However, remember that these cutoffs are for different industries (service versus science). There is no reason to suspect the optimal cutoff of se will be identical for both industries.

[5] As we've previously mentioned, RF models reduce (Gashler et al., 2008; Segal, 2004), but do not eliminate overfitting. This reduction in overfitting is due to the fact that RF models are an "ensemble" method. Ensemble methods combine (aggregate) multiple base models in an attempt to separate the noise from the signal. Also, because RF models internally cross-validate, they can yield an estimate of the cross-validation accuracy. This cross-validation accuracy occurs at the *tree* level. It does not natively cross-validate at the *forest* level. To adequately gauge overfitting would likely require a fresh dataset. Alternatively, one could split the data prior to beginning RF building, build a forest with a training set, and then validate the entire forest on the reserved validation set. Throughout this manuscript, our discussions of OOB and overfitting are referring to the cross-validation that occurs at the tree level.

defaults to a sample size of 67%, this can be modified to include more or less of the sample.

In addition to random sampling individuals, RF will also randomly sample variables. By default, RF generally samples $\sqrt{m}$ variables when the outcome is categorical and $m/3$ when the outcome is numeric, where m is the total number of variables. For example, suppose one has five predictor variables of income: industry, self-efficacy, years of education, parental socioeconomic status (SES), and geographic region. In this case, m, the total number of variables equals five and the algorithm will sample $\sqrt{5} \approx 2$ variables. So, the first tree might sample self-efficacy and parental SES, while the second tree might sample parental SES and geographic region. (This random sampling of variables technically happens at the node level, not at the tree level). The number of variables selected can also be modified to sample more or less variables.

Once variables have been sampled, the algorithm then builds a decision tree from this subset of predictors. As with regular decision trees, the computer determines the hierarchy of the tree, the optimal cutoff values for the node splits, and the predicted scores for each individual. Given each tree contains a different set of predictors, each tree may produce different predictions for each row in the dataset. Additionally, because each tree uses only a subset of the variables, there is never any concern about running out of degrees of freedom to test the model.[6]

Once RF constructs multiple decision trees, one can assess the "variable importance" (VI) of each predictor in the model. Various measures of VI exist, but each attempts to evaluate how the fit of the model is improved by the inclusion of each individual predictor. Once such a measure is the mean decrease in impurity (also known as the "Gini index"). This index simply compares the prediction accuracy of decision trees both before and after the inclusion of the predictor of interest. For example, if we omitted the "Years of programming" node in Fig. 1 and simply categorized all those without statistics anxiety as "Yes," 11 individuals who were classified as "No" would now be classified as "Yes." If we compare the accuracy of this prediction to the prediction where the model includes the "Years of programming" node, that would tell us how important "Years of programming" is in predicting the outcome. If we were then to do that with all variables across all trees (and weight this by the number of times that variable is used for splitting), that would give us the "mean decrease in impurity," or Gini index.

Table 1 Simulated dataset where the self-efficacy scores were permuted (or shuffled). Permutation breaks any association between the other variables (e.g., income, in this case) and the shuffled variable

Income	Self-efficacy (before permutation)	Self-efficacy (after permutation)
45	5	2
57	2	2
64	2	5
40	6	6
61	5	5

While computationally easy to assess, the mean decrease in impurity suffers from a major limitation. Specifically, variables with many possible values (e.g., continuous variables) have inflated estimates of VI relative to variables with few possible values (e.g., one's gender classification, (Strobl, Boulesteix, Zeileis, & Hothorn, 2007)). Put differently, two variables with identical predictive accuracy will have different VI estimates if one has more unique values.

An alternative measure of VI, called "permutation VI" does not suffer from the same bias. This measure is also called "mean decrease in accuracy," but we will call it "permutation VI" so as to not confuse it with mean decrease in impurity. This approach works by randomly shuffling OOB participants' scores for each node in a decision tree. For example, Table 1 shows an example where scores are shuffled. In this case, the first OOB person had a score of five, but after permutation was given the score associated with the second individual (a two). This shuffling removes any correlations between the variables in the dataset. The algorithm can then assess the accuracy of the model before versus after shuffling. If there is a large difference in OOB predictions before versus after shuffling the scores for a particular variable, we can conclude that variable has a strong association with the outcome.

For binary outcomes, the permuted VI score indicates the average change in OOB error (before versus after shuffling) across all trees. For example, if a variable's permutation VI is 0.3, that says that, relative to shuffled scores, the unshuffled scores had OOB scores lower by 30%. For continuous variables, the permutation VI score represents the difference in sum of squared errors between shuffled and unshuffled datasets.

[6] This statement requires some clarification. Degrees of freedom is defined as the number of observations (N) minus the number of parameters estimated. When the number of parameters estimated is equal to the sample size, our degrees of freedom will be zero *and* the model will fit perfectly. For example, suppose one had only two observations. If they were to try to fit a simple regression model, their model would fit a slope and an intercept (two parameters). Also, their model would fit their data perfectly (because the line will pass exactly through the two datapoints). Likewise, if one has three datapoints and attempts to fit a model with two predictors, it too will estimate three parameters (two slopes and an intercept) and will fit perfectly. Perfect fits are problematic for statistical tests because these tests compute a ratio of "signal" to "noise." When the model fits perfectly, there is no noise and this ratio requires division by zero, which is not possible. Consequently, there is no way to determine how well the model fits the data when degrees of freedom are zero. For more information, see (Rodgers, 2019).

Another advantage of the permutation VI measure is that missing data are handled naturally (Hapfelmeier, Hothorn, Ulm, & Strobl, 2012). If an individual is missing a score on a particular variable (say, years of education), the shuffling of scores will simply assign that missing value to another individual.

The disadvantage of permutation VI is that it is computationally expensive. The computer must shuffle scores for every variable, across every node, across every tree in which it appears. Yet this disadvantage becomes increasingly less frustrating as computers become more powerful.

These VI measures allow researchers to winnow down a large list of variables into a smaller subset of contenders for further exploration. This could be done ad hoc (e.g., by choosing to investigate the top three variables), or more concretely. For example, (Genuer, Poggi, & Tuleau-Malot, 2010) utilized mean decrease in impurity VI to develop an objective variable selection algorithm. This algorithm works by generating multiple *forests* so one can estimate variability in VI estimates, which can then be used to essentially determine which variables have VIs that exceed chance. This algorithm can be found in the VSURF package in R (Genuer, Poggi, & Tuleau-Malot, 2019).

In summary, RF models generate hundreds or thousands of decision trees to produce aggregated predictions, all the while natively detecting interactions and nonlinear patterns. Because these models utilize random sampling of variables, the models circumvent the $n < p$ problem and reduce overfitting. These characteristics, we will show, represent a serious advantage for doing many types of research.

Common misconceptions

In the previous section, we outlined several advantages of RF models. Given that psychological research frequently encounters nonlinear patterns (Hayes, Laurenceau, Feldman, Strauss, & Cardaciotto, 2007; Helmich et al., 2020; Lord & Novick, 1968; Mattei, 2014), violated assumptions (Micceri, 1989; Skidmore & Thompson, 2013; Van Horn et al., 2012), and interactions (Cronbach, 1975), it is somewhat surprising to see RF models so infrequently used. We suspect the reason for this is that researchers have common misconceptions about RF models.[7]

The first, and perhaps most common misconception, is that RF models are inappropriate when one is examining a small number of predictor variables. We suspect this misconception stems from the fact that RF was initially designed to handle very large numbers of predictor variables (e.g., more variables than observations as with genetic or biomedical research). While this is true, this does not mean RF models cannot be used with a small number of variables, especially when one wants to leverage RF's ability to detect interactions and/or nonlinear effects.

Another misconception we have encountered is that one must have a large sample size to utilize RF models. In reality, RF models do no worse than traditional models (e.g., regression) and likely do much better. Estimates of cross-validation accuracy (i.e., OOB error) will reflect the uncertainty associated with the smaller sample sizes. Granted, if sample sizes are small enough and/or the signal in the data are weak enough, these OOB estimates will report that the model struggles to predict the outcome of interest. In this sense, RF models do struggle with smaller sample sizes. However, we see this as a feature of RF, not a limitation. For example, in order to determine whether certain variables contribute to a prediction model, it is better for the model to admit difficulty (e.g., through imprecise predictions) than to capitalize on chance patterns as linear models might.[8]

While these common misconceptions can be easily dismissed, there are some genuine limitations of RF models. We will outline and address these limitations in the next section.

Limitations of RF models

RF models are "black box" algorithms

Single decision trees can be easily and intuitively interpreted visually if there are not too many nodes. However, the predictions of individual decision trees are highly unstable. Depending on the number of predictors within an analysis, the predictor variable chosen for the first branching node can be different across two single decision trees. Moreover, this

[7] Most of these misconceptions have come from personal experience of the authors. While we cannot pinpoint exact statements in the literature where people have stated misconceptions, we have encountered resistance from colleagues and reviewers when we have recommended RF models. In this section, we aim to address these misconceptions, which might be preventing the utilization of RF in psychological research.

[8] Sampling variability will affect RF models in different ways. Growing a small number of decision trees increases the variability of VI estimates ((Wang, Yang, & Luo, 2016)). Because it costs the researcher nothing other than computational time, it is generally recommended to grow large forests (e.g., at least 500 trees). On the other hand, having a small sample size cannot be fixed algorithmically. Under extreme cases, the OOB sample will be small and thus OOB/VI estimates will be highly unstable. Likewise, there may not be that many unique values from which to bootstrap. However, once again, this is a problem not unique to RF models. All statistical models will struggle with small sample sizes (unless one has strong priors in a Bayesian analysis. See e.g., Depaoli, 2013). However, splitting the dataset into training and testing might exacerbate the problem of small sample sizes, at least at the individual tree level. However, this problem can be partially offset by generating a large number of trees (Wang et al., 2016). Even still, RF models are preferred to models that do not include cross-validation.

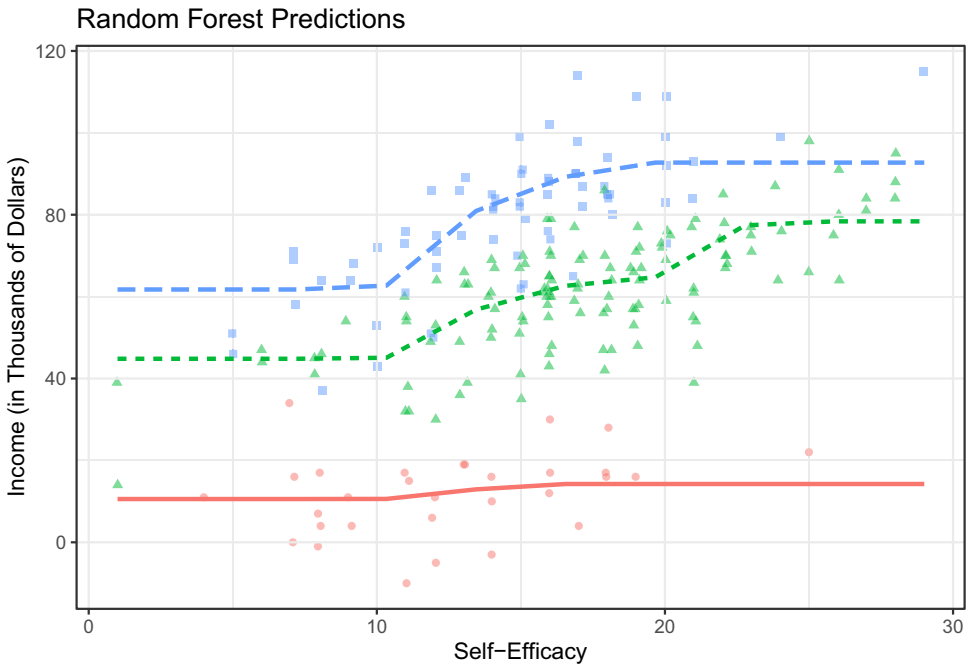

Fig. 4 This shows the same data displayed in Fig. 3, but the fits are from a RF model

difference likely alters the entire structure of the tree and the terminal node prediction between two decision trees might vary drastically (Strobl et al., 2009).

RF addresses this instability by aggregating terminal node predictions across hundreds or thousands of decision trees. Although this process is the very reason for the advantages of RF, limitations remain. Specifically, as RF is a computer-generated machine learning model, it is fundamentally a "black box" algorithm (Breiman, 2001), making it difficult to interpret the model itself. It would be unfeasible, for example, to visually display all decision trees generated from the random forest. More specifically, while regression models yield a simple algebraic equation following analysis, RF does not. In order to predict new observations, one would save the hundreds or thousands of single decision trees into a computer program and feed new data through the saved program.

However, while the fitted algorithm cannot be easily conceptualized, the output of the model can because users can visualize the *predictions* of a RF model. To do so, one could feed the RF model a new dataset containing a wide range of predictor values. RF will then use the forest for the new data to estimate predictions. Of course, many statistical programs will do this sort of prediction automatically. As an example, Fig. 4 displays the same data as shown in Fig. 3, but with the fits of a RF model instead of the fits of a single decision-tree.

One convenient tool for visualizing RF models is Flexplot (Fife, 2021), which is a software application available in R, JASP, and Jamovi. Flexplot is designed to visualize statistical models and has many native functions for visualizing RF models. Flexplot can easily visualize a few variables, though visualizing more than a few variables can be less intuitive. For guidance on visualizing more than a few variables, see (Fife, Longo, Correll, & Tremoulet, 2021) for the JASP version of Flexplot and (Fife, 2021) for the R version. Or, for a systematic strategy for visualizing multivariate data with Flexplot, see (Fife & Mendoza, 2021). An alternative to multivariate visualizations is to plot what are called "marginal" relationships. For example, one might use an added variable plot to visualize the relationship between self-efficacy and income, holding all other variables constant.

In addition to visualizing predictions, one could also conceptualize variable importance metrics. As mentioned previously, VI metrics are intuitive to interpret and do not require users to peek within the black box or to visualize multivariate relationships. For the previous dataset, self-efficacy has a VI of 28.584, which can be roughly interpreted to mean that, relative to a model that excludes industry, the model with industry is approximately $28,584 closer to predicting one's actual income. Additionally, the VI for self-efficacy is 13.544.

RF models lack distributional theory

A second disadvantage of RF is that these algorithms lack statistical distribution theory. While RF's nonparametric approach is advantageous when assumptions are violated, this also makes it difficult to make sophisticated statistical inferences. For example, one cannot derive a p value or a confidence interval from a population distribution. One can obtain substitutes with resampling procedures and from these substitutes derive confidence intervals or make other

sorts of inferences. However, these inferences are tied to the data at hand, though past research has shown that RF predictions have parametric characteristics and can often be used for statistical inferences (McAlexander & Mentch, 2020), including prediction intervals (Zhang, Zimmerman, Nettleton, & Nordman, 2019). Also, RF performs quite well in making inferences beyond the data (Fox et al., 2017; Gao, Wen, & Zhang, 2019; Lu et al., 2016).

Despite this limitation, we rarely see any reason to use RF as the final step in the research process as statistics are tools that allow us to make ever-more-precise mathematical statements about theory. For this reason, nonparametric models are simply "hacks"; they allow us to temporarily acquire an answer to a question in such a way that we don't deceive ourselves (e.g., by violating a statistical assumption). Arguably, the ultimate goal of research is to have precise mathematical parametric models, but often these nonparametric models are a necessary pit stop.[9] RF models, while rarely (if ever) the final destination of a theory's journey, they do assist in moving from imprecise nonparametric answers to specific parametric formulations. We will discuss examples of this process in later sections.

In short, RF models detect interactions, model nonparametric relationships, and they reduce overfitting. While they are considered a "black box" algorithm, their predictions can be easily visualized (e.g., by using Flexplot), particularly when one limits visualizations to only a few variables. Additionally, RF models are best considered a pit stop toward parametric modeling, rather than the final step in theoretical development.

Having covered the strengths, misconceptions, and limitations of RF models, we now turn to the core crux of our paper. As we hope to show, RF models can be used in novel and unique ways. In the following section, we identify a few strategies one might use to leverage the strengths and advantages of RF in psychological research.

Common, uncommon, and novel applications of RF models

The variety of applications of RF continues to expand in psychological research. In this section, we discuss a number of different strategies for using RF models. For each approach, we describe an overall strategy and demonstrate this strategy with applied examples.

Variable selection then parametric modeling

Perhaps the most common reason people use RF models is for variable selection. For example, RF models have been used in psychology to predict correlates of nonsuicidal self-injury (Ammerman, Jacobucci, & McCloskey, 2018), smoking behaviors (Kitsantas, Moore, & Sly, 2007), utilization of psychiatric services (Rossi, Amaddeo, Sandri, & Tansella, 2005), adherence to HIV testing (Pan, Liu, Metsch, & Feaster, 2017), and use of Internet-based psychotherapeutic treatment for depression and anxiety (Wallert et al., 2018).

Very often researchers are not necessarily, or at least presently, interested in developing theoretical explanations of psychological phenomena. Rather, researchers may wish to describe a phenomenon (Mõttus et al., 2020), or to winnow down a large list of candidate variables to a smaller subset of viable predictors. While algorithms exist for doing this in multiple regression (i.e., the various stepwise regression methods), these methods cross-validate poorly (Smith, 2018). Instead, RF can be used. The ensemble of decision trees improves cross-validation accuracy.

One limitation to this approach is that RF treats VI measures as if they represent the final stage of analysis. As we said previously, RF models are best considered as data analytic pit stops; they are powerful tools that offer valuable insight into how we might then utilize parametric models, particularly when paired with visualizations. In the following section, we illustrate how one might leverage VI measures as a means of gaining additional insights.

Example and overall strategy

When attempting to identify a small number of variables from a large set of candidates, we recommend the following strategy:

1. Enter all variables of interest into a RF model and compute variable importance
2. Sort the variables in terms of VI.
3. Select a small number of candidate variables to visualize (e.g., the top four variables as measured by VI).
4. Visualize the RF predictions for this small group of candidate variables. Visualizing multivariate data can be tricky, though we suggest (Fife & Mendoza, 2021) or (Fife, 2021) for simple multivariate visualization strategies.
5. Use the visuals from Step #4 to select appropriate variables for parametric modeling. The preceding step might suggest a certain variable is not helpful to the model,

[9] A similar statement can be made of models aimed to predict versus describe a phenomenon. Generally, classical statistical techniques (e.g., t-tests, ANOVAs, regressions) are used to describe associations, often as part of an effort to explain human behavior. Machine learning techniques, on the other hand, are more often concerned with prediction. However, we see prediction as a stepping-stone toward explanation, much like parametric models are a stepping-stone toward parametric models. In a sense, the maturity of one's science could be said to move from nonparametric/prediction models to parametric/explanatory models. See (Fife & Rodgers, 2021), as well as (Mõttus et al., 2020).

Table 2 Variable importance for the top three variables in the simulated suicide ideology dataset

Variable	Variable. Importance
socialsupport	1.80
age	1.27
depression	0.85

that two variables interact, or a nonlinear pattern exists. That step will guide the choice of parametric model.

When one uses this strategy, they retain many of the benefits of RF (i.e., cross-validation, native interaction detection, and native nonlinear detection), without the disadvantages (i.e., "black box" algorithm).

As an example, we simulated a dataset that contained the outcome variable of suicidal ideation and eight predictor variables: locus of control, depression, age, gender, parental income, grades, social support, and parental history of depression. The data were simulated in such a way that age had a nonlinear relationship with suicidal ideation and depression and social support had an interaction.

We began by computing VI for each of the predictor variables. Table 2 shows VI for the variables social support, age, depression, and locus of control. The next step was to plot these variables using Flexplot. This is an iterative process that may require dozens of plots to disentangle the relationships existent in the data. For example, one might choose to place social support on the *X*-axis in one plot and put depression/LOC in panels. Subsequently, the user might place depression on the *X*-axis and LOC/age in panels. Each visual represents a different "view" or "angle" of the multivariate relationship that might reveal different features of the relationship.

When visualizing these plots, the user is seeking to identify evidence of nonlinear patterns and/or interaction effects since these are most difficult to grasp without visuals. For those interested in understanding how to use multivariate visualizations to detect nonlinear/interaction effects, we recommend (Fife & Mendoza, 2021). The end result of this process yielded three distinct relationships illustrated in Fig. 5. The top plots show the interaction between depression and social support. Notice for those reporting lower levels of depression, there exists a weak relationship between social support and suicidal ideation. However, the relationship is stronger for those reporting more severe levels of depression.

The bottom-left plot shows the nonlinear relationship between suicidal ideation and age. Notice suicidal ideation increases from ages 12 to roughly 18, plateaus until age 22, and then trends downward. Finally, the bottom-right plot shows there is almost no relationship between locus of control and suicidal ideation. Since this had the smallest VI within the top four predictor variables, there is little reason to visualize the remaining variables in the dataset.

The visuals in Fig. 5 suggest the following parametric model:

$$\text{Suicide ideation} = \text{Age} + \text{Age}^2 + \text{Depression} + \text{Social support} + \text{Depression} \times \text{Social support}$$

This parametric model was fit to the data, then visualized in Fig. 6. The red lines show the fit of the regression model, while the blue lines show the fit of the RF model. The two predictions are quite similar, at least near the center of the data, which suggests the parametric model seems to capture the most important elements of the nonparametric model.

Nonparametric modeling

In order to tie statistical models to distributional properties, models make several key assumptions: normality, independence, constant variance, linearity, and homogeneity of regression. The latter two assumptions are particularly problematic for linear models and RF is well equipped to handle these assumptions.[10]

Detecting nonlinearity

Standard statistical models assume linear relationships between the predictors and the outcome. If one encounters a nonlinear relationship, linear models can be rigged to fit some limited nonlinear relationships (e.g., we can add a squared predictor to a linear model to get nonlinear predictions). However, if the appropriate function is not linear (e.g., exponential, logarithmic, logit), linear models will fail.[11] RF models, on the other hand, can fit patterns from any nonlinear relationship: exponential, logarithmic, logistic, polynomial, etc. When researchers begin analyses by visualizing RF models, they can then attempt to identify the appropriate nonlinear function and, if they wish, formalize the mathematical relationship.

[10] RF does not assume linearity or homogeneity of regression. Additionally, RF models don't assume normality or constant variance *per se*; the residuals can be distributed normally, as a Poisson, as an exponential, as a gamma, etc. However, when modeling continuous outcomes, a key component of the RF algorithm is the computation of sum of squared residuals (SSR). SSRs work quite well for symmetric distributions. However, for non-symmetric distributions (outlying data and/or skewness), RF's reliance on SSRs may bias estimates and there are alternative methods that can reduce this bias (Ghosal & Hooker, 2020).

[11] Sometimes one can fix some relationships (e.g., exponential or logarithmic relationship) with transformations. Others (e.g., logit) cannot be fixed with a simple transformation.

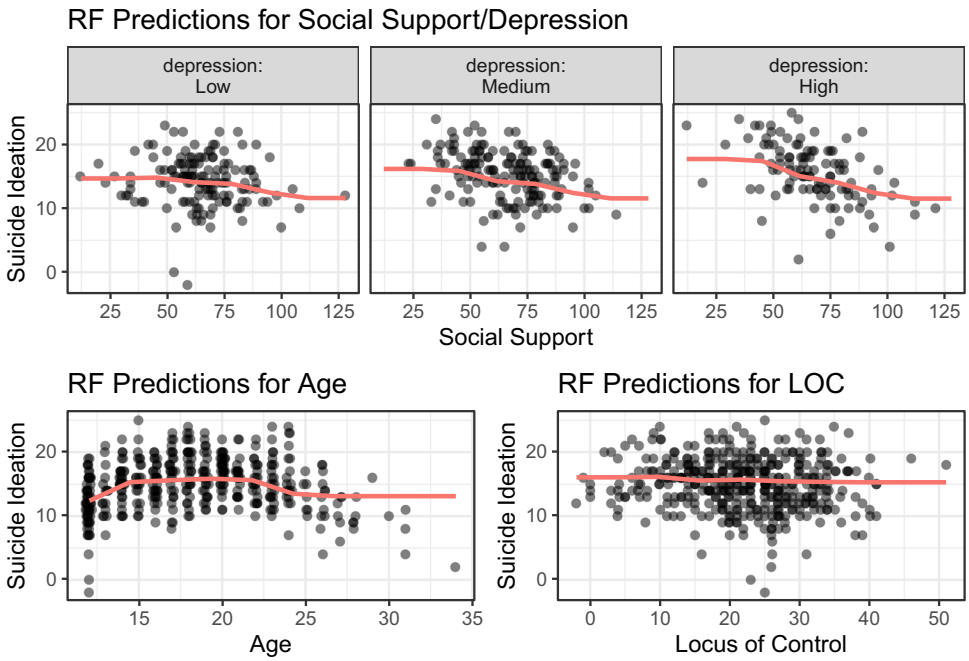

Fig. 5 Fits of the RF model for social support/depression (*top plot*), age (*bottom left*), and locus of control (*bottom right*)

Fife (2020) utilized this exact approach when modeling the relationship between mental illness and psychological distress. Upon visually inspecting RF predictions, Fife was able to identify an appropriate mathematical function, called the Michaelis–Menten or MM equation. The data were refit with a Bayesian MM model and results were replicated using the same MM model on an independent dataset (see video explaining this process at https://youtu.be/5BpmktmvgIA). Figure 7 shows the model fit with RF (left) and the MM equation (right).

To be clear, we reiterate that RF models are rarely an end unto themselves. Rather, they could serve an important step in helping researchers shift from nonparametric to parametric modeling. The results of a RF algorithm can help guide researchers in how to add theoretical and/or statistical precision to their models.

Interaction detection

Another important assumption of traditional statistical models is the assumption of homogeneity of regression slopes. This assumption states that if any interactions do exist between variables, they have been explicitly modeled, see (Fife and Mendoza, 2021; Gelman & Hill, 2006). If interactions exist that have *not* been modeled, estimates will be biased and conclusions gleaned will be misleading. (Gelman & Hill, 2006) noted that violating the assumption of homogeneity of regression (or, "additivity" in their terminology) is one of the most egregious violations.

As an example, suppose a researcher would like to perform a multiple regression. They might evaluate the main effect while also controlling for a number of different variables in order to determine the variance attributable to the main effect. However, as we mentioned previously, multiple regression assumes all "controlled" effects do not interact with the variable of interest. This is termed the "homogeneity of regression" assumption. When violated, any conclusions gleaned from main effects could be extremely misleading. For example, consider the image in Fig. 8, which shows the simulated results of a factorial ANOVA that contains a crossover interaction. If one were to estimate the main effect of treatment, one would conclude that there is *no* effect of treatment. This would be a misleading conclusion. Rather, the main effect of treatment depends on gender.

The homogeneity of regression assumption applies to all linear models (e.g., multiple regression, factorial ANOVA, ANCOVA) and these models are not robust to violations (Fife & Mendoza, 2021; Gelman & Hill, 2006), yet researchers routinely utilize these models without assessing the viability of this assumption (Fife & Mendoza, 2021). This lack of attention to this critical assumption is understandable; for even simple analyses, modeling interactions requires a large number of terms. For example, with only four variables, one would need to model 16 terms (one intercept, four main effects, six two-way interactions, four three-way interactions, and one four-way interaction). Few studies will have large enough sample sizes to precisely estimate the size of these effects. Clearly, it is much easier to assume a main

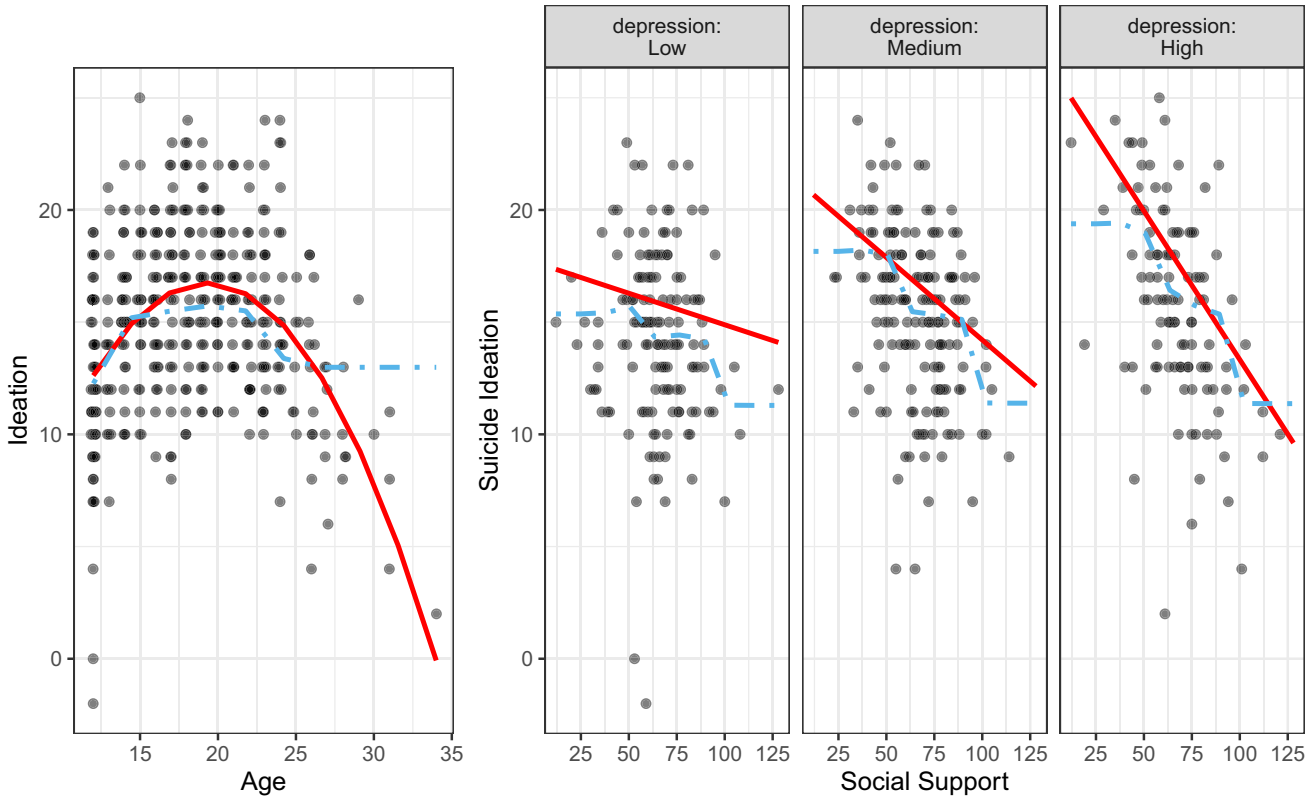

Fig. 6 Fits of a regression model (*red line*) and the RF model (*blue line*) for the age (*left plot*) and social support/depression (*right plot*) relationships

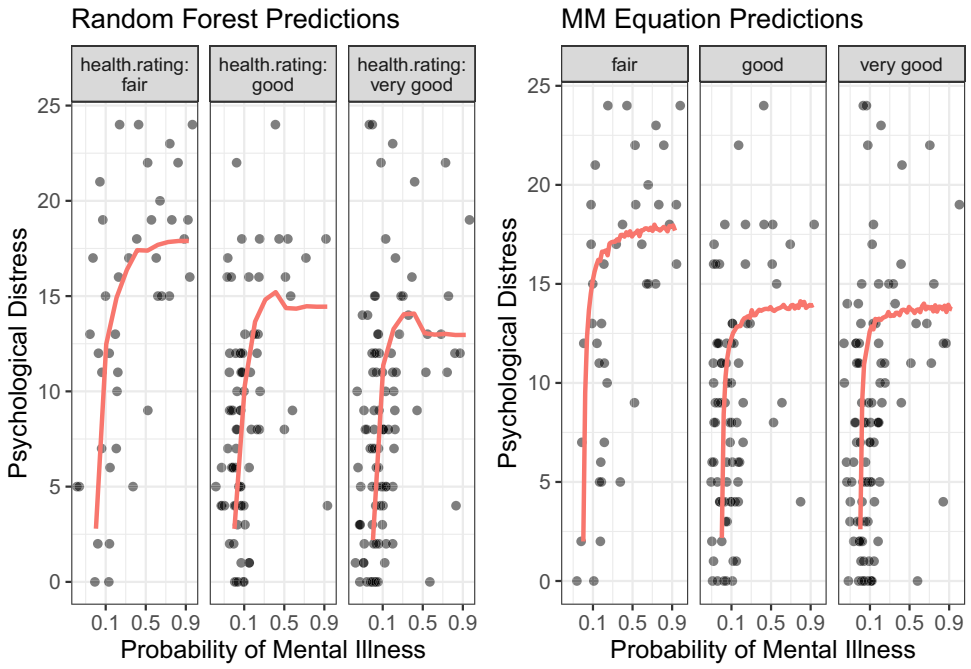

Fig. 7 Random forest predictions (*left*) and predictions from a model using the Michaelis–Menten equation (*right*) of psychological distress. (Fife, 2020) initially used a RF model to identify the nature of the MI/Distress relationship. Using that information, he subsequently fit the data using the MM equation

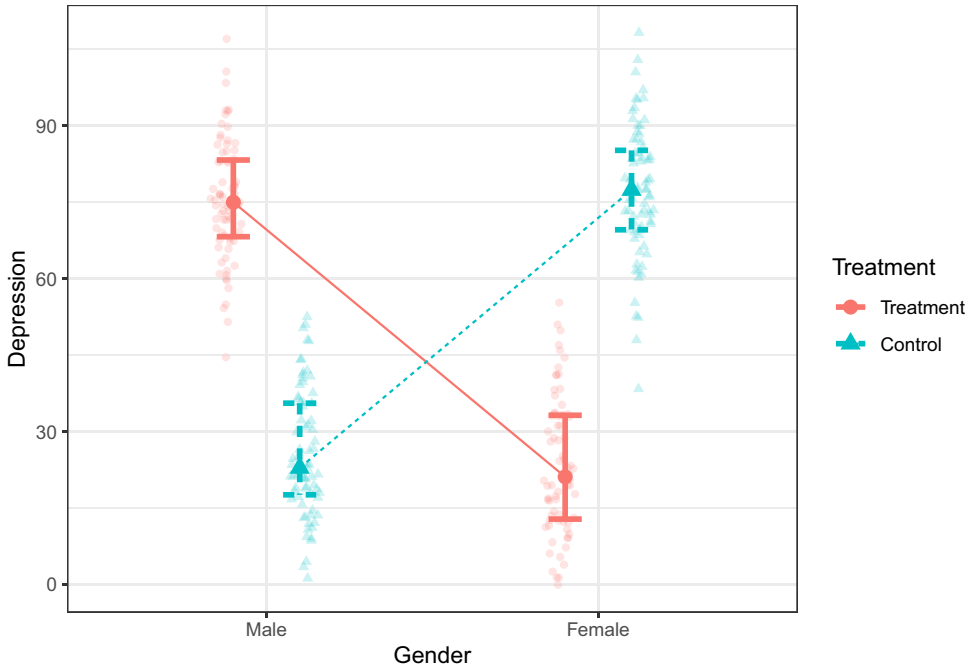

Fig. 8 A crossover interaction of the gender/treatment relationship. If one wished to estimate the effect of treatment on depression, after controlling for gender, the model would suggest there is no treatment effect. Clearly there is a treatment effect; it simply depends on gender. Standard statistical models assume no interactions (except for those that are explicitly modeled). RF models do not make this assumption

effect model and hope there are no interactions present, but the risks are great.

A better alternative is to utilize RF models. RF will natively detect interactions without explicitly modeling them. This saves effort on the part of the researcher, who would otherwise have to, by hand, choose which interactions should be estimated.

Additionally, in traditional regression, when all interaction terms are explicitly modeled, the researcher must then study large tables to determine whether each interaction is worth keeping in the model. The researcher could utilize *p* values to dismiss interaction terms, but because of multiple testing, these *p* values have no probabilistic meaning and are thus prone to bias. Researchers could instead utilize effect size measures (e.g., semi-partial R^2) to make decisions, but these are also prone to overfitting and extremely sensitive to multicollinearity. No matter the metric one uses, the process is time-consuming, difficult, and prone to capitalizing on chance. On the other hand, with RF models, the user need not decide from lengthy tables which terms to keep as the user can inspect VI metrics. If a variable is important, either as a main effect or as an interaction, it will be reflected in the VI metric.

Finally, estimating interactions in traditional regression is very difficult and requires very large sample sizes. Moreover, doing so requires degrees of freedom the researcher may be unable to spare. On the other hand, RF models do not "spend" (Mentch & Zhou, 2019; Rodgers, 2019) degrees of freedom to estimate interactions. Instead, modeling interactions is a natural part of the fitting process.

Overall, linear models are not robust to violations of the assumption of homogeneity of regression and evaluating the viability of this assumption is cumbersome and prone to imprecision. Taken together, we think RF should be the default method for modeling many multivariate relationships, particularly when more than two or three variables are used. Much like residual dependence plots are used for assessing homoscedasticity/linearity in regression, perhaps RF modeling can be a first step in evaluating the homogeneity of regression slopes assumption in multiple regression.

Unfortunately, using RF models for the purpose of interaction detection is neither well known nor common. However, (Kitsantas et al., 2007) utilized classification trees to identify whether a small set of predictors, including social risks, health risks, and peer smoking, interacted with one another in predicting intentions to smoke. They discovered that social and health risks were highly dependent on peer smoking behavior.

Example analysis

We will briefly illustrate this strategy with another simulated dataset. Suppose one were interested in modeling the efficacy of a smoking prevention program on adolescents' intentions to smoke while controlling for peer and parent smoking. Further suppose one simply fit the model without

Table 3 ANOVA summary table of a simulated dataset investigating the effect of treatment on intentions to smoke, after controlling for peer and parent Smoking

	DF	SS	MS	F	*p*
Peer	1	52.87	52.87	195.06	< .001
Parent	1	1.00	1.00	3.68	.055
Treatment	1	0.00	0.00	0	.956
Residuals	896	242.87	0.27		NA

assessing the homogeneity of regression assumption. Table 3 shows an ANOVA summary table of this model, which shows the treatment effect was ineffective at reducing intentions to smoke. However, suppose we modeled the data using a RF algorithm.

Figure 9 outputs predictions from this model. The RF model suggests that, without treatment, adolescents increase their intentions to smoke the more their peers smoke. For the treatment group, on the other hand, peer influence seems to have no effect on intentions to reduce smoking. In other words, the model that (incorrectly) assumed linearity and heterogeneity of regression committed a serious type II error.

As we mentioned previously, RF models best serve as a guide for determining what sort of parametric model is best. For this particular dataset, the RF model suggests we might add a quadratic term to the peer effect as well as an interaction. If we did this, we might plot the fits of the modified regression model, as in Fig. 10.

Assessing parameters of a Monte Carlo

This final strategy is more on the technical side and will likely only be of interest to statisticians. This application uses RF models to identify important parameters in a Monte Carlo simulation. Monte Carlo simulations are powerful techniques often used by statisticians to identify how various statistical procedures perform under different conditions. For example, one might wish to identify how nonlinear parameters bias R^2 values. To do so, a researcher might perform a Monte Carlo simulation, where the researcher simulates different numeric conditions, such as degree of nonlinearity (e.g., by modifying the beta weight of the quadratic term), size of the linear component (e.g., by modifying the beta weight for the linear term), the sample size, and the number of covariates.

When varying these parameters, it is common to pick a few values that span a reasonable range; for example, specifying the degree of nonlinearity as standardized regression weights of either -0.6, -0.3, 0, +0.3, or +0.6. Subsequently, researchers might report results via a large table that shows average bias under every possible condition. Alternatively, some might build an ANOVA that estimates the mean type I error rate from each of the levels of the parameter values,

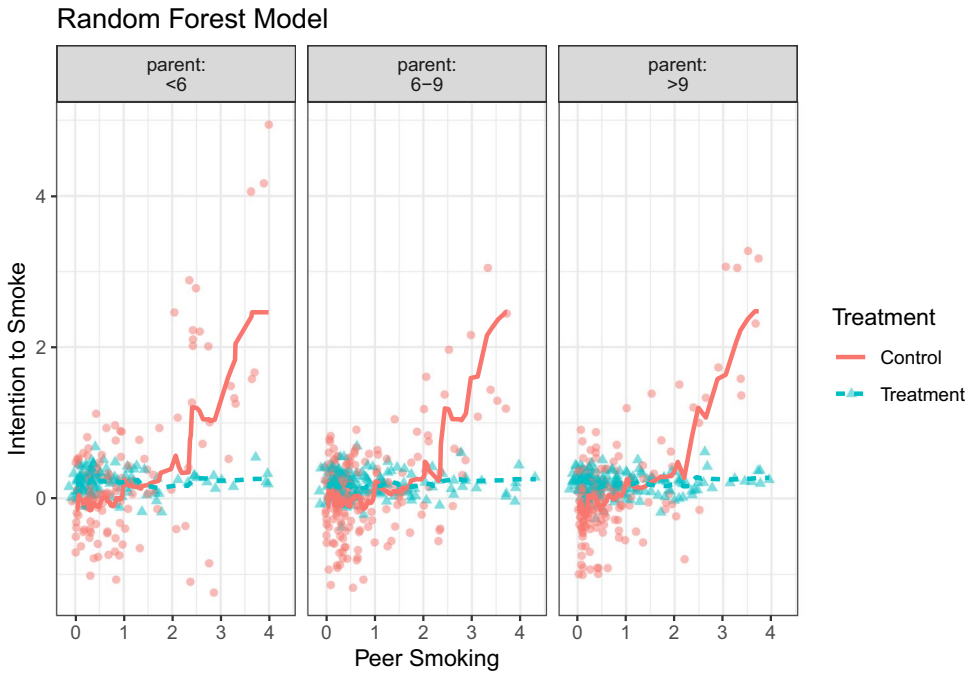

Fig. 9 This plot shows the fits from a RF model that predicts smoking intentions from peer smoking (*X*-axis), parent smoking (*panels*), and treatment condition (*colors*). The RF model picks up on the nonlinear effect of peer on intentions, as well as the interaction between treatment and peer

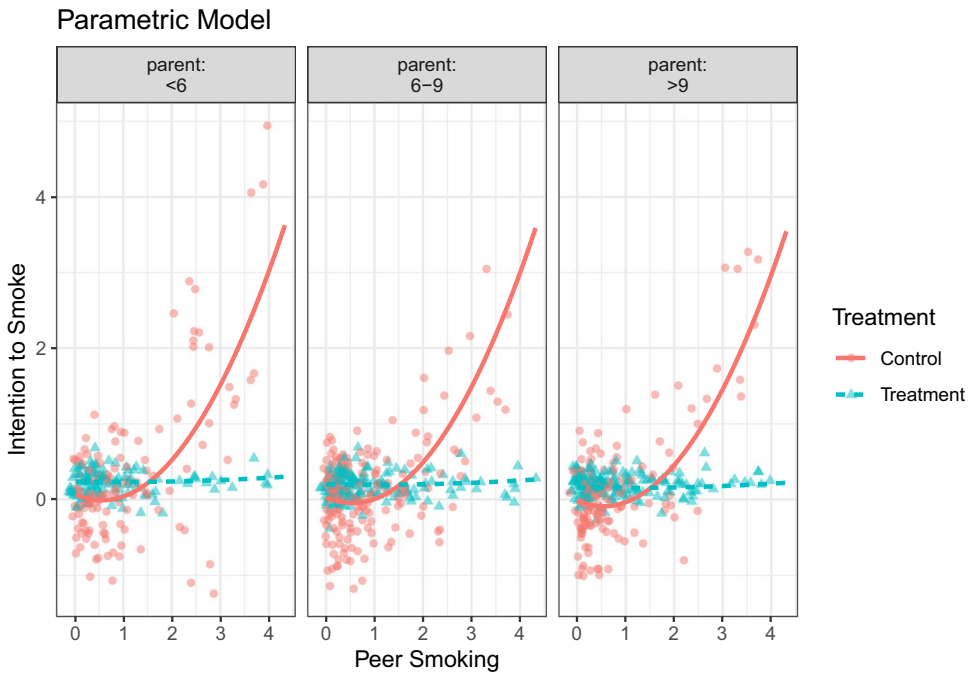

Fig. 10 This plot shows the predictions from a parametric model where a quadratic term was added to the peer/intention to smoke relationship, as well as an interaction term

such as nonlinear component, linear component, sample size, and number of covariates.

A better alternative to this approach is to sample parameters from a liberal range of values. For example, rather than selecting regression weights of -0.6, -0.4, etc., one could instead sample from a uniform distribution that ranges from -0.6 to +0.6. In other words, every iteration of the Monte Carlo will yield unique parameter values. One can then use RF in the same way the ANOVA is commonly used. The advantage of this is RF will identify important parameters and whether their importance derives from nonlinear relationships and/or interaction effects.

As of this writing, we know of no published articles that utilize this strategy. However, the strategy seems promising and may be of use to future researchers.

Discussion

The research landscape is rapidly changing. As a result, analysts are becoming more inclined to expand their statistical toolbox. It is our hope this expansion will include greater use of exploratory data analysis, including the use of RF models. RF models include several advantages that are particularly relevant to psychologists, including the ability to detect interactions and nonlinear effects as well as an impressive ability to avoid overfitting. While RF models are underutilized, we hope this paper has been a step toward more widespread adoption of RF models. In this paper, we have attempted to provide a simple explanation of RF models, addressed common misconceptions, and highlighted the limitations of RF models, including their "black box" nature and their nonparametric assumptions.

We have also provided several applied examples to show how to leverage RF's strengths and overcome its limitations. The key to this endeavor is to visualize the predictions of RF models. Visuals can reveal insights into the statistical information RF models are able to capture as well as suggest parametric alternatives one might pursue. We have provided an appendix (Appendix A) that demonstrates how to perform simple RF analyses and visualizations.

To be clear, we do not suggest RF models replace existing methods. Rather, we suggest the appropriate procedure to use depends on the analyst's intentions. RF models are best suited when researchers have exploratory intentions, in which RF models serve as a pit stop toward more sophisticated confirmatory methods. Alternatively, RF models can be used for more confirmatory research to check the viability of standard parametric models (e.g., linearity and homogeneity of regression). In either case, we hope to see greater use of this powerful tool.

Appendix : Getting started with random forests

To begin, we must install two packages: the `party` package, which contains functions to perform random forest, as well as `flexplot`, which will enable us to visualize the results from an RF analysis. The `party` package is available on CRAN and so can be installed with a simple command:

```
install.packages("party")
```

The `flexplot` package is not currently available on CRAN. To install it, we must first install another package called `devtools`, which allows us to install a package contained on GitHub. Subsequently, we can use the `install_github` command to install `flexplot`:

```
install.packages("devtools")

devtools::install_github("dustinfife/flexplot")
```

Once both packages are installed, we can load each into the working environment:

```
library(flexplot)

library(party)
```

The `flexplot` package comes pre-loaded with several datasets. For this example, we will use the `avengers` dataset, which contains a sample of 812 simulated observations about various characteristics related to the final Avengers battle in the movie Endgame. The variables include a strength score, number of injuries, minutes spent fighting, etc.

Suppose we wish to determine which variables are the best predictors of PTSD (labeled `ptsd` in the dataset). To do so, we could build a random forest model as follows:

```
set.seed(2335)

rf_model_results = cforest(ptsd~., data=avengers)
```

The first line of code (`set.seed(2335)`) simply sets the random number generator seed to a fixed value. This will ensure that those who reproduce our code will get the same results. The second line creates an object called `rf_model_results` that stores the results of a random forest model. The code `ptsd~.` simply tells the function to predict distress from every other variable in the dataset. Please note: this will probably take quite a while to run the analysis.

Subsequently, we could use the `estimates` function in the `flexplot` package to compute variable importance:

```
estimates(rf_model_results)
```

Which will generate the following results:

```
##
##
## Quantiles of absolute value of OOB performance (i.e., abs(predicted - actual)):
##
##    0%   25%   50%   75%  100%
## 0.000 0.128 0.281 0.488 3.402
##
##
## Variable importance (root MSE of predicted versus permuted):
##
##  minutes.fighting          injuries damage.resistance                iq
##             0.394             0.302             0.276             0.210
##       shots.taken       north_south           agility         willpower
##             0.197             0.179             0.179             0.085
##          strength             kills             speed       flexibility
##             0.056             0.052             0.048             0.036
##        superpower              died
##             0.028             0.015
```

The `estimates` function automatically sorts the variables in terms of VI. In this case, minutes.fighting, injuries, and damage.resistance are the top predictors of ptsd.

At this point, we suggest fitting a smaller model that only includes the top predictors. The reason for this is because flexplot struggles to visualize very large models. So, for this dataset, we might fit another smaller model that only includes the top three predictors:

```
rf_reduced = cforest(ptsd~minutes.fighting + injuries +
                       damage.resistance,
                     data=avengers)
```

Now, let's compute predicted values for the RF model. To do so, we can use the `compare.fits` function in flexplot. This computation is fairly computationally intensive. For that reason, we're going to save these predictions in an object called `prediction`. This will allow us to visualize it multiple ways without having to recompute the predictions.

```
predictions =
  compare.fits(ptsd~damage.resistance | minutes.fighting + injuries,
                              data=avengers, model1 = rf_reduced, return.preds=T)
```

Now, we'll take turns visualizing each of the top three variables on the *x*-axis. (Visualizing each variable on the *x*-axis makes the patterns more visually apparent). We'll place the other variables in panels to allow us to detect interactions. The following code produces three plots, one for each predictor variable. The important component is the `prediction = predictions` argument; that will pass the `flexplot` function the predictions from the RF model we computed earlier. These plots are shown in Fig. 11.

```
flexplot(ptsd~damage.resistance | minutes.fighting + injuries,
         data=avengers, prediction = predictions)
flexplot(ptsd~minutes.fighting | damage.resistance + injuries,
         data=avengers, prediction = predictions)
flexplot(ptsd~injuries | damage.resistance + minutes.fighting,
         data=avengers, prediction = predictions)
```

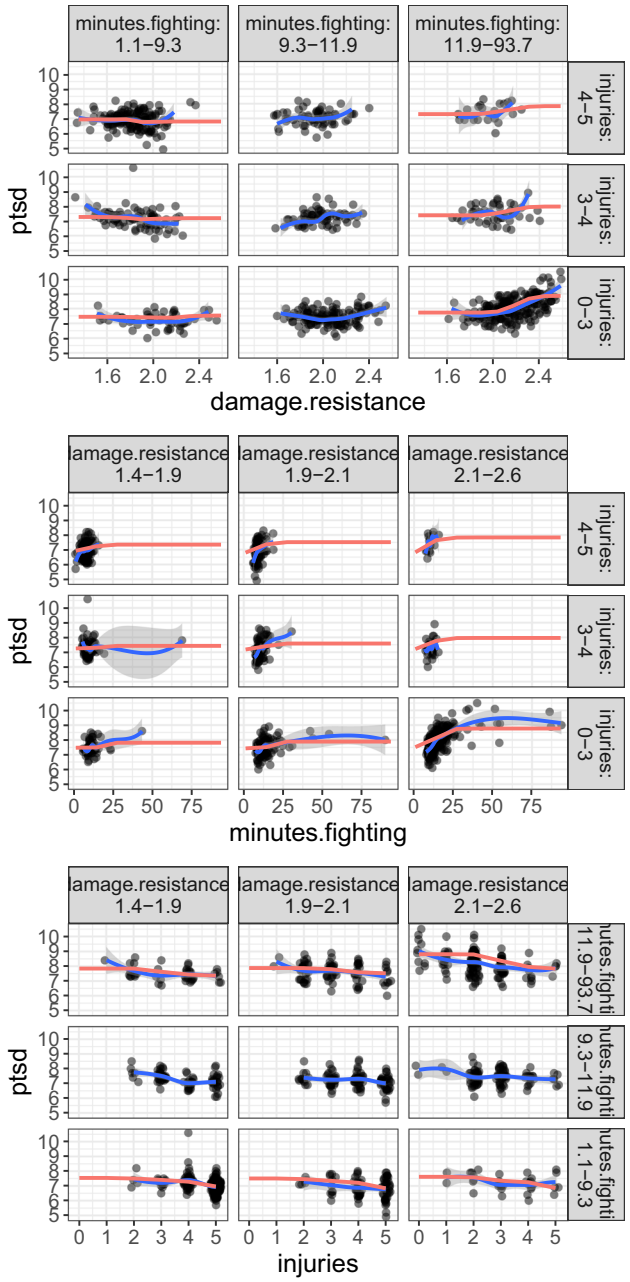

Fig. 11 These plots show the RF predictions of PTSD for various predictor variables

These plots could, of course, be used to identify an appropriate parametric model (e.g., perhaps a model where there's a nonlinear effect for damage.resistance and minutes.fighting, and possibly interactions between damage.resistance and injuries). For more information on identifying appropriate linear models from graphics, see (Fife & Mendoza, 2021).

References

Ammerman, B. A., Jacobucci, R., & McCloskey, M. S. (2018). Using exploratory data mining to identify important correlates of non-suicidal self-injury frequency. *Psychology of Violence*, *8*(4), 515–525. https://doi.org/10.1037/vio0000146

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Couronné, R., Probst, P., & Boulesteix, A. -L. (2018). Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics, 19(1)*. https://doi.org/10.1186/s12859-018-2264-5

Cramer, A. O. J., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P. P. P., & Wagenmakers, E. J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin and Review*, *23*(2), 640–647. https://doi.org/10.3758/s13423-015-0913-5

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. US, American Psychological Association. https://doi.org/10.1037/h0076829

Cumming, G (2014). The New Statistics: Why and How. Psychological Science. https://doi.org/10.1177/0956797613504966

Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychological Methods*, *18*(2), 186.

Fife, D. A. (2020). The eight steps of data analysis: a graphical framework to promote sound statistical analysis. *Perspectives on Psychological Science*, *15*(4), 1054–1075. https://doi.org/10.1177/1745691620917333

Fife, D. A. (2021). Flexplot: Graphical-Based Data Analysis. Psychological Methods. https://doi.org/10.1037/met0000424

Fife, D. A., Longo, G., Correll, M., & Tremoulet, P. (2021). A graph for every analysis: Mapping visuals onto common analyses using flexplot. Behavior Research Methods. https://doi.org/10.3758/s13428-020-01520-2

Fife, D. A., & Mendoza, J. L. (2021). Visual partitioning for multivariate models: An approach for identifying and visualizing complex multivariate dataset. https://doi.org/10.31234/osf.io/avu2n

Fife, D. A., & Rodgers, J. L. (2021). Understanding the Exploratory/Confirmatory Data Analysis Continuum. Moving Beyond the "Replication Crisis". American Psychologist, https://doi.org/10.1037/amp0000886

Fox, E. W., Hill, R. A., Leibowitz, S. G., Olsen, A. R., Thornbrugh, D. J., & Weber, M. H. (2017). Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. Environmental Monitoring and Assessment 189(7)https://doi.org/10.1007/s10661-017-6025-0

Gao, X., Wen, J., & Zhang, C. (2019). An improved random forest algorithm for predicting employee turnover. *Mathematical Problems in Engineering*, 1–12. https://doi.org/10.1155/2019/4140707

Gashler, M., Giraud-Carrier, C., & Martinez, T. (2008). Decision tree ensemble: Small heterogeneous is better than large homogeneous.

Gelman, A., & Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.

Genuer, R., Poggi, J. -M., & Tuleau-Malot, C. (2010). Variable selection using random forests, (Vol. 31. Retrieved from http://www.r-project.org/

Genuer, R., Poggi, J. -M., & Tuleau-Malot, C. (2019). VSURF: Variable selection using random forests. Retrieved from https://CRAN.R-project.org/package=VSURF

Ghosal, I., & Hooker, G. (2020). Boosting random forests to reduce bias; one-step boosted forest and its variance estimate. *Journal of Computational and Graphical Statistics*, 1–10. https://doi.org/10.1080/10618600.2020.1820345

Hapfelmeier, A., Hothorn, T., Ulm, K., & Strobl, C. (2012). A new variable importance measure for random forests with missing data. *Statistics and Computing*, *24*(1), 21–34. https://doi.org/10.1007/s11222-012-9349-1

Hayes, A. M., Laurenceau, J. -P., Feldman, G., Strauss, J. L., & Cardaciotto, L. (2007). Change is not always linear: The study of nonlinear and discontinuous patterns of change in psychotherapy. *Clinical Psychology Review*, *27*(6), 715–723.

Helmich, M. A., Wichers, M., Olthof, M., Strunk, G., Aas, B., & Aichhorn, W. (2020). Sudden gains in day-to-day change: Revealing nonlinear patterns of individual improvement in depression. *Journal of Consulting and Clinical Psychology*, *88*(2), 119.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). Springer.

Kirasich, K., Smith, T., & Sadler, B. (2018). Random forest vs logistic regression: Binary classification for heterogeneous datasets. *SMU Data Science Review*, *1*(3), 9.

Kitsantas, P., Moore, T. W., & Sly, D. F. (2007). Using classification trees to profile adolescent smoking behaviors. *Addictive Behaviors*, *32*(1), 9–23. https://doi.org/10.1016/j.addbeh.2006.03.014

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin and Review, 25(1)*. https://doi.org/10.3758/s13423-016-1221-4

Lord, F. I., & Novick, M. R. (1968) *Statistical theories of mental test scores*. Cambridge: Addison-Wesley.

Lu, R., Munroe, M. E., Guthridge, J. M., Bean, K. M., Fife, D. A., & Chen, H. (2016). Dysregulation of innate and adaptive serum mediators precedes systemic lupus erythematosus classification and improves prognostic accuracy of autoantibodies. *Journal of Autoimmunity*, *74*, 182–193. https://doi.org/10.1016/J.JAUT.2016.06.001

Malley, J. D., Kruppa, J., Dasgupta, A., Malley, K. G., & Ziegler, A. (2012). Probability machines. *Methods of Information in Medicine*, *51*(01), 74–81. https://doi.org/10.3414/me00-01-0052

Matsuki, K., Kuperman, V., & Van Dyke, J. A. (2016). The random forests statistical technique: An examination of its value for the study of reading. *Scientific Studies of Reading*, *20*(1), 20–33.

Mattei, T. (2014). Unveiling complexity: Non-linear and fractal analysis in neuroscience and cognitive psychology. *Frontiers in Computational Neuroscience*, *8*, 17. https://doi.org/10.3389/fncom.2014.00017

McAlexander, R. J., & Mentch, L. (2020). Predictive inference with random forests: A new perspective on classical analyses. *Research & Politics*, *7*(1), 205316802090548. https://doi.org/10.1177/2053168020905487

McArdle, J. J. (2012). Exploratory data mining using CART in the behavioral sciences. In *APA handbook of research methods in psychology, vol 3: Data analysis and research publication. (pp. 405–421). American Psychological Association.* https://doi.org/10.1037/13621-020

McNeish, D. M. (2015). Using lasso for predictor selection and to assuage overfitting: a method long overlooked in behavioral sciences. *Multivariate Behavioral Research*, *50*(5), 471–484. https://doi.org/10.1080/00273171.2015.1036965

Mentch, L., & Zhou, S. (2019). Randomization as regularization: A degrees of freedom explanation for random forest success. 1911.00190

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures, (Vol. 105. Retrieved from https://pdfs.semanticscholar.org/2903/180261ee0d99a27cfe85cde9cf4af74923c6.pdf

Mõttus, R., Wood, D., Condon, D. M., Back, M. D., Baumert, A., Costantini, G., & et al. (2020). Descriptive, predictive and explanatory personality research: Different goals, different approaches, but a shared need to move beyond the big few traits. *European Journal of Personality*, *34*(6), 1175–1201.

Muchlinski, D., Siroky, D., He, J., & Kocher, M. (2016). Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data, (Vol. 24. Retrieved from http://www.jstor.org/stable/24573207

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie Du Sert, N., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. https://doi.org/10.1038/s41562-016-0021

Nelson, L. D., Simmons, J. P., & Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psychology*, *69*, 511–545. https://doi.org/10.1146/annurev-psych-122216

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1708274114

Pan, Y., Liu, H., Metsch, L. R., & Feaster, D. J. (2017). Factors associated with HIV testing among participants from substance use disorder treatment programs in the US: a machine learning approach. *AIDS and Behavior*, *21*(2), 534–546. https://doi.org/10.1007/s10461-016-1628-y.

Rodgers, J. L. (2019). Degrees of freedom at the start of the second 100 years : a pedagogical treatise. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.1177/2515245919882050

Rossi, A., Amaddeo, F., Sandri, M., & Tansella, M. (2005). Determinants of once-only contact in a community-based psychiatric service. *Social Psychiatry and Psychiatric Epidemiology*, *40*(1), 50–56. https://doi.org/10.1007/s00127-005-0845-x

Ryo, M., & Rillig, M. C. (2017). Statistically reinforced machine learning for nonlinear patterns and variable interactions. *Ecosphere*, *8*(11), e01976. https://doi.org/10.1002/ecs2.1976

Schmidt, F. L., & Oh, I. -S. (2016). The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or is it something else?. *Archives of Scientific Psychology*, *4*(1), 32–37. https://doi.org/10.1037/arc0000029

Segal, M. R. (2004). Machine learning benchmarks and random forest regression. Retrieved from http://repositories.cdlib.org/cbmb/bench_rf_regn

Skidmore, S. T., & Thompson, B. (2013). Bias and precision of some classical ANOVA effect sizes when assumptions are violated. *Behavior Research Methods*, *45*(2), 536–546.

Smith, G. (2018). Step away from stepwise. *Journal of Big Data, 5(1)*. https://doi.org/10.1186/s40537-018-0143-6

Steinberg, D., & Colla, P. (1995) *CART: Tree-structured Non-parametric data analysis*. San Diego: Salford Systems.

Strobl, C., Boulesteix, A. -L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348. https://doi.org/10.1037/a0016973

Tay, L., Parrigon, S., Huang, Q., & LeBreton, J. M. (2016). Graphical descriptives: a way to improve data transparency and methodological rigor in psychology. *Perspectives on Psychological Science*, *11* (5), 692–701. https://doi.org/10.1177/1745691616663875

Touw, W. G., Bayjanov, J. R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., & van Hijum, S. A. (2013). Data mining in the life sciences with random forest: A walk in the park or lost in the jungle?. *Briefings in Bioinformatics*, *14*(3), 315–326.

Tukey, J. W. (1986). Analyzing data: Sanctification or detective work?. In L. V. Jones (Ed.) *The collected works of John W. Tukey* (pp. 721–737). London: Chapman & Hall.

Van Horn, M. L., Smith, J., Fagan, A. A., Jaki, T., Feaster, D. J., Masyn, K., & Howe, G. (2012). Not quite normal: Consequences of violating the assumption of normality in regression mixture

models. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(2), 227–249.

Wallert, J., Gustafson, E., Held, C., Madison, G., Norlund, F., Von Essen, L., & Olsson, E. M. G. (2018). Predicting adherence to internet-Delivered psychotherapy for symptoms of depression and anxiety after myocardial infarction: Machine learning insights from the U-CARE heart randomized controlled trial. *Journal of Medical Internet Research, 20(10)*. https://doi.org/10.2196/10754

Wang, H., Yang, F., & Luo, Z. (2016). An experimental study of the intrinsic stability of random forest variable importance measures. *BMC Bioinformatics*, *17*(1), 1–18.

Zhang, H., Zimmerman, J., Nettleton, D., & Nordman, D. J. (2019). Random forest prediction intervals. The American Statistician.