

Visualizing Variable Importance and Variable Interaction Effects in Machine Learning Models

Alan Inglis, Andrew Parnell & Catherine B. Hurley

To cite this article: Alan Inglis, Andrew Parnell & Catherine B. Hurley (2022): Visualizing Variable Importance and Variable Interaction Effects in Machine Learning Models, Journal of Computational and Graphical Statistics, DOI: [10.1080/10618600.2021.2007935](https://doi.org/10.1080/10618600.2021.2007935)

To link to this article: <https://doi.org/10.1080/10618600.2021.2007935>



© 2022 The Author(s). Published with license by Taylor & Francis Group, LLC.



[View supplementary material](#)



Published online: 04 Jan 2022.



[Submit your article to this journal](#)



Article views: 494



[View related articles](#)



[View Crossmark data](#)

Visualizing Variable Importance and Variable Interaction Effects in Machine Learning Models

Alan Inglis^a , Andrew Parnell^b , and Catherine B. Hurley^c 

^aHamilton Institute, Maynooth University, Maynooth, Ireland; ^bHamilton Institute, Insight Centre for Data Analytics, Maynooth University, Maynooth, Ireland; ^cDepartment of Mathematics and Statistics, Maynooth University, Maynooth, Ireland

ABSTRACT

Variable importance, interaction measures, and partial dependence plots are important summaries in the interpretation of statistical and machine learning models. In this article, we describe new visualization techniques for exploring these model summaries. We construct heatmap and graph-based displays showing variable importance and interaction jointly, which are carefully designed to highlight important aspects of the fit. We describe a new matrix-type layout showing all single and bivariate partial dependence plots, and an alternative layout based on graph Eulerians focusing on key subsets. Our new visualizations are model-agnostic and are applicable to regression and classification supervised learning settings. They enhance interpretation even in situations where the number of variables is large. Our R package `vivid` (variable importance and variable interaction displays) provides an implementation. Supplementary files for this article are available online.

ARTICLE HISTORY

Received August 2021
Revised November 2021

KEYWORDS

Black-box; Model explanation; Model visualization

1. Introduction

Visualization is a key tool in understanding statistical and machine learning models. In this article, we present new visualizations to serve two main goals, namely improved model understanding and interpretation. Our new visualizations are based on variable importance and interaction measures, and partial dependence plots (PDPs). A variable importance value is used to express (in a scalar quantity) the degree to which a variable affects the response value through the chosen model. A variable interaction is a scalar quantity that measures the degree to which two (or more) variables combine to affect the response variable. Variable importance and variable interaction (henceforth VImp and VInt; together VIVI) are widely used in many fields to understand and explain the behavior of a model. In biology, they are used to examine gene–gene interactions (e.g., Wang et al. 2012). In high-energy physics VImp can be an important tool in high-dimensional feature selection processes (e.g., Gleyzer and Prosper 2008). In econometrics, they are common tools to evaluate interaction behaviour (e.g., Balli and Sorensen 2010).



Traditional methods of displaying VImp or VInt use variants of line or bar plots, see, for example, Molnar (2019). However, in variable importance plots, there is relatively little emphasis on displaying how pairs of interacting variables may be important in a model. This can be a hindrance to model interpretation, especially if a variable has low importance but a high interaction strength. The inclusion of interacting terms in a model has been shown to affect the prediction performance (Oh 2019).

However, as shown in Wei et al. (2015a), for high-dimensional models that are governed mainly by interaction effects, the performance of certain types of permutation-based variable importance measures will decrease and thereby produce low values of importance. Consequently, viewing the VInt and VImp together provides a more complete picture of the behavior of a model fit.

Our new displays present VInt and VImp jointly in a single plot. We allow for seriation so that variables are reordered with those exhibiting high VIVI grouped together. This assists in interpretation and is particularly useful as the number of variables becomes large. Furthermore, we make use of filtering, so less influential variables can be removed. For our network displays, we use graph clustering to group together interacting variables.


Partial dependence plots were introduced by Friedman (2000) to show how the model's predictions are affected by one or two predictors. In addition to the above, we propose a new display which shows all pairwise PDPs in a matrix-type layout, with a univariate PDP on the diagonal, similar to a scatterplot matrix. With this display, the analyst can explore, at a glance, how important pairs of variables impact the fit. Once again, careful reordering of the variables facilitates interpretation.

Our final display takes the filtering of all pairwise PDPs a step further. We select only those pairwise PDPs with high VInt, and display an Eulerian path visiting these plots by extending the zigzag display algorithm of Hofert and Oldford (2020). We call this a zen-partial dependence plot (ZPDP).

CONTACT Alan Inglis  alan.inglis@mu.ie  Hamilton Institute, Maynooth University, Maynooth, Ireland

¹We use the term “variable” throughout to denote the input to a statistical and machine learning model as this seems to be the most common parlance.

Other terms commonly used include feature, predictor, explanatory variable, independent variable, etc.

 Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

© 2021 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

These new visualizations can be used to explore machine learning models more thoroughly in an easily interpretable way, providing useful insights into variable impact on the fit. This is demonstrated by practical examples. In each plot careful consideration is given to various aspects of the design, including color choices, optimising layouts via seriation, graph clustering, and Euler paths for the ZPDP. Filtering options limit the plots to variables deemed relevant from VImp or VInt scores. Our new displays are appropriate for supervised regression and classification fits, and are model and metric agnostic in that no particular model fit nor importance method is prescribed. The methods described here are implemented in our R package `vivid` (Inglis et al. 2021).

The organization of the article is as follows. In [Section 2](#), we discuss the concepts of VImp and VInt. Then, we describe our new heatmap and network displays of joint variable importance and interaction and demonstrate these on an example. In [Section 3](#), we discuss our new layouts for collections of PDPs, either in a matrix format or zig-zag layout and show their application. In [Section 4](#), we use our new methodology to explore a machine learning fit from a larger dataset. Finally in [Section 5](#), we offer some concluding discussion.

2. Visualizing Variable Importance and Interaction

We begin with a nonexhaustive review of the concepts of VImp and VInt. Though the visualizations, we present are agnostic to the measures used to determine these scalar quantities, some degree of understanding is helpful in interpreting the later plots. We then describe our new visualizations and their design principles and provide illustrations.

2.1. Measuring Variable Importance

A VImp is a scalar measure of a variable's influence on the response. Many techniques have been proposed to calculate variable importance, depending on the type of model. The term “influence” here may encompass changes in the mean response or that of higher order uncertainty. In our work, we focus exclusively on changes in the mean. For a wider review of variable importance techniques, and the different goals that a variety of approaches may achieve see Wei et al. (2015b).

Much of the initial work in VImp focused on estimating the partial derivative of the response with respect to one or two input variables (Frey and Patil 2002). This is a global VImp measure when the model is linear, but perhaps less useful (though still potentially interesting) in nonlinear models where it is often defined as a local importance measure. In high-dimensional settings, these methods can be discretized across a hyper-cube to allow for the identification of, for example, linearity in a nonlinear model (Helton and Davis 2002). Due to their local behavior, we do not incorporate them into our visualizations below.

Some VImp measures arise naturally out of a model structure. The most familiar would be those based on summary statistics created from regression models, such as standardized coefficient values, (partial) correlation coefficients, and R^2 . Many of these can be extended to nonlinear models such as

generalized additive models (Wood 2000), or projection pursuit regression (Friedman and Stuetzle 1981). R^2 in particular seems useful as a VImp measure, as it can be defined for a wide variety of statistical models and can be decomposed into main and potentially high order interaction effects, yielding a VInt measure in addition.

Similarly, other model-structure based methods arise out of now standard machine learning techniques. Random forests, for example, involves the use of the Gini coefficient, and the reduction in mean square error, to catalog a variable's influence on the “purity” of a model output (Breiman 2001). This can naturally be seen as a VImp measure. Others have extended these approaches to introduce conditional and permutation VImp statistics which aim to reduce the bias that may occur due to variable collinearity (see, e.g., Hothorn et al. 2006).

Conditional variants of permutation variable importance were proposed by Strobl et al. (2008) for a random forest. This method examines splits of the trees in a random forest and permutes the variables within these subgroups (see [Section 2.2](#) for more details). Whereas Strobl et al. (2008) relied on the splitting of trees to determine the subgroups, a model-agnostic approach was introduced by Molnar et al. (2020) that builds the subgroups explicitly from the conditional distribution of the variables. In tree-based models such as CART and random forests, Ishwaran et al. (2010) proposed a VImp called minimal depth, which is the proximity of a variable to the root node, averaged across all trees.

Permutation importance was introduced by Breiman (2001) and is measured by calculating the change in the model's predictive performance after a variable has been permuted. The algorithm works by initially recording the model's predictive performance, then, for each variable, randomly permuting a variable and recalculating the predictive performance on the new dataset. The variable importance score is taken to be the difference between the baseline model's performance and the permuted model's performance when a single feature value is randomly shuffled. A similar agnostic permutation concept was developed by Fisher et al. (2019). This method permutes inputs to the overall model instead of permuting the inputs to each individual ensemble member. In situations where no embedded variable importance is available, a model-agnostic approach such as permutation importance is a useful tool.

In theory any of the above global importance measures could be used in our visualizations. However, providing code for each would be a daunting task. Instead, we take a pragmatic approach and use the associated VImp measure with the model that we are fitting. In cases where there is no such obvious method, we use the Fisher et al. (2019) agnostic permutation approach discussed above to measure VImp.²

2.2. Measuring Variable Interaction

Measuring variable interaction in a machine learning model can be considerably harder than estimating marginal importance. Even the definition of the term “interaction” is disputed (Boulesteix et al. 2015). We focus here on bivariate interaction

²In our implementation, any available VImp may be used.

only, though higher order interactions may certainly be present in many situations. Friedman and Popescu (2008) stated that a function $f(\mathbf{x})$ exhibits an interaction between two of its variables x_k and x_l if the difference in the value of a function $f(\mathbf{x})$ as a result of changing the value of x_k depends on the value of x_l . That is, the effect of one independent variable on the response depends on the values of the second independent variable. Often, an interaction is taken to mean a simple multiplication of two (continuous) variables (e.g., Berrington de González and Cox 2007), though in machine learning models much more complex relationships can exist. We follow the definition of Friedman and Popescu (2008) by considering an interaction to be estimated from the difference between joint and marginal partial dependence; a full mathematical definition is given below. Even this definition should not be used without care, as in the case of highly correlated or potentially confounding variables.

In tree-based models such as CART and random forests, much focus has been on measuring interactions via the structure of trees (e.g., Ishwaran et al. 2010; Deng 2019). If two variables are used as splits on the same branch, then this might initially appear like a measure of interaction. However, this does not separate out the interaction from potential marginal effects. The problem is partially overcome by permuting the variables (individually for a VImp, jointly for VInt), to assess the effect on prediction performance. The resulting VInt measure is known as pairwise prediction permutation importance (Wright et al. 2016).

For models that are not tree-based, or when a model-agnostic measure is required, a variety of other methods can be used. Many of these are based on the idea of partial dependence (Friedman 2000). The partial dependence measures the change in the average predicted value as specified feature(s) vary over their marginal distribution. The partial dependence of the model fit function g on predictor variables S (where S is a subset of the p predictor variables) is estimated as

$$f_S(\mathbf{x}_S) = \frac{1}{n} \sum_{i=1}^n g(\mathbf{x}_S, \mathbf{x}_{C_i}), \quad (1)$$

where C denotes predictors other than those in S , $\{\mathbf{x}_{C_1}, \mathbf{x}_{C_2}, \dots, \mathbf{x}_{C_n}\}$ are the values of \mathbf{x}_C occurring in the training set of n observations, and $g()$ gives the predictions from the machine learning model. For one or two variables, the partial dependence functions $f_S(\mathbf{x}_S)$ are plotted (the so-called PDP) to display the marginal fits.

Friedman's H -statistic or H -index (Friedman and Popescu 2008) is a VInt measure created from the partial dependence by comparing the partial dependence for a pair of variables to their marginal effects. Squaring and scaling gives a value in the range (0, 1):

$$H_{jk}^2 = \frac{\sum_{i=1}^n [f_{jk}(x_{ij}, x_{ik}) - f_j(x_{ij}) - f_k(x_{ik})]^2}{\sum_{i=1}^n f_{jk}^2(x_{ij}, x_{ik})}, \quad (2)$$

where $f_j(x_j)$ and $f_k(x_k)$ are the partial dependence functions of the single variables and $f_{jk}(x_j, x_k)$ is the two-way partial dependence function of both variables, where all partial dependence functions are mean-centered.

The H -statistic requires $O(n^2)$ predicts for each pair of variables, and so can be slow to evaluate. Sampling from the training set will reduce the time, though at a cost of increasing the variance of the partial dependence estimates and the H -statistic.

When the denominator in Equation (2) is small, the partial dependence function for variables j and k is flat, and small fluctuations in the numerator can yield spuriously high H -values. Biased partial dependence curves will also lead to inflated H . This occurs in some machine learning approaches which exhibit regression to the mean in their one-way partial dependencies. Furthermore biased partial dependence curves are a particular problem in the presence of correlated predictors. These issues with the H -statistic seem to be not widely known by practitioners (though see Apley and Zhu 2020), and we provide a short illustration of these problems in the appendix.

In our visualizations throughout this article, we use the square-root of the average un-normalized (numerator only) version of Friedman's H^2 for calculating pairwise interactions:

$$H_{jk} = \sqrt{\frac{1}{n} \sum_{i=1}^n [f_{jk}(x_{ij}, x_{ik}) - f_j(x_{ij}) - f_k(x_{ik})]^2}. \quad (3)$$

This reduces the identification of spurious interactions and provides results that are on the same scale as the response (for regression). It does not, however, remove the possibility that some large H -values arise from correlated predictor variables.

We follow the convention of Hastie et al. (2009) by using the logit scale for both the partial dependence and in calculation of the H -statistic when fitting a classification model with a binary response. If the response is multi-categorical a near-logit is used, defined as

$$g_k(x) = \log[p_k(x)] - \frac{1}{K} \sum_{k=1}^K \log[p_k(x)], \quad (4)$$

where $k = 1, 2, \dots, K$ and $p_k(x)$ is the predicted probability of the k th class. PDPs of $g_k(x)$ from Equation 4 can reveal the dependence of the log-odds for the k -th class on different subsets of the input variables.

Alternatives to the H -statistic have been suggested, which could be used in place of the the H -statistic in our visualizations. Hooker (2004) used a functional ANOVA construction to decompose the prediction function into variable interactions and main effects. Greenwell et al. (2018) suggested a partial dependence-based feature interaction which uses the variance of the partial dependence function as a measure of importance of one variable conditional on different fixed points of another.

2.3. Heatmap Visualization With Seriation

Traditionally, variable importance and interaction are displayed separately, with variable interaction itself spread over multiple plots, one for each variable. We direct the reader to Molnar (2019, chap. 8) for examples. We propose a new heatmap display showing VImp on the diagonal and VInt on the upper and lower diagonals. The benefit of such a display is that one can see which

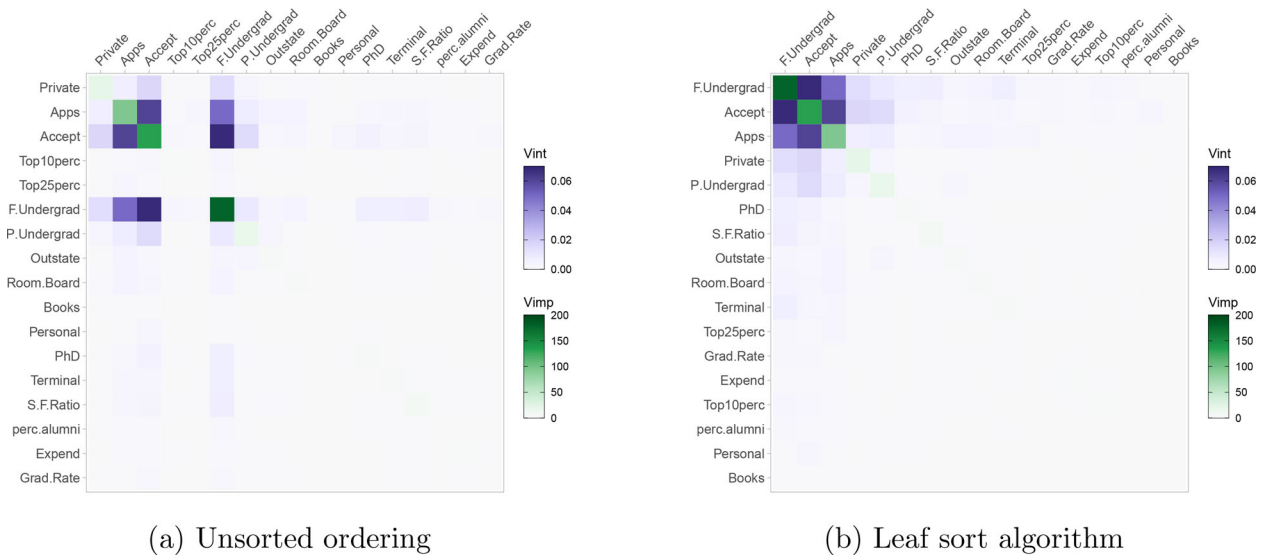


Figure 1. Heatmap from random forest of college application data. In (a) variables are in original order. In (b), the heatmap is re-ordered using leaf sort. In (b) we can see three important and mutually interacting variables, F.Undergrad, Accept and Apps.

variables are important as individual predictors and at the same time see which pairs of variables jointly impact on the response. It also facilitates easy comparison of multiple model fits, which is far less straightforward with separate VImp and VInt displays.

We illustrate the heatmap using a random forest fit to a college applications dataset (American Statistical Association 1995), with Enroll (i.e., the number of new students enrolled) as the response. The data were gathered from 777 colleges across the United States and contains 18 variables ranging from economic factors (such as room and board and book costs) to the number of applications received and accepted. As some of the variables are skewed they are log-transformed prior to building the model. The data was split 70-30 into training and test sets. A value of $R^2 = 0.96$ was obtained for the test set. All plots were made from the training set. See the supplementary materials for a description of the data and transformations.

Figure 1 shows our heatmap with two different orderings. Figure 1(a) has the variables in their original order, while Figure 1(b) uses the leaf-sorting algorithm (described below). The purple color scale used on the off-diagonal shows the Friedman’s H -statistic values (un-normalized) with deeper purple indicating a higher VInt. Similarly, the green color scale on the diagonal represents the level of VImp, here measured using an embedded approach supplied by the random forest (in this case, the increase in node purity). We use colorblind-friendly, single-hued sequential color palettes from Zeileis et al. (2020) going from low to high luminance in both cases, designed to draw attention to high VInt/VImp variables. From the improved ordering in Figure 1(b), there are three clearly important and potentially interacting variables, F.Undergrad (the number of full-time undergraduate students), Accept (the number of applicants accepted), and Apps (the number of applications received), with F.Undergrad having the largest VImp when predicting Enroll.

Many authors have investigated the benefits of re-ordering (also known as seriation) for graphical displays; see, for example, Hurley (2004), Hahsler et al. (2008), and Earle and Hurley

(2015). The benefits of reordering the variables in Figure 1(b) are clear. The right-hand plot lends itself to easy interpretation, whereas the left-hand plot does not.

Most seriation algorithms start with a matrix of dissimilarities or similarities between objects and produce an ordering where similar objects are nearby in the sequence. Our goal here is a little different. As well as placing mutually interacting variables nearby in the sequence, we would like to bring important variables or pairs of variables to the start of the sequence so that the most relevant portion of the heatmap will be in the top-left corner.

We use the leaf sort seriation algorithm from Earle and Hurley (2015). This uses hierarchical clustering followed by a sorting step. Let v_i be a measure of variable importance and s_{ij} be the interaction measure between variables i and j . Treating the matrix of interactions as a similarity matrix, we first construct a hierarchical clustering. This produces a dendrogram, resulting in a variable ordering where high-interacting variables are nearby. Using this ordering in a heatmap generally brings high interactions close to the diagonal, but ignores our goal of placing important variables early in the sequence. For the sorting step, we calculate for each variable a combined measure of its importance and contribution to the interactions, defining these scores as

$$w_i = \lambda_1 v_i + \lambda_2 \max_{j \neq i} s_{ij}.$$

Here, λ_1 and λ_2 are scaling parameters to account for the fact that variable importance and interaction are not measured in the same units. Reasonable choices of λ_1 and λ_2 rescale importance and interaction to, say, unit range or unit standard deviation. We use unit range by default. As there are many possible dendrogram orderings consistent with a hierarchical clustering of the matrix of interactions, the sorting step re-orders the dendrogram leaves so that the weights w_i are generally decreasing.

Sorting the variables in this way will achieve our goals of placing high-interacting pairs of variables nearby in the sequence,

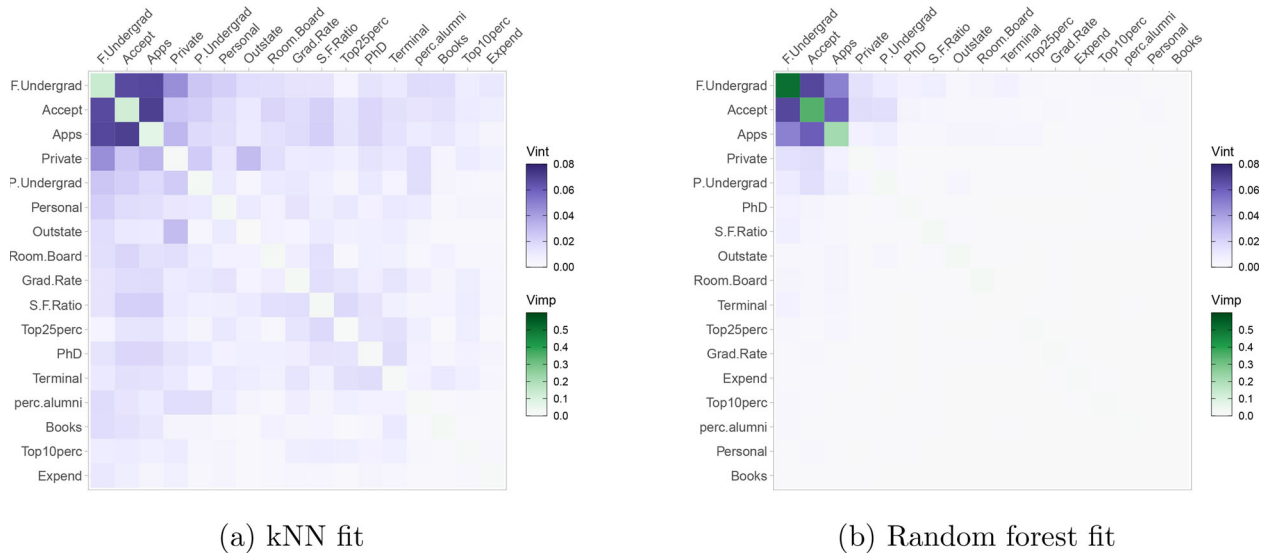


Figure 2. A comparison of a kNN and random forest fit on the college application data. NOTE: Both fits identify F.Undergrad as the most important variable as well as having similar mutual interactions between F.Undergrad, Accept and Apps. The kNN fit identifies many more moderate interactions between variables, especially concerning the variable Private

while simultaneously pulling predictors with high importance and interaction to the top-left of the heatmap, leaving less relevant predictors to the bottom-right. Setting $\lambda_2 = 0$ or $\lambda_1 = 0$ produces plots which sort by descending VImp or max VInt, respectively. For all future heatmap plots, we use the sorting strategy discussed above to optimize the arrangement of variables. After using seriation to re-order the heatmap variables, filtering can be applied to limit the display to the most important or interacting variables; this strategy is especially useful when there are large numbers of predictors.

The heatmap display can be further used to compare different model fits. In Figure 2, we compare the random forest to a k-nearest neighbours (kNN) fit. In the left panel of Figure 2, we have a heatmap of a kNN fit (with $k = 7$ neighbors considered), while the right panel shows the random forest heatmap. To make a direct comparison of the heatmaps, we swap the embedded VImp measures that are available from a random forest fit and instead measure importance with an agnostic permutation approach that allows direct comparison of both the kNN and random forest models. Furthermore, we set both heatmaps to use the same color scale for the VImp and VInt values.

We see in Figure 2 that both the random forest and kNN fit identify F.Undergrad as the most important variable for predicting the number of students enrolled. The top three variables are identical in both models, though the VImp values are much smaller in general across the kNN fit (e.g., the measured VImp for F.Undergrad for the kNN and random forest fits are 0.16 and 0.6, respectively). Both fits show mutual interactions between F.Undergrad, Accept and Apps. However, the kNN fit also suggests a moderate interaction between Private (i.e., whether the university was public or private) and F.Undergrad, which appears somewhat lower in the random forest fit. As Private has a relatively low VImp in both model fits, a simple VImp screening could miss its relevance to the fit. We note though, that this kNN-random forest comparison is for the sake of illustration only, as in this instance the kNN fits poorly by comparison with the random forest, having a test mean square error (MSE) over three times bigger.

2.4. Network Visualization

As our second offering for displaying VIVI, we propose a network plot that shares similar benefits to the heatmap display but differs from it by giving a visual representation of the magnitude of the importance and interaction values not only via color but also by the size of the nodes and edges in a graph. In this plot, each variable is represented by a node and each pairwise interaction is represented by a connecting edge. See Figure 3(a) for an example. The color scales were chosen to match that used in the heatmap, with node size and color luminance increasing with variable importance. Similarly, edge width and color reflects the strength of the VInt. By default we choose a radial layout to display the variables (although this can be changed according to preference) and use the same seriation of variables as the heatmap, with the variables of high importance and high interaction strength placed in a clock-wise arrangement starting at the top. The benefit of such a display is that one can quickly decipher the magnitude of the importance and interactions of the variables as well as seeing which variables both individually and jointly impact on the response.

In Figure 3(a), we again use the random forest fit of the college application data, using the same VImp and VInt measures as in Figure 1. In the network plot, the strong mutual interactions between F.Undergrad, Accept and Apps and are represented by thick, intensely purple lines. F.Undergrad is identified as the most important single predictor and is represented by a large, intensely green node. For settings with large number of predictors, it will be useful to filter the display to focus on high VIVI variables. An additional step groups or clusters the variables according to VImp or VInt values. For example, Figure 3(b) shows a network plot, filtered to display pairs of variables with high VInt and clustered to show groups with mutually similar VInt. Here it is clear that the cluster colored pink contains the variables with the largest VInt scores. In this example, we use hierarchical clustering, but in our implementation, the graph clustering methods provided by the package igraph (Csardi and Nepusz 2006) are directly available.

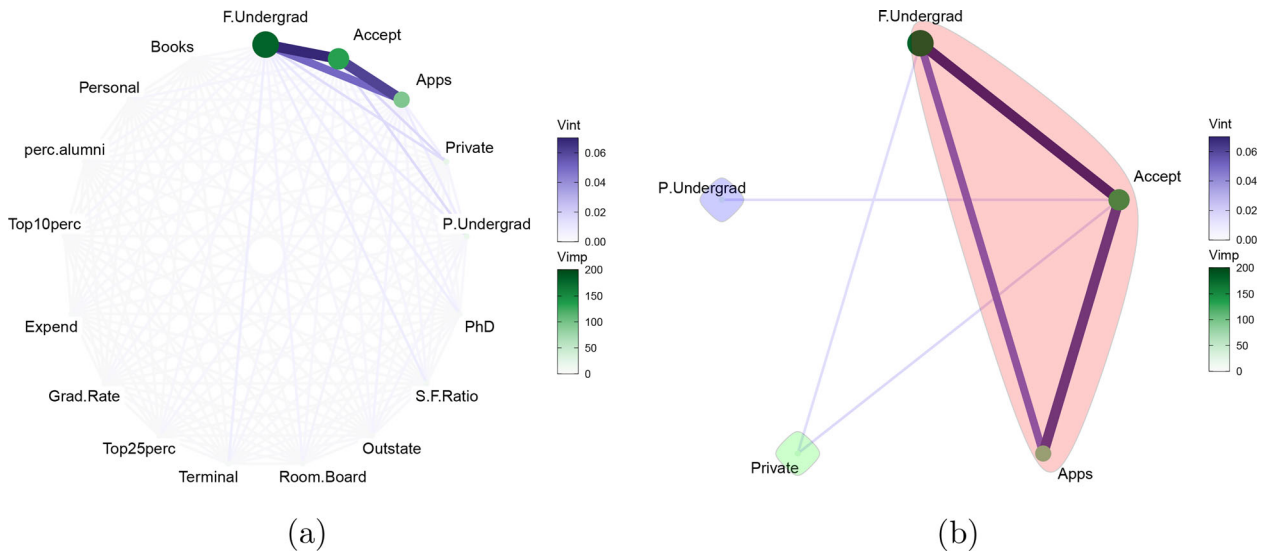


Figure 3. Network plot of a random forest fit on the college application data. Three mutually interacting and important variables can be seen, namely F.Undergrad, Accept and Apps. In (a) all of the variables are displayed. In (b) the network plot has been filtered to display pairs of variables with high VInt and clustered to highlight variables with mutually high VInt.

3. Visualizing Partial Dependence and Individual Conditional Expectation

We introduce new variants of partial dependence and individual conditional expectation plots in two different layouts. With these plots, we can further investigate predictor effects singly and pairwise, especially for those predictors deemed important in our VIVI plots. Additionally, our new plots combine displays of variable pairs, thus highlighting the presence of strong correlations where VInt measures may mislead. Conventionally, PDPs are shown singly or in linear layouts, see Section 8.1 of Molnar (2019) for examples. By comparison, our new displays are more compact, richer, and benefit from seriation.

3.1. Individual Conditional Expectation Curves

Goldstein et al. (2015) described individual conditional expectation (ICE) curves, which are closely related to PDPs. While a PDP shows the average partial relationship between the response and one or two features S , ICE plots display a collection of curves, each showing the estimated relationship between the response and the feature S , at an observed value of other features. Recalling Equation (1), the ICE curves consist of $g(\mathbf{x}_S, \mathbf{x}_{C_i})$ versus \mathbf{x}_S , $i = 1, 2, \dots, n$, while the PDP curve is their average $f_S(\mathbf{x}_S)$. If the ICE curves follow a similar pattern, then the PDP is a useful overall summary, but if the pattern varies, then the feature effect is not homogeneous.

3.2. Generalized Partial Dependence Pairs Plot With ICE Curves

We propose a generalized pairs partial dependence plot (GPDP) with one-way partial dependence and ICE curves with a superimposed partial dependence curve on the diagonal, the bivariate partial dependence on the upper diagonal and scatterplots of raw variable values on the lower diagonal, all of which are colored by the predicted values \hat{y} . Figure 4 provides an example.

With the generalized pairs plot, an analyst can quickly identify which variables singly or jointly impact on the fit. We use a diverging palette so deviations from the average response are emphasized. Here, high values of \hat{y} are shown in dark red and low values are shown in dark blue. Mid-range values are shown in yellow. To avoid interpreting the PDPs where there are no data (and hence potentially spurious H -statistics), we mask out extrapolated areas by plotting the convex hull. For maximum resolution of the bivariate PDPs, the range of the collection of PDP surfaces dictates the limits of the color map. As predictions for individual observations and ice curves are likely to fall beyond these limits, colors are assigned using the closest value in the color map limits.

The ordering of the variables matches that of our heatmap and network plots. The GPDP differs from the previous plots by showing us the distribution of the explanatory variables (lower-diagonal), the exact nature of any linear/nonlinear effects through the use of ICE curves (diagonal), and the average behaviour of the interactions through the use of two-way partial dependence (upper-diagonals). For the ICE curves we have limited the graphic to display a maximum of 30 randomly sampled curves by default, to allow individual ICE curves to be seen. As with the other visualizations, our GPDP can handle both categorical responses and predictors.

Figure 4 shows an example of a GPDP of the college applications data. In the interest of space, we pre-filter this plot to show the seven most influential variables. The bivariate PDPs show the response surface over the convex hull of each variable pair. The lower diagonal plots indicate that F.Undergrad:Accept, F.Undergrad:Apps and especially Accept:Apps are highly correlated with similar increasing marginal effects on the diagonal, suggesting that the high H -values between these variables are likely to be spurious. This is verified by the bivariate linear PDPs for these variables. As Private is a factor with two levels (i.e., yes or no), the partial dependence for each factor level is shown in the upper-diagonal (with yes in red and no in blue). The remaining variables would appear to have little effect either

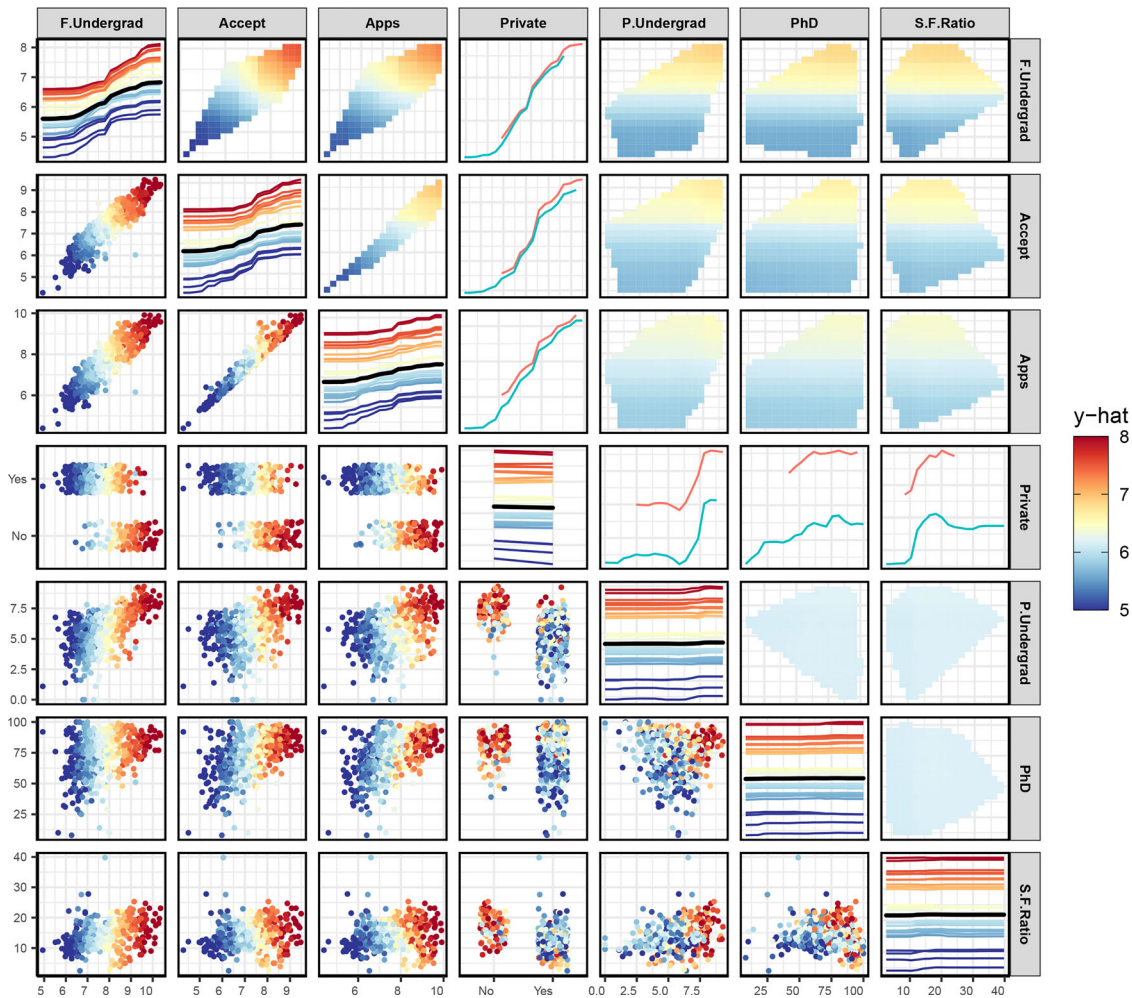


Figure 4. GPDP of a random forest fit on the college data showing the seven most influential variables. From the changing one and two-way partial dependence, we can see that F.Undergrad, Accept and Apps have some impact on the response. However, as they are highly correlated and have similar increasing marginal effects, the potential interactions identified by the H -statistic are likely to be spurious.

singly or jointly on the response. This can be seen from the flat one-way PDP and ICE curves on the diagonal and the flat two-way PDPs.

3.3. Partial Dependence Zenplot

Our final display uses the methods of Hofert and Oldford (2020) to show selected panels of the all-pairs PDP in a space-saving layout, which we call a zen-partial dependence plot (ZPDP). Zenplots (zigzag expanded navigation plots) were designed for showing pairwise plots of high-dimensional data in a zigzag layout. The motivation for zenplots is that they focus on interesting 2D displays, and they permit examination of high-dimensional data. Indeed, Hofert and Oldford (2018) presented an example where they successfully explore pairwise dependence of 465 variables via 164 zenplots. Here, we propose to adapt zenplots for bivariate PDPs.

To describe the construction, consider a network plot showing VImp/VInt such as that in Figure 3(a). Then delete edges with VInt below a threshold, leaving a graph such as that in Figure 3(b). We wish to build partial dependence plots showing pairs of variables with high VInt, that is, visiting each of the edges in our thresholded graph. For a connected graph,

the greedy Eulerian path algorithm of Hurley and Oldford (2011) visits each edge at least once, starting from the highest weighted edge and moving through edges giving preference to the highest-weight available edge. If the graph is not even, then some edges may be visited more than once, or additional edges are visited. If the graph is not connected, then we form sequences for the connected sub-graphs, which are optionally joined into a single sequence.

Zenplots use the zigzag display algorithm of Hofert and Oldford (2020) and allow for the display of high-dimensional data by alternating plot axes in a zigzag-like pattern where adjacent axes share the same variable. We adapt this concept replacing bivariate data plots with bivariate PDPs. As interpretation issues may arise when the distribution of some of the variables is highly skewed, we display a rug plot on each axis to show the distribution of the data. For ease of viewing, the rug plots are a single color and use alpha blending to highlight the distribution. As with our GPDP, there is an option to mask areas where the partial dependence has been extrapolated. The resulting plot displays the most important interacting variables in as small a space as is possible, vastly reducing the number of plots that would be required for interpretation compared with a default matrix scatterplot of PDPs.

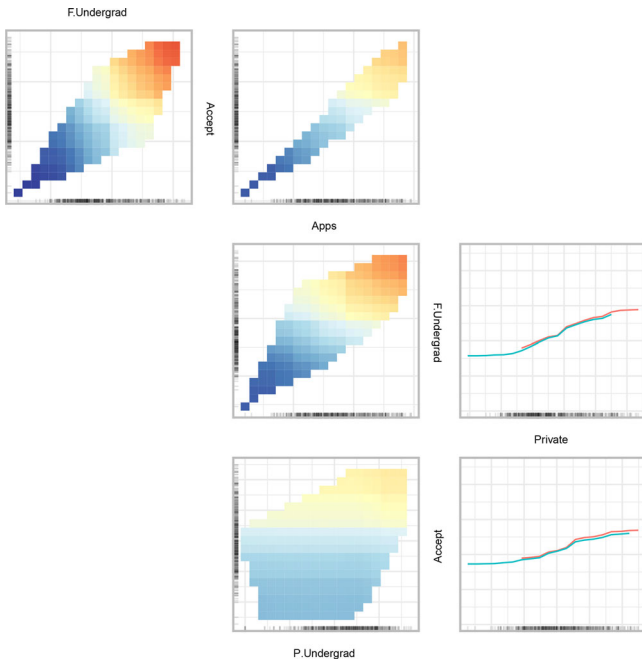


Figure 5. ZPDP of a random forest fit on the college data. We can see that the predicted value for the number of students enrolled increases with each of the variables.

In Figure 5, we show a ZPDP for the random forest fit to the college applications data. The ZPDP shows the bivariate PDPs corresponding to each of the edges of Figure 3. The sequence of plots is obtained from an Eulerian visiting the edges starting with the highest-weight edge, here that is between F.Undergrad and Accept, and following available edges in order of preference by weight thereafter. The resulting Eulerian is F.Undergrad, Accept, Aapps, F.Undergrad, Private, Accept, and P.Undergrad. The plots shown correspond to a subset of those in Figure 4, limited to the more interesting high-interaction pairs. This more compact display helps focus the reader's attention where it is needed, especially as the plots are approximately ordered by decreasing H -index. The variables Private and P.Undergrad show little evidence of marginal importance, and notwithstanding the relatively large H -values, there is not much evidence of interaction with other predictors.

4. Case Study: Cervical Cancer Risk Classification

Cervical cancer remains one of the most prevalent forms of cancer in women globally, ranking fourth in the global cancer incidence in women (Bray et al. 2018). The link between cervical cancer and sexually transmitted diseases (STDs) has been well established. The long-term use of hormonal oral contraceptives is associated with increased risk (Smith et al. 2003). Furthermore, having multiple children has been shown to increase risk (Lukac et al. 2018), particularly in women previously infected with HPV.

Here, we examine and create visualizations for data concerning cervical cancer risk factors (Fernandes et al. 2017). Based on the previous studies, we would expect our visualizations to align with prior identification of important variables, with the addition of gaining new information about how the

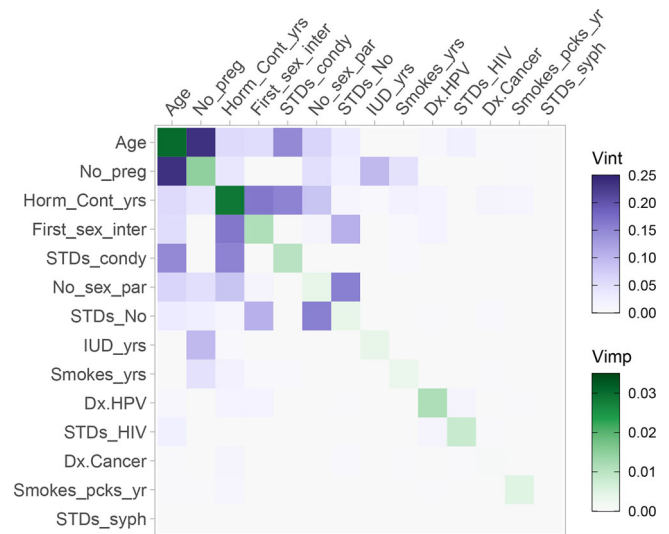


Figure 6. Heatmap of a GBM fit on the cervical cancer data. The first seven variables have the highest VIVI scores. Age and No_preg have the strongest interaction.

variables interact. The data is comprised of historical medical records (such as a patient's STD history, oral contraceptive or intrauterine device [IUD] use) and personal information (such as age and sexual activity). Due to the personal nature of the questions asked for the collection of the data, several patients decided not to answer some of the questions, particularly those concerning STDs. The data have been previously studied (see, e.g., Alsmariy et al. 2020). The full dataset contains 36 variables with 858 observations and uses Biopsy (Healthy or Cancer) as the response.

For this case study, we use a subset of the variables (see supplementary materials for a listing). Preliminary exploration of the data shows that many variables are highly skewed and contain zero values; in this case we use a $\log(x + 1)$ transformation. The data is split 70-30 into training and test sets. We fit a classification gradient boosting machine (GBM) model (Friedman 2000) to the training data, with Biopsy as the response. The accuracy on the test set was measured to be 0.93, and the area under the curve (AUC) was 0.73. All plots were made using the training data, with all PDPs and the H -statistic measured on the logit scale.

Figure 6 displays a heatmap of the GBM fit on the cervical cancer risk data, using a permutation VImp method. Reading from the top-left, the first seven variables have the highest VIVI scores. Overall, Age has the highest importance followed closely by Horm_Cont_yrs (the number of years a patient has taken hormonal contraceptives). This is in agreement with the studies mentioned above. Age also shares the strongest interaction with No_preg (number of pregnancies), which has a medium VImp but is highly important in terms of its interaction. We can see multiple interactions throughout the top seven variables. Of note is the interaction between STDs_No (number of STDs a patient has previously had) and No_sex_par (number of sexual partners). Both of these variables share a strong interaction but have low VImps and they may have been mistakenly eliminated from a model were VImp scores to be used as the sole variable selection metric.

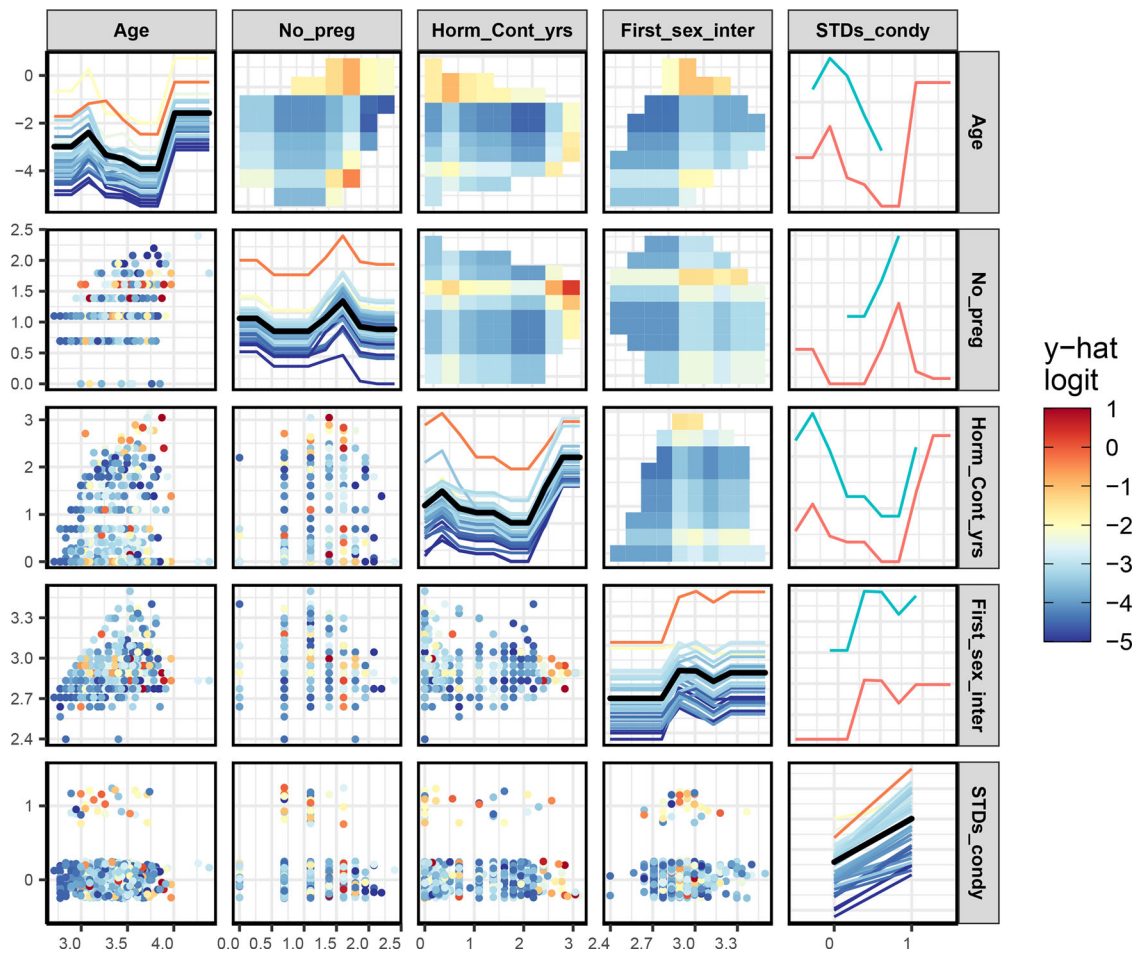


Figure 7. GPDP of a GBM fit on the cervical cancer data. The presence of the STD condylomatosis (STDS_condy) increases the risk of cervical cancer. Risk increases substantially at higher ages and with prolonged use of hormonal contraceptives.

We further explore the impact of the top five variables from Figure 6 on cancer classification in the GPDP plot of Figure 7. To compare response groups, the ICE plots on the diagonal show 25 instances sampled from each of the Cancer and Health groups. The ICE curves are colored according to the predicted log-odds of cancer for that instance. As there is only one red curve, the predicted model accords most observations low cancer probabilities, even for those known to have cancer. The solid black lines on the diagonal of Figure 7 show single variable PDPs. The PDP curve for Age, the single most important predictor, has a mostly decreasing log-odds trend up to an age of 43 (≈ 3.75 on the log scale), with a steep incline thereafter. But we can see from the Age scatterplots there are few cases with ages beyond 43, so the pattern in this area is not supported by much data. The pattern for the Horm_Cont_yrs PDP is similar to that for Age, where log-odds of cervical cancer increases rapidly beyond eight years. In this case though, there are quite a few observations in this region supporting this finding.

According to Figure 6, the predictors No_preg and Age have the strongest interaction. The bivariate PDP plot for No_preg:Age indicates the form of this interaction. A high number of pregnancies is associated with low cancer probability for middle age groups, but is associated with a higher cancer probability for older and, interestingly, younger patients. Note that in the plots with one numeric and one categorical variable, such as the plot for STD_condy (STDs: condylomatosis) and Age, the

numeric variable is always drawn on the x -axis, notwithstanding the label is on the y -axis. This is to allow the plot to be more easily read. In this plot, the bivariate PDP is the same as two PDPs for each level of STDS_condy (where the green curve is for STDS_condy = 1). Although this pair has a relatively high VInt score (as seen in Figure 6), there does not appear to be an interaction present in the bivariate PDP, as the difference between the two curves does not vary with age.

To focus just on predictors with high pairwise interaction scores, we turn to a network plot. Figure 8 displays a network plot of the GBM fit to the cervical cancer risk data, filtered to show pairs of variables with a H -index greater than 0.08 (with the cutoff chosen after inspection of the histogram of H values). The selected variables include the five variables appearing in Figure 7, and three additional variables, namely No_sex_par, STDS_No, and IUD_yrs (number of years with an intrauterine device), with eight relevant interactions between them. This display has some benefits over the heatmap display of Figure 6. First, it focuses directly on pairs of variables with high interaction, particularly with the choice of network layout. Second, in the heatmap plot, even with seriation, some high-interaction pairs of variables may not be positioned nearby which detracts from readability. For example, in Figure 6, associating the relevant variables with the strong interaction for (First_sex_inter, STDS_No) requires considerable effort from the reader.

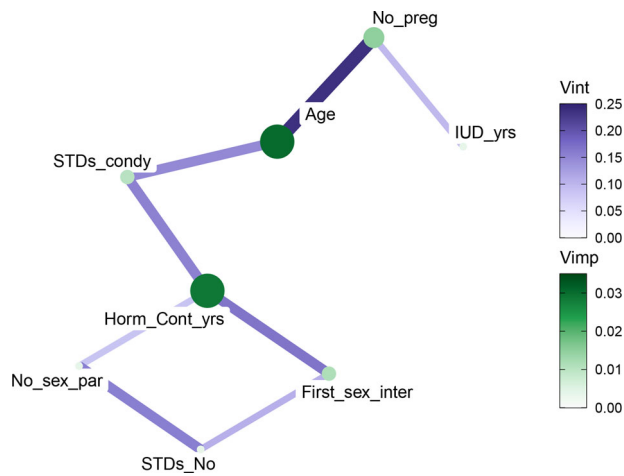


Figure 8. Network graph of a GBM fit on the cervical cancer data, filtered to show pairs of variables with H -index greater than 0.08.

However, this strong interaction is immediately obvious in [Figure 8](#).

To explore these interacting variables further, we use a ZPDP in [Figure 9](#) to show the bivariate PDPs for the eight interactions. The Eulerian path starts with the pair of variables with the highest H -index (here $No_preg:Age$), and from there to $Age:STDs_condy$, ending up at $No_preg:IUD_yrs$. An additional plot is added corresponding to an edge between $Horm_Cont_yrs$ and IUD_yrs to complete the Eulerian. (In this example, it would be possible to construct a ZPDP based

on an Eulerian visiting each edge of the graph in [Figure 8](#) exactly once, but this Eulerian ignores edge weights.) The $STDs_No:No_sex_par$ plot (third row, second column) is a flat surface with no evidence of interaction, despite these variables having a moderate H -index. Interestingly, in the $No_preg:IUD_yrs$ plot (third row, first column), the probability of developing cervical cancer is increasing with IUD_yrs , with a steeper gradient for moderately high No_preg . Further investigation is needed to determine the nature of this effect.

To summarize, we have used our visualizations to identify and examine some clear risk factors associated with developing cervical cancer. Our novel approach allowed us to examine specific pairs of variables that interact and through our use of graphs and PDPs, we can examine how each variable affects the model’s predictions. Specifically, the age of a patient and the number of years of hormonal contraceptive use seem to be important risk factors, agreeing with the previous studies. From [Figure 7](#), the women who took hormonal contraceptives for eight or more years appear to have a higher risk cervical cancer, which is in agreement with the findings of [Smith et al. \(2003\)](#). Surprisingly, as seen in [Figure 6](#), $Dx.HPV$ (i.e., whether the patient has had the previous diagnosis of HPV) was ranked to have middling importance, despite the known link between HPV and cervical cancer. Neither did we see evidence of an interaction between No_preg and $Dx.HPV$, which contrasts with [Lukac et al. \(2018\)](#). These differences may be due to the low frequency of positive cases in the data.

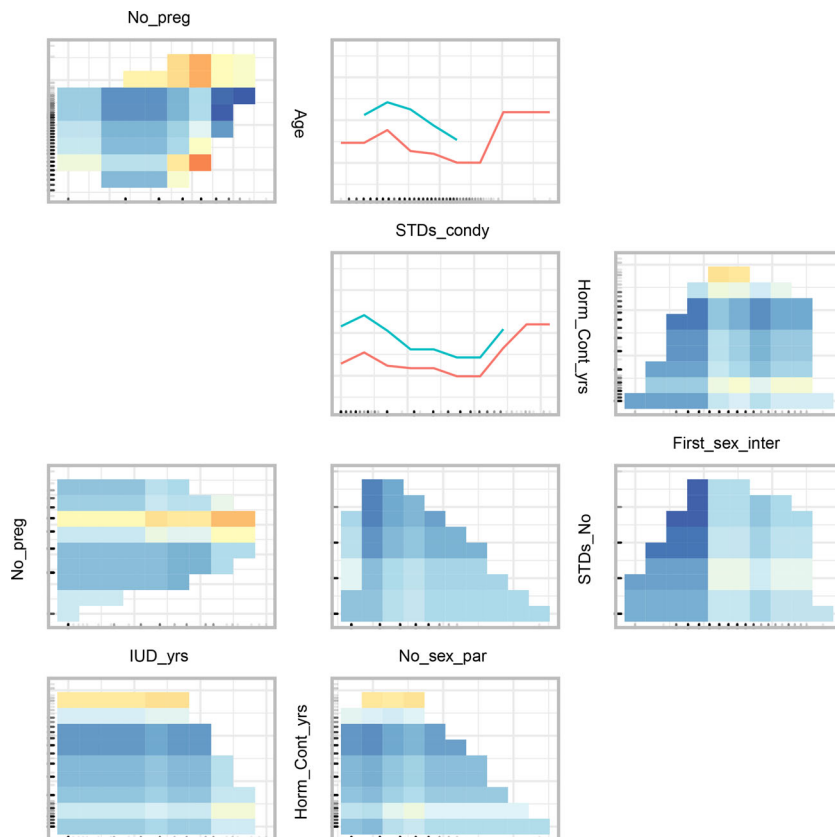


Figure 9. ZPDP of a GBM fit on the cervical cancer risk data. High cancer probability occurs with high number of pregnancies and high age. The color scale is the same as that of [Figure 7](#).

5. Discussion

We have presented innovative and informative methods to visualize the importance and interactions of variables simultaneously from a model. The seriated heatmap of Section 2.3 and the network plot of Section 2.4 are effective in determining which variables have the most impact on the response in a model fit. We view VIVI measures displayed in heatmap and network plots as a starting point for further detailed exploration of the nature of variable effects and interactions in the GPD and ZPD. The ZPD construction is a novel application of the recently proposed zenplots of Hofert and Oldford (2020), which should prove particularly useful to focus exploration on high-VIVI subsets of variables.

Our methods are intuitive, flexible and easily customisable. Built-in or model-agnostic variable importance measures may be used in our heatmap and network displays. In our work to date, we use the model-agnostic H -statistic. Model-agnostic measures are particularly useful when comparing two or more fits. The heatmap and network displays will also be useful for comparing different VIVI measures for the same fit.

As calculation of the VIVI matrix and our visualizations are available for any subset of the data, stratified versions or faceted displays will give insight into higher-order predictor interactions. A drawback to the H -statistic is calculation speed, which is highly model dependent, though sampling and parallel calculation offer useful speed-ups. For example, the 14×14 H -matrix for the GBM fit in Figure 6 computed on 30 randomly selected observations took approximately 16 sec on a MacBook Pro 2.3 GHz Dual-Core Intel Core i5 with 8GB of RAM. Calculation for the 17-predictor random forest fit in Figure 1 is much slower, taking approximately 79 sec, even though here we used just 20 randomly selected observations. A second drawback, we have identified is that high H values can occur in settings where there is no feature interaction, especially in the presence of high variable correlation. The presence and nature of interactions can be further verified in the bivariate PDP, thus avoiding misleading conclusions.

A bivariate importance measure, perhaps obtained by permuting pairs of variables, could be used in place of the H -statistic in the heatmap and network visualizations. It would also

be interesting to explore the interaction measures of Hooker (2004) and Greenwell et al. (2018) in our visualizations, and whether these measures avoid the issues identified with the use of H .

A number of variants of the GPD and ZPD could be investigated in the future work. One possibility for the bivariate PDP, is to subtract the two marginals plotting $f_{jk} - f_j - f_k$, which corresponds directly to the H -statistic. Alternatively, accumulated local effects (ALE) functions (Apley and Zhu 2020) could be used in place of PDPs in our matrix layouts. ALE functions were constructed with the goal of counteracting the bias issues of partial dependence functions. Another option might be to replace the partial dependence f in Equation (3) with the corresponding ALE function, giving a new interaction measure.

Appendix

We explore some limitations of the H -statistic using a simulated dataset. We demonstrate the benefits of the un-normalized version of H , and show how correlated variables can result in spuriously high-interaction measures.

Using the Friedman benchmark equation (Friedman 1991),

$$y = 10 \sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \epsilon \quad (5)$$

where $x_j \sim U(0, 1), j = 1, 2, \dots, 10; \epsilon \sim N(0, 1)$

we simulate 1000 observations and fit a random forest. There are five important variables with an interaction between x_1 and x_2 , and five additional predictors x_6, x_7, \dots, x_{10} unrelated to the response.

In Figure 10(a) and (b), we compare the normalized and un-normalized versions of the H -statistic for the simulated data. Colour legends are not useful here and are omitted. In all cases, the $x_1 : x_2$ interaction is correctly identified. However, in (a) there are numerous spurious strong interactions among the noise variables. In (b) using un-normalized H these spurious strong interactions disappear. The culprit here is the denominator in Equation (2), which for variables x_6, x_7, \dots, x_{10} will be close to zero, thus artificially inflating H . This is the rationale behind our use of the un-normalized H -statistic in our examples throughout.

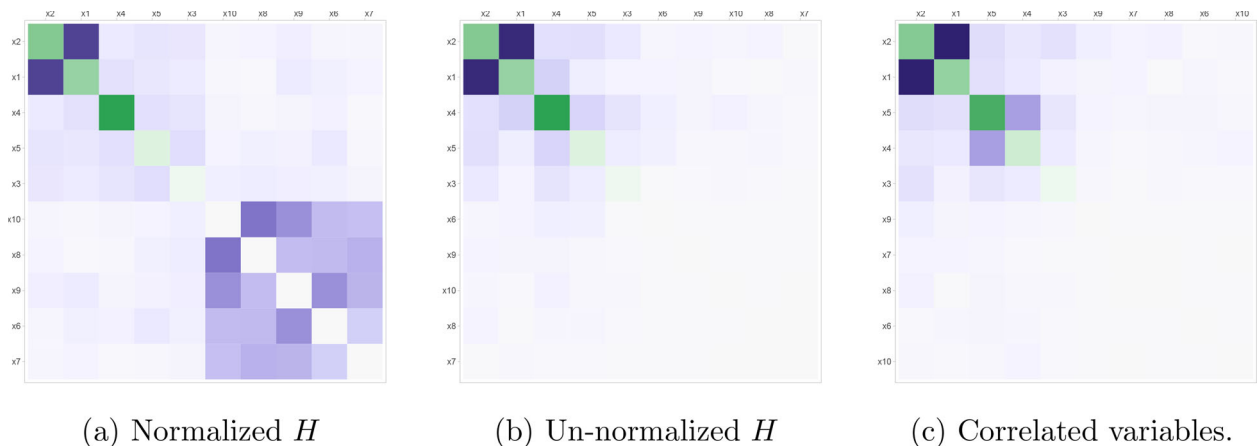


Figure 10. Comparison of the normalized and un-normalized H -statistic and the effect of including correlated variables for a random forest model. In (a) multiple spurious interactions are detected when using the normalized H -statistic. In (b) the spurious interactions have mostly disappeared when using the un-normalized version. In (c) (un-normalized H) a moderate spurious interaction between the correlated variables x_4 and x_5 is detected.

A more subtle cause of spuriously inflated H is due to bias in the partial dependence curve. This is a particular problem in the presence of correlated predictor variables (see, e.g., Apley and Zhu 2020). To demonstrate this, we replace x_5 with $0.3x_5 + 0.7x_4$ in Equation (5) thus introducing a strong correlation (≈ 0.92) between x_4 and x_5 . The resulting VIVI heatmap of the random forest fit in Figure 10(c) shows a moderate x_4 : x_5 interaction which is spurious. Even in the absence of correlation, bias can occur if the fit exhibits regression to the mean. For example, this occurs with tree-based fits such as a random forest, where predictions cannot lie outside the range of training set responses. This bias is evident in Figure 10(b) and (c) as the light purple squares in the top-left section of the heatmaps.

Funding

Alan Inglis and Andrew Parnell's work was supported by a Science Foundation Ireland Career Development Award grant 17/CDA/4695. In addition Andrew Parnell's work was supported by: an investigator award (16/IA/4520); a Marine Research Programme funded by the Irish Government, co-financed by the European Regional Development Fund (Grant-Aid Agreement No. PBA/CC/18/01); European Union's Horizon 2020 research and innovation programme InnoVar under grant agreement No 818144; SFI Centre for Research Training in Foundations of Data Science 18CRT/6049, and SFI Research Centre awards I-Form 16/RC/3872 and Insight 12/RC/2289_P2. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

ORCID

Alan Inglis  <http://orcid.org/0000-0002-1151-6657>

Andrew Parnell  <http://orcid.org/0000-0001-7956-7939>

Catherine B. Hurley  <http://orcid.org/0000-0003-2758-5531>

Supplemental Material

Datasets: Description in datasets.pdf.

College Code for examples of Section 2 and 3 in college.Rmd.

Cervical Code for examples of Section 4 in cervical.Rmd.

hstat: Code for example of Appendix in hstat.Rmd.

References

- Alsmariy, R., Healy, G., and Abdelhafez, H. (2020), "Predicting Cervical Cancer using Machine Learning Methods," *International Journal of Advanced Computer Science and Applications*, 11. [8]
- American Statistical Association (1995), "College Data Contains Statistics for a Large Number of US Colleges from the 1995 Issue of US News and World Report. The dataset was used in the American Statistical Association A.S.A Statistical Graphics Section's 1995 Data Analysis Exposition." [4]
- Apley, D. W., and Zhu, J. (2020), "Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models," *Journal of the Royal Statistical Society, Series B*, 82, 1059–1086. [3,11,12]
- Balli, H., and Sorensen, B. (2010), "Interaction Effects in Econometrics," *C.E.P.R. Discussion Papers, CEPR Discussion Papers*, 45. [1]
- Berrington de González, A., and Cox, D. (2007), "Interpretation of Interaction: A Review," *The Annals of Applied Statistics*, 1, 371–385. [3]
- Boulesteix, A. L., Janitz, S., Hapfelmeier, A., Steen, K. V., and Strobl, C. (2015), "Letter to the Editor: On the Term 'Interaction' and Related Phrases in the Literature on Random Forests," *Briefings in Bioinformatics*, 16, 338–345. [2]
- Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R., Torre, L., and Jemal, A. (2018), "Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries: Global Cancer Statistics 2018," *CA: A Cancer Journal for Clinicians*, 68, 394–424. [8]
- Breiman, L. (2001), "Random Forests," *Machine Learning*, 45, 5–32. [2]
- Csardi, G., and Nepusz, T. (2006), "The igraph Software Package for Complex Network Research," *InterJournal, Complex Systems*, 1695. [5]
- Deng, H. (2019), "Interpreting Tree Ensembles With inTrees," *International Journal of Data Science and Analytics*, 7, 277–287. [3]
- Earle, D., and Hurley, C. (2015), "Advances in Dendrogram Seriation for Application to Visualization," *Journal of Computational and Graphical Statistics*, 24, 1–25. [4]
- Fernandes, K., Cardoso, J. S., and Fernandes, J. (2017), "Transfer Learning With Partial Observability Applied to Cervical Cancer Screening," in *Iberian Conference on Pattern Recognition and Image Analysis*, New York: Springer, pp. 243–250. [8]
- Fisher, A. J., Rudin, C., and Dominici, F. (2019), "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously," *Journal of Machine Learning Research*, 20, 177:1–177:81. [2]
- Frey, H., and Patil, S. (2002), "Identification and Review of Sensitivity Analysis Methods," *Risk Analysis*, 22, 553–578. [2]
- Friedman, J. H. (1991), "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, 19, 1–67. [11]
- (2000), "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, 29. [1,3,8]
- Friedman, J. H., and Popescu, B. E. (2008), "Predictive Learning Via Rule Ensembles," *The Annals of Applied Statistics*, 916–954. [3]
- Friedman, J. H., and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association*, 76, 817–823. [2]
- Gleyzer, S., and Prosper, H. (2008), "PARADIGM, a Decision Making Framework for Variable Selection and Reduction in High Energy Physics," *PoS*. [1]
- Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2015), "Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation," *Journal of Computational and Graphical Statistics*, 24, 44–65. [6]
- Greenwell, B. M., Boehmke, B. C., and McCarthy, A. J. (2018), "A Simple and Effective Model-Based Variable Importance Measure," arXiv:1805.04755. [3,11]
- Hahsler, M., Hornik, K., and Buchta, C. (2008), "Getting Things in Order: An Introduction to the R Package Seriation," *Journal of Statistical Software*, 25, 1–34. [4]
- Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.), Springer Series in Statistics. New York: Springer-Verlag. [3]
- Helton, J. and Davis, F. (2002), "Illustration of Sampling-Based Methods for Uncertainty and Sensitivity Analysis," *Risk Analysis: An Official Publication of the Society for Risk Analysis*, 22, 591–622. [2]
- Hofert, M., and Oldford, W. (2018), "Visualizing Dependence in High-Dimensional Data: An Application to S&P 500 Constituent Data," *Econometrics and Statistics*, 8, 161–183. [7]
- (2020), "Zigzag Expanded Navigation Plots in R: The R Package Zenplots," *Journal of Statistical Software*, 95, 1–44. [1,7,11]
- Hooker, G. (2004), "Discovering Additive Structure in Black Box Functions," *KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 575–580. [3,11]
- Hothorn, T., Hornik, K., and Zeileis, A. (2006), "Unbiased Recursive Partitioning: A Conditional Inference Framework," *Journal of Computational and Graphical Statistics*, 15, 651–674. [2]
- Hurley, C. (2004), "Clustering Visualizations of Multidimensional Data," *Journal of Computational and Graphical Statistics*, 13, 788–806. [4]
- Hurley, C., and Oldford, R. (2011), "Eulerian Tour Algorithms for Data Visualization and the PairViz Package," *Computational Statistics*, 26, 613–633. [7]
- Inglis, A., Parnell, A., and Hurley, C. (2021), *vivid: Variable Importance and Variable Interaction Displays*, R package version 0.1.0. [2]

- Ishwaran, H., Kogalur, U., Gorodeski, E., Minn, A., and Lauer, M. (2010), “High-Dimensional Variable Selection for Survival Data,” *Journal of the American Statistical Association*, 105, 205–217. [2,3]
- Lukac, A., Sulovic, N., Smiljic, S., Ilic, A., and Saban, O. (2018), “The Prevalence of the Most Important Risk Factors Associated With Cervical Cancer,” *Materia Socio Medica*, 30, 131. [8,10]
- Molnar, C. (2019), *Interpretable Machine Learning*, lulu.com. Available at: <https://christophm.github.io/interpretable-ml-book/>. [1,3,6]
- Molnar, C., König, G., Bischl, B., and Casalicchio, G. (2020), “Model-agnostic Feature Importance and Effects with Dependent Features—A Conditional Subgroup Approach,” arXiv:2006.04628. [2]
- Oh, S. (2019), “Feature Interaction in Terms of Prediction Performance,” *Applied Sciences*, 9, 5191. [1]
- Smith, J., Green, J., Berrington de González, A., Appleby, P., Peto, J., Plummer, M., Franceschi, S., and Beral, V. (2003), “Cervical Cancer and Use of Hormonal Contraceptives: A Systematic Review,” *Lancet*, 361, 1159–67. [8,10]
- Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008), “Conditional Variable Importance for Random Forests,” *BMC Bioinformatics*, 9, 307. [2]
- Wang, M., Lo, S.-H., Zheng, T., and Hu, I. (2012), “Interaction-Based Feature Selection and Classification for High-Dimensional Biological Data,” *Bioinformatics (Oxford, England)*, 28, 2834–42. [1]
- Wei, P., Zhenzhou, L., and Song, J. (2015a), “A Comprehensive Comparison of Two Variable Importance Analysis Techniques in High Dimensions: Application to an Environmental Multi-Indicators System,” *Environmental Modelling & Software*, 70, 178–90. [1]
- (2015b), “Variable Importance Analysis: A Comprehensive Review,” *Reliability Engineering System Safety*, 142, 399–432. [2]
- Wood, S. (2000), “Modelling and Smoothing Parameter Estimation With Multiple Quadratic Penalties,” *Journal of the Royal Statistical Society, Series B*, 62, 413–428. [2]
- Wright, M. N., Ziegler, A., and König, I. (2016), “Do Little Interactions Get Lost in Dark Random Forests?” *BMC Bioinformatics*, 17, 145. [3]
- Zeileis, A., Fisher, J. C., Hornik, K., Ihaka, R., McWhite, C. D., Murrell, P., Stauffer, R., and Wilke, C. O. (2020), “colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes,” *Journal of Statistical Software, Articles*, 96, 1–49. [4]