Using Data Mining Tools of Decision Trees in Quality and Reliability Applications

Brief Example on Modern Engineered Wood

Hyunjoong (June) Kim, Frank Guess and Timothy Young
Yonsei University, Seoul, Korea
University of Tennessee, Knoxville

2006 Joint Research Conference on Statistics
in Quality, Industry and Technology
Knoxville, TN

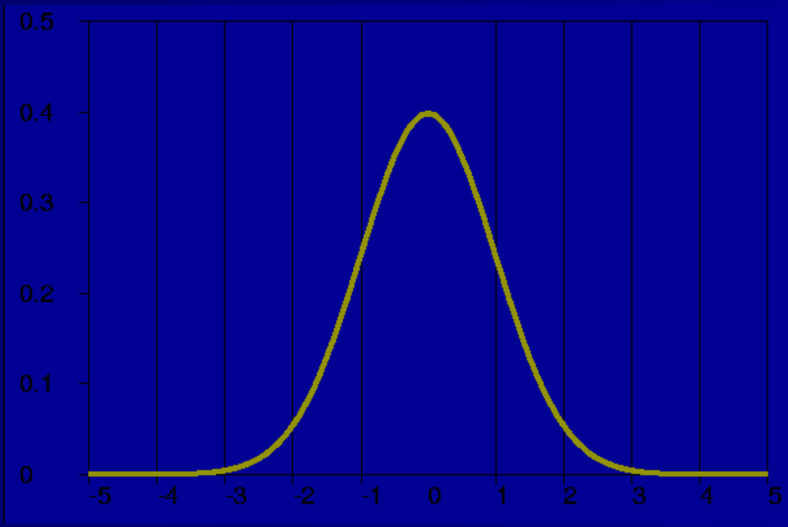
Paper link: http://stat.bus.utk.edu/techrpts/index.html
USA's NSF & Korean sponsored
Reliability & Application Workshop paper

Data Mining (DM)

An Important Tool for Industrial Processes and Real-Time Decision-Making

Data Mining (DM)
Philosophy of Reducing Variation

- □ If you cannot quantify variation, how will you reduce variation?



- □ Once variation is quantified, what are sources of variation?
- □ DM and Decision Trees powerful methods for identifying sources of variation.

Decision Trees (DT)

"Real-time Supervised Classification Learning"

Philosophy of Data Mining

- DM tools find hidden structures & helpful patterns
- DM provides exploratory data analysis with little human interactions
- Is that good or bad?
- Yes, that is good & it is bad! Both … ! Need balance of both automation & human interactions.
- Win/win - old time Deming strategy and modern process of thinking deeper… .
- DT is aligned well with the DM philosophy

Data Mining (DM) & Trees

- Decision trees (classification and regression) are a popular method among DM tools
- Quick ref: Guidici (2003) *Applied Data Mining: Stat Methods for Bus. & Indus.*
- We will discuss results of a case study using "GUIDE", one of many DT methods.

GUIDE

(Generalized, Unbiased, Interaction Detection and Estimation)

Loh (2002)

http://www.stat.wisc.edu/~loh/guide.html

Fit one regression model at each node – multiple regression, stepwise, etc.

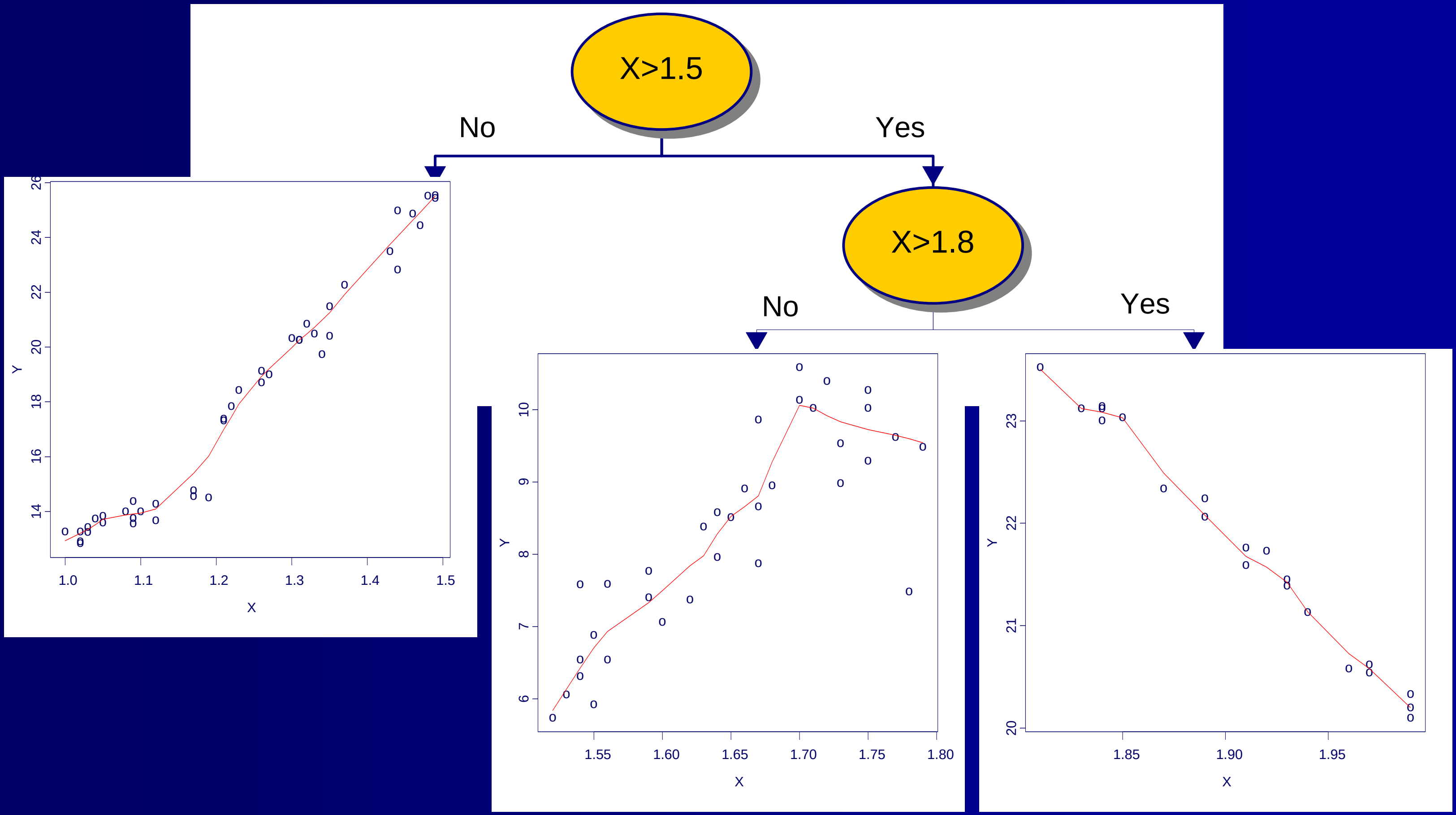
1. Use residuals to select split variable (negligible bias)
2. Select split point or split set
3. Prune tree as in CART

See also Kim & Loh (2003) take a "CRUISE" with another DT tool

http://web.utk.edu/~hjkim/ & 2004 work …

GUIDE

(Key features: sensitivity to curvature and interactions)



X>1.5

No

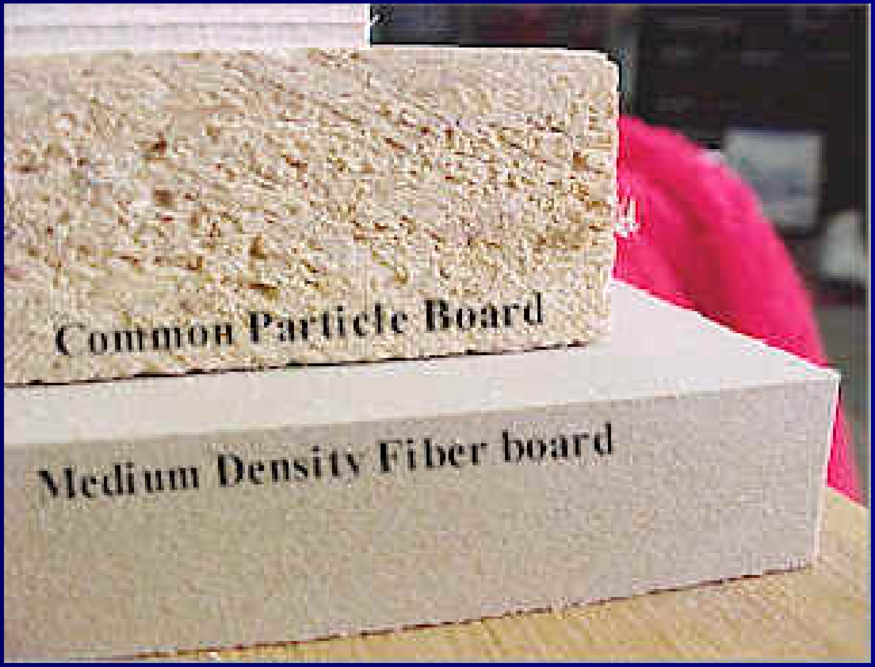
Yes

X>1.8

No

Yes

Case Study
Medium Density Fiberboard (MDF)



Case Study
MDF

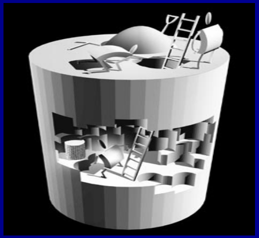
- Medium density fiberboard (MDF): a highly used engineered wood composite
- Interested in the tensile strength (psi) or "internal bond" (IB) of MDF
- Destructive testing is performed during the manufacturing process (Goal: maximize & improve product quality and reliability)
- Prevention of unacceptable reliability can result in millions of dollars saving
- Young & Guess (2002) & Guess et al. (2003) - *IJRA*

Predictors and Response (IB)

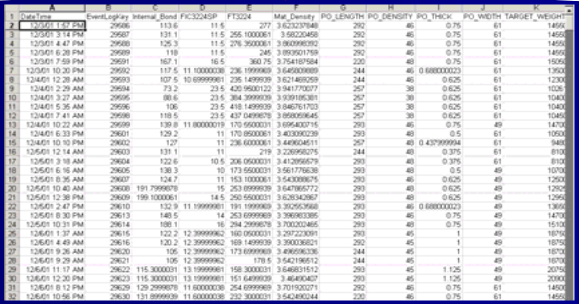
- Predictors: describe types and manufacturing conditions
  - panel density (lbs/ft^3)
  - panel thickness (inches)
  - and others (moisture, line speed, temperature, etc.)
  - day of the week (Sunday through Saturday)
  - shift (morning, afternoon, night)
  - week of the month (first through fourth week)
- Response variable is the strength of IB
- Next comparing regression only vs. GUIDE?

Automated Real-Time Relational Databases

Lab – Destructive Testing
Event Data



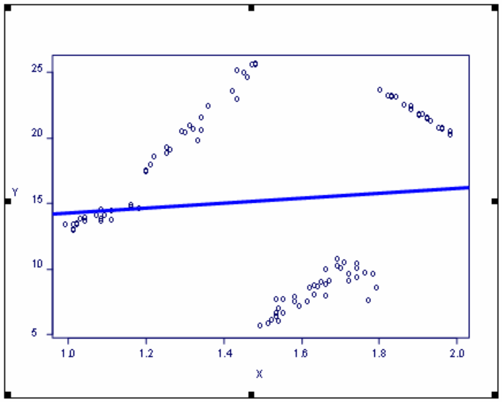
Data Warehouse
Real-Time Data



	A	B	C	D	E	F	G	H	I	J	K
1	DateTime	EventLogKey	Internal_Bond	FIC3224SP	FT3224	Mat_Density	PO_LENGTH	PO_DENSITY	PO_THICK	PO_WIDTH	TARGET_WEIGHT
2	12/3/01 1:57 PM	29586	113.6	11.5	277	3.623237848	292	46	0.75	61	1455
3	12/3/01 3:14 PM	29587	131.1	11.5	255.1000061	3.58220458	292	46	0.75	61	1455
4	12/3/01 4:47 PM	29588	125.3	11.5	276.3500061	3.860998392	292	46	0.75	61	1455
5	12/3/01 6:28 PM	29589	118	11.5	245	3.893501759	292	46	0.75	61	1455
6	12/3/01 7:59 PM	29591	167.1	16.5	360.75	3.754187584	220	48	0.75	61	1505
7	12/3/01 10:20 PM	29592	117.5	11.10000038	236.1999969	3.645809889	244	46	0.688000023	61	1350
8	12/4/01 12:28 AM	29593	107.5	10.69999981	235.1499939	3.621469259	244	46	0.625	61	1230
9	12/4/01 2:29 AM	29594	73.2	23.5	420.9500122	3.941770077	257	38	0.625	61	1026
10	12/4/01 3:27 AM	29595	88.6	23.5	384.3999939	3.939185381	257	38	0.625	61	1040
11	12/4/01 5:35 AM	29596	106	23.5	418.1499939	3.846761703	257	38	0.625	61	1040
12	12/4/01 7:41 AM	29598	118.5	23.5	437.0499878	3.858059645	257	38	0.625	61	1045
13	12/4/01 10:22 AM	29599	139.8	11.80000019	170.6500031	3.695400715	293	46	0.75	49	1470
14	12/4/01 6:33 PM	29601	129.2	11	170.8500061	3.403090239	293	48	0.5	61	1050
15	12/4/01 10:10 PM	29602	127	11	236.6000061	3.449604511	257	48	0.4379999994	61	948
16	12/5/01 12:14 AM	29603	131.1	11	219	3.226958275	244	48	0.375	61	810
17	12/5/01 3:18 AM	29604	122.6	10.5	206.0500031	3.412856579	293	48	0.375	61	810
18	12/5/01 6:16 AM	29605	138.3	10	173.5500031	3.561776638	293	48	0.5	49	1070
19	12/5/01 8:35 AM	29607	124.7	11	153.1000061	3.543088675	293	46	0.625	49	1250
20	12/5/01 10:40 AM	29608	191.7999878	15	253.8999939	3.647865772	293	48	0.625	49	1292
21	12/5/01 12:38 PM	29609	199.1000061	14.5	250.5500031	3.628342867	293	48	0.625	49	1295
22	12/5/01 2:47 PM	29610	132.9	11.19999981	191.1999969	3.392553568	293	46	0.688000023	49	1365
23	12/5/01 8:30 PM	29613	148.5	14	253.6999969	3.396983385	293	46	0.75	49	1470
24	12/5/01 10:31 PM	29614	188.1	16	294.2999878	3.700202465	293	48	0.75	49	1510
25	12/6/01 1:37 AM	29615	122.2	12.39999962	160.0500031	3.297223091	293	45	1	49	1875
26	12/6/01 4:49 AM	29616	120.2	12.39999962	169.1499939	3.390036821	292	45	1	49	1875
27	12/6/01 9:26 AM	29620	105	12.39999962	173.6999969	3.496596336	244	45	1	49	1870
28	12/6/01 9:29 AM	29621	105	12.39999962	178.5	3.542196512	244	45	1	49	1870
29	12/6/01 11:17 AM	29622	115.3000031	13.19999981	158.3000031	3.646831512	293	45	1.125	49	2075
30	12/6/01 12:20 PM	29623	115.3000031	13.19999981	151.6499939	3.46490407	293	45	1.125	49	2090
31	12/6/01 8:12 PM	29629	129.2999878	11.60000038	254.6999969	3.701920271	292	46	0.75	61	1450
32	12/6/01 10:56 PM	29630	131.8999939	11.60000038	232.3000031	3.542490244	220	46	0.75	61	1455

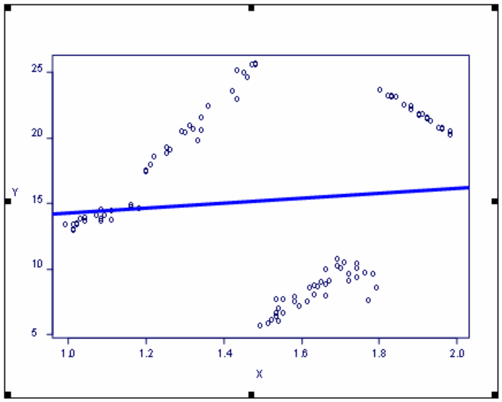
- □ Mills collect a lot of data, how much knowledge is gained?

“Regression Only” Analysis



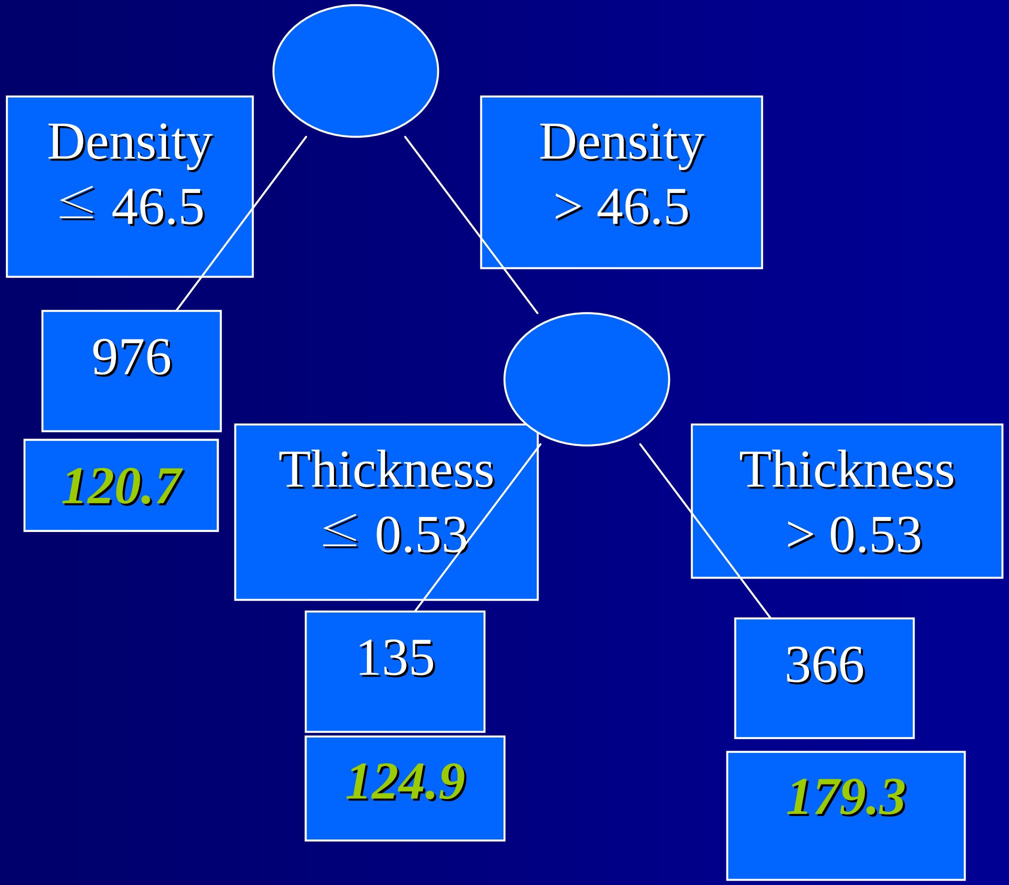
Source		Coef	S.E.	Source		Coef	S.E.
length		0.003	0.015	Day	Sunday	0.873	1.171
density		21.371	0.476		Monday	5.123	1.168
thickness		70.701	4.317		Tuesday	1.522	1.216
width		-0.063	0.106		Wednesday	-3.264	1.201
week	1	3.021	0.799		Thursday	-1.846	1.369
	2	-2.841	0.797		Friday	-4.266	1.202
	3	0.736	0.735	Shift	Morning	-1.845	0.632
					Afternoon	1.535	0.639

“Regression Only” Analysis



Source		Coef	S.E.	Source		Coef	S.E.
length		0.003	0.015	Day	Sunday	0.873	1.171
density		21.371	0.476		Monday	5.123	1.168
thickness		70.701	4.317		Tuesday	1.522	1.216
width		-0.063	0.106		Wednesday	-3.264	1.201
week	1	3.021	0.799		Thursday	-1.846	1.369
	2	-2.841	0.797		Friday	-4.266	1.202
	3	0.736	0.735	Shift	Morning	-1.845	0.632
					Afternoon	1.535	0.639

GUIDE Regression Tree



Density ≤ 46.5

Density > 46.5

976

120.7

Thickness ≤ 0.53

Thickness > 0.53

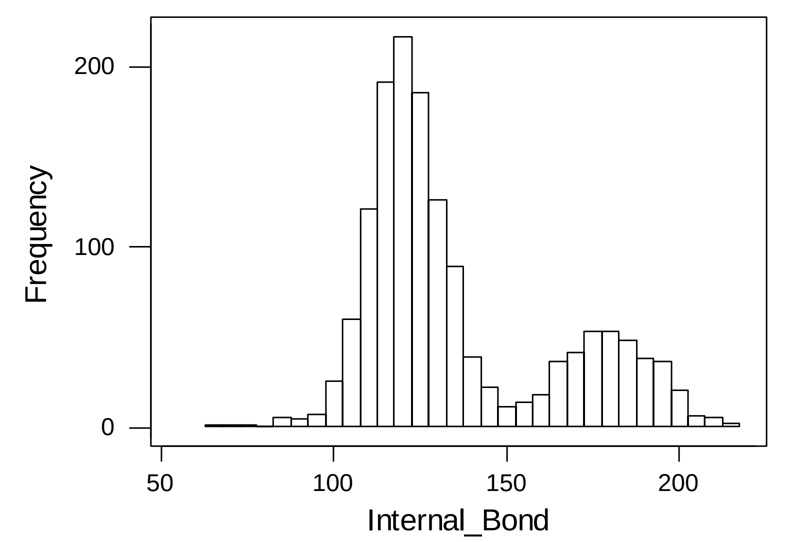
135

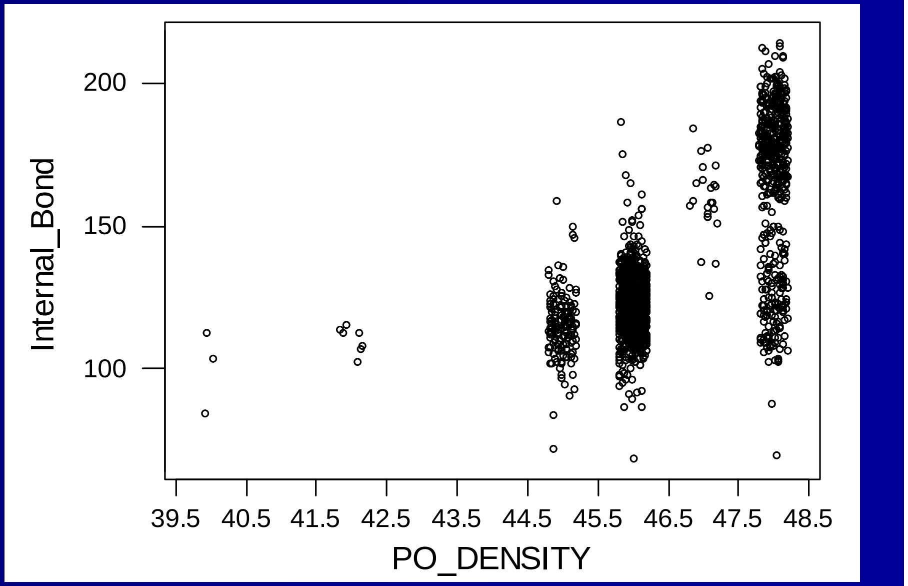
124.9

366

179.3

Why split, e.g., panel density?

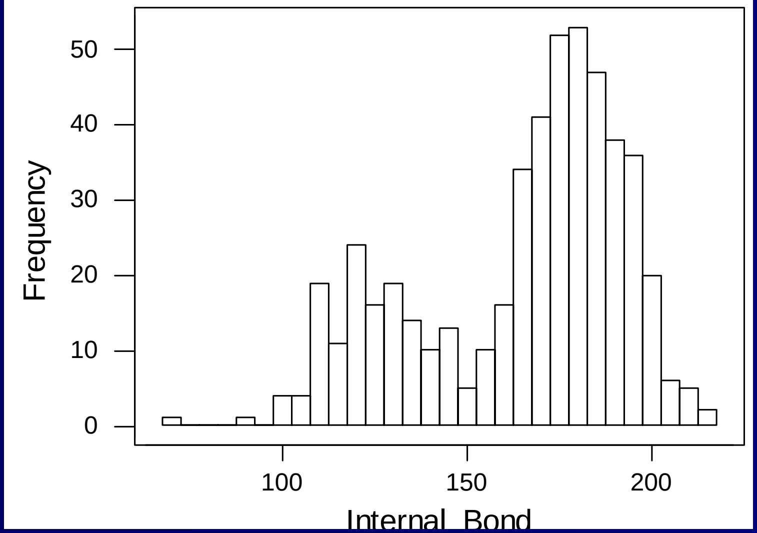


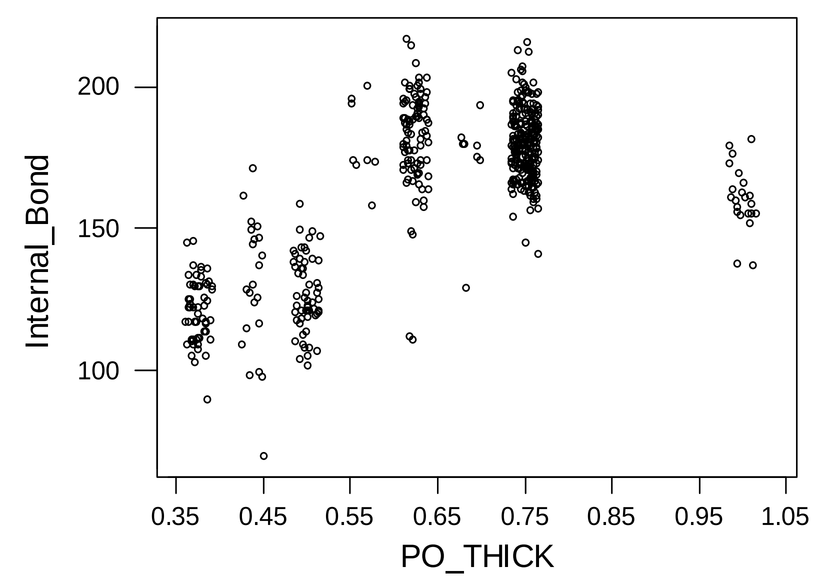


GUIDE identifies location of split very easily and quickly

(not simply along pre-defined product type set points)

Why split, e.g., panel thickness?

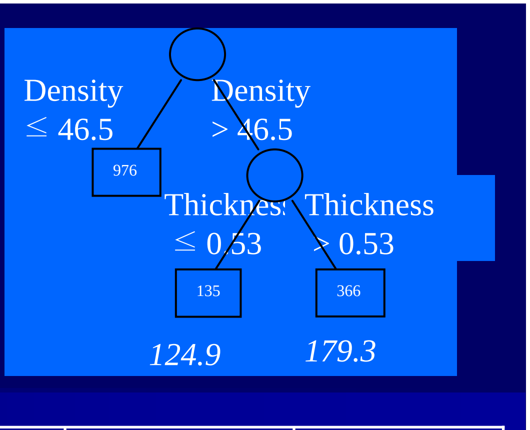




GUIDE identifies location of split very easily and quickly

(not simply along pre-defined product type set points)

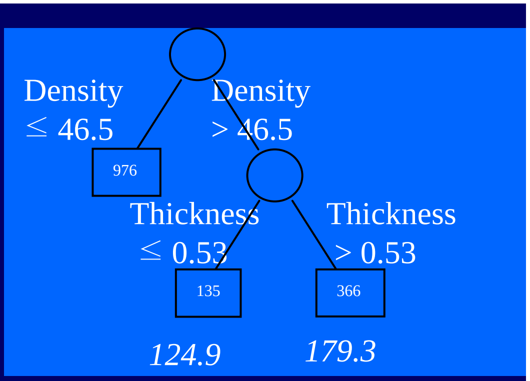
Density ≤ 46.5



Source		Coef	S.E.	Source		Coef	S.E.
length		0.009	0.012	Day	Sunday	0.037	0.903
density		3.889	0.708		Monday	-1.770	0.971
thickness		-11.063	3.765		Tuesday	-2.466	0.892
width		0.0362	0.087		Wednesday	0.806	0.997
week	1	1.762	0.629		Thursday	0.455	1.062
	2	-1.516	0.625		Friday	1.754	0.995
	3	0.286	0.569	Shift	Morning	-1.035	0.495
					Afternoon	0.922	0.497

Not desirable to have “Day”, “Shift” as source of variation!

Density > 46.5 and Thickness > 0.53



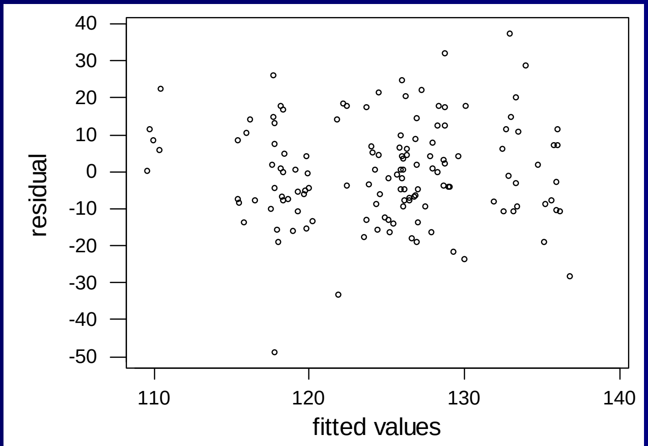
Source		Coef	S.E.	Source		Coef	S.E.
length		-0.019	0.022	Day	Sunday	0.692	1.803
density		22.561	4.834		Monday	-1.866	1.829
thickness		1.320	14.81		Tuesday	3.100	3.411
width		0.089	0.181		Wednesday	1.603	2.602
week	1	3.614	1.245		Thursday	0.269	2.808
	2	-4.913	1.270		Friday	-0.331	1.928
	3	-0.936	1.153	Shift	Morning	-0.777	0.990
					Afternoon	-0.171	0.990

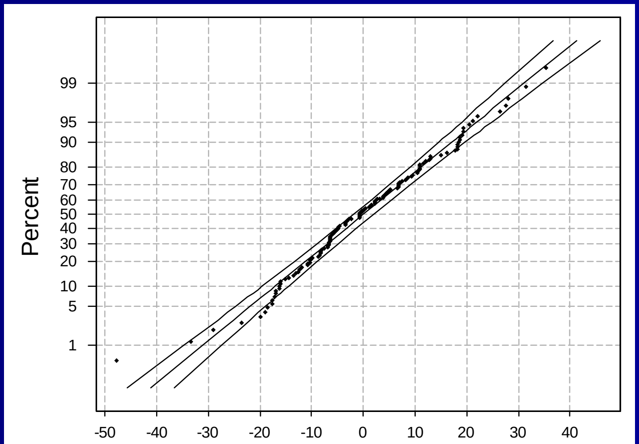
Again, not desirable to have “Day” as source of variation!

GUIDE vs. Regression Only

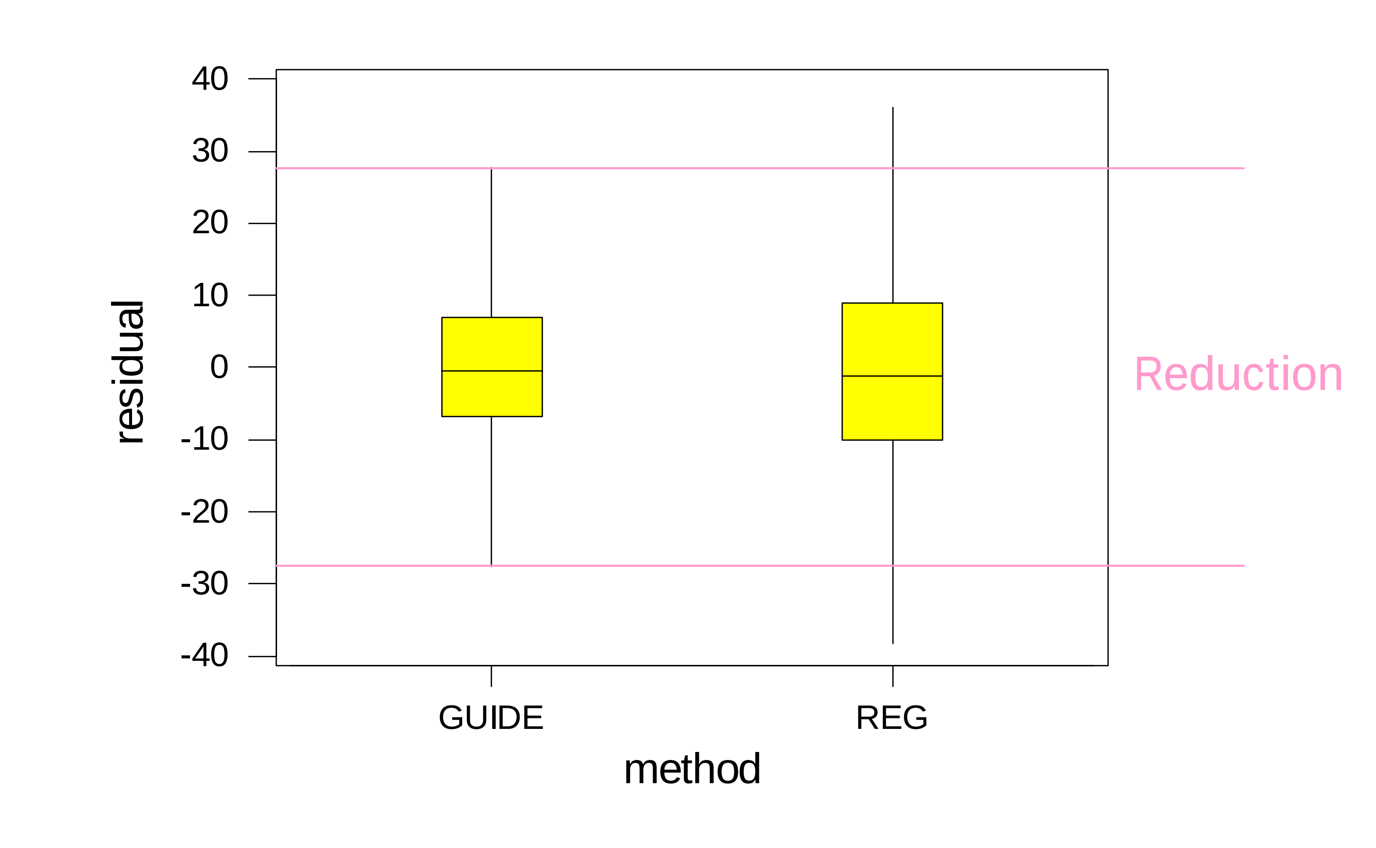
"Regression Only" Coefficient of Determination: R^2= 62.4%

"Guide" Coefficient of Determination; R^2 = 83.0%





GUIDE vs. Regression only

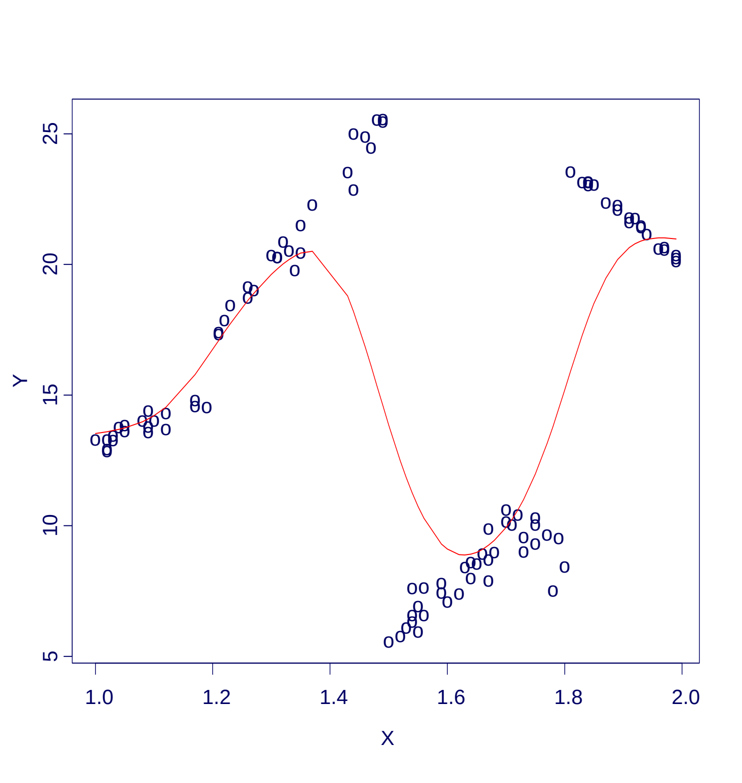


Note

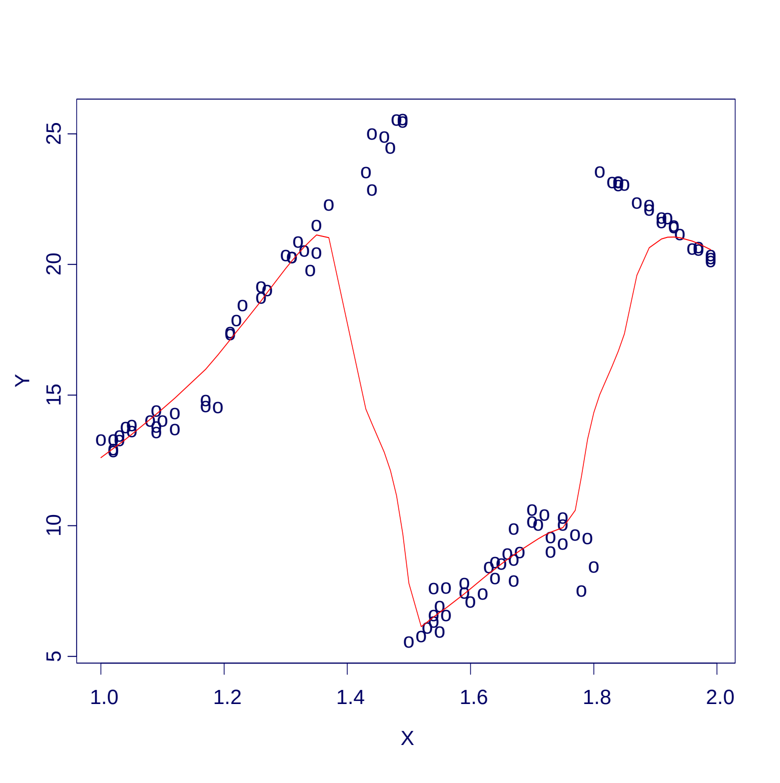
- GUIDE detected useful partitions
- Automatic procedure – little human interaction: wonderful exploratory tool!
- Easy Interpretation
  - follow the tree !
  - important variables appear in the tree
- Better fitted model & better predictions
- Missing data, simpler, interactions, etc.

Decision Trees may be more helpful in analyzing industrial processes than "Global Modeling"

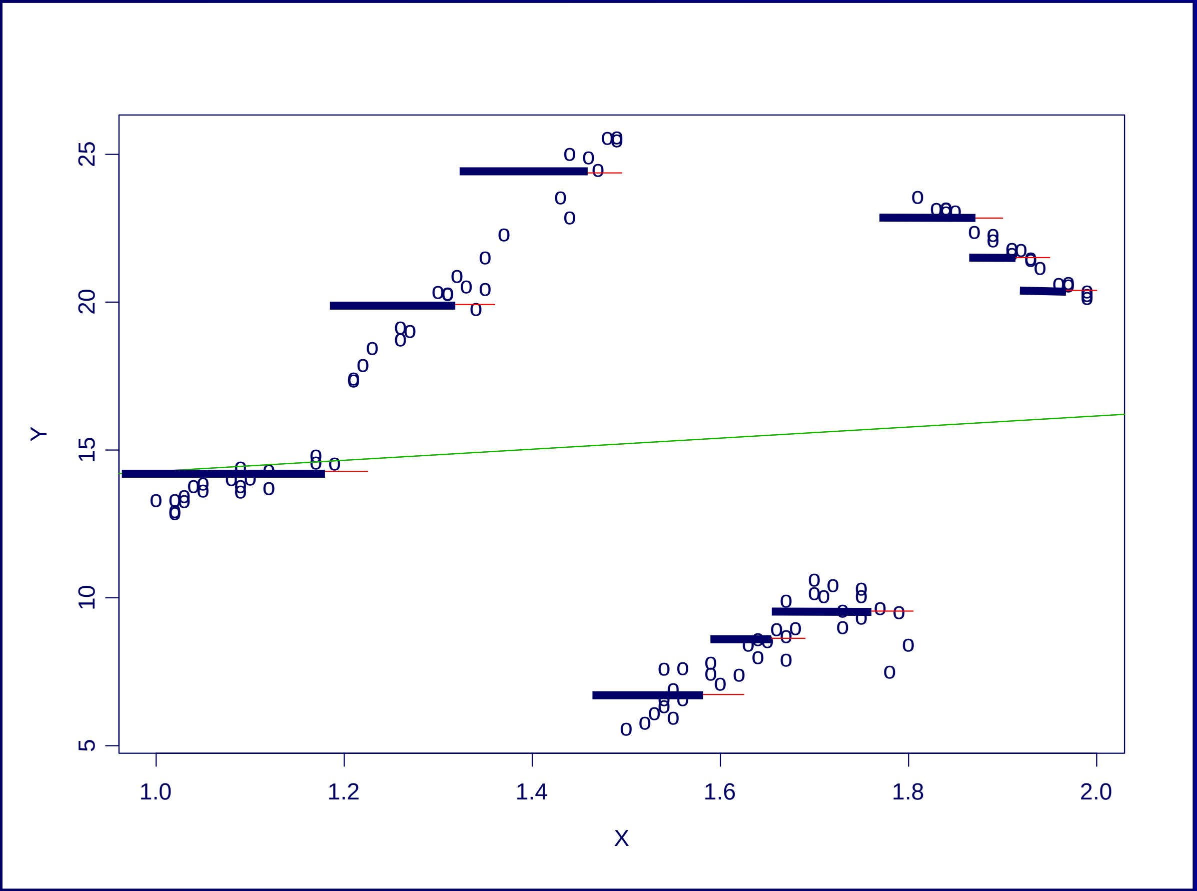
Smoothing Spline



Loess



CART (Breiman *et al.* 1984) “Classification and Regression Tree”



Some Limitations of CART: mean function fit

DT Modeling

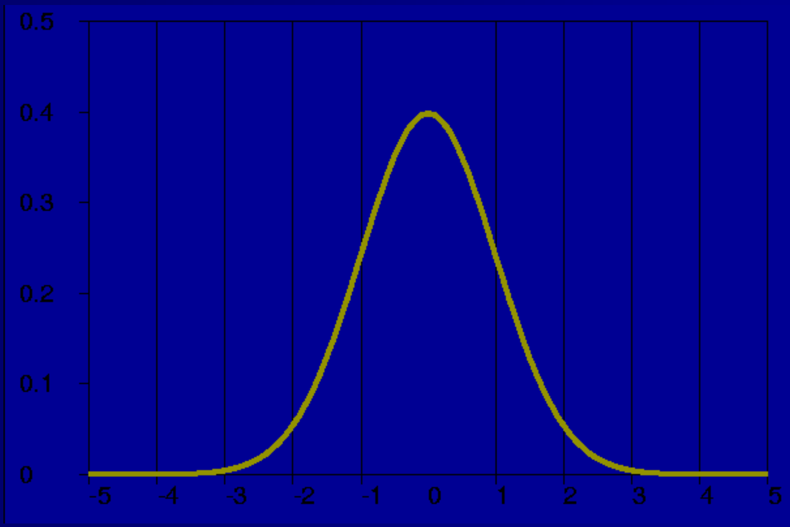
- Assume different (multiple) correlation structure exists within database
- Tree-based model = data partitioning + statistical modeling + pruning
- Intermediate nodes = partition data space
- Terminal nodes = fit the Node model (example: linear regression, loess, mean)
- Pruning = prevent over-fitting
- "Given a Node model, how to find the partition (heterogeneity) of the data space?"

Conclusions

- Decision Trees, e.g., GUIDE, etc., helpful
- Note there are data mining approaches other than decision trees, such as neural nets, genetic algorithms, etc. (dependent on problem and data heterogeneity)
- Decision Trees are attractive because they show clearly how to reach a decision, and easy to interpret by practitioners

Decision Trees (DT)
Philosophy of Reducing Variation

- □ If you cannot quantify variation, how will you reduce variation?



- □ Once variation is quantified, what are sources of variation?
- □ DT is a useful method for identifying sources of variation.

Links & Future work

GUIDE:

http://www.stat.wisc.edu/~loh/guide.html

CRUISE:

http://web.utk.edu/~hjkim/

Paper:

http://stat.bus.utk.edu/techrpts/index.html