Proceedings of the of the 17th Conference on Computer Science and Intelligence Systems pp. 489–497
DOI: 10.15439/2022F265
ISSN 2300-5963 ACSIS, Vol. 30

A short note on post-hoc testing using random forests algorithm: Principles, asymptotic time complexity analysis, and beyond

Lubomír Štěpánek
Department of Statistics and Probability
Faculty of Informatics and Statistics
University of Economics
nám. W. Churchilla 4, 130 67 Prague, Czech Republic
lubomir.stepanek@vse.cz
&
Institute of Biophysics and Informatics
First Faculty of Medicine
Charles University
Salmovská 1, Prague, Czech Republic
lubomir.stepanek@lf1.cuni.cz

Filip Habarta, Ivana Malá, Luboš Marek
Department of Statistics and Probability
Faculty of Informatics and Statistics
University of Economics
nám. W. Churchilla 4, 130 67 Prague
Czech Republic
{filip.habarta, malai, marek}@vse.cz

***Abstract*—When testing whether a continuous variable differs between categories of a factor variable or their combinations, taking into account other continuous covariates, one may use an analysis of covariance. Several post-hoc methods, such as Tukey's honestly significant difference test, Scheffé's, Dunn's, or Nemenyi's test are well-established when the analysis of covariance rejects the hypothesis there is no difference between any categories. However, these methods are statistically rigid and usually require meeting statistical assumptions. In this work, we address the issue using a random forest-based algorithm, practically assumption-free, classifying individual observations into the factor's categories using the dependent continuous variable and covariates on input. The higher the proportion of trees classifying the observations into two different categories is, the more likely a statistical difference between the categories is. To adjust the method's first-type error rate, we change random forest trees' complexity by pruning to modify the proportions of highly complex trees. Besides simulations that demonstrate a relationship between the tree pruning level, tree complexity, and first-type error rate, we analyze the asymptotic time complexity of the proposed random forest-based method compared to established techniques.**

I. Introduction

COMPARING a continuous variable's means of two or more categories (or their combinations) of one or more factor variables and detecting significant mutual differences, if any, is very common in applied statistics. Particularly when the dependent variable needs to be adjusted by other continuous covariates, an analysis of covariance (ANCOVA) is a tool of choice [1].

Since the analysis of covariance tests whether there is, in general, a difference between at least two categories of a given factor, the big question is to determine where exactly the statistical difference is, i. e., which two (or more) exact categories of the factor are those the significant difference arises from.

For this reason, post-hoc tests are usually applied to identify the significantly different categories of their combinations. Some of them are quite established, for instance, Tukey honestly significant difference (HSD) test [2], Scheffé's test [3], Dunn's test [4], or, if needed, Nemenyi's test with a reduced amount of assumptions required to be met [5].

However, the covariance analysis and the post-hoc tests are limited by relatively tough statistical assumptions, usually in terms of normality of independence of observation subsamples. Furthermore, empirically, when there are multiple methods for one task, that usually implies each method is limited somehow and, consequently, there is no "apriori first choice" method routinely working in all situations.

This work introduces a new post-hoc method based on a random forest algorithm to overcome the mentioned. Each classification tree the random forest model consists of has got its complexity, i. e. number of leaf nodes, by which it can classify into only one or more categories of observations than only one. The continuous dependent variable and continuous covariates are the variables by which an entry sample of observations is split into subsamples, using logical formulas with the variables and searched cut-offs. Considering the categories given by the factor variable (or more factor variables) entering the analysis of covariance, these may be refined as an output for the random forest algorithm, not only as an input for the analysis of covariance. If the number of trees in the random forest model with sufficient complexity, i. e. classifying into two or more factor categories or their combinations, is high enough, then the hypothesis that there is no statistical difference between the two categories is hardly

likely. As a tuning parameter, the pruning level may affect how complex the trees in a random forest are.

After the well-established methods revisiting, we describe principles behind the random forest-based algorithm for post-hoc testing, derive the asymptotic time complexity of the proposed method, and estimate a feasible number of trees in the random forest model regarding the number of other factors and continuous covariates. Eventually, we do simulations to compare the new method to others, i. e. established ones, and, particularly, describe a relationship between the random forest tree pruning level and tree complexity and the model's first-type error rate.

II. Principles and Assumptions of Analysis of Covariance and Post-Hoc Tests Revisited

In this section, we recapitulate basic principles of analysis of covariance and commonly used post-hoc tests to refresh their logic and mention their assumptions and limitations.

A. Analysis of covariance (ANCOVA) – principles, assumptions, and limitations

Principles of ANCOVA. Analysis of covariance is a linear model standing in between analysis of variance (ANOVA) [6] and linear regression [1], [7]. While the analysis of variance assumes there is a continuous dependent variable and independent categorical factors, linear regression allows for independent covariates as the analysis of covariance. However, compared to the linear regression, it estimates effect sizes as excesses above or below covariate variable average and enables to elegantly estimate an explained variability proportion of the continuous dependent variable by covariates. Furthermore, analysis of variance is performed particularly when continuous covariates are not of much interest compared to the factors. That being said, inference tests for coefficients of the covariates are usually skipped.

A model of the analysis of covariance, including $k \in \mathbb{N}$ categorical factor variables and $m \in \mathbb{N}$ continuous covariates, is for i-th observation from $n \in \mathbb{N}$ observations in total, as follows,

$$y_i = \mu + \sum_{j=1}^{k} \delta_j + \sum_{l=1}^{m} \beta_l x_{i,l} + \varepsilon_i, \tag{1}$$

where y_i is a value of the dependent continuous variable for i-th observation, μ is a grand total mean of the dependent variable, δ_j is an effect of j-th factor on i-th observation, with $\forall j \in \{1, 2, \ldots, k\}$, β_l is a coefficient (slope) of l-th covariate, with with $\forall l \in \{1, 2, \ldots, m\}$, $x_{i,l}$ is a value of l-th covariate for i-th observation, and ε_i is a residual term of i-th observation, respectively.

Firstly, coefficients β_l for $\forall l \in \{1, 2, \ldots, m\}$, listed in a vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)^T$, are estimated as those minimizing the sum of residuals [8] from formula (1), ignoring (not yet estimated) effects δ_j of the factor variables, thus,

$$\begin{aligned}\boldsymbol{\beta} = \arg\min_{\beta_l \in \mathbb{R}^m} \left\{ \sum_{i=1}^{n} \varepsilon_i^2 \right\} \Bigg|_{\forall j \in \{1,2,\ldots,k\}: \delta_j = 0} = \\ = \arg\min_{\beta_l \in \mathbb{R}^m} \left\{ \sum_{i=1}^{n} \left(y_i - (\mu + \sum_{l=1}^{m} \beta_l x_l) \right)^2 \right\} = \\ = \arg\min_{\beta_l \in \mathbb{R}^m} \left\{ \sum_{i=1}^{n} \left(y_i - \mu - \sum_{l=1}^{m} \beta_l x_l \right)^2 \right\}. \end{aligned} \tag{2}$$

So far, considering formula (2) the analysis of covariance is similar to the linear regression. Secondly, once the vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)^T$ of linear coefficients is estimated, a part close to multifactorial analysis of variance follows. A total sum of squares, $\mathrm{SS}_{\text{tot}} = \sum_{i=1}^{n} (y_i - \bar{y})^2$, describing total variability of the dependent variable [9], is corrected (reduced) by variability explained by the continuous covariates, SS_{β_l}, getting $\mathrm{SS}^*_{\text{tot}}$, so

$$\begin{aligned}\mathrm{SS}^*_{\text{tot}} = \mathrm{SS}_{\text{tot}} - \mathrm{SS}_{\beta_l} = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{l=1}^{m} \frac{\mathrm{cov}(\boldsymbol{y}, \boldsymbol{x}_l)^2}{\mathrm{var}(\boldsymbol{x}_l)} = \\ = \sum_{i=1}^{n} (y_i - \bar{y}^*)^2, \end{aligned} \tag{3}$$

considering a vector of the dependent variable $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$, a vector of l-th covariate $\boldsymbol{x}_l = (x_{1,l}, x_{2,l}, \ldots, x_{n,l})^T$, and grand mean $\bar{y}^*$ adjusted by the correction.

Finally, k-way analysis of variance for the coefficients δ_j estimation is applied. The null hypothesis claiming that j-th factor does not affect the dependent continuous variable, i. e. $H_0 : \delta_j = 0$, is formulated k-times and tests using the adjusted sum of squares, $\mathrm{SS}^*_{\text{tot}}$, decomposed into component for the factors and for the residuals. Using formula (3) and assuming j-th factor has got n_j categories and c-th category has got $n_{j,c}$ observations, the decomposition is as follows [9],

$$\begin{aligned}\mathrm{SS}^*_{\text{tot}} = \sum_{i=1}^{n} (y_i - \bar{y}^*)^2 = \sum_{j=1}^{k} \sum_{c=1}^{n_j} \sum_{i=1}^{n_{j,c}} (y_i - \bar{y}^*)^2 = \\ = \sum_{j=1}^{k} \sum_{c=1}^{n_j} \sum_{i=1}^{n_{j,c}} (y_i - \bar{y}_{j,c} + \bar{y}_{j,c} - \bar{y}^*)^2 = \\ = \sum_{j=1}^{k} \sum_{c=1}^{n_j} \sum_{i=1}^{n_{j,c}} (y_i - \bar{y}_{j,c})^2 + \sum_{j=1}^{k} \sum_{c=1}^{n_j} (\bar{y}_{j,c} - \bar{y}^*)^2 = \\ = \sum_{j=1}^{k} \mathrm{SS}_{j\text{-th factor}} + \mathrm{SS}_{\boldsymbol{\varepsilon}} \end{aligned} \tag{4}$$

where $\mathrm{SS}_{j\text{-th factor}}$ is sum of squares for j-th factor, $\mathrm{SS}_{\boldsymbol{\varepsilon}}$ is sum of squares for residuals, and $\bar{y}_{j,c}$ is an average of all values that belong to c-th category of j-th factor. Consequently, using

formula (4), the null hypothesis $H_0 : \delta_j = 0$ for j-th factor is rejected on confidence level $1 - \alpha$ if and only if

$$F = \frac{\mathrm{SS}_{j\text{-th factor}}/(n_j - 1)}{\mathrm{SS}_{\boldsymbol{\varepsilon}}/(n + k - 1 - \sum_{j=1}^{k} n_j)} \geq \\ \geq F_{1-\alpha}\left(n_j - 1, n + k - 1 - \sum\nolimits_{j=1}^{k} n_j\right), \quad (5)$$

where $F_{1-\alpha}(df_1, df_2)$ is $(1-\alpha)$-th quantile of Fisher-Snedecor distribution with df_1 and df_2 degrees of freedom, respectively. Rejecting the null hypothesis for j-th factor does not determine which categories of the factor mutually differ significantly, though.

Assumptions and limitations of ANCOVA. Analysis of covariance assumes that residuals are independent, i. e. for each $r, s \in \{1, 2, \ldots, n\}$ so that $r \neq s$ is $\mathrm{cov}(\varepsilon_r, \varepsilon_s) = 0$, and of the same variance, i. e. for each $r \in \{1, 2, \ldots, n\}$ is $\varepsilon_r = \sigma^2 < 0$. Moreover, the residuals should be normally distributed, i. e. for each $r \in \{1, 2, \ldots, n\}$ is $\varepsilon_r \sim \mathcal{N}(0, \sigma^2)$ [1].

B. Post-hoc tests – principles, assumptions, and limitations

Assuming the null hypothesis has been rejected for j-th factor, one would like to determine which exact two or more categories of the factor significantly differ. Let us mark average values of observations that belong to categories c_r and c_s of j-th factor, with $r, s \in \{1, 2, \ldots, n_j\}$, as μ_r and μ_s. Usually, the categories c_r and c_s of j-th factor significantly differ when some inequality using data parameters or estimates holds, as showed below applying the mathematical notation from formulas (4) and (5). The decision process may be repeated for each pair of categories r, s of j-th factor to research all possible differences.

1) Tukey honestly significant differences (HSD) test: Based on Tukey, averages of the categories c_r and c_s significantly differ if

$$\frac{|\mu_r - \mu_s|}{\hat{\sigma}\sqrt{2/n}} \geq q(\alpha, k, n - k), \quad (6)$$

where $q(\alpha, k, n-k)$ is studentized critical value for confidence level α and σ^2 is residuals' variance, n is sample size and k is the number of factors.

Tukey HSD method assumes that subsamples for compared categories are independent, of the same variability (*homoskedasticity*), and follow normal distribution [2].

2) Scheffé's test: Following Scheffé's (unweighted) test, averages of the categories c_r and c_s significantly differ if

$$|\mu_r - \mu_s| \geq \sqrt{(k-1) \cdot F_{1-\alpha}\left(df_1, df_2\right) \cdot \mathrm{SS}_{j\text{-th factor}}}, \quad (7)$$

where $df_1 = n_j - 1$ and $df_2 = n + k - 1 - \sum_{j=1}^{k} n_j$.

Scheffé's test is less limited than Tukey's HSD test since there is no explicit assumption of any normal distribution of observations; however, it has lower statistical power, though [3].

3) Dunn's test: Transforming values of dependent variable y_i that belong to the categories c_r and c_s from initial continuous ones to their ranks, we get their averages $\bar{w}_r$ and $\bar{w}_s$. Dunn's test recommends considering the categories c_r and c_s as different when

$$|\bar{w}_r - \bar{w}_s| \geq z_{1-\alpha/2} \cdot \sqrt{\frac{\frac{n(n+1)}{12} + \sum_{t \in |\bar{w}_r - \bar{w}_s|}\left(n_t^3 - \frac{n_t}{12(n-1)}\right)}{\left(\frac{1}{n_{c_r}} + \frac{1}{n_{c_s}}\right)}}, \quad (8)$$

where t is a possible tied value of the ranks w_r and w_s, n_t is a count of tied ranks at value t, and n_{c_r} and n_{c_s} are numbers of observations in categories c_r and c_s, respectively.

While assumption-free, Dunn's test may fail to identify significant differences between categories due to its low statistical power [4].

4) Nemenyi's test: Similarly to Dunn's test, assuming average ranks $\bar{w}_r$ and $\bar{w}_s$ of the categories c_r and c_s, these significantly differ if

$$|\bar{w}_r - \bar{w}_s| \geq q(\alpha, k, n-k) \cdot \sqrt{\frac{n(n+1)}{24} \cdot \left(\frac{1}{n_{c_r}} + \frac{1}{n_{c_s}}\right)}, \quad (9)$$

where $q(\alpha, k, n-k)$ is studentized critical value for confidence level α, n is sample size, k is the number of factors, and n_{c_r} and n_{c_s} are numbers of observations in categories c_r and c_s, respectively.

Nemeyi's test is nonparametric and robust enough, but may suffer from low statistical power, though [5].

III. Principles and Assumptions of the Random Forests

In advance of the proposed method introduction we shortly point out important pieces of knowledge about classification trees and random forests.

A. Classification trees – principles and assumptions

Classification trees from the CART family of trees (classification and regression trees) split a hyperspace of $k \in \mathbb{N}$ explanatory variables (continuous or categorical) into disjunctive hyper-rectangles, fitting simple (constant) models there by minimizing a given criterion [10].

An observation given by a vector of values $\boldsymbol{x_i} = (x_{i,1}, x_{i,2}, \ldots, x_{i,k})^T$ is classified into one of m classes of a target variable by a set of rules that comes from node formulas, created throughout the tree is growing, as described in Fig. 1. Initially, the root covers all observations till a node rule, i. e. a found explanatory variable and a cut-off value minimizing the given criterion partitions the dataset into two parts. Each part is then again split by a new rule set for a child node. The process is recursively repeated by growing the tree, by which a set of node rules successively splits the input dataset into more parts that are mutually more and more different. The process of the tree growing, called a top-down induction of a decision tree (TDIDT), is stopped by a stopping criterion, e. g. maximum of leaf (ending) nodes, maximum tree deepness level, etc.

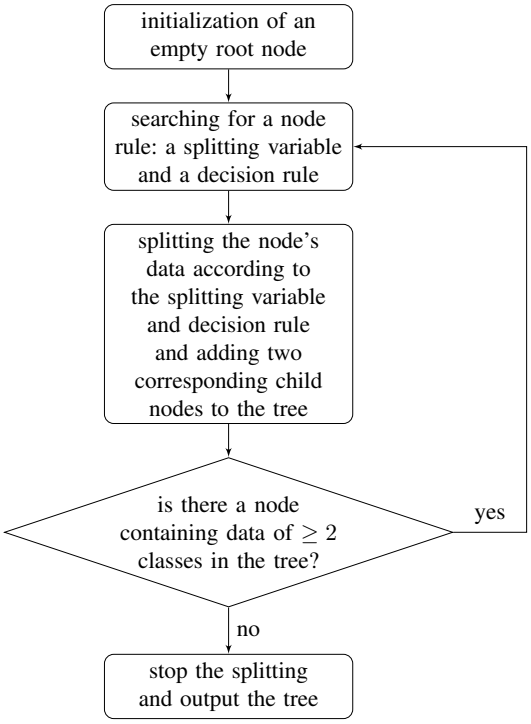

Fig. 1. A top-down induction of a decision tree (TDIDT).

Let $\sigma(\bullet)_j$ be a proportion of all observations that belong – by rules of all nodes from root to leaf one – to a target class j. A leaf node n_t classifies into the class c_f^* if $c_f^* = \operatorname{argmax}_{f\in\{1,2,\ldots,n_j\}}\{\sigma(\bullet)_f\}$. Since each node is through the tree growing a leaf one (for a limited time), the criterion has to be minimized in searching for node n_t rule. There are several commonly used criteria, also called *impurity measure*, $Q_{n_t}(T)$, such as misclassification error (10), Gini index (11), or deviance (cross-entropy) (12),

$$Q_{n_t}(T) = 1 - \sigma(\bullet)_f, \tag{10}$$

$$Q_{n_t}(T) = \sum_{j=1}^{m} \sigma(\bullet)_f(1 - \sigma(\bullet)_f), \tag{11}$$

$$Q_{n_t}(T) = -\sum_{j=1}^{m} \sigma(\bullet)_f \cdot \log \sigma(\bullet)_f. \tag{12}$$

One can easily see that the lower the impurity measure is, the higher $\sigma(\bullet)_f$, i. e. a proportion of a target class f in the node n_t, has to be, as expected.

Classification trees, as depicted in Fig. 1, in order to minimize the leaf nodes impurity, tend to overfit the node rules on a given dataset, which is done by the tree's typical "overgrowing", i. e. high complexity. To avoid this, besides some other naive approaches, *pruning* is commonly applied. Firstly, let us use usually defined *cost–complexity function*,

$$C_\kappa(T) = \sum_{n_t\in\{\boldsymbol{n}_t\}} |\{\boldsymbol{x}_{n_t}\}| \cdot Q_{n_t}(T) + \kappa \cdot |\{\boldsymbol{n}_t\}|, \tag{13}$$

where $\{\boldsymbol{n}_t\}$ is a set of leaf nodes of the tree and $\{\boldsymbol{x}_{n_t}\}$ is a set of all observations constrained by rules coming from the root till the node n_t. The idea of the pruning is to find a subtree T_κ so that $T_\kappa \subset T$ for a given κ that minimizes the statistics $C_\kappa(T)$, i. e. $T_\kappa = \operatorname{argmin}_T \left\{\sum_{n_t\in\{\boldsymbol{n}_t\}} |\{\boldsymbol{x}_{n_t}\}| \cdot Q_{n_t}(T) + \kappa \cdot |\{\boldsymbol{n}_t\}|\right\}$.

The $\kappa \geq 0$ is a tuning parameter that governs the trade-off between a high tree complexity and size (for low values of κ) and tree parsinomity and reproducibility to other datasets (for large values of κ).

B. Principles of the random forests

Random forests are finite sets of (distinct) classification trees, described in detail above, each classifying a k-dimensional observation, $\boldsymbol{x_i} = (x_{i,1}, x_{i,2}, \ldots, x_{i,k})^T$, into one of $m \in \mathbb{N}$ target classes [11]. The final class $c_f^* \in \{1, 2, \ldots, n_j\}$ of a k-dimensional observation is the one the largest subset of the random forest's trees classifies it into[1], i. e., $\overset{*}{f} = \operatorname{argmax}_{f\in\{1,2,\ldots,n_j\}}\{\#\text{ of trees classifying into the class } c_f\}$.

What is worth to be mentioned is that only $k^* < k$ variables are considered as possible partitioning variables in node rules. The subset of k^* variables from the original k explanatory variables is selected randomly using bootstrapping; that ensures the pre-selected k^* variables are mutually independent enough. A flowchart of the random forest model building is in Fig. 2.

Neither classification trees nor random forests have important assumptions or limitations worth speaking off.

IV. The Proposed Method for Post-Hoc Testing

In this section, we introduce a novel alternative for post-hoc testing based on a random forest algorithm. Considering the ANCOVA notation, categories of a factor that contains statistically different effects on the continuous dependent variable are leaf node classes each tree of a random forest classifies into. The dependent variable and the covariates, and other factor variables, if any, are entry variables that serve for node rules if needed. Each tree of the random forest model can either classify only into one category (as a root note tree) or into two or more categories, based on its complexity (size). For details, see Fig. 3.

The more trees of sufficient complexity can classify into the classes (categories) in the forest, the more likely we can reject the null hypothesis that there is no difference between the effects of the factor's categories on the dependent variable. This is formally done by ANCOVA, too. What is more, if a proportion of trees classifying into two given categories is large enough, considering all trees, the given two categories seem to be of statistically different effect on the dependent variable [12].

A proportion of trees classifying into two or more categories to all trees in the random forest is close to a point estimate

[1]In case of a tie, i. e. there are two or more target classes the maximum forest's trees classify the observation into, one of them is picked randomly.

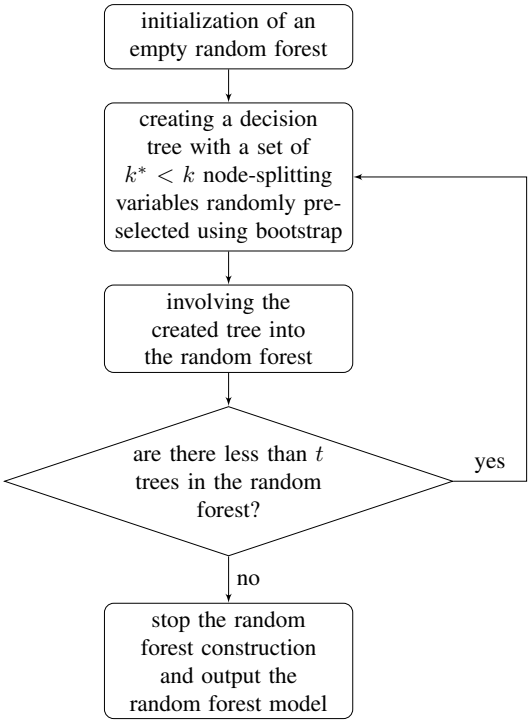

Fig. 2. A construction of the random forest model involving t decision trees.

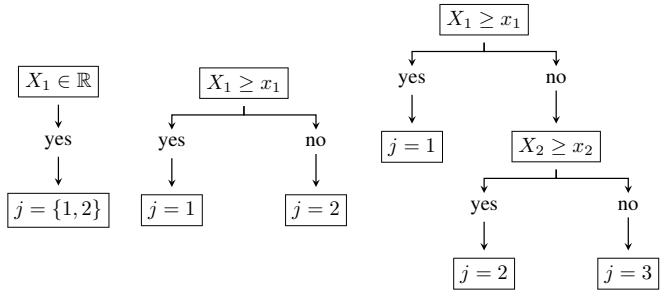

Fig. 3. An example of a root node tree (on the left) not able to classify into any class unambiguously, an example of a tree with sufficient complexity (in the middle) able to classify into two classes ($j = 1$ and $j = 2$), and an example of a tree with sufficient complexity (on the right) able to classify into three classes ($j = 1$, $j = 2$, and $j = 3$).

of the p-value. The p-value is the probability we incorrectly reject the null hypothesis of no different effects of the factor categories on the dependent variable, assuming the null hypothesis is true. Thus, the method also provides statistical inference as a post-hoc test. Since we could modify a random forest's tree complexity (size), i. e. also tendencies to classify either only into one or into two or more classes, by pruning and the tuning parameter κ, we may control the first-type error rate, i. e. the incorrect rejection of the null hypothesis when it is true, of the random forest model as inferential post-hoc test. The proposed method is due to the random forest algorithm behind almost assumption-free.

Besides the derivations of the inferential properties of the method, we also discuss the method's asymptotic time complexity and do a simulation study with varying κ tuning parameters to describe a relationship between the parameter and the method's first-type error rate.

A. Statistical inference behind the proposed method

Using the mathematical notation from previous sections, let us assume the ANCOVA already rejected the null hypothesis that the j-th factor does not affect the dependent variable. Thus, the question is what two (or more) categories of j-th factor are significantly different so that the factor influences the dependent variable[2].

Intuitively, when a large number of the (appropriately pruned) trees of the random forest model can classify into two given classes, i. e. categories, then one can hardly suppose the categories are statistically without a difference.

Similarly to the post-hoc tests, let the null hypothesis H_0 claim that there is no statistical difference between the given two categories c_r and c_s of j-th factor. The alternative hypothesis H_1 claims the contradiction, so

H_0 : *No statistical difference between categories* c_r *and* c_s.
H_1 : *Statistical difference between categories* c_r *and* c_s.

Whenever a post-hoc test rejects the null hypothesis H_0 in favor of the alternative hypothesis H_1, the case is equivalent to a situation the test's p-value is lower than or equal to a prior set significance level α, usually equal to 0.05.

By definition, the p-value is a probability of gaining data at least as extreme as the data actually observed, assuming the null hypothesis is true. Let t_c be a number of trees in the random forest model that are in contradiction to the null hypothesis (under the null hypothesis assumption). Then, the value of t_c is equal to the number of all trees classifying, besides other classes, into given two classes (categories) c_r and c_s; showing that there is a difference between the two classes. Let the $\mathcal{I}_{c_r,c_s}(\tau)$ be an identifier function returning 1 if and only if the tree τ classifies into the classes (categories) c_r and c_s (regardless whether it classifies into other classes, too), thus,

$$\mathcal{I}_{c_r,c_s}(\tau) \begin{cases} 1, & \text{tree } \tau \text{ classifies into categories } c_r \text{ and } c_s, \\ 0, & \text{otherwise.} \end{cases}$$

We can derive

$$t_c = \sum_{\forall \tau \in \text{ random forest}} \mathcal{I}_{c_r,c_s}(\tau),$$

[2]Supposing all categories of j-th factor are similar and mutually without significant differences, then the categorization of j-factor would not result in null hypothesis rejection about no effect of j-th factor on the dependent variable. That is a contradiction, so, there should be two or more different categories of j-th factor.

and assuming the random forest model contains exactly $t \in \mathbb{N}$ trees, and all trees are induced randomly regardless of their complexity[3], the p-value is estimated by $\hat{p}$ as

$$\begin{aligned}\hat{p} &= P(\text{getting data at least as extreme as the observed} \mid H_0) = \\ &= P\left(\sum_{\forall \tau \in \text{ random forest}} \mathcal{I}_{c_r,c_s}(\tau) \geq t_c \mid H_0\right) = \\ &= P\left(\sum_{\forall \tau \in \text{ random forest}} \mathcal{I}_{c_r,c_s}(\tau) \in \{t_c, t_c+1, \ldots, t\}\right) = \\ &= \frac{|\{t_c, t_c+1, \ldots, t\}|}{t} = \\ &= \frac{t - t_c + 1}{t} = \\ &= 1 - \frac{t_c - 1}{t}. \end{aligned} \tag{14}$$

Thus, formula (14) shows that the p-value's estimate is equal to the fraction of $1 - \frac{t_c-1}{t}$. Intuitively, supposing the initial number t_c of trees in the random forest model that are complex enough to classify, besides others, to categories c_r and c_s is generally low. In that case, such a model is not "much" in contradiction to the null hypothesis about no differences between the two categories. Thus, when the t_c is relatively low, the fraction p-value $= 1 - \frac{t_c-1}{t}$ is relatively high and close to 1, so, unlikely lower than $\alpha (= 0.05)$. The null hypothesis probably fails to be rejected. However, for high values of t_c, i. e. when there are many trees in the forest with sufficient complexity classifying into the two categories c_r and c_s (thus, in contradiction to the null hypothesis), then – since the high value of t_c – the fraction p-value $= 1 - \frac{t_c-1}{t}$ is relatively low and perhaps below the α level. Consequently, the null hypothesis is likely rejected.

Indeed, the κ parameter determines how complex the trees in the random forest are or how radical the pruning of the trees is. Investigating formula (13), one can realize that if $\kappa = 0$, then there is no penalization for large tree complexity, so trees in the random forest are generally very complex (i. e., of large size). So, whenever there are at least two observations, one from c_r category and the other from c_s category of j-th factor, all trees in the forest would classify those observations into their categories, i. e. that for each tree τ is $\mathcal{I}_{c_r,c_s}(\tau) = 1$, which results into $t_c = \sum_{\forall \tau \in \text{ random forest}} \mathcal{I}_{c_r,c_s}(\tau) = t$, and, thus, p-value estimate is p-value $= 1 - \frac{t_c-1}{t} = 1 - \frac{t-1}{t} = \frac{1}{t} \approx 0$. Finally, if p-value ≈ 0, then also p-value $\approx 0 < \alpha$ which, consequently, results into the null hypothesis rejection. However, when the null hypothesis rejection is often, it is also very likely a *false* rejection, that increases the first-type error rate. High chance of the null hypothesis rejection means also the high statistical power, though, i. e. the case when the incorrect null hypothesis is *correctly* rejected.

For $\kappa > 0$, the penalization for tree complexity (size) is applied, so, the trees' complexity (size) decreases, and, thus, if not all, many of the trees do not classify into both c_r and c_s categories. This means that there are trees τ in the random forest so that $\mathcal{I}_{c_r,c_s}(\tau) = 0$, and, finally, $t_c = \sum_{\forall \tau \in \text{ random forest}} \mathcal{I}_{c_r,c_s}(\tau) < t$. So, p-value estimate is p-value $= 1 - \frac{t_c-1}{t} > 0$, and it could be below or above the α level.

[3]This is ensured by the bootstrapped selection of $k^* < k$ node rule variables.

B. A feasible low bound of the number of trees t in the random forest

Adopting the ANCOVA mathematical notation, when two categories c_r and c_s of j-th factor are compared using the random forest-based method, other $k-1$ factors, together with m covariates and the originally dependent continuous variable, play as input variables for node decision rules. Assuming that ℓ-th factor contains $n_\ell \geq 2$ categories and cut-offs for the covariates are usually estimated as midpoints of covariates' ranges, splitting the ranges into a number of categories to be classified into, i. e. n_c, we may estimate a minimum number of mutually different trees. Each mentioned feature could be or could not be included in a tree; that being said, the number of all combinations of $k-1$ factor node rules is at least $2^{\prod_{\ell \in \{1,2,\ldots,k\} \setminus j} n_\ell} \geq 2^{2^{k-1}}$, and the number of all combinations of m covariates and the dependent variable, split into n_j intervals, is at least n_j^{m+1}. Thus, the minimum number of mutually different trees and, thus, the feasible low bound of the random forest trees' number is

$$t > 2^{\prod_{\ell \in \{1,2,\ldots,k\} \setminus j} n_\ell \cdot n_j^{m+1}} \geq 2^{2^{k-1} \cdot n_j^{m+1}}.$$

Furthermore, since the number of trees t determines decimal precision of p-value estimate based on formula 14, if we ask for decimal precision of $d \in \mathbb{N}$ digits, then the minimum number of trees in the random forest is about

$$t > 10^{d+1},$$

to ensure feasible precision for d-th decimal digit.

C. A brief asymptotic time complexity analysis of the proposed method

The random forest model consists of classification trees as atomic units, constructed following the flowchart 1 and algorithm 1. A decision tree is induced until the moment all its leaf nodes include only observations of one category of j-th factor, i. e. a sequence of node rules coming from root note till the leaf one successively limits the entry dataset to one-class observations [13]. When the categories are well balanced across the dataset, each binary partitioning halves them, and the tree average depth is around $\log_2 n$ levels; thus, the asymptotic time complexity is also $\Theta(\log_2 n)$, assuming one split of a node takes one time atomic unit. However, if the categories are totally unbalanced, each splitting cuts the current dataset of size n^* into 1 and $n^* - 1$ observations, which takes n steps in total. Then, tree depth is n, so the asymptotic time complexity is $\Theta(n)$. Assuming there are $k-1$ factors, m covariates and the originally dependent variable, i. e. $k - 1 + m + 1 = k + m$ variables, searched through averagely $\frac{n}{2}$ observations within each node splitting, the asymptotic time complexity of one tree induction $\Theta(\dagger)$

is therefore between $\Theta(\log_2 n)$ (best-case scenario) and $\Theta(n)$ (worst-case scenario),

$$\Theta\left((k+m)\frac{n}{2}\log_2 n\right) \leq \Theta(\dagger) \leq \Theta\left((k+m)\frac{n}{2}n\right),$$
$$\Theta\left(\frac{(k+m)n}{2}\log_2 n\right) \leq \Theta(\dagger) \leq \Theta\left(\frac{(k+m)n^2}{2}\right),$$
$$\Theta\left((k+m)n\log_2 n\right) \lesssim \Theta(\dagger) \lesssim \Theta\left((k+m)n^2\right). \quad (15)$$

Algorithm 1: The top-down induction of decision trees (TDIDT) following the logic of the flowchart 1

Data: a $n \times (m+k)$ dataset of n observations, with j-th target factor, $k-1$ factors, m covariates, and one dependent continuous variable
Result: a classification tree

$T = (\{\boldsymbol{n}\})$ // a tree T with a set ;
// of nodes $\boldsymbol{n}$;
$\{\boldsymbol{n}\} = \{\texttt{root}\}$ // initially, the tree T ;
// is a root;
$\sigma(\bullet)_j$ // a node criterion;
while $\exists$ *a node* $\in \{\boldsymbol{n}\}$ *so that data constrained by all node rules coming from root to the node belong to* ≥ 2 *classes* **do**
  find for the node a splitting variable and splitting point minimizing the $\sigma(\bullet)_j$;
  add to the node two child nodes n_{left} a n_{right};
  $\{\boldsymbol{n}\} := \{\boldsymbol{n} \cup \{n_{\text{left}}, n_{\text{right}}\}\}$;
  $T := (\{\boldsymbol{n}\})$ // update the tree using;
  // the new node set $\boldsymbol{n}$;
end
a completely induced tree T;

Since a random forest contain t trees, each built in $\Theta(\dagger)$ time by (15), the entire random forest asymptotic time complexity construction $\Theta(\ddagger)$ is

$$\Theta\left(t\frac{(k+m)n}{2}\log_2 n\right) \leq \Theta(\ddagger) \leq \Theta\left(t\frac{(k+m)n^2}{2}\right),$$
$$\Theta\left(t(k+m)n\log_2 n\right) \lesssim \Theta(\ddagger) \lesssim \Theta\left(t(k+m)n^2\right). \quad (16)$$

One model of the random forest provides one (point) estimate of the p-value using the formula (14), enabling to statistically distinguish between two categories. In comparison, the estimation of the decision rules for post-hoc tests using formulas (6), (7), (8), and (9) usually take only several linear steps, assuming the $\text{SS}_{j\text{-th factor}}$ in (7) term is precalculated. Fortunately, the time complexity (16) is still polynomial. Furthermore, since the building of the random forest with the complexity of (16) is based on independent trees induction, it could be parallelized; then, if the random forest building would be parallelized into $\pi \leq t$ independent slave processes each inducting a bunch of $\frac{t}{\pi} \in \mathbb{N}$ trees, the time complexity (16) would be reduced to

$$\Theta\left(t\frac{(k+m)n}{2\pi}\log_2 n\right) \leq \Theta(\ddagger) \leq \Theta\left(t\frac{(k+m)n^2}{2\pi}\right),$$
$$\Theta\left(\frac{t}{\pi}(k+m)n\log_2 n\right) \lesssim \Theta(\ddagger) \lesssim \Theta\left(\frac{t}{\pi}(k+m)n^2\right).$$

V. Simulation study

To compare the established post-hoc tests with the proposed method, particularly its first-type error rate, we run a simulation study generating many $n \times (k+m+1)$ datasets with n observations, k factor variables, m covariates with various relationships between the variables, and, lastly, with the continuous dependent variable. For each post-hoc test, i. e. Tukey HSD test, Scheffé's test, Dunn's test, Nemenyi's test, and random forest-based method, we compare two categories of a selected factor so that the categories have significantly non-different averages within the continuous dependent variable and check how many times the methods claim there is a significant difference. Thus, in theory, we measure the first-type error rate. The simulation was repeated for different κ parameter values to illustrate how the value of κ determines the first-type error rates in the new method, i. e., what are ideal κ values to control the first-type error rate on a feasible level.

The datasets were generated as follows. One of the k factors, let's say the j-th one, contained two categories, c_r and c_s, following a normal distribution with the same average of the continuous dependent variable, i. e. $\mathcal{N}(0, 1^2)$. Other $k-1$ factors always split the dependent continuous variable into 2 to 4 categories with random averages from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, where $\boldsymbol{\mu} \in \langle -1, 1\rangle$ and $\boldsymbol{\sigma}^2 \in \langle 0, 2\rangle$. Furthermore, the covariates also followed normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, where $\boldsymbol{\mu} \in \langle -1, 1\rangle$, $\boldsymbol{\sigma}^2 \in \langle 0, 2\rangle$, and correlations $\boldsymbol{r}$ between the dependent continuous variable and covariates were randomly from $\boldsymbol{r} \in \langle -0.5, 0.5\rangle$. The continuous variable was dependent for all post-hoc tests with exception of the proposed method, where j-th factor is as dependent one.

There were $\eta = 1000$ datasets, as depicted above, generated in total, and for each $\kappa \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The number of trees in each random forest was always $t = 1000$. Numbers of cases where p-value was lower than or equal to $\alpha = 0.05$ regardless of the post-hoc test were summed up, indicating the point estimates of the first-type error rates, as illustrated in table I. The simulation study was performed using R programming language and environment [14]. There are more numerical applications of R language to various fields in [15]–[23].

While the established post-hoc tests returned point estimates of the first-type error rate about 0.050 (regardless of κ since their formulas (6), (7), (8), and (9) are not functions of the κ), point estimates of the first-type error rates output by the introduced method progressively decreased with increasing value of κ, see table I.

However, since the proposed random forest-based algorithm for categories' averages comparison is data-determined and heuristic, it is nontrivial to suggest specific values of the

TABLE I
POINT ESTIMATES OF THE FIRST-TYPE ERROR RATES FOR POST-HOC TESTS, I. E. TUKEY HSD TEST (T-HSD), SCHEFFÉ'S TEST (SCHT), DUNN'S TEST (DT), NEMENYI'S TEST (NT), AND THE PROPOSED METHOD (RF-T) FOR DIFFERENT VALUES OF TUNING PARAMETER κ, BASED ON THE SIMULATION DESCRIBED ABOVE.

	method					
	T-HSD	SchT	DT	NT	RF-T	κ
# of cases in total	1000	1000	1000	1000	1000	0.1
$\#\{p\text{-value} \leq 0.05\}$	56	51	45	42	66	
first-type error rate	0.056	0.051	0.045	0.042	0.066	
# of cases in total	1000	1000	1000	1000	1000	0.3
$\#\{p\text{-value} \leq 0.05\}$	54	51	45	42	58	
first-type error rate	0.056	0.051	0.045	0.042	0.058	
# of cases in total	1000	1000	1000	1000	1000	0.5
$\#\{p\text{-value} \leq 0.05\}$	60	58	45	51	49	
first-type error rate	0.060	0.058	0.045	0.041	0.049	
# of cases in total	1000	1000	1000	1000	1000	0.7
$\#\{p\text{-value} \leq 0.05\}$	50	56	41	50	38	
first-type error rate	0.050	0.056	0.041	0.050	0.038	
# of cases in total	1000	1000	1000	1000	1000	0.9
$\#\{p\text{-value} \leq 0.05\}$	47	53	48	46	29	
first-type error rate	0.047	0.053	0.048	0.046	0.029	

pruning parameter κ to reach a given level of the first-type error rate.

Still, as indicated by the derived theory and simulation study, the higher the pruning parameter κ is, the higher penalization for too complex trees in a random forest is. Thus, the less complex the trees in a random forest are, which results in lower trees' ability to classify into two or more classes of the factor variable and, consequently, the lower first-type error rate of the random forest as an inferential algorithm.

VI. CONCLUSION

When searching for statistical differences between categories' averages of a given factor, once analysis of covariance is performed, post-hoc tests may identify which two or more categories have significantly different impacts on the dependent continuous variable average. However, the post-hoc tests are usually limited by rigid statistical assumptions.

In this work, we introduced a novel method for post-hoc testing based on a random forest algorithm. Rather than a statistical comparison of dependent variable's averages for two factor categories and evaluation of its effect size, the proposed technique refines the logic of testing. The factor with compared categories becomes an output variable, i. e., its categories populate leaf nodes of the model trees, and other factors, initially dependent continuous variable and covariates, if any, serve input variables, i. e., as quantities in node rules. The higher the random forest model trees' complexity, i. e., size is, the more likely the trees classify into (besides others) the two compared categories, and, thus, the null hypothesis claims there is no statistical difference between the compared categories' averages is more likely to be rejected. Furthermore, since trees' pruning level determines the trees in the random forest model complexity, a tuning parameter affecting the significance of the pruning also changes trees' complexity and, consequently, the probability of correctly rejecting the null hypothesis. Finally, the tree pruning may modify the first-type error rate, too. The asymptotic time complexity of the random forest-based post-hoc method is usually higher than the complexities of the established procedures but is still polynomial and might be parallelized.

Therefore, the introduced random forest-based method seems to be a valid alternative to other, commonly used post-hoc tests.

VII. ACKNOWLEDGEMENT

This paper is supported by the grant OP VVV IGA/A, CZ.02.2.69/0.0/0.0/19_073/0016936 with no. 18/2021, which has been provided by the Internal Grant Agency of the Prague University of Economics and Business.

REFERENCES

[1] Geoffrey Keppel and Thomas D Wickens. *Design and analysis*. en. 4th ed. Upper Saddle River, NJ: Pearson, Jan. 2004.

[2] John W. Tukey. "Comparing Individual Means in the Analysis of Variance". In: *Biometrics* 5.2 (June 1949), p. 99. DOI: 10.2307/3001913. URL: https://doi.org/10.2307/3001913.

[3] H Scheffe. *The analysis of variance*. en. Wiley Classics Library. Nashville, TN: John Wiley & Sons, Feb. 1999.

[4] Olive Jean Dunn. "Multiple Comparisons among Means". In: *Journal of the American Statistical Association* 56.293 (Mar. 1961), pp. 52–64. DOI: 10.1080/01621459.1961.10482090. URL: https://doi.org/10.1080/01621459.1961.10482090.

[5] Myles Hollander and Douglas Alan Wolfe. *Nonparametric Statistical Methods*. en. 2nd ed. Wiley series in probability & statistics: applied section. Nashville, TN: John Wiley & Sons, Feb. 1999.

[6] Ellen R Girden. *ANOVA: Repeated measures*. 84. Sage, 1992. ISBN: 0803942575.

[7] Kenneth L. Lange, Roderick J. A. Little, and Jeremy M. G. Taylor. "Robust Statistical Modeling Using the t Distribution". In: *Journal of the American Statistical Association* 84.408 (Dec. 1989), p. 881. DOI: 10.2307/2290063. URL: https://doi.org/10.2307/2290063.

[8] A. Charnes, E. L. Frome, and P. L. Yu. "The Equivalence of Generalized Least Squares and Maximum Likelihood Estimates in the Exponential Family". In: *Journal of the American Statistical Association* 71.353 (Mar. 1976), pp. 169–171. DOI: 10.1080/01621459.1976.10481508. URL: https://doi.org/10.1080/01621459.1976.10481508.

[9] R A Bailey. *Cambridge series in statistical and probabilistic mathematics: Design of comparative experiments series number 25*. Cambridge, England: Cambridge University Press, Apr. 2008.

[10] Leo Breiman. *Classification and regression trees*. New York: Chapman & Hall, 1993. ISBN: 9780412048418.

[11] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/a:1010933404324. URL: https://doi.org/10.1023/a:1010933404324.

[12] Lubomír Štěpánek, Filip Habarta, Ivana Malá, and Luboš Marek. "Analysis of asymptotic time complexity of an assumption-free alternative to the log-rank test". In: *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems*. IEEE, Sept. 2020. DOI: 10.15439/2020f198. URL: https://doi.org/10.15439/2020f198.

[13] Kawther Hassine, Aiman Erbad, and Ridha Hamila. "Important Complexity Reduction of Random Forest in Multi-Classification Problem". In: *2019 15th International Wireless Communications Mobile Computing Conference (IWCMC)*. 2019, pp. 226–231. DOI: 10.1109/IWCMC.2019.8766544.

[14] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017. URL: https://www.R-project.org/.

[15] Lubomír Štěpánek, Pavel Kasal, and Jan Měšťák. "Evaluation of facial attractiveness for purposes of plastic surgery using machine-learning methods and image analysis". In: *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, Sept. 2018. DOI: 10.1109/healthcom.2018.8531195. URL: https://doi.org/10.1109/healthcom.2018.8531195.

[16] Lubomír Štěpánek, Pavel Kasal, and Jan Měšťák. "Machine-learning at the service of plastic surgery: a case study evaluating facial attractiveness and emotions using R language". In: *Proceedings of the 2019 Federated Conference on Computer Science and Information Systems*. IEEE, Sept. 2019. DOI: 10.15439/2019f264. URL: https://doi.org/10.15439/2019f264.

[17] Lubomír Štěpánek, Pavel Kasal, and Jan Měšťák. "Machine-Learning and R in Plastic Surgery – Evaluation of Facial Attractiveness and Classification of Facial Emotions". In: *Advances in Intelligent Systems and Computing*. Springer International Publishing, Sept. 2019, pp. 243–252. DOI: 10.1007/978-3-030-30604-5_22. URL: https://doi.org/10.1007/978-3-030-30604-5_22.

[18] Lubomír Štěpánek, Pavel Kasal, and Jan Měšťák. "Machine-learning at the service of plastic surgery: a case study evaluating facial attractiveness and emotions using R language". In: *Proceedings of the 2019 Federated Conference on Computer Science and Information Systems*. IEEE, Sept. 2019. DOI: 10.15439/2019f264. URL: https://doi.org/10.15439/2019f264.

[19] Lubomír Štěpánek, Pavel Kasal, and Jan Měšťák. "Evaluation of Facial Attractiveness after Undergoing Rhinoplasty Using Tree-based and Regression Methods". In: *2019 E-Health and Bioengineering Conference (EHB)*. IEEE, Nov. 2019. DOI: 10.1109/ehb47216.2019.8969932. URL: https://doi.org/10.1109/ehb47216.2019.8969932.

[20] Lubomír Štěpánek, Filip Habarta, Ivana Malá, Luboš Marek, and Filip Pazdírek. "A Machine-learning Approach to Survival Time-event Predicting: Initial Analyses using Stomach Cancer Data". In: *2020 International Conference on e-Health and Bioengineering (EHB)*. IEEE, Oct. 2020. DOI: 10.1109/ehb50910.2020.9280301. URL: https://doi.org/10.1109/ehb50910.2020.9280301.

[21] Lubomír Štěpánek, Filip Habarta, Ivana Malá, Luboš Marek, and Filip Pazdírek. "A random forest-based approach for survival curves comparing: principles, computational aspects and asymptotic time complexity analysis". In: *Proceedings of the 16th Conference on Computer Science and Intelligence Systems*. IEEE, Sept. 2021. DOI: 10.15439/2021F89.

[22] Lubomír Štěpánek, Filip Habarta, Ivana Malá, Luboš Marek, and Filip Pazdírek. "Data envelopment analysis models connected in time series: A case study evaluating COVID-19 pandemic management in some European countries". In: *2021 International Conference on e-Health and Bioengineering (EHB)*. Iasi, Romania: IEEE, Nov. 2021. DOI: 10.1109/EHB52898.2021.9657597.

[23] Owen Jones, Robert Maillardet, and Andrew Robinson. *Introduction to Scientific Programming and Simulation Using R*. Chapman and Hall/CRC, Mar. 2009. DOI: 10.1201/9781420068740. URL: https://doi.org/10.1201/9781420068740.