

# **Interactive Data Mining and DOE: Tools for the Next Wave in Six Sigma**

## **7th Annual Six Sigma Summit**

### **January 25, 2006**

Marie Gaudard, Ph.D., Phil Ramsey, Ph.D., Mia L. Stephens, MS

North Haven Group

<http://www.northhavengroup.com/>

[pramsey@northhavengroup.com](mailto:pramsey@northhavengroup.com), 603-672-5651

[mstephens@northhavengroup.com](mailto:mstephens@northhavengroup.com), 207-363-5739

[mgaudard@northhavengroup.com](mailto:mgaudard@northhavengroup.com), 352-560-0312

# Talk Outline

- DMAIC, Data Mining, and DOE
- What is Data Mining?
- JMP<sup>®</sup> and Data Mining
- The Press Band Data - Partitioning
- JMP<sup>®</sup> DOE Platforms
- The Press Band Data – DOE
- Case Study: Defect Reduction
- Summary

# DMAIC, Data Mining, and DOE

In both transactional and manufacturing Six Sigma situations, large observational data sets relating to the processes of interest are often available.

These data sets can be “mined” in order to:

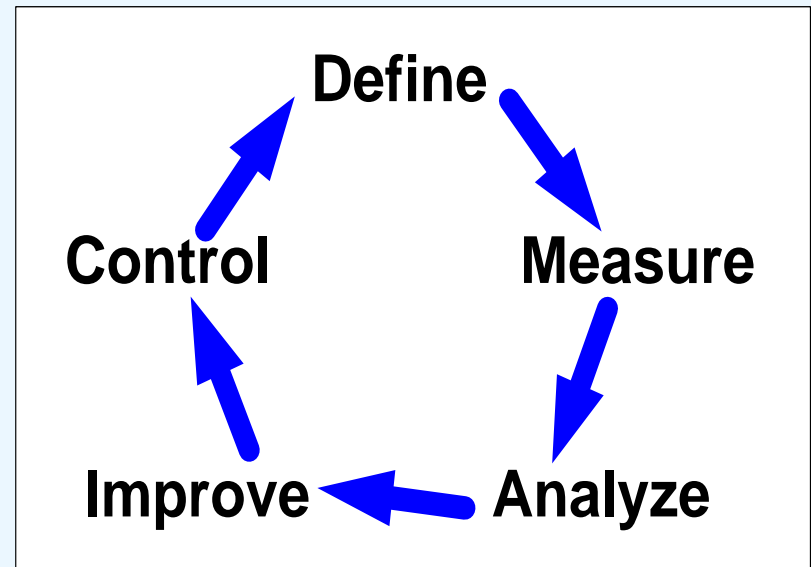
- Identify well-scoped Six Sigma projects;
- Provide background information on relationships between predictors and responses, prior to further data collection or simply as background knowledge;
- Suggest causal relationships and potential solutions;
- Identify anomalies;
- Reduce the number of predictors to be studied.

# DMAIC, Data Mining, and DOE

As such, data mining can support the Define, Measure, Analyze, and Improve phases of DMAIC.

This talk will focus on the use of a data mining technique, called recursive partitioning, in the Improve Phase.

Specifically, we will see how partitioning and design of experiments synergize.



# What Is Data Mining?

**Data Mining** is the analysis of large observational data sets with the goal of finding unsuspected relationships.

By “large” data sets, we mean either a **large number of records**, or a **large number of variables measured on each record**.

The other key word in the definition is “observational”.

- Often, data sets used in data mining were collected for purposes other than those of the data mining study.
- Consequently, data mining data sets often lack the integrity and relevance found in data sets collected as part of a well-defined study.

# What Is Data Mining?

Types of algorithms typically associated with data mining include:

- Multiple linear and logistic regression,
- Classification and regression trees,
- Neural nets,
- Clustering algorithms, and
- Association rules.

However, other techniques are often used as well, such as:

- Extensive display and visualization tools,
- Variable reduction techniques, and
- Bayesian methods.

# JMP<sup>®</sup> and Data Mining

JMP<sup>®</sup> provides a number of data mining tools for **modeling**:

- Multiple linear and logistic regression
- Classification and regression trees (the Partition platform)
- Neural nets
- Cluster analysis

In terms of **visualization**, almost all JMP analyses are supported by extensive display and visualization tools.

JMP also provides convenient methods for **preprocessing** data.

# JMP<sup>®</sup> and Data Mining

**JMP's Neural Net platform** fits a neural net with one hidden layer to a continuous or categorical response.

**JMP's Partition platform** is a classification and regression tree-fitting methodology.

Other tree-fitting methodologies, found in high-end and very expensive data mining packages, are CART<sup>™</sup>, CHAID<sup>™</sup>, C5.0.

Since convenience data sets are often messy and unruly, JMP's display and preprocessing capabilities support the user in data-cleaning.

# JMP<sup>®</sup> and Data Mining

**JMP's Partition platform** is a version of Classification and Regression Tree Analysis.

Both response and factors can be either continuous or categorical.

**Continuous factors** are **split** into **two partitions** according to **cutting values**.

**Categorical factors** are **split** into **two groups of levels**.

If the **response is continuous**, the fitted values are the **means** within groups.

If the **response is categorical**, the fitted values are the **response rates** (estimated probabilities) within groups.

# JMP<sup>®</sup> and Data Mining

The splits are determined by maximizing the LogWorth criterion, which reflects the degree of separation for a potential split.

For a continuous response, the LogWorth is related to the sums of squares due to the differences between means.

For a categorical response, it is related to the likelihood ratio chi-square statistic.

JMP's Partition platform is extremely useful **for both exploring relationships and for modeling.**

JMP's Partition platform provides only a minimal criterion to determine when to stop building a tree (a **stopping rule**). We have found this advantageous.

# The Press Band Data - Partitioning

We will illustrate JMP's partition platform using a data set from the rotogravure printing business.

In this printing process:

- An engraved copper cylinder is rotated in a bath of ink,
- Excess ink is removed,
- Paper is pressed against the inked image,
- Once the job is complete, the engraved image is removed from the cylinder, and the cylinder is reused.

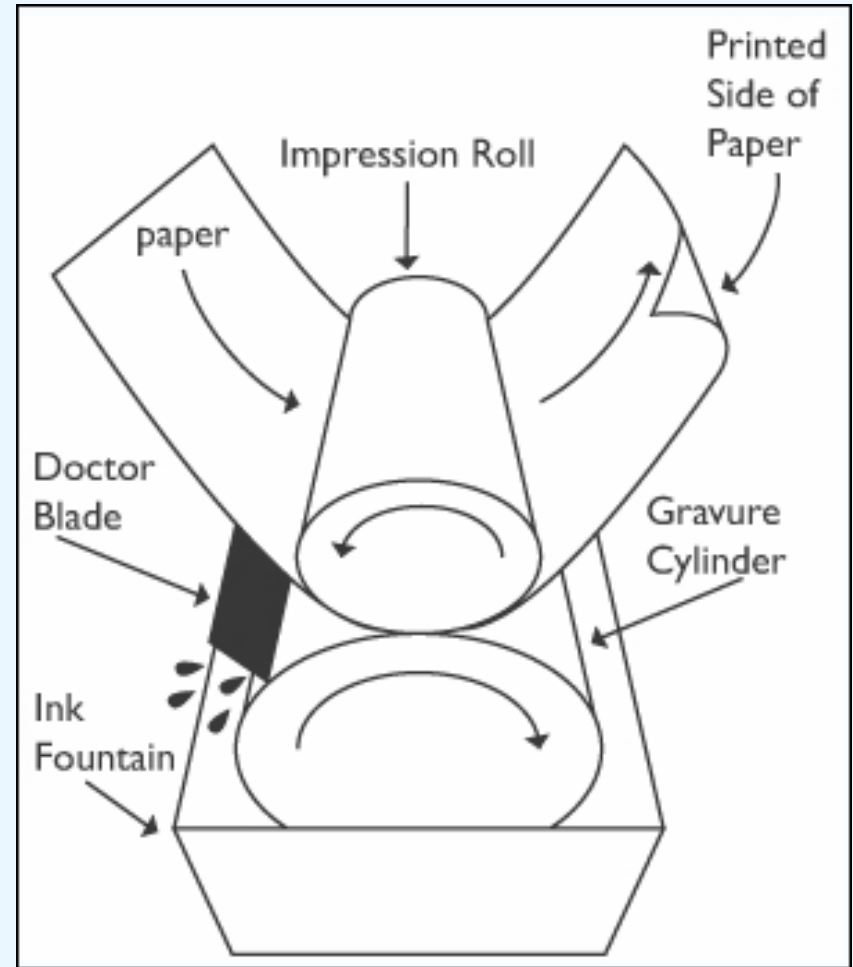
# The Press Band Data - Partitioning

The image to be printed is incised into the cylinder.

The cylinder is rotated through the ink fountain.

The doctor blade acts as a squeegee, removing excess ink.

The paper is pressed against the cylinder by the impression roll, transferring the image.



# The Press Band Data - Partitioning

A defect called **banding** can sometimes occur, ruining the product.

Banding consists of grooves that appear in the cylinder at some point during the print run.

Once detected, the run is halted, and the cylinder is removed and repaired (repolished or replated).

This process can take from one-half to six hours.

Understanding the conditions that lead to banding is critical and could save a printer enormous amounts of money.

# The Press Band Data - Partitioning

We will use a data set that contains observational data on banding.

This data set can be found at <http://ftp.ics.uci.edu/pub/machine-learning-databases/> and is called *cylinder-bands*.

The target variable is “Band Occurred?”, whose values are BAND and NOBAND.

We will tell the story of a fictional Six Sigma project team, which is charged with reducing banding defects.

# The Press Band Data - Partitioning

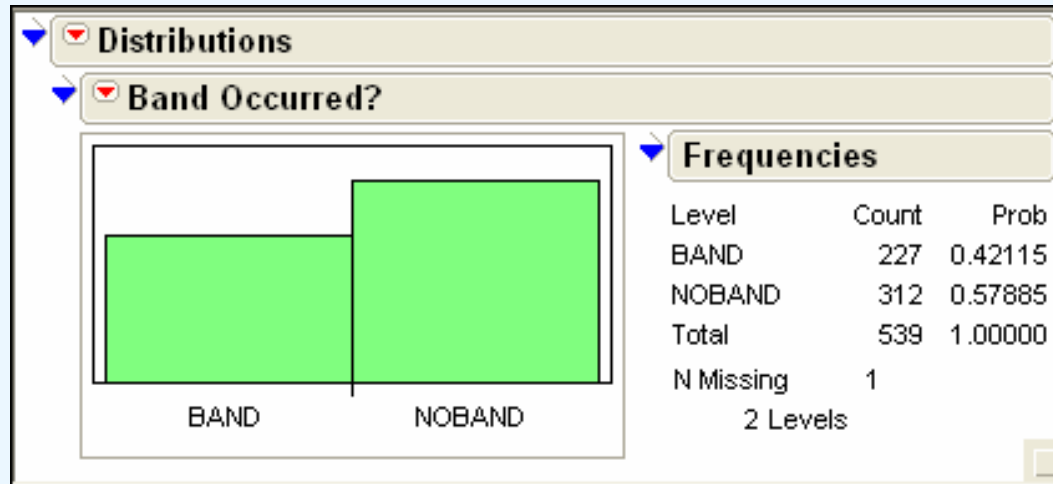
A portion of the Press Band data set.

There are 540 records and 39 variables.

PressBandData									
PressBandData		Date	Date M/Y	Job Number	Cylinder No.	Customer	grain screened	proof on ctd ink	paper type
Partition	1	03/30/1990	03/1990	23040	X750	GUIDEPOSTS	YES	YES	UNCOATED
Columns (39/0)	2	04/09/1990	04/1990	34683	G467	ECKERD	NO	YES	COATED
solvent type	3	04/09/1990	04/1990	25416	X203	TVGUIDE	YES	YES	UNCOATED
type on cylinder	4	04/14/1990	04/1990	34545	O21	TARGET	NO	YES	COATED
press type	5	04/17/1990	04/1990	36858	T313	EXXON	YES	YES	UNCOATED
press	6	04/18/1990	04/1990	36053	J68	WARDS	NO	YES	COATED
cylinder size	7	04/18/1990	04/1990	36053	J42	WARDS	NO	YES	COATED
location	8	04/18/1990	04/1990	36858	F329	EXXON	YES	YES	UNCOATED
plating tank	9	04/25/1990	04/1990	34664	G496	BURDINES	YES	YES	UNCOATED
proof cut	10	04/26/1990	04/1990	34545	O6	TARGET	NO	YES	UNCOATED
viscosity	11	04/26/1990	04/1990	34545	O14	TARGET	NO	YES	UNCOATED
caliper	12	05/05/1990	05/1990	47103	T244	MODMAT	YES	YES	UNCOATED
ink temperature	13	05/07/1990	05/1990	47103	M93	MODMAT	NO	YES	COATED
humidity	14	05/07/1990	05/1990	47103	M260	MODMAT	YES	YES	UNCOATED
roughness	15	05/07/1990	05/1990	47103	T383	MODMAT	NO	YES	COATED
blade pressure	16	05/07/1990	05/1990	47103	T78	MODMAT	YES	YES	UNCOATED
varnish pct	17	05/07/1990	05/1990	47103	M4	MODMAT	YES	YES	UNCOATED
press speed	18	05/07/1990	05/1990	36928	M432	HOMESHOPPIN	NO	YES	COATED
ink pct	19	05/07/1990	05/1990	36928	M257	HOMESHOPPIN	NO	YES	COATED
solvent pct	20	05/09/1990	05/1990	47103	F242	MODMAT	YES	YES	UNCOATED
ESA Voltage	21	05/10/1990	05/1990	47103	F672	MODMAT	YES	YES	UNCOATED
wax	22	05/11/1990	05/1990	47103	M260	MODMAT	YES	YES	UNCOATED
hardener	23	05/14/1990	05/1990	47103	F679	MODMAT	YES	YES	UNCOATED
roller durometer	24	05/17/1990	05/1990	36054	X400	WARDS	NO	YES	COATED
current density	25	05/17/1990	05/1990	34752	X776	TOYSRUS	NO	NO	COATED
anode space ratio	26	05/17/1990	05/1990	34752	X713	TOYSRUS	NO	NO	COATED
chrome content	27	05/18/1990	05/1990	34402	I331	AUSTADS	YES	YES	UNCOATED
Band Occurred?	28	05/24/1990	05/1990	36648	F227	JAMESWAY	NO	YES	UNCOATED
Rows	29	06/02/1990	06/1990	36859	F590	NATLWILDLIFE	YES	YES	UNCOATED
All rows	30	06/03/1990	06/1990	36859	F670	NTLWILDLIFE	YES	YES	UNCOATED
Selected	31	06/06/1990	06/1990	36859	F331	NATLWILDLIFE	YES	YES	UNCOATED
Excluded	32	06/06/1990	06/1990	36859	F571	NATLWILDLIFE	YES	YES	UNCOATED
Hidden									
Labelled									

# The Press Band Data - Partitioning

The team used JMP's Distribution Platform to generate the analysis below, which indicates that banding occurred in 42% of jobs.



The team would like to understand the root causes of banding.

# The Press Band Data - Partitioning

In the Measure phase, a Six Sigma team often constructs a cause and effect diagram to determine potential root causes of the defect (banding); the team then collects data, and constructs Pareto charts to determine where to focus efforts.

**But Pareto charts overlook complex relationships and interactions** among possible explanatory variables.

In our example, the Six Sigma team has a large observational data set available, and the team will use it to construct some knowledge about the process.

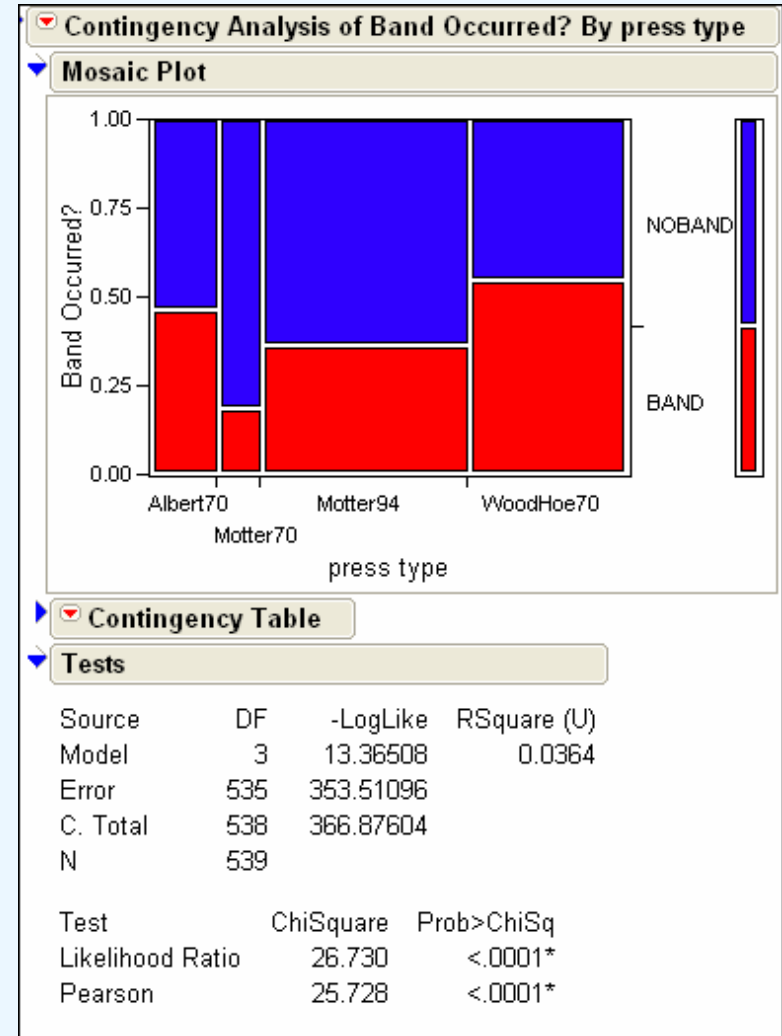
The available predictors for banding consist of 11 categorical variables and 18 continuous variables.

# The Press Band Data - Partitioning

The team can explore two-way relationships between the predictors and “Band Occurred?” using:

- Mosaic plots for categorical predictors;
- Comparison boxplots for continuous predictors.

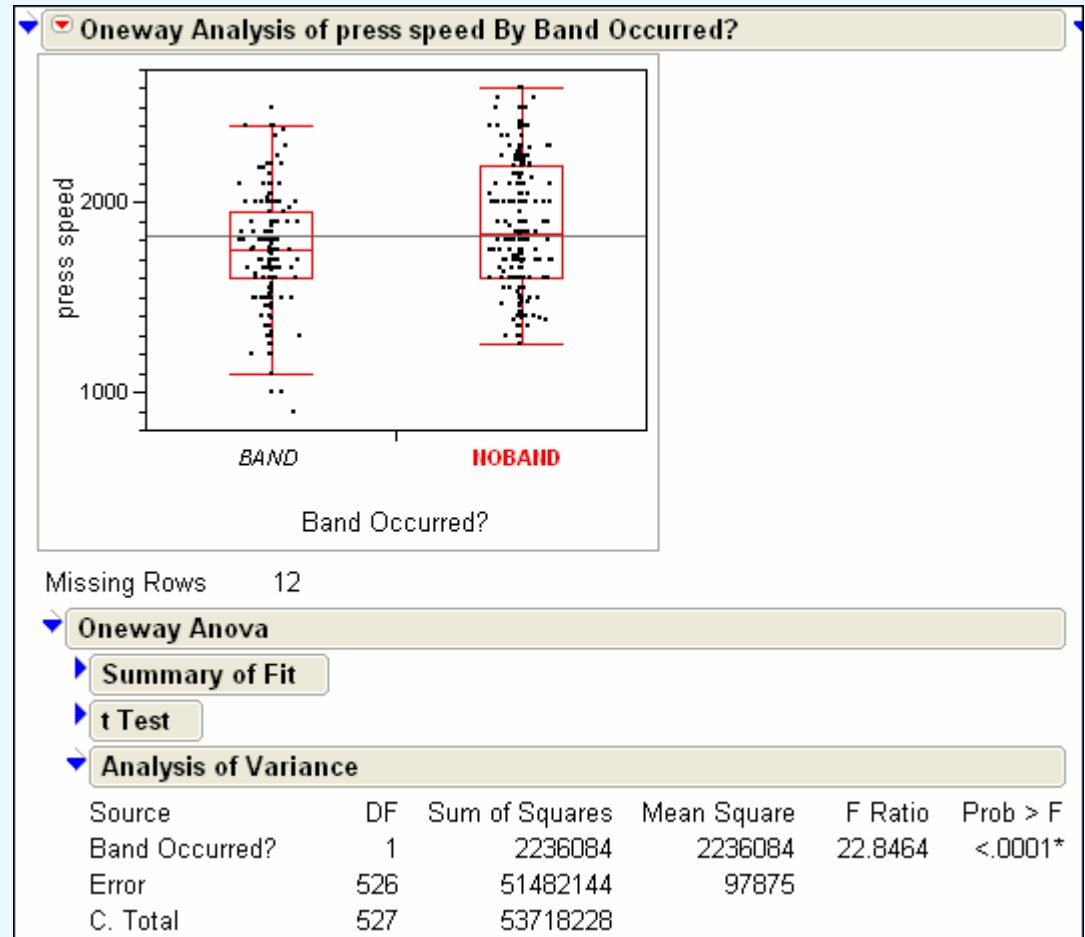
A Contingency Analysis, with a mosaic plot, is shown to the right. The red bars indicate the proportion of runs with banding.



# The Press Band Data - Partitioning

A Oneway Analysis, showing comparison boxplots, is shown to the right.

*But*, although both analyses show significant relationships, complex interactions are ignored.



# The Press Band Data - Partitioning

The team could attempt to fit a logistic regression model, but including all the predictors is not reasonable.

This is because of the many categorical variables, and the fact that numerous cells are not populated.

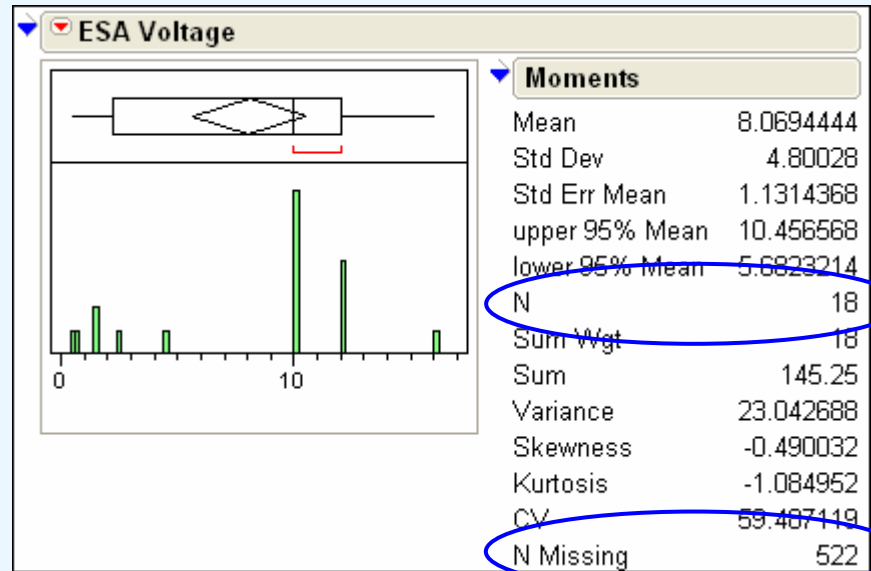
However, **the team can easily construct a classification tree**, with “Band Occurred?” as target.

This analysis will provide rich information on the conditions that lead to banding.

# The Press Band Data - Partitioning

We note, at this point, that observational data sets must always be examined for data integrity before they are analyzed.

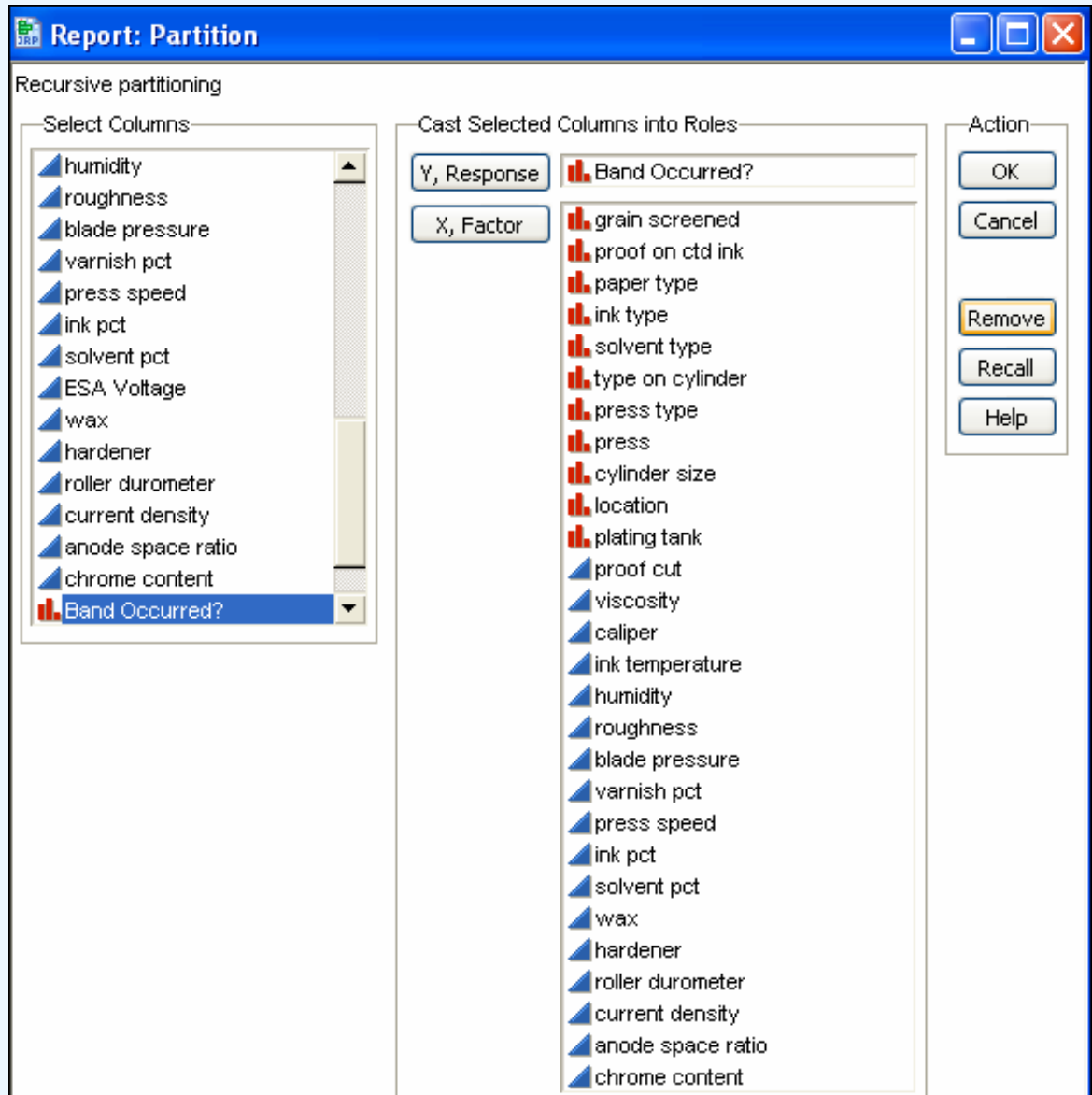
For example, the variable ESA Voltage is missing for all but 18 records:



It is unreasonable to include this variable in modeling efforts, and the team will not include it in its classification model.

A classification model is requested using the Partition menu in JMP.

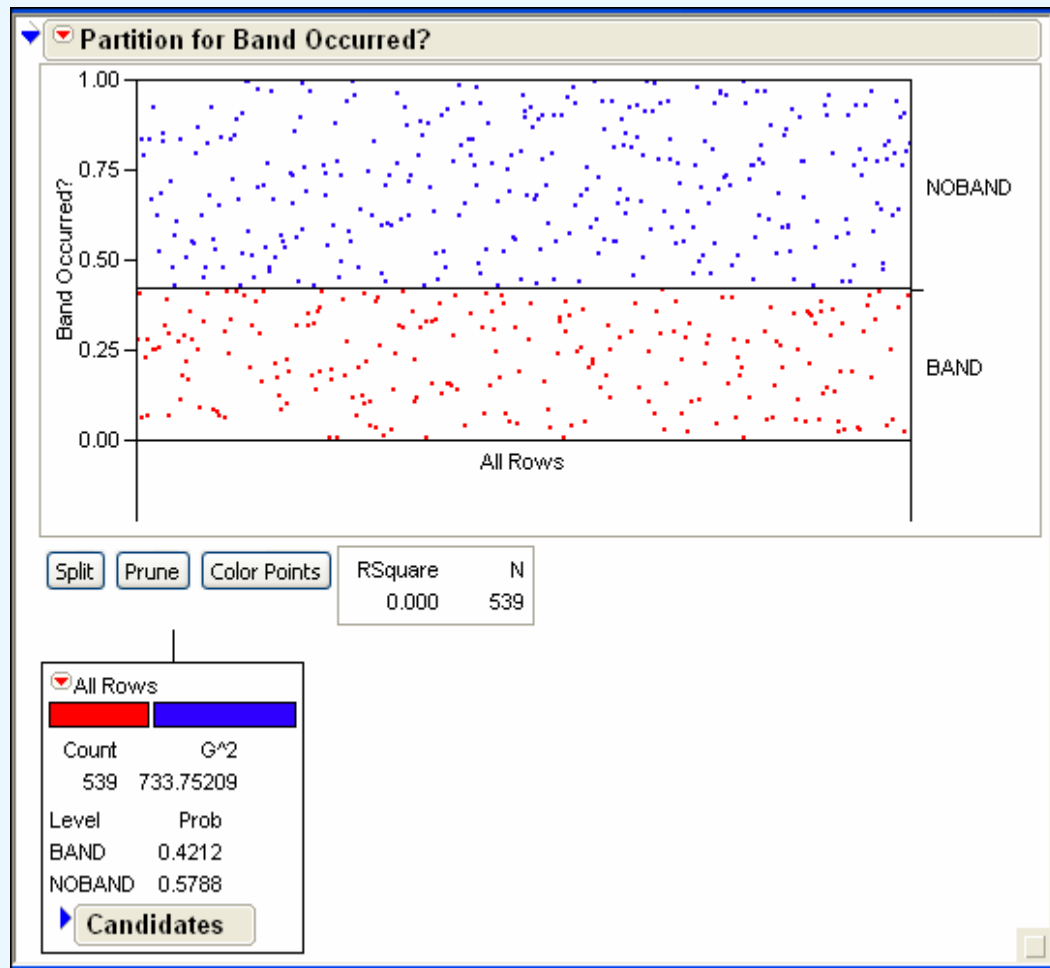
“Band Occurred?” is the response, and 28 variables are input as candidate predictors.



The Partition report to the right opens.

Points corresponding to the runs are jittered in such a way that runs with banding are red, and are in the area of the graph beneath the horizontal divider at 42.12%.

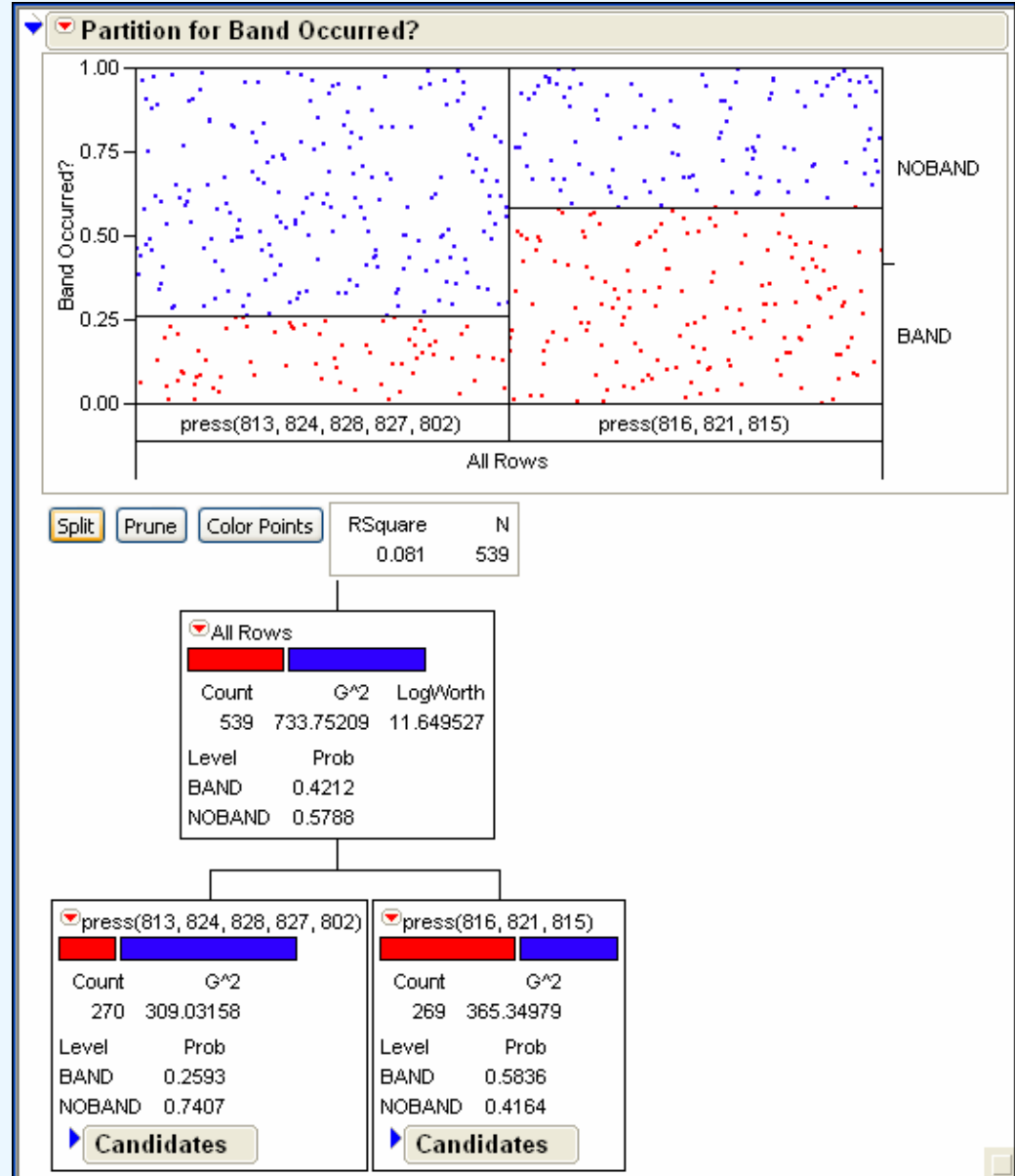
Blue points (no banding) are shown above the line.



We split once.

Note that JMP chooses the variable “press” as the splitting variable.

The split places four presses in a group where NOBAND is predicted, and the three other presses in a group where BAND is predicted.



Based on this split, the team learns that some presses are more affected by banding than others.

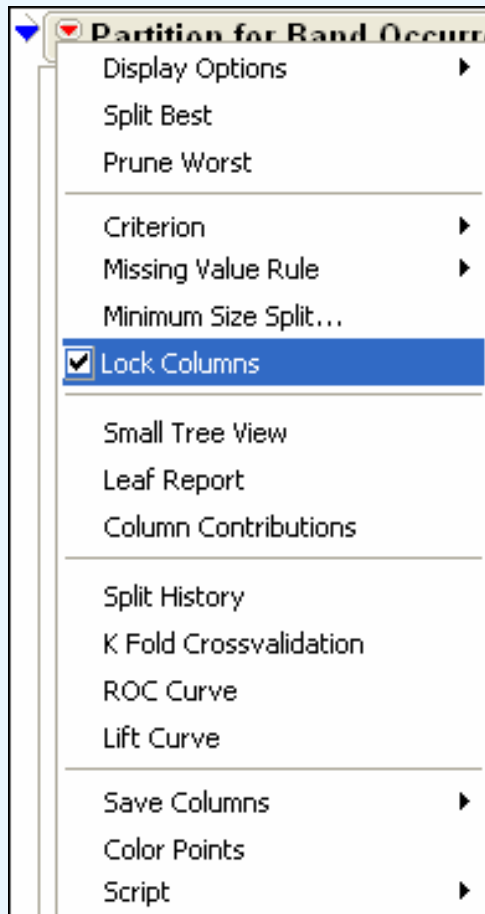
But this does not help the team in identifying root causes of banding.

When splits occur on variables that are only tangentially useful relative to process improvement actions, one can **force attention to more useful predictors** by excluding the tangential variables from the partitioning algorithm.

This is done by selecting the **Lock columns** option in the Partition platform menu.

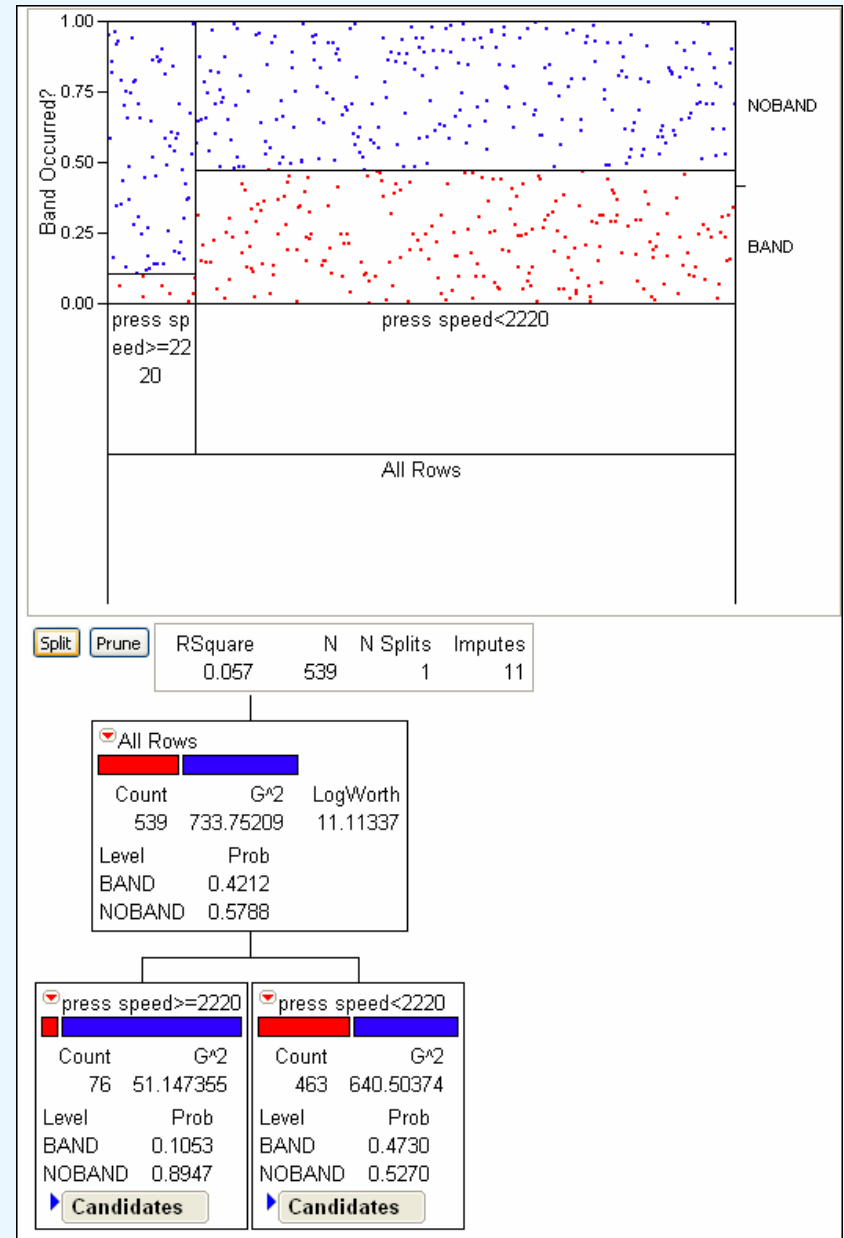
The team “prunes” back the initial split, and then locks the “press” column to prevent it from being used as a partition variable.

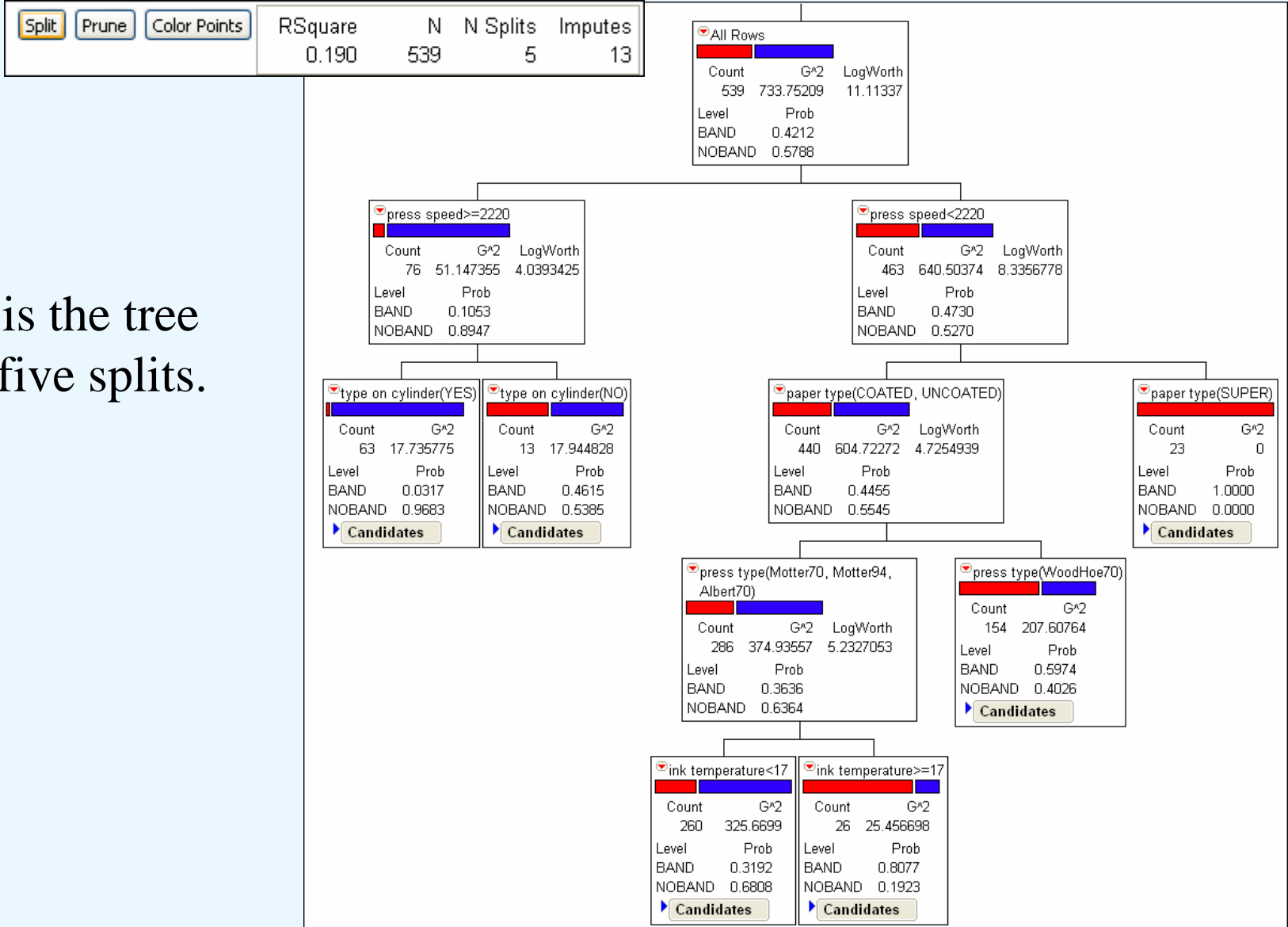
To lock the “press” column, first select Lock columns, then select “press” from the column list.



Term	Lock	Term	Lock
grain screened	<input type="checkbox"/>	roughness	<input type="checkbox"/>
proof on ctd ink	<input type="checkbox"/>	blade pressure	<input type="checkbox"/>
paper type	<input type="checkbox"/>	varnish pct	<input type="checkbox"/>
ink type	<input type="checkbox"/>	press speed	<input type="checkbox"/>
solvent type	<input type="checkbox"/>	ink pct	<input type="checkbox"/>
type on cylinder	<input type="checkbox"/>	solvent pct	<input type="checkbox"/>
press type	<input type="checkbox"/>	wax	<input type="checkbox"/>
press	<input checked="" type="checkbox"/>	hardener	<input type="checkbox"/>
cylinder size	<input type="checkbox"/>	roller durometer	<input type="checkbox"/>
location	<input type="checkbox"/>	current density	<input type="checkbox"/>
plating tank	<input type="checkbox"/>	anode space ratio	<input type="checkbox"/>
proof cut	<input type="checkbox"/>	chrome content	<input type="checkbox"/>
viscosity	<input type="checkbox"/>		
caliper	<input type="checkbox"/>		
ink temperature	<input type="checkbox"/>		
humidity	<input type="checkbox"/>		

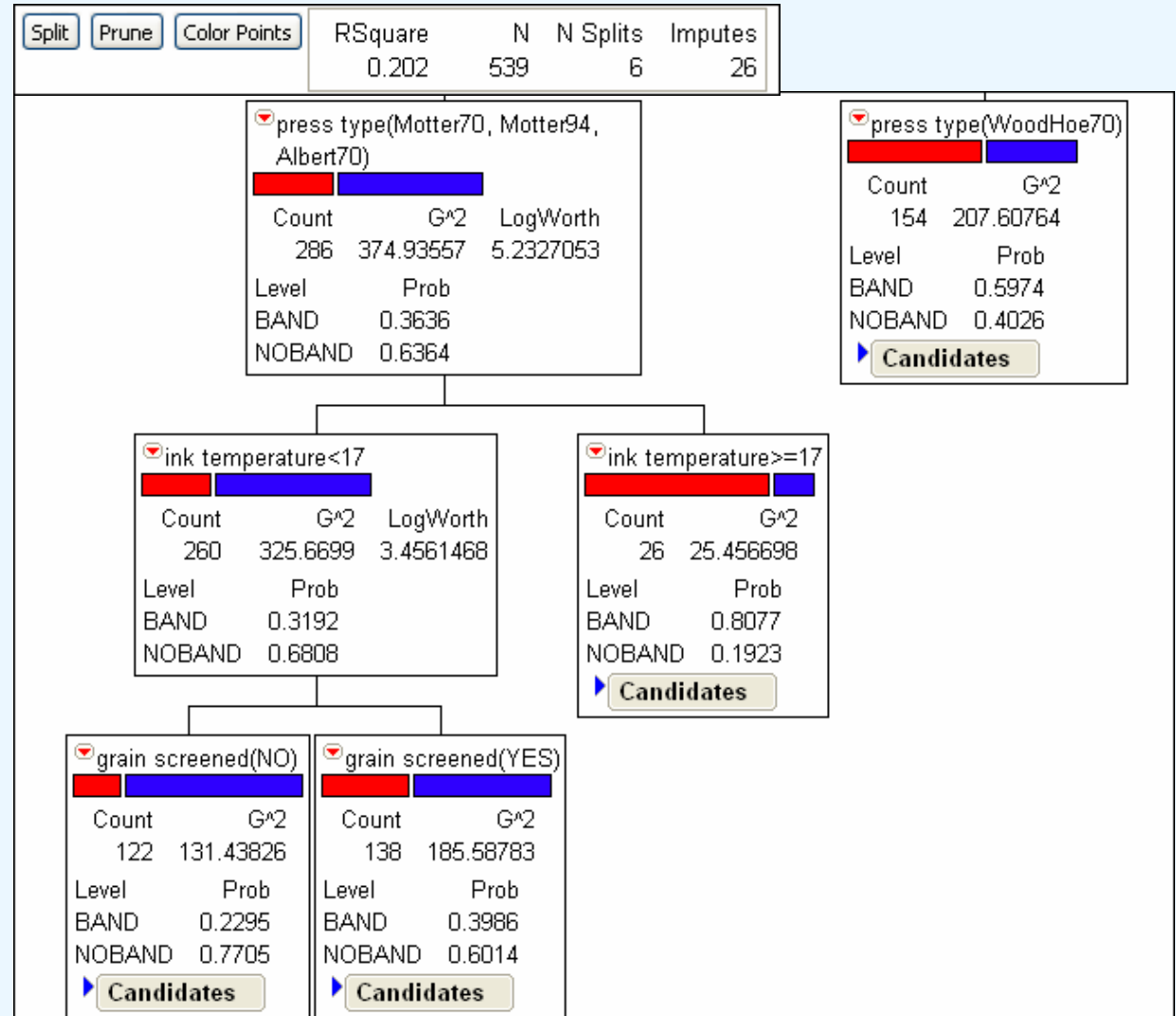
With the “press” variable locked, the first split now occurs on the variable “press speed”.





Here is the tree after five splits.

A sixth split selects “grain screened” as the partitioning variable in the “ink temperature < 17” branch.



# The Press Band Data - Partitioning

Note that splits have occurred both on categorical and continuous predictors.

Note also that 26 values have been imputed.

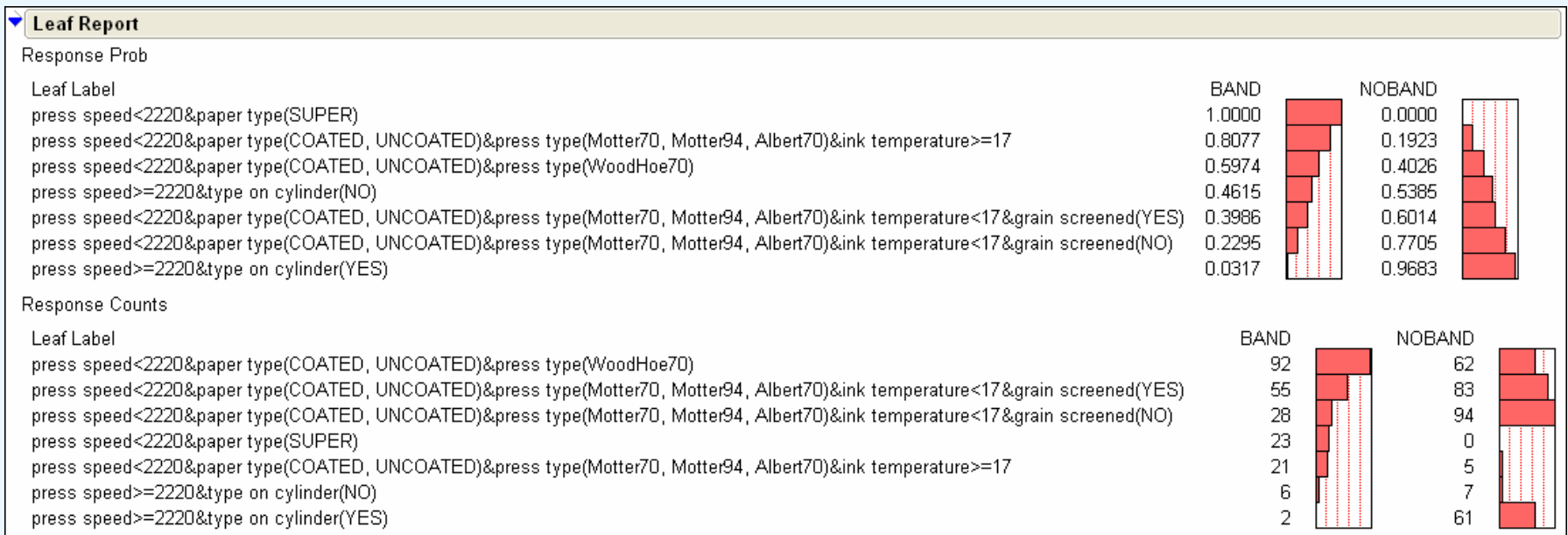
This means that certain variables used in the analysis had missing values, and so values, computed according to a selected rule, have been substituted.

The user controls splitting; at each split, JMP provides the best splitting variable and grouping of levels for that variable.

# The Press Band Data - Partitioning

As trees get large, they become visually intractable.

JMP provides a Leaf Report, which gives the rule set and a graphic display of the terminal nodes' discriminatory ability.



# The Press Band Data - Partitioning

Formulas for the predicted probabilities, leaf numbers, and leaf labels (rule set) can be saved to columns in the JMP data table.

(Note that the leaf labels have been truncated.)

	Prob(Band Occurred?==BAND)	Prob(Band Occurred?==NOBAND)	Leaf Label
1	0.39855072	0.60144928	press speed<2220&paper type(COATED, UNCOATED)&press type(Motter70, Motter94, Albert70)&ink temperature<1
2	0.5974026	0.4025974	press speed<2220&paper type(COATED, UNCOATED)&press type(WoodHoe70)
3	0.80769231	0.19230769	press speed<2220&paper type(COATED, UNCOATED)&press type(Motter70, Motter94, Albert70)&ink temperature>=
4	0.2295082	0.7704918	press speed<2220&paper type(COATED, UNCOATED)&press type(Motter70, Motter94, Albert70)&ink temperature<1
5	0.39855072	0.60144928	press speed<2220&paper type(COATED, UNCOATED)&press type(Motter70, Motter94, Albert70)&ink temperature<1
6	0.03174603	0.96825397	press speed>=2220&type on cylinder(YES)
7	0.03174603	0.96825397	press speed>=2220&type on cylinder(YES)
8	0.39855072	0.60144928	press speed<2220&paper type(COATED, UNCOATED)&press type(Motter70, Motter94, Albert70)&ink temperature<1
9	0.5974026	0.4025974	press speed<2220&paper type(COATED, UNCOATED)&press type(WoodHoe70)
10	0.2295082	0.7704918	press speed<2220&paper type(COATED, UNCOATED)&press type(Motter70, Motter94, Albert70)&ink temperature<1
11	0.2295082	0.7704918	press speed<2220&paper type(COATED, UNCOATED)&press type(Motter70, Motter94, Albert70)&ink temperature<1
12	0.39855072	0.60144928	press speed<2220&paper type(COATED, UNCOATED)&press type(Motter70, Motter94, Albert70)&ink temperature<1



# The Press Band Data - Partitioning

JMP provides a number of guides that aid the user in deciding when to stop splitting. These include:

- $R^2$
- Column contributions analysis
- Minimum split size
- Lift curves
- ROC curves

We will not pursue these here, but suffice it to say that the user has complete flexibility in deciding when to stop splitting.

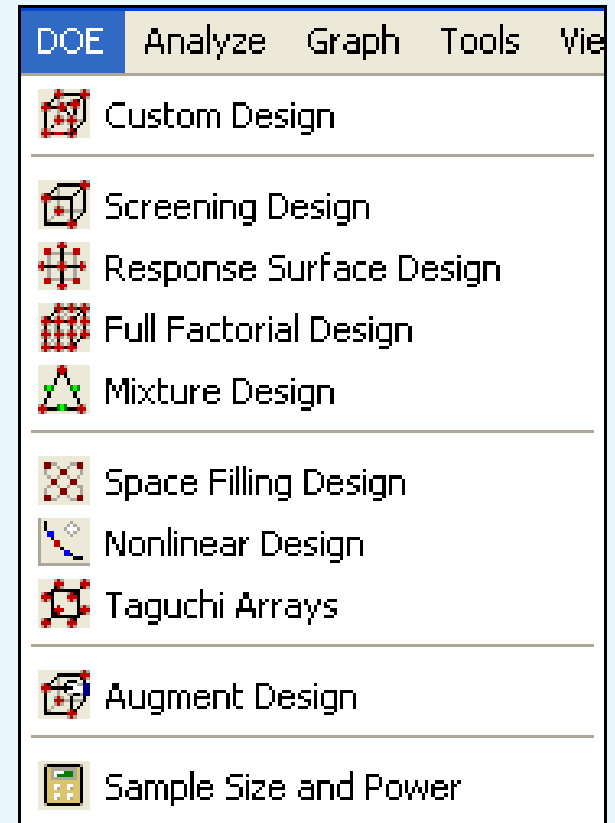
# JMP® DOE Platforms

JMP 6 provides the DOE options shown to the right.

Screening Design allows the user to define standard two-level full and fractional factorial designs, as well as Plackett-Burman designs.

Full Factorial Design allows the user to design multiple-level full factorials with categorial or continuous factors.

JMP also provides Response Surface, Mixture, and other design platforms.



# JMP<sup>®</sup> DOE Platforms

A new feature in JMP 6 is a greatly enhanced **Custom Design** platform.

This is an incredibly flexible structure for designing simple to complex experiments. It accommodates:

- Continuous and categorical factors with arbitrary numbers of levels;
- Hard/easy-to-change factors, mixture factors;
- Inequality constraints on factors;
- Covariates and uncontrollable variables;
- User-specified lists of interactions and polynomial terms to be estimated.

# The Press Band Data - DOE

Now our Six Sigma team begins to address root causes.

The predictors that appear to be of interest as a result of the six-split partition analysis are:

- Press type
- Type on cylinder
- Paper type
- Grain screened
- Press speed
- Ink temperature

# The Press Band Data - DOE

Although the partition analysis suggests an association of these predictors with banding, the team realizes that association is not causality.

The team decides to run a designed experiment to determine if these factors and their interactions have a **causal effect** on banding.

A big challenge facing the team is to **define a continuous measure for degree of banding**.

This is because an experiment based on a categorical response, such as BAND or NOBAND, will require a large (often prohibitive) number of runs to detect factor level differences.

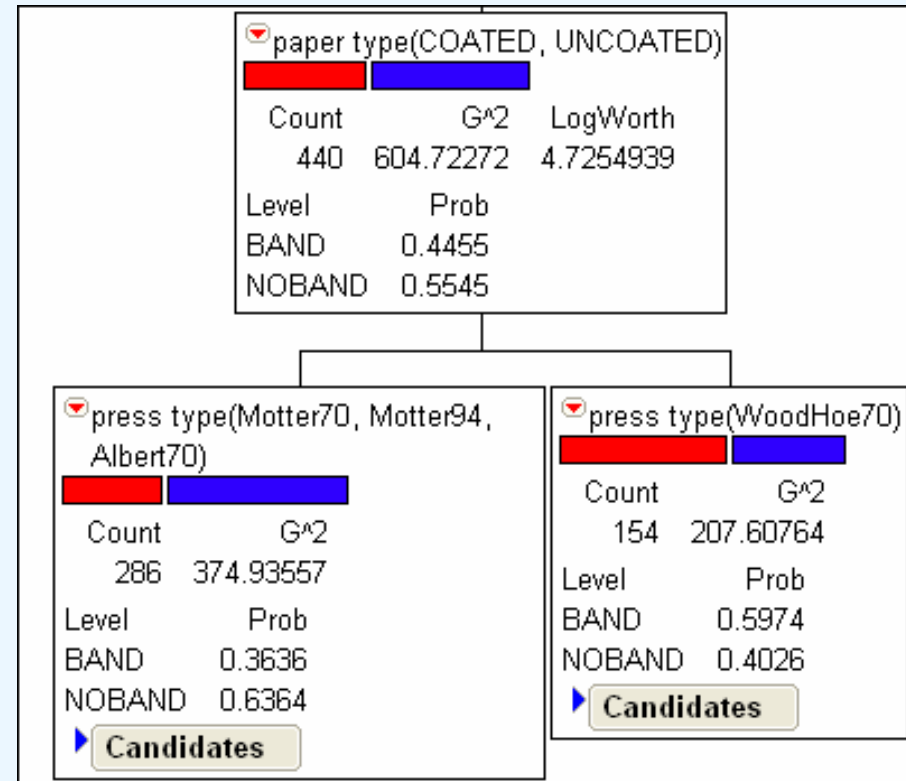
# The Press Band Data - DOE

A second challenge to the team is to **determine factor level settings**.

Here, the partition platform continues to be of value.

Press Type is one of the factors of interest.

Note that there are four Press Types, and that in our six-split partition model, these are split into two nodes as shown.



# The Press Band Data - DOE

Which Press Types should be included in the experiment?

At each node, the red arrow contains options relating to further splits at that node.

At the Press Type (Motter70, Motter94, Albert70) node, the team chooses Prune Below to undo splitting beyond this node.

Now the team chooses Split Specific to choose a further split on Press Type at the optimal split value.

The screenshot shows a JMP software interface. On the left, a menu is open for the variable 'press type(Motter70, Motter94, Albert70)'. The menu options are: Split Best, Split Here, Split Specific, Prune Below, Prune Worst, Select Rows, Show Details, and Candidates. The 'Candidates' button is highlighted. On the right, a 'Candidates' table is displayed with the following data:

Count	G^2
154	207.60764
Level	Prob
BAND	0.5974
NOBAND	0.4026

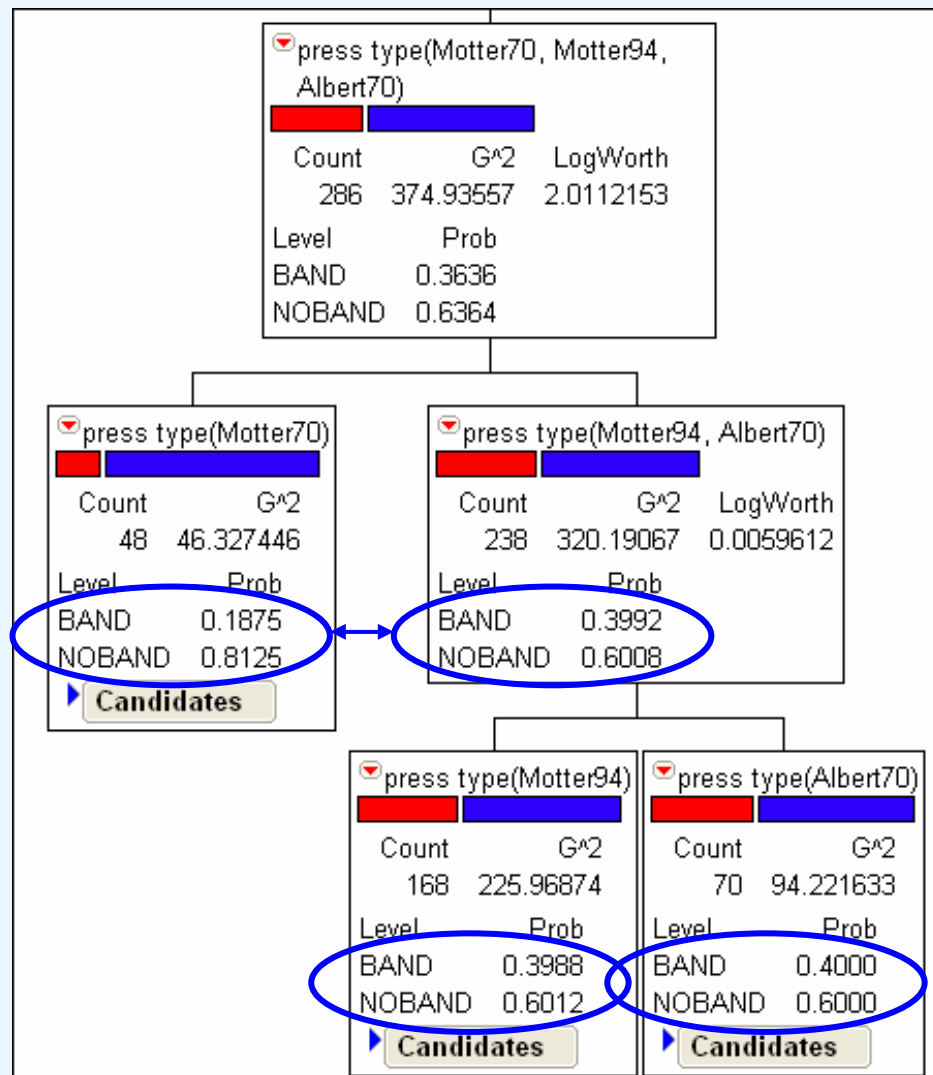
Below the table, there is a 'Candidates' button with a red arrow pointing to it.

The screenshot shows the 'Specify Split' dialog box in JMP. The list of variables includes: solvent type, type on cylinder, press type (highlighted), press, cylinder size, location, plating tank, proof cut, viscosity, caliper, and ink temperature. The 'Split at:' dropdown menu is set to 'optimal value'. The 'OK' and 'Cancel' buttons are visible on the right side of the dialog.

# The Press Band Data - DOE

This shows that Motter 70 presses and Motter 94 and Albert 70 presses are different relative to their effects on BAND at this point in the tree.

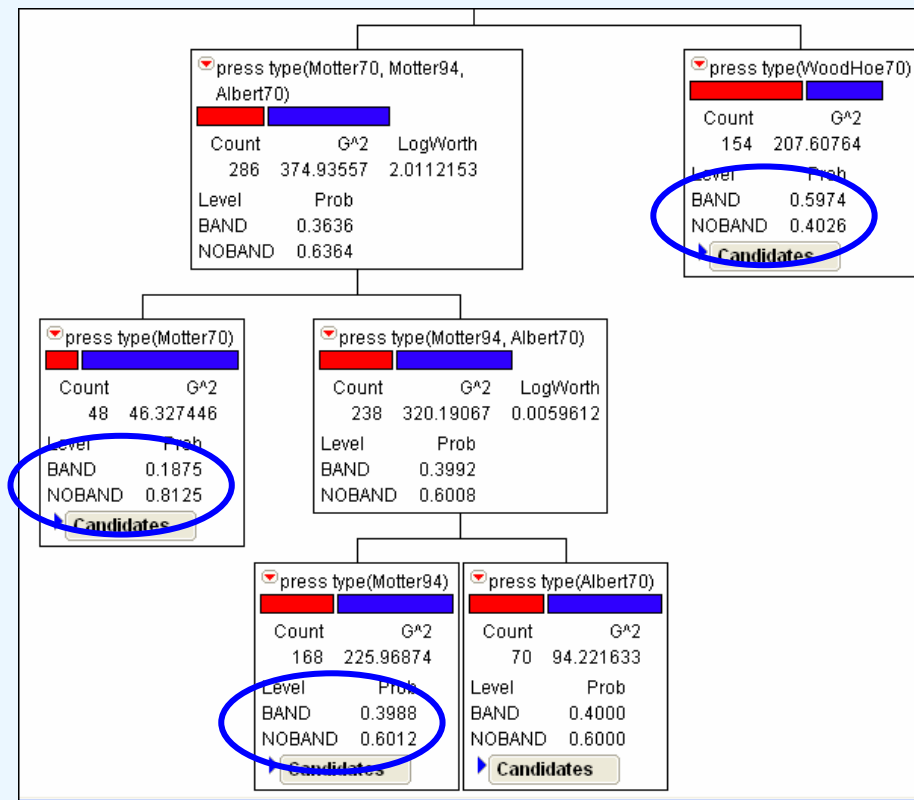
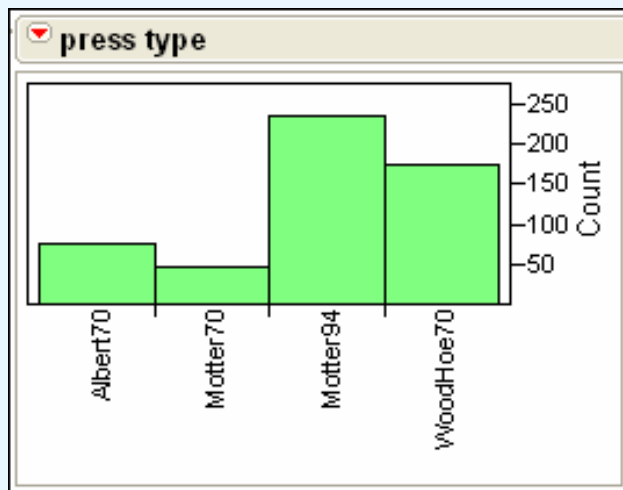
A further Split Specified at the Motter 94 and Albert 70 node indicates that these two press types appear to have a similar effect on BAND.



# The Press Band Data - DOE

Given this information and the bar chart below, the team decides on **three levels for Press Type**:

- Woodhoe 70,
- Motter 70,
- Motter 94.



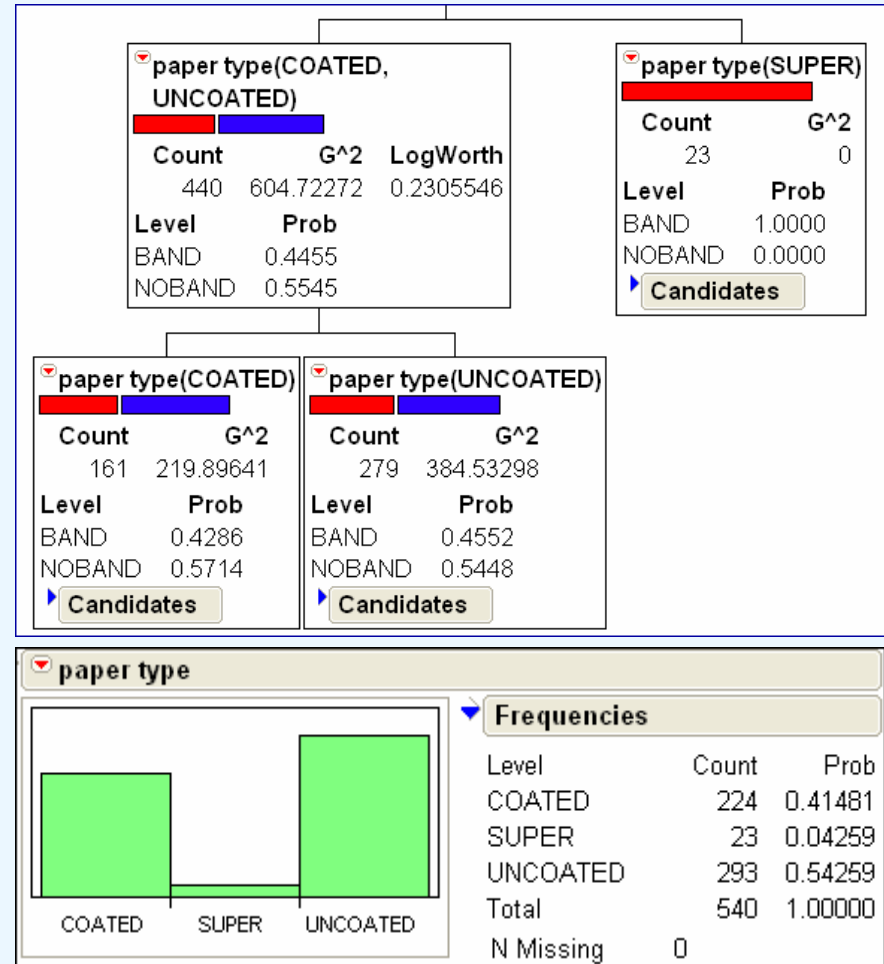
# The Press Band Data - DOE

The team now turns its attention to Paper Type.

Super paper is rarely used, but it is always affected by banding. However, the team learns that its use is being phased out.

Coated and Uncoated paper seem to be affected at about the same rates.

The team decides to **hold Paper Type constant at Uncoated** during the experiment.



# The Press Band Data - DOE

Once factor levels have been determined for all factors, the team must **determine the randomization scheme** for the experiment.

The **experimental factors** are:

- Press type
- Type on cylinder
- Grain screened
- Press speed
- Ink temperature

Complete randomization would require that factor-level settings be assigned randomly to runs, and that equipment be reset from scratch for each run.

# The Press Band Data - DOE

However, factors that involve the press setup will be difficult and time-consuming to change, while factors that can be manipulated within a press run will be easier to change.

The team determines that the **difficult to change** factors are:

- Press type
- Type on cylinder
- Grain screened

And that the **easy to change** (within press run) factors are:

- Press speed
- Ink temperature

# The Press Band Data - DOE

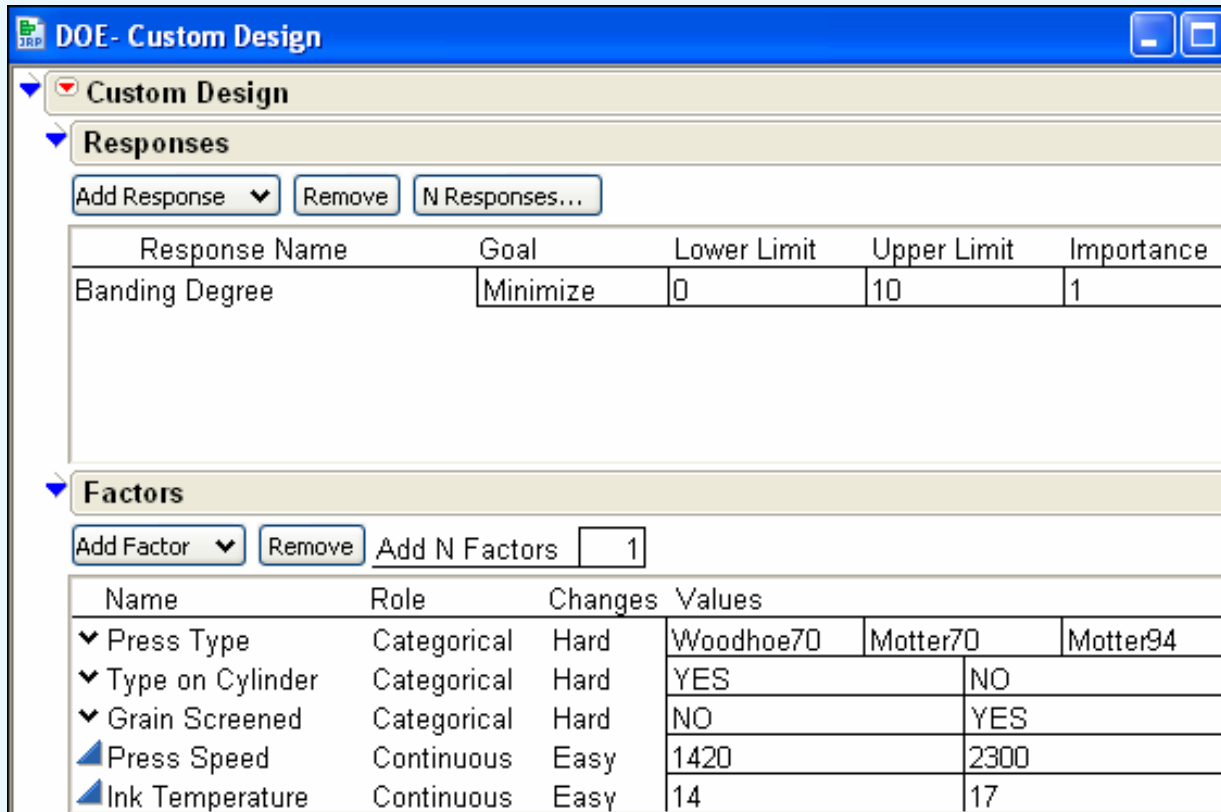
Note that the team faces a fairly complex design problem:

- A combination of continuous and categorical factors,
- Multiple-level factors,
- Hard and easy to change factors,
- Interactions that must be estimated.

The team uses the JMP Custom Design platform to facilitate the design process.

# The Press Band Data - DOE

The response and factors are added to the Custom Design list as shown.



The screenshot shows the 'DOE - Custom Design' window. It is divided into two main sections: 'Responses' and 'Factors'.

**Responses Section:**

- Buttons: Add Response (dropdown), Remove, N Responses...
- Table:

Response Name	Goal	Lower Limit	Upper Limit	Importance
Banding Degree	Minimize	0	10	1

**Factors Section:**

- Buttons: Add Factor (dropdown), Remove, Add N Factors (input: 1)
- Table:

Name	Role	Changes	Values
▼ Press Type	Categorical	Hard	Woodhoe70 Motter70 Motter94
▼ Type on Cylinder	Categorical	Hard	YES NO
▼ Grain Screened	Categorical	Hard	NO YES
▲ Press Speed	Continuous	Easy	1420 2300
▲ Ink Temperature	Continuous	Easy	14 17

# The Press Band Data - DOE

The team decides on a design that estimates **all two-way interactions**.

Note that the default design will require 24 runs, of which 12 will require changes to the press setup.

**Model**

Main Effects Interactions RSM Cross Powers Remove Term

Name	Estimability
Intercept	Necessary
Press Type	Necessary
Type on Cylinder	Necessary
Grain Screened	Necessary
Press Speed	Necessary
Ink Temperature	Necessary
Press Type*Type on Cylinder	Necessary
Press Type*Grain Screened	Necessary
Press Type*Press Speed	Necessary
Press Type*Ink Temperature	Necessary
Type on Cylinder*Grain Screened	Necessary
Type on Cylinder*Press Speed	Necessary
Type on Cylinder*Ink Temperature	Necessary
Grain Screened*Press Speed	Necessary
Grain Screened*Ink Temperature	Necessary
Press Speed*Ink Temperature	Necessary

**Design Generation**

Number of Whole Plots: 12

Number of Runs: 24

Minimum 21  
 Default 24  
 Compromise 36  
 Grid 48  
 User Specified .

# The Press Band Data - DOE

The design that the team will use is shown below.

Design							
Run	Whole Plots	Press Type	Type on Cylinder	Grain Screened	Press Speed	Ink Temperature	Banding Degree
1	1	Motter70	NO	NO	1420	17	.
2	1	Motter70	NO	NO	2300	14	.
3	2	Motter94	NO	NO	2300	17	.
4	2	Motter94	NO	NO	1420	14	.
5	3	Woodhoe70	NO	NO	2300	17	.
6	3	Woodhoe70	NO	NO	1420	14	.
7	4	Woodhoe70	YES	YES	1420	14	.
8	4	Woodhoe70	YES	YES	2300	17	.
9	5	Motter70	NO	YES	2300	17	.
10	5	Motter70	NO	YES	1420	14	.
11	6	Motter70	YES	YES	1420	17	.
12	6	Motter70	YES	YES	2300	14	.
13	7	Motter94	YES	YES	2300	17	.
14	7	Motter94	YES	YES	1420	14	.
15	8	Motter70	YES	NO	1420	14	.
16	8	Motter70	YES	NO	2300	17	.
17	9	Motter94	NO	YES	1420	17	.
18	9	Motter94	NO	YES	2300	14	.
19	10	Woodhoe70	YES	NO	2300	14	.
20	10	Woodhoe70	YES	NO	1420	17	.
21	11	Woodhoe70	NO	YES	1420	17	.
22	11	Woodhoe70	NO	YES	2300	14	.
23	12	Motter94	YES	NO	1420	17	.
24	12	Motter94	YES	NO	2300	14	.

# The Press Band Data - DOE

JMP conveniently saves the model that will be used to analyze the experiment in the data table.

When the team has entered responses from the experiment, the team will simply run this model.

The screenshot displays the JMP software interface for fitting a model. On the left, the 'Custom Design' table is visible, with the 'Model' column circled in blue. The main window shows the 'Report: Fit Model' dialog box. The 'Model Specification' section is expanded, showing the following details:

- Select Columns:** Whole Plots, Press Type, Type on Cylinder, Grain Screened, Press Speed, Ink Temperature, Banding Degree.
- Pick Role Variables:** Y (Banding Degree), Weight (optional Numeric), Freq (optional Numeric), By (optional).
- Personality:** Standard Least Squares
- Emphasis:** Minimal Report
- Method:** REML (Recommended)
- Unbounded Variance Components
- Estimate Only Variance Components
- Buttons:** Help, Run Model, Remove
- Construct Model Effects:**
  - Add: Whole Plots & Random
  - Cross: Press Type, Type on Cylinder, Grain Screened, Press Speed, Ink Temperature
  - Nest: Press Type\*Type on Cylinder, Press Type\*Grain Screened, Press Type\*Press Speed, Press Type\*Ink Temperature, Type on Cylinder\*Grain Screened, Type on Cylinder\*Press Speed, Type on Cylinder\*Ink Temperature
  - Macros: Grain Screened\*Press Speed, Grain Screened\*Ink Temperature, Press Speed\*Ink Temperature
  - Attributes: Degree (2)
  - Transform: (None)
  - No Intercept

# Case Study: Defect Reduction

A Six Sigma team was addressing the occurrence of a product defect.

Although the occurrence rate was small (4.5%), occurrence costs exceeded \$10,000 per incident.

A large number of processing factors and raw material factors were suspected of causing the defect.

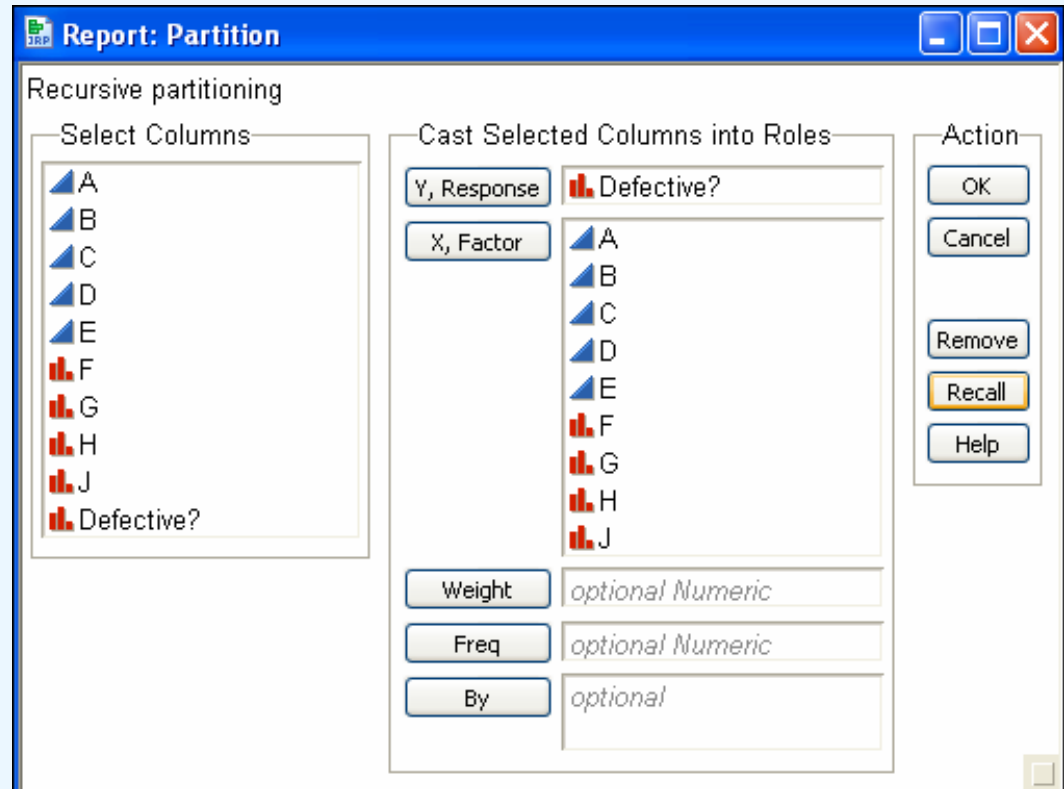
To obtain information on which factors might be associated with the defect, the team used the partition platform to analyze a large observational database, containing process and quality information, for the product of interest.

# Case Study: Defect Reduction

The database contained  
6253 records.

Nine process and raw  
material factors (five  
continuous, four  
categorical), were used  
as inputs to the Partition  
analysis.

(The factors are generically  
named to preserve  
confidentiality.)

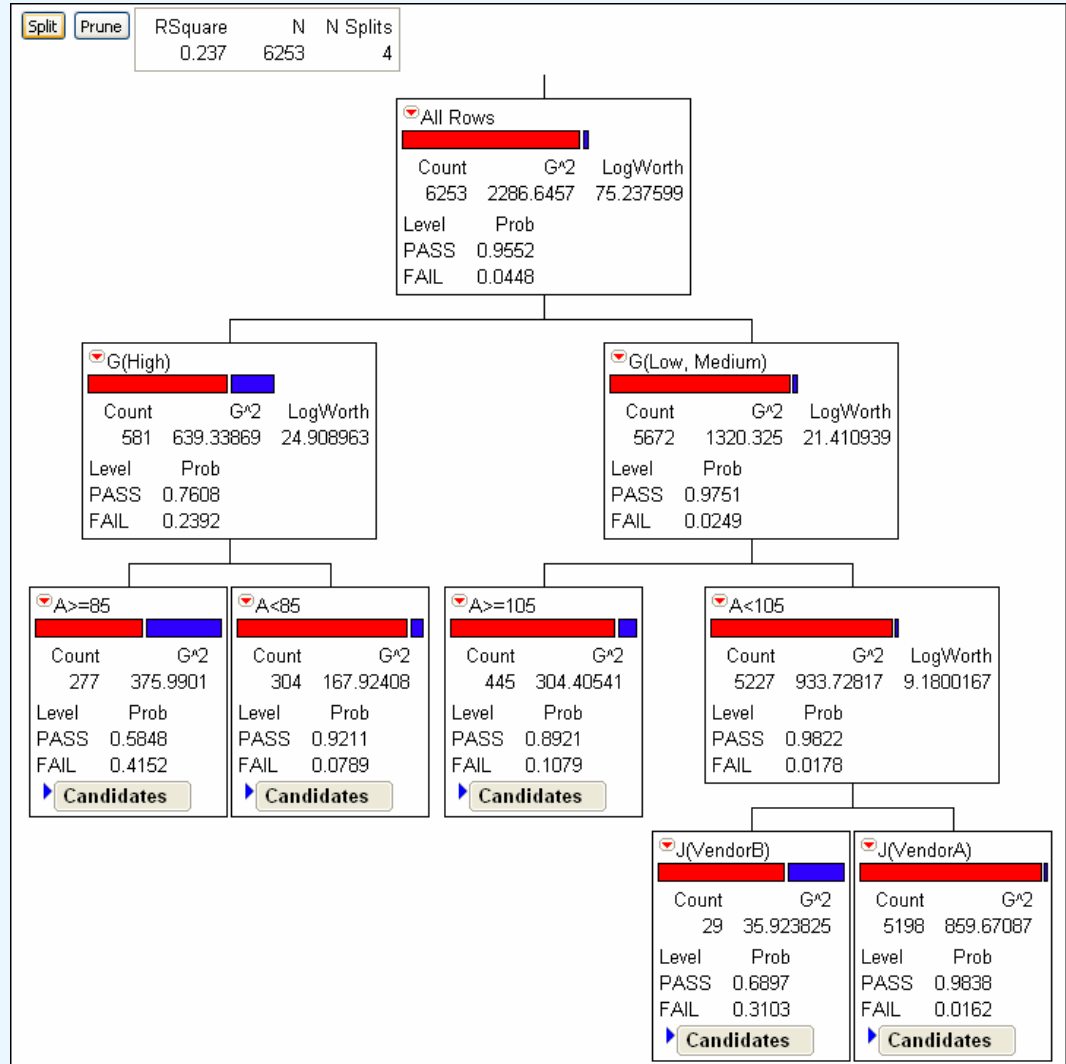


# Case Study: Defect Reduction

The Partition analysis, based on four splits, is shown to the right.

Three factors are involved:

G, A, and J.

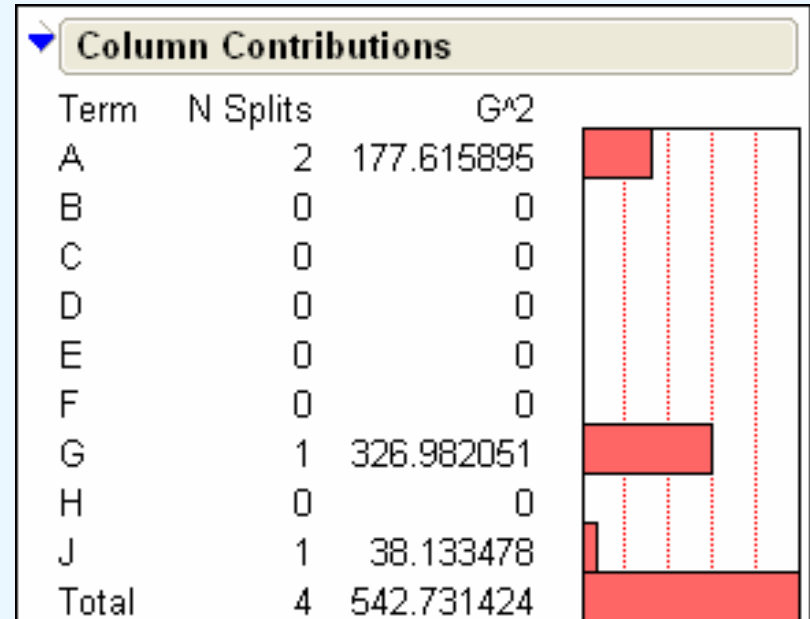


# Case Study: Defect Reduction

The Column Contributions report suggests that these three factors explain a large amount of the “variation” in the response.

Based on this analysis, the team performed a **2<sup>3</sup> factorial experiment** with factors G, A, and J.

The experiment led to **root cause identification and elimination of the defect.**



# Summary

This talk has discussed the use of partition analysis in supporting variable selection for design of experiments.

Although both examples were classification tree models, regression trees can be used in a similar fashion.

We have found this pairing extremely valuable in our Six Sigma project work and training.

We introduce the Partition platform in both our Green Belt and Black Belt training.

We are in the process of integrating JMP 6's enhanced Custom Design platform in our Green Belt and Black Belt training.

# Summary

Partitioning overcomes some of the shortcomings of multiple linear and logistic regression (traditional regression).

Multiple linear regression modeling works well when the predictors and the response are linearly related.

However, relationships are seldom linear.

Traditional regression can be adversely affected by outliers and unruly distributions, both for the predictors and response.

Traditional regression does not deal well with categorical predictors that have many levels (for example, Part Number, Distribution Center, Sales Region).

# Summary

## Partition methods:

- Assist in data exploration;
- Help with variable reduction;
- Inform variable recoding (grouping levels of categorical variables into fewer categories);
- Often allow the building of better models than would traditional regression methods;
- Are intuitive and easily understood by Six Sigma project team members;
- Combined with DOE, can greatly enhance project success.

# Summary

We provide the following guidance to Six Sigma project teams:

A Six Sigma project team should consider using tree-based methods when any of the following hold:

- There is a large observational data set to explore;
- The team's data contains multi-level categorical variables;
- The data is unruly (many outliers, missing data);
- The data may contain complex interactions.