Searching for Dusty Corners: Understanding the Prediction of the Cross Section of Returns

Carlos Carvalho, John Cochrane, Juhani Linnainmaa, Rob McCulloch

1. Goals
2. Predictability
3. Fit-the-Fit, Where are the nonlinearities and interactions ??

1. Goals

Predict

Simple approach to predicting the monthly cross section of firm returns using variables obtained in the previous month.

Focus on methods that could work well with little tuning:

- linear
- simple trees
- random forests
- BART

Our goal is not to get the ultimate return predictor.

Our goal is is to use relatively straighforward predictive methods than can capture nonlinearities and then interpret the results.

Interpret

fit the fit

Data: (x,R), R: Return on a given firm for a given month, x: characteristics of the firm, preceeding month.

Summarize $E(R) = \hat{f}(x)$ by searching for simple fits of the fit.

In this talk we fit a simple tree to the fit, that is our data is $(x, \hat{R})$ where $\hat{R}$ is an in-sample fit or a prediction.

Carvalho and Hahn have emphasized the effectiveness of the fit-the-fit approach in a series of recent papers.

See Carvalho, Hahn, McCulloch for variable selection.

Variable selection:
Can I find a function of a subset of the variables x_S so that $g(x_S)$ approximates $\hat{f}(x)$ well.

Dusty Corners:

We think there are small parts of the predictor space where interesting nonlinearities kick in.

We will try to indentify variables that contribute to nonlinearity and interactions in the dusty corners.

For example, a dusty corner corresponds the prediction of unusually small returns.

Interesting non-linearites kick in for extreme returns, not so much in the middle.

We just want simple ways of seeing this.

Data:

```
> length(dates)
[1] 1253753
> dates[1]
[1] 196306
> dates[1253753]
[1] 202005
```

- 684 months of data, June 1963 to May 2020.
- Each month we have a cross section of firm returns, and 62 firm characteristics measured in the previous month.
- 1,253,753 total observations on a firm return and vector of characteristics.
- threw out "tinies"
- on a monthly basis express each x as a quantile in $(0, 1)$.
- regression imputation of missing values
- monthly demean returns, so we are predicting amount above average

```
> dim(ddf)
[1] 1114198      63

> names(ddf)
 [1] "accruals"      "assetgrowth"   "BE/ME"         "cashflowtoequ"
 [5] "aturnoverchg"  "debtissuance"  "earningsprice" "EM"
 [9] "grossprofitab" "inventorygrow" "herfindahl"    "netoperatinga"
[13] "piotroskiF"    "abnormalinves" "leverage"      "accruals_beme"
[17] "netwcap_chang" "oscore"        "profitmargin"  "profitability"
[21] "returnonequit" "salesgrowth"   "salestoprice"  "sustainablegr"
[25] "total_xfin"    "zscore"        "indadjcapxgro" "salesminusinv"
[29] "investmenttoc" "invgrowthrate" "I/A"           "qmj_profitabi"
[33] "chs_distress"  "ffprofitabili" "organizationc" "advertising"
[37] "opleverage"    "rd"            "tax"           "Profitability"
[41] "dssdur"        "disttohigh"    "amihud"        "marketbeta"
[45] "firmage"       "Volatility"    "ME"            "LT_reversals"
[49] "maxdailyret"   "r_12,2"        "intmomentum"   "Price"
[53] "seasonality"   "seasonality_o" "ST_reversals"  "Volume"
[57] "divmonth"      "sharevolume"   "coskewness"    "divyield"
[61] "momreversal"   "BE/ME*"        "R"
```

Some Key Predictor Variables

For example, these variables turn out to be particularly interesing:

ST_reversals:
prior one month return. "short term reversals".

r_12,2:
prior one year return, skipping a month. "momentum effect".

Volatility

disttohigh

Panel B: List of return predictors

Category	Predictor	Category	Predictor
Investment,	Asset growth	Price-scaled	Book-to-market
growth, and	Inventory growth	predictors	Monthly book-to-market
duration	Sales growth		Cash-flow to equity
	Sustainable growth		Enterprise multiple
	Ind.-adjusted CAPX growth		Sales to price
	Growth in sales-inventory		
	Investment to capital	Financing	Total external financing
	Investment growth rate	and	Dividend month
	Investment to assets	payouts	Dividend yield
	Abnormal investment		Debt issuance
	Equity duration		
		Expenses	Advertising
Operational	Accruals		R&D
efficiency,	Change in asset turnover		Taxes
earnings	Accruals and book-to-market		
quality, and	Changes in net working capital	Momentum,	Distance to high
industry	Net operating assets	reversals,	Long-term reversals
	Industry concentration	and	Maximum daily return
		seasonality	Momentum
Profitability	Profit margin		Intermediate momentum
	Return on assets		Seasonality
	Return on equity		Seasonal reversals
	Gross profitability		Short-term reversals
	Earnings to price		Momentum and reversals
	QMJ profitability		
	Operating profitability	Other price	Amihud's illiquidity
	Cash-based profitability	and volume	Market beta
			Firm age
Distress	Z-score		Idiosyncratic volatility
and	O-score		Firm size
leverage	Financial distress		Nominal price
	Piotroski's F-score		High-volume return premium
	Operating leverage		Share volume
	Leverage		Coskewness

R: cross section of returns, each month t, each firm in that month.
x: predictor variables used for R (measured at time $t-1$).

Approach:

Our overall approach is the following:

- For each month t fit a model giving $\hat{R} = \hat{f}_t(x)$.
- Roll the fitted models: $\hat{f}_t^P(x) = \sum_{j=1}^{\nu} w_j \, \hat{f}_{t-j}(x)$.
- Check that $\hat{f}_t^P(x)$ has reasonable predictive performance.
- Inspect $\{\hat{f}_t^P\}$ to learn about the relationship, (e.g., what variables are used).
- Also consider $\hat{f}^A(x) = \frac{1}{N} \sum_{t=1}^{N} \hat{f}_t(x)$.

For example, we often use $\nu = 120$ (10 years), $w_j = 1/120$.

Choice of "Learner"

We have to fit a model each month so we want to use approaches that do not require a lot of tuning. In addition, our x variables are "messy" so we need methods that perform well in this case.

We focus on methods based on trees and ensembles of trees:

- Trees are capable of uncovering any kind of non-linearity and interaction.
- Trees handle messy x variables: they are invariant to monotonic transformations of the predictor variables.
- Single trees partition the x space into rectangular subsets somewhat reminiscent of what you obtain by sorting stocks into portfolios
- Ensembles of trees, in which many trees are combined to get an overall fit, are the best "off-the-shelf" models.
- We will use Random Forests and BART (Bayesian Additive Regression Trees) which is an ensemble method related to boosting. Generally, BART requires less tuning than other boosting type approaches. Random Forests is well known for performing well with minimal tuning.

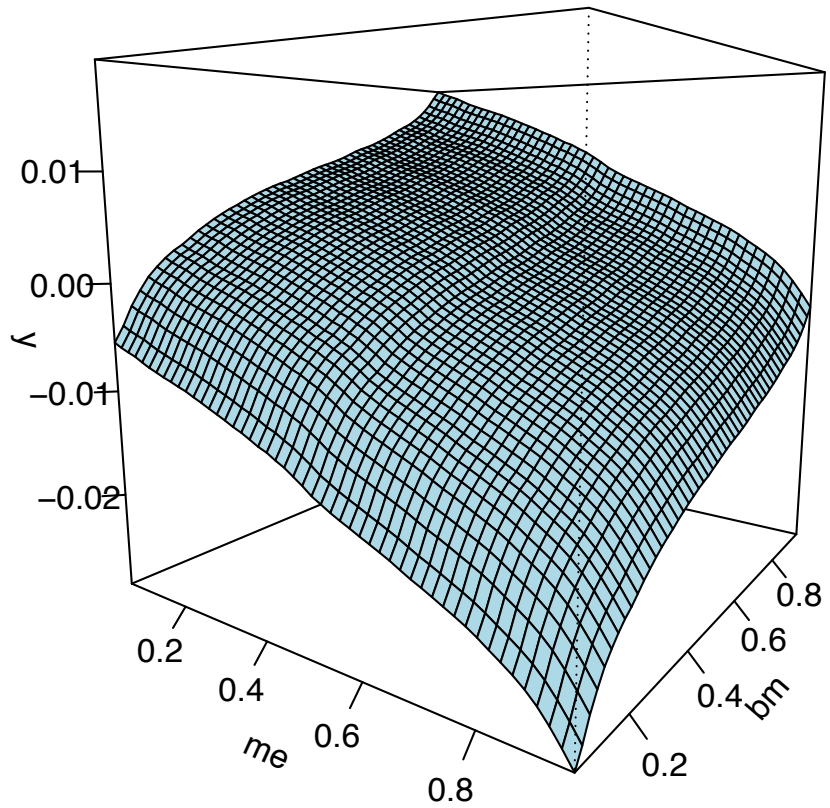
We ran default BART and default random forests.

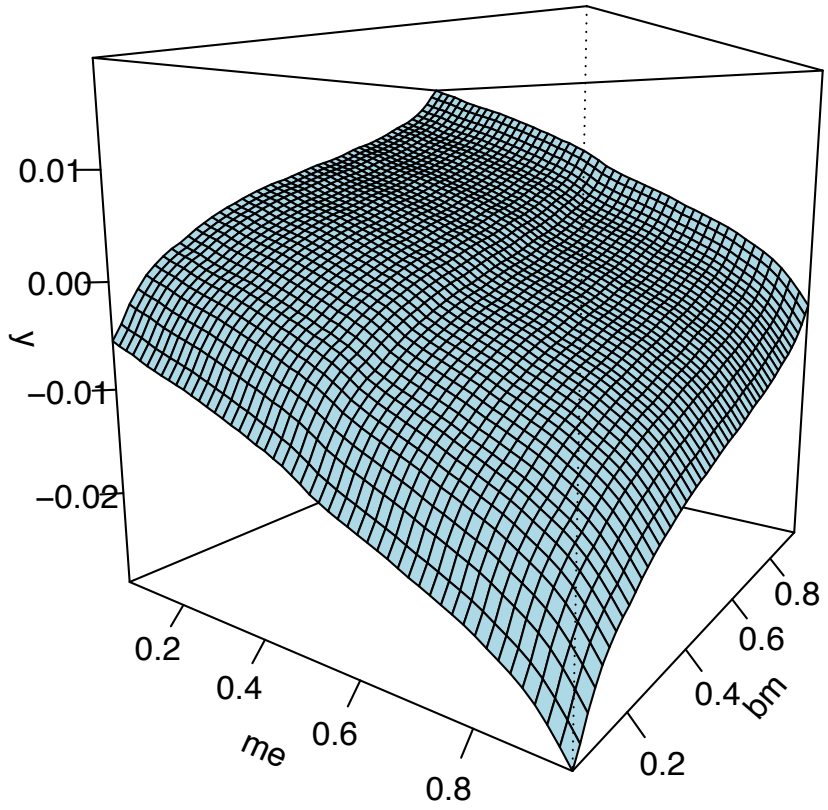
Our goal is to have some understanding of what the non-linear fitted relationship is.

With a two-dimensional x, we can plot.

E(R) vs
x1: me = market equity
x2: bm =book-to-market.

Hard in high dimensions!!!





- Looks pretty linear for most of the middle of the x=(me,bm) predictor space.
- big me and small bm really interact to give you low returns.
- A little non-linear upturn for big bm, especially at small me.
  At big me, small bm , there is a *dusty corner*.
- nonlinearity for large bm across a range of return values.

Note:

Most of the methods could be used with estimates of $E(R \mid x)$ from any learner.

For example, Gu, Kelly and Xiu have some interesting results with neural nets.

Most of our results just examine the fit $E(R \mid x)$, but we are working on capturing the uncertainty.

2. Predictability

Is there any predictive ability?

Are the Machine Learners any better than linear?

Stacked Correlations

Stack all the R for each month and all the out-of-sample $\hat{R}$ for each month and compute the simple pearson correlations.

rf is Random Forests.
bart is BART.

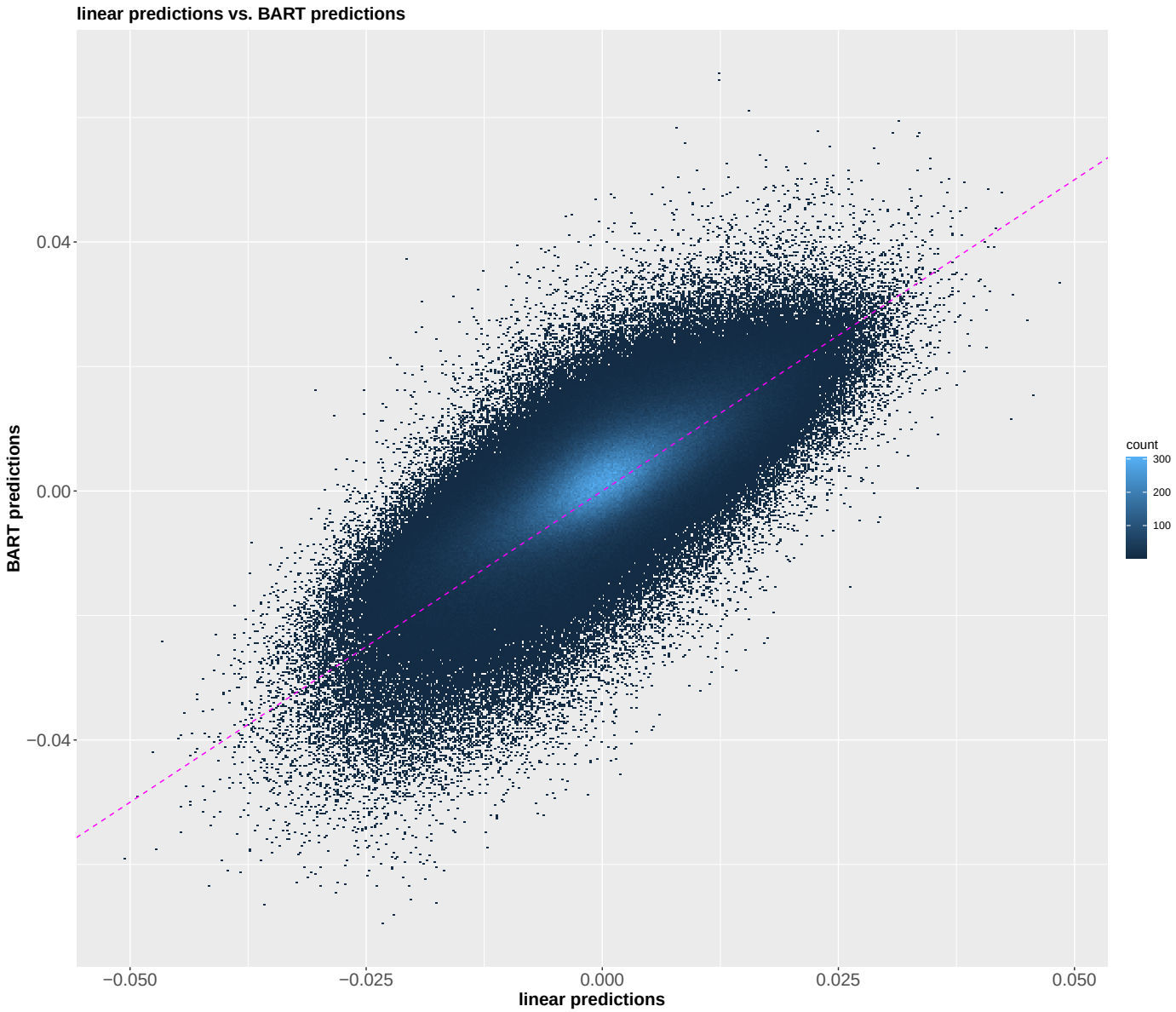
*15 uses just 15 variables we got from our variable selection.

*3 just uses "ME" "BE.ME" "r_12.2"

```
> round(cor(omat),3)
              oR olinear3  orf3 obart3 olinear   orf obart olinearv15 orfv15 obartv15
oR         1.000    0.027 0.024  0.027   0.048 0.040 0.054      0.046  0.051    0.056
olinear3   0.027    1.000 0.781  0.809   0.461 0.475 0.377      0.578  0.494    0.444
orf3       0.024    0.781 1.000  0.899   0.371 0.465 0.369      0.456  0.496    0.441
obart3     0.027    0.809 0.899  1.000   0.393 0.456 0.393      0.479  0.487    0.469
olinear    0.048    0.461 0.371  0.393   1.000 0.609 0.753      0.889  0.704    0.737
orf        0.040    0.475 0.465  0.456   0.609 1.000 0.698      0.631  0.805    0.672
obart      0.054    0.377 0.369  0.393   0.753 0.698 1.000      0.713  0.737    0.787
olinearv15 0.046    0.578 0.456  0.479   0.889 0.631 0.713      1.000  0.767    0.791
orfv15     0.051    0.494 0.496  0.487   0.704 0.805 0.737      0.767  1.000    0.851
obartv15   0.056    0.444 0.441  0.469   0.737 0.672 0.787      0.791  0.851    1.000
```

BART is more like linear and rf!!??

BART predictions compared to linear:



We looked at various ways to check there is some "BART predictability".

```
> lmf = lm(oR ~ olinear + obart + orf,omatDf)
summary(lmf)

Call:
lm(formula = oR ~ olinear + obart + orf, data = omatDf)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0820 -0.0563 -0.0038  0.0509  6.3056

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0008179  0.0001372  -5.960 2.52e-09 ***
olinear      0.2212042  0.0196362  11.265  < 2e-16 ***
obart        0.4710455  0.0193945  24.288  < 2e-16 ***
orf          0.0462427  0.0231940   1.994   0.0462 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1153 on 1114194 degrees of freedom
Multiple R-squared:  0.002995,Adjusted R-squared:  0.002993
F-statistic:  1116 on 3 and 1114194 DF,  p-value: < 2.2e-16
```

Juhani:

```
The results show that adding the BART factor always helps.
This is not too surprising given the results in (1)
| those results show that the BART factor
contains information not found in the other factors.
```

Rememer, out goal is not to find *the* killer return predictor, but to simply find a reasonable one.

3. Fit-the-Fit

We want to understand the function $\hat{R} = \hat{f}(x)$.

We create the data

$$(x, \hat{R} = 100\,\hat{f}(x))$$

Usually, we use all the x from all observations but we will also consider using "dusty corner" x, which give high or low returns.

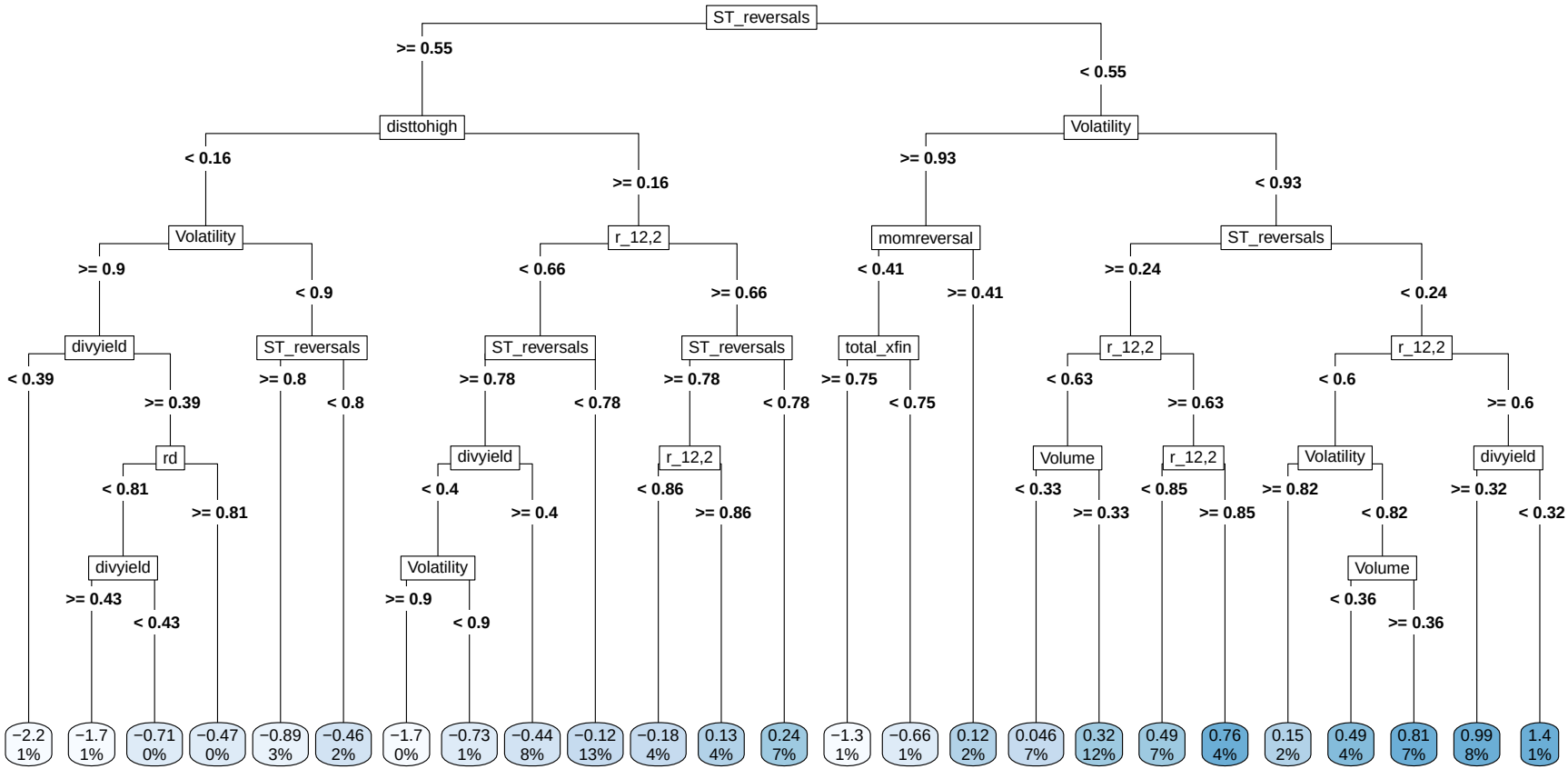
$\hat{f}$ could be $\hat{f}^P$ or $\hat{f}^A$.

For $\hat{f}^P$ we have 1,114,198 $(x, \hat{f})$ pairs, for $\hat{f}^A$ we have 1,253,753.

We use all 62 x predictor variables or just a subset of 15 we found by Carvalho, Hahn, McCulloch variable selection.

We then fit a single tree to this data and use the tree to interpret $\hat{f}$.

$\hat{f}^P$, Fit-fit tree with 25 bottom nodes.
Recall that each predictor is in (0,1).



Dusty corners:

```
-2.201 when
ST_reversals >=  0.55 & Volatility >=  0.90 & disttohigh <  0.16 & divyield <  0.39

1.417 when
ST_reversals <  0.24 & Volatility <  0.93 & r_12,2 >= 0.60 & divyield <  0.32
```

visualizing the tree:

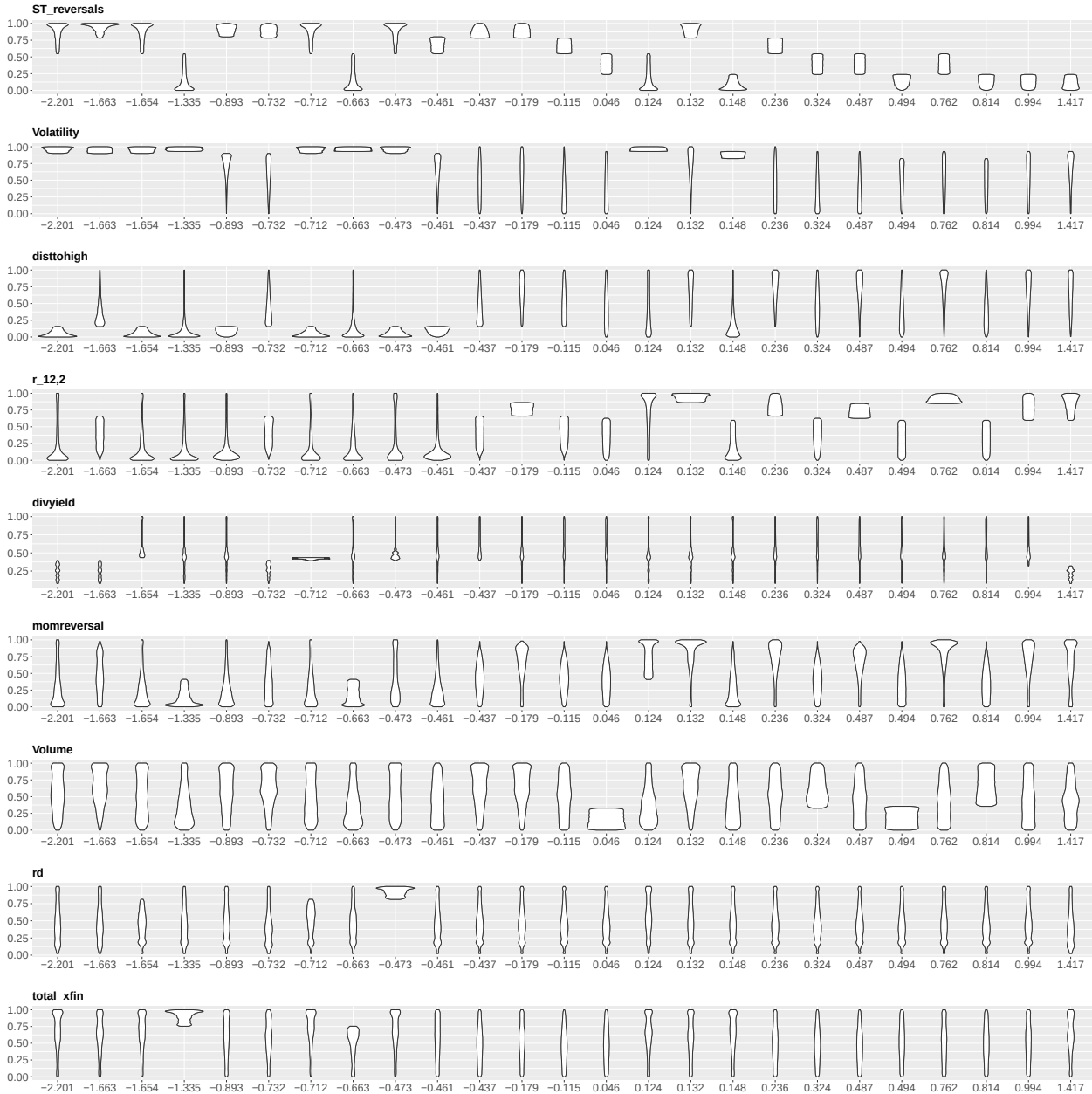
$\hat{f} = \hat{f}^P$,
Fit-fit tree with 25 bottom nodes.

Each row corresponds to a predictor used in the tree.
Only 9 of the 62 variables are used.

Each columns corresponds to a bottom node of the tree.

Each violin depicts the distribution of the predictor in a bottom node.

Bottom node labeled by mean of 100 $\hat{f}$ over observations in that bottom node.



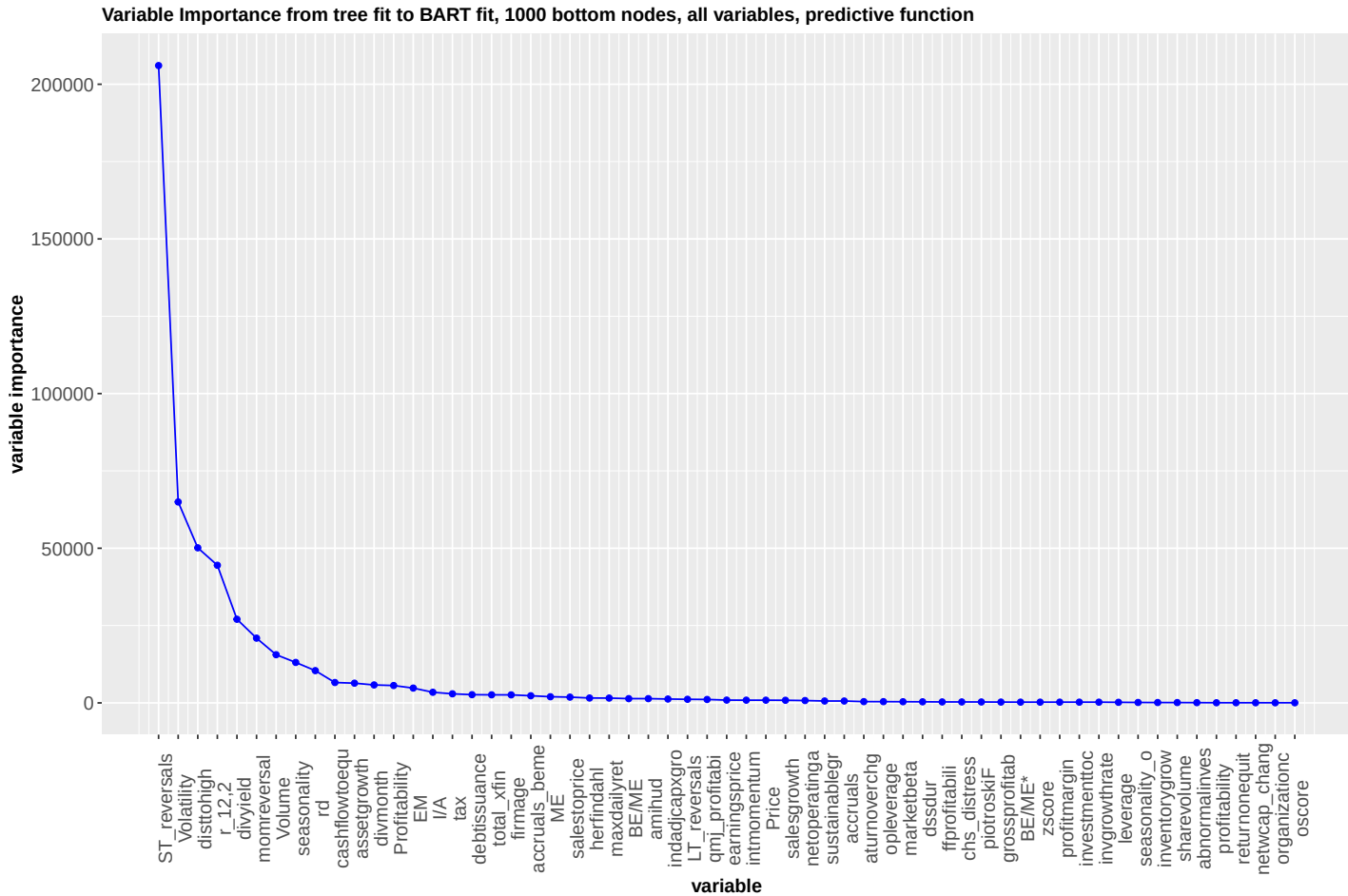
visualizing the tree:

Of course the tree with just 25 bottom nodes may be a crude summary of the fit.
Let's look at variable importance from a tree with 1,000 bottom nodes.

$\hat{f} = \hat{f}^P$

Fit-fit tree with 1,000 bottom nodes.

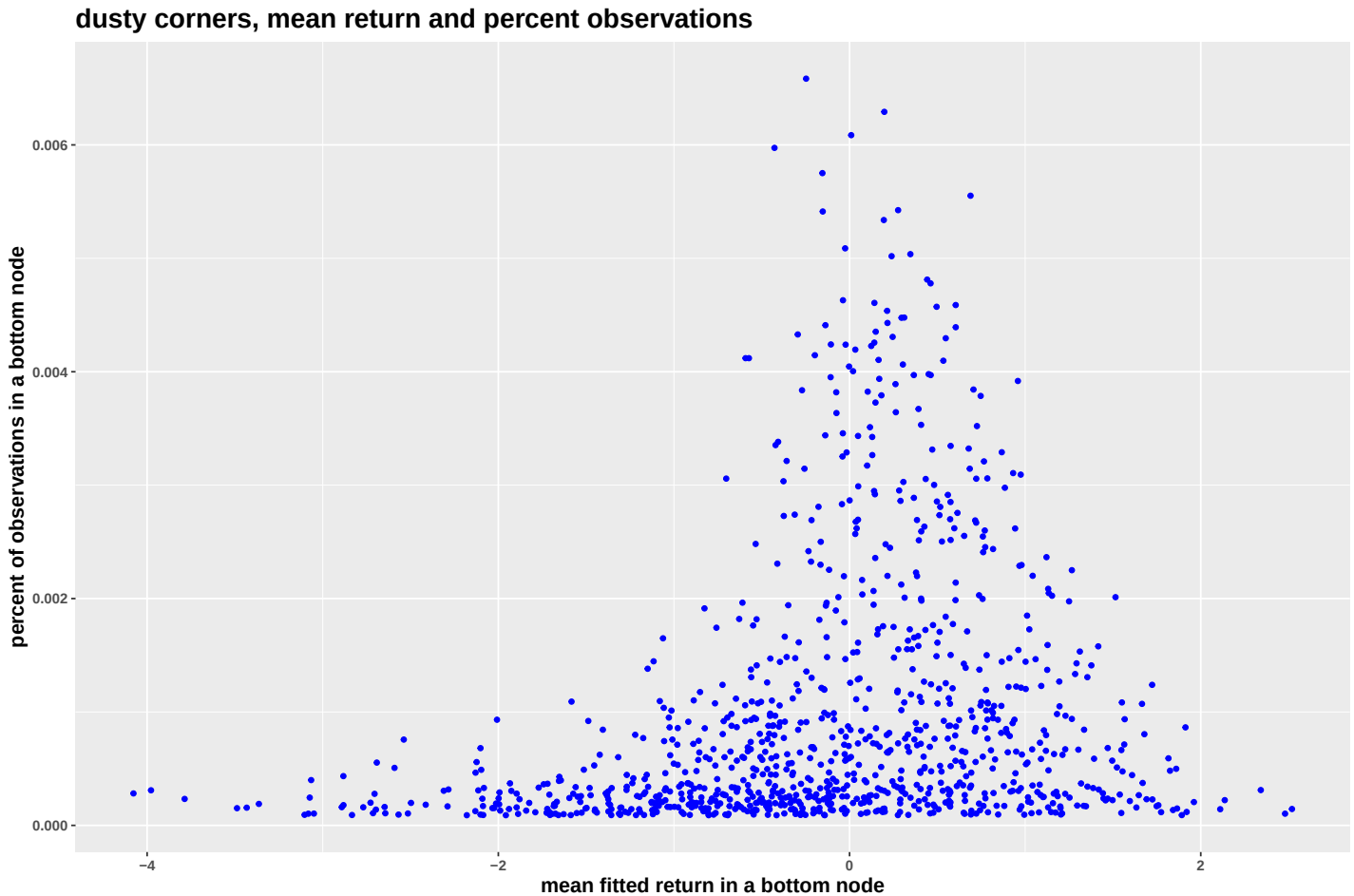
Variable importance.



Dusty corners:

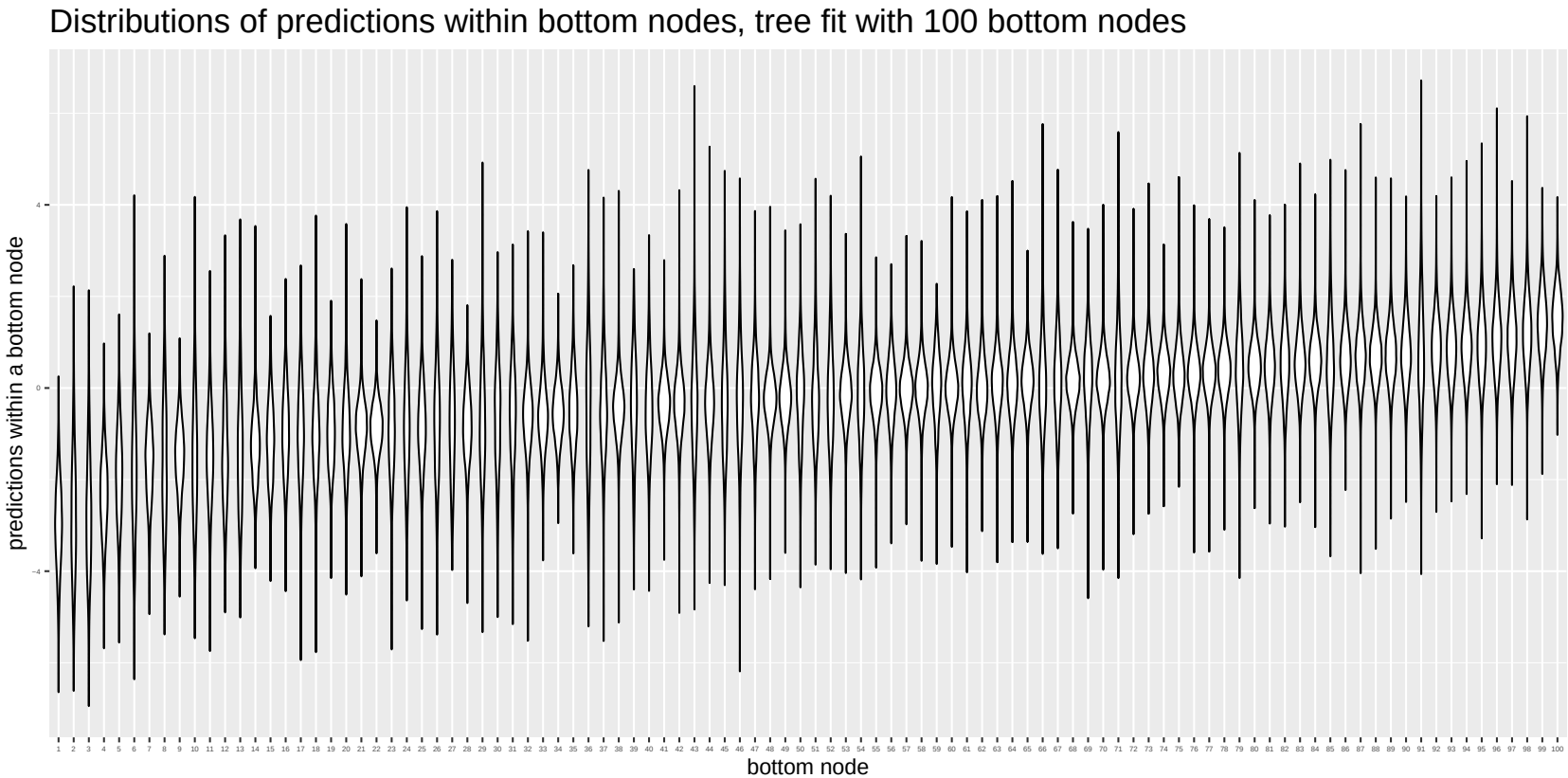
$\hat{f} = \hat{f}^P$, tree with 1,000 bottom nodes.

For each bottom node get the mean of $\hat{f}(x)$ and percent observations in the bottom node.

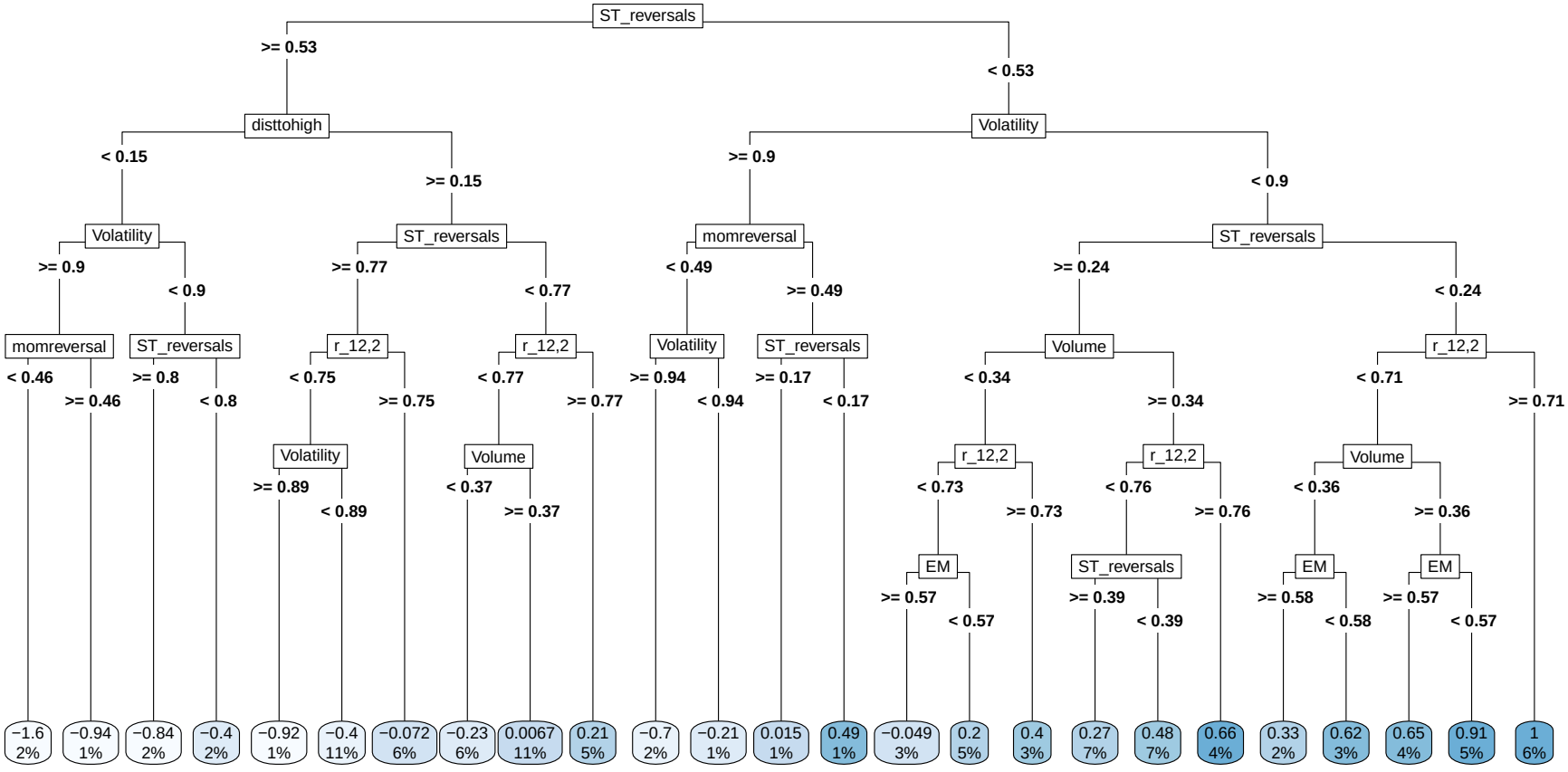


Distribution of 100 $\hat{f}^P(x)$ in each bottom node.

Tree with 100 bottom nodes.



$\hat{f}^A$, Fit-fit tree with 25 bottom nodes.
Recall that each predictor is in (0,1).



Dusty corners:

```
-1.5804 when
ST_reversals >= 0.53 & Volatility >= 0.90 & disttohigh <  0.15 & momreversal <  0.46
1.0083 when
ST_reversals <  0.24  & Volatility <  0.90  & r_12,2 >= 0.71
```

visualizing the tree:

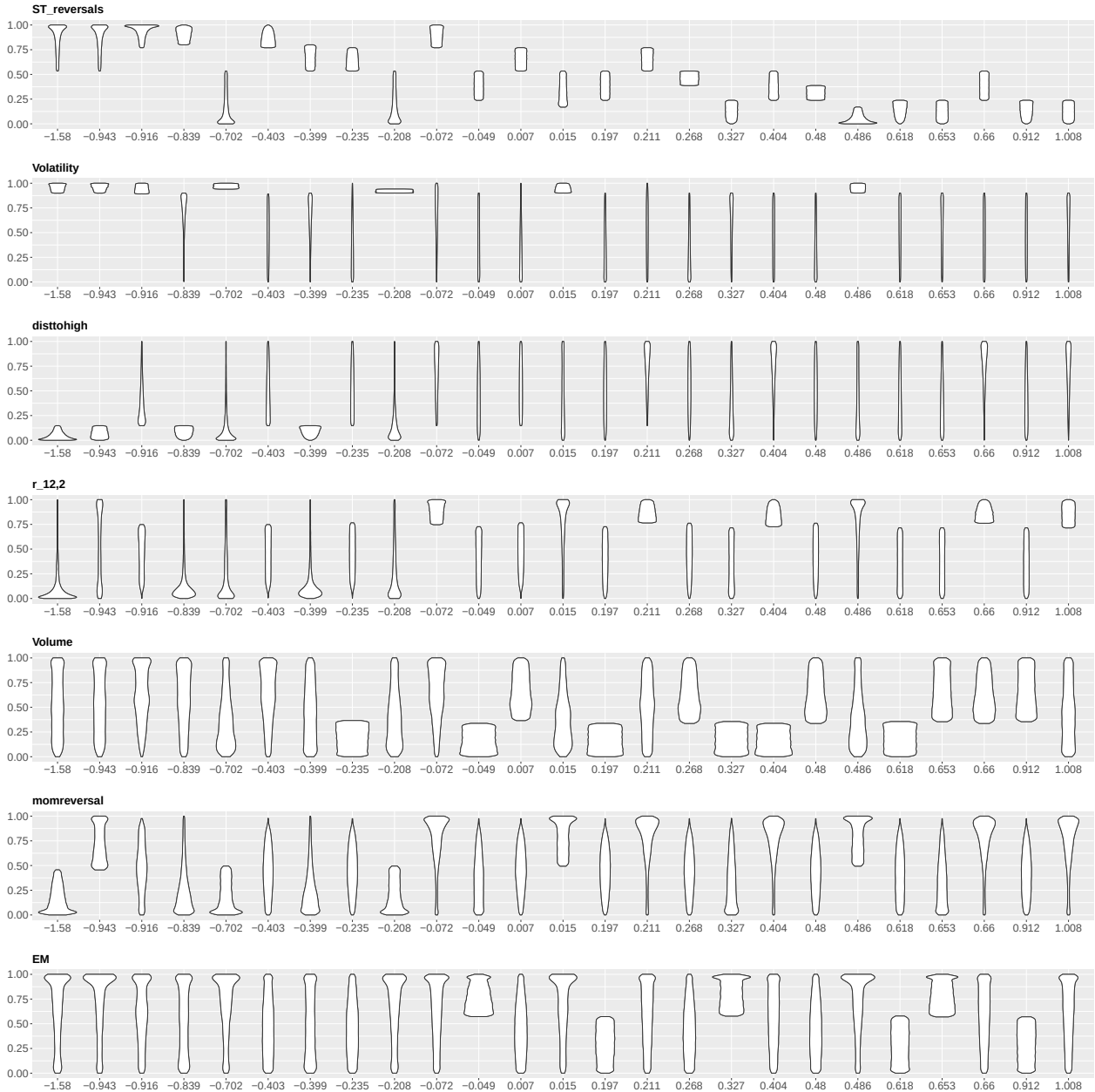
$\hat{f} = \hat{f}^A$,
Fit-fit tree with 25 bottom nodes.

Each row corresponds to a predictor used in the tree.
Only ?? of the 62 variables are used.

Each columns corresponds to a bottom node of the tree.

Each violin depicts the distribution of the predictor in a bottom node.

Bottom node labeled by mean of 100 $\hat{f}$ over observations in that bottom node.

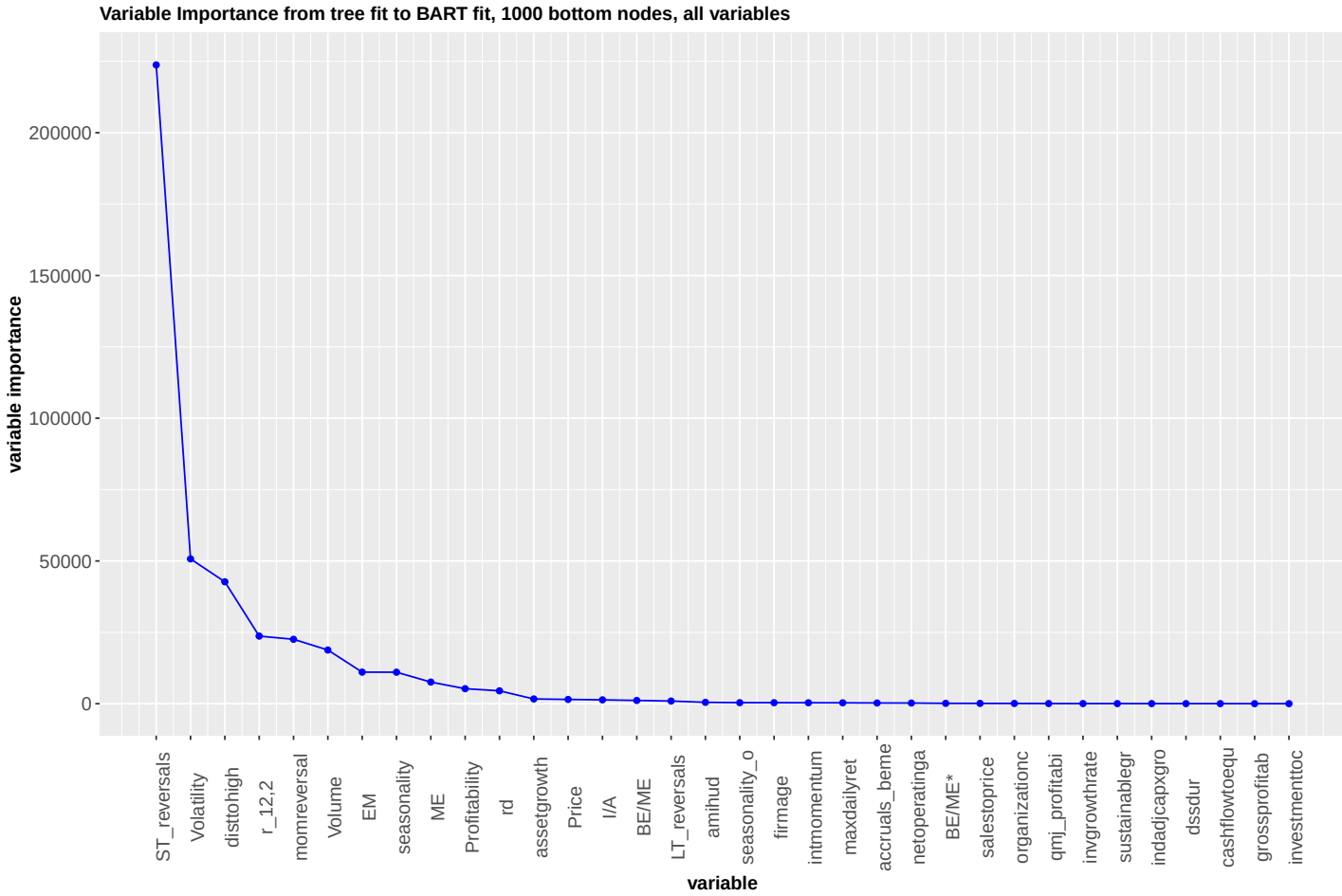


visualizing the tree:

$\hat{f} = \hat{f}^A$

Fit-fit tree with 1,000 bottom nodes.

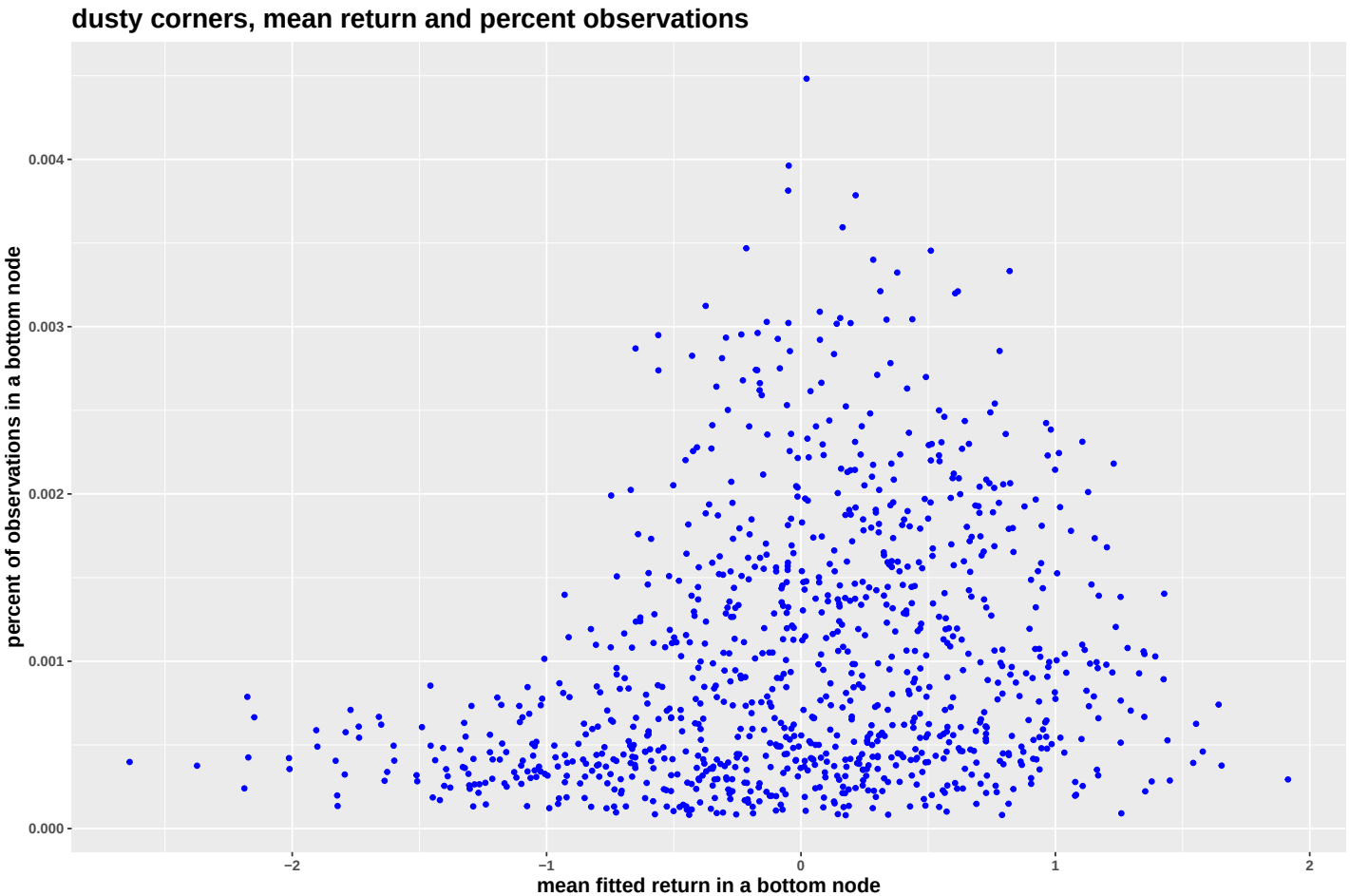
Variable importance.



Dusty corners:

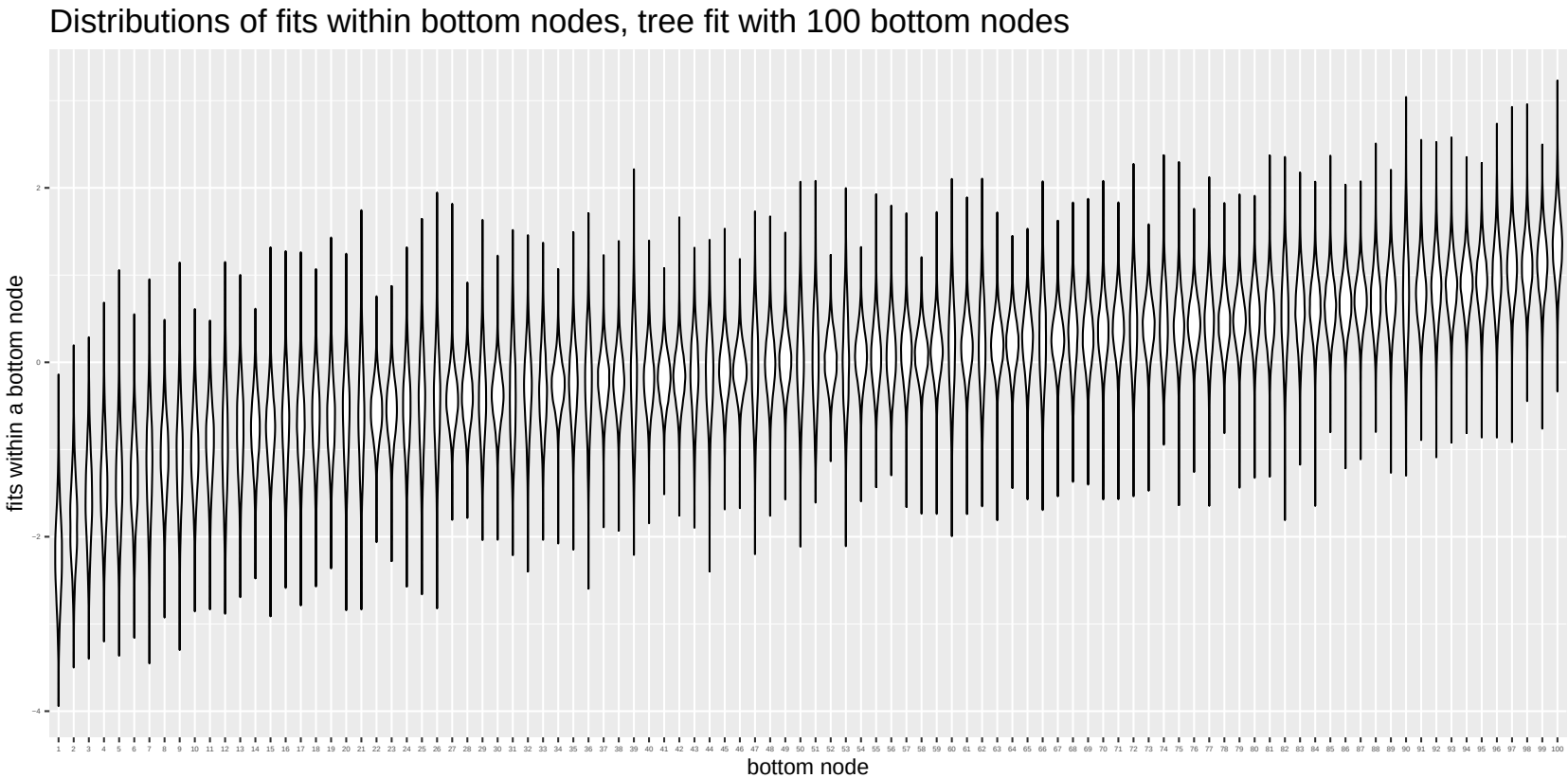
$\hat{f} = \hat{f}^A$, tree with 1,000 bottom nodes.

For each bottom node get the mean of $\hat{f}(x)$ and percent observations in the bottom node.



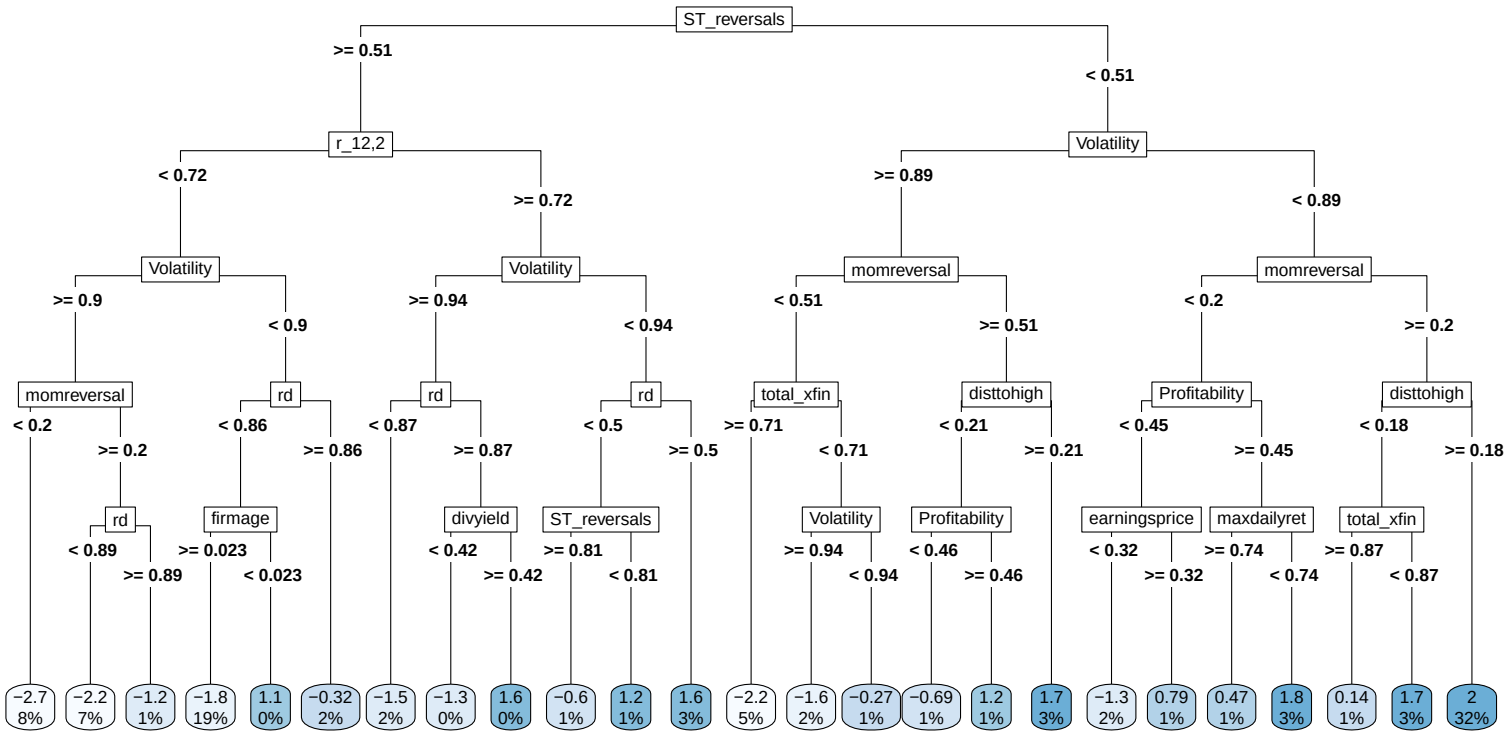
Distribution of 100 $\hat{f}^A(x)$ in each bottom node.

Tree with 100 bottom nodes.



Only use observations giving bottom and top 5% of $\hat{f}(x)$ and then fit the tree.

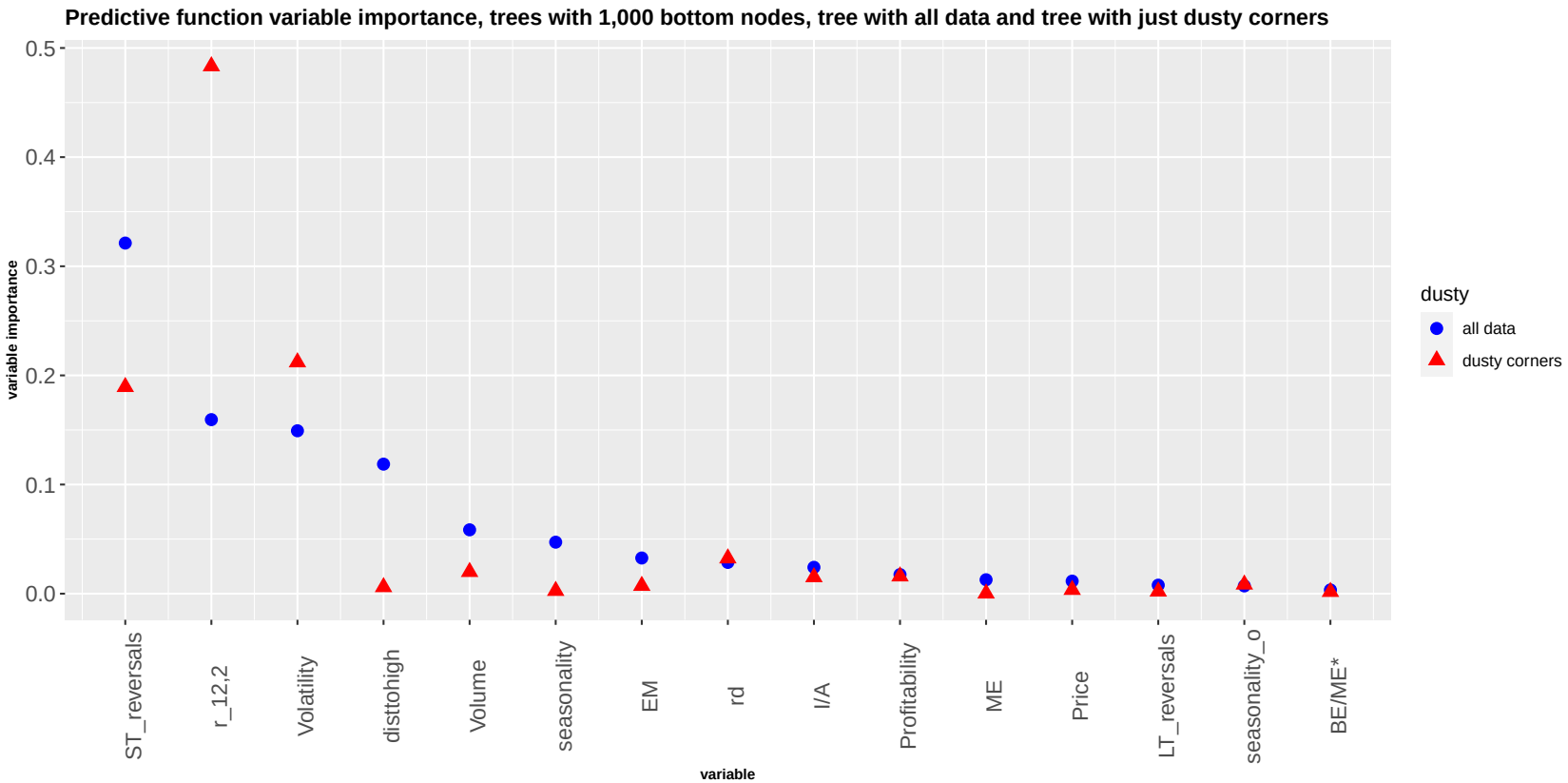
$\hat{f} = \hat{f}^P(x)$



Only use observations giving bottom and top 5% of $\hat{f}(x)$ and then fit the tree with 1,000 bottom nodes.

$\hat{f} = \hat{f}^P(x)$

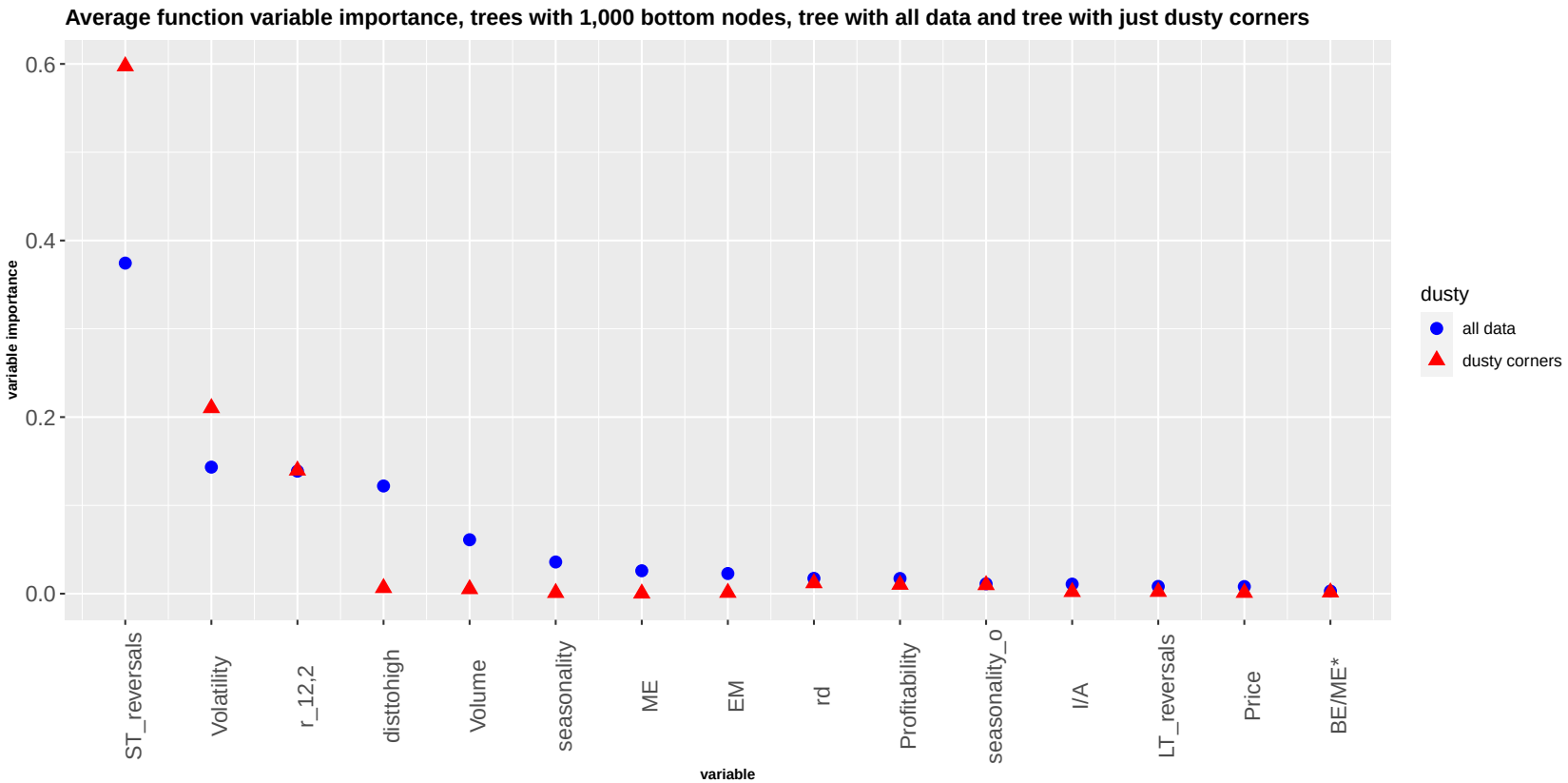
Just used 15 variables.



Only use observations giving bottom and top 5% of $\hat{f}(x)$ and then fit the tree with 1,000 bottom nodes.

$\hat{f} = \hat{f}^A(x)$

Just used 15 variables.



Looking for Non-linearities: Fit-the-Fit, Linear residuals

Looking long and hard at the trees can give you a sense of the relationship, but figuring out what is linear and not, is hard.

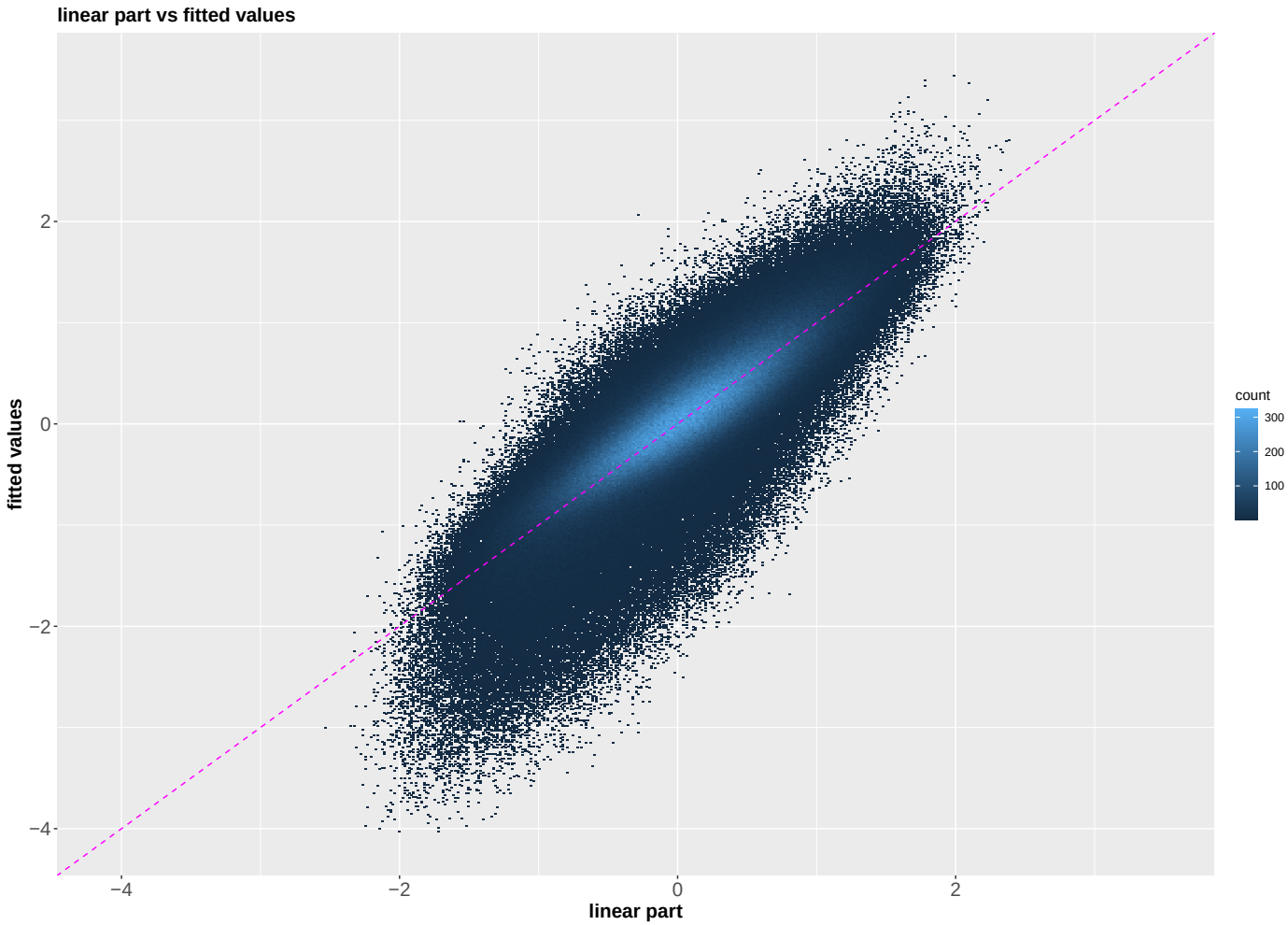
Our idea is that *mostly* the fit $\hat{R}$ is well approximated by a linear fit.

But, there are important "dusty" corners where there are departures from linearity.

To find the dusty corners, we regress the fit $\hat{R}$ on x and then seek to understand the residuals.

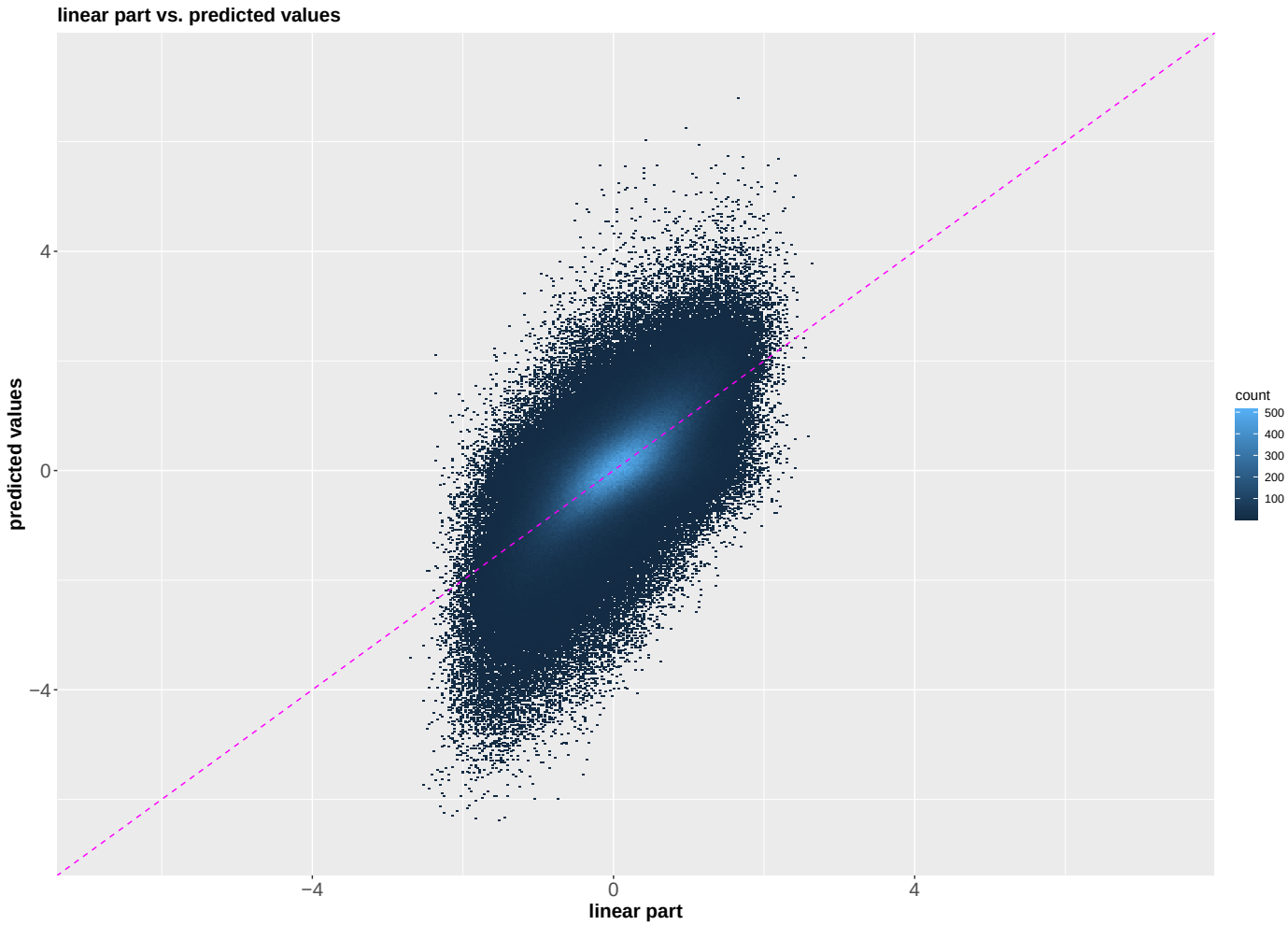
$\hat{f} = \hat{f}^A$, 15 variables.

Linear part vs. $100\,\hat{f}^A(x)$.



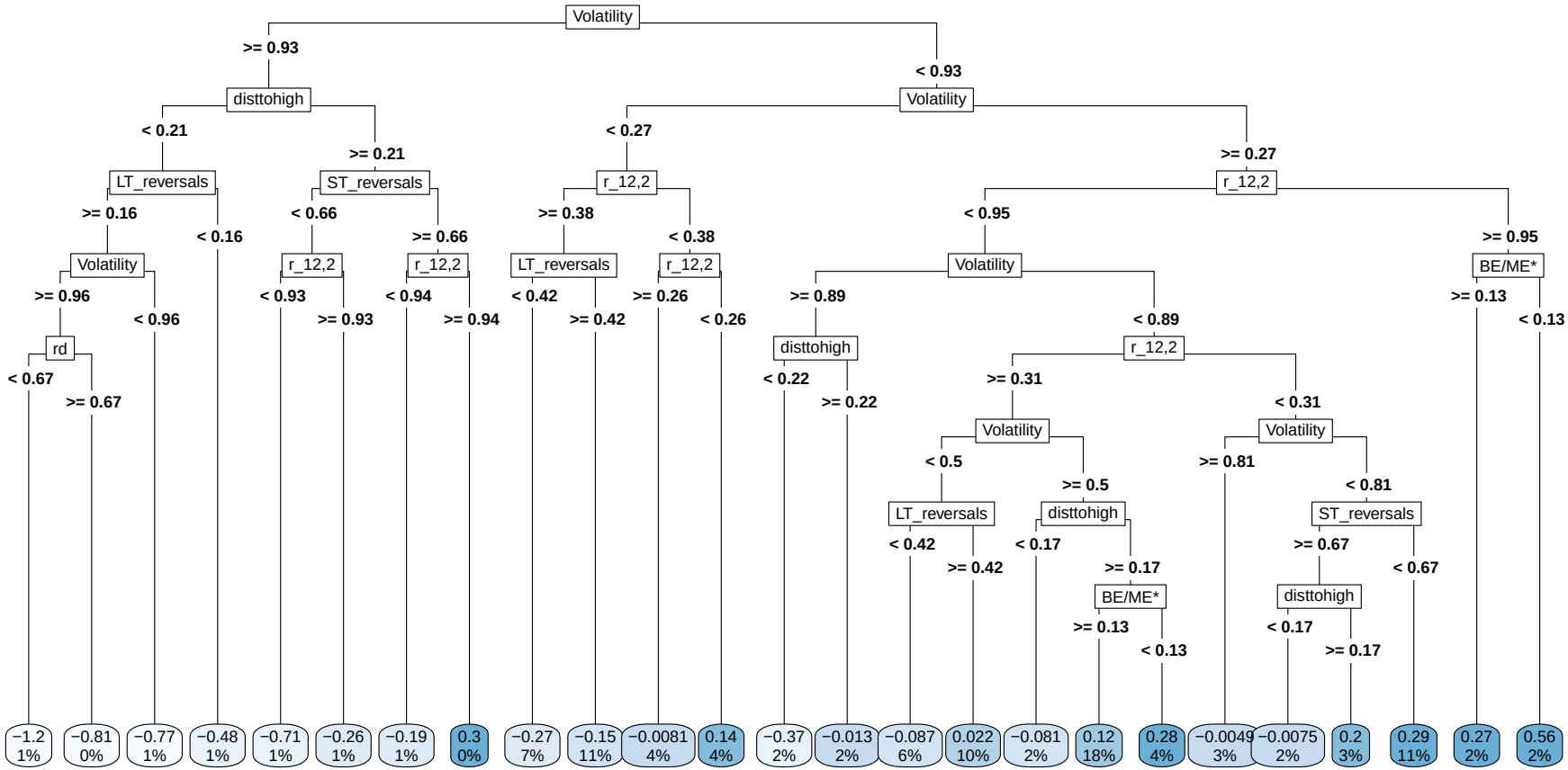
$\hat{f} = \hat{f}^P$, 15 variables.

Linear part vs. $100\,\hat{f}^P(x)$.



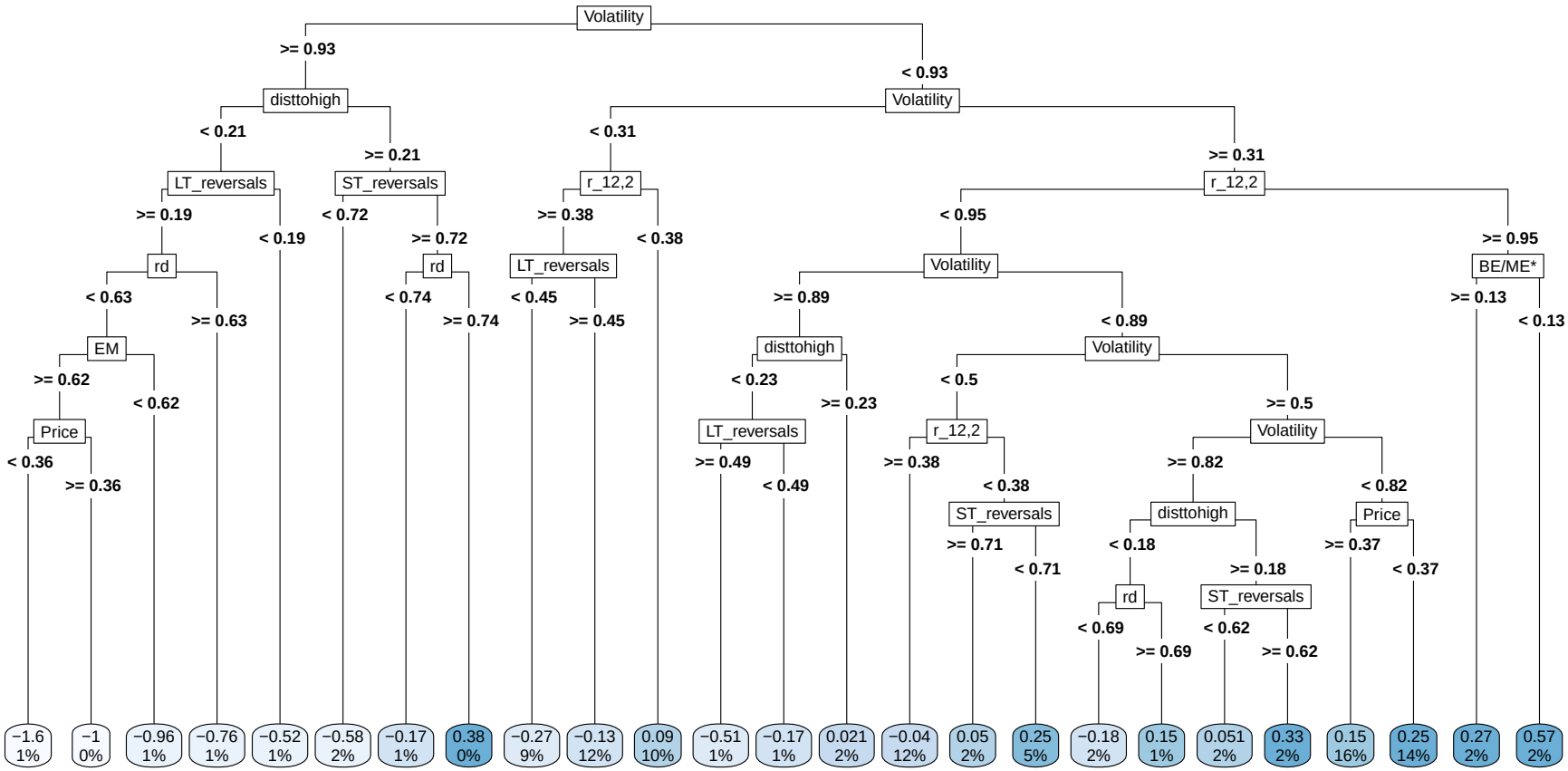
$\hat{f} = \hat{f}^A$, 15 variables.

Tree fit to residuals.



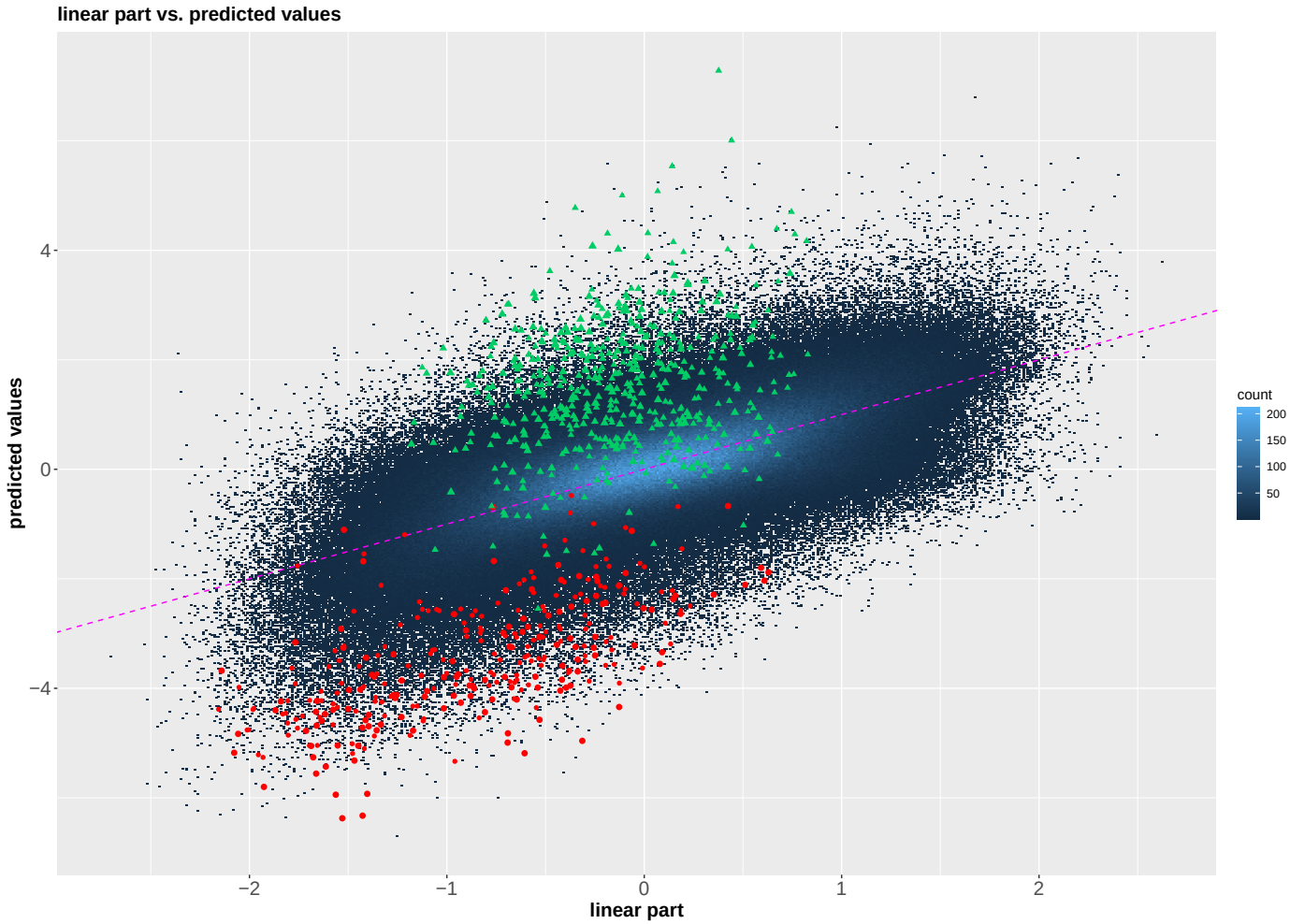
$\hat{f} = \hat{f}^P$, 15 variables.

Tree fit to residuals.

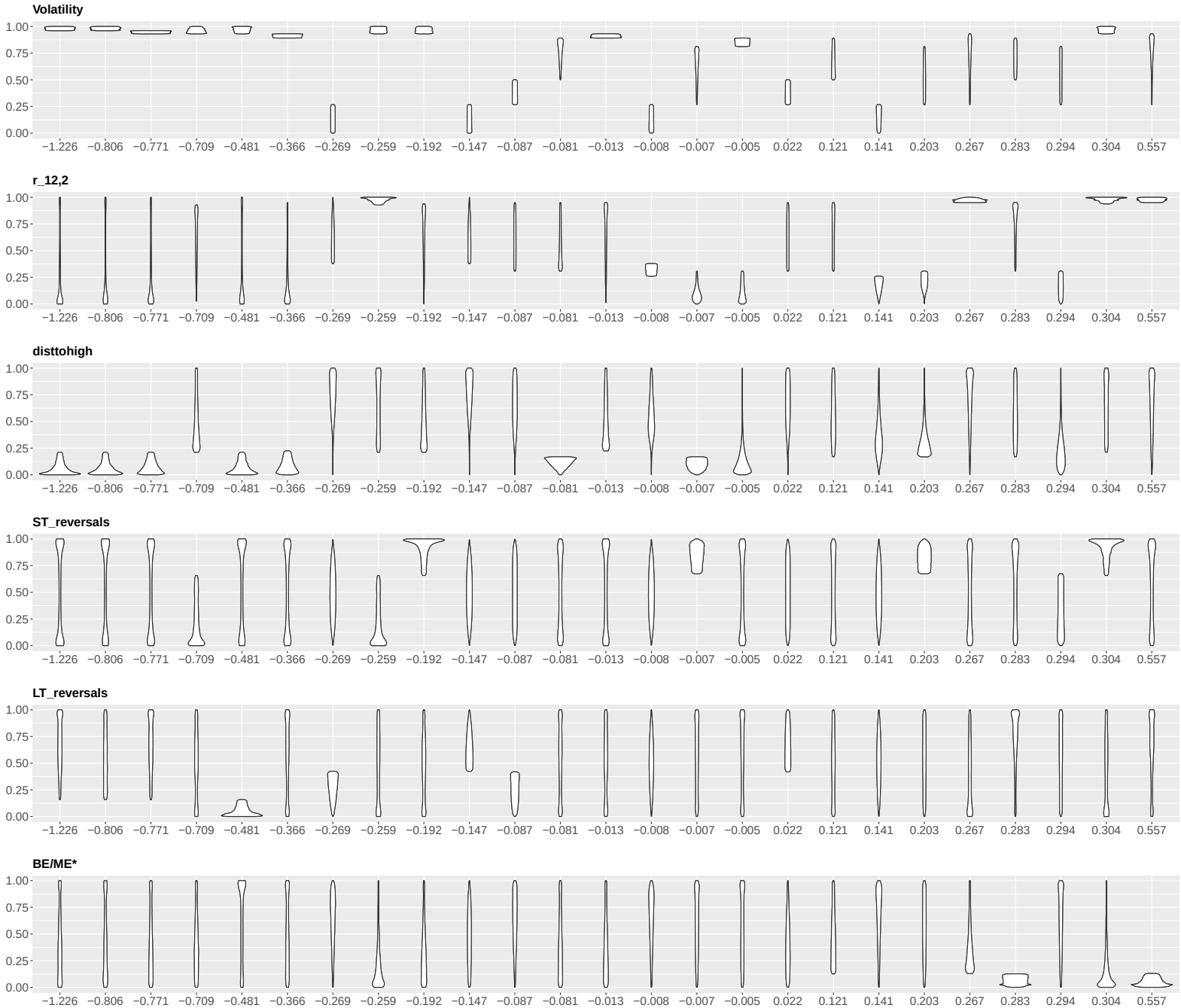


$\hat{f} = \hat{f}^P$, 15 variables.

Linear part versus predicted values using f_P and the 15 selected variables.
Predicted non-linear values from bottom nodes of 1,000 bottom node color coded.
Red means low non-linear values from tree and green means high nonlinear values
with size of plot symbol related to size of nonlinear value.

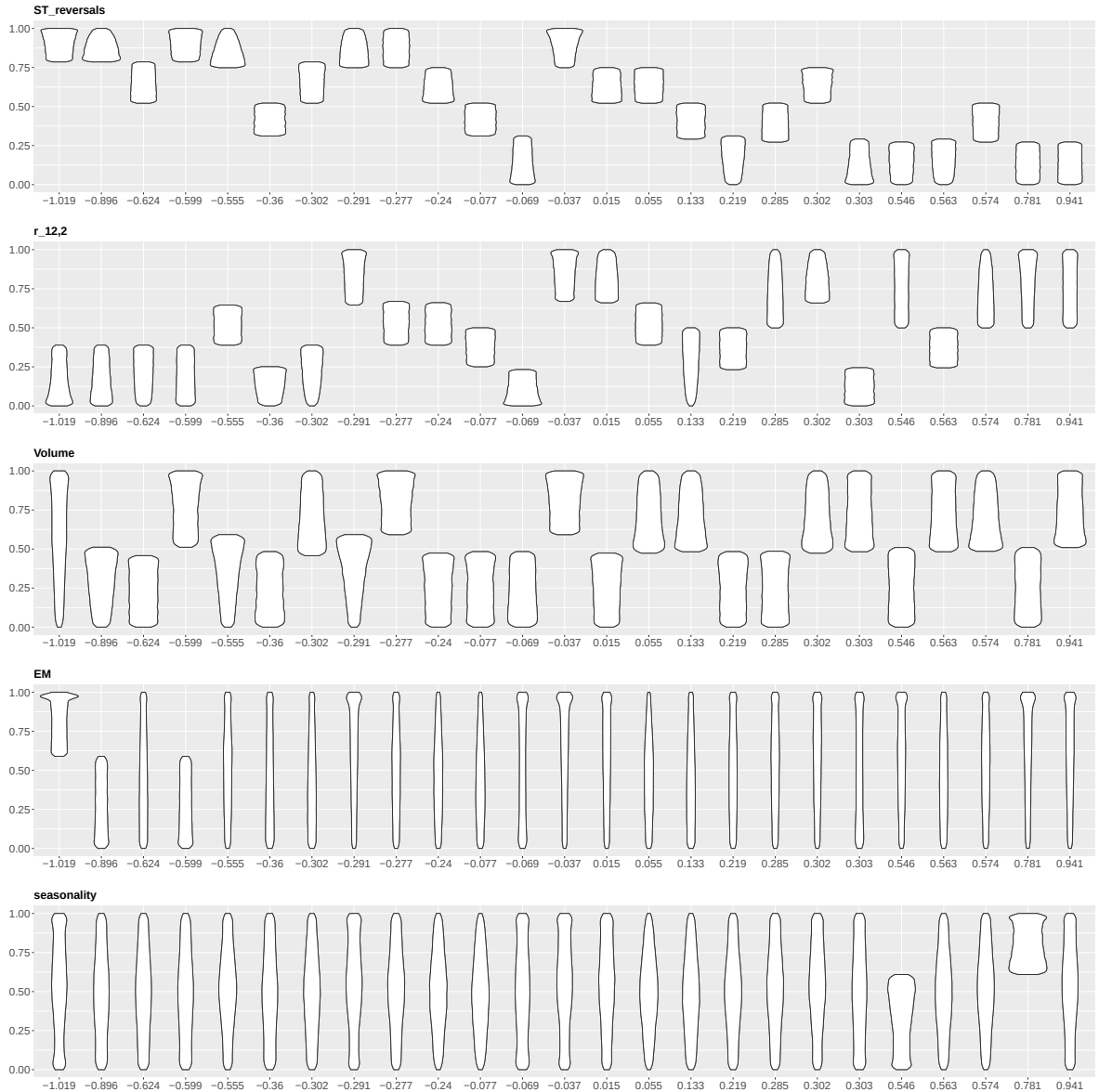


$\hat{f} = \hat{f}^A$, 15 variables.



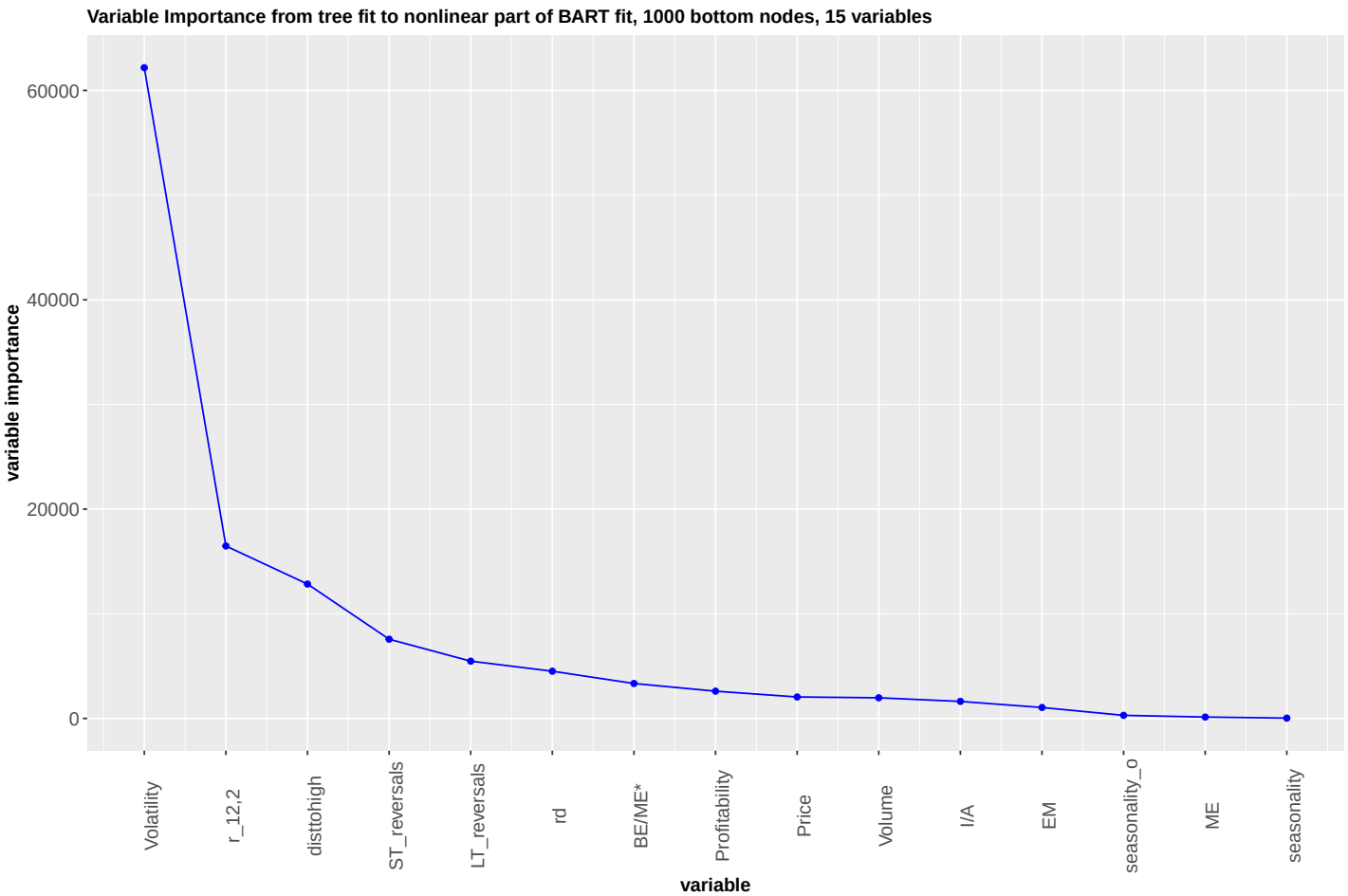
$\hat{f} = \hat{f}^A$, 15 variables.

Tree fit to *linear part*.



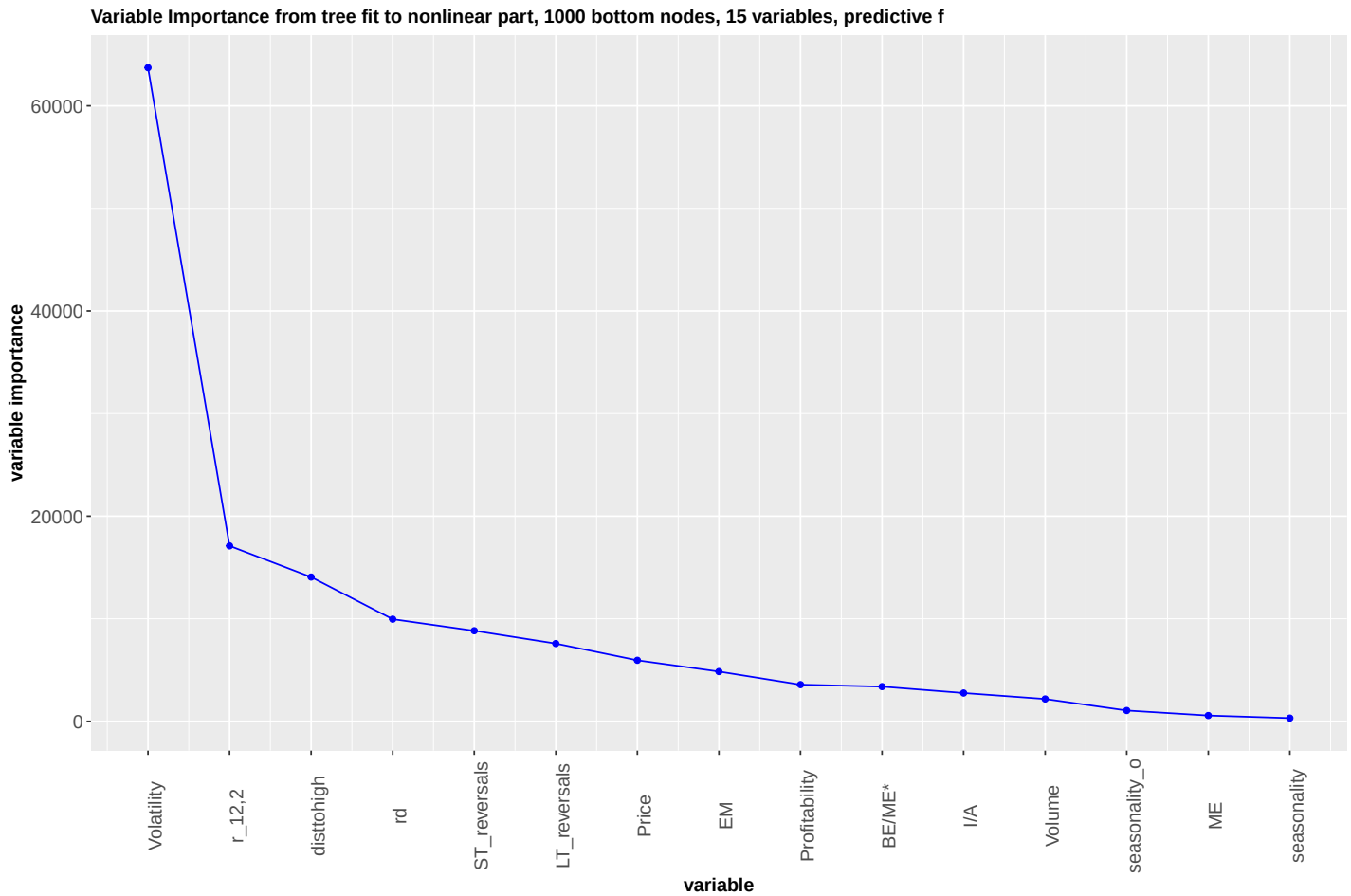
$\hat{f} = \hat{f}^A$, 15 variables.

Tree with 1,000 bottom nodes.



$\hat{f} = \hat{f}^P$, 15 variables.

Tree with 1,000 bottom nodes.



Looking for Interactions: Fit-the-Fit, GAM Residuals

We have found the parts of predictor space where the nonlinear fit seems to be different from the linear fit.

But *how* are they different??

Something we often think about are *interactions*.

Do certain variables *combine* to produce an effect.

We will pull out a GAM fit and look at the resids to find the interactions.

What is a GAM?

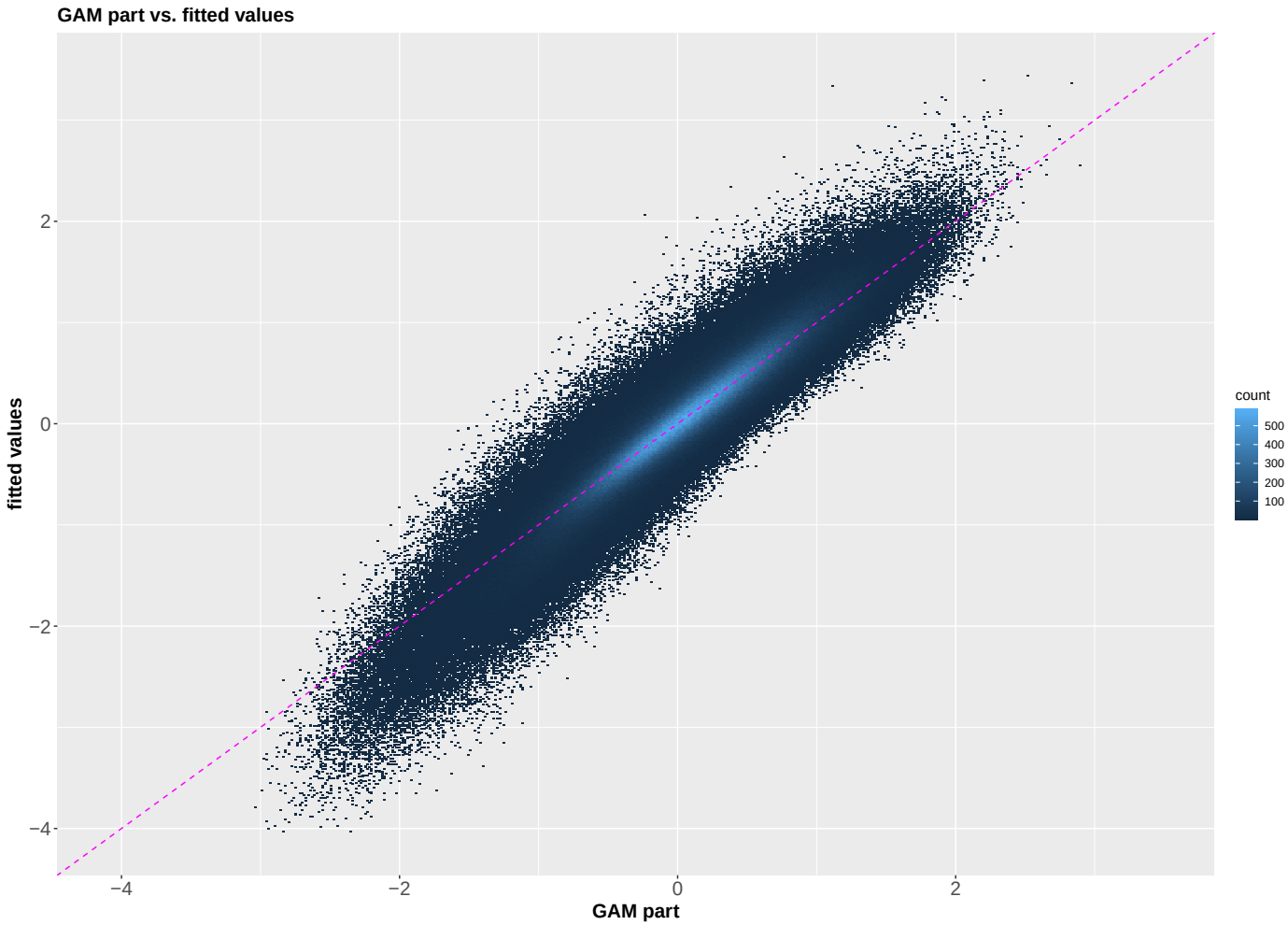
$$f(x_1, x_2, \ldots, x_p) = \sum_{j=1}^{p} f_j(x_j).$$

where we are very flexible in the fitting of each f_j.

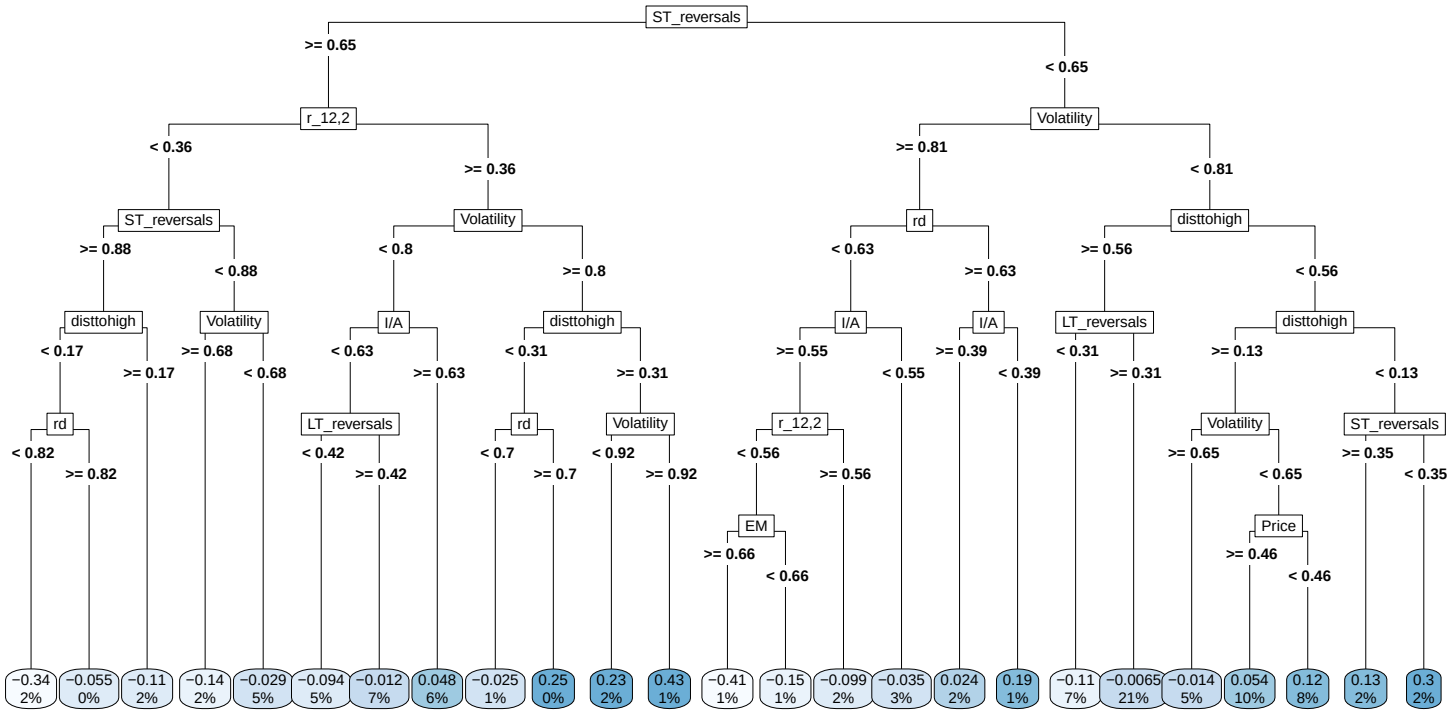
So we can be as nonlinear as we like in each variable, but there are no interactions.

Pretty popular in applied statistics.

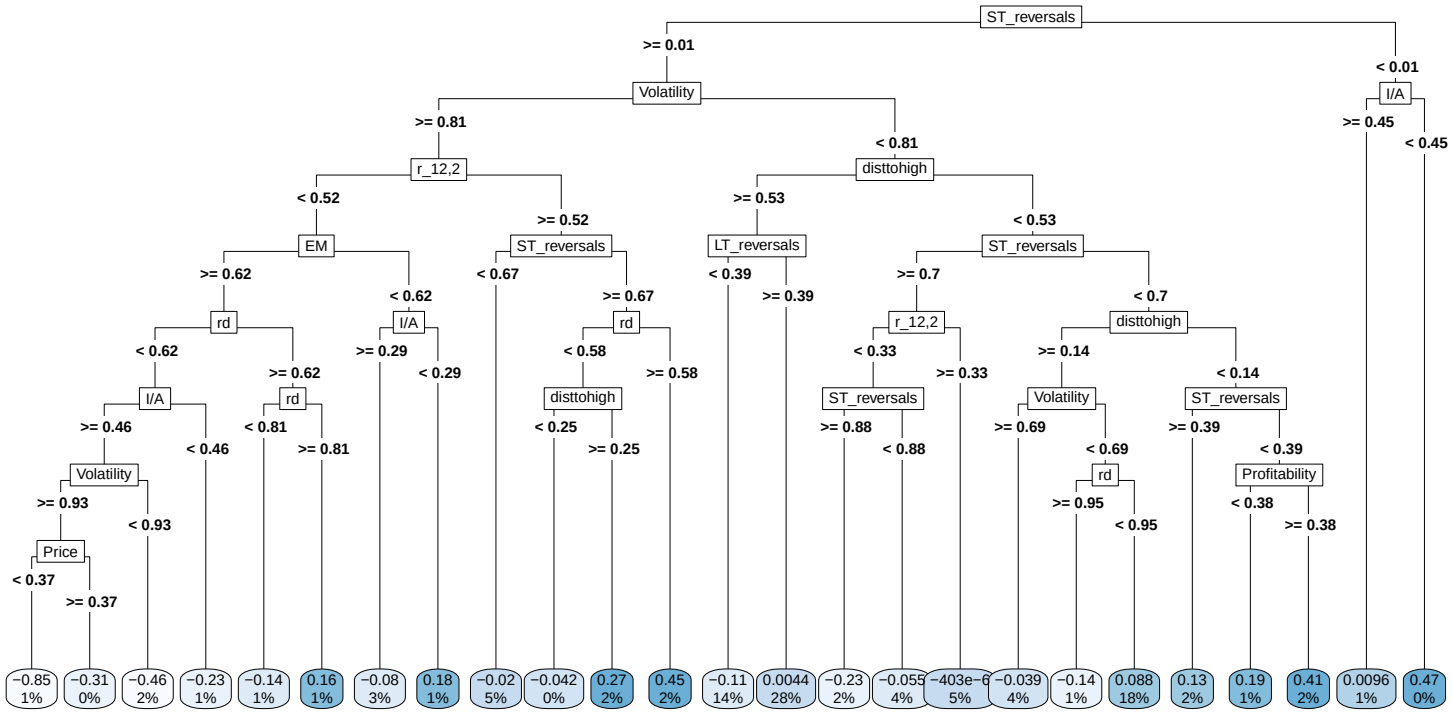
$\hat{f}^A$, 15 x.


GAM part vs. fitted values
fitted values
GAM part
count
500
400
300
200
100
2
0
−2
−4
−4
−2
0
2

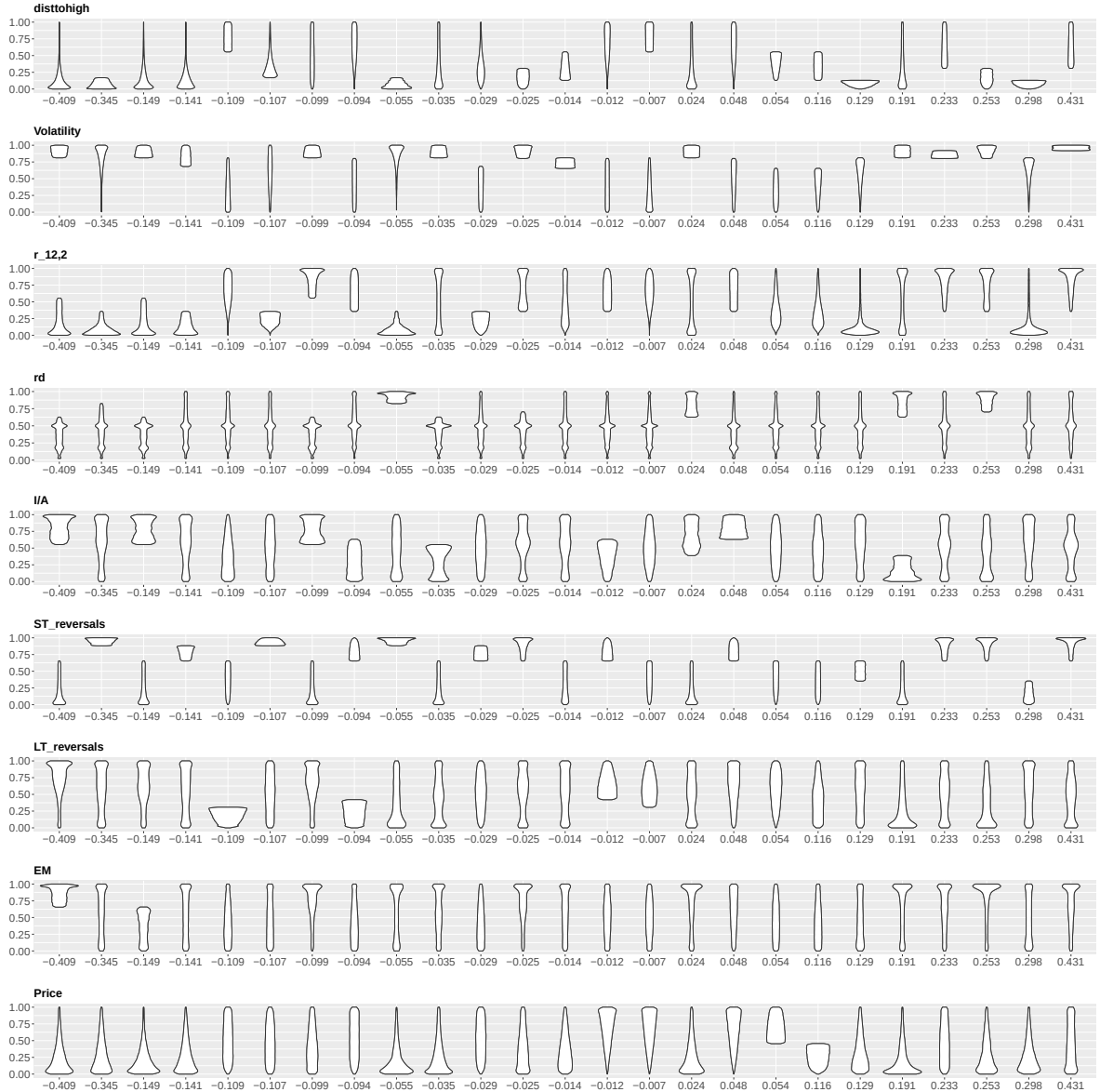
$\hat{f} = \hat{f}^A$, 15x.



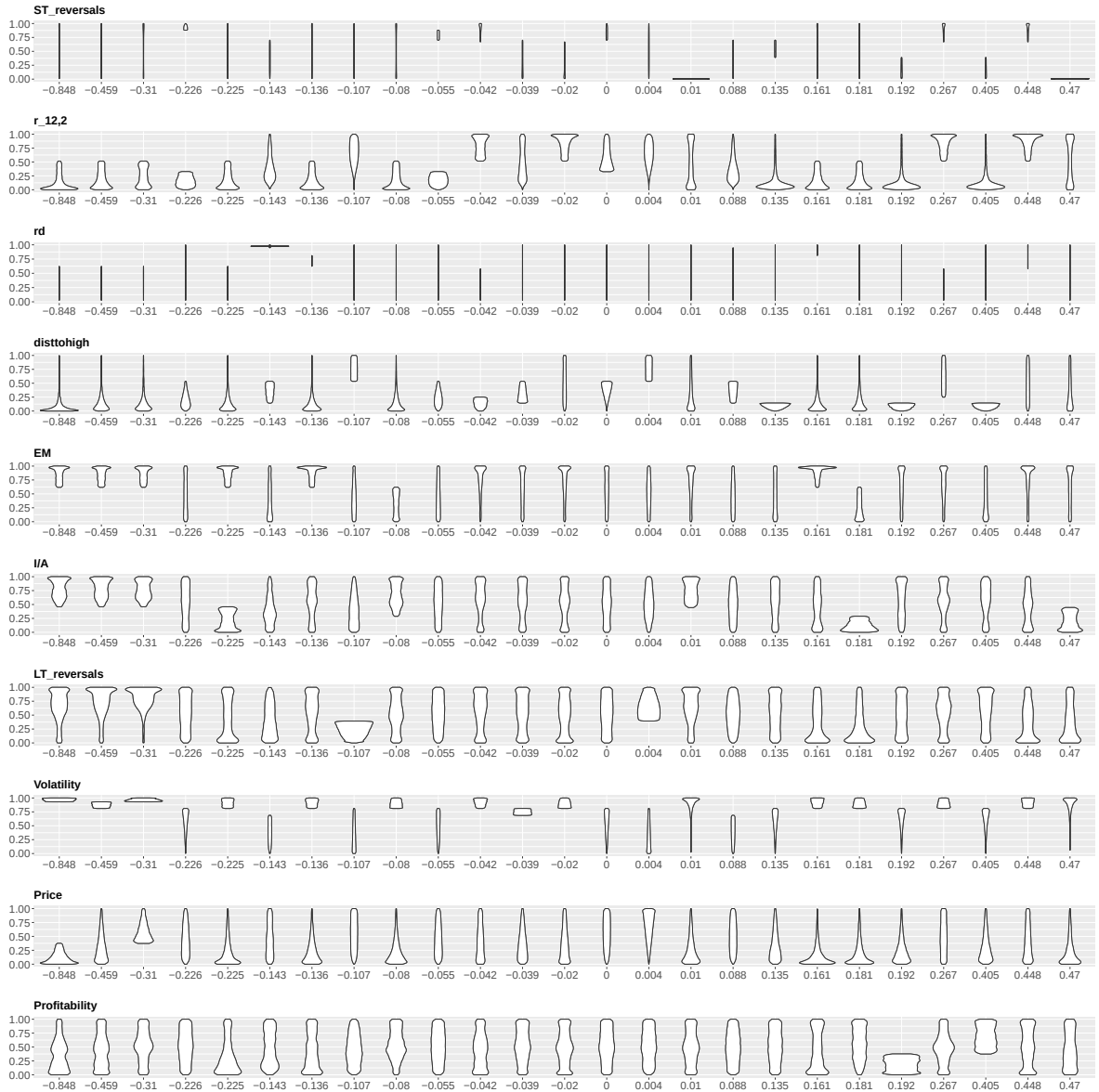
$\hat{f} = \hat{f}^P$, 15x.



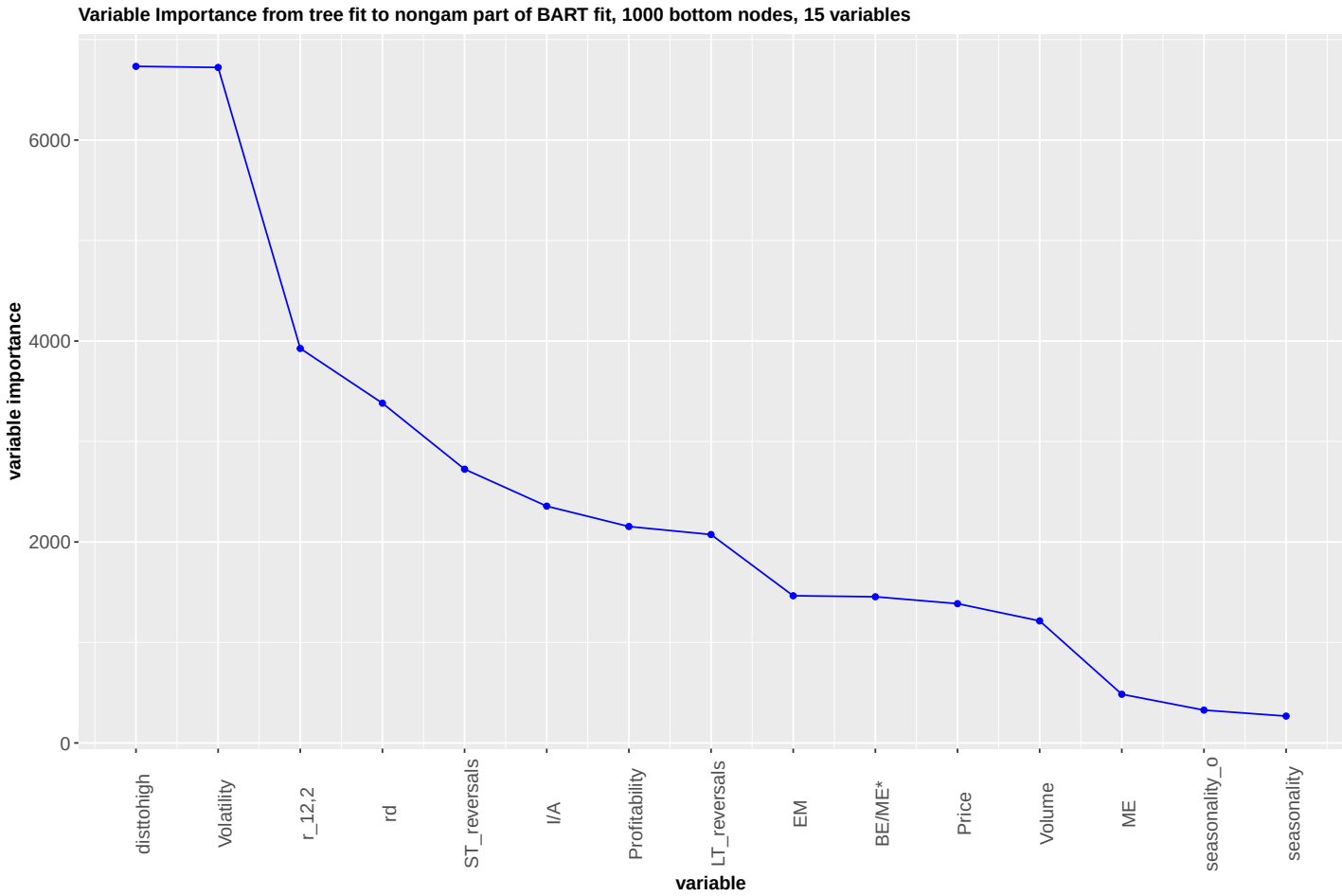
$\hat{f} = \hat{f}^A$, 15x.



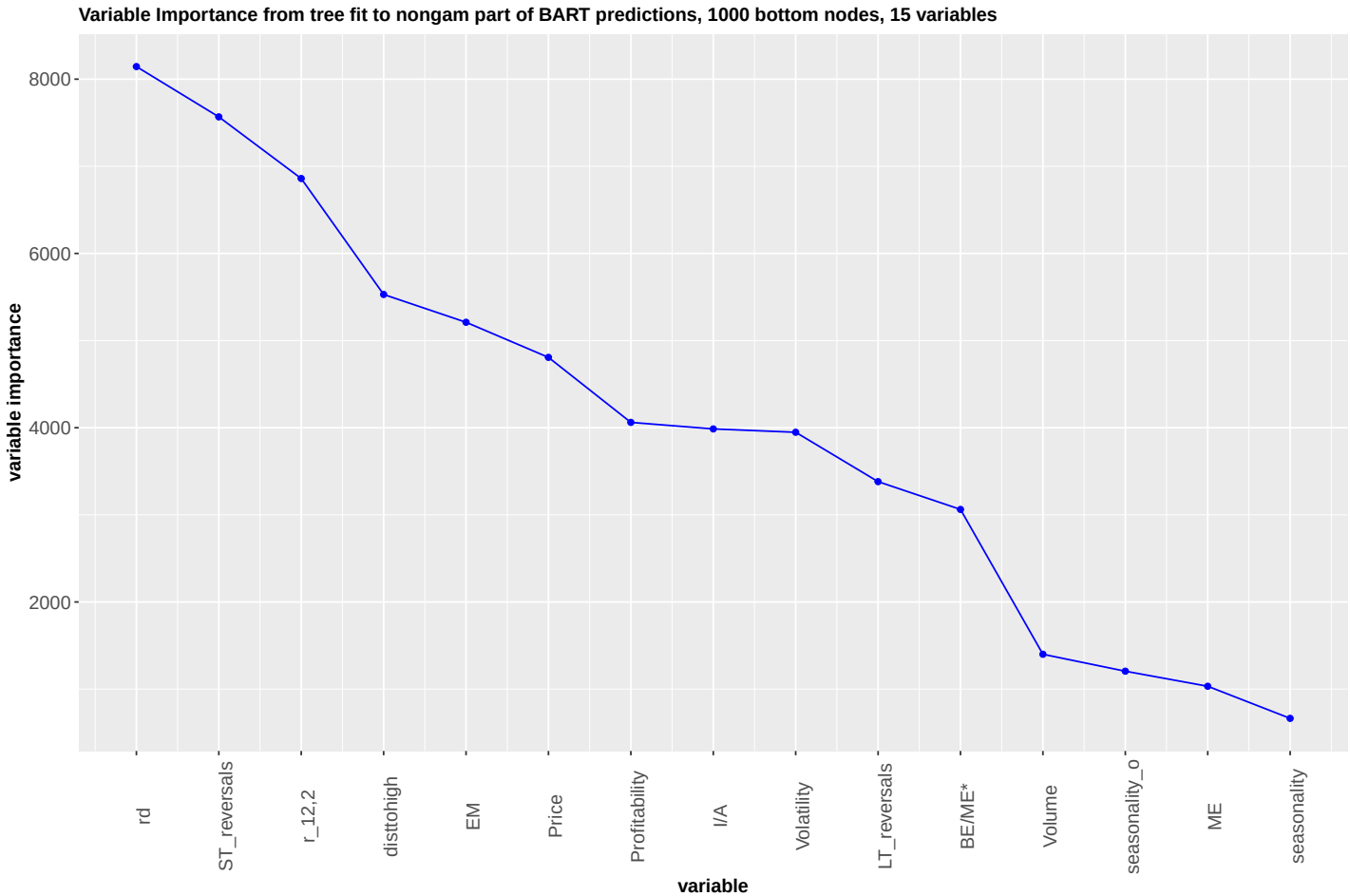
$\hat{f} = \hat{f}^P$, 15x.



$\hat{f} = \hat{f}^A$, 15x.

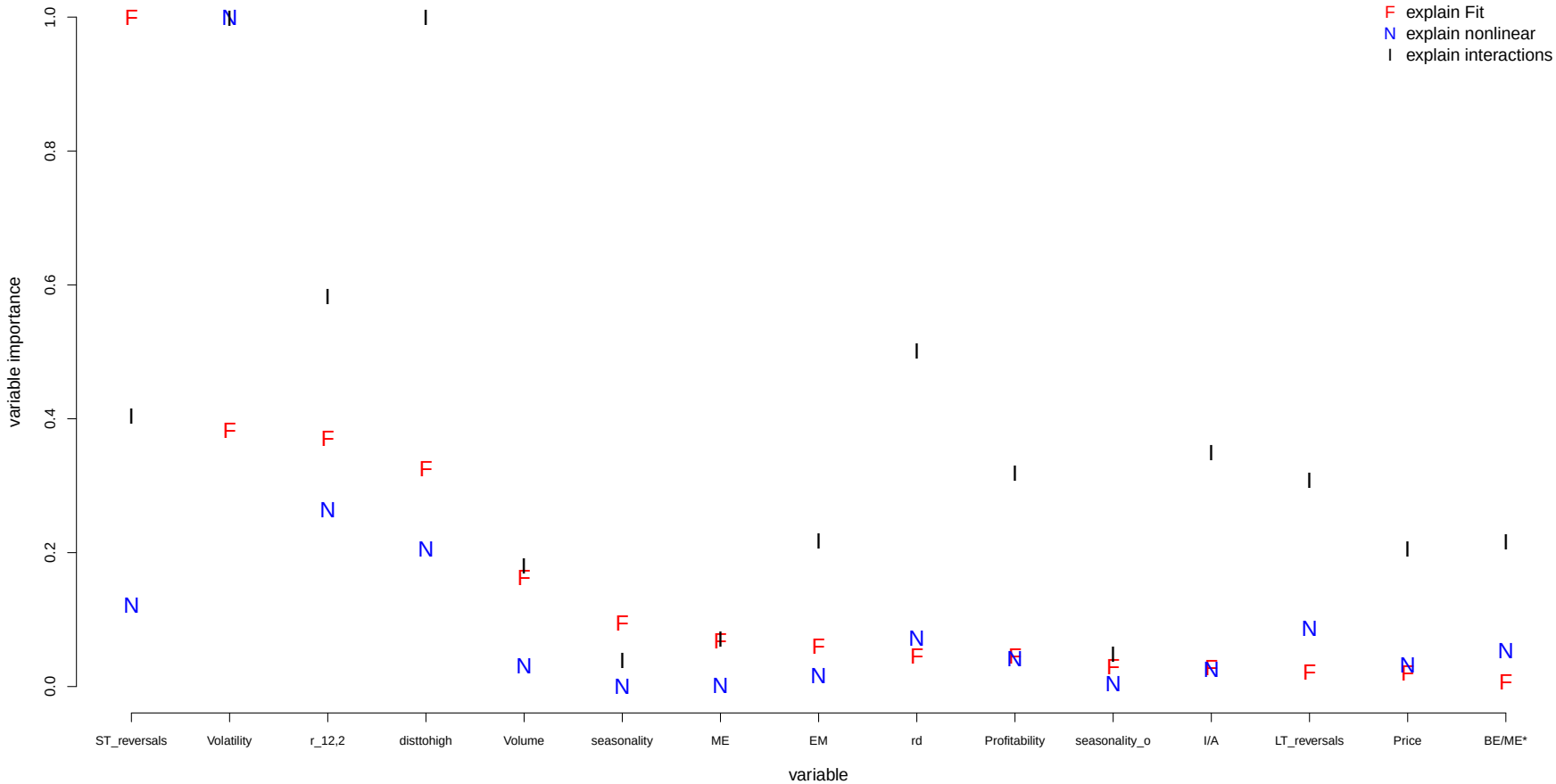


$\hat{f} = \hat{f}^P$, 15x.



$\hat{f} = \hat{f}^A$, 15x.

fit, nonlinear, and non-gam variable importance from 1,000 bottom node tree.



3. Concluding Remarks

a quote from Gu, Kelly, Xiu:

```
"The most successful predictors are
   price trends, liquidity, and volatility."
```

So, big picture we agree with Gu et. al. but add a few more.

Nice confirmation since much of what we done is different *and* we have much more of a feeling for what kinds or roles the key variables play.