Advancing Inference in Supervised Learning Procedures via Permutation Tests and Importance Sampling, with Applications to Environmental Science

by

Timothy Coleman

Bachelor of Arts, Colgate Univeristy, 2016

Master of Arts, University of Pittsburgh, 2018

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Timothy Coleman

It was defended on

November 19th, 2020

and approved by

Lucas Mentch, University of Pittsburgh Department of Statistics

Satish Iyengar, University of Pittsburgh Department of Statistics

Larry Wasserman, Carnegie Mellon University, Department of Statistics and Data Science

& Department of Machine Learning

Kehui Chen, University of Pittsburgh Department of Statistics

Advancing Inference in Supervised Learning Procedures via Permutation Tests and Importance Sampling, with Applications to Environmental Science

Timothy Coleman, PhD

University of Pittsburgh, 2020

Random forests, since being proposed by Breiman [2001a], have become popular supervised regression and classification techniques. Their popularity stems from being easy to implement - the default hyper-parameter settings are often not far from optimal and are often competitive with more involved supervised models [Fernández-Delgado et al., 2014a]. While random forests are complex, they are not completely impenetrable to theoretical analysis. In this thesis, we present several contributions to random forest methodology. First, we provide a motivating application of random forests to ornithological data, where we develop a novel hypothesis test for testing equality of distribution of random forest curves [Coleman et al., 2017]. Then, we refine an observation made during that application into a means of testing hypotheses about the validation error of random forests, allowing for computationally efficient tests that are analogous to the F-test for linear regression. Finally, we propose a means of accounting for a discrepancy in test and training distributions, motivated by the problem of forecasting power outages from hurricanes.

Table of Contents

List of Tables

List of Figures

Preface

This thesis would not have been possible without all the people that have given me a chance. This path was not always obvious, and so without these people I would not have had the time to follow it:

- *Dr. Adam Burnett, Colgate University*
- *Dr. Jens Christensen, Colgate University*
- *Dr. Ahmet Ay, Colgate University*
- *Dr. Mike Loranty, Colgate University*
- *Dr. James McCollough, Air Force Research Lab*
- *Dr. Siddharth Manay, Lawrence Livermore National Lab*
- *Dr. Ryan Goldhahn, Lawrence Livermore National Lab*
- *Dr. Kimberly Kaufeld, Los Alamos National Lab*
- *Dr. Mary Frances Dorn, Los Alamos National Lab*

Additionally, I thank my committee members, including a special thanks to Dr. Lucas Mentch and Dr. Satish Iyengar, for all their kindness, understanding, and guidance for the duration of my PhD.

1.0 Introduction

Random forests are a computationally efficient, easily implemented method of supervised learning, and are able to model complex nonlinear interactions. As such, they are a natural tool to use in regression/classification problems in the environmental sciences, where large observational datasets of complex phenomena are commonplace. Drawing statistically valid conclusions in these settings remains challenging, and developing tools that accomplish this is the main goal of this thesis. For clarity, in the introduction, we first include some definitions of random forests and existing tools for analyzing their outputs. These definitions are repeated and refined in the subsequent chapters, but we believe that including the context of this dissertation is to the benefit of the reader. Relevant literature for each chapter is presented in the introduction of the chapter. We then present a structural overview of the rest of the thesis.

1.1 Some Definitions

Random forests, broadly speaking, are ensemble learners, that aggregate the predictions of many "weaker" learners (trees) to generate an overall prediction (the forest). Random forests are used when one has data of the structure, $\mathcal{D}_n = \{Z_1, Z_2, ..., Z_n\}$, with $Z_i = (X_i, Y_i)$ consisting of observations on covariates $X = (X_1, ..., X_p) \in \mathcal{X}$ and a response $Y \in \mathcal{Y}$. In the regression context, $\mathcal{Y}$ is often an uncountable subset of the real line, whereas in the classification context $\mathcal{Y}$ is some finite set. In this paper, the overwhelming focus is on the regression context, because often classification problems can be formulated as regressions of probabilities. For regression problems, we assume that $Y = m(X) + \epsilon$ where $m(X) = \mathbb{E}(Y|X = X)$ and ϵ is an independent noise process, typically with $\mathbb{E}(\epsilon) = 0$ and $\text{Var}(\epsilon) < \infty$. The goal of the random forest procedure is to accurately estimate $m(X)$. Each tree in a random forest is constructed by drawing resamples of size $k_n \leq n$, from $\mathcal{D}_n$, drawing a randomization parameter ξ from some distribution Ξ, and constructing a randomized decision

tree. The algorithm for generating the random tree is left up to the user, but popular methods include the CART algorithm [Breiman et al., 1984] or the conditional inference tree algorithm [Hothorn et al., 2006]. This process is repeated B times and the random forest prediction at some point $X \in \mathcal{X}$ is given by

$$RF_B(X) = \frac{1}{B}\sum_{j=1}^{B} T_j(X; \xi_j; \mathcal{D}_n).$$

To evaluate the RF prediction accuracy at a test location X with true response value y, we can measure the mean squared error

$$MSE_{RF}(X; y, \mathcal{D}_n) = \left(\left(\frac{1}{B}\sum_{j=1}^{B} T_j(X)\right) - y\right)^2.$$

Similarly, we can write the MSE of a forest at a collection of N_t test points $\mathcal{T}$ as $MSE_{RF}(\mathcal{T}) = \frac{1}{N_t}\sum_{\ell=1}^{N_t} MSE_{RF}(X_\ell; Y_\ell, \mathcal{D}_n)$.

MSE_{RF} can be adapted into a variable importance metric via the out-of-bag technique, which notes that each data point in $\mathcal{D}_n$ is excluded from some proportion of the trees in the forest. Specifically, consider that each of the B resamples generates a data matrix, $\mathcal{D}_j^*$, $j = 1, ..., B$. Then, let $B_i = \sum_{j=1}^{B} I(X_i \notin \mathcal{D}_j^*)$, i.e. the number of resamples that do not contain (X_i, Y_i), so that we can write the out-of-bag (oob) error as

$$\text{OOB}_B = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{B_i}\sum_{k=1}^{B} T(X_i; \xi_k) I(X_i \notin \mathcal{D}_k^*) - Y_i\right)^2.$$

OOB metrics can be used to estimate the importance of a variable $j \in \{1, ..., p\}$ by setting $\text{Imp}_j = \text{OOB}_{m,B}^{\pi_j} - \text{OOB}_{m,B}$, where π_j refers to the permuting (or random shuffling) of variable j in the out of bag sample. These metrics are often quick to calculate and are implemented in many popular statistical packages for fitting random forests, and thus are used quite regularly. However, they have several statistically undesirable properties, which are discussed throughout this thesis.

1.2 Structure of the Thesis

We begin with a motivating application from the ornithological community, using methods developed in Mentch and Hooker [2016a] and data from the eBird project [Sullivan et al., 2009a] to test for the importance of maximum temperature in north american tree swallow migrations. We additionally develop a functional permutation test for evaluating the hypotheses that 2008 and 2009 were especially anomalous years in tree swallow occurrence. One of the key observations of this work is that permuting the base learners of an ensemble may also lead to a valid statistical procedure. This work is under revision as Coleman et al. [2017].

From this intuition, we next develop a test for feature importance that is analogous to an F-test in simple linear regression. In particular, if a feature is unimportant, an ensemble trained with the feature should be as accurate as one without the feature. As such, we propose a test that permutes the individual trees, creating pseudo-forests, and then recording the difference in mean squared error. We prove that the base learners in a bagged model are exchangeable, and then appeal to theorems about exchangeability to prove that the permutation distribution and the sampling distribution are asymptotically equivalent. This avoids the challenging variance estimation problem of Mentch and Hooker [2016a], Wager and Athey [2018], and enjoys provable statistical validity, in contrast to some of the innovations on the out-of-bag metrics proposed in works like Janitza et al. [2016], Altmann et al. [2010], Ishwaran and Lu [2019]. Simulations showing the power and type I error control of the procedure are provided. Further, applications to the eBird data and a wildfire dataset [Cortez and Morais, 2007] are provided. Additionally, we demonstrate an application of the procedure outside of the ecological domain, to data from a cohort of patients diagnosed with irritable bowel syndromes (IBDs).

Finally, we present another methodological modification to random forests that is motivated by an environmental science application. Standard supervised learning procedures are validated against a test set that is assumed to have come from the same distribution as the training data. However, in the context of climate change, hurricane intensity has (and will continue to be) amplified beyond what is captured in the historical record, leading to record

damages from hurricane such as Irma, Sandy, and Harvey. Forecasting the power outages from these storms is especially challenging, given their severity with respect to the historical record. The extreme nature of the most devastating hurricanes means that typical validation set ups will provide severe underestimates of validation error. Our method provides a data-driven means of adapting a machine learning method to deal with extreme events. We consider the case of having many labeled (with continuous, numeric outcomes) observations from one distribution, P_1, and training a model to make predictions at unlabeled points that come from P_2, where P_1 and P_2 are absolutely continuous with respect to each other. We combine the high predictive accuracy of random forests with an importance sampling scheme, where the splits and predictions of the base-trees account for the weight assigned to each training observations. These weights correspond to a non-parametric likelihood ratio estimate, which is also estimated via a random forest, avoiding a costly high-dimensional density estimation problem. We also provide methods for imputing missing data in a way that respects the assumptions of the procedure, and for consistent tuning using a weighted out-of-bag error metric.

In the proposal of this thesis, the basic groundwork of each topic had been developed. Since then, much of the dissertation work has been on filling in the details for this work. Examples include the theoretical justification of the asymptotic validity of the permutation test. The approach presented in the proposal seemed promising, but turned out to be unfruitful, with the associated theorems only holding true on a set of measure 0. The updated theory, which uses a delta method argument, has been included in this final thesis. We have elected to include the full written manuscripts for each work, as those manuscripts provide all the relevant details to the proposed methods. Moreover, the relevant literature is presented in the introductory section of each chapter. The materials in Chapter 2 are available as Coleman et al. [2017], while the materials of Chapters 3 and 4 are on the arXiv as Coleman et al. [2019b] and Coleman et al. [2019a], respectively. Each of chapters 2-4 is self-contained, with relevant literature and definitions presented in their introductions. The supplementary material for each chapter is listed in the appendix, including more detailed simulation results and technical proofs.

2.0 Detecting Tree Swallow Population Anomalies Using Random Forests

2.1 Introduction

Tree Swallows (*Tachycineta bicolor*) are migratory aerial insectivores. In a recent breeding season study, Winkler et al. [2013] suggested that maximum daily temperature during the breeding season had a significant effect on the abundance of the flying insects that are the primary food source of Tree Swallows. This local-scale study conducted in upstate New York established how cold snaps, defined as two or more consecutive days when the maximum temperatures did not exceed 18.5°C, can result in a diminished food supply, thereby suggesting an indirect link between lower temperatures and lower fledgling success. Other work supports the hypothesis that migratory birds, like Tree Swallows, have breeding patterns that are affected by climate change, [Dunn and Winkler, 1999, Hussell, 2003]. While these papers focus heavily on breeding success and food availability, in this work, we investigate the associations of temperature on regional and local patterns of species occurrence during the autumn migration.

Most ornithological studies rely on controlled, local or regional level studies during a single season of the year, limiting the spatial and temporal scope of the analysis. The eBird project [Sullivan et al., 2009b, 2014b] hosted by the Cornell University Lab of Ornithology is a global bird monitoring project that allows for analysis on a much larger scale. This citizen science project compiles crowd-sourced observations of bird sightings, opening the door for a more data-driven approach to formally investigate scientific questions of interest. The eBird project harnesses the efforts of the bird-watching community by encouraging bird-watchers (birders) to record checklists of the species they encountered on each outing. These data have been used in a range of applications such as describing bird distribution across broad spatiotemporal extents [Fink et al., 2010, 2018], prioritizing priority habitat to conservation [Johnston et al., 2015], and identifying continental-scale constraints on migratory routes [La Sorte et al., 2016].

We study Tree Swallow populations during the autumn migration. During this time,

the species are believed to be *facultative migrants*. Facultative migration is an opportunistic migration strategy where individuals migrate in response to local conditions, such as the prevailing food supplies or weather conditions. Specifically, we study the low elevation New England / Mid-Atlantic Coast stretching north from the Chesapeake Bay to Boston, known as Bird Conservation Region 30 (BCR30) [Sauer et al., 2003] that forms the northern extent of the Tree Swallow winter range in Eastern North America.

Anecdotal accounts from bird watchers in this region suggest that Tree Swallows inhabit this region for prolonged autumn periods only during relatively warm winters. Though never formally documented or proven, in the years 2008 and 2009 it was widely believed in the ornithological community that the species did not linger in the region as late into the autumn as usual. Alternatively, mortality in the northern parts of the range in those years may have been higher. Thus, our primary objectives in this work are twofold: (1) to formally test whether the temporal pattern of Tree Swallow occurrence during the autumn migrations of 2008 and 2009 were substantially different than what would be expected during a typical autumn migration and (2) if there are differences, to investigate the association between local-scale patterns of occurrence and daily maximum temperature across broad geographic extents.

2.1.1 Challenges in Modelling Tree Swallows

While the ecological questions in the previous section are relatively straightforward to pose, providing accurate answers and provably valid statistical inference is challenging. In general, we expect that as daily temperatures decrease, the occurrence rate of Tree Swallows should also decrease as the species gravitates towards regions with more plentiful food or suffers higher mortality where food availability has been driven down by cold maximum temperatures. However, there are many strong sources of variation affecting the observed local-scale spatiotemporal patterns of species occurrence during the migration that can modify and mask the local-scale predictive utility of temperature. Ecological patterns of local-scale occurrence are affected by elevation, land cover types (e.g. open fields vs forests), and weather. Because of the difficulty finding and identifying birds in the field, variation in detection rates

further complicates modeling and inference about the underlying ecological processes. Based on previous work (see, for example, Zuckerberg et al. [2016]), we expect such associations to appear as complex, high-order interactions among the available covariates and that many of these associations and interactions will vary throughout the autumn migration.

Thus, one of the main analytical challenges is to develop models that can exploit rich covariate information to account for varied and complex sources of variation while facilitating statistical inference about potentially complex, local-scale associations. The large amount of available data, together with the presence of both nonlinear and high-order interactions, complicates the use of most traditional parametric and semiparametric models for this task. Thus, we rely on the more flexible alternative offered by random forests [Breiman, 2001b]. Random forests have a well-documented history of empirical success and are considered to be among the best "off-the-shelf" supervised learning methods available [Fernández-Delgado et al., 2014b]. This strong track record of predictive accuracy makes them an ideal "black-box" model for complex natural processes. Furthermore, tree-based methods have also proven very successful in other eBird projects [Robinson et al., 2018, Fink et al., 2018].

Though black-box model are not easily amenable to statistical inference, recent asymptotic results from Mentch and Hooker [2016b], Peng et al. [2019] and Wager and Athey [2018] on the distribution and variance estimation of predictions resulting from RF models provide a formal statistical framework for addressing our primary questions of interest. Moreover, as we demonstrate, traditional non-parametric inferential procedures can also be used to help draw inferences from these complex models.

In this paper, we begin with a brief overview of the data and available covariate information in Section 2.2. In Section 2.3, we provide further evidence for use of random forests to answer the questions posed earlier. We then construct preliminary RF models to assess the influence of temperature and produce maps of prediction differences between models to understand the spatial patterns in the association with maximum daily temperature. In Section 2.4, we develop a permutation-style test to investigate how unusual the 2008 and 2009 migration patterns appear to be by treating the RF predictions over time as functional data. Finally, In Section 2.4.3, we make use of recent asymptotic results to test the

significance of maximum daily temperature at a variety of local test locations throughout the region of interest, BCR30. Throughout this work, the predictive associations uncovered should not be interpreted as causal effects. Indeed, due to the structure of the tree swallow data, it is more likely that the causal relationship between maximum daily temperature and `occurrence` is indirect, as maximum daily temperature affects food availability or other local resources upon which Tree Swallows depend.. However, we recognize the important work done in extending random forests to estimation of causal effects, and as such, a section implementing the causal forests of Wager and Athey [2018], Athey et al. [2019] is provided in the supplementary material.

2.2 Data Overview

The eBird data is accumulated on a per-birder outing basis. During each outing, the birder records the species of birds observed. Each species observed is recorded as a presence observation while unobserved species are marked absent. The outing is then referenced with environmental, spatial, temporal, and user information. This last set of predictors is included in order to account for variation in detection rates, a potential confounder when making inference about species distributions. Our outcome of interest is the probability that at least one Tree Swallow is observed given the spatial, temporal, and detection process information. We refer to this probability as *occurrence*, that is

$$O = P(\text{Tree Swallow is Observed} \mid X = x)$$

where X denotes the covariate information. Because we are interested in the eastern autumn migration, we restrict our attention to eBird observations located in the BCR30 region that were recorded on or after the 200th day of the year between the years 2008-2013. In total, the full dataset contains 173002 observations on 30 variables, with occurrences of tree swallows in 10.8% of the observations.

Spatial information is captured by land cover and elevation data. To account for habitat-selectivity each eBird location has been linked to the remotely-sensed MODIS global land

cover product (MCD12Q1) [Friedl et al., 2010]. Here we use the 2011 MODIS land cover data as a static snapshot of the landcover. These landcover predictors were associated with eBird observations collected from 2004 to 2012. Finally, we use the University of Maryland (UMD) classification scheme [Hansen et al., 2000] to classify each 500m $\times$ 500m pixel (25 hectare) as one of 14 classes, including classes such as water, evergreen needleleaf forest, and grasslands. We summarized the land cover data as the proportion of each land cover class within a 3.0km $\times$ 3.0km (900 hectare) pixel centered at each location using FRAGSTATS [McGarigal et al., 2012].

Temporal information is included at three resolutions. At the finest temporal resolution, the time of the day at which the observation was made is used to model variation in availability for detection; e.g., diurnal variation in behavior [Diefenbach et al., 2007] may make species more or less conspicuous. For our purposes, we restrict our attention to the day of year (`DoY`) and the year itself, corresponding to our interest in anomalies in the fall migration.

Temperature data was collected from the DayMet project, hosted by Oak Ridge National Lab [Thormton et al., 2017]. The data includes daily maximum (`max_temp`), minimum, and mean temperature for each day in the training period. We also estimated an expected daily maximum temperature for each day by taking the mean daily maximum temperature for each eBird location from 1980-2007. The anomaly relative to this expected maximum (`max_temp_anomaly`, defined as `max_temp` minus the 1980-2007 normal `max_temp`) is of particular interest since `max_temp` alone is strongly correlated with `DoY`. Each eBird location is further associated with the 30m gridded elevation from the ASTER Global Digital Elevation Model Version 2.

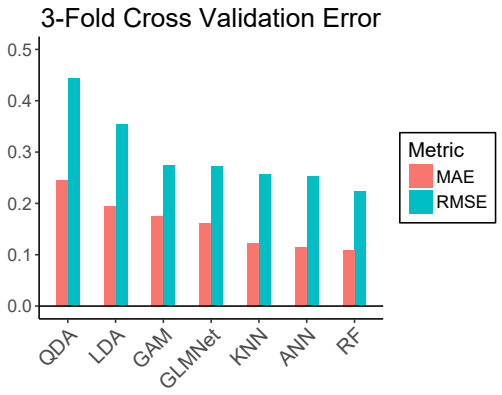
Finally, there are three user effort variables included in the model to account for variation in detection rates: the hours spent searching for species (`eff_hours`), the length of transects traveled during the search (`eff_dist`), and the number of people in the search party (`n_obs`). In addition, an indicator of observations made under the "traveling count" protocol was included to allow the model to capture systematic differences in species detection between the the counts recorded by traveling and stationary birders.

2.3 Preliminary Models

In this section, we provide further evidence for the use of random forests to model Tree Swallow migration. As noted earlier, random forests are typically used in datasets with many observations on many predictors, whose effect on the response may be a complex, nonlinear function of the predictors. As demonstrated in Fernández-Delgado et al. [2014b], though random forests do not universally dominate other methods, they are often exceptionally accurate and robust supervised learners.

Fundamentally, however, our interest in this work is in scientific understanding and statistical inference and certainly there are numerous alternative statistical models that provide a more direct means of accomplishing this. However, to trust such inference, we must trust that the model selected is able to accurately capture the complex underlying mechanisms. This suggests the question: are there notable gains in accuracy by using random forests, or would a more straightforward statistical model suffice? To answer this using the eBird data, we use cross validation (CV) to measure the predictive accuracy of a variety of popular modeling techniques for binary outcomes. In particular, we train a random forest with `mtry = 5` (not chosen by cross validation) and 500 trees, a k-nearest-neighbors (KNN) regression model with k chosen from $\{5, 7, .., 21, 23\}$, a 3-layer artificial neural network (ANN) with the number of neurons chosen from $\{15, 30, 100\}$ at each layer [Bergmeir and Benítez, 2012], Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and a Generalized Additive Model (GAM) with degrees of freedom chosen from $\{1, 6, 11, ..., 21, 26\}$ [Hastie, 2017]. Finally, we train an elastic-net penalized logistic regression (GLMNet) model [Friedman et al., 2010], with weights $\alpha \in \{0, 1\}$, (0 corresponds to the Lasso, 1 corresponds to Ridge Regression), with cost parameter $\lambda \in \{0, .01, ..., .15\}$. This model fits coefficients to all covariates and also every two-way interaction, to parsimoniously select the strongest interaction models. These models are trained using the `caret` package [Kuhn, 2017] in `R`, and, with the exception of random forests, the parameters chosen reflect those which lead to the smallest CV estimate of Root Mean Squared Error (RMSE). We also report the CV estimate of the Mean Absolute Error (MAE).

The results of this analysis are displayed in Figure 2.3.1. Even without tuning, the



Model	RMSE	MAE
Random Forest	0.22280	0.10863
ANN (100, 30, 15)	0.25300	0.11465
KNN ($k = 11$)	0.25558	0.12195
GLMNet ($\lambda = 0$)	0.27260	0.16163
GAM ($df = 26$)	0.27380	0.17412
LDA	0.35350	0.19400
QDA	0.44427	0.24541

Figure 2.3.1: 3-fold CV estimates of RMSE and MAE for various predictive models, plotted in descending RMSE order, and tabulated with optimal hyper-parameters.

off-the-shelf random forest model attains the lowest RMSE and MAE scores with the other flexible models, such as KNN and the ANN, not far behind. We see that the GAM and GLMNet models are similar in performance, with LDA and QDA lagging severely behind. Notably, the GLMNet model selected a tuning parameter that maximized model complexity, i.e. $\lambda = \alpha = 0$, even with all two-way interactions considered. This further suggests that a parametric model is unreasonable due to the complex interactions and functions of the covariates that go into predicting occurrence. The strong predictive performance of random forests, combined with the recent advancements in inference for random forests, make them an ideal model for drawing conclusions about tree swallow migrations. As a first step analysis, we make use of traditional means for drawing inferences from black-box methodology of RFs with tools such as partial effect plots and conclude with a spatial analysis of the effect of `max_temp`. These analyses provide heuristic and provisional answers to the study questions, and motivate the more formal testing procedures developed and executed in later sections.

2.3.1 Inspecting Annual Migration Differences

Recall that the initial motivation for this study was a widely perceived difference in Tree Swallow autumn distribution patterns in the years 2008 and 2009. In particular, it was believed that Tree Swallows had remained in the northern regions for a shorter period in the fall during 2008-2009, but the ornithological community was unsure of the mechanism(s)

behind this earlier departure/decline. Accordingly, we begin by partitioning the BCR30 Tree Swallow data into two training samples: one containing observations from 2008-2009 and the other containing the observations from 2010-2013. Formally, denote the entire training set as $\mathcal{D}$ so that we can write our partitioned training datasets as $\mathcal{D}_{08-09}$ and $\mathcal{D}_{10-13}$, respectively, with $\mathcal{D} = \mathcal{D}_{08-09} \cup \mathcal{D}_{10-13}$. For each day of the year beginning with `DoY` 200, 100 points were selected at random from $\mathcal{D}$ to serve as a validation set. These 16600 points were then removed from the corresponding training set.

We first construct a RF on each of these temporally divided training sets. It is important to note however, that the eBird project has grown substantially in popularity since its inception in 2002 and thus later years contain many more observations than earlier years. In particular, $\mathcal{D}_{08-09}$ contains a total of 21,907 observations, while $\mathcal{D}_{10-13}$ contains 151,095. Because a RF trained on a larger dataset may be more stable, any differences observed between predictions generated by the two datasets may be partially explained by the difference in data sizes. To account for this, we also selected (uniformly at random, without replacement) a subsample of size 21,907 from the $\mathcal{D}_{10-13}$ training data and with it, constructed a third RF. Predictions were made at all points in the validation set, and averaged by day. The results are shown in Figure 2.3.2.

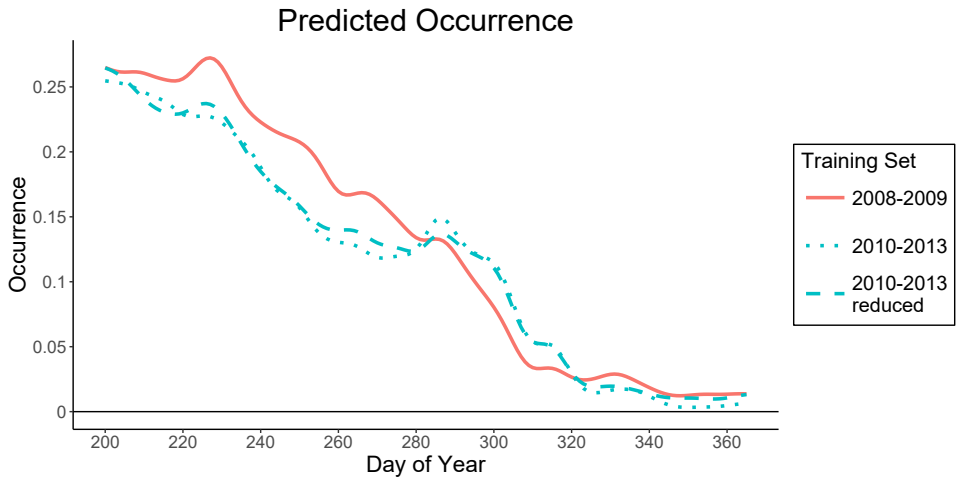


Figure 2.3.2: Predicted occurrence by random forests trained on $\mathcal{D}_{08-09}$, $\mathcal{D}_{10-13}$, and a subsample of $\mathcal{D}_{10-13}$. Predictions shown are kernel smoothed estimates of the prediction surface using a Gaussian kernel with a bandwidth of 5.

From Figure 2.3.2, we see that the RF trained on $\mathcal{D}_{08-09}$ predicts the largest occurrence until approximately `DoY` 285, after which the 2010-2013 forests are higher until approximately `DoY` 320, from which point the differences appear negligible. This seems to support the hypothesis that during the years 2008 and 2009, Tree Swallows remained in northern regions longer before departing more quickly. Importantly, the predictions from the RF trained on the reduced dataset from 2010-2013 forest differs only very slightly from those generated by the RF trained on the full $\mathcal{D}_{10-13}$ data, suggesting that the more substantial departures observed between predictions generated by RFs trained on $\mathcal{D}_{08-09}$ and $\mathcal{D}_{10-13}$ have little to do with the differing training sample sizes.

An additional focus of our work is to investigate the significance and impact of maximum temperature (`max_temp`) in predicting occurrence. Using the entire training dataset, we estimated a partial effect function of `max_temp` with `DoY` removed as a covariate (Figure 2.3.3), to account for any confounding effects between `DoY` and `max_temp`. These functions are estimated by discretizing `max_temp` in the training data, $\mathcal{D}$, into a grid. For each grid point, predictions are made at each observation whose discretized `max_temp` corresponds to the grid value. The partial effect value at the grid point is then the average of all the predictions at that grid point. As expected, we see a steady increase in occurrence with `max_temp` starting around 7°C which appears to begin leveling off around 32°C. As an interesting side note, the sharpest increase appears to occur around 15°C, which corresponds to a period of heightened insect activity suggested in Winkler et al. [2013].

Finally, recall that out-of-bag (oob) variable importance measures are a popular *ad hoc* measure that usually accompanies random forest predictions. Breiman [2001b] introduced these oob measures as a means of quickly assessing variable importance by calculating the decline in prediction accuracy observed when the values of a particular variable are permuted amongst the oob samples. According to this metric, `max_temp` was determined to be the most important covariate, though we note this with caution as the oob measures are often unreliable. A substantial amount of previous work — see Strobl et al. [2007b], Nicodemus et al. [2010], Toloşi and Lengauer [2011b], Hooker [2007] for popular examples — has demonstrated serious flaws with such measures, most notably that they tend to inflate the importance of groups of correlated covariates. This is especially problematic in our context,

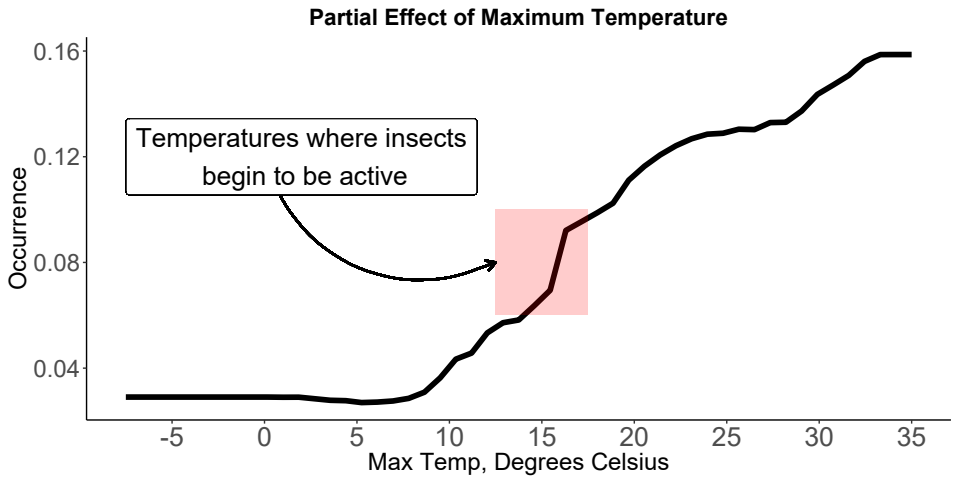


Figure 2.3.3: Partial effects plot for `max_temp` for a RF trained on the entire dataset. Shaded region corresponds to range of temperatures where potentially flying insects shift from inactive to flying [Winkler et al., 2013].

where daily maximum temperature is one of several highly correlated spatial covariates of varying importance to predicting occurrence.

2.3.2 Visualizing the Spatial Effects of Maximum Temperature

Here, we examine the spatiotemporal effect of `max_temp` on occurrence throughout the Northeast. We construct two RFs, one with the original data and one with `max_temp` permuted, and compare the predictions generated by these two models. To remove variation associated with the detection process, a nuisance when investigating the effect of maximum temperature, we set the observer characteristics (`n_obs`, `eff_hours`, and `eff_dist`) to 1 which coincides with typical levels for many ebird observers. The test set consists of the points in the 3km grid within the $[-78°, -68°] \times [37°, 44°]$ longitude/latitude region, where landcover characteristics (UMD classes, elevation) are concatenated with `max_temp`. The values assigned to `max_temp` in the test set were imputed from the 2014 DayMet observations, providing temperature information that was collected independently of the `max_temp` values in $\mathcal{D}$.

We then make predictions using both forests and calculate the difference in predictions.

Formally, for a test point in the grid X_{ij}, define the difference in predictions between the original and permuted forest as

$$\hat{d}_{ij} := R(\mathcal{D}; X_{ij}) - R(\mathcal{D}^{\pi}; X_{ij}) = R(X_{ij}) - R_{\pi}(X_{ij})$$

where $\mathcal{D}$ denotes the original training data, $\mathcal{D}^{\pi}$ denotes the training data with `max_temp` permuted, and R and R_{π} denote the RFs trained on such data, respectively. In order to examine the temporal dynamics, we create 9 test grids, each 20 days apart, throughout the fall.

The resulting heat maps of prediction differences demonstrating the effects of `max_temp` are shown in Figure 2.3.4. Red indicates the predictions of the original RF being higher than the permuted RF; blue indicates that the permuted forest made larger predictions. We see that earlier in the fall $R > R_{\pi}$ followed by roughly equal predictions onward from day 101 of the fall. Under the assumption that `max_temp` is unrelated to the response and therefore simply noise, we might expect that the differences in predictions between the original and permuted RFs across space are also simply random, uncorrelated noise. A Moran's I test, (see Appendix A), provides strong evidence that the differences plotted in Figure 2.3.4 exhibit spatial autocorrelation. The purpose of this test is to search for local effects of `max_temp` in RF predictions; if `max_temp` is meaningful in predicting occurrence, we would expect the differences in predictions between two points near each other to be more strongly correlated than two points further away. This provides statistical backing to what is clear from Figure 2.3.4: there is certainly a "clumping" among the differences between the random forests, suggesting that `max_temp`'s effect has local homogeneity.

The results of the Moran's I test should be interpreted with some caution; the predictions from the forests R and R_{π} are functions of the test set `max_temp` which itself has strong spatial correlation. Thus, the predictions and their differences may exhibit correlation regardless of the true association between occurrence and maximum temperature. A more direct approach to assessing the local-scale importance of maximum daily temperature is provided in Section 2.4.3.

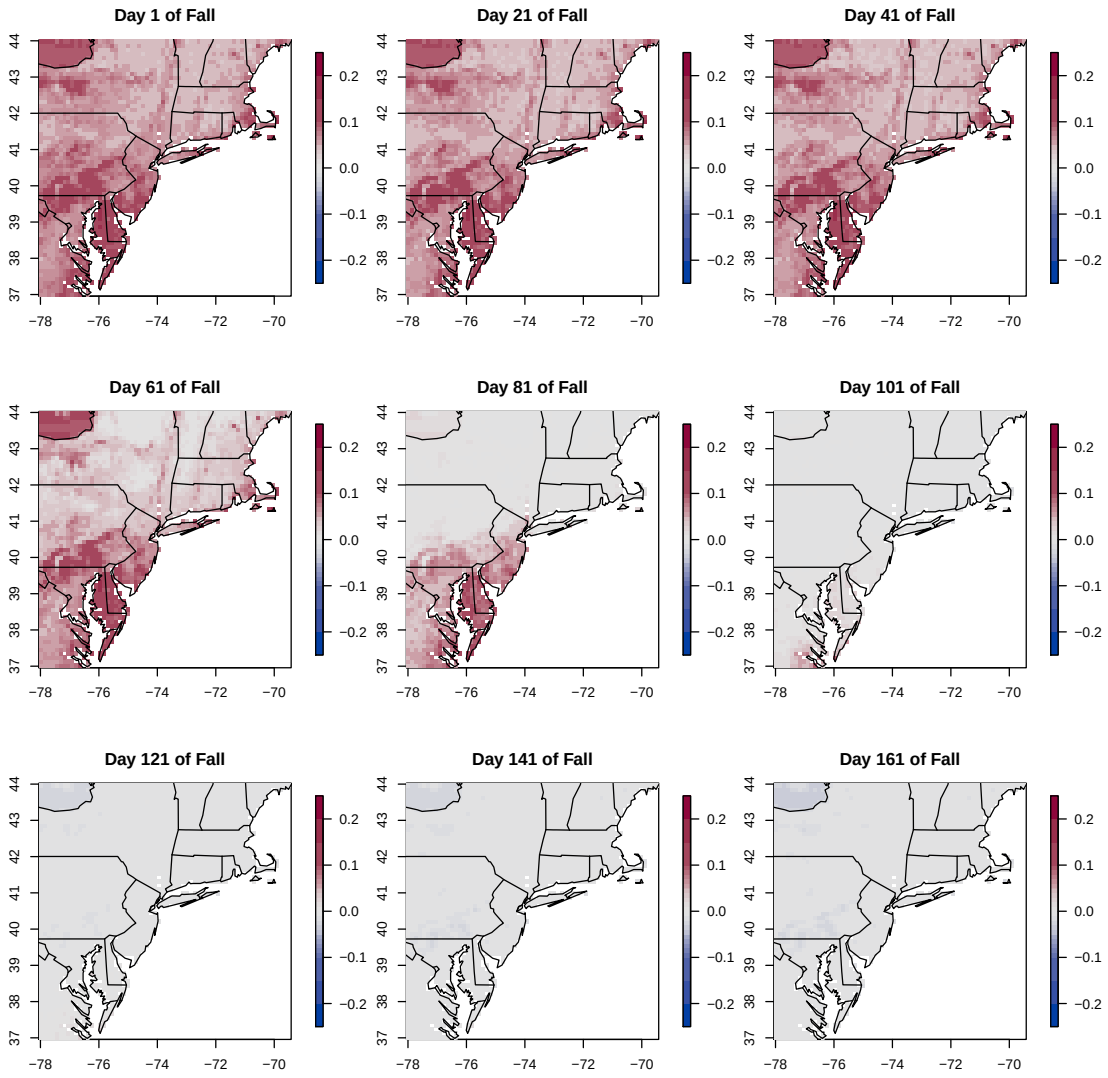


Figure 2.3.4: Predicted occurrence differences ($\hat{d}$) between original and permuted RFs calculated at 9 time points throughout the fall. Red indicates larger predictions from the original RF; grey indicates roughly equal predictions

2.4 Testing for Regional Differences in Occurrence

We now devise and implement a permutation test to explicitly assess whether the prediction curves in Figure 2.3.2 exhibit differences that could plausibly be due to chance. Here we consider *regional* hypotheses, meaning that we investigate differences in species occurrence throughout the entire BCR30 region as opposed to at a specific location or set of locations within that region.

2.4.1 Testing Procedure and Data

Our strategy here is to use a permutation test to investigate hypotheses about the distribution of occurrence in the 2008-2009 and 2010-2013 groupings. Permutation tests, in addition to maintaining exact control of the Type I error rate for distributional hypotheses, have the advantage of being completely distribution free, regardless of the test statistic used. If we let $D_{08-09}(X, Y)$ denote the joint distribution of the covariates and occurrence for the 2008-2009 data, and similarly define $D_{10-13}(X, Y)$, we then want to test

$$\begin{aligned} \mathrm{H}_0 &: D_{08-09}(X, Y) = D_{10-13}(X, Y) \\ \mathrm{H}_1 &: D_{08-09}(X, Y) \neq D_{10-13}(X, Y). \end{aligned} \tag{2.4.1}$$

To account for differing training set sizes and also in the interest of both computational efficiency and being conservative in our testing procedures, we now construct a reduced training set from the $\mathcal{D}_{10-13}$ data containing the same number of observations as $\mathcal{D}_{08-09}$. We construct this reduced set by taking each observation in $\mathcal{D}_{08-09}$ and drawing a radius around it in both space (0.2 decimal degrees in both latitude and longitude, an area of approximately 352 km^2) and time (2 days). We then locate all observations from $\mathcal{D}_{10-13}$ within this radius and select from these an observation uniformly at random, without replacement. This produces a "nearest neighbor" training set, $\mathcal{D}^{NN}_{10-13}$, with roughly the same spatiotemporal distribution of observations, allowing us to more closely examine the influence of the other covariates on the functional observations. By enforcing spatio-temporal uniformity between the datasets, we are controlling for differences in eBird user behavior between the two groups. As such, any difference observed is more attributable to year to year changes in ecological variables

(such as land cover characteristics) or occurrence itself. The first stage of calculating our test statistic is to train a random forest on both $\mathcal{D}_{08-09}$ and $\mathcal{D}^{NN}_{10-13}$. We then use these forests to make predictions at fixed test points, from which several summary statistics are calculated.

Our test set consists of 166×1000 points, with 1000 points taken for each day in the fall. We construct this test set by sampling 1000 locations from a 3km$\times$3km grid covering the east coast, referenced with their land cover and elevation characteristics, as well as `max_temp` for that day. The maximum temperature information included is the expected daily maximum temperature, estimated from the 1980-2007 temperature information provided by DayMet. The variables associated with the eBird user (e.g. `eff_dist`, `eff_hours`, and `n_obs`) are set to 1 uniformly, to again represent typical eBird user levels.

Let $R_{08-09}(\cdot)$ denote the prediction function of the RF trained on $\mathcal{D}_{08-09}$. Then f_{08-09} is defined as

$$f_{08-09}(t) := \frac{1}{1000} \sum_{k=1}^{1000} R_{08-09}(X_{k,t}), \quad t \in \{200, ..., 365\}$$

where $X_{k,t}$ is a point in the test set corresponding to time t. Thus, since the test points are stratified by time, $f_{08-09}(t)$ denotes the average over predictions made at all 1000 test points on each day and therefore represents a time-averaged version of the raw RF prediction function. The function $f_{10-13}(t)$ is defined in exactly the same fashion for a RF trained on $\mathcal{D}^{NN}_{10-13}$.

Recall that the original hypothesis was that Tree Swallows remained in BCR30 longer in 2008-2009 than in 2010-2013, followed by a sharp decline in numbers. Preliminary analysis in Figure 2.3.2 supports this hypothesis, so we now evaluate the statistical significance of the evidence. Formally, in early fall (`DoY` 200-264), it appears that $f_{08-09} > f_{10-13}$. Later in the fall (`DoY` 265-310), it appears that $f_{08-09} < f_{10-13}$ and finally as winter sets in (`DoY` 311-365), we see $f_{08-09} \approx f_{10-13}$. We therefore partition our time frame into three disjoint time periods, $T_1 = \{200, ..., 264\}$, $T_2 = \{265, ..., 310\}$, and $T_3 = \{311, ..., 365\}$, and let $I_{T_i}(t)$ be an indicator function for each period. We then consider the restricted functional observations

$$f^{(i)}_{08-09}(t) := f_{08-09}(t) I_{T_i}(t) \quad \text{for } i = 1, 2, 3$$

which are defined in the same fashion for $f^{(i)}_{10-13}(t)$. Each of these restricted functional observations is then incorporated into test statistics to evaluate following sets of hypotheses

$$\begin{aligned} \mathrm{H}_{0,i} &: D^{(i)}_{08-09}(X,Y) = D^{(i)}_{10-13}(X,Y) \\ \mathrm{H}_{1,i} &: D^{(i)}_{08-09}(X,Y) \neq D^{(i)}_{10-13}(X,Y). \end{aligned} \tag{2.4.2}$$

To evaluate these hypotheses, we begin by calculating the prediction functions over time using the original datasets and then, for each of many iterations, we permute the `year` covariate, re-partition the data into the two groups consisting of data from 2008-2009 and 2010-2013, and construct the new RF prediction functions.

Permutation tests for hypotheses of this form reject H_0 if the test statistic, T_0, calculated on the original data, falls in the extreme (upper or lower $\alpha/2$) quantile of the permutation distribution of test statistics. Formally, given two sets of data $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^m$, let $\mathcal{G}$ be the group of all permutations of the indices $1, ..., m+n$. Then, consider a statistic of the form $T = T(Z_1, ...Z_{m+n})$, and let T_0 be the statistic calculated on the original data. A p-value for the hypothesis the null hypothesis $H_0 : D(X) = D(Y)$ is given by

$$p = \frac{1}{|\mathcal{G}|} \sum_{\pi \in \mathcal{G}} I(|T_0| > |T(Z_{\pi(1)}, ..., Z_{\pi(m+n)})|).$$

Note that $|\mathcal{G}| = \binom{m+n}{n}$ which is quite large, so we instead sample 1000 draws from the permutation distribution uniformly at random, which maintains the size of the test at α [Lehmann and Romano, 2006]. Permutation tests offer flexibility in the choice of test statistic, and different test statistics offer different levels of power. As such, we consider the following two measures of functional distance

$$\begin{aligned} KS &= \sup_{t \in T_1 \cup T_2 \cup T_3} |f_{08-09}(t) - f_{10-13}(t)| \\ \Delta_i &= \frac{1}{|T_i|} \sum_{t \in T_i} (f^{(i)}_{08-09}(t) - f^{(i)}_{10-13}(t)), \quad i = 1, 2, 3. \end{aligned} \tag{2.4.3}$$

The first measure in Equation 2.4.3 refers to the Kolmorgorov-Smirnov statistic, traditionally used to test hypotheses about distribution functions. The use of this statistic for two sample functional testing procedures was studied by Hall and Van Keilegom [2007]. KS is calculated across the full time period, and then used for testing for an overall difference in the underlying distributions. In contrast, our raw distance measures $\Delta_1, \Delta_2,$ and Δ_3 are designed to test for equality of the underlying distributions only in time periods $T_1, T_2,$ and T_3, respectively.

Based on the visual evidence in Figure 2.3.23, we may expect to see a difference during the first two time periods, but likely not during the third time period.

2.4.1.1 A Computationally Efficient Alternative Testing Procedure

The procedure described above maintains many of the desirable statistical properties of permutation tests, such as exactness under any distribution. As such, we refer to it as the *canonical* permutation test. However, permutation tests were developed for situations where test statistics are easily calculated. Because our test statistic involves training of two random forests, there is substantial computational burden incurred in conducting each test. Indeed, running the full test requires constructing $2 \times N_{\text{Perm}} \times B$ decision trees, where B is the size of each forest. As such, we now propose a computationally efficient alternative.

Random forest predictions can be written as a function of the training data, the test point, and a collection of randomization parameters, $\xi = \{\xi_1, ..., \xi_B\}$, which dictate the feature subsetting and resampling used in each tree. For a given test point X, the random forest prediction $R(X; \mathcal{D}, \boldsymbol{\xi})$ can be written as

$$R(X; \mathcal{D}, \boldsymbol{\xi}) = \frac{1}{B} \sum_{k=1}^{B} T(X; \mathcal{D}, \xi_k)$$

where $T(\cdot; \mathcal{D}, \xi)$ is a standard CART decision tree trained on $\mathcal{D}$ using randomization ξ. In a random forest, the randomization parameters are drawn in an iid fashion, so that for any point X and any number of trees B, $\{T(X; \mathcal{D}, \xi_k)\}_{k=1}^{B}$ is an iid sequence conditional on $\mathcal{D}$. Now, we can appeal to the classical De Finetti's Theorem [De Finetti, 1937] for infinitely exchangeable random variables, which states that *a sequence of infinitely exchangeable random variables is exchangeable if and only if it is iid conditional on some other random variable.* The sequence of trees used in a random forest are iid, conditional on the data, and therefore are infinitely exchangeable. Moreover, suppose we partition a collection of B trees into k subgroups, each consisting of B/k trees, and form k random forests from these trees. Then, the same argument gives that $R_1(X; \mathcal{D}), ..., R_k(X; \mathcal{D})$ is infinitely exchangeable, and further that the functional observations (like those used in the test statistics in Equation 2.4.3) are realizations of an infinitely exchangeable sequence of functions. As such, if we train B trees,

and then randomly stratify the trees into k forests of equal size, we have an exchangeable sequence of functions.

Exchangeability is fundamental to the exactness of permutation tests. In fact, a permutation test is fundamentally a test of *exchangeability* - for two groups of data $X_i \overset{iid}{\sim} P$ and independently, $Y_i \overset{iid}{\sim} Q$, the data are exchangeable if and only if $P \equiv Q$. To see this, note that under an exchangeability assumption:

$$D_0 := D(X_1, ..., X_n, Y_1, ..., Y_m) = D(Z_{\pi(1)}, ..., Z_{\pi(m+n)}) := D_{\boldsymbol{Z}_\pi} \quad \forall \pi \tag{2.4.4}$$

where $\boldsymbol{Z}_\pi$ is any permutation of the X_i and Y_i. Thus, for any given test statistic (i.e. a function of the $m + n$ observations), the quantile of the observed test statistic across all possible permutations should approximately follow a uniform distribution [Pesarin and Salmaso, 2010]. For iid observations, Equation 2.4.4 factors as

$$D(X_1, ..., X_n, Y_1, ..., Y_m) = \prod_{i=1}^{n} P(X_i) \times \prod_{i=1}^{m} Q(Y_i) \overset{\text{exchangeable}}{\equiv} \prod_{i=1}^{n+m} P(Z_\pi(i)).$$

Thus for iid data, the finite sample permutation test provides an exact test for hypotheses about equality of distribution. Indeed, as a result of Equation 2.4.4, for any statistic $T(\cdot)$, $T(\boldsymbol{Z}) \overset{d}{=} T(\boldsymbol{Z}_\pi)$. A more rigorous argument for the validity of the tests is presented in Lehmann and Romano [2006].

In the set up described, we sample 20 exchangeable functions, $\{f_{k,(08-09)}\}_{k=1}^{20}$ and $\{f_{k,(10-13)}\}_{k=1}^{20}$ each using 50 identically trained decision trees. Treating the observed functions f_{08-09} and f_{10-13} as observations from functional distributions $\mathcal{F}_{08-09}$ and $\mathcal{F}_{10-13}$, our goal is to determine whether these distributions that generated our observed prediction functions are, in fact, the same. More explicitly, we consider hypotheses of the form

$$\begin{aligned} \mathrm{H}_0^f &: \mathcal{F}_{08-09} = \mathcal{F}_{10-13} \\ \mathrm{H}_1^f &: \mathcal{F}_{08-09} \neq \mathcal{F}_{10-13}. \end{aligned} \tag{2.4.5}$$

where the H^f notation is to distinguish these hypotheses from those in Equation 2.4.1. It should be noted that even under H_0^f, the forests are not exactly exchangeable between groups since the conditioning random variable (the datasets, $\mathcal{D}_{08-09}, \mathcal{D}_{10-13}^{NN}$) are different. As such, there will be stronger dependence within the groups of trees. To ameliorate this,

we impose an additional condition on the construction of the random forests. In particular, instead of bootstrapping, we now subsample observations, i.e. each tree is trained on $k < n$ observations, without replacement. We use a dynamic subsampling rate, with $k_n = n^p$ for some $p \in (0, 0.5)$, so that $\lim_{n\to\infty} k_n/n = 0$. This ensures that the decision trees used are asymptotically independent, which means that the dependence between tree predictions dies off as $n \to \infty$. Thus, the within group and between group dependences approach each other as $n \to \infty$. We note that this is a standard requirement imposed upon random forest construction in the random forest theory, such as in Mentch and Hooker [2016a], Wager and Athey [2018], Scornet et al. [2015b]. The choice of mini-ensembles of size 50 is done to balance predictive accuracy and the higher within sample dependence.

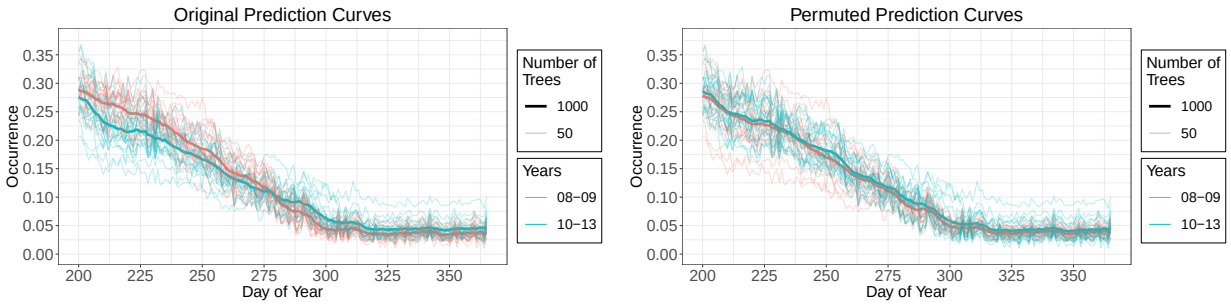
While H_0^f does not imply H_0, rejecting H_0^f supports the notion that migration patterns in the years 2008 and 2009 differed significantly from those observed from 2010-2013. Note that because a random forest is simply an average of decision trees, we can reformulate each of the statistics in Equation 2.4.3 by substituting in $f_{(08-09)}(t) = \frac{1}{20}\sum_{k=1}^{20} f_{k,(08-09)}(t)$ and likewise for $f_{(10-13)}(t)$. Then we shuffle the functional observations between the 2008-2009 group and the 2010-2013 group many times, at each stage calculating the statistics in Equation 2.4.3. That is, to form a permuted random forest, we permute groups of decision trees rather than the data itself, so that we now only have to train $2B$ trees. Similarly, for the temporally segmented test, each restricted functional observation is from some distribution $\mathcal{F}^{(i)}_{08-09}$ or $\mathcal{F}^{(i)}_{10-13}$, leading naturally to hypotheses of the form

$$\begin{aligned} \mathrm{H}^f_{0,i} &: \mathcal{F}^{(i)}_{08-09} = \mathcal{F}^{(i)}_{10-13} \\ \mathrm{H}^f_{1,i} &: \mathcal{F}^{(i)}_{08-09} \neq \mathcal{F}^{(i)}_{10-13}. \end{aligned} \tag{2.4.6}$$

We take advantage of the bagging structure inherent to random forests, but the same framework, which we refer to as the *functional* permutation test, could be applied to any bagged learner.

2.4.2 Global Test Results

To implement the canonical permutation test, we utilize the `randomForest` package in `R` to calculate the RF predictions [Liaw and Wiener, 2002]. As in Section 2.3.1, the RFs

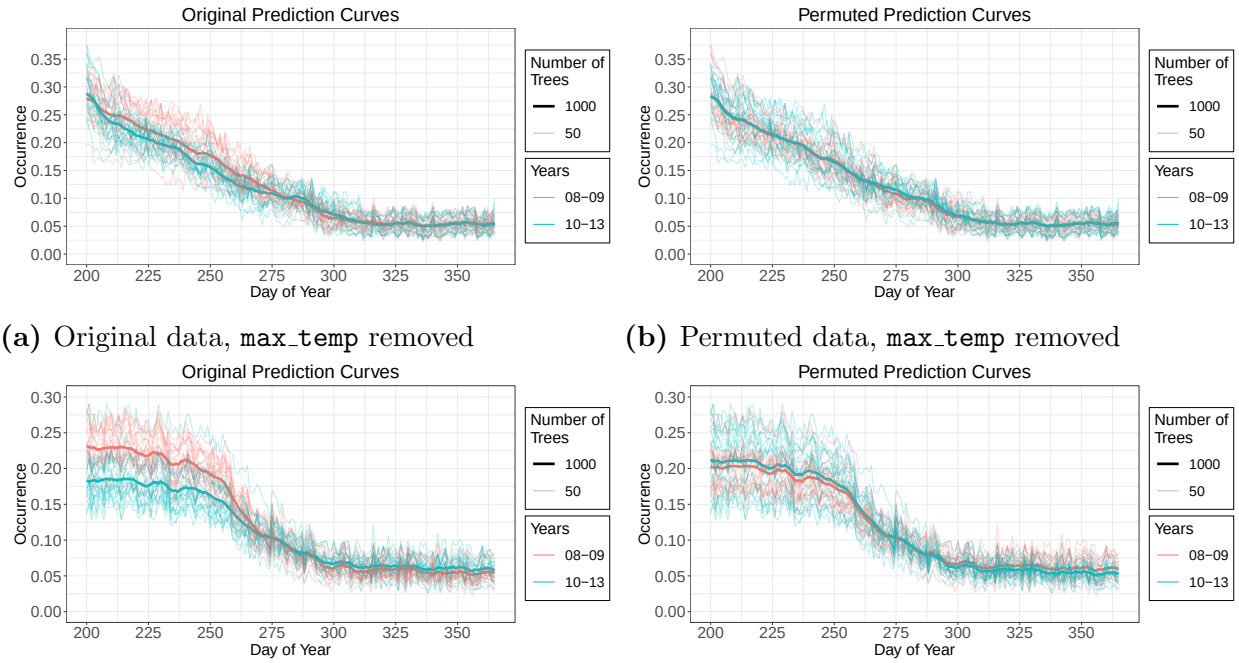


(a) f_{08-09}, f_{10-13}, functional data unpermuted **(b)** f_{08-09}, f_{10-13}, functional data permuted

Figure 2.4.1: Prediction curves generated by RFs which use all covariates, including `max_temp` and `DoY`. These are used in the functional permutation test. Lighter lines show the collection of functional data, darker lines show average that forms the full RF function.

are trained on the entire set of predictors, including both `DoY` and `max_temp`. To account for the correlation between `max_temp` and `DoY`, we conduct two additional followup versions of the original permutation tests: one with `DoY` included and `max_temp` removed and one with `max_temp` included and `DoY` removed. The functional permutation test is conducted using 20 functional observations in each group, using a subsampling rate of $a_n = n^{0.55}$ and setting `mtry` $= 7$. These constraints on the tree construction worsens the predictions of the individual models, but further weakens the dependence between the functional data.

The p-values obtained from the canonical permutation test are shown in the second row of Table 2.4.2. Based on these results alone, there does not appear to be strong evidence of a difference in the underlying functional distributions, even early in the migration period. However, a more compelling story appears in the results of functional permutation test. The associated functions, $\{f_{k(08,09)}\}_{k=1}^{20}$ and $\{f_{k,(10-13)}\}_{k=1}^{20}$, along with their averages, are shown in Figure 2.4.1, along with an example of a permutation of the functional data. Based on a visual inspection, it appears that for RFs trained on the original 2008-2009 data ($\mathcal{D}_{08-09}$) and the reduced nearest neighbor 2010-2013 data ($\mathcal{D}_{10-13}^{NN}$), $f_{08-09} > f_{10-13}$ until around `DoY` 280, with negligible differences thereafter. The p-values from the functional permutation test are presented in Table 2.4.2, and provide strong evidence for a difference in migration patterns. In particular, we are able to reject H_0^f, for the full feature and `max_temp` models at any reasonable level α.



(a) Original data, `max_temp` removed

(b) Permuted data, `max_temp` removed

(c) Original data, `DoY` removed

(d) Permuted data, `DoY` removed

Figure 2.4.2: Prediction curves generated by RFs with `DoY` included and `max_temp` removed (top row, (a) and (b)) and with `max_temp` included and `DoY` removed (bottom row, (c) and (d)). Lighter lines show collection of functional data, darker lines show average that forms the full RF function.

The RF prediction functions corresponding to the `max_temp` only and `DoY` only models, with both the original and (randomly selected) permuted datasets, are shown in Figure 2.4.2. Here we begin to see evidence for the importance of `max_temp`: when only `DoY` is used as a predictor, the prediction functions trained on the original datasets closely resemble those trained on the permuted data. However, when `max_temp` is included and `DoY` is removed, we see a clear difference in predicted occurrence until midway into the migration season, a story which closely matches the anecdotal accounts from the ornithological community. Moreover, the raw statistic values in Table 2.4.1 show that the greatest differences (however those differences are measured) are consistently observed in the `max_temp` model.

Test Statistic	KS	Δ_1	Δ_2	Δ_3
`DoY` & `max_temp` included	0.05112	0.02419	**-0.00741**	**-0.00816**
`DoY` included; `max_temp` removed	0.02718	0.01530	0.00037	0.00096
`max_temp` included; `DoY` removed	**0.05329**	**0.03651**	-0.00170	-0.00533

Table 2.4.1: Observed values for each of the test statistics in Equation 2.4.3. Bolded values are the largest magnitude differences for each test.

Test Statistic	KS	Δ_1	Δ_2	Δ_3	KS	Δ_1	Δ_2	Δ_3
Null Hypothesis Tested	H_0	$H_{0,1}$	$H_{0,2}$	$H_{0,3}$	H_0^f	$H_{0,1}^f$	$H_{0,2}^f$	$H_{0,3}^f$
`DoY` & `max_temp` included	0.196	0.075	0.768	0.829	0.002	0.007	0.103	0.954
`DoY` included; `max_temp` removed	0.122	0.034	0.544	0.163	0.565	0.182	0.676	0.677
`max_temp` included; `DoY` removed	0.001	0.000	0.299	0.775	0.070	0.055	0.020	0.711

Table 2.4.2: P-values for the canonical permutation test (left) and functional test (right); Tests are done with all covariates included in the datasets (row three), `DoY` included and `max_temp` removed (row four), and `max_temp` included and `DoY` removed (row five).

The p-values resulting from the followup tests appear to tell a similar story, in both the functional and canonical permutation test. From Table 2.4.2 we see that when `max_temp` is removed, the smallest p-value is only 0.182 corresponding to the test for raw differences in time period one as measured with Δ_1. However, when `max_temp` is included and `DoY` removed, the largest p-value among the first four tests is only 0.07. The p-value from the final test for raw differences in the third time period (measured by Δ_3) is large at 0.711, but recall that this is what was expected as the prediction curves appear very similar in all cases late in the season. A similar pattern appears in the p-values in Table 2.4.2.

Before continuing with the localized tests the in the following section, we acknowledge that the p-values from the canonical permutation tests, though reasonably small and substantially lower than in the other tests, fail to surpass the commonly accepted $\alpha = 0.05$ threshold in all but one instance and too large to conclude that migration patterns differed significantly in the two sets of years. However, note that these tests are conservative in two ways. First and most obviously, permutation tests themselves suffer lower power than

their parametric counterparts. More subtly, the RF prediction curves generated by the data from 2010-2013 were not trained on the full available dataset, but were trained on a carefully selected subset $\mathcal{D}^{NN}_{10-13}$ designed to spatiotemporally mimic the observations collected in 2008-2009. Given this, it is reasonable to interpret the results of the canonical test as providing at least moderate evidence for a difference in migration patterns that is influenced by `max_temp`. The same patterns appear in the functional test results, in greater magnitude, providing stronger evidence of a yearly difference in occurrence patterns. The localized tests in the following section allow for a more direct means of measuring the precise questions of interest and provide more decisive evidence.

2.4.3 Random Forests as U-Statistics

Our localized tests rely on the work of Mentch and Hooker [2016a] from which we now briefly review some key results. Suppose we have a training sample $\mathcal{D} = \{Z_1, ..., Z_n\}$ consisting of n iid observations from some distribution F_Z with which we construct a (possibly randomized) ensemble consisting of m base learners, each built with a subsample of size k, and use this ensemble to predict at some location X. Denote each base learner by h so that we can write the expected prediction as $\theta_k = \theta_k(X) = \mathbb{E}h(X; Z_1, ..., Z_k)$ and the (empirical) ensemble prediction as $\hat{\theta}_k = \hat{\theta}_k(X) = \frac{1}{m}\sum_{i=1}^{n} h(X; Z_1^*, ..., Z_k^*)$ where each collection $(Z_1^*, ..., Z_k^*)$ represents a subsample of size k from $\mathcal{D}$. Then, under some regularity conditions introduced in Mentch and Hooker [2016a] later weakened in Peng et al. [2019],

$$\frac{\sqrt{m}(\hat{\theta}_k - \theta_k)}{\sqrt{\frac{k^2}{\alpha}\zeta_1 + \zeta_k}} \xrightarrow{d} \mathcal{N}(0, 1) \tag{2.4.7}$$

where $\alpha = \lim_{n\to\infty} n/m$ and the other variance parameters are of the form

$$\zeta_c = \text{cov}(h(Z_1, ..., Z_c, Z_{c+1}, ..., Z_k), h(Z_1, ..., Z_c, Z'_{c+1}, ..., Z'_k)) \tag{2.4.8}$$

for $1 \leq c \leq k$ and where $Z'_{c+1}, ..., Z'_k$ denote additional iid observations from F_Z.

Importantly, this result can be utilized to construct formal hypothesis tests of variable importance. Suppose we have p covariates $X_1, ..., X_p$ and we want to test the predictive

importance (significance) of X_1. Let $\hat{d}_i = R(X_i) - R_{\pi_1}(X_i)$ denote the difference in predictions between a forest trained on $\mathcal{D}$ and another trained on $\mathcal{D}_{\pi_1}$ in which X_1 is permuted.

Consider N such prediction points with differences denoted by $\hat{d} = (\hat{d}_1, ..., \hat{d}_N)^T$. Mentch and Hooker [2016a] show that when the same subsamples are used to construct the trees in each random forest, the differences are infinite-order U-statistics and thus follow the asymptotic distribution in Equation 2.4.7. Let $\hat{\Sigma}_d$ be the estimated covariance matrix of the $\hat{d}_i$. Then, given our vector of pointwise differences, $\hat{d}^T \hat{\Sigma}_d^{-1} \hat{d} \sim \chi^2_N$ and we can use this as a test statistic to formally evaluate the hypotheses

$$\begin{aligned} \mathrm{H}_0 : &\quad \mathbb{E}R(\boldsymbol{x_i}) = \mathbb{E}R_{\pi_1}(\boldsymbol{x_i}) \quad \text{for all } i \in \{1, ..., N\} \\ \mathrm{H}_1 : &\quad \mathbb{E}R(X) \neq \mathbb{E}R_{\pi_1}(X) \quad \text{for some } i \in \{1, ..., N\} \end{aligned} \tag{2.4.9}$$

where the expectation is taken with respect to the training data and randomization. This procedure naturally extends to the more general case where any subset of the features is tested for significance by simply permuting that entire subset of features. Furthermore, this procedure remains valid whenever those features are simply removed from the alternative random forest instead of being permuted, though the permutation-based approach is generally considered more robust and reliable [Mentch and Hooker, 2016a].

2.4.4 Local Influence of Maximum Temperature

We return now to the question of determining whether maximum daily temperature can partially explain the different Tree Swallow patterns of occurrence observed in 2008-2009. The global tests in the previous section suggested that `max_temp` may provide information about the interannual variation in occurrence beyond what is provided by seasonal effects alone captured by `DoY`. We therefore want to distill the predictive influence of `max_temp` from that of `DoY`. In this section, we consider testing for the variable importance of `max_temp_anomaly`.

To fit this into the hypothesis testing framework described above, we calculate one subsampled random forest with the original data and another with a permuted version of `max_temp_anomaly`. Note that because the hypotheses in Equation 2.4.9 are evaluated at only fixed test points, careful selection of these points is important. Since we are interested in

evaluating the hypotheses across a variety of locations and times, we stratify our test points by location and conduct 6 different tests. The training and test set used here are selected from points inside wildlife refuge areas. Wildlife refuges are of particular interest because they include areas that are resistant to local environmental changes due to the environmental protections in place, helping to isolate the predictive influence of regional temperature fluctuations on Tree Swallow occurrence. In total, we select 6 groups of 25 test points each, which are subsequently removed from the training set. These 6 groups and points are shown in Figure 2.4.3. Spatial centers for each of these regions were selected based on a high density of observations and the test points were selected uniformly at random from within a 0.3 decimal degree radius. The final training set consists of 25727 observations. We apply the above hypothesis testing procedure at each of our 6 regional test locations, building separate ensembles for each location. As in Section 2.4, we make all features available for splitting at each node in each tree so that our random forest procedure reduces to subsampled bagging (subbagging). These tests are implemented using the `rpart` package in `R` to construct the regression trees [Therneau et al., 2017]. Keeping with the recommendations of Mentch and Hooker [2016a], we take our subsample size to be $k = 160 \approx \sqrt{25727}$; in general, larger subsamples can be used if base learners are constructed in an alternative fashion to comply with honesty and regularity conditions [Wager and Athey, 2018]. We build 1.25×10^7 trees for each ensemble, to attain high precision in the estimation of the covariance matrix, $\hat{\Sigma}_d$.

Table 2.4.3 summarizes the test statistics and p-values obtained from the tests in each region. These local tests for the significance of `max_temp_anomaly` suggest that the anomaly is predictive of occurrence in testing locations 2-5, with less significance in location 1 and no significance in location 6. These suggest a transition zone within BCR30 between testing locations 1 & 6 where `max_temp_anomaly` is important in predicting occurrence. North of this zone, temperatures may be too cold to allow insect activity in the fall and south of this zone, temperatures may be warm enough to allow insect activity year round.

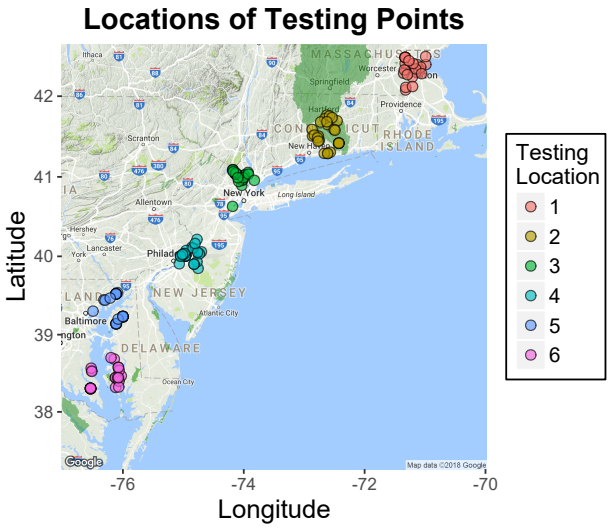


Figure 2.4.3: Test points selected for study. Light green overlay indicates Fish and Wildlife Service wildlife refuges.

Testing Location	Test Statistic	p-value
1	38.07	0.04553
2	58.12	1.889E-04
3	58.21	1.835E-04
4	62.14	5.275E-05
5	59.93	1.068E-04
6	28.44	0.2880

Table 2.4.3: Test statistics and p-values for the hypotheses in Equation 2.4.9 at the points show in Figure 2.4.3.

2.5 Discussion

2.5.1 Ornithological Implications

Our goal in this work was to thoroughly examine Tree Swallow migration patterns from recent years and to examine the role temperature changes may have had in explaining differences among years. The global hypothesis tests evaluated over the entire BCR30 region in Section 2.4 provided evidence for the hypothesis that the seasonal patterns of distributions indeed differed in the years 2008 and 2009. The fact that this difference no longer seemed apparent whenever `max_temp` was excluded as a predictor supports the hypothesis that temperature plays an important role explaining year-to-year variation in occurrence. While these conclusions examine only the average region-wide effect, the corresponding localized hypothesis tests carried out at specific locations along the Tree Swallow migration route in Section 2.4.4 provide formal justification for the importance of maximum temperature beyond being merely a correlate of some other seasonally varying effect. These results are

especially important in the context of climate change, providing the first statistically sound evidence from eBird data that variation in ambient temperatures is related to the mortality and/or migration of a wild bird, supporting conclusions of other ecological work La Sorte et al. [2016].

2.5.2 Methodological Discussion

Finally, it is important to note that the procedures we implemented in addressing our hypotheses of interest are very general. Exactly the same approach could be taken to investigate distribution dynamics for any species as well as for far more general problems completely outside of an ecological context. Fundamentally, our problem involved assessing and characterizing the significance of a subset of available predictor variables in which the underlying regression function was believed to consist of nonlinear and complex interaction terms which, localized in predictor space, precluded the *a priori* specification of a suitable parametric model for which traditional inference methods may have been available. We present two forms of black box inference: development of non-parametric permutation-style tests that are agnostic to the underlying procedure, and asymptotic results about the predictions of ensemble learners. The recent asymptotic results for infinite-order U-statistics allowed us to model the data through a series of flexible but complex black-box models – in our case, regression trees – while retaining the ability to formally characterize results. We also present a classical statistical argument for the validity of the functional permutation test. To our knowledge, this is the first time the connection between bagged models and exchangeability has been noted, and used to form a framework for valid hypothesis testing. This procedure maintains asymptotic validity (in terms of controlling Type I error), and requires training substantially fewer trees than the canonical permutation test. We also note that this procedure, despite being non-parametric, requires far fewer trees than even the theoretical results of Mentch and Hooker [2016a] and Wager and Athey [2018], making it a much more practical tool for our case study. In the next chapter, we discuss an extension of this procedure to testing other hypotheses, which marks the main methodological contribution of this thesis.

3.0 Permutation Tests For Ensemble Methods

3.1 Introduction

Advances in computing power and big data collection have produced numerous situations in which complex supervised learning methods can drastically outperform more rigid classical statistical models in terms of predictive accuracy. Despite these advances, many such models and algorithms are largely impenetrable to traditional statistical analysis. The random forests algorithm [Breiman, 2001a] is among the relatively few supervised procedures for which formal statistical properties have recently been developed, paving the way for inference procedures. As detailed below, however, methods proposed to this point for assessing variable importance have either been *ad hoc* and susceptible to producing misleading and inconsistent results even in simple settings or have come with severe restrictions on the testing framework while incurring extreme computational overhead. The primary goal of this paper is to develop a formal, statistically valid hypothesis test approach that maintains high power with orders of magnitude fewer required computations that scales naturally and efficiently to large data settings where supervised learning tools like random forests are most likely to be employed in practice.

Our work builds directly on the foundation of permutation tests, which have their roots in the work of Fisher [1937] using contingency tables. Classical work on permutation tests from Hoeffding [1952] and Lehmann et al. [1949] demonstrates the convergence of the permutation distribution to the sampling distribution for a wide variety of test statistics. Much of the modern work has focused on extending permutation tests to situations where the data may not be iid or even exchangeable (e.g. Romano [1990]). Studentization is typically proposed as a means of forcing the sampling distribution of a statistic to converge to a normal distribution to which it is then shown that the permutation distribution also converges. This idea has underpinned results in Neuhaus [1993] and Janssen [2005], who provide various sufficient conditions for the convergence to the unconditional distribution.

Permutation tests are exact tests for hypotheses of equal distribution under the assump-

tion of iid sequences, but are not necessarily valid for more general hypotheses. Convergence to the unconditional distribution ensures that the permutation distribution can be used for a finite sample exact test of equality of distribution and an asymptotically valid test for more general hypotheses. In this work, we prove results regarding the asymptotic validity of our procedure for more general hypotheses. The individual models (base learners) in supervised ensembles, such as decision trees in a random forest, naturally lend themselves to the permutation framework by being exchangeable in many practical cases.

3.1.1 Related Work on Random Forests

Decision trees recursively partition the covariate space and generate predictions by fitting some simple model – often an average or majority vote – within each resulting region. Of particular interest are the classical **C**lassification **A**nd **R**egression **T**rees [Breiman et al., 1984]. CART procedures often have low bias, but can overfit the data without careful pruning. Bagging stabilizes the variance by training many individual learners on bootstrap samples. Random forests [Breiman, 2001a] augment the bagging procedure by introducing auxiliary randomness in the construction of each individual learner, leading to trees with a lower degree of dependence but higher individual variances. Since their introduction, random forests have sustained a long track-record of empirical success in terms of predictive accuracy; see Fernández-Delgado et al. [2014a] for a recent large-scale comparison in which random forests outperform nearly all competitors.

Recent years have seen something of a surge in the development of formal statistical analyses of random forests. Wager et al. [2014b] applied the infinitesimal jackknife variance estimate developed in Efron [2014] to produce closed form variance estimates for random forest predictions. Scornet et al. [2015a] provided the first consistency results for Breiman's original random forest procedure for additive regression functions. Mentch and Hooker [2016a] derived the closed form asymptotic distribution for random forest predictions under restrictions on subsample size. Wager and Athey [2018] proved both consistency and asymptotic normality for subsampled random forests whenever trees are restricted to being built according to honesty and regularity conditions and large numbers of trees are constructed.

The original random forest formulation has also been extended to various setups including quantile regression [Meinshausen, 2006], survival analysis [Ishwaran and Lu, 2008, Cui et al., 2017], reinforcement learning [Zhu et al., 2015], and a generalized framework allowing random forests weights to be used for general local parameter estimation [Athey et al., 2016].

In addition to their robust history of empirical success and these newly-developed statistical properties, the availability of *ad hoc* tools for evaluating variable importance has also been a major contributing factor to their continued widespread practical use. Among these, thanks in large part to their computational feasibility, the out-of-bag (oob) measures proposed by Breiman [2001a] remain the most popular with versions of this measure available in nearly every major statistical software. Unfortunately, in the decades since their introduction, a substantial amount of literature has repeatedly demonstrated their inadequacy and inconsistency; see Strobl et al. [2007a] and Toloşi and Lengauer [2011a] as popular, representative examples. Among the issues with oob measures are a tendency to inflate the relative importance of categorical covariates with many levels as well as those with high correlation to others. The latter issue is particularly problematic as variables deemed most important may have relatively little impact on the response but be highly dependent only on each other. Recent work by Hooker and Mentch [2019] gives an explanation for this behavior based on extrapolation.

In light of these issues, recent work has sought to cast the issue of variable importance more formally in a classical hypothesis testing framework. Notably, Mentch and Hooker [2016a] showed an equivalence between subsampled random forests and infinite-order U-statistics, allowing for asymptotic normality to be established and a hypothesis testing procedure for evaluating variable importance to be proposed. This test, though valid, is quite computationally prohibitive. The hypotheses are presumed to be evaluated at predefined test locations in some test set $\mathcal{T}$ and whenever $|\mathcal{T}| = N_t > 1$, calculating the test statistic involves estimating an $N_t \times N_t$ covariance matrix. Accurate estimation of the covariance necessitates constructing a very large number of trees – exponentially more than would be required for the construction of the random forests themselves – and thus becomes computationally infeasible for more than a few dozen test points, even when the original dataset is relatively small. Mentch and Hooker [2017] extend the procedure to tests for additivity

and provide an alternative approximate test involving random projections that allows the procedure to scale up slightly but with additional computational overhead. Even employing the potentially more efficient infinitesimal jackknife variance estimate utilized in Wager et al. [2014b] and Wager and Athey [2018] requires the number of trees constructed be at least on the order of n to be valid. Thus, while these kinds of procedures can be shown to successfully alleviate the troublesome issues with the classical oob measures, their computational complexity precludes their use in the vast majority of practical settings.

In contrast with these previous approaches, this work develops a formal testing framework for variable importance that is both computationally efficient and statistically valid. In particular, our procedure is almost entirely computationally agnostic to the number of test points utilized. The permutation scheme we employ avoids the need for an explicit covariance estimation and thus does not require a larger number of trees for larger datasets. Instead, our hypothesis tests provide valid p-values for the predictive importance of any given subset of covariates while maintaining the same order of computational complexity as the original random forest procedure. Put simply, if the size and structure of the available data allows for a random forest model to be constructed, our testing procedure can be readily employed. We note also that while our focus here is on random forests, only a small portion of the theory we provide is tree-specific and thus ensembles consisting of other kinds of base learners easily fit within this testing framework as well.

The remainder of this paper is laid out as follows. In Section 4.3, we give an overview of the testing procedure, and further highlight its benefits over existing methods. In Section 3.3, we present results regarding the statistical properties of the proposed test, namely that it attains validity for the desired hypotheses. In Section 3.4, we present simulation studies of the testing procedure for a variety of underlying regression functions, as well as a comparison with two different knockoff statistics. In Section 3.6, we apply our procedure to multiple ecological datasets where random forests have been successfully employed in recent applied work. In addition to the main text, all technical proofs are provided in Appendix B.1, and additional simulations demonstrating the robustness of the proposed procedure are presented in Appendix B.2.

3.2 Overview of the Testing Procedure

Consider a sample $\mathcal{D}_n = \{Z_1, Z_2, ..., Z_n\}$, with $Z_i = (X_i, Y_i)$ consisting of observations on covariates $X = (X_1, ..., X_p) \in \mathcal{X}$ and a response $Y \in \mathcal{Y}$. In this work, it is assumed that $Z_k \overset{iid}{\sim} F$ where F is some distribution with support on $\mathcal{X} \times \mathcal{Y}$. In the regression context, we assume that $Y = m(X) + \epsilon$ where $m(X) = \mathbb{E}(Y|X = X)$ and ϵ is an independent noise process, typically with $\mathbb{E}(\epsilon) = 0$ and $\text{Var}(\epsilon) < \infty$. The goal of the random forest procedure is to accurately estimate $m(X)$. Each tree in a random forest is constructed by drawing subsamples of size $k_n < n$, from $\mathcal{D}_n$, drawing a randomization parameter ξ from some distribution Ξ, and constructing a randomized decision tree. This process is repeated B times and the random forest prediction at some point $X \in \mathcal{X}$ is given by

$$RF_{B,k_n}(X) = \frac{1}{B}\sum_{j=1}^{B} T_{k_n,j}(X; \xi_j; \mathcal{D}_n). \tag{3.2.1}$$

To evaluate the RF prediction accuracy at a test location X with true response value y, we can measure the mean squared error

$$MSE_{RF}(X; y, \mathcal{D}_n) = \left(\left(\frac{1}{B}\sum_{j=1}^{B} T_{k_n,j}(X)\right) - y\right)^2$$

where we have suppresed some notation for convenience. Similarly, we can write the MSE of a forest at a collection of N_t test points $\mathcal{T}$ as $MSE_{RF}(\mathcal{T}) = \frac{1}{N_t}\sum_{\ell=1}^{N_t} MSE_{RF}(X_\ell; Y_\ell, \mathcal{D}_n)$.

Let RF^π be defined similarly to Equation 3.2.1, but with $\mathcal{D}_n$ replaced by $\mathcal{D}_n^\pi$, where $\mathcal{D}_n^\pi$ replaces some subset of features with an alternate copy drawn independent of Y given the rest of the covariates. To make this concrete, suppose that this subset consists of just a single feature X_j. We can then evaluate whether X_j is important by conducting a test of the following hypotheses

$$\begin{aligned} H_0^j &: \mathbb{E}(MSE_{RF}(\mathcal{T})) = \mathbb{E}(MSE_{RF^\pi}(\mathcal{T})) \\ H_1^j &: \mathbb{E}(MSE_{RF}(\mathcal{T})) < \mathbb{E}(MSE_{RF^\pi}(\mathcal{T})) \end{aligned} \tag{3.2.2}$$

where the expectation is taken over the training data and auxiliary randomness. Though conditional on $\mathcal{T}$, we stress that the computationally complexity of the testing procedure we

employ is almost entirely immune to the size of this test set, effectively allowing practitioners to evaluate the hypothesis at as many locations as are desired. We call X_j important if we are able to reject H_0^j, and correspondingly measure its importance as the difference in MSEs, $MSE_{RF^\pi}(\mathcal{T}) - MSE_{RF}(\mathcal{T})$. This definition of importance is model based and therefore different than alternative definitions such as that utilized in the recent *knockoff* literature [Barber et al., 2015, Candes et al., 2016], where a variable X_j is deemed unimportant if

$$(Y \perp\!\!\!\perp X_j) \mid X_{-j}.$$

The standard knockoff procedure controls the False Discovery Rate (FDR) for hypotheses about each of the covariates for arbitrary distributions over (X, Y). It should be noted that conditional independence of X_j and Y is neither necessary nor sufficient for H_0^j. However, in practice, the test statistic utilized in the knockoff procedure is generally taken as the difference in importance measures between original and knockoff variables and thus the outcome of the procedure itself remains highly model dependent. We also note that our procedure, while it could use knockoff variables as the alternate random copies in $\mathcal{D}_n^\pi$, does not require knowledge of the distribution of the covariates to maintain validity.

3.2.1 Testing Procedure

Intuitively, if two randomized ensemble methods produce predictions that are similarly accurate, then the permutation distribution of discrepancies in accuracy should be centered around 0. In our particular setting for testing feature significance, we compare the accuracy of two ensembles built on different data. For a given (original) dataset $\mathcal{D}_n$, we first construct $\mathcal{D}_n^\pi$ in such a way so as to remove any dependence of response on these features. However, rather than permuting the data and retraining entire random forests, we first train trees on both $\mathcal{D}_n$ and $\mathcal{D}_n^\pi$ separately, record predictions at the test locations, and then permute the predictions (trees) between the forests. The new forests formed at each iteration thus consist of some trees built on the original data and some built with the permuted counterpart. In this light, the testing procedure can be seen as directly analogous to a classic permutation test to evaluate equality in distribution across two groups. Importantly, this procedure requires only $2B$ trees, regardless of the size of the test set.

Pseudo-code for the permutation test is provided in Algorithm 1. We use $\oplus$ to denote concatenation of data matrices by column, $\uplus$ to denote concatenation by row, and $\ominus$ to denote the removal of columns from a dataset. In order to prevent p-values exactly equal to 0, we add 1 to the numerator and denominator, ensuring that under H_0 the p-values are stochastically larger than uniform random variables. This suffices to make the testing procedure slightly more conservative, but more amenable to potential p-value transforming procedure, like an FDR filter; see, for example, Phipson and Smyth [2010] for a more thorough discussion. Crucially, note that this procedure requires no explicit variance estimation of the N_t predictions made by individual forests and thus requires only $2B$ trees regardless of the size of the test set. This provides a very dramatic computational speed-up over existing parametric approaches [Mentch and Hooker, 2016a, 2017] that require the estimation of a $N_t \times N_t$ covariance matrix, which, in turn, requires the construction of exponentially more trees beyond what is needed to construct the original forests.

Algorithm 1: Permutation test pseudocode for variable importance

Data: Training data $\mathcal{D}_n$ test sample ($\mathcal{T} = [(X_1, y_1), ..., (X_{N_t}, y_{N_t})]$), specified feature(s) of interest, X_S, N_0 number of permutations to evaluate

Result: p-value, $\tilde{p}$ for importance of X_S at points in $\mathcal{T}_n$

SET number of permutations n_{perm}, subsample size k_n, and $n_{tree} = B$;

DEFINE X_S^π by permuting the rows of $\mathcal{D}_n$ and selecting the columns corresponding to X_S ;

DEFINE $\mathcal{D}_n^\pi = \mathcal{D}_n \ominus X_S \oplus X_S^\pi$;

for i *in* $\{1, ..., B\}$ **do**

  SAMPLE k_n rows from $\mathcal{D}_n$: $\mathcal{D}_i^* = \{Z_{i,1}^*, ..., Z_{i,k_n}^*\}$;

  SAMPLE k_n rows from $\mathcal{D}_n^\pi$: $\mathcal{D}_i^{*\pi} = \{Z_{i,1}^{*\pi}, ..., Z_{i,k_n}^{*\pi}\}$;

  TRAIN trees $T_i(\cdot)$ on $\mathcal{D}_{i,k_n}^*$ and $T_i^\pi(\cdot)$ on $\mathcal{D}_{i,k_n}^{*\pi}$;

  PREDICT at $\mathcal{T}_n$ using T_i, T_i^π, generating $\boldsymbol{T}_i = [T_i(X_1), ..., T_i(X_{N_t})]$ and $\boldsymbol{T}_i^\pi = [T_i^\pi(X_1), ..., T_i^\pi(X_{N_t})]$

end

CALCULATE $MSE_0 = \frac{1}{N_t}\big|\big|\frac{1}{B}\sum_{i=1}^B \boldsymbol{T}_i - \boldsymbol{y}\big|\big|_2^2$ and $MSE_0^\pi = \frac{1}{N_t}\big|\big|\frac{1}{B}\sum_{i=1}^B \boldsymbol{T}_i^\pi - \boldsymbol{y}\big|\big|_2^2$;

for j *in* $\{1, ..., N_0\}$ **do**

  SAMPLE $\boldsymbol{T}_{j,1}^*, ..., \boldsymbol{T}_{j,B}^*$ from $\{\boldsymbol{T}_1, ...\boldsymbol{T}_B, \boldsymbol{T}_1^\pi, ..., \boldsymbol{T}_B^\pi\}$ without replacement, call the B remaining trees $\boldsymbol{T}_{j,1}^{*\pi}, ..., \boldsymbol{T}_{j,B}^{*\pi}$;

  CALCULATE $MSE_j^* = \frac{1}{N_t}\big|\big|\frac{1}{B}\sum_{l=1}^B \boldsymbol{T}_{j,l}^* - \boldsymbol{y}\big|\big|_2^2$ and $MSE_j^{*\pi} = \frac{1}{N_t}\big|\big|\frac{1}{B}\sum_{l=1}^B \boldsymbol{T}_{j,l}^{*\pi} - \boldsymbol{y}\big|\big|_2^2$

end

CALCULATE $\tilde{p} = \frac{1}{N_0+1}\left[1 + \sum_{j=1}^{N_0} I\big((MSE_0^\pi - MSE_0) \leq (MSE_j^{*\pi} - MSE_j^*)\big)\right]$

3.3 Establishing Statistical Validity

We now develop the theoretical backing for the hypothesis testing procedure outlined above. We note upfront that while the idea behind the procedure – shuffling trees between forests to carry a test analogous to a two-group permutation test – is relatively straightforward, a substantial amount of technical derivation is required to establish its validity and thus we now provide something of a roadmap for the following subsections. In Subsection 3.3.1, we make explicit the connection between bagged models and exchangeable random variables and build upon this in Subsection 3.3.2 to establish asymptotic normality for subsampled random forest predictions. Asymptotic normality of individual predictions is then used to establish a central limit theorem in Subsection 3.3.3 for the difference in MSEs between two forests. In theory, knowledge of this sampling distribution is sufficient to formally evaluate hypotheses of the form in Equation 3.2.2. However, as already alluded to, estimating the variance of that distribution would require the construction of enormous numbers of additional trees, becoming computationally infeasible even for relatively small test sets. Thus, in Subsection 3.3.4, we show that our proposed permutation test is asymptotically equivalent to this parametric alternative, thereby allowing for formal hypothesis tests for feature importance to be carried out while maintaining the same order of computational magnitude as the construction of a typical random forest. For readability, technical discussions and proofs are reserved for Appendix B.1.

3.3.1 Exchangeable Random Variables & Permutation Tests

Recall that a sequence of random variables $X_1, X_2, ...$ is exchangeable if $(X_{i_1}, X_{i_2}, ..., X_{i_k}) \stackrel{d}{=} (X_{\pi(1)}, X_{\pi(2)}, ..., X_{\pi(k)})$ for every finite sub-collection indexed by $i_1, ..., i_k$ and every permutation of the indices $\pi(\cdot)$. Permutation tests naturally lend themselves to exchangeable data by providing a means of evaluating the hypothesis that the joint distribution of a collection of random variables is invariant under permutations. They maintain exactness for the null hypothesis whenever $X_i \stackrel{iid}{\sim} P$ and independently $Y_j \stackrel{iid}{\sim} Q$ because the

joint measure of the data factorizes as

$$\mu(X_1, ..., X_n, Y_1, ..., Y_m) = \prod_{i=1}^{n} P(X_i) \prod_{j=1}^{m} Q(Y_j)$$

which is invariant to permutations of observations if and only if $P = Q$.

Modern work for permutation tests has focused largely on modifications needed to account for violations of the exchangeability assumption. Chung and Romano [2013] propose a studentization of the permutation test statistic when conducting inference on a functional of two distributions. Consider, for example, a two sample problem, with $X_1, ..., X_n \overset{iid}{\sim} P_X = \mathcal{N}(0, 5)$ and independently let $Y_1, ..., Y_m \overset{iid}{\sim} P_Y = \mathcal{N}(0, 1)$. Clearly, median$(P_X)$ = median(P_Y), but the data are no longer exchangeable and so an unstudentized permutation test of H_0 : median(P_X) = median(P_Y) is no longer valid at a pre-specified level. However, note that exchangeability is violated only because the data are no longer identically distributed; permutation tests can remain valid for data that are correlated but identically distributed so long as the pairwise dependence is constant. The upshot of this is that random forest ensembles possess this property, and thus can be shown to be exchangeable as formalized in Theorem 1.

Theorem 1. *Denote a sequence of (potentially randomized) subsampled trees as* $\{T_k(\cdot)\}_1^\infty$. *Under the conditions outlined above, the residuals at* $\boldsymbol{Z}^* = (X^*, Y^*) \sim F$ *given by*

$$r_k = T_k(X^*) - Y^*$$

form an infinitely exchangeable sequence of random variables.

In the case of a single random forest, exchangeability is readily apparent as the order in which trees are trained has no bearing on their structure. Indeed, Theorem 1 can be extended to any bagged learning method.

Given a dataset $\mathcal{D}_n$ with $n \times p$ design matrix X, let $\mathcal{S} \subset \{1, ..., p\}$ and define $X_\mathcal{S} = \{X_j : j \in \mathcal{S}\}$ and $X_{-\mathcal{S}} = \{X_j : \ j \notin \mathcal{S}\}$ where $X_\mathcal{S}$ consists of the covariates that we seek to test for importance. We then create a randomized version of $X_\mathcal{S}$ independent of Y, denoted by $X_\mathcal{S}^\pi$. Note in particular that when the entire joint density $P(X)$ of the covariates is known, Algorithm 1 of Candes et al. [2016] can be used to generate the knockoffs that make up

$X_{\mathcal{S}}^{\pi}$ which then ensures that $[X_{-\mathcal{S}}, X_{\mathcal{S}}] \stackrel{d}{=} [X_{-\mathcal{S}}, X_{\mathcal{S}}^{\pi}]$. By construction, $X_{\mathcal{S}}^{\pi}$ is independent of $Y|X_{-\mathcal{S}}$ and consequently, if we replace $X_{\mathcal{S}}$ with $X_{\mathcal{S}}^{\pi}$ in the design matrix to form a new training dataset $\mathcal{D}_n^{\pi}$, then the trees trained on $\mathcal{D}_n^{\pi}$ inherit the conditional independence so that $T(X; \mathcal{D}_n^{\pi})$ is independent of $Y|\ X_{-\mathcal{S}}$, allowing for the testing of a null hypothesis of conditional independence.

3.3.2 Asymptotic Behavior of Trees

Within-forest exchangeability is not sufficient to justify the proposed testing procedure at the nominal level. Instead, we need to establish sufficient conditions to justify exchanging trees between forests. An important step in this direction is to establish the existence of a limiting sequence of subsampled trees that behave like an iid sequence.

Condition 1. *There exists a random function* T_∞ *such that* $\lim_{n\to\infty} T_{k_n} \stackrel{d}{=} T_\infty$

Later, we provide sufficient conditions for this to hold. We note that this condition is similar in spirit to Assumption 15.7.1 in Lehmann and Romano [2006], which is fundamental to the validity of subsampling based intervals for model parameters.

In practice, we would like to establish results for random forests trained on growing subsamples. If we insist that the subsample size k_n grow slower than $\sqrt{n}$, we obtain the following intuitive result.

Lemma 1. *Consider a collection of* B_n *trees* $\{T_{j,k_n}\}_{j=1}^{B_n}$ *built from a training set of size* n *on subsamples of size* k_n *satisfying Condition 1. Then, as long as* $k_n/\sqrt{n} \to 0$ *and*

$$\binom{B_n}{2} \log\left[\frac{\binom{n-k_n}{k_n}}{\binom{n}{k_n}}\right] \to 0$$

the infinite sample sequence of trees $\{T_{1,\infty,k_\infty}, ..., T_{B,\infty,k_\infty}, ...\}$ *is an infinite sequence of pairwise independent random functions.*

The condition on the number of trees B_n is likely not of much practical importance. For finite B_n, the probability sequence has the form of a_n^K, so because $a_n \to 1$, a_n^K also converges to 1. However, if we let B_n grow with n, the number of trees may overwhelm the

independence induced by subsampling. Thus, we must let the log probability of an individual pair being independent go to 0 faster than $\binom{B_n}{2} \approx B_n^2/2$ goes to infinity.

Lemma 1 establishes asymptotic pairwise independence, but not that the limiting sequence is iid. For this, we turn to a result from Aldous [1985].

Lemma 2. *[Aldous, 1985] Let $Z_1, Z_2, ...$ be an infinitely exchangeable sequence. If $Z_i \perp\!\!\!\perp Z_j, i \neq j$, then $Z_1, Z_2, ...$ is a sequence of iid random variables.*

An immediate consequence of the preceding lemmas is the following corollary.

Corollary 1. *Let $\{T_{j,k_n}\}_{j=1}^{B_n}$ be a collection of B_n trees trained on subsamples from $\mathcal{D}_n$, satisfying the conditions of Lemma 1. Then, $\{T_{j,\infty}\}_{j=1}^{\infty} := \lim_{n\to\infty}\{T_{j,k_n}\}_{j=1}^{B_n}$ is an iid sequence of functions.*

The infinite sequence of subsampled trees enjoys many properties that the finite sequence does not. In particular, we can obtain the following pointwise central limit theorem.

Corollary 2. *Let $\{T_{j,k_n}\}_{j=1}^{B_n}$ be a sequence of trees on subsamples from $\mathcal{D}_n$, satisfying Condition 1 and the conditions of Lemma 1. Further, assume $X \in \mathcal{X}$ is such that $0 < Var(T_\infty(X)) = \sigma^2(X) < \infty$. Then as $n \to \infty$*

$$\sqrt{B_n}\left[\frac{1}{B_n}\sum_{i=1}^{B_n} T_{i,k_n}(X) - \mathbb{E}\left(\frac{1}{B_n}\sum_{i=1}^{B_n} T_{i,k_n}(X)\right)\right] \xrightarrow{d} \mathcal{N}(0, \sigma^2(X)). \tag{3.3.1}$$

Corollary 2 follows directly from applying the Central Limit Theorem to the sequence of univariate random variables $\{T_{j,\infty}(X)\}_{j=1}^{\infty}$, which are iid by Corollary 1.

Remark. For a collection of test points, $X_1, ..., X_{N_t}$, we can also consider the sequence of vectors $\boldsymbol{T}_{i,k_n} = [T_{i,k_n}(X_1), ..., T_{i,k_n}(X_{N_t})]^T$, which are iid by Corollary 1. If we assume that $\Sigma = \mathbb{E}\big[(\boldsymbol{T}_{i,k_n} - \mathbb{E}(\boldsymbol{T}_{i,k_n}))(\boldsymbol{T}_{i,k_n} - \mathbb{E}(\boldsymbol{T}_{i,k_n}))^T\big]$ has finite entries, the multivariate central limit theorem gives that as $n \to \infty$

$$\sqrt{B_n}\left[\frac{1}{B_n}\sum_{i=1}^{B_n} \boldsymbol{T}_{i,k_n} - \mathbb{E}\left(\frac{1}{B_n}\sum_{i=1}^{B_n} \boldsymbol{T}_{i,k_n}\right)\right] \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

Remark. We can generalize the independence results to a collection of two sets of trees. In particular, suppose that we now train $B_n/2$ trees on $\mathcal{D}_n = \{Z_i\}_{i=1}^n$ and $\mathcal{D}_n^\pi = \{Z_i^\pi\}_{i=1}^n$, where $Z_i^\pi = ([X_\mathcal{S}, X_{-\mathcal{S}}^\pi]_i, Y_i)$. Note that $Z_i^\pi \perp\!\!\!\perp Z_j, \forall\ i \neq j$, so there is the same independence structure between the datasets as within. Thus, the probability that a pair of trees trained on subsamples of size k_n, one from $\mathcal{D}_n$ and one from $\mathcal{D}_n^\pi$, are independent is the same as the probability that a pair of trees within forest are independent. As such, $\{T_{i,k_n}(X)\}_{i=1}^{B_n}$ and $\{T_{i,k_n}^\pi(X)\}_{i=1}^{B_n}$, where B_n, k_n satisfy the conditions of Lemma 1, behave like two independently iid samples.

We intentionally leave $\sigma(X)$ as an abstraction in Corollary 2 since estimation of $\sigma(X)$ is not straightforward. Instead, this result will be used as the basis for asymptotic validity of our permutation test which, uncharacteristically, is far more computationally efficient. Going forward, we consider the asymptotic case, so that the sequence of tree predictions behaves like an iid sequence. Further, in the infinite sample case, the number of trees can be made arbitrarily large, and so we allow B to go to infinity with the understanding that it does so in such a way that respects the requirements of Lemma 1. This is largely a matter of notational convenience; we could explicitly include the dependence on n in each of the following statements and stress that the limiting distributions only hold as $n \to \infty$.

3.3.3 Asymptotic Distribution of MSEs

The previous subsection established asymptotic normality of subsampled random forest predictions. Here we build upon those results to establish asymptotic normality for MSEs of random forest predictions as well as the difference in MSEs between two random forests. We conclude the subsection by providing conditions under which trees conform to the conditions necessary to obtain that asymptotic normality. In the following we provide a high-level discussion in addition to the key results. A more technical discussion is reserved for the appendix.

To begin, consider a single test point (X, y). We can write the MSE as

$$MSE_{RF}(X; y) = g\left(\frac{1}{B}\sum_{i=1}^{B} T_i(X), y\right) \tag{3.3.2}$$

where $g(a,b) = (a-b)^2$. The asymptotic distribution of the MSE can be derived via the delta method. Appeals to the mean value theorem and the law of large numbers gives that the MSE is asymptotically a linear function of the random forest prediction. These derivations, discussed in more detail in Appendix B.1, yield the following result, given in terms of a general function g.

Lemma 3. *Assume the conditions needed from Corollary 2. Additionally, assume that g has at least k derivatives for some $k \geq 3$, and that $g^{(k)}(x) < \infty$ for all x. Further, assume that $\mathbb{E}|T_i(X)|^k < \infty$. Then,*

$$\sqrt{B}\left[\mathbb{E}g(RF_B(X)) - g(\mathbb{E}RF_B(X))\right] = \frac{g''(\mathbb{E}RF_B(X))\sigma^2}{2\sqrt{B}} + o(B^{-3/2}) = o(1).$$

Since the MSE function defined as $g(RF_B(X)) = (RF_B(X) - y)^2$ satisfies the conditions posited by Lemma 3, we can conclude that

$$\sqrt{B}\left[g\left(RF_B(X)\right) - \mathbb{E}g\left(RF_B(X)\right)\right] \xrightarrow{d} \mathcal{N}\left(0, g'(\mathbb{E}RF_B(X))^2\sigma^2\right).$$

Remark. Corollary 2 is not a necessary prerequisite for obtaining the asymptotic normality of the MSE. In fact, a similar argument could be used to justify the asymptotic normality of the MSE for any random forest who satisfies a central limit theorem and a law of large numbers (with respect to its own expectation), such as the results in Mentch and Hooker [2016a] and Wager and Athey [2018].

We can extend this result to the two forest case, where we compare the MSE of $RF_B(X)$ against that of $RF_B^{\pi}(X)$. In particular, if $\mathbb{E}MSE_{RF}(X;y) = \mathbb{E}MSE_{RF^{\pi}}(X;y)$, we see that

$$\sqrt{B}\left[MSE_{RF}(X;y) - MSE_{RF^{\pi}}(X;y)\right] \xrightarrow{d} \mathcal{N}\left(0, g'(\mathbb{E}RF_B(X))^2\sigma^2 + g'(\mathbb{E}RF_B^{\pi}(X))^2\sigma_{\pi}^2\right) \tag{3.3.3}$$

where $\sigma_{\pi}^2 = \mathrm{Var}(T^{\pi}(X))$. For a test set $\mathcal{T}$ with N_t points, we can calculate the pointwise squared errors as $MSE_{RF}(\mathcal{T}) = [(RF_B(X_i) - y_i)^2]_{i=1}^{N_t}$.

Finally, to connect back to our procedure, we now derive the asymptotic distribution of the differences in MSE between two forests. As above, let $MSE_{RF}(\mathcal{T})$ be the MSE of a random forest evaluated on a test set $\mathcal{T}$ and let $MSE_{RF^{\pi}}(\mathcal{T})$ denote the MSE of a forest

trained on the partially randomized data. By the results above, under the null hypothesis that $\mathbb{E}MSE_{RF}(\mathcal{T}) = \mathbb{E}MSE_{RF^\pi}(\mathcal{T})$, we have that as $B \to \infty$,

$$\sqrt{B}\mathbf{1}/\boldsymbol{N}_{\boldsymbol{t}}^T(MSE_{RF}(\mathcal{T}) - MSE_{RF^\pi}(\mathcal{T})) \xrightarrow{d} \mathcal{N}(0, \tau^2)$$

for some $\tau^2 > 0$. Appendix B.1 provides a technical derivation of the precise form of τ. We reiterate however that our permutation test approach, by design, avoids the need to compute this complex variance and so we do not discuss it further here.

Until now, our discussion has remained largely agnostic to the type of base-learners employed, subject to the regularity conditions needed for asymptotic normality. We turn now to establishing that the trees typically grown in a random forest satisfy such conditions. The following result follows a similar strategy as Lemma 2 in Meinshausen [2006] with regularity conditions similar to those imposed in Wager and Athey [2018].

Proposition 1. *Assume that $Y = m(X) + \epsilon$, where $m(\cdot)$ is continuous on the unit cube. Let $\mathcal{X} = [0,1]^p$, and assume that $X_{i,j} \overset{iid}{\sim} Unif(0,1)$ for $i = 1, ..., n$ and $j = 1, ..., p$. Then, let $T_n(X)$ be a tree trained on iid pairs $(X_1, Y_1), ..., (X_n, Y_n)$ such that each leaf of the tree contains a single observation. Further, assume the trees satisfy the following two conditions:*

(i) $\exists\ \gamma > 0$ *such that* $P($*variable* j *is split on*$) > \gamma$ *for* $j \in \{1, ..., p\}$

(ii) Each split leaves at least γn *observations in each node.*

Then, for each $X \in \mathcal{X}$

$$T_n(X) \xrightarrow{d} Y|X = X \text{ as } n \to \infty.$$

The tree predictions thus asymptotically behave like the conditional samples of Y and as a result, should have finite non zero variance. Note that Breiman [2001a] recommends building trees to full depth in which case Condition 1 is automatically satisfied.

3.3.4 Extension to Permutation Tests

The results in the previous subsection established that the sampling distribution of MSE differences between forests was asymptotically normal, but with a computationally intractable variance. Here we conclude our theoretical discourse by showing that the permutation distribution converges to that sampling distribution. As a result, the permutation test proposed in Section 3.2 is asymptotically equivalent to the standard parametric hypothesis test for variable importance but without the additional computational overhead. We begin by restating a classical theorem from Hoeffding.

Theorem 2. *[Hoeffding, 1952] For a sequence of data* $\{X_i\}_{i=1}^N$ *and a statistic* $S : \mathbb{R}^N \to \mathbb{R}$*, define the permutation distribution function as*

$$\hat{J}_N(t) = \frac{1}{|\mathcal{G}_N|} \sum_{\pi \in \mathcal{G}_N} I\big\{S(X_{\pi(1)}, ..., X_{\pi(N)}) \leq t\big\}$$

where $\mathcal{G}_N$ *is the group of all permutations of* $\{1, ..., N\}$*. Let* π, π' *be two permutations drawn independently and uniformly over* $\mathcal{G}_N$*, and suppose that as* $N \to \infty$

$$\big(S(X_{\pi(1)}, ..., X_{\pi(N)}), S(X_{\pi'(1)}, ..., X_{\pi'(N)})\big) \xrightarrow{d} (S, S') \tag{3.3.4}$$

where S, S' *are iid with cdf* $R(\cdot)$*. Then for all* t *at which* $R(\cdot)$ *is continuous,* $\hat{J}_N(t) \xrightarrow{p} R(t)$*.*

Direct application of Theorem 2 is often challenging. Suppose $\{X_i\}_{i=1}^n \overset{iid}{\sim} P_X$ and independently $\{Y_i\}_{i=1}^m \overset{iid}{\sim} P_Y$, and we calculate the statistic $\sqrt{n+m}\,[S(X_1, ..., X_n) - S(Y_1, ..., Y_m)]$, and further define $p = \lim_{n\to\infty} \frac{n}{n+m}$. Theorem 2.1 of Chung and Romano [2013] states that if there exists a function ψ_{P_Z} (which may depend on the distribution of the data, P_Z) such that

$$\sqrt{N}\,[S(Z_1, ..., Z_N) - \mathbb{E}S(Z_1, ..., Z_N)] = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi_{P_Z}(Z_i) + o_{P_Z}(1) \tag{3.3.5}$$

(i.e. the statistic is asymptotically linear), then the permutation distribution of the aforementioned statistic is asymptotically normal with mean 0 and variance given by

$$\tau^2 = \frac{1}{p(1-p)} \mathrm{Var}(\psi(Z)) = \frac{1}{p(1-p)} \left[p\mathrm{Var}(\psi(X)) + (1-p)\mathrm{Var}(\psi(Y))\right] \tag{3.3.6}$$

where $Z \sim pP_X + (1-p)P_Y$. A key challenge is that τ^2 is often not equal to the variance of the unconditional distribution without additional assumptions on P_X and P_Y.

The goal here is thus to provide a general result combining the delta method with the results of Chung and Romano [2013].

For a given test point (X, y), it can be shown that the MSE satisfies Equation 3.3.5 for

$$\psi(T(X)) = g'(\mathbb{E}RF_B(X))\left[T(X) - \mathbb{E}RF_B(X)\right]$$
$$\psi^\pi(T^\pi(X)) = g'(\mathbb{E}RF_B^\pi(X))\left[T^\pi(X) - \mathbb{E}RF_B^\pi(X)\right]$$

and thus the conditions needed to apply Theorem 2.1 of Chung and Romano [2013] are satisfied. The derivation of the permutation distribution variance and resulting formalization of the validity of our testing procedure for a single test point are given in Appendix B.1.

To extend this result to the more general case with a test set $\mathcal{T}$ consisting of N_t points, recall that the multipoint MSE can be broken down into a sum of iid components. In particular, results from in Subsection 3.3.3 give that

$$\sqrt{B}\left[MSE_{RF}(\mathcal{T}) - \mathbb{E}MSE_{RF}(\mathcal{T})\right] = \frac{1}{\sqrt{B}}\sum_{i=1}^{B}\bar{T}_i + o_P(1)$$

where $\bar{T}_i$ is an iid sequence of mean 0 random variables. Thus, the scaled and centered MSE satisfies the linearity condition in Equation 3.3.5. In particular, $\bar{T}_1, ..., \bar{T}_B \overset{iid}{\sim} P$ and $\bar{T}_1^\pi, ..., \bar{T}_B^\pi \overset{iid}{\sim} P^\pi$ and we are testing $H_0 : \mathbb{E}\bar{T}_i = \mathbb{E}\bar{T}_i^\pi$. Thus, because each is calculated with B trees, the same results hold and the test is asymptotically valid at multiple test points. This leads naturally to the following culminating theorem.

Theorem 3. *Let $T_{1,k_n}, ..., T_{B,k_n}$ and $T^\pi_{1,k_n}, ..., T^\pi_{B,k_n}$ be two collections of trees satisfying the conditions of Lemmas 1 and 3, and fix a collection of test points $\mathcal{T}$. Consider a test of the null hypothesis*

$$H_0 : \mathbb{E}\left[MSE_{RF}(\mathcal{T}) \;\middle|\; \mathcal{T}\right] = \mathbb{E}\left[MSE_{RF^\pi}(\mathcal{T}) \;\middle|\; \mathcal{T}\right]$$

using the statistic $\hat{\Delta} = MSE_{RF}(\mathcal{T}) - MSE_{RF^\pi}(\mathcal{T})$. Then under H_0, the permutation distribution of $\sqrt{B}\hat{\Delta}$ converges to a normal distribution with mean 0 and variance the same as that of the unconditional distribution of $\sqrt{B}\hat{\Delta}$, as $n \to \infty$. Thus, the permutation test attains the asymptotic Type I error rate.

3.3.4.1 Beyond the iid Approximation

As a final concluding remark related to the technical details of this permutation procedure, we note that the conditions of Lemma 1 are likely far stronger than needed to attain the ultimate result in Theorem 3. The proofs of validity for the permutation tests rely on projecting the random forest (which is a correlated sum $\frac{1}{B}\sum_{i=1}^{B} T_i(X)$) onto a sum of iid random variables, $\sum_{i=1}^{n} \psi_n(Z_i)$ for some function ψ_n, to which a central limit theorem can then apply. Indeed, this is exactly the approach of the Hájek projection and H-decomposition used respectively by Mentch and Hooker [2016a] and Wager and Athey [2018]. In these works, roughly speaking, it is shown that under constraints on the forest construction, the random forest prediction at a point X satisfies

$$\frac{1}{\sqrt{B}}\sum_{i=1}^{B}[T_i(X) - \mathbb{E}RF_B(X)] = \sum_{i=1}^{n}\psi_n(Z_i) + o_P(1).$$

For example, if the Hájek projection is used, $\psi_n(Z_i) = \sqrt{B}\mathbb{E}\left[RF_B(X) \mid Z_i\right] - \mathbb{E}RF_B(X)$. Moreover, as mentioned in the remark following Lemma 3, the fact that the MSE is asymptotically linear is independent of the iid approximation, and thus the MSE for these forests is also asymptotically linear.

3.4 Simulations

We now apply our testing procedure in a number of settings with varying regression functions and covariate structures. We simulate data from four models summarized in Table 3.4.1, with covariate structures summarized in Table 3.4.2. For each of our simulations, we train random forests using the `randomForest` package in R [Liaw and Wiener, 2002] using the default `mtry` parameters.

Model #	Data Generating Model	Covariate Structure
1	$Y = \beta X_1 + \beta I(X_6 = 2) + \epsilon$	M1
2	$Y = \beta \sin(\pi I(X_7 = 2) X_1) + 2\beta(X_3 - .05)^2 + \beta X_4 + \beta X_2 + \epsilon$	M1
3	$P(Y = 1\|X) = \text{expit}\left[\beta \sum_{j=2}^{5} X_j\right]$	M2
4	$Y = RF_{\texttt{eBird}}(X) + \epsilon$	`eBird`

Table 3.4.1: Distributions of $Y|X$ for each model. $\text{expit}(z) = \frac{1}{1+e^z}$.

Model #	Covariate Structure
M1	$X_1, ..., X_5 \overset{iid}{\sim} Unif(0,1)$, $X_6, ..., X_{10} \overset{iid}{\sim}$ Multinomial$(1, [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]^T)$
M2	$X_1, ..., X_{500} \sim \text{AR}_1(0.15)$
`eBird`	Data from Coleman et al. [2017] - 12 variables + 2 proxy variables

Table 3.4.2: Distribution of X for various simulation studies.

3.4.1 Power and Error Control

Model 1 is a standard ANCOVA model, which is intended to include both an important discrete and continuous predictor, to demonstrate the robustness of the proposed procedure to covariate type. Here we test the importance of (X_1, X_6, X_2, X_7) where X_1, X_6 are important, X_1, X_2 are continuous, and X_6, X_7 are categorical. Model 2 resembles the MARS data generating model [Friedman, 1991] commonly used in random forest studies, but with a modification to include an important discrete covariate. In both settings, we draw $n = 2000$ points from the joint distribution of (X, Y), subsample sizes of $k_n = n^{0.6} \approx 95$, and build $B = 125$ trees in each forest. Predictions were made at $N_t = 100$ test points, each drawn from the same joint distribution as the training data. Note that the null hypothesis, as defined in Equation 3.2.2, is conditional on the test points used. These simulations change the null hypothesis each time, because the validation set changes. Thus, the simulations mimic the common practice of random splitting the data into a training and validation fold.

For Models 1 and 2, we focus on a marginal signal to noise ratio, which is controlled by the parameters β and σ. We fix $\beta = 10$ across all simulations let $\sigma = 10/j$ where j takes 9

equally spaced values between 0.005 and 2.25 so that for small k, the signal to noise ratio (SNR) is small. The results are shown in Figure 3.4.1. We see that the test maintains the nominal type I error rate and attains high power for marginal SNRs near 1 for all variables except X_7 in Model 2. Note also that the type I error rate appears insensitive to the covariate structure. In the MARS model, we see that the test has more power against X_3 than X_7, because X_7 is only important insofar as it interacts with X_1.

Model 3 is an adaptation of the model used in Candes et al. [2016] for high-dimensional correlated data. Here we test for the significance of X_2, which is important, and also X_1 and X_{500}, which are unimportant, but X_1 is highly correlated with X_2 and X_{500} is much more weakly correlated. Candes et al. [2016] demonstrated that the standard logistic regression p-values in this situation are far from uniform under H_0, so that standard parametric inference may not be valid. Random forests, on the other hand, have been shown [Biau, 2012, Scornet et al., 2015a] to be largely insensitive to the dimension of the ambient feature space, and instead sensitive only to the "strong" feature space. This setting helps to explore the utility of our method in the high dimensional sparse signal case.

We limit $n = 600$ so that p/n is not small, though the dimension of the strong features is still small relative to n. We let $k_n = n^{0.6} \approx 46$, $B = 125$, $N_t = 100$, and vary the β coefficient according to 8 equally spaced values between 0.01 and 2.5 and also for 7 equally spaced values between 5 and 20. The results are shown in the bottom panel of Figure 3.4.1. Note that the test resolves the biased p-value issue associated with the standard glm procedure and is still able to attain reasonable power for the effect of X_2. The power is likely limited by the fact that for large β, the change in the marginal effect of each covariate only changes $P(Y = 1|X)$ slightly due to the rapidly decaying first derivative of the expit(z) function.

Finally, we turn to Model 4 where the true data generating model is a random forest. We utilize a dataset from Coleman et al. [2017] describing the occurrence of tree swallows and to construct $RF_{\texttt{eBird}}$, we draw 5000 points from the data, and train $RF_{\texttt{eBird}}$, a random forest with `mtry` $= 9$ and 1000 total trees. To simulate from this model, we draw (without replacement) samples of size n from the remaining 20727 points, predict at them using $RF_{\texttt{eBird}}$, and add Gaussian noise. We test for the effect of two variables: `eff.hours`, which corresponds to the number of hours a user expended upon a hike, and `dfs`, which is a fractional mea-

surement of day of year. We further include two proxy variables (not used to train $RF_{\texttt{eBird}}$), defined as $\texttt{eff.hours.proxy} = \frac{\texttt{eff.hours}+Z_{0.5}}{\sqrt{\text{Var}(\texttt{eff.hours})+0.5}}$ and $\texttt{dfs.proxy} = \frac{\texttt{dfs}+Z_{0.025}}{\sqrt{\text{Var}(\texttt{dfs})+0.025}}$ where Z_σ is a standard normal random variable with variance σ^2. The purpose of this construction is that the proxy variables' relationship with Y is solely dictated by their dependence on their original copy. In Model 4, we let $n = 2000$, $k_n = n^{0.6}$, $B = 125$, $N_t = 100$, and let $\sigma = e^{-j}$ for 10 values of j equally spaced between 1 and 5. The results of this simulation are show in Figure 3.4.1. We see that again the test maintains the nominal type I error rate with modest power for signal variables. Moreover, the procedure correctly identifies the true variables as important over their highly correlated proxies.

3.4.2 Normality of Permutation Distributions

One of the central claims of our work is that the permutation distribution is asymptotically Gaussian. To demonstrate this, we now provide a concise simulation demonstrating that the permutation approximation of the Gaussian proposed in Theorem 3 is valid in practice. We simulate $n = 2000$ training observations from Model 2 (with $\beta = 10, \sigma = 10$), along with $N_t = 100$ test observations and apply our procedure to test for the significance of X_3 (important) and X_5 (unimportant). The random forests each consist of $B = 200$ trees trained on subsamples of size $k_n = n^{0.6}$, with $\texttt{mtry} = 3$. The resulting permutation distributions are shown in Figure 3.4.2.

These plots demonstrate that the permutation distributions do approximate a Gaussian distribution. Moreover, in the null case, the observed Δ_B lies squarely in the center of the distribution, while in the alternative case, Δ_B lies far away from the center. Next, we more formally investigate the power/validity of the testing procedure.

3.4.3 Formal Comparison with Knock-offs

In this section, we formally compare the proposed procedure with several implementations of the knockoff framework [Barber et al., 2015, Candes et al., 2016], an exciting new method for statistically valid variable selection in a model-free way. As noted in Section 3.2, the null hypothesis tested by knockoffs is slightly different from that in our procedure. To

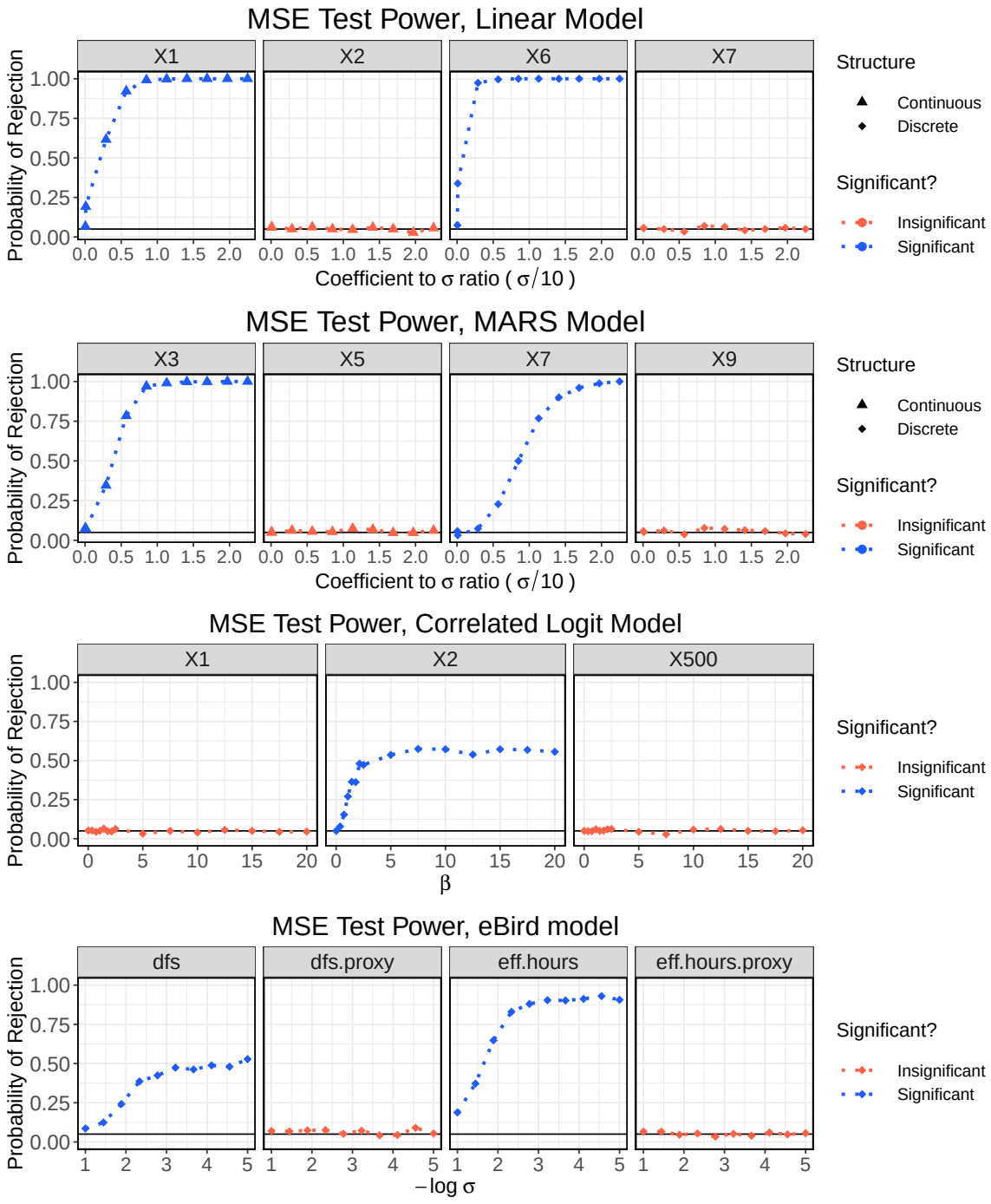
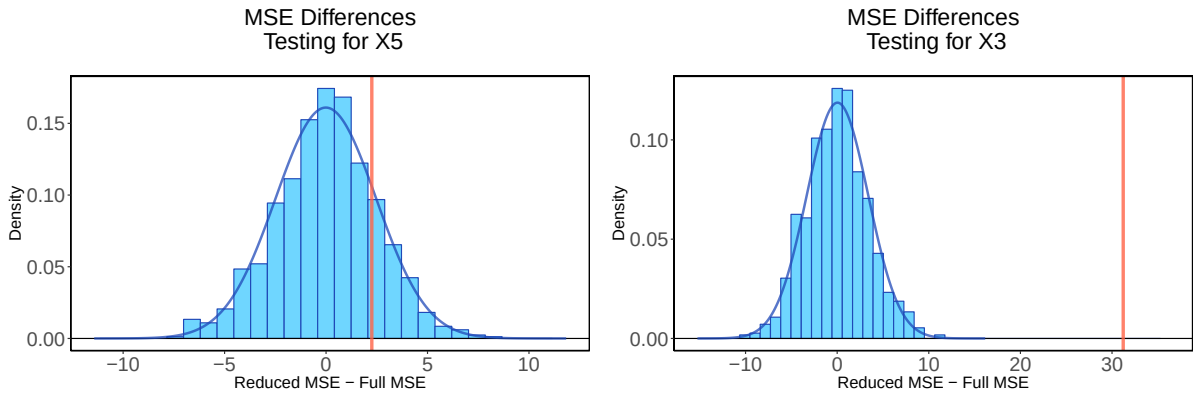


Figure 3.4.1: Simulation results for each of the models from Table 3.4.1. Black line corresponds to $\alpha = 0.05$, the nominal level



(a) Under H_0 **(b)** Under H_1

Figure 3.4.2: Permutation distributions of Δ_B in the simulation from subsection 3.4.2. Red line indicates observed value, and histograms are overlayed by an estimated normal density.

a practitioner, however, the procedures would likely be used in a similar way, and as such we leave this subtlety out of our subsequent discussion. In general, we conclude that our method is largely complementary to knockoffs.

A key assumption of the knockoff framework is that the distribution of the covariates X is known (also referred to as the model-X assumption) , which, crucially, our method does not require. Candes et al. [2016] proposes a second order method for generating knockoffs via a Gaussian analogue for X (i.e. a Gaussian random vector with the same covariance and mean as X). As of now, it is unclear how well knockoffs perform, both in terms of power and Type I Error control, when an approximation is employed. Finally, our method is designed to be powerful in situations where the response has a complex relationship with the data. To tackle these diverse scenarios, we use the following simulation set-up, with 4 different pairings:

- We fix $p = 25$, and generate covariates according to the following data distributions, one where the model-X assumption is satisfied, and one where it is not:

  Gaussian $X \sim \mathcal{N}(0, \Sigma)$ where $\Sigma_{ij} = \rho^{|i-j|}$ and we choose $\rho = 0.25$.

  Fish Toxicity We simulate X from the UCI fish toxicity data set provided by Cassotti et al. [2015], which comes with $n = 908$ observations on 6 covariates with information regarding chemicals that are believed to be toxic to a species of minnow. These co-

variates are quite non-Gaussian. To fill in the remaining 19 covariates, we randomly sample 19 columns (with replacement) from the original 6, and then sample rows of those 19 columns from the original data, so that there is no replication between the original 6 columns and the 19 synthetic columns. X is also scaled and centered, to account for differing units.

- Our responses are generated according to the following two regression functions. In both cases, $\epsilon \sim \mathcal{N}(0, 1/\text{SNR})$.

  Linear $Y = \sum_{j=1}^{s} X_j + \epsilon$

  Flattened Sine $Y = \frac{1}{\sqrt{\sum_{j=1}^{s} X_j^2}} \sin\left(\pi \sqrt{\sum_{j=1}^{s} X_j^2}\right) + \epsilon$. Note that in this set up, each variable has little linear effect but quite a strong nonlinear joint effect.

- For the responses, we vary the parameters s and SNR, to respectively control the density of the model (in terms of the number of important features) and the strength of the signal present in the data.

Each set up is evaluated at 6 different values of s, spaced evenly between 2 and 25, for a very sparse to a dense model, and 10 different signal to noise ratios, spaced evenly from 0.5 to 5, for a total of 60 simulation pairs. We apply the knockoff filter with the both standard lasso coefficient difference statistic and random forest out-of-bag importance statistic, to use both a linear and nonlinear statistic. We apply our procedure and the two knockoff approaches to 100 repetitions of $n = 908$ observations from the 4 different model set ups listed above. For our procedure, we build 125 trees, holdout 90 observations at random for testing, and take subsamples of size $k = \sqrt{908} \approx 30$. All tests here are conducted with respect to evaluating the marginal importance of X_1, which is important (in the sense of conditional independence) in each scenario. We define the power of the knockoff procedure to be $\frac{1}{100} \sum_{l=1}^{100} I(X_1 \in \mathcal{S}_l)$, where $\mathcal{S}_l$ is the selection of variables produced by the knockoff filter.

The results are plotted in Figure 3.4.3. Several patterns are shared across the plots. First, the proportion of important variables appears to be more important for attaining good power than the SNR in both our method and the knockoff procedure. However, the directionality is inverted - our procedure performs much better in sparse models, while knockoffs seem to require a dense model to select any variables. Next, the Gaussian flattened sine presents a challenge for both procedures, but our procedure is able to attain good power in all other

scenarios, while knockoffs really only succeed (with either statistic) in the linear Gaussian case. While throughout these simulations the FDR is controlled at the nominal level, a steep price is paid in terms of power for losing the knowledge of the distribution of X. Both knockoff statistics exhibit almost identical performance, which suggests further that the oob importance measures are unlikely to be useful as a nonlinear test statistics.

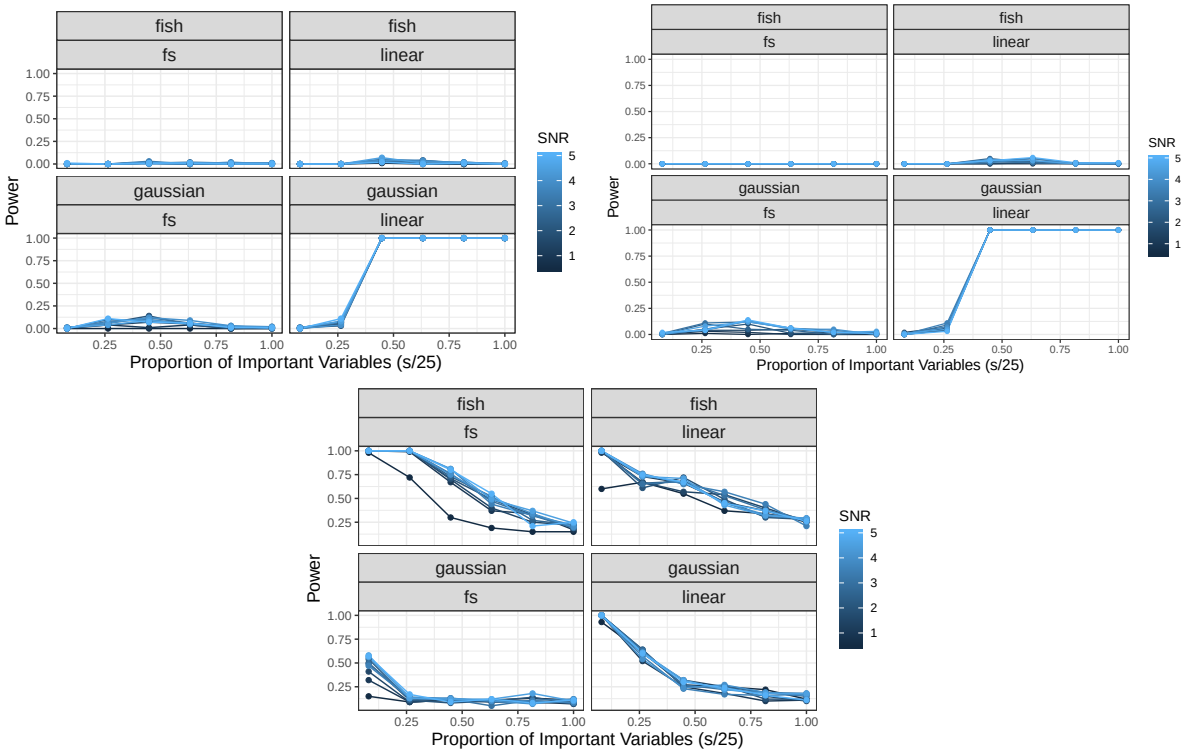


Figure 3.4.3: Simulation results for the knockoff comparison, showing the associated power curves, calculated with respect to a nominal Type I error rate of $\alpha = 0.10$. The knockoff procedure is run with the FDR threshold set to α. Light shades of blue indicate more powerful signal. `fs` refers to the flattened sine model. Bottom: our procedure. Top left: Knockoffs with the lasso statistic. Top right: Knockoffs with the random forest out-of-bag importance statistic.

We conclude that knockoffs are a powerful method when there are many covariates suspected to be important to the response. In these cases, the knockoff procedure can efficiently identify a dense model. However, the overwhelming dependence on s and the model-X assumption being satisfied suggests the need for more direct alternatives like that proposed here. Our procedure exhibits qualitatively similar behavior in 3/4 of the set ups,

attaining good power even for signal to noise ratios below 1 in the sparse model. Knockoffs maintain a computational edge over our method, needing only a single model fit to provide FDR controlled variable selection, while a naive implementation of our method would require $2p$ model fits, followed by a FDR filter such as Benjamini-Hochberg that accepts p-values [Benjamini and Hochberg, 1995].

3.5 Discussion

3.5.1 Validity of the Central Limit Theorem for Random Forests

The work above hinges on the conclusion of Equation 3.3.1 - that because subsampled tree predictions behave like an iid sequence in the limit, their sample means (i.e. random forest predictions) are approximately normally distributed. Recall that this conclusion was the result of Lemma 1 which shows that every pair of subsampled trees is asymptotically independent. We now present a pair of central limit theorems for exchangeable random variables that weaken this condition.

Theorem 4. *[Chow and Teicher, 2012] Let* $\{X_n\}$, $n \in \mathbb{N}$, *be an infinitely exchangeable sequence of random variables with* $\mathbb{E}(X_n) = 0$ *and* $Var(X_n) = 1$. *If* $Cov(X_i, X_j) = Cov(X_i^2, X_j^2) = 0$ *for all* $i \neq j$, *then* $\frac{1}{\sqrt{B}} \sum_{n=1}^{B} X_n \xrightarrow{d} Z$ *where* $Z \sim \mathcal{N}(0,1)$, *as* $B \to \infty$.

The implication of this theorem is that the trees need not be independent in order to obtain asymptotic normality. Using the variance calculation in Friedman et al. [2010], this requires a random forest with variance on the order of $\mathcal{O}(1/B)$ variance. Further, Chow and Teicher [2012] prove the converse - that is, the only way for random forests, as currently considered, to produce asymptotically normal predictions is for the correlations between trees to die out. This suggests that a bootstrapped forest will not be asymptotically normal as currently considered, as the correlations persist, even in the infinite sample case, and the tree predictions are exchangeable.

The above conclusion is quite disturbing - it may seem that there is no hope for asymptotically normality in the bootstrapping case, as bootstrapped trees are also exchangeable by

Theorem 1. However, we now state a second result, referenced in Klass and Teicher [1987]:

Theorem 5. *Let* $\{X_n\},\ n \in \mathbb{N}$ *be an infinitely exchangeable sequence with* $\mathbb{E}(X_n) = 0$ *and* $Cor(X_i, X_j) \equiv \rho > 0$. *Then,* $\frac{1}{B\rho}\sum_{n=1}^{B} X_n \xrightarrow{d} Z \sim \mathcal{N}(0,1)$ *if and only if* $\mathbb{E}(\prod_{i=1}^{n} X_i)$ *exists and is equal to* $\mathbb{E}(Z^n)$ *for all* n.

The main qualitative difference is the rescaling factor - for the correlated case is $\frac{1}{B}$, not $\frac{1}{\sqrt{B}}$. This is corroborated by the variance calculation from Friedman et al. [2010] - the variance of a bootstrapped forest, in our context, no longer vanishes because the variance of individual trees does not vanish. Thus, a more punitive rescaling is necessary to attain asymptotic stability. The condition on the moment matching can be interpreted as a more restrictive version of the canonical condition that each random variable has variance 1. This condition, however, is necessary for a CLT to hold, and as such could be an emphasis of future research into the limiting distribution of bootstrapped forests.

Under the assumption that the regularity conditions of Theorem 5 are met, in finite sample situations, we are always technically in the regime of Theorem 5, so that a central limit theorem holds, but with a difficult to estimate variance that may not line up with the variance estimate suggested by the permutation distribution. However, we argue that the limiting distribution suggested by each theorem is similar, so long as the number of trees built B is small relative to the correlation between the trees. In particular, assume that the tree predictions at X have variance σ^2, and pairwise correlation ρ. Then, recalling the result from Friedman [2001], the variance of the random forest prediction at X is given by

$$\text{Var}\left[\frac{1}{B}\sum_{i=1}^{B} T_i(X)\right] = \sigma^2\frac{B}{B^2} + 2\frac{\frac{B^2-B}{2}}{B^2}\rho\sigma^2 = \sigma^2\left(\frac{1+(B-1)\rho}{B}\right)$$

For $\rho = 0$ (the variance implied by Theorem 4), we see that the random forest variance simply $\frac{\sigma^2}{B}$. This is the variance used in the permutation test theory above - i.e. the variance that arises by treating the trees as iid. Thus, the difference between the true variance and the theoretical variance is $\sigma^2\left(\frac{(B-1)\rho}{B}\right)$. Clearly, as B gets large, this term approaches $\rho\sigma^2$, which is the maximum difference between the variances. However, for small B, i.e. $1/\rho \gg B$, we see that the difference term is very near 0, so the iid approximation is reasonable. Recall that both ρ and B depend on n, so that a more reasonable requirement (for an iid-like CLT

to hold) on the number of trees may be $\rho_n = o(1/B_n)$, rather than the condition laid out in Lemma 2.

To demonstrate this effect, we conduct an auxillary simulation, similar to the simulations from Mentch and Hooker [2016a]. We simulate 1000 datasets of varying sizes from the MARS model (Equation 3.5.1) and make predictions at $X_1 = X_2 = ... = X_5 = 0.5$.

$$Y = 10\sin(\pi X_1 X_2) + 20(X_3 - 0.05)^2 + 10X_4 + 5X_3 + \epsilon \tag{3.5.1}$$

where $X_1, ..., X_5 \overset{iid}{\sim} Unif(0,1)$ and $\epsilon \sim \mathcal{N}(0, 10)$. We simulate the unconditional distribution of three different random forest predictions. We first consider the full bootstrapped forest, then a forest with large subsamples, and a forest with small subsamples. Theorem 5 suggests that if the tree predictions can be rescaled to have moments of a standard normal random variable, then the limiting distribution of random forest predictions should be asymptotically normal.

We also estimate the relevant terms in the variances associated with both central limit theorems. We estimate the correlation between the tree predictions (and the squared tree predictions) by training two trees on iid resamples from each simulated dataset. Then, the correlation between the vector of tree predictions is recorded. We also record the average MSE of each forest, the variance of the forest predictions, the variance of the tree predictions, and the variance of the MSEs. The squared error terms are recorded with respect to a fixed test outcome (i.e. are conditional on a particular realization of $Y|X = X$).

These results are presented in Table 3.5.1. As expected, the smaller the subsample size is relative to the overall subsample size, the lower the correlation between the trees. Somewhat surprisingly, the bootstrapped forest exhibits higher tree variance (≈ 90.2) than the subsampled forests ($\approx 82.1, 84.0$), leading to a much more variable overall forest when combined with the higher tree correlations. Interestingly, the variance of the random forest predictions are much smaller than the average MSE, indicating that the squared bias term is dominating the error. This may be due to the conditional nature of the squared error terms - essentially, the bias is a function of both the underlying bias of the random forest and the realized error.

n	k_n	$\mathbb{E}(MSE_{RF})$	$\text{Var}(RF(X))$	$\text{Var}(MSE_{RF})$	$\text{Var}(T_{k_n}(X))$	$\rho_{k_n}(X)$	$\rho_{2,k_n}(X)$
2250	2250	178.94386	5.08631	3690.87766	90.23730	0.04146	0.02635
2250	481	111.81116	2.06818	932.02915	82.07984	0.02329	0.02683
10000	1585	27.80852	1.46533	157.52904	84.03259	0.00808	0.01743

Table 3.5.1: Simulation results for the figures plotted in Figure 3.5.1. We let $\rho_{k_n}(X) = \text{Cor}(T_{k_n}(X), T'_{k_n}(X))$ and similarly $\rho_{2,k_n}(X) = \text{Cor}(T^2_{k_n}(X), T^{2'}_{k_n}(X))$. All expectations are with respect to the distribution of the training data and conditional on the test point.

The resulting distributions are plotted in Figure 3.5.1. Notably, the random forest prediction densities are well approximated by a Gaussian density in each case, including the bootstrapped forest. To our knowledge, this is the first evidence of asymptotic normality of the bootstrapped forest presented. The MSE distributions remain somewhat skewed, but are generally well approximated by normal densities in each case.

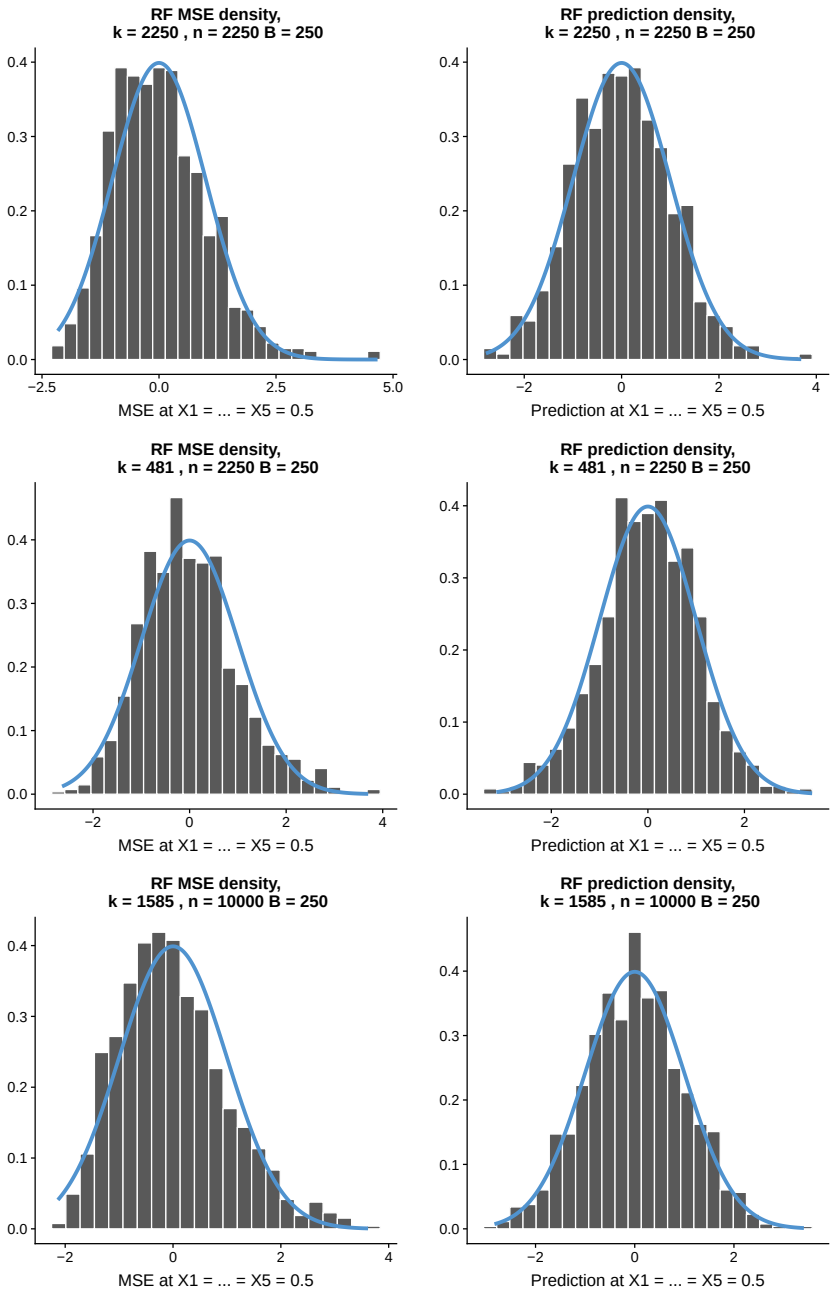


Figure 3.5.1: Various distributions of random forest predictions from data from the MARS model. Subsample size (k), sample size (n), and number of trees (B) shown in subtitle. Results are studentized (i.e. have mean 0 and variance 1) – blue overlay shows standard normal distribution.

3.5.2 Variable Importance

Random forest practitioners are often interested in the marginal importance of all variables. Out-of-bag and Gini impurity measures are typically reported for all variables. In contrast, the permutation test proposed here is only a test for marginal importance. However, due to the computational efficiency of our method, we can run a marginal test for each covariate, with respect to the original forest. This requires building $p+1$ forests - one for a permutation of each covariate, and one original forest. Note that simply running the test for each covariate would result in $2p$ forests. Then, each covariate returns a p-value $\tilde{p}_1, ..., \tilde{p}_p$, and our importance score could be simply to order the p-values in ascending order. However, in practice, because of the finite number of permutations, strongly significant variables often obtain $p_{min} = 1/(N_{perm}+1)$. To alleviate ties in p-values, we instead recommend a metric that does not rely on the p-values directly:

$$\text{Imp}_j = \frac{\Delta_j(\mathcal{D}_n) - \mathbb{E}_\pi(\Delta_j(\mathcal{D}_n))}{\sqrt{\text{Var}_\pi \Delta_j(\mathcal{D}_n)}}$$

where $\Delta_j(\mathcal{D}_n) = MSE_0^\pi - MSE_0$ is the observed MSE differences, $\mathbb{E}_\pi(\Delta_j(\mathcal{D}_n))$ is the mean of all MSE differences possible when permuting the $2B$ trees on $\mathcal{D}_n$, and $\text{Var}_\pi \Delta_j(\mathcal{D}_n)$ is similarly defined. Intuitively, the variable importance is just the (unitless) z-score of the observed MSE differences with respect to the permutation distribution. Because the permutation distribution of Δ_j is asymptotically Gaussian (as shown in later sections), this metric is closely linked with the p-value. and in fact, (letting $\Phi(\cdot)$ be the standard normal distribution function) $\text{Imp}_j \approx \Phi^{-1}(\tilde{p}_j)\sqrt{\text{Var}_\pi \Delta_j(\mathcal{D}_n)} + \mathbb{E}_\pi \Delta_j(\mathcal{D}_n)$, assuming that $\tilde{p}_j$ is calculated with respect to enough permutations.

Simply applying Algorithm 1 to each variable in the model still may be unrealistic computationally in high-dimensional situations - each test requires building $2B$ trees, and with p tests, in all $2pB$ trees are needed. For large p, this may be substantially larger than just $2B$. As such, we propose a computational speedup which only requires building one (potentially larger) forest. The goal is to obtain a collection of decision trees $\boldsymbol{T}_1, ..., \boldsymbol{T}_B$ in which there are some trees that have not been trained with X_j, and some trees that have been trained using X_j.

Recall that random forests are built using decision trees that consider only a random subset of possible splits at each node during tree construction. Typically, these subsets are all possible splits among $m < p$ features, where the m features are chosen uniformly at random. This means that inevitably (unless $m \ll p$ or the trees are very shallow), most trees will be have at least considered all features in their construction. However, consider restricting the features available for the entire construction of the tree. That is, each tree is trained on a resample of the rows of the design matrix and on a subsample of the columns. Ho [1998] analyzed subsampling features in a decision tree ensemble, and the idea has since been expanded upon in methodological updates such as rotation forests, introduces a step of performing on the subsampled features [Rodriguez et al., 2006]. These methods attempt improving predictive accuracy, but simultaneously introduce a convenient mechanism for evaluating variable importance. In particular, for each variable X_j and each decision tree $\boldsymbol{T}_k$, we can introduce the variables $\nu_{j,k}$, where:

$$\nu_{j,k} = \begin{cases} 1 & \text{If variable } X_j \text{ was used in training tree } \boldsymbol{T}_k \\ 0 & \text{otherwise} \end{cases}$$

Note that each $\nu_{j,k}$ is, marginally, a Bernoulli random variable with mean m/p and $\nu_{j,k} \perp\!\!\!\perp \nu_{j,k'}$ for $k \neq k'$, so that in a collection of B trees, on average, $\sum_{k=1}^{B} \mathbb{E}(\nu_{j,k}) = Bm/p$ will contain variable X_j and $B(p-m)/p$ will exclude X_j. We then see two natural partitions of the collection of trees into forests:

$$RF_j = \{\boldsymbol{T}_k : \nu_{j,k} = 1\}$$
$$RF_{-j} = \{\boldsymbol{T}_k : \nu_{j,k} = 0\}$$

We can then apply the second loop of Algorithm 1 to the trees of RF_j, RF_{-j} for each $j \in \{1, ..., p\}$ and obtain importance scores and p-values without retraining entire forests each time. This process is summarized in Algorithm 2.

Results in the previous sections demonstrate that the testing procedure provides marginal coverage, but invariably a multiple testing procedure must be employed for high-dimensional problems. Type I error control across all covariates results in overly conservative methods, and instead we suggest two methods for instead controlling the False Discovery Rate (FDR)

Algorithm 2: Holdout Importance method for calculating all importance scores at once

Data: Training data $(\mathcal{D}_n)$ test sample $(\mathcal{T} = [(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_{N_t}, y_{N_t})])$

Result: P-values and importance scores for $X_j; j \in \{1, ..., p\}$ at test points $\mathcal{T}_n$

SET n_{perm} number of permutations, subsample size k_n, and $\min_{tree}$ in each forest ;

SET $B_j = 0$ for $j = 1, ..., p$ and $\ell = 0$;

while ANY$(B_j < \min_{tree})$ **do**

  SAMPLE $[\nu_{\ell,1}, ..., \nu_{\ell,p}] \in \{0, 1\}^p$ where $\sum_j \nu_{\ell,j} = m$ and $P(\boldsymbol{\nu}) \equiv \frac{1}{\binom{p}{m}}$;

  SET $\mathcal{M}_\ell = \{j : \nu_{\ell,j} = 1\}$;

  SAMPLE $\mathcal{D}^*_{\ell,k_n}$ from the rows of $\mathcal{D}_n \ominus \boldsymbol{X}_{-\mathcal{M}_\ell}$;

  TRAIN $\boldsymbol{T}_\ell(\cdot)$ on $\mathcal{D}^*_{\ell,k_n}$;

  UPDATE $\ell \leftarrow \ell + 1$ & $B_j \leftarrow B_j + \nu_{\ell,j}$, $j \in \{1, ..., p\}$

end

PREDICT at $\mathcal{T}_n$ using $\boldsymbol{T}_\ell(\cdot)$, generating $\boldsymbol{T}_\ell = [T_i(\boldsymbol{x}_1), ..., T_i(\boldsymbol{x}_{N_t})]$;

for $j \in \{1, ..., p\}$ **do**

  SET $RF_j = \{\boldsymbol{T}_k : \nu_{j,k} = 1\}$ $RF_{-j} = \{\boldsymbol{T}_k : \nu_{j,k} = 0\}$;

  if $B_j \neq \ell - B_j$ **then**

    SAMPLE $m_j = \min\{B_j, \ell - B_j\}$ trees uniformly from RF_j, RF_{-j}

  end

  CALCULATE $MSE_j = \left|\left|\frac{1}{m_j} \sum_{\boldsymbol{T} \in RF_j} \boldsymbol{T} - \boldsymbol{y}\right|\right|^2$ and $MSE_{-j} = \left|\left|\frac{1}{m_j} \sum_{\boldsymbol{T} \in RF_{-j}} \boldsymbol{T} - \boldsymbol{y}\right|\right|^2$;

  for k *in* $\{1, ..., n_{perm}\}$ **do**

    SAMPLE $RF^*_{k,j}$ uniformly w/o replacement from $RF_j \cup RF_{-j}$;

    SET $RF^*_{k,-j} = \{RF_j \cup RF_{-j}\} \setminus RF^*_{k,j}$;

    CALCULATE $\Delta_{j,k}(\mathcal{D}_n) = \left|\left|\frac{1}{m_j} \sum_{\boldsymbol{T} \in RF^*_{k,-j}} \boldsymbol{T} - \boldsymbol{y}\right|\right|^2 - \left|\left|\frac{1}{m_j} \sum_{\boldsymbol{T} \in RF^*_{k,j}} \boldsymbol{T} - \boldsymbol{y}\right|\right|^2$

  end

  CALCULATE $\tilde{p}_j = \frac{1}{N_0+1}\left[1 + \sum_{k=1}^{N_0} I\big(MSE_{-j} - MSE_j > \Delta_{j,k}(\mathcal{D}_n)\big)\right]$;

  CALCULATE $\text{Imp}_j = \frac{\Delta_j(\mathcal{D}_n) - \mathbb{E}_\pi(\Delta_j(\mathcal{D}_n))}{\sqrt{\text{Var}_\pi \Delta_j(\mathcal{D}_n)}}$

end

(see Benjamini and Hochberg [1995], Hochberg and Benjamini [1990] for more details on FDR):

1. Applying the Benjamini-Hochberg procedure to the original p-values
2. Using the knock-off procedure of Candes et al. [2016] with the Imp_j importance measures.

We leave a study of which method is preferrable to future work. As demonstrated in the simulations for Model 4 in Section 3.4, the p-values generated by Algorithm 1 are valid even in the high-dimensional case, so that the standard p-value based procedures like that in Benjamini and Hochberg [1995] could be readily applied.

3.5.3 Null Hypothesis Considerations

Besides its feasibility, this permutation approach also offers some flexibility in the kinds of problems open to investigation by practitioners. Consider, for example, the mediator detection problem arising frequently in medical studies wherein a covariate X_1 is a *mediator* for another covariate X_2 whenever the effect of X_2 on the response is nullified (or substantially lessened) by including X_1 in the model. The same two-step process often employed with linear models can be carried out with random forests using the tests developed here: first determine whether X_2 is significant without X_1 in the model, then test whether the significance of X_2 disappears whenever X_1 is included. Moreover, our procedure attains good power in a wide variety of model set ups, and as such is likely usable off-the-shelf by practitioners interested in the nonlinear regression inference problem.

The primary goal of this work is to identify covariates that produce statistically significant improvements in model accuracy. To assess this, we considered building two forests, one on the original dataset $\mathcal{D}_n$ and another on a second dataset $\mathcal{D}_n^\pi$ wherein the covariate(s) of interest X_S are rendered independent of Y, conditional on the rest of the features. This muting of X_S can be achieved in various ways:

- Outright exclusion: X_S is simply removed from the second training dataset.
- Random permutation: Each covariate in X_S is randomly shuffled so that X_S is replaced by some permuted alternative X_S^π in the second training dataset.

- Knockoffs: Each covariate in X_i in X_S is replaced by some knockoff alternative X_i^π sampled from the distribution of $X_i|X_{-i}$ so that X_S is replaced by a randomized alternative X_S^π in the second training dataset. See Candes et al. [2016] for details.

The manner in which covariates are randomized or muted will subtly alter the underlying null hypotheses in Equation 3.2.2, but crucially, the Type I error control of our procedure holds for each of these null hypotheses. Indeed, it is possible to reject the null because of artifacts in the covariate distribution, rather than a notion of conditional independence. Assuming $X_\mathcal{S} \perp\!\!\!\perp Y|X_{-\mathcal{S}}$, we would expect predictions from trees trained on $\mathcal{D}_n$ to have the same distribution as those generated from trees trained on $\mathcal{D}_n^\pi$. In this case, a rejection of the null hypothesis of equal MSE's suggests that $X_\mathcal{S}$ and Y are not conditionally independent. This is the case if the distribution of X is known (or can be estimated easily), so that a knockoff version of $X_\mathcal{S}$ can be employed. However, practically speaking, our method provides valid model based inference even without any knowledge of the covariate distribution. In such cases, formally investigating whether particular covariates significantly improve predictive accuracy beyond permuted analogues, for example, can still provide valuable insight into their relative value and utility.

3.6 Application to Ecological Data

We now apply the above methodology to two ecological datasets where random forests have been shown to perform well in recent work.

1. The eBird data used to train $RF_{\texttt{eBird}}$, also analyzed in Coleman et al. [2017]. The task here is to predict tree swallow `occurrence` during the fall in Bird Conservation Region (BCR) 30. Features include information about latitude, longitude, time of year, user characteristics (as described in Section 3.4), and environmental characteristics, such as temperature anomaly and land cover features. The data consists of $n = 25727$ observations on $p = 23$ features, gathered between 2008 and 2013.
2. Forest fire data from Cortez and Morais [2007], where the task is to predict $\log(1 + \text{area})$

burned by several fires in northern Portugal using information about location, time of year, and local weather characteristics. The data contains $n = 537$ observations on $p = 13$ features.

For both applications, we first apply the hold-out procedure (Algorithm 2) to identify marginally important variables, and then a localized MSE test to the variables found most important by the hold-out procedure. These procedures are done by splitting the data into 85%/15% training/test splits at random, with the same splits (i.e. training and test data) used for both procedures. The random forests were trained with the `randomForest` package using the default `mtry` parameters, and $k_n = n^{0.6}$, $B = 250$. A partial effect plot of the rain covariate in the forest fire data is shown in Figure 3.6.1, demonstrating the non-linear relationship between the outcome and rain - suggesting that a linear model is likely inappropriate. Similar results are suggested for the eBird data in Coleman et al. [2017].

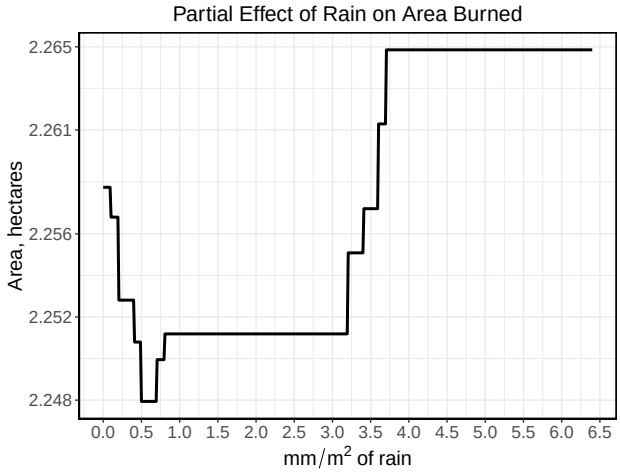
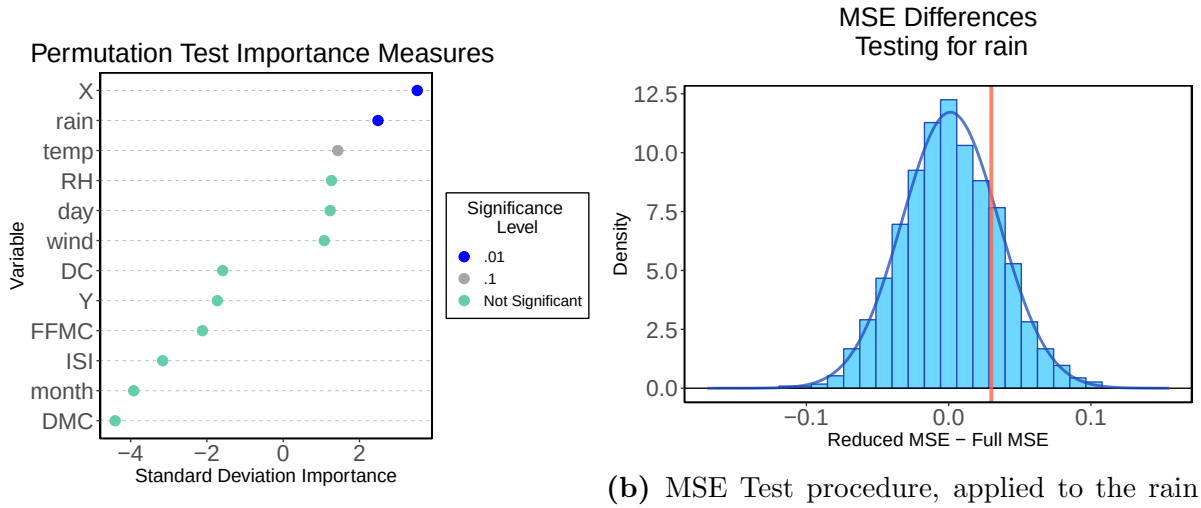


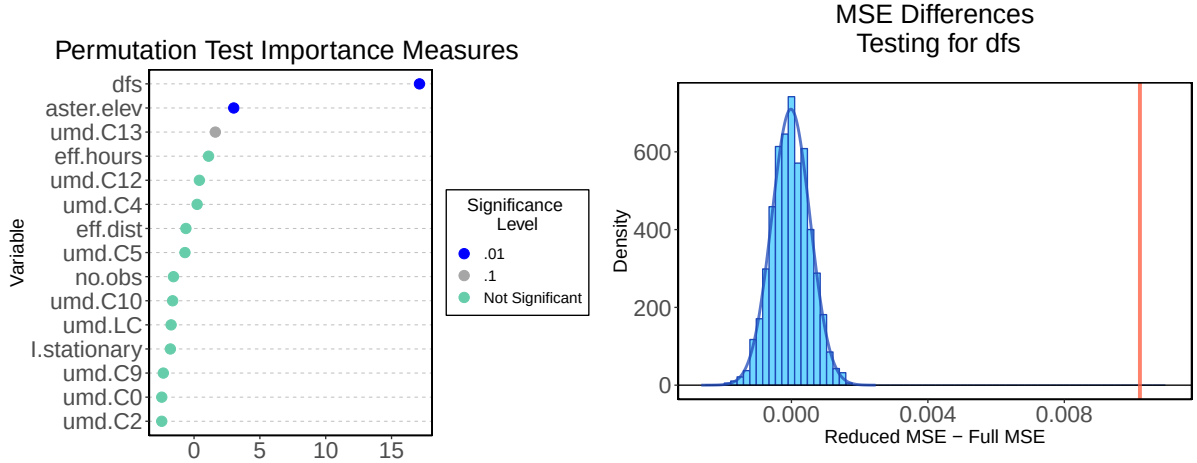
Figure 3.6.1: Partial effect plot of a random forest to predict the area burned by forest fires



(a) Holdout procedure

(b) MSE Test procedure, applied to the rain covariate

Figure 3.6.2: Results on the forest fire data from Cortez and Morais [2007]. Red line indicates observed value, and histograms are overlayed by an estimated normal density.



(a) Holdout procedure

(b) MSE Test procedure, applied to the dfs (day of year) covariate

Figure 3.6.3: Results for the procedure applied to the eBird data [Sullivan et al., 2009a, 2014a].

We see in Figure 3.6.2 that the X (which corresponds to longitude) ($\text{Imp}_j = 3.515$), rain ($\text{Imp}_j = 2.486$), and temp ($\text{Imp}_j = 1.431$) variables are most important, and that the rain variable is somewhat less important ($\tilde{p} = 0.206$) when tested for outside of the hold-out procedure. The story is more clear in Figure 3.6.3, where the day of year covariate (dfs) is hugely important ($\text{Imp}_j = 17.12$), and shows up highly important in its marginal test

($\tilde{p} = 0.001$.)

3.7 Additional Applications

A main motivation for the permutation test was practical applications, and to this effect we now detail some additional collaborations where the methodology has been applied. In particular, we present a collaboration with clinicians about evaluating the predictive importance of *mobile health* (mHealth) variables, which are gathered using fitness trackers. The main conclusions of this work were reached using the permutation test methodology proposed earlier. The purpose of this section is to demonstrate an interesting application of the methodology in the spirit of the proposed method. We also note that the permutation test procedure has been implemented as an `R` package, `RFTest`, available at `https://github.com/tim-coleman/RFtest`.

The study focused on patients with Inflammatory Bowel Diseases (IBDs) who also own fitness tracking devices, such as those produced by FitBit and Garmin. Participants report longitudinal survey data on outcomes such as disease activity scores along with self-reported measures such as sleep disturbance while also contributing mHealth lifestyle data from wearable devices and apps, covering 24 different device types from multiple wearable brands. IBD patients have extremely heterogeneous phenotypes with symptoms that fluctuate. IBDs patients broadly can be discretized into two categories, those diagnosed with Crohn's Disease (CD) and Ulcerative/Indeterminate Colitis (UCIC). These diseases have their own

Several studies suggest a relationship between self reported mHealth variables and disease symptoms [Ananthakrishnan et al., 2013, Jones et al., 2015]. Our mHealth dataset contains numerous features describing physical activity and sleep in addition to a number of other lifestyle characteristics, allowing for a broad, large-scale analysis of the features most associated with IBD disease activity and symptoms. In all, we studied eight outcomes, including two disease scores and six quality of life outcomes. We study the relationship of these eight outcomes with mHealth features that encompass thirteen different categories, such as sleep and steps taken. In all, our data consisted of 539 observations on 127 features,

represetning summarized data from 539 inter-survey periods in the longitudinal data.

The complex underlying relationships in the data suggest that machine learning (ML) approaches are preferred over parametric models with more rigid structure. Using the permutation test methodology developed in Section 3.2, we assessed the predictive relevance of mHealth data in forming more accurate predictive models than could be obtained with survey data alone, and we further analyzed which specific mHealth variables are most predictive of outcomes for patients with IBDs. We compared models which were trained using only baseline data, which includes patient demographic information as well as prior survey results, against those trained using the baseline data and mHealth data. For each test, a random 15% of the data were held out as a validation set.

Because the permutation test methodology is agnostic to base learner, under the assumption that the base learners are pairwise independent, we elected to apply the method to bagged conditional inference trees (cForest) [Hothorn et al., 2006] and bagged elastic net models, which are penalized linear models [Zou and Hastie, 2005], which were shown in another analysis to be the most accurate models, as selected by cross-validation. We conducted several hypothesis tests, testing for both the overall effect of any mHealth variables as well as the effect of individual groups of variables. The results of the test for the overall effect of mHealth variables are presented in Table 3.7.1. The results for the tests of the effect of the groupings of the mHealth variables are presented in Figure 3.7.1 and Figure 3.7.2.

	anxiety	depression	fatigue	sleep	social	scai	scdai	pain
C-Forest Test	0.977	0.802	0.624	0.012	0.253	0.214	0.096	0.032
Elastic Net Test	0.899	0.008	0.426	0.666	0.069	0.089	0.288	0.030

Table 3.7.1: Permutation test p-values from applying the permutation test procedure. The top and bottom rows show the results of the test conducted with conditional inference trees and elastic net models, respectively.

The overall tests for significance indicate that, that in aggregate, mHealth data was predictive of pain interference in both models, with more modest evidence for an effect on SCDAI disease activity, social relationship, and depression scores. The tests for predictive significance of the groups of mHealth features provide modest evidence active duration (time

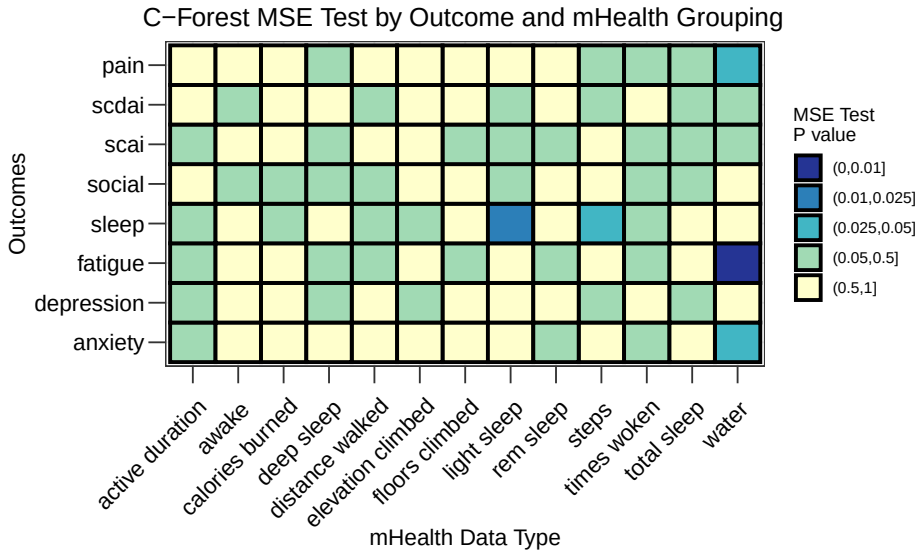


Figure 3.7.1: Permutation test with conditional inference trees.

spent in moderate-to-vigorous activity (active duration)) was predictive of pain interference and disease activity for patients with either CD and UC (Figure 1) in the elastic net models (SCDAI and SCAI). The tests on the elastic net model also also suggests that distance traveled throughout the day was predictive of disease activity for UC, sleep disturbance, fatigue, and depression scores. While total sleep was only predictive only of for UC disease activity for patients with UC, it was also predictive for depression and pain overall. Total number of Steps per day were was only strongly predictive for CD disease activity in patients with CD. Tests conducted on the cForest model detected fewer significant results, with water consumption being the only mHealth feature consistenly shown to improve predictions across outcomes. We note that the p-values presented here have not been adjusted for multiple testing purposes, though the results are amenable to many multiple-testing adjustments, which is beyond the scope of this work.

Beyond the clinical implications of these results, this analysis raises several avenues for improvement of the original procedure. In particular, there is a fair amount of inconsistency between the p-values reported in Figure 3.7.1 and Figure 3.7.2. In particular, The inconsistency could be due to the difference in base learner, but more likely is that the discrepancy is due to randomness in the sample splitting. Indeed, for small sample sizes, differing sam-

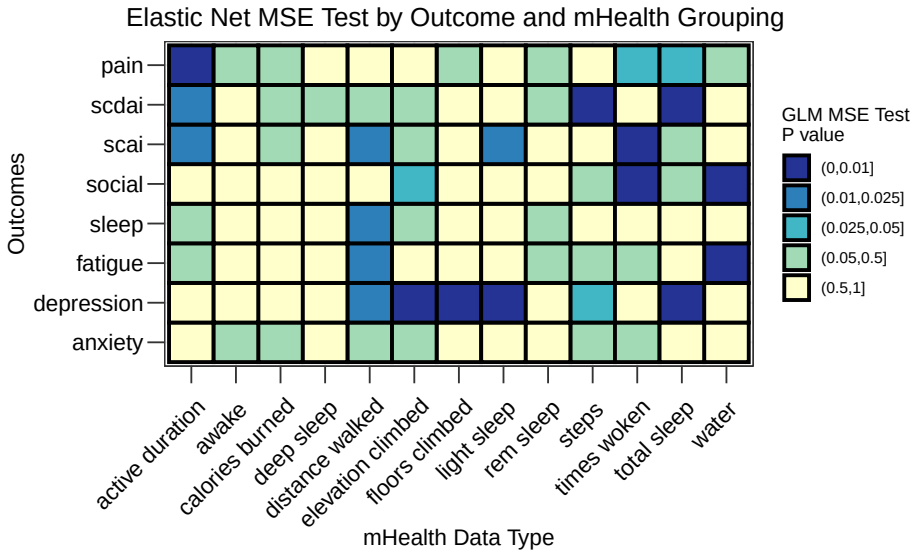


Figure 3.7.2: Permutation test with elastic net base models.

ples can produce differing results, because it may be that there is considerable variability in the test-conditional MSE, $\mathbb{E}[MSE_{RF}(\mathcal{T})|\mathcal{T}]$, which in turn affects the null hypotheses tested by the permutation procedure. Lei [2019] discusses this phenomenon in the context of cross-validation, but a similar conclusion likely applies here. For larger test sets, the test-conditional MSE is close to the expected MSE, so that the null hypothesis presented in Theorem 3 is more similar a test about the unconditional MSE, $\mathbb{E}[MSE_{RF}(\mathcal{T})]$.

Sample splitting has been employed to develop statistically valid procedures, such as for developing a hypothesis test for LASSO coefficients [Wasserman and Roeder, 2009], or in the conformal inference framework, where sample splitting is key to many of the guarantees of that procedure [Shafer and Vovk, 2008, Lei et al., 2018]. A clear philosophical drawback of sample splitting procedures is that different auxillary randomizations can lead to different conclusions, independent of the randomness in the data. As such, there has been work on aggregating p-values from multiple splits of data, as in Meinshausen et al. [2009]. Layering an aggregation scheme on top of the permutation test procedure, and analyzing the theoretical qualities of such a scheme, remain an interesting and promising avenue for methodological development.

3.8 Conclusion

We have proposed a new avenue of theoretical analysis of bagged models, by noting that bagged models can be seen as sums of exchangeable random variables. The deep connection between exchangeability and permutation tests further motivates usage of a permutation test that permutes the base learners between ensemble methods. This test uses the flexibility of random forest models to conduct inference, and as such attain good power, but also maintain Type I error control. This procedure uses permutation distributions to avoid the computational cost of variance estimation of a random forest.

4.0 Forecasting the Damages of the Hundred Year Storm: Importance Forest

4.1 Introduction

In machine learning, it is often assumed, implicitly or explicitly, that data used in training and data held out for prediction follow the same distribution. As such, models find an approximating function $\hat{f}$ that minimizes the *global generalization error*, which for a loss function $L(\hat{f}(X,Y)$ is defined as $\mathbb{E}_{(X,Y)}L(\hat{f}(X),Y)$, where the expectation is taken with respect to the distribution of both X and Y. However, it may be that

$$\mathbb{E}_{(X,Y)\sim P_{\text{train}}}L(\hat{f}(X),Y) \neq \mathbb{E}_{(X,Y)\sim P_{\text{test}}}L(\hat{f}(X),Y)$$

because $P_{\text{train}} \neq P_{\text{test}}$. As such, minimizing the left hand side may not yield an estimator that minimizes the second quantity. This idea of utilizing knowledge of where predictions will be sought as part of the training process is a natural fit in areas such as personalized medicine [Liu and Meng, 2016], for example, where physicians may often seek the most accurate predicted outcomes for particular patients, rather than a global minimizer. Powers et al. [2015] make use of this notion of *customized training* to cluster pixels from mass spectrometric images taken from lung cancer patients in order to fit more precise individual models to each cluster.

To formalize the above framework, consider covariates X which take values in some p dimensional space $\mathcal{X} \subset \mathbb{R}^p$ and a response Y which takes values in $\mathcal{Y} \subset \mathbb{R}$. Suppose we have two sets of data $\mathcal{D}$ and $\mathcal{D}'$ where $\mathcal{D} = (X_i, Y_i)_{i=1}^n \overset{iid}{\sim} P_1$ and $\mathcal{D}' = (X_i', Y_i')_{i=1}^m \overset{iid}{\sim} P_2$. where P_i is a probability measure on $\mathcal{X} \times \mathcal{Y}$ for $i = 1, 2$. Furthermore, assume that the Y_i' are unavailable. The goal is to attain accurate point estimates and prediction intervals for Y_i'. Now, suppose P_1, P_2 satisfy

$$\begin{aligned} P_1(X,Y) &= P(Y|X)P_1^*(X) \\ P_2(X,Y) &= P(Y|X)P_2^*(X) \end{aligned} \tag{4.1.1}$$

so that the conditional distribution of the target is the same for both datasets - the change is in the covariate distribution. This model is commonly referred to as the *covariate shift* model, and has been the source of intense research in recent decades [Shimodaira, 2000, Sugiyama and Müller, 2005, Sugiyama et al., 2007, Reddi et al., 2015]. The issue arises when P_2^* and P_1^* concentrate mass in different areas of $\mathcal{X}$. In this case, standard guarantees about the effectiveness of many regression estimates of $P(Y|X = X)$ are invalid for X in areas of low mass of P_1^*, even as $n \to \infty$. This is especially problematic if the low mass areas of P_1^* have high mass in P_2^*. To resolve this, we propose learning a mapping between P_1^* and P_2^* by estimating the likelihood ratio function $\ell(X) = \frac{dP_2^*(X)}{dP_1^*(X)}$. Note we have assumed that P_1^* and P_2^* are absolutely continuous with respect to each other, i.e. for all measurable A, $P_1^*(A) > 0 \iff P_2^*(A) > 0$. In essence, we want to calculate the likelihood ratio, $\Lambda = \frac{dP_2^*}{dP_1^*}$, without necessarily specifying the form of P_1^* and P_2^*. This precludes the use of typical parametric likelihood functions. Moreover, the high dimension of many problems means that the naive approach of estimating two densities will be quite unstable.

4.1.1 A Motivating Example: Hurricane Power Outages

One of the most damaging effects of hurricanes is the loss of power for many people in the storm track. Forecasting these outage counts is a direct way of quantifying the damage done by a hurricane, whereas meteorological forecasts, such as of windspeed and storm surge, tend to focus less on the human impact of the storm. Advances in machine learning have led to large improvements in predictive modeling of power outages that result from tropical storms and hurricanes. These models typically take in two sets of covariate information: (1) Information about the storm, such as windspeed expected in each study unit (2) information about each study unit, such as the soil types and demographics of the unit.

The focus of this paper is to develop a method for accurately forecasting outages during storms across a wide variety of geographic extents, using only inputs available on such a geographic scale. Effectively, this means we cannot use information about the power-grid itself due to limited coverage, resolution, and types of information reported about each local grid. Several challenges are inherent to this problem.

Data Availability The National Hurricane Center [Landsea and Franklin, 2013] only provides full data for storms from 1995 onwards.

Rarity of Severe Events Severe storms are, by definition, anomalous, and therefore are potentially underrepresented in the available data. Moreover, they may be overrepresented for particular areas of interest due to chance.

Interest in Severe Events Forecasting less severe outages is inherently less useful to practitioners - often, the interest is in whether or not the forecasts for the big storms are accurate.

Outage data is provided by the EAGLE-I system, which aggregates national information about the power-grid. Power outages are clearly dynamic throughout the storm - in our dataset, outages are reported every 15 minutes for each county affected for the duration of the storm. For simplicity, we summarize the outage extent in the following way: (1) We record a running minimum outage $M_{i,t} = \min\{O_{i,k} : \ k \in [t, t+8)\}$, where $O_{i,k}$ is the time series of power outages in county i; (2) We let $Y_i = \log_{10}(\max_t M_{i,t})$. This quantity serves as our response variable, and is referenced with the predictors listed in the supplemental material and in Pasqualini et al. [2017]. Taking the logarithm of the outages helps to alleviate the heavytailed nature of the response, and further its interpretation can help quantify the magnitude of the expected effect [Tokdar and Kass, 2010, Willoughby et al., 2007].

In all, the data contains outage counts from 17 hurricanes and tropical storms between 2011 and 2017, for a total of 5015 observations, on 75 predictors. Given a county in a storm with covariates X, we want to estimate the conditional distribution $Y|X = X$ of county level outages, with emphasis on point estimates and prediction intervals. Moreover, we are typically interested in making forecasts for the entire affected region of a hurricane at once.

4.2 Related Work

To fit a random forest into this framework, one solution would be to implement a weighted bootstrap in the resampling phase of the forest. Canonically, each observation has probability of being selected $p_i \equiv 1/n$, under the weighted scheme, $p_i \propto w_i$, where w_i are some

weights obtained *a priori*. This approach was considered by Xu et al. [2016], who proposed a weighting scheme where the weights measure the importance of each training point relative to a *single test point*. We note that this approach is well-suited for making predictions at a single point, i.e. where P_2^* is a degenerate distribution with all of its mass concentrated at x_0. However, the weights used change from test point to test point, meaning that a new weighting scheme must be used for each point, and thus a new random forest must be trained for each test point, leading to a total of $|\mathcal{D}'| + 1$ forests needed. This may incur needless computational cost. A speed-up could be to cluster the test points and then apply the above scheme to the centroids of the clusters to get a weighting scheme for all points within the cluster. This is quite similar to the approach suggested by Powers et al. [2015]. In contrast to these procedures, we want to use distributional information about the covariates in our weighting. Moreover, for practical purposes, we seek a method with minimal additional computational overhead.

4.2.1 Related Hurricane Outage Work

Liu et al. [2005] used negative binomial regression to forecast outages during three storms during the 1990's. Guikema and Quiring [2012] found that generalized linear models lacked sufficient flexibility to accurately forecast power outages, and instead turned to non-parametric models, such as random forests and gradient boosting. More recently, Wanik et al. [2015] used a combined random forest, gradient boosting, and a single decision tree to forecast outages. He et al. [2017] used quantile regression forests [Meinshausen, 2006] to provide prediction intervals, in addition to point estimates, for power outage forecasts. Quantile based methods may be preferable due to the heavy-tailed nature of power outage distributions - the averaging used in conditional mean estimation can lead to severe over/under estimates of power outages. Moreover, practitioners are likely more interested in a prediction interval than a confidence interval, as a prediction interval can inform evacuations/preparations. As such, much of the recent work in random forest inference, such as Wager et al. [2014a], Mentch and Hooker [2016a], Wager and Athey [2018], Coleman et al. [2019b], Peng et al. [2019] is of less interest because of their focus on conditional mean

estimation.

In our case, assume we train a model only on data from P_1, which may be data from several hurricanes in years prior, and then use it to make predictions about data that come from P_2, such as the outages for a yet-observed hurricane, whose characteristics may be quite different than storms previously recorded. Table 4.2.1 shows the result of this procedure for 6 hurricanes between 2012 and 2017. In particular, each model is tuned by minimizing the out-of-bag error for each parameter configuration, and the optimal model is then used to make predictions for the held-out storm. For example, to forecast Hurricane Arthur, we use data from the 16 other storms to train a random forest, which is then used to learn $f(x) = \mathbb{E}(Y|X = x)$, and $Q_\alpha(x) = F_Y^{-1}(\alpha|X = x)$ for $\alpha = 0.1, 0.5, 0.9$. Thus, the forest predicts the conditional mean, the conditional median, and a conditional 80 % prediction interval. If the covariate structure was similar for each storm, we would anticipate seeing roughly similar error metrics across storms, especially seeing as the sample size is similar across each iteration. Rather, we see that three storms (Harvey, Nate, Matthew) have similar error metrics, while Arthur, Irma, and Sandy are much higher. It is not surprising that these are the storms that are most difficult to forecast - Irma and Sandy in particular were historically damaging storms [Cangialosi et al., 2018]. Perhaps more telling is that the prediction intervals for the higher error storms provide much poorer coverage. Meinshausen [2006] showed that, under regularity assumptions, the conditional quantiles estimated by a quantile regression forest are consistent - as such, we would expect prediction intervals to maintain near the nominal coverage level. However, Harvey shows minor departures from this coverage level and Irma, Sandy, and Arthur shows a extreme departure from this level. To summarize performance, we also report a "score" metric, which is defined as

$$\text{Score} = \left(\frac{1}{\text{MAE}} + \frac{1}{\text{RMSE}} + \frac{4}{\text{IntWidth}} \right) \frac{\text{Covg}}{1 - \alpha} \tag{4.2.1}$$

so that the score is penalized for higher loss (MAE, RMSE), for wider intervals, and for lower coverage %. This is not a formal loss function, but an attempt to quantify overall predictive performance. We note that Irma and Sandy have the lowest scores by far - again suggesting the difficulty in forecasting the damage from these storms.

Storm	`mtry`	`nodesize`	MAE	RMSE	Covg	Interval Width	Score
Matthew-2016	50	5	0.6269	0.7861	0.8898	2.6946	4.3021
Nate-2017	40	5	0.6727	0.8124	0.8759	2.5094	4.1960
Harvey-2017	50	5	0.7509	0.9026	0.7632	2.4214	3.4695
Arthur-2014	45	5	0.8498	1.0322	0.6862	2.2623	2.9839
Sandy-2012	40	10	0.9817	1.2197	0.5781	2.2376	2.3293
Irma-2017	45	5	1.1846	1.4044	0.3706	2.4051	1.3258

Table 4.2.1: Tuned random forest results for 6 storms in the hurricane dataset. "Covg" and "Interval Width" refer to 80% prediction intervals estimtaed via quantile regression forests.

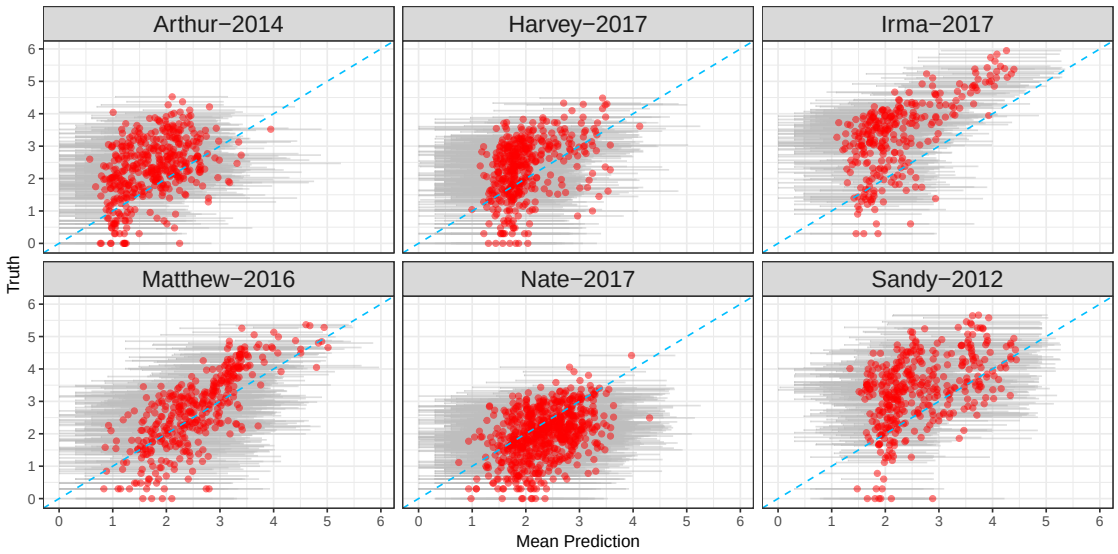


Figure 4.2.1: Fitted vs Predicted for each storm-holdout model. Blue line represents perfect prediction, and grey bars represent 80% prediction intervals.

4.3 Methods

We begin with a brief summary of importance sampling. Importance sampling refers to weighting observations to either reduce the variance of some point estimate or to "tilt" a sample observed from P_1 to be similar to P_2. Canonically, if the goal is to estimate $\mu := \mathbb{E}(f(X)) < \infty$ where $X \sim P$ for some function f, one would draw a sample $X_1, ..., X_n \overset{iid}{\sim} P$ and use $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$. Importance sampling instead weights each observation by how

"useful" it is to measuring μ. The idea is that observations in regions where $|f(x)| \approx 0$ are ultimately not useful to calculating μ. A canonical example is, for $X \sim N(0,1)$, calculating $\mathbb{E}(I(X > 30))$ - because $P(X > 30) \propto e^{-30^2/2}$, sampling uniformly at random will require a tremendous number (practically infinite) of samples to visit the region of interest. However, if we had samples from $P_2 = N(30,1)$, we would sample the region of interest quite often.

As such, importance sampling seeks to weight each X by how much it resembles a sample from P_2. The idea is to replace the observations X with $X^* = Xw(X)$, for $w(x) = \frac{P_2(x)}{P_1(x)}$. We then let $\tilde{\mu} = \frac{1}{n}\sum_{i=1}^n X_i^*$. If, P_1 and P_2 are known, the $w(X_i)$ are already normalized (in the sense that they sum to 1). In our case, we know neither distribution, and can only calculate an un-normalized likelihood ratio between the two. As such, the self-normalized importance sampling estimate $\tilde{\mu} = \frac{\sum_{i=1}^n w_i f(X_i)}{\sum_{j=1}^n w_j}$ is of more use. The problem is to construct a random forest using data from P_1 as if the data had come from P_2. We propose a two stage procedure to solve this problem:

1. First, we train a model to learn $\ell(x) = \frac{dP_2^*(x)}{dP_1^*(x)}$, the ratio of the data densities at x. We then estimate $\ell(x_1), ..., \ell(x_n)$ for each point in $\mathcal{D}$.
2. We construct a randomized tree using an importance weighted criterion for both the splits and the predictions.

Tree-based models are constructed by recursively partitioning the feature space. Partitioning takes a rectangular subspace A and partitions it into two further rectangular subspaces A_L, A_R, where $A_L = \{X \in A : X_i^{(j)} < z\}$, $A_R = A \setminus A_L$, and $x^{(j)}$ represents the j^{th} coordinate of an observation. In the context of a continuous feature space (i.e. no categorical predictors), the quality of a split is assessed by

$$L(j,z) = \frac{1}{N_n(A)}\sum_{i=1}^n (Y_i - \bar{Y}_A)^2 I(X_i \in A) - \frac{1}{N_n(A)}\sum_{i=1}^n \left(Y_i - \bar{Y}_{A_L} I(X_i^{(j)} < z) - \bar{Y}_{A_R} I(X_i^{(j)} \geq z)\right)^2 I(X_i \in A) \quad (4.3.1)$$

where $N_n(A)$ indicates the number of observations in the original sample that lie in region A and $\bar{Y}_A$ is the sample mean of the response over all observations who lie in region A. This

criterion is typically evaluated at all possible split points, and the split selected satisfies $(A_L, A_R) = \text{argmax}_{(j,z)} L(j,z)$. This process is initialized with $A = \mathcal{X}$, and then repeated recursively until the trees reach a specified depth or terminal node size. The trees output a rectangular partition, $A_1, ..., A_m$ where m is the number of terminal nodes in the tree, and where $\mathcal{X} = \cup_{i=1}^m A_i$. Let $A^*(x)$ be the partition segment containing x, so that the prediction at x is given by

$$T(x; \mathcal{D}) = \sum_{i=1}^{n} \frac{I(x_i \in A^*(x))}{N_n(A^*(x))} Y_i.$$

The construction of the trees above can be seen as repeated calculation of different statistical functionals. For a given probability measure P supported on a set A, consider a rectangular partition of A into A_L and A_R, such that $A_L = \{x \in A : x^{(j)} < z\}$ and $A_R = A \setminus A_L$. Define $P_L = \frac{1}{P(A_L)} PI(x \in A_L)$, normalizing so that P_L is a valid probability measure. We can then define the functionals

$$\begin{aligned} T_1(P) &= \int y dP(y) \\ T_{j,z}(P) &= \int (y - T_1(P))^2 dP(y) - \\ &\quad \int \left[(y - T_1(P_L))^2 I(x \in A_L) + (y - T_1(P_R))^2 I(x \in A_R) \right] dP(y). \end{aligned}$$

In the above, the functionals are calculated only with respect to the response coordinate - i.e. they are scalars, not vectors. For a given node A, define $\hat{P}_A = \frac{1}{N_n(A)} \sum_{i=1}^n \delta_{(x_i, Y_i)} I(x_i \in A)$, where $\delta_{(x_i, Y_i)}$ places mass 1 at the pair (x_i, Y_i). We can redefine Equation 4.3.1 in terms of functionals of empirical distributions as

$$L(j,z) = T_{j,z}(\hat{P}_A).$$

The prediction stage can similarly be seen as $T(x; \mathcal{D}) = T_1(\hat{P}_{A^*(x)})$. The main innovation we propose here is to replace $\hat{P}_A$, which may estimate the training data distribution well, with another estimate $\tilde{P}_A$ that well approximates the distribution of the test data. Then, the functionals described above are calculated over $\tilde{P}_A$ for both the structure and prediction

stages of the tree construction. In practice, we use the following formulation of $\tilde{P}_A$, which depends on a weight vector $\boldsymbol{w}$

$$\tilde{P}_{A,\boldsymbol{w}} = \sum_{i=1}^{n} \frac{w_i I(x_i \in A)}{\sum_{j=1}^{n} w_j I(x_j \in A)} \delta_{(x_i, Y_i)}. \tag{4.3.2}$$

We thus replace the factor $1/N_n(A)$ with a value proportional to w_i. We use $\boldsymbol{w} = \{\ell(x_1), ..., \ell(x_n)\}$, so that $T(\tilde{P}_{\boldsymbol{w}})$ is an approximation to $T(P_2)$ rather than $T(P_1)$. Tree construction proceeds by recursively maximizing $T_{j,z}(\tilde{P}_{A,\boldsymbol{w}})$ over each node, until the control parameters of the tree are met. As in the unweighted case, we can restrict the set of possible splits randomly at each node, such as only allowing $\texttt{mtry} < p$ features available for splitting, which can provide a forest variance reduction by decorrelating the trees. Then, the weighted tree predictions are given as

$$T_{\boldsymbol{w}}(x; \mathcal{D}) = T_1(\tilde{P}_{A^*(x),\boldsymbol{w}}).$$

Finally, a forest is created by resampling the data many times and training a randomized tree on each data. The forest prediction, like in standard random forests (which estimate the conditional mean fuction) is given by

$$m_{B,\boldsymbol{w}}(x; \mathcal{D}) = \frac{1}{B} \sum_{k=1}^{B} T_{\boldsymbol{w}}(x; \mathcal{D}, \xi_k)$$

where ξ_k are iid randomization parameters determining the resamples and available features for splitting at each node. These procedures are summarised in the two subsequent algorithms.

Algorithm 1 Weighted Regression Tree

procedure WEIGHTEDTREE($\mathcal{D}, \boldsymbol{w}, \xi, m_n$) ▷ $\boldsymbol{w}$ are weights, ξ is randomization, m_n is maximum number of terminal nodes
  Set $\mathcal{P}_0 = \{\mathcal{X}\}$ ▷ The root node is the entire feature space
  Set $t = 1$ ▷ Counter for number of terminal nodes
  For all $1 \leq k \leq \texttt{nrow}(\mathcal{D})$ set $\mathcal{P}_k = \varnothing$
  Set $d = 0$ ▷ Tree depth counter
  while $t < m_n$ **do**
    if $\mathcal{P}_d = \varnothing$ **then**
      $d \leftarrow d + 1$
    else
      Set A as the first element in $\mathcal{P}_d$ ▷ P_A is the within-node distribution
      Let $\mathcal{M}_{\xi,d} \subset \{1, .., p\}$ be features available for splitting
      Evaluate $T_{j,z}(P_A)$ $\forall z$ and for all $j \in \mathcal{M}_{\xi,d}$
      Set $A_L^* = \{\boldsymbol{X} \in A : X^{(j^*)} < z^*\}$ where $z^*, j^* = \text{argmax}_{j,z}(P_A)$ and set $A_R^* = A \setminus A_L^*$
      Set $\mathcal{P}_d \leftarrow \mathcal{P}_d \setminus \{A\}$
      Set $\mathcal{P}_{d+1} \leftarrow \mathcal{P}_{d+1} \cup \{A_L^*\} \cup \{A_R^*\}$
      Set $t \leftarrow t + 1$
  Prediction at point $\boldsymbol{x}$ is made by $T_1(P_{A(\boldsymbol{x})})$ where $A(\boldsymbol{x}) \in \mathcal{P}_d$ is the node containing $\boldsymbol{x}$

Algorithm 1 Weighted Random Forest

procedure WEIGHTEDRF($\mathcal{D}_{\text{TRAIN}}, \mathcal{D}_{\text{TEST}}, \texttt{REPLACE}, k_n, B$)
  For all $\boldsymbol{X} \in \{\mathcal{D}_{\text{TRAIN}}, \mathcal{D}_{\text{TEST}}\}$, set $Z = I(\boldsymbol{X} \in \mathcal{D}_{\text{TEST}})$
  Run a random forest, RF_ℓ, which estimates $\pi = P(Z = 1|\boldsymbol{X})$
  Evaluate $RF_\ell(\boldsymbol{X}_i) = \hat{\pi}_i$ for all $\boldsymbol{X}_i \in \mathcal{D}_{\text{TRAIN}}$
  Set $\hat{\ell}_i = \frac{\hat{\pi}_i}{1-\hat{\pi}_i} + \frac{1}{\texttt{nrow}(\mathcal{D}_{\text{TRAIN}})}$ ▷ Second term is to prevent 0 weights
  Let $\boldsymbol{\ell} = \{\hat{\ell}_1, ..., \hat{\ell}_n\}$
  for $k \in \{1, ..., B\}$ **do** ▷ B is total number of trees to be trained
    if `REPLACE` **then**
      Draw k_n observations w/ replacement, with $P(\boldsymbol{X}_i \text{ Selected}) \propto \hat{\ell}_i$
    else
      Draw k_n observations w/o replacement, with $P(\boldsymbol{X}_i \text{ Selected}) \propto \hat{\ell}_i$
    Let $\mathcal{D}_{k,k_n}$ be the resampled data, and $\boldsymbol{\ell}_{k,k_n}$ be the resampled weights
    Set $T_k \leftarrow$ WEIGHTEDTREE($\mathcal{D}_{k,k_n}, \boldsymbol{\ell}_{k,k_n}, \xi_k$) ▷ ξ_k controls other randomization
  return $\{T_1, ..., T_B\}$ ▷ Collection of trees

4.3.1 Weighted Quantile Regression

Recall that a major interest in the forecasting problem are prediction intervals, and quantile regression forests [Meinshausen, 2006] provide a natural means of non-parametric quantile regression. As such, we propose a means of using the importance forest procedure for quantile regression. As Meinshausen [2006] notes, a random forest estimate can be reformulated as a weighted mean of the observations, as opposed to the sample mean of the

trees. For a prediction point x and a point in the training set X_i, a decision tree (constructed using prior weights $\boldsymbol{w}$) drawn with parameter ξ induces the following weights:

$$t_i(x;\xi,\boldsymbol{w}) = I(X_i \in A^*_\xi(x))\frac{w_i}{\sum_{j=1}^n w_j I(X_j \in A^*_\xi(x))}$$

Then, given B trees trained using randomization parameters $\xi_1, ..., \xi_B$, we can define the random forest weights by:

$$r_{i,B}(x;\boldsymbol{w}) = \frac{1}{B}\sum_{k=1}^{B} t_i(x;\xi_k,\boldsymbol{w})$$

Following Meinshausen [2006], we can then use these weights to get an estimate of $F(y|X = x) = P(Y \leq y|X = x)$ as

$$\tilde{F}_{\boldsymbol{w}}(y|X = x) = \sum_{i=1}^{n} r_{i,B}(x;\boldsymbol{w})I(Y_i \leq y).$$

We can similarly define a quantile function

$$\tilde{Q}_{p,\boldsymbol{w}}(x) = \inf\{y : \ \tilde{F}_{\boldsymbol{w}}(y|X = x) \geq p\}.$$

Note that $\tilde{F}_{\boldsymbol{w}}(y|X = x)$ only takes on $n + 1$ values, so evaluation of $\tilde{Q}_{p,\boldsymbol{w}}(x)$ amounts to a grid search over these $n + 1$ values. For a provided quantile, p, we see that $\tilde{Q}_{p,\boldsymbol{w}}(x) = Y_{k^*}$, where $k^* = \min_k \sum_{i=1}^k r_{(i),B}(x;\boldsymbol{w}) \geq p$, where the notation $r_{(i),B}(x;\boldsymbol{w})$ indicates that the RF weights are now ordered corresponding to response value, i.e. $i > k \iff Y_i \geq Y_k$.

4.3.2 Learning ℓ

Each element of the weight vector $\ell(X_i)$ is a ratio of densities of two different covariate distributions. These densities are unknown and are over high dimensional feature space. As such, many traditional density estimation tools are unlikely to be effective. We briefly describe a method from Kanamori et al. [2009] below, with a more thorough discussion of its advantages reserved for the appendix. We describe two candidate procedures for density estimation, probabilistic classification and kernel moment matching. We argue that the probabilistic classification approach, while simple to implement, may be unstable in high dimensions.

4.3.2.1 Probabilistic Classification

We can use the favorable properties of tree based density estimates in high dimensions to learn ℓ. The algorithm of Breiman [2001a] can be used for unsupervised learning, by returning measures of adaptive distance between observations. Crucially, this procedure relies on the creation of a synthetic covariate dataset, and then learning the probability that a particular observation came from the true or synthetic dataset. The synthetic dataset is created by drawing n observations (with replacement) uniformly and independently from each covariate, destroying any dependencies between the observations. The idea is that if there is high-dimensional structure, the model should easily discriminate between the two datasets. In the covariate shift literature, this procedure is referred to as a *probabilistic classification* method, as it transform the density ratio estimation problem into a classification problem [Barber et al., 2019].

To formalize the above, we impose another assumption about the distribution of test and training. For all $X_i \in \{\mathcal{D}, \mathcal{D}'\}$, we assume that $X_i \overset{iid}{\sim} P(X_i) = \alpha P_1^*(X_i) + (1-\alpha)P_2^*(X_i)$, where $\alpha \in (0,1)$. In the canonical machine learning context, $\alpha \equiv 1$ (without loss of generality), which covers the situation where the test and training covariates have the same distribution. We introduce the synthetic response $Z = I(X \sim P_2^*)$. For every observation in $X_i \in \{\mathcal{D}, \mathcal{D}'\}$, this amounts to $Z_i = I(X_i \in \mathcal{D}')$, where $I(\cdot)$ is an indicator function. We then want to learn $P(Z=1|X)$, i.e. the probability that an observation came from one dataset or another. Note that this relies on the density discrepancy between P_1^* and P_2^*, which may be a nonlinear function of complex interactions between each feature. Then, it follows that

$$P(Z_i = 1|X_i) = \frac{P(X_i|Z_i=1)P(Z_i=1)}{P(X_i)} = \frac{dP_2^*(X_i)P(Z_i=1)}{P(X_i)}$$

and thus

$$\frac{P(Z_i=1|X_i)}{P(Z_i=0|X_i)} = \frac{\frac{dP_2^*(X_i)P(Z_i=1)}{P(X_i)}}{\frac{dP_1^*(X_i)P(Z_i=0)}{P(X_i)}} = \ell(X_i)\frac{P(Z_i=1)}{P(Z_i=0)}.$$

We only require our importance sampling weights to be proportional to $\ell(X_i)$, so that any information placed in the marginal distribution of Z_i is accounted for in the normalization. Using the random forest estimates $\hat{\pi}_i$ of $\pi_i = P(Z_i=1|X_i)$, we let $w(X_i) := w_i = \frac{\hat{\pi}_i}{1-\hat{\pi}_i}$ be our estimate of the appropriate weighting scheme. To ameliorate dividing by 0, in practice, a

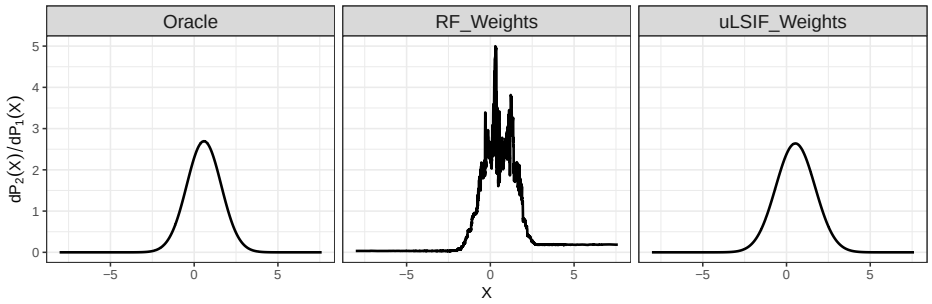


Figure 4.3.1: Comparison of estimated density ratios between an inverted random forest classifier and the uLSIF method of Kanamori et al. [2009]. In this example, $P_1^*(X) = \mathcal{N}(0, 2.5^2)$ and $P_2^*(X) = \mathcal{N}(0.5, 0.95^2)$, and models were learned with $n = 1500$ examples from each. In the above example, the RF attained RMSE of 0.355 while the uLSIF method attained an RMSE of 0.139.

small constant δ is often added to both the numerator and denominator. We now provide an approximate error estimate of the classifier-inverted ratio weight. We can write $\hat{\pi}_i = \pi_i + \epsilon_i$ for some error term ϵ_i which we assume has finite variance σ_ϵ^2. Then, the ratio weights are given by

$$w_i = \frac{\hat{\pi}_i}{1 - \hat{\pi}_i} = \frac{\pi_i + \epsilon_i}{1 - \pi_i - \epsilon_i} := g_i(\epsilon_i)$$

where g_i is a differentiable function with derivative $g_i'(x) = (1 - \pi_i - x)^{-2}$. Then, assuming that ϵ_i satisfies both a central limit theorem and a law of large numbers (asymptotic in N) we see that

$$\text{Var}\left(\sqrt{N} w_i\right) \approx g'\left(\mathbb{E}\epsilon_i\right)^2 \sigma_\epsilon^2 = \frac{\sigma_\epsilon^2}{(1 - \pi_i - \mathbb{E}\epsilon_i)^4}$$

so that if the asymptotic bias ($\mathbb{E}\epsilon_i$) is small or 0, the variance of the weight estimates scales as $O\left((1 - \pi_i)^{-4}\right)$. This can lead to severe instability in the probabilistic classifier estimate, if the underlying conditional probabilities are close to 1. The effect of this instability is shown in Figure 4.3.1, where even in a simple univariate case, the probabilistic classifier picks up on the general trend of the density ratios, but has high variance. As such, an alternative method of estimating the likelihood ratio weights is needed.

4.3.2.2 Least Squares Importance Fitting

Another method for estimating density ratios that has been explored is Least Squares Importance Fitting, developed by Kanamori et al. [2009]. The approach essentially reduces down to modelling the ratio as a linear output

$$\ell(x) = \sum_{k=1}^{b} \alpha_k K_\sigma(x, X_k)$$

where $\alpha_k \geq 0$ for all k, X_k are centroid points, σ is a bandwidth parameter, and $K_\sigma(\cdot, \cdot)$ is a Gaussian kernel. The authors recommend using the points in $\mathcal{D}'$ as the centroids. The model fitting proceeds by minimizing the objective function

$$L_\lambda(\alpha) = \left[\frac{1}{2n} \alpha^T \left[\sum_{i=1}^{n} K_\sigma(X_i, X_k) K_\sigma(X_i, X_j) \right]_{k,j=1}^{k,j=n} \alpha - \left[\frac{1}{m} \sum_{k=1}^{m} K_\sigma(X_i, X_k) \right]_{i=1,\ldots,n}^{T} \alpha + \lambda ||\alpha||_1 \right] \tag{4.3.3}$$

where λ is a tuning parameter, and the first term uses observations from the training data, while the second term uses observations from the test data. The tuning parameters (σ, λ), are selected by leave one out cross validation, whose analytic form is provided by Kanamori et al. [2009]. Minimizing Equation 4.3.3 subject to $\hat{\alpha}_k \geq 0$ for all k can be computationally expensive, so in practice, Kanamori et al. [2009] recommends using an unconstrained approximation which is provably close to the constrained estimates. Then, ratio estimates are made by calculating $w(X) = \sum_{k=1}^{m} \hat{\alpha}_k K_\sigma(X, X_k)$. This approach inherits many of the favorable properties of regularized least squares models, and is computationally efficient. The efficacy of this model is demonstrated in the rightmost panel of Figure 4.3.1, where the estimated weights are almost indistinguishable from the ground truth.

4.3.2.3 Weight Regularization

In practice, p is large in many problems. Thus, the weights are likely to be either quite small or quite large, inappropriately concentrating mass on only a few points. A typical quantifier of this effect is *effective* sample size, which is defined as

$$n_{\text{eff}} = \frac{\left(\sum_{i=1}^{n} w(X_i)\right)^2}{\sum_{i=1}^{n} w(X_i)^2}.$$

To understand effective sample size, it is useful to look at the two extreme scenarios: 1) If all the weights are uniform, then $n_{\text{eff}} = n$ and 2) if the weights are a 1-hot vector, i.e. all weights are 0 except for a single entry, then $n_{\text{eff}} = 1$. Thus, the more evenly distributed the weights, the higher n_{eff}, so that effective sample size is an estimate of the equivalent sample size if all the data came from P_2.

Under large magnitude covariate shifts, the relative influence of certain points in the training set can grow, meaning a low effective sample size and model instability [Shimodaira, 2000]. To combat this, a common technique is to introduce a smoothing parameter $\gamma \in (0, 1]$, and to use weights $w(x; \gamma) = w(x)^\gamma$, which has the effect of shrinking all the weights, but shrinking the large weights more severely. Selecting γ via typical procedures such as cross-validation is challenging, because such procedures suffer from the same flaws illustrated in Section 4.1. As such, we instead suggest the following heuristic. First, fix $n_0 \in (1, n)$, typically as a fraction of the overall sample size. Then, select γ such that $n_{\text{eff}} = n_0$ when using weights $w(x)^\gamma$. This is equivalent to finding the roots of

$$f(\lambda) = \frac{\left(\sum_{i=1}^n w(X_i)^\gamma\right)^2}{\sum_{i=1}^n w(X_i)^{2\gamma}} - n_0$$

which can be calculated quickly in many software packages. In works such as Sugiyama et al. [2007], the authors recommend using importance weighted cross validation to select γ. However, this weighted cross validation is calculated only with respect to $\gamma = 1$, so that the cross validation estimate may inherit some of the undesirable properties of non-regularized weights, e.g. instability and high variance. As such, we suggest *a priori* selection of γ, which is then used in estimation of both the weighted random forest and the weighted model.

4.3.3 Tuning the Model

A key part of any predictive analysis is estimation of generalization error. Typically, this is done through methods such as repeated training/test splits, cross validation, or bootstrapping. These procedures repeatedly use uniform resampling to create training/test splits, and loss is calculated by making predictions on the held out set using a model trained on the training split. The hyper-parameters associated with the optimal score are then recorded,

and a final model is trained with those parameters. This framework is appropriate when the test set and training set are assumed to have come from the same distribution - a random sample from the empirical distribution is an unbiased approximation to a random sample from the population. The same is not true under covariate shift, but we would still like a method of tuning a model, with the goal of minimizing the generalization error under P_2, as in Sugiyama et al. [2007].

Random forests (and other bagging methods) have an additional means of estimation of the generalization error: the out-of-bag (oob) error. Each base learner is trained on only a fraction of the unique instances in the training set, creating a natural training/test split. For each split, the oob error is usually calculated as the mean squared error on the held out set, and the overall oob error is given averaging across resamples. Friedman et al. [2001] note that the oob error can be reformulated as the error associated with taking each observation (X_i, Y_i) and constructing a random forest using only trees in which (X_i, Y_i) did not appear in the sample, and then recording the loss when making a prediction at X_i using this forest. Let $B_i = \sum_{j=1}^{B} I(X_i \notin \mathcal{D}_j^*)$, i.e. the number of resamples that do not contain (X_i, Y_i), so that we can write the oob error as

$$\text{OOB}_{m,B} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{B_i}\sum_{k=1}^{B} T(X_i;\xi_k)I(X_i \notin \mathcal{D}_k^*) - Y_i\right)^2. \tag{4.3.4}$$

Because $\lim_{B\to\infty} B_i = \infty$, we can construct an infinite random forest for each point, so that by the law of large numbers, $\lim_{B\to\infty} \text{OOB}_{m,B} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}_\xi T(X_i;\xi,\mathcal{D}_{-i}) - Y_i\right)^2$. Thus, as $B \to \infty$, Equation 4.3.4 approaches the n-fold cross validation error, which is then used as an estimate of the generalization error of the forest. Similarly, we define the weighted oob error as

$$\text{OOB}_{m,B}^{\boldsymbol{w}} = \frac{1}{\sum_{j=1}^{n} w_j}\sum_{i=1}^{n} w_i\left(\frac{1}{B_i}\sum_{k=1}^{B} T_{\boldsymbol{w}}(X_i;\xi_k)I(X_i \notin \mathcal{D}_k^*) - Y_i\right)^2. \tag{4.3.5}$$

In what follows, we let $m_{B_i}(X_i) = \frac{1}{B_i}\sum_{k=1}^{B} T_{\boldsymbol{w}}(X_i;\xi_k)I(X_i \notin \mathcal{D}_k^*)$ be the random forest trained using only trees that did not see observation (X_i, Y_i). The utility of this weighted metric is a result of the following proposition.

Proposition 2. *Let* $\{Z_i\}_{i=1}^N \overset{iid}{\sim} Bernoulli(\alpha)$*, and let* $(X_i, Y_i)_{i=1}^{n+m} | Z_i \overset{iid}{\sim} Z_i P_2 + (1 - Z_i) P_1$*, where* P_1 *and* P_2 *satisfy Equation 4.1.1. Define* $m = \sum_{i=1}^N Z_i$*. Assume that* $Y_i \geq 0$ *almost surely,* $\sup_X \mathbb{E}(Y^4 | X = X) < K$ *for some constant* K*, and that*

$$\rho_n^* = \max_{k=1,2} \max_{i \neq j} Cor_{P_k} \left[(m_{B_i}(X_i) - Y_i)^2, (m_{B_j}(X_j) - Y_j)^2 \right] \to 0$$

as $n \to \infty$*. Further, assume that for all* $X \in \mathcal{X}$*,* $w_N(X)$ *is consistently proportional to the likelihood ratio,* $\ell(X) = \frac{dP_2^*(X)}{dP_1^*(X)}$*, so that* w_N *satisfies*

$$w_N(X) = c \frac{dP_2^*(X)}{dP_1^*(X)} + \epsilon_N(X) \ \forall \ X \in \mathcal{X}$$

where c *is a constant that does not depend on* X*, and* $\epsilon_N(X)$ *is a sequence of random variables satisfying* $P(\sup_X |\epsilon_N(X)| < \eta_N) = 1$*, where* $\eta_N \to 0$ *as* $N \to \infty$*. Let* $\theta_{P_2} = \mathbb{E}_{P_2}(\lim_{B \to \infty} OOB_{m,B})$*. Then, as* $B, n \to \infty$

$$OOB_{m,B}^{\boldsymbol{w}} \xrightarrow{p} \theta_{P_2}.$$

Sugiyama et al. [2007] showed that the weighted n-fold CV is *almost* unbiased for the true validation error under P_2, so that often $\theta_{P_2} = \mathbb{E}_{(X,Y) \sim P_2}(m_B(X) - Y)^2$. The upshot of this result is that we can use the weighted oob error as a consistent metric of the generalization error for data from P_2, and so minimizing the weighted oob error in training should produce a good model for data from P_2.

4.3.4 Dealing with Missing Data

A challenge of using a dataset agglomerated from many diverse sources are missing observations. Discarding missing observations is obviously unsatisfactory, but imputation should be done in a careful manner. In particular, because the procedure above relies on the training data all coming from one distribution, standard imputation procedures (such as mean imputation) effectively impose a new distribution on the missing covariates. To overcome this, we propose the following iterative procedure:

1. Let $\mathcal{M}_0 \subset \{1, ..., p\}$ denote the column indices of covariates with missing observations, and let $X_{\mathcal{M}_0} = \{X^{(j)} : \ j \in \mathcal{M}_0\}$, and similarly let $X_{-M_0} = \{X^{(j)} : \ j \notin \mathcal{M}_0\}$
2. Sample a covariate $X^{(j)}$ from the columns of $X_{\mathcal{M}_0}$ randomly. Train a random forest with $X^{(j)}$ as the response, using only data from $X_{-\mathcal{M}_0}$. This requires subsetting the dataset to $\{X_i : X_i^{(j)}$ is not missing$\}$.
3. For each $\{X_i : X_i^{(j)}$ is missing$\}$ sample $U_i \sim Unif(0, 1)$ and set $X_i^{(j)} = \hat{Q}_{U_i}(X_{i,-\mathcal{M}_0})$. Set $\mathcal{M}_1 = \mathcal{M}_0 \setminus \{j\}$.
4. Repeat steps (2)-(3), at each stage sampling covariate j_k from $\mathcal{M}_k$ to serve as the response, where $\mathcal{M}_k = \mathcal{M}_{k-1} \setminus \{j_k\}$ for $k = 1, ..., |\mathcal{M}_0|$.

This procedure is, at first glance, similar to the `missForest` procedured proposed by Stekhoven and Bühlmann [2011], who use a standard regression/classification forest to impute the missing values. These essentially use conditional mean imputation, e.g. imputation of $\mathbb{E}(X^{(j)}|X_{-j})$. However, a degenerate distribution at the conditional mean is not the same as the full conditional distribution of $X^{(j)}|X_{-j}$, and thus is incompatible with the likelihood procedure described earlier.

The process of using quantile regression for imputation is studied in Chen [2014], who studies the properties of using parametric and semi-parametric quantile regression for response imputation in a regression context. Now we make the following assumptions, which are motivated by similar assumptions and results in Meinshausen [2006].

(A1) Continuous, strictly increasing CDF Let $F_j(x|X_{-j} = x_{(-j)}) = P(X^{(j)} \leq x|X_{-j} = x_{(-j)})$ be the conditional distribution function of each covariate. Then, we assume that $x_1 > x_0 \implies F_j(x_1) > F_j(x_0)$, and that $F_j(x)$ is continuous for every $x \in \mathbb{R}$.

(A2) Access to consistent CDF estimator Assume that $\hat{F}_j(x|X_{-j} = x_{(-j)})$ satisfies

$$\hat{F}_j(x|X_{-j} = x_{(-j)}) \xrightarrow{p} F_j(x|X_{-j} = x_{(-j)}) \text{ for all } x \in \mathbb{R}, \text{ as } n \to \infty.$$

Any distribution satisfying (A1) will have a well-defined conditional quantile function, $Q_p(x_{-j}) = F_j^{(-1)}(p|X_{-j} = x_{-j})$; further, the conditional quantile function will be continuous. While the empirical CDF is not everywhere-continuous, we can still define $\hat{F}_j^{(-1)}(p) = \inf\{x : \hat{F}_j(x) \geq p\}$. Then, (A2) implies that $F_j(\hat{F}_j^{(-1)}(p)) \xrightarrow{p} \hat{F}_j(\hat{F}_j^{(-1)}(p)) = p$ for all $p \in (0, 1)$. Because $F_j^{(-1)}$ is continuous, the continuous mapping theorem gives that

$$F_j^{(-1)}(F_j(\hat{F}_j^{(-1)}(p))) = \hat{F}_j^{(-1)}(p) \xrightarrow{p} F_j^{(-1)}(p) \text{ as } n \to \infty \ \forall p \ \in (0, 1). \tag{4.3.6}$$

Equation 4.3.6 holds uniformly for p in the unit interval, so it will also hold for $U \sim Unif(0, 1)$. The probability integral transform gives that $F_j^{-1}(U)$ is a random variable with CDF F_j. The quantile regression forests of Meinshausen [2006] satisfy (**A2**) for a wide class of distributions, and so the upshot of this result is that the imputation scheme suggested above provides a consistent way of generating imputations that follow P_1^*. Thus, this imputation scheme is compatible, asymptotically, with the likelihood ratio procedure described earlier.

4.4 Simulations

We now provide a variety of simulations to demonstrate the utility of our proposed method in various settings.

4.4.1 An Illustrative Regression Example

We begin with a simple example of a covariate shifted model, and demonstrate that the weighted forest can indeed pick up on local behavior. The model for the simulation is given by $Y|X \sim \mathcal{N}(\varphi(X), 0.5)$, for $\varphi(X) = \max\left\{\frac{e^X}{1+e^X}\sin(X), \frac{e^{-X}}{1+e^{-X}}\sin(-X)\right\}$, where $\varphi(x)$ has considerable local structure.

To simulate covariate shift, we draw training data according to $P_1(X) = \mathcal{N}(-4, 3.5^2)$ and testing data according to $P_2(X) = \mathcal{N}(3.5, 1.5^2)$. The training distribution is quite dispersed, whereas the test distribution concentrates mass around a particular region of the real line. We implement locally optimized random forests using two sources of weights: 1) Learned weights from the method of Kanamori et al. [2009] and 2) oracle weights, corresponding to $\ell(X) \propto \frac{\phi\left(\frac{X-3.5}{1.5}\right)}{\phi\left(\frac{X+4}{3.5}\right)}$, where $\phi(\cdot)$ is the standard normal density function. We draw $n = 500$ and $n_{\text{test}} = 250$ points from the shift model as the validation set. Results are shown in Figure 4.4.1. We see that the unweighted forest struggles to pick up on the main signal in the test area, while the oracle weighted and learned weighted forests come much closer to the true signal. The unweighted forest fits a constant function on the high mass regions of P_2, whereas the oracle/learned weight forests are much closer to the truth. Note that this improvement comes at the cost of decreased performance in the region around $X = 0$, but this area is does not contribute much mass to the RMSE under P_2. The learned weights are approximately correct until around $X = 3$, at which point the lack of data in this region leads to a decline in weight performance. Running this simulation over 150 runs, we see that on average the ranger model has $RMSE = 0.2440$, the learned weighted model has $RMSE = 0.1565$ and the oracle weighted model has $RMSE = 0.1133$. While model performance is more than just RMSE, we see a convincing case that the weighted forest is able to adapt to a specified region of interest.

4.4.2 High Dimensional Simulation

We now compare our procedure against a baseline random forest. The random forest models used are trained using the `ranger` package [Wright and Ziegler, 2015]. For computational efficiency, the resampling is done without replacement so that each tree is trained on

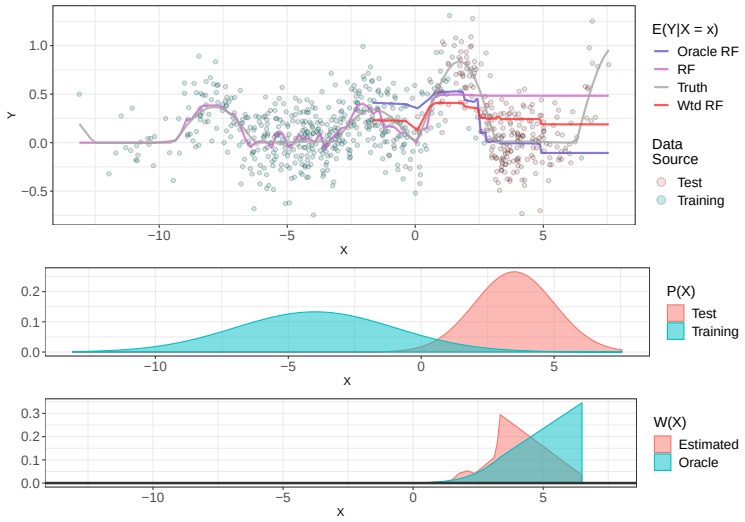


Figure 4.4.1: *Top:* Fitted functions according to the three tested models, along with an overlay of the training points. *Center:* The training and test densities used. *Bottom:* Estimated density ratio terms and true density ratio terms.

$k_n < n$ unique observations. Since approximately 63% of the dataset is represented in a given bootstrap resample, so we take $k_n = 0.6n$. Implementation of the weighted forest is done using the `rpart` package using the weights option [Therneau et al., 1997]. For each model, we build $B = 500$ trees. As an additional point of comparison, we apply the customized forest method of Powers et al. [2015] with 5 clusters, generating 5 random forests.

We draw 150 datasets of size $n = 1000$ with $p = 31$ covariates along with $n_{test} = 200$ points to be used as a validation set. The covariate distribution is given by

$$[X^{(1)}, ..., X^{(6)}] \sim \text{Dirichlet}(\boldsymbol{\alpha})$$
$$X^{(7)}, ..., X^{(31)} \overset{iid}{\sim} \text{Uniform}(0, 1)$$

where $\boldsymbol{\alpha}$ is a pre-specified parameter. For the training set, we use $\boldsymbol{\alpha}_1 = \lambda^{[1,2,3,4,5,6]}$ and for the test set, we use $\boldsymbol{\alpha}_2 = \lambda^{[6,5,4,3,2,1]}$, where $\lambda > 0$ is a parameter that controls how disparate the densities are (higher λ leads to higher discrepancy). In these simulations, we use $\lambda \in \{$`1, 1.07, 1.14, 1.21, 1.29, 1.36, 1.43, 1.5`$\}$ - noting that $\lambda = 1$ is the case where $P_1 = P_2$.

Model #	Data Generating Model
1	$Y = 5X^{(1)} + \epsilon$
2	$Y = 5\sin(\pi X^{(1)}) + \epsilon$
3	$Y = 10\sin(\pi X^{(1)} X^{(2)}) + 20(X^{(3)} - 0.5)^2 + 10X^{(4)} + 5X^{(5)} + \epsilon$
4	$Y = 5e^{2\sqrt{X^{(1)} X^{(2)}} + X^{(6)}} + \epsilon$
5	$Y = 5\sum_{j=1}^{5} \left(X^{(j)}\right)^2 + \epsilon$

Table 4.4.1: Distributions of $Y|X$ for each model used in the simulation. In each case, ϵ is mean 0, Gaussian noise with $\mathbb{E}(\epsilon^2) = 0.25$.

Note that P_2 concentrates much more density on $X^{(5)}$ and $X^{(6)}$ than P_1, but they still have the same support. The inclusion of 25 predictors whose distribution does not change is to reflect the fact that P_1 and P_2 may include the same marginal distribution for many covariates. We simulate a response, Y, using several different response functions, summarized in Table 4.4.1.

Model 1 is intended to demonstrate a situation where the marginal distribution of Y may vary dramatically between P_1 and P_2. Model 2 shows a situation where the conditional mean is a periodic function of $X^{(1)}$, so discrepancies in the magnitude of $X^{(1)}$ should affect the response less adversely. Model 3 is the popular MARS simulation model [Friedman, 1991], which has been used as a stand-in for a complex regression function in previous work [Mentch and Hooker, 2016a, Xu et al., 2016]. Model 4 similarly represents a complex function with a discontinuity. Finally, Model 5 represents a model where the marginal distribution of Y is agnostic to changes between P_1 and P_2.

4.4.3 Simulation Results

We analyze simulation results over both the data generating model and over the λ parameter which controls the discrepancy in P_1 and P_2. The resulting scores (calculated according to Equation 4.2.1), RMSEs, and coverage probabilities are shown in Figure 4.4.2. Tables of results are ommitted from the main text for conciseness, and instead are available in the supplemental materials.

In general, according to the score metric, the weighted forest performs better than the unweighted forest in Models 1 and 2. Moreover, performance is stronger in models 3 and

4 until a certain point, when the shift becomes too large. In model 5, unsurprisingly, the weighted and unweighted forest perform near identically, because the marginal distribution of Y is not changing drastically. Further, looking at the RMSE plots, we see that the weighted forest is consistently able to attain a lower error rate than the unweighted forest in Models 1-4, with some breakdown at high λ. The one area where performance of the weighted model is somewhat worse than unweighted model is in coverage percentage, where the prediction intervals have slightly lower coverage in many of the situations. However, we note that the weighted procedure still maintains the nominal coverage in all cases for small values of λ. Moreover, in Models 1 and 2, the shift affects the weighted forest less severely than in Models 3 and 4. Finally, results presented in the appendix show that the weighted forest incurs much smaller prediction intervals than those of the unweighted procedure. Thus, the weighted forest sacrifices some small coverage probability (and often does not drop below the nominal level) in exchange for much narrower prediction intervals. Other than in Model 5, the customized procedure lags well behind both the weighted and unweighted random forests using the score metric, particularly when considering the quantile regression coverage. The customized procedure does provide competitive RMSE's

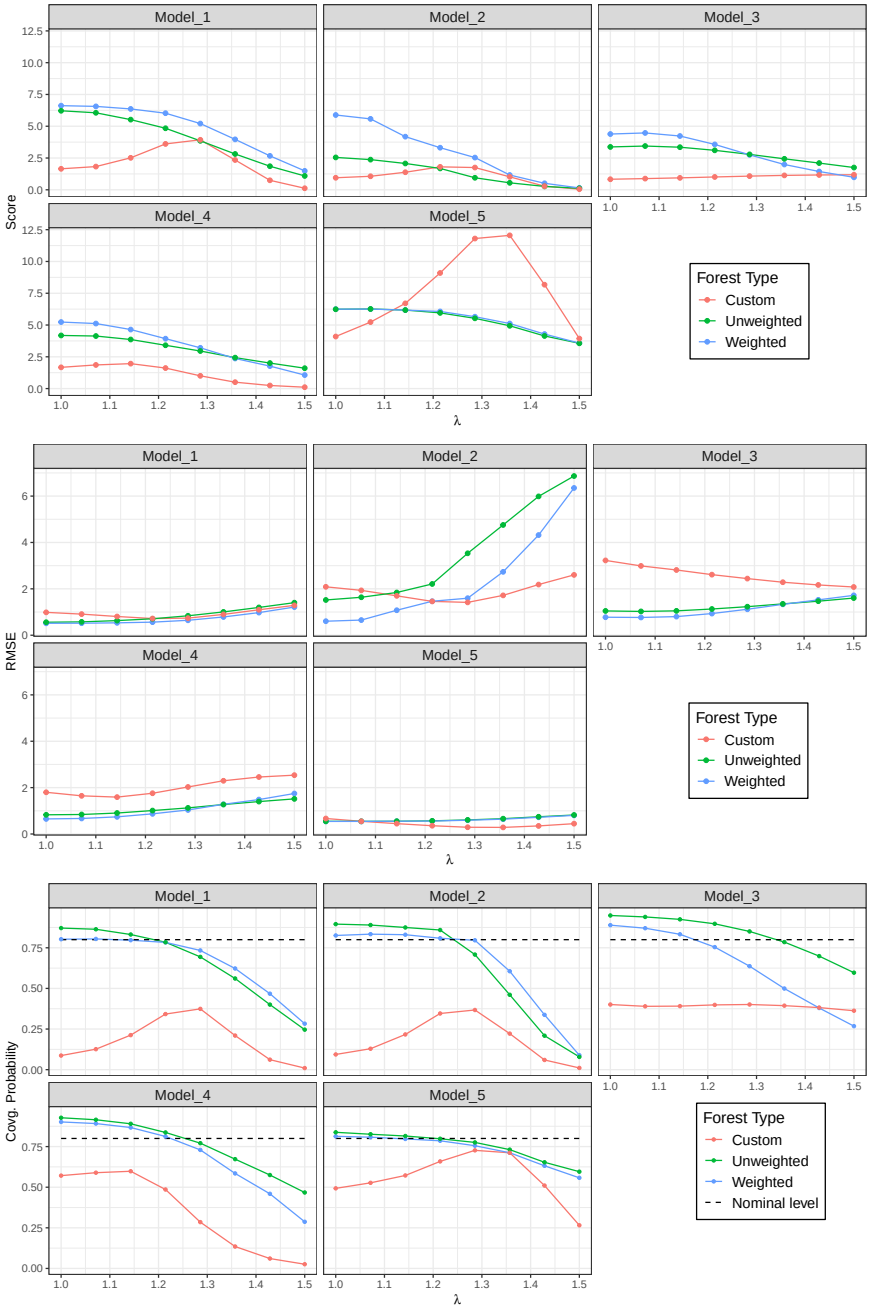


Figure 4.4.2: Results for the Score (top), RMSE (center), and Coverage probabilities (bottom) from the simulation study from subsection 4.4.2. The dashed line in the bottom indicates the nominal coverage level, 0.80.

4.5 Application to Hurricanes

We now turn to the problem of forecasting hurricane power outages. To begin, we apply this procedure described in subsection 4.3.4 to impute the missing values in the training data. In total, 26 columns had missingness and there were a total of 12244 observations that needed imputation, a non-negligible portion of the dataset. We note that because of how the training/test splits overlap from storm to storm, the imputation procedure covers both the training and test sets. We fit a weighted forest and a random forest with `mtry` $= 50$ and `nodesize` $= 5$, corresponding to the parameters suggested from Table 4.2.1. For the weighted model, we again use the method of Kanamori et al. [2009] to estimate the weights. Moreover, we fix the minimum effective sample size at $n_0 = 0.75n$ and run the optimization procedure from subsection 4.3.2 to estimate the weight regularization λ. The results are presented in Table 4.5.1. We see that the performance in general is similar between the weighted and unweighted models, but the weighted model provides slight improvements in Harvey, Irma, and Matthew, in terms of the score metric.

Storm	Model	λ	RMSE	MAE	Covg	Interval Width	Score
Harvey-2017	Weighted	0.1305	0.9097	0.7327	0.8014	2.5033	3.6171
Harvey-2017	Unweighted	0.1305	0.9069	0.7467	0.7679	2.4358	3.4848
Irma-2017	Weighted	0.0084	1.4021	1.1608	0.4615	2.4658	1.6394
Irma-2017	Unweighted	0.0084	1.4111	1.1786	0.3776	2.3777	1.3592
Sandy-2012	Weighted	1.0000	1.2286	1.0357	0.5391	2.1310	2.1901
Sandy-2012	Unweighted	1.0000	1.2204	0.9876	0.5521	2.2075	2.2353
Nate-2017	Weighted	0.3602	0.8355	0.7225	0.8528	2.6684	3.8660
Nate-2017	Unweighted	0.3602	0.8154	0.6746	0.8615	2.4930	4.1285
Matthew-2016	Weighted	1.0000	0.7932	0.6193	0.8898	2.5561	4.3897
Matthew-2016	Unweighted	1.0000	0.7943	0.6298	0.8924	2.6867	4.2988
Arthur-2014	Weighted	1.0000	1.0724	0.8634	0.6721	2.3111	2.8540
Arthur-2014	Unweighted	1.0000	1.0616	0.8432	0.6745	2.2994	2.8983

Table 4.5.1: Model performance by storm, with weighted and unweighted storms fitted. Bolded values represent the better of the two by storm and loss function. λ value reported is selected by the effective sample size calculation from subsection 4.3.2.

As a followup, we additionally implemented a study of tuning the model using the weighted out-of-bag metric from subsection 4.3.3. To do this, we tune the `mtry` parameter over a grid consisting of $\mathcal{M} = \{27, 39, 51, 63, 75\}$ for both an unweighted and weighted random forest. For the weighted forest, we record $\text{OOB}^{\boldsymbol{w}}_{m,B}$ and the weighted RMSE, and

$\text{OOB}_{m,B}$ and the unweighted RMSE. The results are shown in Figure 4.5.1, where the out-of-bag error for each mtry value is plotted against the RMSE of that model. For all storms except Hurricane Nate, we see that both $\text{OOB}_{m,B}$ and $\text{OOB}^{\boldsymbol{w}}_{m,B}$ dramatically underestimate the holdout RMSE, with the weighted out-of-bag error providing a slightly less biased estimate. However, in the context of model selection, typically the model with the lowest out-of-bag error (and thus lowest estimated generalization error) is selected. Thus, for model selection purposes, the generalization error estimate is less important than the ranking. We see that the weighted oob error selects an optimal model for Hurricane Matthew, and a near optimal model for hurricanes Irma and Sandy, while the unweighted model selects an optimal model for Hurricane Sandy, and a near optimal model for Irma, Nate, and Matthew. Moreover, for Hurricane Matthew, the OOB-RMSE rankings are recovered exactly, and for Hurricane Irma the same is true with the exception of one `mtry` value. In the unweighted case, there are no such clear stories.

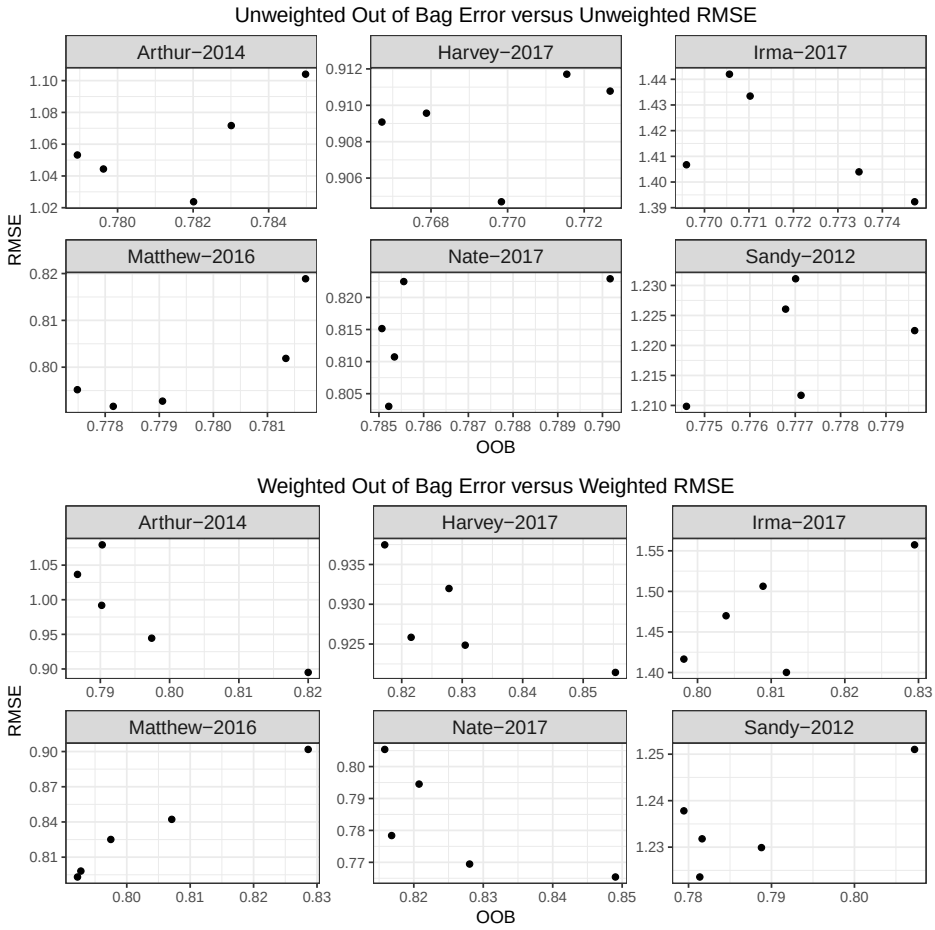


Figure 4.5.1: Out-of-bag error versus holdout RMSE. **Top:** Results for the unweighted forest. **Bottom:** Results for the weighted forest.

4.6 Conclusion

We proposed a modification to the random forest algorithm to account for distributional changes between test and training sets, which often arise in practice. We accomplish this by imposing a covariate shift assumption, and then using existing density ratio methods to estimate the likelihood ratio weights, $\ell(X) \propto \frac{dP_2(X)}{dP_1(X)}$. We moreover provided methods for imputing missing data and tuning the model in ways that respect the statistical assumptions associated with the problem. The simulation study clearly demonstrates the utility of the proposed method - the importance weighted forest typically outperforms a standard random

forest in the covariate shift case. However, importance weighting is only able to address small changes in covariate distribution. Indeed, in Figure 4.4.2 it was shown that both the weighted and unweighted forest perform worse as the magnitude of the shift grows.

5.0 Summary and Additional/Future Work

Random forests are accurate, computationally efficient machine learning tools that are widely used in practice because of their ease of implementation, particularly in applications to medicine and environmental sciences. Despite the ease of training a random forest, interpreting the outputs remains challenging, with flawed, *ad hoc* tools remaining quite popular, such as the out-of-bag metrics. In chapters 2 and 3, we proposed tools for random forest interpretation that are statistically meaningful. The tools in chapter 2 were catered to the tree swallow migration problem, while the tools in chapter 3 are quite general. Additionally, in chapter 4, we presented an importance-sampling modification to the random forest algorithm to account for covariate shift, and demonstrated the predictive utility of the modification via simulations and the application to hurricane outage forecasting.

There are several avenues for methodological improvement, which have been presented in the conclusions of each chapter. Future work could incorporate the methodological suggestions from these sections. However, more promising are further applications to environmental and climate science. Climate change, in particular, and its human impacts, are of great interest given the current trajectory of carbon emissions. Presently, some of our work is on applications to studying the influencing factors of permafrost thaw in the Arctic. Permafrost thaw is of significant research interest because it presents a possible positive feedback loop in the climate system: warming temperatures due to greenhouse gas emissions melts permafrost, potentially releasing more greenhouse gasses. The magnitude and nature of this feedback loop is the subject of intense research currently, e.g. [Lawrence and Slater, 2008, Koven et al., 2011, Meehl et al., 2012, Lawrence et al., 2015]. Analyzing the dynamics of the feedback loop via machine learning models, such as random forests, is especially exciting.

Appendix A Chapter 2 Appendix

A.1 Moran's I

We now describe the formal procedure used to evaluate the significance of spatial autocorrelation in Figure 2.3.4, which was mentioned in section 2.3. In order to formally test the following hypotheses

$$\begin{aligned} \mathrm{H}_0 &: \mathrm{cov}(d_{ij}, d_{i'j'}) = 0 \quad \text{for all } \; ij \neq i'j' \\ \mathrm{H}_1 &: \mathrm{cov}(d_{ij}, d_{i'j'}) \neq 0 \quad \text{for some } \; ij \neq i'j' \end{aligned}$$

we first must define a distance matrix between points in the grid. Here we calculate pairwise distances

$$\delta_{ij,i'j'} := \sqrt{(\mathrm{lat}_{ij} - \mathrm{lat}_{i'j'})^2 + (\mathrm{lon}_{ij} - \mathrm{lon}_{i'j'})^2}$$

as the Euclidean distances between the latitude/longitude coordinates in the grid and utilize an inverse distance weighting scheme

$$w_{ij,i'j'} := \frac{1}{\delta_{ij,i'j'}}.$$

The distance between a point and itself is 0, but we assign $w_{ij,ij} = 0$, as is standard practice. For computational feasibility, we make these calculations on only a sub-grid consisting of every sixth point in the original test grid. To evaluate the hypotheses, we use the test statistic [Moran, 1948]

$$I_{\mathrm{obs}} = \frac{N \sum_{ij,i'j'} w_{ij,i'j'}(d_{ij} - \bar{d})(d_{i'j'} - \bar{d})}{W \sum_{ij}(d_{ij} - \bar{d})^2}$$

as our test statistic, where W is the sum of all entries in the weight matrix and N is the number of grid points (in our case, 5822). We then calculate the standardized statistic $Z^* = (I_{\mathrm{obs}} - \mathbb{E}_0 I_{\mathrm{obs}})/\hat{\sigma}(I_{\mathrm{obs}})$ which is asymptotically standard normal under H_0. The calculated I_{obs} values are reported in table A.1.1. Note that $\mathbb{E}_0 I_{\mathrm{obs}} = -1/(N-1) \neq 0$.

Day of Year	1	21	41	61	81	101	121	141	161
I_{obs}	0.1782	0.1594	0.1609	0.2150	0.3291	0.1870	0.1325	0.1119	0.1281
Z^*	477	427	431	576	881	502	355	300	343
P-value	0	0	0	0	0	0	0	0	0

Table A.1.1: Moran's I test for spatial autocorrelation for each of the days for which a prediction map was generated, see Figure 2.3.4

A.2 S.3 A Causal Inference Analysis

We now implement an analysis to estimate the effect of `max_temp_anomaly` on `occurrence`. We note that the quantity of interest here is distinct from that in the main text, and so we begin with a brief overview of the potential outcomes framework and causal random forests. For a continuous treatment W, the average treatment effect at a point x is defined as

$$\tau(x) = \frac{\text{Cov}(Y, W|X = x)}{\text{Var}(W|X = x)}$$

so that $\tau(x)$ measures the average *linear* effect of W on Y given covariates $X = x$. This is an extension of the potential outcomes framework [Rubin, 2005], which defines the counterfactual treatments, $Y^{(w)}$ for each w in the support of W. The fundamental goal of causal random forests is to estimate $\tau(x)$ [Wager and Athey, 2018]. Causal trees, for continuous treatments, proceed by recursively partitioning the feature space until some stopping criterion is met and then performing local linear regression of Y on W within the terminal nodes. Then, a prediction of $\tau(x)$ is given by the estimated treatment effect within the terminal node containing x. The forest is generated by repeatedly creating randomized trees, and averaging the estimated treatment effect from each tree. Wager and Athey [2018] showed that if several regularity conditions are enforced upon training of the trees, then $\hat{\tau}(x)$ is asymptotically consistent to the true effect.

To interpret $\hat{\tau}(x)$ as an estimate of a causal effect, one must place an additional assumption on the distribution of the data, namely *unconfoundedness*, which states that

$$Y_i^{(w)} \perp\!\!\!\perp W_i \mid X_i \quad \forall \quad w.$$

Essentially, the response, given a particular treatment assignment w, needs to be locally (in X) independent of the process by which treatments are assigned. The assumption can also be viewed as stating that the covariates X are a sufficient *adjustment set* to infer the causal effect of W on Y.

In our application, Y is `occurrence`, W is `max_temp_anomaly`, and X are the remaining covariates such as land cover, time of year, user characteristics, etc. Thus, the unconfoundedness assumption translates to assuming that the distribution of temperature anomalies is independent of the distribution of potential occurrences for all possible temperatures, conditional on the other covariates. We believe that the unconfoundedness assumption is likely unrealistic in this situation. Consider that both `occurrence` and `max_temp_anomaly` are both realizations of time series with high serial dependence. In Figure A.2.1, we present two different directed acyclic graph (DAG) representations of the time series structure of the data, fixed at a location x. In the left panel, if we assume that W_t are sequentially independent, but perhaps day-to-day occurrences are dependent, then unconfoundedness holds. However, in the right panel, we model the more realistic scenario, where both the treatment and outcome are serially dependent, so that $Y_t^{(w_t)}$ and W_t are confounded by W_{t-1}. As such, we have reason to doubt that $\tau(x)$ is identifiable from this data. Further, we do not observe W_{t-1}, and so cannot include it in an adjustment set. The problem becomes more intractable when one considers that the causal mechanism between W and Y is primarily insect activity, as described in the introduction of the main text, which acts as an unobserved variable. Because insect activity is *also* serially dependent and unobserved, it can effectively make the adjustment set require infinite history of the time series observations, making causal inference impossible [Malinsky and Spirtes, 2018].

With these caveats in mind, we now apply the causal forest algorithm to the data used in Section 5 of the main text. We use the same training and test points described in the main text and apply the causal forest algorithm implemented in the `grf` package [Athey et al., 2019] with the default parameters (including tree honesty) enabled. Then, predictions $\hat{\tau}(x)$ were recorded for each of the stratified test points.

It is hard to discern any spatial trend in $|\hat{\tau}(x)|$ from the left panel of Figure A.2.2. However, the time series plots tell an intuitive story - negative temperature anomalies lead

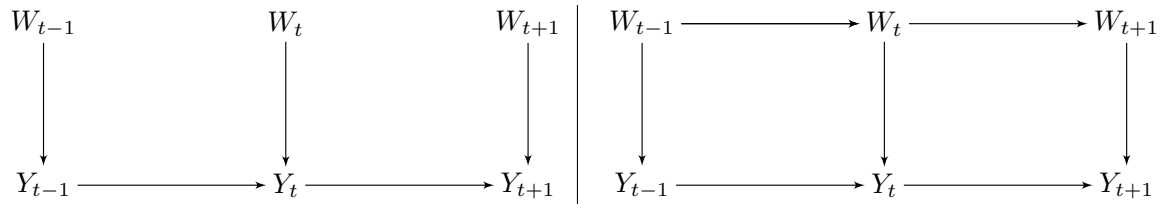


Figure A.2.1: Two different directed acyclic graphs (DAGs) describing the relationship between W_t and Y_t. **Left:** A treatment scheme that would satisfy unconfoundedness. **Right:** A more likely DAG, for which unconfoundedness does not hold.

to reduced occurrence earlier in the season (particularly in the the northern testing zones), followed by a flattening of the effect later in the migration season. Notice that the flattening occurs the earliest in Zone 1 (around `DoY` 270), and latest in Zone 6 (around `DoY` 320), which again coincides with spatial differences in seasonality. We note that these results are in consensus with the formal testing from Section 4 of the main text, where differences between the regression functions died down later in the year.

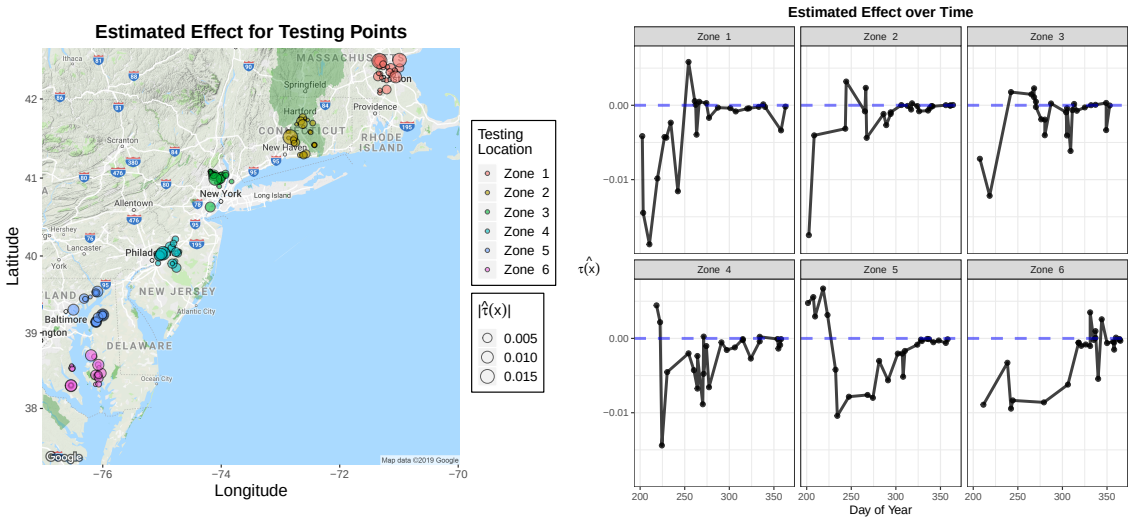


Figure A.2.2: Left: Absolute causal forest estimates $|\hat{\tau}(x)|$ for each point in the test set used in Section 5 of the main text. Size of the circle corresponds to magnitude of the estimated treatment effect. **Right:** A plot of $\hat{\tau}(x)$ over time in each zone.

Appendix B Chapter 3 Appendix

B.1 Proofs of Technical Results

We now provide the technical details and proofs for theoretical discussion in Section 3. For completeness, theorems and lemmas are restated.

Theorem 1. *Under the exchangeability conditions outlined in Section 3.1, denote a sequence of (potentially randomized) trees trained on subsamples from $\mathcal{D}_n$ as $\{T_k(\cdot)\}_1^\infty$. Moreover, consider an independently drawn test point, $\boldsymbol{Z}^* = (X^*, Y^*) \sim F$. Then, the residuals*

$$r_k = T_k(X^*) - Y^*$$

form an infinitely exchangeable sequence of random variables.

Proof. Let Ξ be the distribution of randomization parameters, and let $\mathcal{S}_{k_n}(\mathcal{D}_n)$ be the distribution of subsamples of size k_n drawn uniformly from the original data. Then, to construct a tree, we have the following procedure:

1. Draw $\mathcal{D}^*_{k_n} \sim \mathcal{S}_{k_n}(\mathcal{D}_n)$
2. Draw $\xi \sim \Xi$
3. Draw $\boldsymbol{Z}^* \sim F$
4. Construct a tree according to some combining function, say ϕ , of $\xi, \mathcal{D}^*_{k_n}$, i.e. $T = \phi(\xi, \mathcal{D}^*_{k_n})$.

Each draw is done independent of the other draws. Repeating (1) and (2) independently gives iid sequences $\{\mathcal{D}^*_{l,k_n})\}_{l=1}^\infty$ and $\{\xi_j\}_{j=1}^\infty$. Then, the sequence

$$T_1 = \phi(\xi_1, \mathcal{D}^*_{1,k_n}),\ T_2 = \phi(\xi_2, \mathcal{D}^*_{2,k_n}), ...$$

is a mixture of iid sequences, where the mixture is directed (in the sense of Aldous [1985]) by $\mathcal{D}_n$. So, $\{T_l \mid \mathcal{D}_n\}$ is exactly an iid sequence of functions. Further, $\{r_l \mid \mathcal{D}_n, \boldsymbol{Z}^*\}$ is an iid sequence of random variables, and thus the conclusion follows from the converse of DeFinetti's Theorem. □

Aldous [1985] page 29 provides more details on this construction and the implications of De Finetti's Theorem. We turn now to Lemma 1 from Section 3.2, which establishes asymptotic pairwise independence of subsampled decision trees.

Lemma 1. *Consider a collection of B_n trees built from a training dataset of size n on subsamples of size k_n, say $\{T_{j,k_n}\}_{j=1}^{B_n}$, satisfying Condition 1. Then, as long as $k_n/\sqrt{n} \to 0$ and*

$$\binom{B_n}{2} \log\left[\frac{\binom{n-k_n}{k_n}}{\binom{n}{k_n}}\right] \to 0$$

the infinite sample sequence of trees, $\{T_{1,\infty,k_\infty}, ..., T_{B,\infty,k_\infty}, ...\}$ is an infinite sequence of pairwise independent random functions.

Proof. Condition 1 guarantees the existence of a limiting random variable. It is sufficient to show that asymptotically, the trees are trained using independent training samples, because we have assumed that our original data are iid. Define the indices of a subsample in the following way:

$$\text{ind}(\mathcal{D}^*_{k_n}) := \{j \in \{1, ..., n\} : Z_j \in \mathcal{D}^*_{k_n}\}.$$

Then, by the assumption that the Z_k are independent,

$$\mathcal{D}^*_{k_n,j} \perp\!\!\!\perp \mathcal{D}^*_{k_n,l} \iff |\text{ind}(\mathcal{D}^*_{k_n,j}) \cap \text{ind}(\mathcal{D}^*_{k_n,l})| = 0$$

so, it is sufficient to show that

$$\lim_{n\to\infty} P(|\text{ind}(\mathcal{D}^*_{k_n,j}) \cap \text{ind}(\mathcal{D}^*_{k_n,l})| = 0) = 1, \ \forall\ j \neq l.$$

Note that if $k_n \geq n/2$, this event has probability 0, so choose n so that $n > 2k_n$. Then

$$\begin{aligned}
P(|\text{ind}(\mathcal{D}^*_{k_n,j}) \cap \text{ind}(\mathcal{D}^*_{k_n,l})| = 0) &= \frac{\binom{n-k_n}{k_n}}{\binom{n}{k_n}} \\
&= \frac{((n-k_n)!)^2}{n!(n-2k_n)!} \\
&= \frac{(n-k_n)!}{n!} \times \frac{(n-k_n)!}{(n-2k_n)!} \\
&= \frac{(n-k_n)(n-k_n-1)...(n-2k_n+1)}{n(n-1)...(n-k_n+1)}.
\end{aligned}$$

There are k_n terms in both the numerator and denominator here, so we can separate the product in the term above as

$$\begin{aligned}
P(|\text{ind}(\mathcal{D}^*_{k_n,j}) \cap \text{ind}(\mathcal{D}^*_{k_n,l})| = 0) &= \frac{n-k_n}{n} \times \frac{n-k_n-1}{n-1} \times ... \times \frac{n-2k_n+1}{n-k_n+1}. \\
&\geq \left(\frac{n-2k_n+1}{n}\right)^{k_n} \\
&= \left(1 - \frac{2k_n+1}{n}\right)^{k_n} \\
&= \exp\left[k_n \log\left(1 - \frac{2k_n+1}{n}\right)\right] \\
&\approx \exp\left[k_n\left(-\frac{2k_n+1}{n}\right) - \frac{k_n}{2}\left(\frac{2k_n+1}{n}\right)^2\right] \\
&\approx \exp\left[-\frac{2k_n^2+k_n}{n}\right] \\
&\approx 1
\end{aligned}$$

where $a_n \approx b_n$ means that $\lim_{n\to\infty} a_n/b_n = 1$, and we have used the Taylor expansion of $\log(1-x)$ in the above.

This means that two pre-specified subsamples will be independent in the limit. Next, we need to ensure that this holds for all subsamples, i.e.

$$P\left(\bigcap_{j\neq l}\{|\text{ind}(\mathcal{D}^*_{k_n,j}) \cap \text{ind}(\mathcal{D}^*_{k_n,l})| = 0\}\right) \to 1.$$

For B_n trees, there are $\binom{B_n}{2}$ subsample pairings, each drawn independently. Thus

$$\begin{aligned}
P\left(\bigcap_{j\neq l}\{|\text{ind}(\mathcal{D}^*_{k_n,j}) \cap \text{ind}(\mathcal{D}^*_{k_n,l})| = 0\}\right) &= \prod_{j\neq l} P(|\text{ind}(\mathcal{D}^*_{k_n,j}) \cap \text{ind}(\mathcal{D}^*_{k_n,l})| = 0) \\
&= \left(\frac{\binom{n-k_n}{k_n}}{\binom{n}{k_n}}\right)^{\binom{B_n}{2}}.
\end{aligned}$$

Next, by assumption,

$$\log P\left(\bigcap_{j\neq l}\{|\text{ind}(\mathcal{D}^*_{k_n,j}) \cap \text{ind}(\mathcal{D}^*_{k_n,l})| = 0\}\right) = \binom{B_n}{2} \log\left[\frac{\binom{n-k_n}{k_n}}{\binom{n}{k_n}}\right] \to 0$$

so that the probability of this event goes to 1. □

Now, we more formally derive the asymptotic distribution of the MSE statistic, using a delta method argument. This discussion derives the results in subsection 3.3.3. In what follows, we suppress the dependence on y, writing just $MSE_{RF}(X;y) = g\left(RF_B(X)\right)$. We derive the asymptotic distribution of the MSE via the delta method, which we belabor here for its intuitive value. We can then appeal to the mean value theorem to say

$$g(RF_B(X)) = g(\mathbb{E}RF_B(X)) + g'(\tilde{R}_B(X))[RF_B(X) - \mathbb{E}RF_B(X)]$$

where $\tilde{R}_B(X)$ is a random quantity bounded between $RF_B(X), \mathbb{E}RF_B(X)$. The law of large numbers gives that $RF_B(X) = \mathbb{E}RF_B(X) + o_P(1)$ and further $\tilde{R}_B(X) \xrightarrow{p} \mathbb{E}RF_B(X)$. Next, continuity of g' gives that $g'(\tilde{R}_B(X)) \xrightarrow{p} g(\mathbb{E}RF_B)$. Thus,

$$\begin{aligned}\sqrt{B}\left[g(RF_B(X)) - g(\mathbb{E}RF_B(X))\right] &= g'(\tilde{R}_B(X))\sqrt{B}\left[RF_B(X) - \mathbb{E}RF_B(X)\right] \\ &\xrightarrow{d} \mathcal{N}\left(0, g'(\mathbb{E}RF_B(X))^2\sigma^2\right) \\ &\stackrel{d}{=} \mathcal{N}\left(0, 4(\mathbb{E}RF_B(X) - y)^2\sigma^2\right) \text{ for } g(z) = (z-y)^2.\end{aligned}$$

The calculation above is more informative - we see that the MSE is asymptotically a linear function of the random forest prediction. An issue is that the above quantity is centered around $g(\mathbb{E}RF_B(X))$ rather than $\mathbb{E}g(RF_B(X))$, which we now address. In particular, suppose we begin by centering around $\mathbb{E}g(RF_B(X))$ rather than $g(\mathbb{E}RF_B(X))$. Then,

$$\begin{aligned}&\sqrt{B}\left[g\left(RF_B(X)\right) - \mathbb{E}g\left(RF_B(X)\right)\right] = \\ &\qquad \sqrt{B}\left[g\left(RF_B(X)\right) - g(\mathbb{E}RF_B(X))\right] + \sqrt{B}\left[g\left(\mathbb{E}RF_B(X)\right) - \mathbb{E}g\left(RF_B(X)\right)\right] \quad \text{(B.1.1)}\end{aligned}$$

so that if $\sqrt{B}\left[g\left(\mathbb{E}RF_B(X)\right) - \mathbb{E}g\left(RF_B(X)\right)\right] = o(1)$, then the same distributional result holds. This is shown in Lemma 3.

After Lemma 1, we next need to prove Lemma 3, whose purpose is to show that the observed MSE is asymptotically centered around its own expectation.

Lemma 3 *Assume the conditions needed from Corollary 2. Additionally, assume that g has at least k derivatives for some $k \geq 3$, and that $g^{(k)}(x) < \infty$ for all x. Further, assume that $\mathbb{E}|T_i(X)|^k < \infty$. Then,*

$$\sqrt{B}\left[\mathbb{E}g(RF_B(X) - g(\mathbb{E}RF_B(X))\right] = \frac{g''(\mathbb{E}RF_B(X))\sigma^2}{2\sqrt{B}} + o(B^{-3/2}).$$

Proof. We rely on a result presented in Oehlert [1992], which states that under the conditions presented in the lemma statement,

$$\mathbb{E}g(RF_B(X)) = g(\mathbb{E}RF_B(X)) + \frac{g''(\mathbb{E}RF_B(X))\sigma^2}{2B} + o(B^{-2}). \tag{B.1.2}$$

Thus, the result follows from multiplying both sides of Equation B.1.2 by $\sqrt{B}$ and rearranging terms. □

As a quick note about the result above, recall that application of the mean value theorem requires that $g'(\mathbb{E}RF_B(X)) \neq 0$, which occurs if and only if $\mathbb{E}RF_B \neq y$. The expected prediction can be written as $\mathbb{E}RF_B(X) = m(X) + \delta(X)$, where $\delta(X)$ is the pointwise bias of the random forest. Recalling that the response is given by $Y = m(X) + \epsilon$, if it holds for all X that $P(\epsilon \neq \delta(X)) = 1$, then the result holds for the squared error calculated with respect to almost all Y and thus is trivially satisfied for continuous errors. A similar result could be applied to any continuously differentiable loss function $g(\cdot, \cdot)$, again under the condition that g' is almost surely non zero.

Next, we include details about the limiting distribution of the MSE at multiple points. To calculate τ^2, for the MSE at each point in $\mathcal{T}$, let $g_j(RF_B(X_j)) = (RF_B(X_j) - Y_j)^2$, by continuity, $g_j'(\tilde{R}_B(X_j)) = g_j'(\mathbb{E}RF_B(X_j)) + o_P(1)$. Thus, we see that

$$\begin{aligned}
MSE_{RF}(\mathcal{T}) &= \frac{1}{N_t}\sum_{j=1}^{N_t} MSE_{RF}(X_j, Y_j) \\
&= \frac{1}{N_t}\sum_{j=1}^{N_t} g_j'(\mathbb{E}RF_B(X_j))\left[RF_B(X_j) - \mathbb{E}RF_B(X_j)\right] + o_P(1) \\
&= \frac{1}{N_t}\sum_{j=1}^{N_t} g_j'(\mathbb{E}RF_B(X_j))\left[\frac{1}{B}\sum_{i=1}^{B}\left[T_i(X_j) - \mathbb{E}RF_B(X_j)\right]\right] + o_P(1) \\
&= \frac{1}{B}\sum_{i=1}^{B}\underbrace{\frac{1}{N_t}\sum_{j=1}^{N_t} g_j'(\mathbb{E}RF_B(X_j))\left[T_i(X_j) - \mathbb{E}RF_B(\boldsymbol{X_j})\right]}_{\bar{T}_i} + o_P(1)
\end{aligned}$$

where $g_j(\cdot)$ is used to suggest that the squared difference is calculated with respect to Y_j. $\bar{T}_i$ is an iid sequence, so that $\sqrt{B}[MSE_{RF}(\mathcal{T}) - \mathbb{E}MSE_{RF}(\mathcal{T})]$ is asymptotically an iid sum with mean 0 and variance $\sigma^2_{\bar{T}}$ given by

$$\sigma^2_{\bar{T}} = \frac{1}{N_t}\sum_{j=1}^{N_t}\sigma_j^2\left(g_j'(\mathbb{E}RF_B(X_j))\right)^2 + \frac{2}{N_t}\sum_{i<j} g_j'(\mathbb{E}RF_B(X_j))g_i'(\mathbb{E}RF_B(X_i))\rho_{ij} \qquad \text{(B.1.3)}$$

where $\rho_{ij} = \text{Cov}(T(X_i), T(X_j))$ and $\sigma_j^2 = \text{Var}(T(X_j))$. We can obtain a similar variance ($\sigma^2_{\bar{T}^\pi}$) for $MSE_{RF^\pi}(\mathcal{T})$, so that under the hypothesis that $\mathbb{E}MSE_{RF}(\mathcal{T}) = \mathbb{E}MSE_{RF^\pi}(\mathcal{T})$, τ^2 can be seen to be

$$\tau^2 = \sigma^2_{\bar{T}} + \sigma^2_{\bar{T}^\pi}.$$

That the $\bar{T}_i$ and $\bar{T}_i^\pi$ are two indenpendently iid sequences follows from Lemma 2. Independence of the two samples follows from a similar argument to the second remark after Corollary 2. Crucially, there are many complicated quantities in this Equation B.1.3, i.e. $\sigma_j^2, \sigma_{\pi,j}^2, \rho_{ij}, \rho_{ij}^\pi$, for which there are not obvious estimators available and thus this result alone is not clearly practical. In the following sections, we verify the validity of our proposed permutation procedure, which avoids the necessary explicit estimation of these quantities.

Next, we move on to the proof of Proposition 1, which gives that the trees typically trained in a random forest obey the necessary regularity conditions for Corollary 2.

Proposition 1. *Assume that $Y = m(X) + \epsilon$, where $m(\cdot)$ is continuous on the unit cube. Let $\mathcal{X} = [0,1]^p$, and assume that $X_{i,j} \overset{iid}{\sim} Unif(0,1)$ for $i = 1, ..., n$ and $j = 1, ..., p$. Then, let $T_n(X)$ be a tree trained on iid pairs $(X_1, Y_1), ..., (X_n, Y_n)$ such that each leaf of the tree contains a single observation. Further, assume the trees satisfy the following two conditions:*

(i) *$\exists \gamma > 0$ such that $P($variable j is split on$) > \gamma$ for $j \in \{1, ..., p\}$*

(ii) *Each split leaves at least γn observations in each node.*

Then, for each $X \in \mathcal{X}$

$$T_n(X) \xrightarrow{d} Y|X = X \text{ as } n \to \infty$$

Proof. Each tree divides $\mathcal{X}$ into a partition of rectangular subspaces, corresponding to leaves of the tree. Following Meinshausen [2006], for each point X (with coordinates $[x_1, ..., x_p]$), let $\ell(X)$ denote the unique leaf of the tree containing X. Let $R_\ell(X)$ be the rectangular

subspace of $[0,1]^p$ corresponding to a particular leaf $\ell(X)$. The rectangular nature of the subspaces means that for each input feature, R_ℓ can be expressed as

$$R_\ell(X) = \bigotimes_{i=1}^{p} [a(X,i), b(X,i)]$$

where $0 \leq a(X,i) \leq x_i \leq b(X,i) \leq 1$ are scalars inducing an interval in dimension i. Then, the tree (by the existence of the requisite γ) satisfies the conditions of Lemma 2 in Meinshausen [2006], so that $\max_i |a(X,i) - b(X,i)| \xrightarrow{p} 0$. Let $\boldsymbol{a}(X) = [a(X,1), ..., a(X,p)]$ and similarly define $\boldsymbol{b}(X)$, so that the previous sentence implies: $\boldsymbol{a}(X) \xrightarrow{p} \boldsymbol{b}(X)$. We therefore also see that $a(X,i), b(X,i) \xrightarrow{p} x_i$ for all i.

The trees are fully grown, so the tree prediction at the point X is given by

$$T_n(X) = \sum_{k=1}^{n} I(X_k \in R_\ell(X)) Y_k$$

i.e. the response for the observation whose leaf contains X. As such, let k^* be the index corresponding to the observation who shares a leaf with X, so that $T_n(X) = Y_{k^*}$. We can deconstruct the event $X_{k^*} \in R_\ell(X)$ as

$$\{X_{k^*} \in R_\ell(X)\} = \bigcap_{i=1}^{p} \{a(X,i) \leq X_{i,k^*} \leq b(X,i)\}.$$

Thus, in the limit, $a(X,i), b(X,i) \xrightarrow{p} X_{i,k^*}$, and so $X_{i,k^*} \xrightarrow{p} x_i$ for all i. Further, continuity of m yields that $m(X_{k^*}) \xrightarrow{p} m(X)$. Thus, we see that, in the limit

$$Y_{k^*} = m(X) + \epsilon_{k^*} \stackrel{d}{=} m(X) + \epsilon \stackrel{d}{=} Y|X = X$$

because ϵ_{k^*} is independent of the location of X. □

Next, we provide more details about the permutation distribution derived in subsection 3.3.4. Recall that the calculation of the permutation distribution variance follows immediately from Equation 3.3.6; the permutation distribution of the statistic

$$\sqrt{2B}\left[MSE_{RF}(X;y) - MSE_{RF^\pi}(X;y)\right]$$

converges to a normal distribution with mean 0 and variance

$$\tau^2 = \frac{1}{1/4}\left[\frac{1}{2}\text{Var}(g'(\mathbb{E}RF_B(X))T(X)) + \frac{1}{2}\text{Var}(g'(\mathbb{E}RF_B^\pi(X))T^\pi(X))\right].$$

This is double the variance of Equation 3.3.3, because the previous calculations were done for a $\sqrt{B}$ rescaling, and so the conditional and unconditional variances agree. Because the ensemble sizes used in Algorithm 1 are assumed to be the same, $p = \frac{1}{2}$, so that the permutation test for equivalence of forest predictions is automatically valid in the sense of matching the permutation and unconditional distributions. This argument is formalized in the following result.

Theorem 6. *Let $T_{1,k_n}, ..., T_{B,k_n}$ and $T^\pi_{1,k_n}, ..., T^\pi_{B,k_n}$ be two collections of trees satisfying the conditions of Lemmas 1 and 3, and fix a test point with location X and response Y. Consider a test of the null hypothesis*

$$H_0 : \mathbb{E}\left[MSE_{RF}(X;Y)\middle|\ X, Y\right] = \mathbb{E}\left[MSE_{RF^\pi}(X;Y)\middle|\ X, Y\right]$$

using the statistic $\hat{\Delta} = MSE_{RF}(X;Y) - MSE_{RF^\pi}(X;Y)$. Then under H_0, the permutation distribution of $\sqrt{B}\hat{\Delta}$ converges to a normal distribution with mean 0 and variance

$$\tau^2 = g'(\mathbb{E}RF_B(X))^2\sigma^2 + g'(\mathbb{E}RF_B^\pi(X))^2\sigma_\pi^2$$

which is also the variance of the unconditional distribution of $\sqrt{B}\hat{\Delta}$, as $n \to \infty$. Thus, the permutation test attains the asymptotic Type I error rate.

Proof. The only claim that remains to be verified is that the permutation test attains the Type I error rate. Let $\Phi(\cdot)$ be the standard normal cdf, and let $\hat{J}_B(t)$ be the (random) cdf of the permutation distribution, with corresponding quantile function $\hat{J}_B^{-1}(q)$. By the argument preceding the theorem statement, we have that $\sup_t |\hat{J}_B(t) - \Phi(t/\tau)| \xrightarrow{p} 0$. Then, by Lemma 11.2.1 of Lehmann and Romano [2006], for any number $q \in (0,1)$, $\hat{J}_B^{-1}(q) \xrightarrow{p} \tau\Phi^{-1}(q)$. In particular, for a given significance level α, the 1-sided permutation test of H_0 at the level α has a critical value $\hat{J}_B^{-1}(1-\alpha)$ which converges in probability to $\tau\Phi^{-1}(1-\alpha)$. Thus, as $B \to \infty$,

$$P(\sqrt{B}\hat{\Delta} \geq \hat{J}_B^{-1}(1-\alpha)|H_0) \to P(\sqrt{B}\hat{\Delta} \geq \tau\Phi^{-1}(1-\alpha)|H_0) \to \alpha.$$

□

B.2 Additional Simulations

We include some additional simulations here to demonstrate the following points.

1. The accuracy of the permutation distribution approximation of the Gaussian. The theory outlined in Section 3.3 establishes that the difference in MSEs between forests is asymptotically Gaussian but the difficulty in estimating the resulting variance largely restricts its direct usage in practical settings. We go on to demonstrate that the permutation distribution approaches this distribution, thereby circumventing the need for a direct variance estimate. The simulations below present empirical evidence that this approximation is reasonable in practical settings.
2. The instability of the variance estimation procedures laid out in Wager et al. [2014b] and Mentch and Hooker [2016a]. Clearly variance estimation is useful for developing confidence intervals about random forest predictions, which in the case of pointwise consistency (as in the honest trees proposed by Wager and Athey [2018]), are also valid for the underlying regression function. However, in the hypothesis testing framework, these estimates are useful only insofar as they allow for calculation of a test statistic. These variance estimates, such as the infinitesmal jackknife of Wager et al. [2014b], recommend

building $B = \mathcal{O}(n^{\beta})$ trees where $\beta \geq 1$. We demonstrate that this recommendation cannot be violated.

3. The robustness (and potential weaknesses) of the proposed procedure to different random forest implementations. In particular, we want to study the effect of larger subsamples/more trees. The theoretical results presented in Section 3.3 rely on treating the tree predictions as iid. Clearly, this is never true in practice, and some theoretical justification for the effects of this being small were presented in Section 3.5.

B.2.1 Variance Estimation Instability

Here, we use the infinitesmal jackknife (IJ), as implemented in the `ranger` package [Wright and Ziegler, 2015], to estimate the variance of a random forest prediction at a given point. In particular, we simulate data from Model 2 from Table 3.4.1, train a subsampled random forest, and record the IJ variance estimate of random forest prediction at $X_1 = ... = X_5 = 0.5$ and $X_6 = ... = X_{10} = 2$. We use $n = 2000$, $k_n = n^{0.5} \approx 44$, and vary the number of trees. Often times, the IJ variance estimate is negative, leading to a `NaN` output from the IJ software. These instances represent a case when the IJ estimate is useless to a practitioner, and as such, we report the percentage of times that a `NaN` output is returned for each number of trees. For each number of trees, we repeat the simulation 100 times, and results are shown in Figure B.2.1.

The IJ estimate provides overwhelmingly large variance estimates for small numbers of trees, leading to overly conservative confidence intervals and tests with exceptionally low power. Moreover, the ribbon remains quite wide until around $B = 2000$ trees, suggesting that at least $\mathcal{O}(n)$ trees are necessary to attain a stable variance estimate. A similar number of trees is necessary to ensure that a `NaN` is never returned. We should note that this is the simplest possible case of variance estimation, i.e. the estimation is only at a single point. The problem grows exponentially more complex as more test points are considered and covariance estimates are needed. Mentch and Hooker [2016a] note that the procedure is infeasible for more than 20-30 test points. The authors demonstrate in follow-up work [Mentch and Hooker, 2017] that an approximate test can be produced by utilizing random

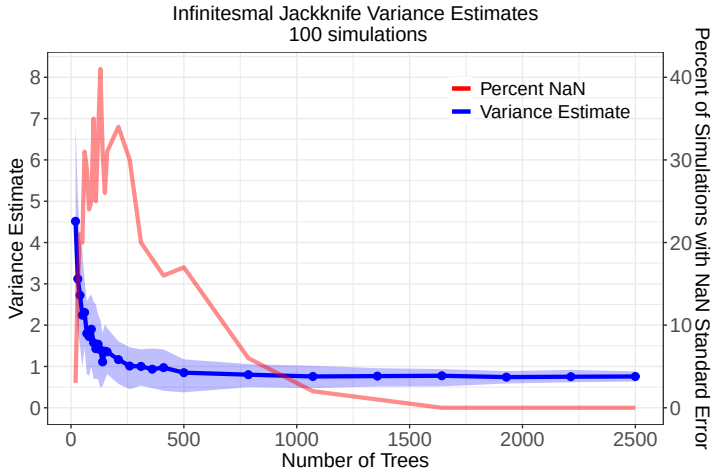


Figure B.2.1: `ranger` IJ variance estimate. Blue ribbon plot indicates central 90% of variance estimates (corresponds to left axis), and red line (corresponds to right axis) represents percentage of runs that return `NaN`.

projections which allows for slightly larger test sets but at the cost added computational strain. In contrast, besides the minimal overhead required to form the additional predictions, the testing procedure proposed here is almost entirely immune to the number of points in the test set. Once the initial predictions are formed, the only remaining work is to shuffle predictions (trees) and re-compute the difference in MSE between forests.

B.2.2 Test Robustness

We now present more figures similar to the power curves presented in Section 3.4. The goal here is to present the proposed procedure's robustness to the number of trees B and the subsample size k_n. To do so, we modify the simulation study plotted in the second panel of Figure 3.4.1. Here, we fix the error variance at $\sigma^2(\epsilon) = 16$, and again simulate $n = 2000$ training observations and $N_t = 100$ test observations. First, we vary the number of trees built, according to

$$B \in \{20, 50, 75, 125, 250, 375, 500, 750, 1000\}$$

and let $k_n = n^{0.6}$. The resulting simulations are plotted in Figure B.2.2.

Two clear patterns are clear in the figure - the power and type I error rate of the test

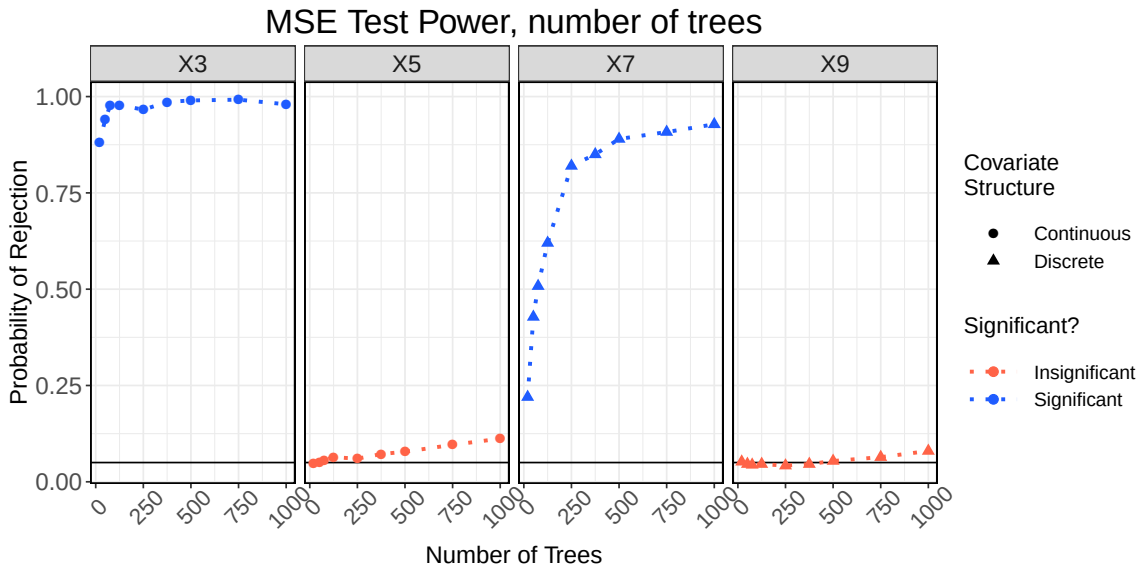


Figure B.2.2: Model 2 power curves for 500 simulations, by number of trees. The Y-axis represents $P(\tilde{p} \leq \alpha)$ where $\alpha = 0.05$ and is shown as the horizontal line across the bottom of the plots.

both increase as the number of trees grows. However, the rate of growth for each of them is markedly different - the test attains high power around $B \approx 250$ trees, but deviations from the nominal level are only noticeable around $B \approx 500$ trees. Even when $B = 1000$, the observed level is still within nearly 5% of the baseline. Thus, while the level of the test may be slightly inflated for large numbers of trees, the procedure remains valid for limited, but realistic tree sizes.

Recall that the subsample size is a key limiting factor of Lemma 1 - it is required that $k_n = o(\sqrt{n})$ - to establish asymptotic normality. Other work [Wager and Athey, 2018] weakens these conditions, but places explicit restrictions on the types of trees allowed in the ensemble. We now examine the behavior of our procedure under larger sample sizes. We use the same simulation parameters as in Figure B.2.2, but now fix $B = 125$ and let $k_n = n^p$, and we vary p at 10 equally spaced values between 0.1 and 0.99.

The resulting simulation is shown in Figure B.2.3. We see that for $p \leq 0.75$, the Type I error rate is maintained, but for larger subsamples, we begin to see a severe deviation. Though severe, this is not necessarily surprising as such large subsampling rates correspond

directly to a more severe violation of the iid approximation.

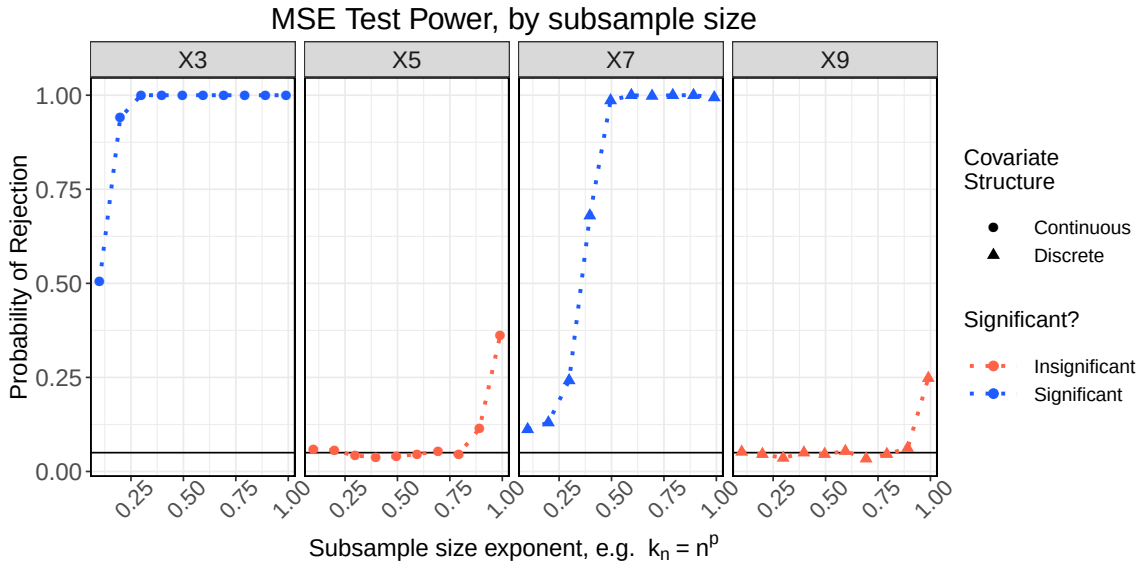


Figure B.2.3: Model 2 power curves for 500 simulations, by subsample exponent. The Y-axis represents $P(\tilde{p} \leq \alpha)$ where $\alpha = 0.05$ and is shown as the horizontal line across the bottom of the plots.

Appendix C Chapter 4 Appendix

C.1 Proof of Proposition 3

Here, we prove Proposition 3, which is restated below followed by its proof.

Proposition 3. *Let* $\{Z_i\}_{i=1}^N \overset{iid}{\sim} Bernoulli(\alpha)$*, and let* $(x_i, Y_i)_{i=1}^{n+m} | Z_i \overset{iid}{\sim} Z_i P_2 + (1 - Z_i) P_1$*, where* P_1 *and* P_2 *satisfy the covariate shift assumption (Equation 1 in the main text). Define* $m = \sum_{i=1}^N Z_i$*. Assume that* $Y_i \geq 0$ *almost surely,* $\sup_x \mathbb{E}(Y^4 | x = x) < K$ *for some constant* K*, and that*

$$\rho_n^* = \max_{k=1,2} \max_{i \neq j} Cor_{P_k} \left[(m_{B_i}(x_i) - Y_i)^2, (m_{B_j}(x_j) - Y_j)^2 \right] \to 0$$

as $n \to \infty$*. Further, assume that for all* $x \in \mathcal{X}$*,* $w_N(x)$ *is consistently proportional to the likelihood ratio,* $\ell(x) = \frac{dP_2^*(x)}{dP_1^*(x)}$*, i.e.* w_N *satisfies*

$$w_N(x) = c \frac{dP_2^*(x)}{dP_1^*(x)} + \epsilon_N(x) \ \forall \ x \in \mathcal{X}$$

where c *is a constant that does not depend on* x*, and* $\epsilon_N(x)$ *is a sequence of random variables satisfying* $P(\sup_x |\epsilon_N(x)| < \eta_N) = 1$*, where* $\eta_N \to 0$ *as* $N \to \infty$*. Let* $\theta_{P_2} = \mathbb{E}_{P_2}(\lim_{B \to \infty} OOB_{m,B})$*. Then, as* $B, n \to \infty$

$$OOB_{m,B}^{\boldsymbol{w}} \xrightarrow{p} \theta_{P_2}.$$

Proof. To show this, we use a standard trick in the importance sampling literature to rewrite $\text{OOB}_{m,B}^{\boldsymbol{w}}$ as

$$\text{OOB}_{m,B}^{\boldsymbol{w}} = \frac{\sum_{i=1}^n w_i (m_{B_i}(x_i) - Y_i)^2}{\sum_{j=1}^n w_j} = \frac{\frac{1}{n} \sum_{i=1}^n w_i (m_{B_i}(x_i) - Y_i)^2}{\frac{1}{n} \sum_{j=1}^n w_j} \ . \tag{C.1.1}$$

An important point of clarification is that we use N to be the total sample size, n to be the size of the training set, and m be the size of the test set. Because $n \sim \text{Binomial}(N, \alpha)$, $\lim_{N \to \infty} n = \infty$ (and similarly for m) almost surely. Thus, we use $n \to \infty$, $m \to \infty$,

and $N \to \infty$ interchangeably. The weak law of large numbers gives that as $n \to \infty$, the denominator of Equation C.1.1 obeys

$$\frac{1}{n}\sum_{j=1}^{n} w_j \xrightarrow{p} c\mathbb{E}_{x\sim P_1^*}\left[\frac{dP_2^*(x)}{dP_1^*(x)}\right] = c\int_{\mathcal{X}} dP_2^*(x) = c.$$

By assumption, $w_i = c\frac{dP_2^*(x_i)}{dP_1^*(x_i)} + \epsilon_N(x_i)$, so that the numerator of Equation C.1.1 can be expressed as

$$\frac{1}{n}\sum_{i=1}^{n}\left[c\frac{dP_2^*(x_i)}{dP_1^*(x_i)} + \epsilon_N(x_i)\right]\left(\frac{1}{B_i}\sum_{k=1}^{B_i} T_{\boldsymbol{w}}(x_i;\xi_k) - Y_i\right)^2.$$

Now, we want to show that this converges in probability to $c\theta_{P_2}$. We do this by analyzing the variance of the numerator of Equation C.1.1. Note that we have

$$\begin{aligned}
&\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n}\left(c\frac{dP_2^*(x_i)}{dP_1^*(x_i)} + \epsilon_N(x_i)\right)\left(\frac{1}{B_i}\sum_{k=1}^{B_i} T_{\boldsymbol{w}}(x_i;\xi_k) - Y_i\right)^2\right] \\
&= \mathrm{Var}\Bigg[\underbrace{c\frac{1}{n}\sum_{i=1}^{n}\left(\frac{dP_2^*(x_i)}{dP_1^*(x_i)}\right)\left(\frac{1}{B_i}\sum_{k=1}^{B_i} T_{\boldsymbol{w}}(x_i;\xi_k) - Y_i\right)^2}_{S_{1,n}} \\
&\qquad\qquad + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\epsilon_N(x_i)\left(\frac{1}{B_i}\sum_{k=1}^{B_i} T_{\boldsymbol{w}}(x_i;\xi_k) - Y_i\right)^2}_{S_{2,N}}\Bigg].
\end{aligned}$$

We approximate $\mathrm{Var}(S_{1,n} + S_{2,N})$ as $\mathrm{Var}(S_{1,n}) + \mathrm{Var}(S_{2,N})$, because $\mathrm{Cov}(S_{1,n}, S_{2,N}) \to 0$ as $N \to \infty$. To see this last fact, note that $S_{2,N}$ satisfies

$$|S_{2,N}| < \frac{\eta_N}{n}\sum_{i=1}^{n}\left(\frac{1}{B_i}\sum_{k=1}^{B_i} T_{\boldsymbol{w}}(x_i;\xi_k) - Y_i\right)^2 \tag{C.1.2}$$

and that the quantity on the right hand side is integrable, so that by dominated convergence, $\mathbb{E}(S_{2,N}) \to 0$. Moreover, by assumption, the squared out-of-bag residuals are bounded in probability (because they are assumed to have finite mean/variance). Thus, the cross-term can be controlled as

$$\begin{aligned}
&\mathbb{E}\left[S_{2,N} \times \frac{c}{n}\sum_{i=1}^{n}\left(\frac{dP_2^*(x_i)}{dP_1^*(x_i)}\right)\left(\frac{1}{B_i}\sum_{k=1}^{B_i} T_{\boldsymbol{w}}(x_i;\xi_k) - Y_i\right)^2\right] \\
&\quad < \mathbb{E}\left[\frac{\eta_N}{n}\sum_{i=1}^{n}\left(\frac{1}{B_i}\sum_{k=1}^{B_i} T_{\boldsymbol{w}}(x_i;\xi_k) - Y_i\right)^2 \times \frac{c}{n}\sum_{i=1}^{n}\left(\frac{dP_2^*(x_i)}{dP_1^*(x_i)}\right)\left(\frac{1}{B_i}\sum_{k=1}^{B_i} T_{\boldsymbol{w}}(x_i;\xi_k) - Y_i\right)^2\right]
\end{aligned}$$

which, again by dominated convergence, converges to 0.

Now, we want to show that the variance of $S_{2,N}$ vanishes. Recall that by hypothesis, $P(\lim_{N\to\infty} S_{2,N} = 0) = 1$, and so it follows that $P(\lim_{N\to\infty} S_{2,N}^2 = 0) = 1$. Then, again we can appeal to dominated convergence (using the quantity in Equation C.1.2 squared as our upper bound) to get that $\mathrm{Var}(S_{2,N}) \to 1$ as $N \to \infty$. All that remains to show is that $\mathrm{Var}(S_{1,n}) \to 0$ as $n \to \infty$. The variance of $S_{1,n}$ can be expressed as

$$\begin{aligned}
\mathrm{Var}(S_{1,n}) &= \mathrm{Var}\left[\frac{c}{n}\sum_{i=1}^{n}\left(\frac{dP_2^*(x_i)}{dP_1^*(x_i)}\right)(m_{B_i}(x_i) - Y_i)^2\right] \\
&= \frac{c^2}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left[\left(\frac{dP_2^*(x_i)}{dP_1^*(x_i)}\right)(m_{B_i}(x_i) - Y_i)^2\right] \\
&+ \frac{2c^2}{n^2}\sum_{1\le i<j\le n}\mathrm{Cov}\left[\left(\frac{dP_2^*(x_i)}{dP_1^*(x_i)}\right)(m_{B_i}(x_i) - Y_i)^2, \left(\frac{dP_2^*(x_j)}{dP_1^*(x_j)}\right)(m_{B_j}(x_j) - Y_j)^2\right].
\end{aligned}$$

Because Y_i is almost surely positive, and $m_{B_i}(\cdot)$ is an average of positive random variables, both are positive almost surely. Also, note that the likelihood ratio term is also positive, so that the whole quantity $\left(\frac{dP_2^*(x_i)}{dP_1^*(x_i)}\right)(m_{B_i}(x_i) - Y_i)^2 > 0$ almost surely. Then, we make use the fact that for positive random variables W, Z,

$$\begin{aligned}
\mathrm{Var}_{W,Z\sim P}\big[(W-Z)^2\big] \le \mathbb{E}_{W,Z\sim P}\big[(W-Z)^4\big] &= \mathbb{E}_{W,Z\sim Q}\left[\frac{dP(W,Z)}{dQ(W,Z)}(W-Z)^4\right]] \\
&\le \max\big(\mathbb{E}_P(W^4), \mathbb{E}_P(Z^4)\big).
\end{aligned}$$

Note that in the above, we use $\mathbb{E}_P(W^4)$ to indicate integration over the marginal distribution of W under joint distribution P. Because m_{B_i} is a weighted sum of random variables with bounded 4th moments, it also has a bounded 4th moment. Letting $\kappa = \max\{\max_i \mathbb{E}_{P_1}(m_{B_i}(x_i)^4), K\}$, we see that

$$\begin{aligned}
\mathrm{Var}(S_{1,n}) &\le \frac{c^2 n\kappa}{n^2} + \frac{2c^2}{n^2}\sum_{1\le i<j\le n}\mathrm{Cov}\left[\left(\frac{dP_2^*(x_i)}{dP_1^*(x_i)}\right)(m_{B_i}(x_i) - Y_i)^2, \left(\frac{dP_2^*(x_j)}{dP_1^*(x_j)}\right)(m_{B_j}(x_j) - Y_j)^2\right] \\
&\le \frac{\kappa c^2}{n} + \frac{2c^2}{n^2}n^2\kappa\rho_n^* \\
&= \frac{\kappa c^2}{n} + 2\kappa c^2\rho_n^*.
\end{aligned}$$

The above goes to 0 by hypothesis, and noting that $\mathbb{E}S_{1,n} = c\theta_{P_2}$, we can apply Chebyshev's inequality to conclude that

$$\frac{1}{n}\sum_{i=1}^{n} w_i \left(\frac{1}{B_i} \sum_{k=1}^{B_i} T_{\boldsymbol{w}}(x_i; \xi_k) - Y_i \right)^2 \xrightarrow{p} c\theta_{P_2} \text{ as } N \to \infty.$$

Finally, Slutsky's Lemma gives that $\text{OOB}^{\boldsymbol{w}}_{m,B} \xrightarrow{p} \theta_{P_2}$ as $N, B \to \infty$. □

C.2 Detailed Simulation Results

The purpose of this section of the appendix is to provide specific results for the simulation from the high dimensional simulation from Chapter 4 in the form of tables. For each model described in the high dimensional simulation section of the main text, we provide the full results for each λ value. We also provide plots similar to those from the main text for the MAE and Interval Width statistics, for completeness in Figure C.2.1.

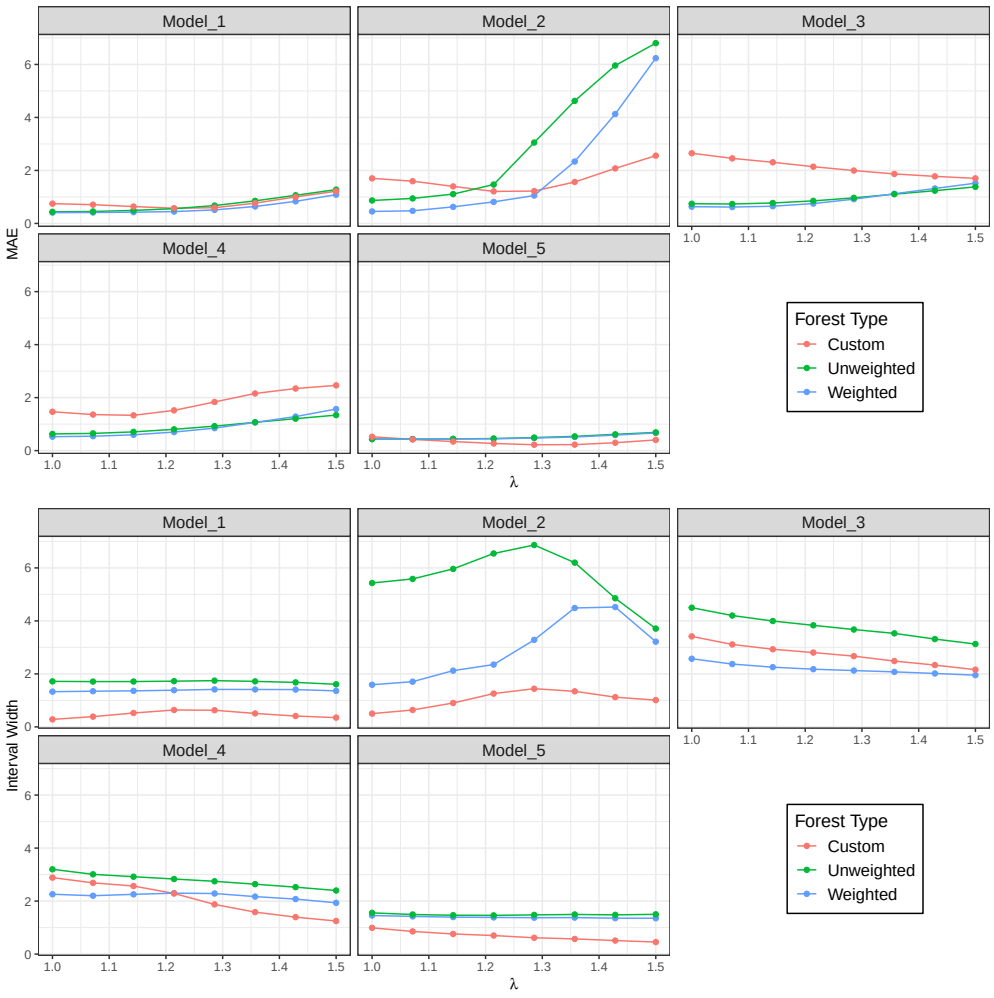


Figure C.2.1: Results from the high dimensional simulation for MAE (top) and Interval Width (bottom)

lambda	Model	RMSE	MAE	Covg	Interval Width	Score
1.000	Weighted	0.514	0.409	0.803	1.329	6.619
1.000	Unweighted	0.560	0.437	0.871	1.717	6.217
1.000	Custom	0.986	0.747	0.087	0.283	1.652
1.071	Weighted	0.518	0.413	0.805	1.345	6.562
1.071	Unweighted	0.576	0.451	0.864	1.710	6.058
1.071	Custom	0.910	0.709	0.126	0.386	1.826
1.143	Weighted	0.534	0.425	0.797	1.360	6.361
1.143	Unweighted	0.631	0.493	0.832	1.711	5.521
1.143	Custom	0.807	0.638	0.213	0.524	2.515
1.214	Weighted	0.564	0.448	0.783	1.384	6.026
1.214	Unweighted	0.708	0.557	0.784	1.725	4.840
1.214	Custom	0.720	0.574	0.342	0.637	3.609
1.286	Weighted	0.643	0.511	0.734	1.414	5.211
1.286	Unweighted	0.838	0.675	0.694	1.745	3.852
1.286	Custom	0.737	0.594	0.374	0.629	3.932
1.357	Weighted	0.787	0.639	0.623	1.412	3.968
1.357	Unweighted	1.004	0.852	0.561	1.717	2.818
1.357	Custom	0.897	0.766	0.210	0.507	2.344
1.429	Weighted	0.975	0.835	0.467	1.408	2.670
1.429	Unweighted	1.198	1.063	0.400	1.679	1.854
1.429	Custom	1.098	1.001	0.062	0.410	0.747
1.500	Weighted	1.214	1.087	0.283	1.360	1.489
1.500	Unweighted	1.404	1.277	0.246	1.607	1.090
1.500	Custom	1.287	1.216	0.010	0.350	0.121

Table C.2.1: Simulation results for Model 1.

lambda	Model	RMSE	MAE	Covg	Interval Width	Score
1.000	Weighted	0.604	0.453	0.826	1.590	5.884
1.000	Unweighted	1.520	0.867	0.896	5.430	2.547
1.000	Custom	2.087	1.701	0.094	0.500	0.947
1.071	Weighted	0.653	0.476	0.834	1.706	5.579
1.071	Unweighted	1.636	0.945	0.890	5.584	2.375
1.071	Custom	1.934	1.594	0.129	0.638	1.063
1.143	Weighted	1.077	0.626	0.831	2.123	4.181
1.143	Unweighted	1.843	1.107	0.875	5.962	2.071
1.143	Custom	1.698	1.398	0.217	0.902	1.381
1.214	Weighted	1.465	0.813	0.809	2.352	3.311
1.214	Unweighted	2.210	1.468	0.859	6.544	1.681
1.214	Custom	1.459	1.209	0.346	1.258	1.804
1.286	Weighted	1.596	1.054	0.796	3.284	2.536
1.286	Unweighted	3.533	3.053	0.708	6.864	0.946
1.286	Custom	1.416	1.222	0.367	1.441	1.751
1.357	Weighted	2.734	2.338	0.606	4.488	1.162
1.357	Unweighted	4.756	4.625	0.461	6.199	0.548
1.357	Custom	1.718	1.565	0.222	1.343	1.021
1.429	Weighted	4.319	4.129	0.337	4.523	0.511
1.429	Unweighted	5.987	5.960	0.209	4.854	0.266
1.429	Custom	2.185	2.077	0.060	1.122	0.281
1.500	Weighted	6.351	6.236	0.089	3.215	0.148
1.500	Unweighted	6.865	6.803	0.079	3.709	0.119
1.500	Custom	2.601	2.558	0.011	1.010	0.051

Table C.2.2: Simulation results for Model 2.

lambda	Model	RMSE	MAE	Covg	Interval Width	Score
1.000	Weighted	0.776	0.633	0.889	2.572	4.386
1.000	Unweighted	1.047	0.744	0.949	4.499	3.372
1.000	Custom	3.226	2.650	0.401	3.415	0.830
1.071	Weighted	0.765	0.619	0.871	2.372	4.473
1.071	Unweighted	1.026	0.735	0.940	4.202	3.442
1.071	Custom	2.989	2.456	0.390	3.111	0.881
1.143	Weighted	0.803	0.650	0.833	2.253	4.231
1.143	Unweighted	1.050	0.772	0.925	3.995	3.351
1.143	Custom	2.814	2.308	0.391	2.929	0.939
1.214	Weighted	0.934	0.751	0.754	2.177	3.570
1.214	Unweighted	1.130	0.852	0.898	3.833	3.105
1.214	Custom	2.613	2.142	0.399	2.805	1.010
1.286	Weighted	1.119	0.916	0.637	2.126	2.750
1.286	Unweighted	1.230	0.965	0.851	3.674	2.786
1.286	Custom	2.442	1.997	0.401	2.670	1.076
1.357	Weighted	1.331	1.122	0.499	2.076	1.990
1.357	Unweighted	1.351	1.105	0.785	3.529	2.437
1.357	Custom	2.285	1.868	0.394	2.484	1.133
1.429	Weighted	1.525	1.320	0.380	2.016	1.439
1.429	Unweighted	1.467	1.236	0.699	3.315	2.103
1.429	Custom	2.168	1.777	0.382	2.333	1.167
1.500	Weighted	1.719	1.519	0.268	1.952	0.982
1.500	Unweighted	1.603	1.381	0.597	3.127	1.752
1.500	Custom	2.079	1.701	0.363	2.159	1.183

Table C.2.3: Simulation results for Model 3.

lambda	Model	RMSE	MAE	Covg	Interval Width	Score
1.000	Weighted	0.650	0.527	0.902	2.258	5.234
1.000	Unweighted	0.829	0.630	0.927	3.199	4.180
1.000	Custom	1.799	1.467	0.571	2.885	1.671
1.071	Weighted	0.670	0.544	0.892	2.202	5.117
1.071	Unweighted	0.842	0.650	0.915	3.010	4.135
1.071	Custom	1.647	1.361	0.589	2.688	1.860
1.143	Weighted	0.740	0.598	0.867	2.254	4.644
1.143	Unweighted	0.905	0.708	0.890	2.920	3.862
1.143	Custom	1.591	1.332	0.598	2.568	1.962
1.214	Weighted	0.870	0.701	0.812	2.298	3.928
1.214	Unweighted	1.011	0.804	0.837	2.833	3.407
1.214	Custom	1.757	1.519	0.486	2.288	1.615
1.286	Weighted	1.037	0.848	0.730	2.286	3.208
1.286	Unweighted	1.126	0.924	0.771	2.746	2.957
1.286	Custom	2.028	1.836	0.285	1.874	1.004
1.357	Weighted	1.280	1.067	0.585	2.169	2.367
1.357	Unweighted	1.271	1.069	0.673	2.638	2.436
1.357	Custom	2.296	2.157	0.135	1.584	0.504
1.429	Weighted	1.485	1.286	0.459	2.075	1.768
1.429	Unweighted	1.398	1.206	0.575	2.526	2.008
1.429	Custom	2.458	2.344	0.061	1.393	0.244
1.500	Weighted	1.747	1.569	0.287	1.934	1.064
1.500	Unweighted	1.515	1.340	0.467	2.398	1.603
1.500	Custom	2.537	2.463	0.026	1.248	0.109

Table C.2.4: Simulation results for Model 4.

lambda	Model	RMSE	MAE	Covg	Interval Width	Score
1.000	Weighted	0.543	0.434	0.813	1.454	6.247
1.000	Unweighted	0.547	0.437	0.838	1.555	6.241
1.000	Custom	0.668	0.523	0.493	0.990	4.096
1.071	Weighted	0.544	0.435	0.808	1.422	6.255
1.071	Unweighted	0.545	0.436	0.826	1.493	6.263
1.071	Custom	0.548	0.423	0.527	0.854	5.232
1.143	Weighted	0.550	0.441	0.797	1.394	6.173
1.143	Unweighted	0.551	0.443	0.815	1.466	6.176
1.143	Custom	0.447	0.343	0.572	0.759	6.705
1.214	Weighted	0.558	0.446	0.786	1.383	6.068
1.214	Unweighted	0.568	0.457	0.798	1.461	5.954
1.214	Custom	0.357	0.270	0.659	0.700	9.094
1.286	Weighted	0.594	0.476	0.755	1.373	5.654
1.286	Unweighted	0.607	0.492	0.776	1.478	5.527
1.286	Custom	0.291	0.223	0.727	0.615	11.804
1.357	Weighted	0.639	0.514	0.712	1.376	5.117
1.357	Unweighted	0.661	0.538	0.731	1.495	4.940
1.357	Custom	0.285	0.226	0.712	0.572	12.054
1.429	Weighted	0.720	0.587	0.632	1.354	4.287
1.429	Unweighted	0.743	0.615	0.653	1.480	4.140
1.429	Custom	0.349	0.298	0.510	0.510	8.173
1.500	Weighted	0.803	0.666	0.558	1.353	3.578
1.500	Unweighted	0.820	0.688	0.596	1.499	3.561
1.500	Custom	0.446	0.404	0.266	0.454	3.935

Table C.2.5: Simulation results for Model 5.

Bibliography

David J Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198. Springer, 1985.

André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

Ashwin N Ananthakrishnan, Millie D Long, Christopher F Martin, Robert S Sandler, and Michael D Kappelman. Sleep disturbance and risk of active disease in patients with crohn's disease and ulcerative colitis. *Clinical Gastroenterology and Hepatology*, 11(8):965–971, 2013.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *arXiv preprint arXiv:1610.01271*, 2016.

Susan Athey, Julie Tibshirani, Stefan Wager, et al. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

Rina Foygel Barber, Emmanuel J Candès, et al. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction under covariate shift. *arXiv preprint arXiv:1904.06019*, 2019.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

Christoph Bergmeir and José M. Benítez. Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. *Journal of Statistical Software*, 46(7):1–26, 2012. URL `http://www.jstatsoft.org/v46/i07/`.

GÃŠrard Biau. Analysis of a random forests model. *Journal of Machine Learning Research*, 13(Apr):1063–1095, 2012.

L Breiman, JH Friedman, R Olshen, and CJ Stone. Classification and regression trees. 1984.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001a.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001b.

Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection. *arXiv preprint arXiv:1610.02351*, 2016.

John P. Cangialosi, Andrew S. Latto, and Robbie Berg. Hurricane irma. In *National Hurricane Center Tropical Cyclone Report*. 2018. URL `https://www.nhc.noaa.gov/data/tcr/AL112017_Irma.pdf`.

M Cassotti, D Ballabio, R Todeschini, and V Consonni. A similarity-based qsar model for predicting acute toxicity towards the fathead minnow (pimephales promelas). *SAR and QSAR in Environmental Research*, 26(3):217–243, 2015.

Senniang Chen. *Imputation of missing values using quantile regression*. PhD thesis, Iowa State University, 2014.

Yuan Shih Chow and Henry Teicher. *Probability theory: independence, interchangeability, martingales*. Springer Science & Business Media, 2012.

EunYi Chung and Joseph P Romano. Exact and asymptotically robust permutation tests. *The Annals of Statistics*, pages 484–507, 2013.

Tim Coleman, Lucas Mentch, Daniel Fink, Frank La Sorte, Giles Hooker, Wesley Hochachka, and David Winkler. Statistical inference on tree swallow migrations. *arXiv preprint arXiv:1710.09793*, 2017.

Tim Coleman, Kimberly Kaufeld, Mary Frances Dorn, and Lucas Mentch. Locally optimized random forests. *arXiv preprint arXiv:1908.09967*, 2019a.

Tim Coleman, Wei Peng, and Lucas Mentch. Scalable and efficient hypothesis testing with random forests. *arXiv preprint arXiv:1904.07830*, 2019b.

Paulo Cortez and Aníbal de Jesus Raimundo Morais. A data mining approach to predict forest fires using meteorological data. 2007.

Yifan Cui, Ruoqing Zhu, Mai Zhou, and Michael Kosorok. Some asymptotic results of survival tree and forest models. *arXiv preprint arXiv:1707.09631*, 2017.

Bruno De Finetti. La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68, 1937.

Duane R Diefenbach, Matthew R Marshall, Jennifer A Mattice, and Daniel W Brauning. Incorporating availability for detection in estimates of bird abundance. *The Auk*, 124(1): 96–106, 2007.

Peter O Dunn and David W Winkler. Climate change has affected the breeding date of tree swallows throughout North America. *Proceedings of the Royal Society of London B: Biological Sciences*, 266(1437):2487–2490, 1999.

Bradley Efron. Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109(507):991–1007, 2014.

Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems. *Journal of Machine Learning Research*, 15(1):3133–3181, 2014a.

Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems. *Journal of Machine Learning Research*, 15(1):3133–3181, 2014b.

Daniel Fink, Wesley M Hochachka, Benjamin Zuckerberg, David W Winkler, Ben Shaby, M Arthur Munson, Giles Hooker, Mirek Riedewald, Daniel Sheldon, and Steve Kelling. Spatiotemporal exploratory models for broad-scale survey data. *Ecological Applications*, 20(8):2131–2147, 2010.

Daniel Fink, Tom Auer, Viviana Ruiz-Gutierrez, Wesley M Hochachka, Alison Johnston, Frank A La Sorte, and Steve Kelling. Modeling avian full annual cycle distribution and population trends with citizen science data. *bioRxiv*, page 251868, 2018.

Ronald Aylmer Fisher. *The design of experiments*. Oliver And Boyd; Edinburgh; London, 1937.

Mark A Friedl, Damien Sulla-Menashe, Bin Tan, Annemarie Schneider, Navin Ramankutty, Adam Sibley, and Xiaoman Huang. MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote sensing of Environment*, 114(1): 168–182, 2010.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:, 2001.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

Jerome H Friedman. Multivariate adaptive regression splines. *The annals of statistics*, pages 1–67, 1991.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Seth D Guikema and Steven M Quiring. Hybrid data mining-regression for infrastructure risk assessment based on zero-inflated data. *Reliability Engineering & System Safety*, 99: 178–182, 2012.

Peter Hall and Ingrid Van Keilegom. Two-sample tests in functional data analysis starting from discrete data. *Statistica Sinica*, pages 1511–1531, 2007.

MC Hansen, RS DeFries, John RG Townshend, and Rob Sohlberg. Global land cover classification at 1 km spatial resolution using a classification tree approach. *International Journal of Remote Sensing*, 21(6-7):1331–1364, 2000.

Trevor Hastie. gam: Generalized additive models. 2017. URL `https://CRAN.R-project.org/package=gam`. R package version 1.14-4.

Jichao He, David W Wanik, Brian M Hartman, Emmanouil N Anagnostou, Marina Astitha, and Maria EB Frediani. Nonparametric tree-based predictive modeling of storm outages on an electric distribution network. *Risk Analysis*, 37(3):441–458, 2017.

Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(8):832–844, August 1998. ISSN 0162-8828. doi: 10.1109/34.709601.

Yosef Hochberg and Yoav Benjamini. More powerful procedures for multiple significance testing. *Statistics in medicine*, 9(7):811–818, 1990.

Wassily Hoeffding. The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, pages 169–192, 1952.

Giles Hooker. Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732, 2007.

Giles Hooker and Lucas Mentch. Please stop permuting features: An explanation and alternatives. *arXiv preprint arXiv:1905.03151*, 2019.

Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674, 2006.

David JT Hussell. Climate change, spring temperatures, and timing of breeding of tree swallows *(Tachycineta bicolor)* in southern Ontario. *The Auk*, 120(3):607–618, 2003.

Hemant Ishwaran and Min Lu. Random survival forests. *Wiley StatsRef: Statistics Reference Online*, pages 1–13, 2008.

Hemant Ishwaran and Min Lu. Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in medicine*, 38 (4):558–582, 2019.

Silke Janitza, Ender Celik, and Anne-Laure Boulesteix. A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification*, pages 1–31, 2016.

Arnold Janssen. Resampling student'st-type statistics. *Annals of the Institute of Statistical Mathematics*, 57(3):507–529, 2005.

Alison Johnston, Daniel Fink, Mark D. Reynolds, Wesley M. Hochachka, Brian L. Sullivan, Nicholas E. Bruns, Eric Hallstein, Matt S. Marrifield, Sandi Matsumoto, and Steve Kelling.

Abundance models improve spatial and temporal prioritization of conservation resources. *Ecological Applications*, 25(7):1749–1756, 2015. doi: 10.1890/14-1826.1.

Patricia D Jones, Michael D Kappelman, Christopher F Martin, Wenli Chen, Robert S Sandler, and Millie D Long. Exercise decreases risk of future active disease in patients with inflammatory bowel disease in remission. *Inflammatory bowel diseases*, 21(5):1063–1071, 2015.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(Jul):1391–1445, 2009.

Michael Klass and Henry Teicher. The central limit theorem for exchangeable random variables without moments. *The Annals of Probability*, pages 138–153, 1987.

Charles D Koven, Bruno Ringeval, Pierre Friedlingstein, Philippe Ciais, Patricia Cadule, Dmitry Khvorostyanov, Gerhard Krinner, and Charles Tarnocai. Permafrost carbon-climate feedbacks accelerate global warming. *Proceedings of the National Academy of Sciences*, 108(36):14769–14774, 2011.

Max Kuhn. caret: Classification and regression training. 2017. URL `https://CRAN.R-project.org/package=caret`. R package version 6.0-78.

Frank A La Sorte, Wesley M Hochachka, Andrew Farnsworth, André A Dhondt, and Daniel Sheldon. The implications of mid-latitude climate extremes for North American migratory bird populations. *Ecosphere*, 7(3), 2016.

C. W. Landsea and J. L. Franklin. Atlantic Hurricane Database Uncertainty and Presentation of a New Database Format. *Monthly Weather Review*, 141:3576–3592, October 2013. doi: 10.1175/MWR-D-12-00254.1.

David M Lawrence and Andrew G Slater. Incorporating organic soil into a global climate model. *Climate Dynamics*, 30(2-3):145–160, 2008.

David M Lawrence, Charles D Koven, S Cl Swenson, William J Riley, and AG Slater. Permafrost thaw and resulting soil moisture changes regulate projected high-latitude co2 and ch4 emissions. *Environmental Research Letters*, 10(9):094011, 2015.

Eric L Lehmann, Charles Stein, et al. On the theory of some non-parametric hypotheses. *The Annals of Mathematical Statistics*, 20(1):28–45, 1949.

Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

Jing Lei. Cross-validation with confidence. *Journal of the American Statistical Association*, pages 1–20, 2019.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2 (3):18–22, 2002. URL `http://CRAN.R-project.org/doc/Rnews/`.

Haibin Liu, Rachel A Davidson, David V Rosowsky, and Jery R Stedinger. Negative binomial regression of electric power outages in hurricanes. *Journal of infrastructure systems*, 11 (4):258–267, 2005.

Keli Liu and Xiao-Li Meng. There is individualized treatment. why not individualized inference? *Annual Review of Statistics and Its Application*, 3:79–111, 2016.

Daniel Malinsky and Peter Spirtes. Causal structure learning from multivariate time series in settings with unmeasured confounding. In *Proceedings of 2018 ACM SIGKDD Workshop on Causal Discovery*, pages 23–47, 2018.

K McGarigal, SA Cushman, and E Ene. Fragstats v4: spatial pattern analysis program for categorical and continuous maps. university of massachusetts, amherst, massachusetts, usa. *goo.gl/aAEbMk*, 2012.

Gerald A Meehl, Warren M Washington, Julie M Arblaster, Aixue Hu, Haiyan Teng, Claudia Tebaldi, Benjamin N Sanderson, Jean-Francois Lamarque, Andrew Conley, Warren G Strand, et al. Climate system response to external forcings and climate change projections in ccsm4. *Journal of Climate*, 25(11):3661–3683, 2012.

Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7 (Jun):983–999, 2006.

Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

Lucas Mentch and Giles Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research*, 17(1):841–881, 2016a.

Lucas Mentch and Giles Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research*, 17(1):841–881, 2016b.

Lucas Mentch and Giles Hooker. Formal hypothesis tests for additive structure in random forests. *Journal of Computational and Graphical Statistics*, pages 1–9, 2017.

Patrick AP Moran. The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):243–251, 1948.

Georg Neuhaus. Conditional rank tests for the two-sample problem under random censorship. *The Annals of Statistics*, pages 1760–1779, 1993.

Kristin K Nicodemus, Joseph H Callicott, Rachel G Higier, Augustin Luna, Devon C Nixon, Barbara K Lipska, Radhakrishna Vakkalanka, Ina Giegling, Dan Rujescu, David St Clair, et al. Evidence of statistical epistasis between DISC1, CIT and NDEL1 impacting risk for schizophrenia: biological validation with functional neuroimaging. *Human genetics*, 127 (4):441–452, 2010.

Gary W Oehlert. A note on the delta method. *The American Statistician*, 46(1):27–29, 1992.

Donatella Pasqualini, Kimberly Kaufeld, and Mary Frances Dorn. Electric power outage forecasting model. Technical report, Los Alamos National Laboratory, 11 2017.

Wei Peng, Tim Coleman, and Lucas Mentch. Asymptotic distributions and rates of convergence for random forests and other resampled ensemble learners. *arXiv preprint arXiv:1905.10651*, 2019.

Fortunato Pesarin and Luigi Salmaso. *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons, 2010.

Belinda Phipson and Gordon K Smyth. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, 9(1), 2010.

Scott Powers, Trevor Hastie, Robert Tibshirani, et al. Customized training with an application to mass spectrometric imaging of cancer tissue. *The Annals of Applied Statistics*, 9 (4):1709–1725, 2015.

Sashank Jakkam Reddi, Barnabas Poczos, and Alex Smola. Doubly robust covariate shift correction. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Orin J Robinson, Viviana Ruiz-Gutierrez, and Daniel Fink. Correcting for bias in distribution modelling for rare species using citizen science data. *Diversity and Distributions*, 24:460–472, 2018.

Juan José Rodriguez, Ludmila I Kuncheva, and Carlos J Alonso. Rotation forest: A new classifier ensemble method. *IEEE transactions on pattern analysis and machine intelligence*, 28(10):1619–1630, 2006.

Joseph P Romano. On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, 85(411):686–692, 1990.

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

John R Sauer, Jane E Fallon, and Rex Johnson. Use of North American Breeding Bird Survey data to estimate population change for bird conservation regions. *The Journal of wildlife management*, pages 372–389, 2003.

Erwan Scornet, Gérard Biau, Jean-Philippe Vert, et al. Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741, 2015a.

Erwan Scornet, Gérard Biau, Jean-Philippe Vert, et al. Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741, 2015b.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(Mar):371–421, 2008.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2011.

Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25, 2007a.

Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25, 2007b.

Masashi Sugiyama and Klaus-Robert Müller. Input-dependent estimation of generalization error under covariate shift. *Statistics & Decisions*, 23(4/2005):249–279, 2005.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert MÃžller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8 (May):985–1005, 2007.

Brian L Sullivan, Christopher L Wood, Marshall J Iliff, Rick E Bonney, Daniel Fink, and Steve Kelling. ebird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142(10):2282–2292, 2009a.

Brian L Sullivan, Christopher L Wood, Marshall J Iliff, Rick E Bonney, Daniel Fink, and Steve Kelling. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142(10):2282–2292, 2009b.

Brian L Sullivan, Jocelyn L Aycrigg, Jessie H Barry, Rick E Bonney, Nicholas Bruns, Caren B Cooper, Theo Damoulas, André A Dhondt, Tom Dietterich, Andrew Farnsworth, et al. The ebird enterprise: an integrated approach to development and application of citizen science. *Biological Conservation*, 169:31–40, 2014a.

Brian L Sullivan, Jocelyn L Aycrigg, Jessie H Barry, Rick E Bonney, Nicholas Bruns, Caren B Cooper, Theo Damoulas, André A Dhondt, Tom Dietterich, Andrew Farnsworth, et al. The eBird enterprise: an integrated approach to development and application of citizen science. *Biological Conservation*, 169:31–40, 2014b.

Terry Therneau, Beth Atkinson, and Brian Ripley. rpart: Recursive partitioning and regression trees. 2017. URL `https://CRAN.R-project.org/package=rpart`. R package version 4.1-11.

Terry M Therneau, Elizabeth J Atkinson, et al. An introduction to recursive partitioning using the rpart routines, 1997.

M.M. Thormton, P.E. Thornton, Y. Wei, R.S. Vose, and A.G. Boyer. Daymet: Station-Level Inputs and Model Predicted Values for North America, Version 3, 2017. URL `https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1391`.

Surya T Tokdar and Robert E Kass. Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60, 2010.

Laura Toloşi and Thomas Lengauer. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, 27(14):1986–1994, 2011a.

Laura Toloşi and Thomas Lengauer. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, 27(14):1986–1994, 2011b.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Stefan Wager, Trevor Hastie, and Bradley Efron. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research*, 15(1):1625–1651, 2014a.

Stefan Wager, Trevor Hastie, and Bradley Efron. Confidence intervals for random forests: the jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15 (1):1625–1651, 2014b.

DW Wanik, EN Anagnostou, BM Hartman, MEB Frediani, and M Astitha. Storm outage modeling for an electric distribution network in northeastern usa. *Natural Hazards*, 79(2): 1359–1384, 2015.

Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.

Hugh E Willoughby, EN Rappaport, and FD Marks. Hurricane forecasting: The state of the art. *Natural Hazards Review*, 8(3):45–49, 2007.

David W Winkler, Miles K Luo, and Eldar Rakhimberdiev. Temperature effects on food supply and chick mortality in tree swallows *(Tachycineta bicolor)*. *Oecologia*, 173(1):129–138, 2013.

Marvin N Wright and Andreas Ziegler. Ranger: a fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*, 2015.

Ruo Xu, Dan Nettleton, and Daniel J Nordman. Case-specific random forests. *Journal of Computational and Graphical Statistics*, 25(1):49–65, 2016.

Ruoqing Zhu, Donglin Zeng, and Michael R Kosorok. Reinforcement learning trees. *Journal of the American Statistical Association*, 110(512):1770–1784, 2015.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

Benjamin Zuckerberg, Daniel Fink, Frank A La Sorte, Wesley M Hochachka, and Steve Kelling. Novel seasonal land cover associations for eastern north american forest birds identified through dynamic species distribution modelling. *Diversity and Distributions*, 22 (6):717–730, 2016.