Kuhn, M. (2008). Building predictive models in R using the `caret` package. *J. Stat. Software*, **28**(5), 1–26.
Vukićević, M., Jovanović, M., Delibašić, B., Išljamović, S. & Suknović, M. (2012). Reusable component-based architecture for decision tree algorithm design. *Int. J. Artif. Intell. Tools*, **21**(05).
Wolpert, D.H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Comput.*, **8**(7), 1341–1390.
Wu, X., Kumar, V., Quinlan, R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G., Ng, A., Liu, B., Philip, Y., Zhou, Z.-H., Steinbach, M., Hand, D. & Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowl. Inf. Syst.*, **14**(1), 1–37.

[*Received April 2014, accepted April 2014*]

International Statistical Review (2014), 82, 3, 367–370 doi:10.1111/insr.12057

Rejoinder

Wei-Yin Loh

Department of Statistics, University of Wisconsin, Madison, WI 53706, USA
E-mail: loh@stat.wisc.edu

I thank the discussants for their thoughtful comments, which helped to fill in some gaps and expand the scope of the review. I will address each one below.

Carolin Strobl wonders when it is helpful to create separate nodes for missing values. CHAID seems to be the only algorithm to do this, but it has a procedure to merge some of the nodes before they are split further. Its effectiveness has not been studied. GUIDE treats missing values in a categorical variable as a separate category but does not assign them to a separate node. If there are missing values at a split on an ordered variable, GUIDE sends them to the same left or right child node, depending on which split yields the greater decrease in node impurity. Ding & Simonoff (2010) studied a simpler version of this technique, where missing ordered values are mapped to infinity and hence are always sent to the right child node. Using only binary-valued variables with training and test sets having missing values where missingness in a predictor variable depends on the values of the response variable (MAR), they found that this technique is better than case deletion, variable deletion, grand mean/mode imputation, surrogate splits (as used in RPART) and fractional weights (as used in C4.5). Case deletion and grand mean/mode imputation tend to be worst, a finding supported by Twala (2009), who considered missingness dependent on other predictor variables (MAR) and missingness due to truncation (MNAR) but not missingness dependent on the response variable. He found that the best method was an ensemble of classification trees constructed by multiple imputation of the missing values with the expectation–maximization algorithm (Dempster *et al.*, 1977; Rubin, 1987). It is not clear whether this is either due to multiple imputation or ensemble averaging. Note that because both studies employed C4.5 and RPART exclusively as the base classifiers, it is unknown if the conclusions extend to other methods. Further, some other missing value techniques, such as nodewise mean and mode imputation (FACT and QUEST) and alternative surrogate split methods (CRUISE), were not included.

Strobl wonders whether ignorance is the reason that biassed recursive partitioning methods continue to be used frequently. Many people still associate the term ‘classification and regression trees’ with CART and its software. Commercial software publishers perpetuate this misconception by largely basing their offerings on CART. The availability of RPART also

encouraged the use and extension of CART (e.g. MVPART). On the other hand, selection bias may not cause serious harm if a tree model is used for prediction but not interpretation, in some situations. Selection bias can increase the likelihood of spurious splits on irrelevant variables, but if the sample size is large and there are not too many such variables, a correspondingly large tree may subsequently split on the important variables. If the spurious splits survive after pruning, they simply stratify the data into two or more subsets each having its own subtree, and overall prediction accuracy may be preserved; see Martin (1997) for a related discussion. If the sample size is small, however, then the spurious splits will increase the frequency of trivial pruned trees.

Strobl's comments on importance scores bring us back to the meaning of the 'importance' of a variable. Because it is not well defined, the concept has produced a plethora of importance measures. CART has a measure based on the efficiency of surrogates splits, and FACT has a measure based on F-statistics. At that time, both seemed reasonable as there was not much need for either, due to data sets being small and variables few. Now that data sets can contain thousands and even millions of variables, the situation is different. Further, there is evidence (e.g. Doksum *et al.*, 2008 and Loh, 2012) that when the number of variables is very large, some sort of preliminary variable selection can substantially improve the prediction accuracy of a model. Because importance measures are well suited (and are being used) for this task, it is time to examine the notion more carefully. As Strobl observes, some people consider the importance of a variable 'more or less on its own' [e.g. random forest (RF)], whereas others think of it as the residual effect after other variables are accounted for in a model (e.g. linear regression). Although the latter point of view is more specific, it lacks a sense of universality, because a variable can be important for one model but not for another. On the other hand, perhaps universality (i.e. being model free) is not attainable. Nonetheless, there is one property that every importance measure ought to have, namely, unbiasedness. Strobl *et al.* (2007) showed that RF is biased in the 'null' sense that, if all the variables are independent of the response, the frequencies with which they appear in the trees depend on their types.

A more general definition of unbiasedness, applicable to non-forest methods as well, is that under this null scenario, all variables are ranked equally on average. To see how RPART, RF, Cforest (CF, from the PARTY package) and GUIDE perform by this criterion, I simulated 5000 data sets with each set consisting of 100 observations on seven mutually independent predictor variables and one normally distributed response variable. Three predictor variables are continuous (normal, uniform and chi-squared with two degrees of freedom) and four are categorical with 2, 4, 10 and 20 equiprobable categories; all are independent of the response variable. Figure 3 shows the average rank of each variable for each method (rank 1 is most important and rank 7 the least). RPART and RF tend to find the variables with 10 and 20 categories the most important (although they differ on the *most* important) and the binary predictor the least. CF has a slight bias towards ranking the binary variable most important and the two variables with 10 and 20 categories least; GUIDE has a smaller bias towards ranking ordinal variables more important than categorical variables. The biases of CF and GUIDE are negligible, however, compared with that of the other two.

Antonio Ciampi touches on several philosophical issues. I will offer my take on some of his main points. The 'dilemma' between interpretability and accuracy is a result of the human mind's limitations in understanding complex structure. Fortunately, if the structure is comprehensible, the mind is exceptionally good at drawing insights that no machine can match. Therefore, rather than a dilemma, it is really a choice: construct simple models that humans can use to enhance their understanding of the problem or build complex models for automatic and accurate predictions. Both are laudable goals.

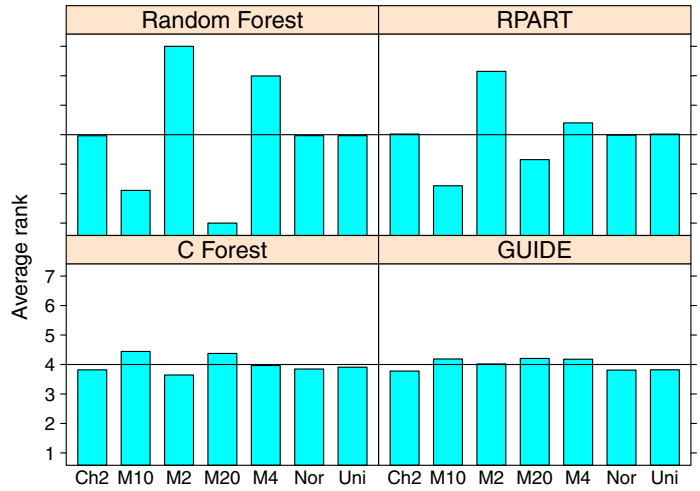

Figure 1. *Average ranks (with 1 being most important and 7 least) of variables for sample size 100 over 5000 simulation iterations. Simulation standard errors are less than 0.03. 'Ch2', 'Nor' and 'Uni' denote* χ^2_2, $N(0,1)$ *and* $U(0,1)$ *variables. 'Mk' denotes a multinomial (categorical) variable with k equiprobable levels. The response variable is* $N(0,1)$ *and all variables are mutually independent. A method is unbiased if each variable has average rank 4.0, which is marked by horizontal lines.*

Ciampi mentions several ways to improve the accuracy of single trees, such as using global model search (Bayesian and genetic algorithms) and probabilistic splits (soft nodes). The accuracy of Bayesian trees comes from model averaging; there is no evidence that the tree with the largest posterior probability has comparable accuracy. Global search techniques inevitably produce randomised solutions that may be undesirable in some applications. It is harder to follow the path of an observation in a model with probabilistic splits than it is in a model with conventional (hard) splits.

I agree that univariate (monothetic) splits are not the only ones that are interpretable. Ciampi's example of a (polythetic) split on the number of symptoms possessed by a patient is certainly interpretable and can be implemented as sums of indicator variables. But because there are numerous combinations of variables that can form the sums, this approach invites computational and selection bias problems. His idea of hierarchical tree structures is intriguing, particularly if the predictor variables are naturally clustered.

Ciampi notes that data have become more complex. In business, biology, medicine and other fields, predictor and response variables are increasingly observed as longitudinal series. Owing to difficulties caused by the number and location of the observation 'time' points varying between subjects, small number of subjects relative to number of time points, large number of baseline covariates and occurrence of missing values, the traditional approach of fitting parametric stochastic models to the processes is seldom feasible. A more practical solution may be a non-parametric approach that treats the longitudinal series as random functions (Loh & Zheng, 2013).

I thank Hongshik Ahn for his review of some of the more recent ensemble methods. The fact that the accuracy of an ensemble increases as the dependence among the component classifiers decreases, provided that the latter are equally accurate, motivates the construction of ensembles where each classifier is built from a mutually exclusive subset of predictors. But it is difficult to do this without destroying the requirement of equally accurate classifiers. This is obvious when there is exactly one informative predictor variable and many irrelevant ones. Then, all but one classifier (the one involving the informative predictor) do nothing except to dilute the accuracy

of the ensemble. On the other hand, the classifier containing the informative variable should be more accurate than the one built with all the variables. This may explain the behaviour of CERP and LORENS. The WAVE method of adaptively assigning weights to classifiers seems to be a promising direction.

I thank Chi Song and Heping Zhang for the references to genetic applications. Subgroup identification, a key part of personalised medicine, is rapidly gaining attention. The goal is to find patient subgroups, defined by measurable patient characteristics (such as demographic, phenotype, genotype and protein biomarkers) prior to treatment, that respond differentially to treatment. Negassa *et al.* (2005), Su *et al.* (2009), Foster *et al.* (2011), Lipkovich *et al.* (2011) and Dusseldorp & Van Mechelen (2013) have proposed tree-based solutions. Alternatives that do not have selection bias have been implemented in the GUIDE software.

While I agree with Thomas Rusch and Achim Zeileis that in an ideal world, all published algorithms would be accompanied by free software, there are reasons why this does not always occur in practice. Quite often, the author of the software is a student who is not the architect of the algorithm. When the student graduates, there is no one to distribute and maintain the code. This was the case with the SUPPORT algorithm, although its best features have since been incorporated in GUIDE. Then, there is the author who plans only to publish a paper and move on to other problems, with no intention to distribute and maintain the software. As a result, the latter is developed only as far as it is needed for the examples and simulations in the paper.

Acknowledgements

This rejoinder was prepared with partial support from the National Science Foundation grant DMS-1305725.

References

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **39**, 1–38.

Doksum, K., Tang, S. & Tsui, K.-W. (2008). Nonparametric variable selection: The EARTH algorithm. *J. Amer. Statist. Assoc.*, **103**, 1609–1620.

Dusseldorp, E. & Van Mechelen, I. (2013). Qualitative interaction trees: A tool to identify qualitative treatment–subgroup interactions. *Stat. Med.*, **33**, 219–237, DOI 10.1002/sim.5933.

Foster, J.C., Taylor, J.M.G. & Ruberg, S.J. (2011). Subgroup identification from randomized clinical trial data. *Stat. Med.*, **30**, 2867–2880.

Lipkovich, I., Dmitrienko, A., Denne, J. & Enas, G. (2011). Subgroup identification based on differential effect search–a recursive partitioning method for establishing response to treatment in patient subpopulations. *Stat. Med.*, **30**, 2601–2621.

Martin, J.K. (1997). An exact probability metric for decision tree splitting and stopping. *Mach. Learn.*, **28**, 257–291.

Negassa, A., Ciampi, A., Abrahamowicz, M., Shapiro, S. & Boivin, J.R. (2005). Tree-structured subgroup analysis for censored survival data: Validation of computationally inexpensive model selection criteria. *Stat. Comput.*, **15**, 231–239.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Su, X., Tsai, C.L., Wang, H., Nickerson, D.M. & Bogong, L. (2009). Subgroup analysis via recursive partitioning. *J. Mach. Learn. Res.*, **10**, 141–158.

Twala, B. (2009). An empirical comparison of techniques for handling incomplete data using decision trees. *Appl. Artificial Intell.*, **23**, 373–405.

[*Received March 2014, accepted March 2014*]