Interactive Data Mining and Design of Experiments: the JMP® Partition and Custom Design Platforms

Marie Gaudard, Ph. D., Philip Ramsey, Ph. D., Mia Stephens, MS

North Haven Group

March 2006

Statistical Discovery.™ From SAS.

Table of Contents

Abstract

This paper describes an application of one of JMP software's data mining techniques, called recursive partitioning, to a manufacturing data set. This data set is available on www.jmp.com, so the reader can use the data set to follow the development in this paper, if so desired.

The paper describes how a fictional Six Sigma project team uses partitioning to narrow down the list of potential experimental factors. The team then constructs an experimental design using the JMP Custom Design platform. The paper also summarizes a real case study that illustrates the synergy between partitioning and design of experiments.

The purpose of the paper is to illustrate the value, for Six Sigma projects, of analyzing historical manufacturing data to inform the choice of factors and levels for statistically designed experiments. The paper is in the form of a tutorial for the relevant JMP analyses.

1. Data Mining

1.1. What Is Data Mining?

The term *data mining* refers to the analysis of large observational data sets with the goal of finding unsuspected relationships. A data set can be "large" either in the sense that it contains a large number of records or that a large number of variables is measured on each record.

Data mining techniques are often applied to data sets that were collected for purposes other than those of the data mining study. The data sets employed are often transaction logs, such as records of all credit card purchases over a month. So the data sets used in data mining often consist of observational and convenience samples rather than random samples. These data sets also tend to be messy; they tend to include outliers, missing values, sparsely populated variables, and unruly data distributions.

In its infancy, data mining was used in customer research to answer simple questions, such as, "Who buys what?" It was also used in market basket analysis to make associations, for example, "If a customer buys Product X, is she likely to buy Product Y?" As data mining has evolved, so have its applications. In the field of biological research, for example, data mining techniques are extremely useful in analyzing microarrays, which result in data sets that have large numbers of variables — sometimes hundreds of thousands.

Today, data mining techniques are widely used in market research, analysis of customer satisfaction surveys, and in many areas where large databases are available. For example, data mining is used in the credit industry to decide which applicants are good credit risks. It is also used in fraud detection, for example, to identify instances of credit card or insurance fraud.

We have found data mining techniques valuable for quality improvement initiatives in Six Sigma programs. 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, either before further data collection or simply as background knowledge.
- Suggest causal relationships and potential solutions.
- Identify anomalies.
- Reduce the number of predictors to be studied.

As such, data mining can be used to support the Define, Measure, Analyze, and Improve phases of the DMAIC cycle. It can also be used to support Design for Six Sigma (DFSS) projects.

1.2. Data Mining Techniques

The types of relationships that one seeks to discover or model in a data mining study can be categorized into two main structures: global models and local patterns.

A global model defines a structure that applies globally to all points in the data set. Typical examples include predictive and classification models. Also of interest is anomaly detection, which consists of detecting deviations from the general, and is useful in fraud studies.

A local pattern is a relationship that applies in a restricted region of the variable values. For example:

- In a marketing study, researchers might learn that 90% of customers who buy a high-end yogurt product also buy high-end ice cream.
- In a study of accounts receivable data, researchers might learn that a certain group of customers do not fit the general pattern in terms of payment and returns.

Data mining is associated with a large collection of modeling techniques. Some classical statistical methods, such as multiple linear regression and logistic regression, are sometimes included in the data mining arsenal. Other data mining methods include neural nets, classification and regression trees, clustering algorithms, and association rules.

Because data sets used in data mining tend to be messy, preprocessing tools that facilitate data exploration, visualization, and validation are useful in the data mining process. Data visualization methods are also critical in validating and understanding models that are derived using data mining techniques.

1.3. Traditional Statistics versus Data Mining

As mentioned earlier, classical modeling methods are considered useful in data mining applications. However, modeling techniques such as classification, regression tree analysis, and neural nets differ from classical techniques in a fundamental way. Classical techniques assume an underlying model. This model is fitted to the data, the model fit is evaluated, and if the fit is considered adequate, hypothesis tests for the effect of predictors are performed in order to identify significant predictors. Overfitting is prevented through the use of statistical tests and diagnostics based on the underlying model assumptions. The quality of model predictions is assessed using prediction intervals.

In contrast, techniques such as classification, regression tree analysis, and neural nets do not assume an underlying model, and so do not accommodate hypothesis testing. Models derived using these techniques are usually validated on independent data. Often, the complete data set is split into a *training*, or *development*, data set and an *evaluation* data set. Models are built using the development data, and they are evaluated on the evaluation data.

In large data sets, which are often characterized by complex observations, it is easy to model noise. Since most data mining analyses are used for predictive purposes, it is important not to model idiosyncrasies of the training data. The practitioner must always be aware of the tension between modeling the underlying structure and modeling the noise (underfitting and overfitting).

1.4. JMP and Data Mining

JMP software provides a comprehensive and interactive environment for exploring and visualizing data, modeling relationships, and designing experiments. JMP is a desktop statistics package that is suited for all users, including every level of Six Sigma practitioner — from Green Belts to Master Black Belts.

JMP provides the user with a number of data mining tools, including:

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

Host JMP analyses are supported by extensive display and visualization tools. Rows in the data table are dynamically linked to graphs. These links make it easy, for example, to locate outliers in the data table, highlight groups of points in graphs that seem anomalous, and color points in graphs according to the levels of a selected nominal variable. Because data mining data sets are often messy and unruly, the display capabilities in JMP support the user in data cleaning and data exploration, and later in the visualization of model results.

The JMP Neural Net platform fits a neural net with one hidden layer to a continuous or nominal response. The JMP Partition platform is a classification and regression tree methodology. Other tree-fitting methodologies, found in high-end and very expensive data mining packages, are CART®, CHAID, and C5.0.

2. Recursive Partitioning

2.1. Partitioning

The JMP Partition platform is a version of classification and regression tree analysis.

Both response and factors (predictors) can be either continuous or categorical. Continuous factors are split into two partitions according to cutting values. Categorical factors (factors that are nominal or ordinal) are split into two groups of levels.

If the response is continuous, the sum of squares due to the differences between means is a measure of the difference in the two groups. Both the variable to be split at a given level and the cutting value for the split are determined by maximizing a quantity, called the *LogWorth*, which is related to the p-value associated with the sum of squares due to the difference in means. In the case of a continuous response, the fitted values are the means within the two groups.

If the response is categorical, the splits are determined by maximizing a LogWorth statistic that is related to the likelihood ratio chi-square statistic, reported in the JMP output as “G^2”. In this case, the fitted values are the estimated proportions, or response rates, within groups.

The JMP Partition platform is extremely useful for both exploring relationships and for modeling. It is very flexible, allowing a user to find not only splits that are optimal in a global sense, but also node-specific splits that satisfy various criteria. The platform provides only a minimal *stopping rule* — that is, a criterion to end splitting. This rule is based on a user-defined minimum node size. The platform does not incorporate any other stopping rules; this is advantageous in that it enhances flexibility.

2.2. The Press Band Data

To illustrate the JMP Partition platform, consider this example from the rotogravure printing business. In the printing process:

1. An engraved copper cylinder is rotated in a bath of ink.
2. Excess ink is removed.
3. Paper is pressed against the inked image.
4. The engraved image is removed from the cylinder once the job is complete.
5. The cylinder is re-used.

A defect called banding — which consists of grooves that appear in the cylinder at some point during the print run — can sometimes occur, ruining the product. When banding is detected, the run is halted, and the cylinder is removed and repaired. This process can take anywhere from 30 minutes to six hours. So, understanding the conditions that lead to banding is critical and could save a printer enormous amounts of money.

We will utilize a set of observational data on banding. This data can be found at **http://ftp.ics.uci.edu/pub/machine-learning-databases/** and is called “cylinder-bands”. The following image captures part of the data. The data set contains 540 records and 39 variables. The target variable is “Band Occurred?” and its values are “BAND” and “NOBAND”.

PressBandingData

PressBandingData
Partition

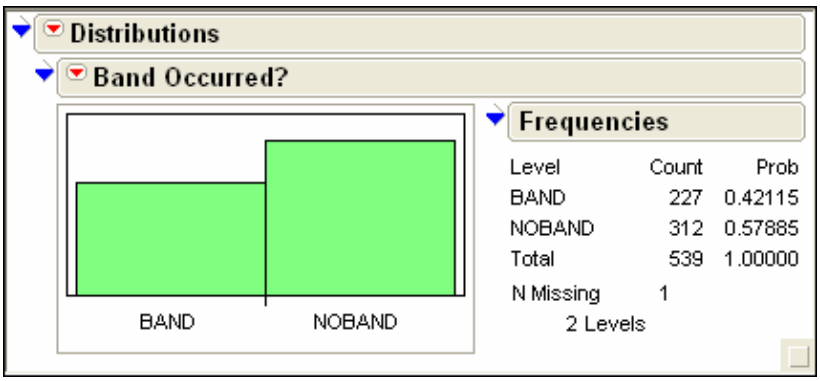
Columns (39/0)
solent type
type on cylinder
press type
press
cylinder size
location
plating tank
proof cut
viscosity
caliper
ink temperature
humidity
roughness
blade pressure
varnish pct
press speed
ink pct
solvent pct
ESA Voltage
wax
hardener
roller durometer
current density
anode space ratio
chrome content
Band Occurred?

Rows
All rows 540
Selected 0
Excluded 0
Hidden 0
Labelled 0

350 Cols	Date	Date M/Y	Job Number	Cylinder No.	Customer	grain screened	proof on ctd ink	paper type
1	03/30/1990	03/1990	23040	X750	GUIDEPOSTS	YES	YES	UNCOATED
2	04/09/1990	04/1990	34683	G467	ECKERD	NO	YES	COATED
3	04/09/1990	04/1990	25416	X203	TVGUIDE	YES	YES	UNCOATED
4	04/14/1990	04/1990	34545	O21	TARGET	NO	YES	COATED
5	04/17/1990	04/1990	36858	T313	EXXON	YES	YES	UNCOATED
6	04/18/1990	04/1990	36053	J68	WARDS	NO	YES	COATED
7	04/18/1990	04/1990	36053	J42	WARDS	NO	YES	COATED
8	04/18/1990	04/1990	36858	F329	EXXON	YES	YES	UNCOATED
9	04/25/1990	04/1990	34664	G496	BURDINES	YES	YES	UNCOATED
10	04/26/1990	04/1990	34545	O6	TARGET	NO	YES	UNCOATED
11	04/26/1990	04/1990	34545	O14	TARGET	NO	YES	UNCOATED
12	05/05/1990	05/1990	47103	T244	MODMAT	YES	YES	UNCOATED
13	05/07/1990	05/1990	47103	M93	MODMAT	NO	YES	COATED
14	05/07/1990	05/1990	47103	M260	MODMAT	YES	YES	UNCOATED
15	05/07/1990	05/1990	47103	T383	MODMAT	NO	YES	COATED
16	05/07/1990	05/1990	47103	T78	MODMAT	YES	YES	UNCOATED
17	05/07/1990	05/1990	47103	M4	MODMAT	YES	YES	UNCOATED
18	05/07/1990	05/1990	36928	M432	HOMESHOPPIN	NO	YES	COATED
19	05/07/1990	05/1990	36928	M257	HOMESHOPPIN	NO	YES	COATED
20	05/09/1990	05/1990	47103	F242	MODMAT	YES	YES	UNCOATED
21	05/10/1990	05/1990	47103	F672	MODMAT	YES	YES	UNCOATED
22	05/11/1990	05/1990	47103	M260	MODMAT	YES	YES	UNCOATED
23	05/14/1990	05/1990	47103	F679	MODMAT	YES	YES	UNCOATED
24	05/17/1990	05/1990	36054	X400	WARDS	NO	YES	COATED
25	05/17/1990	05/1990	34752	X776	TOYSRUS	NO	NO	COATED
26	05/17/1990	05/1990	34752	X713	TOYSRUS	NO	NO	COATED
27	05/18/1990	05/1990	34402	I331	AUSTADS	YES	YES	UNCOATED
28	05/24/1990	05/1990	36648	F227	JAMESWAY	NO	YES	UNCOATED
29	06/02/1990	06/1990	36859	F590	NATLWILDLIFE	YES	YES	UNCOATED
30	06/03/1990	06/1990	36859	F670	NTLWILDLIFE	YES	YES	UNCOATED
31	06/06/1990	06/1990	36859	F331	NATLWILDLIFE	YES	YES	UNCOATED
32	06/06/1990	06/1990	36859	F571	NATLWILDLIFE	YES	YES	UNCOATED

2.3. Formation of the Six Sigma Team

An analysis of the data, using the JMP Distribution platform, indicates that banding occurred in 42% of press runs:



This validates the formation of a Six Sigma team charged with reducing or eliminating banding defects. We will tell the story of this fictional Six Sigma project team.

2.4. The Measure Phase

To fulfill its mission, the team must identify the root causes of banding. To identify root causes of a problem, Six Sigma teams often construct cause-and-effect diagrams and begin collecting data on potential root causes. Then they construct Pareto charts in an effort to find root causes. But Pareto charts overlook complex relationships and interactions among possible explanatory variables.

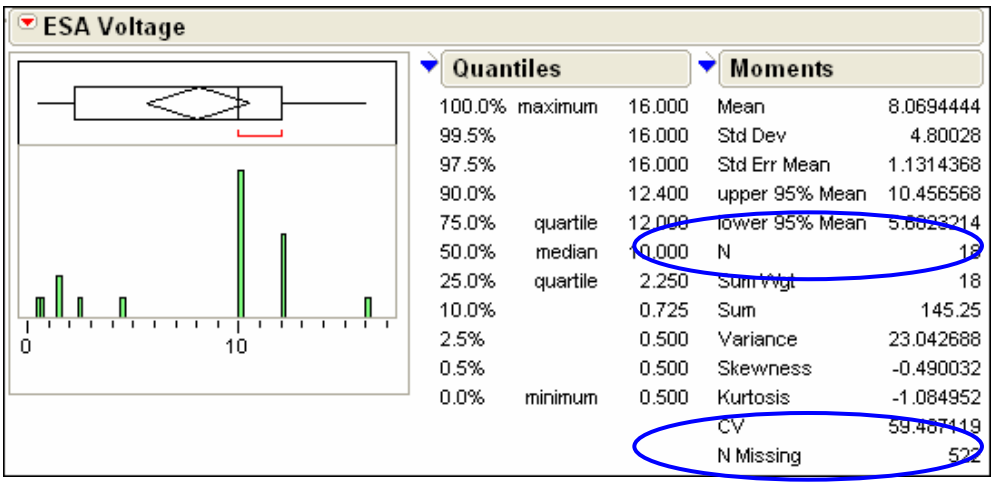
However, our Six Sigma team has a large historical data set available, and it makes sense to see what can be learned from this data before proceeding to further data collection. The available predictors for banding consist of 11 categorical variables and 18 continuous variables. The team could, at this point, explore two-way relationships between the predictors and the nominal response “Band Occurred?”. This exploration would consist of mosaic plots and contingency tables for categorical predictors, and comparison boxplots and ANOVA analyses for continuous predictors.

But such pairwise analyses will, necessarily, ignore complex interactions. The team could employ logistic regression, with “Band Occurred?” as the response and all relevant predictors included (excluding “ESA Voltage”, which is discussed in the following section). But it is not reasonable to fit such a model because of the many categorical predictors, and the fact that only 60 rows contain non-missing entries for all predictors (“ESA Voltage” excluded).

However, the team can easily construct a classification tree with “Band Occurred?” as the response. As we will see, this analysis provides rich information about the conditions that lead to banding.

2.5. Data Validation

Observational data sets must always be examined for data integrity before they are analyzed. For example, the following graphic indicates that the variable “ESA Voltage” is missing for all but 18 records:



The number of missing rows by variable is shown in the following table. This table is easily obtained under Tables/Summary.

	Label	Number Missing
1	N Missing(ESA Voltage)	522
2	N Missing(varnish pct)	281
3	N Missing(hardener)	219
4	N Missing(location)	156
5	N Missing(wax)	146
6	N Missing(roughness)	107
7	N Missing(blade pressure)	64
8	N Missing(proof on ctd ink)	57
9	N Missing(ink pct)	56
10	N Missing(solvent pct)	56
11	N Missing(solvent type)	55
12	N Missing(proof cut)	55
13	N Missing(roller durometer)	55
14	N Missing(grain screened)	49
15	N Missing(caliper)	28
16	N Missing(plating tank)	19
17	N Missing(type on cylinder)	18
18	N Missing(press speed)	12
19	N Missing(current density)	8
20	N Missing(anode space ratio)	8
21	N Missing(viscosity)	6
22	N Missing(cylinder size)	4
23	N Missing(chrome content)	4
24	N Missing(ink temperature)	3
25	N Missing(humidity)	2
26	N Missing(Band Occurred?)	1
27	N Missing(paper type)	0
28	N Missing(ink type)	0
29	N Missing(press type)	0
30	N Missing(press)	0

The team decides not to include “ESA Voltage” in their partition analysis. However, the team uses all other potential predictors, even though some of these are about 50% missing. The partition algorithm imputes — that is, randomly assigns — values for the missing values, and this allows the variables that are poorly populated to be noticed, if they indeed help explain banding.

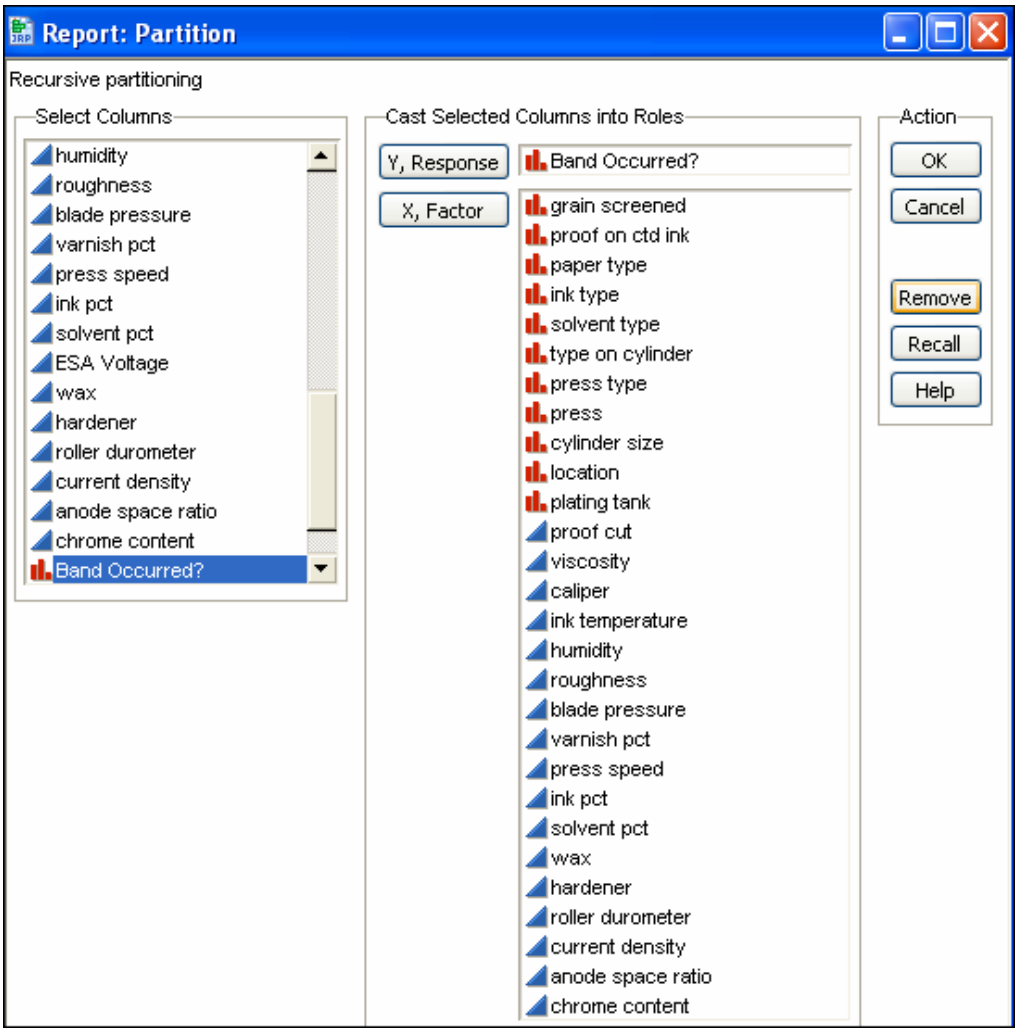
Note that JMP 6 provides a Missing Data Pattern platform that helps identify patterns in missing data. The following table shows an analysis with all predictors (other than “ESA Voltage”) and the response “Band Occurred?” included. A “1” in the “Patterns” column indicates that there are missing values in the variable that was entered in that ordered position. For example, row 14 indicates missing values on ten variables — the variables entered first, second, fifth, tenth, etc. These variables are listed to the right of the “Patterns” column and display a “1” to indicate missing values. A total of nine rows in the original data table have the missing value pattern described in row 14 of the Missing Value Pattern table.

	Count	N Cols Missing	Patterns	grain screene...	proof on ctd ink	paper type	ink type	solvent type	type on cylinder	press type
1	60	0	00000000000000000000000000000	0	0	0	0	0	0	0
2	46	1	00000000000000000010000000000	0	0	0	0	0	0	0
3	30	2	00000000000000000010000100000	0	0	0	0	0	0	0
4	29	1	00000000000000000000000100000	0	0	0	0	0	0	0
5	28	1	00000000000000000000001000000	0	0	0	0	0	0	0
6	22	3	00000000000000000010001100000	0	0	0	0	0	0	0
7	22	1	00000000010000000000000000000	0	0	0	0	0	0	0
8	17	2	00000000000000000000001100000	0	0	0	0	0	0	0
9	15	2	00000000010000000010000000000	0	0	0	0	0	0	0
10	14	2	00000000000000000010001000000	0	0	0	0	0	0	0
11	13	3	00000000000000001010000100000	0	0	0	0	0	0	0
12	9	2	00000000000000001000000100000	0	0	0	0	0	0	0
13	9	2	00000000000000011000000000000	0	0	0	0	0	0	0
14	9	10	11001000010100000010110110000	1	1	0	0	1	0	0
15	7	3	00000000010000000000001100000	0	0	0	0	0	0	0
16	7	3	00000000010000000010000100000	0	0	0	0	0	0	0
17	7	9	11001000010100000010110010000	1	1	0	0	1	0	0
18	6	8	01001000010000000010110110000	0	1	0	0	1	0	0

Note that only 60 rows have non-missing values for all variables entered. Using the designated predictors, a classical procedure such as logistic regression would utilize only 60 rows of the data.

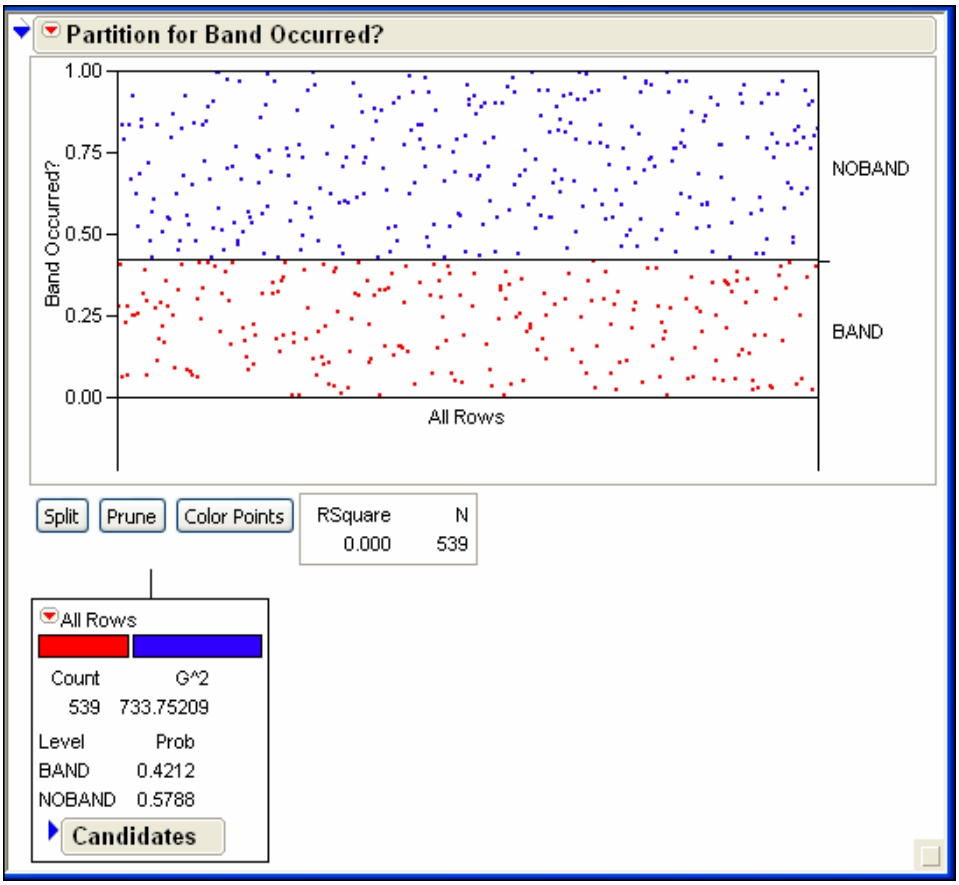
2.6. The Classification Model

Our Six Sigma team proceeds to fit a classification model using the Partition menu in JMP. The response is “Band Occurred?”, and the 28 variables are input as candidate predictors.



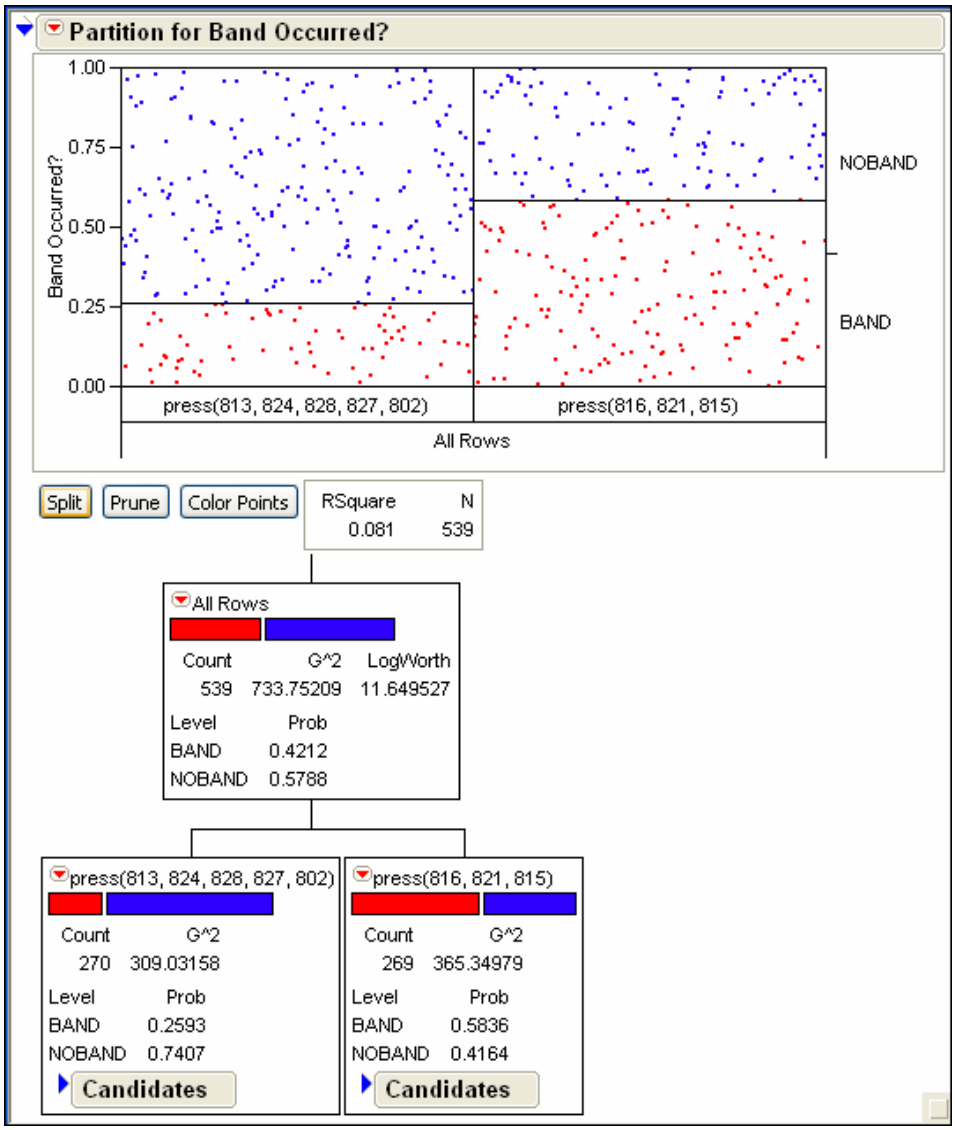
2.7. The Partition Report

The following partition report opens. Points corresponding to the runs are jittered in such a way that runs with banding are shown in red and are plotted in the area of the graph beneath the horizontal divider at 42.12%. Blue points, which represent no banding, are shown above the line.



2.8. Time to Split

Now the team performs the first split. JMP chooses the variable “press” as the splitting variable. The graph updates to the following figure. The split places five presses in a node in which “NOBAND” is more likely, and the three other presses in a node in which “BAND” is more likely.



2.9. The Splitting Criterion

If we had opened the Candidates list in the initial node before splitting, we would have seen Candidate G^2 and LogWorth values.

Take the variable “humidity” as an example. The partition algorithm obtains all possible splits of “humidity”. For each possible split, the likelihood ratio chi-square value for a test of independence of “Band Occurred?” versus the two “humidity” groupings is obtained. The G^2 value for “humidity” that is shown in the Candidates list is the largest possible one, and so corresponds to the likelihood ratio chi-square value for the best split, based on G^2.

All Rows

Count	G^2
539	733.75209

Level	Prob
BAND	0.4212
NOBAND	0.5788

Candidates

Term	Candidate G^2		LogWorth
grain screened	15.26317452		4.02908560
proof on ctd ink	1.84018649		0.75713796
paper type	41.19685301		9.09263749
ink type	23.36747644		5.47317019
solvent type	0.12447377		0.03691410
type on cylinder	6.74183269		2.02605134
press type	19.21404721		4.05141802
press	59.37071878	*	11.64952731
cylinder size	0.65493009		0.17425329
location	7.48196255		1.17734960
plating tank	0.77105555		0.42034133
proof cut	10.64919339		1.90329222
viscosity	11.42286651		2.01782601
caliper	2.43769360		0.24417435
ink temperature	13.62015533		2.47926676
humidity	17.09997024		3.44683674
roughness	9.47150999		1.80321356
blade pressure	7.18955741		1.04488922
varnish pct	5.01276163		0.28628205
press speed	42.28087082		11.11336985
ink pct	17.17727140		3.38586103
solvent pct	10.72213806		1.45229444
wax	6.74466686		1.01416710
hardener	8.87487939		1.50836683
roller durometer	17.99548697		3.99976992
current density	16.98014698		3.89053048
anode space ratio	2.78235275		0.11752418
chrome content	8.38271198		2.19021045

The LogWorth values are the logs of adjusted *p*-values for the chi-square test of independence. These are adjusted to account for the number of ways that splits can occur. For a particular variable, such as “humidity”, the LogWorth value corresponding to the split that gives the largest such value is the one shown in the table.

Note that, in the preceding candidates list, the variable “press” shows the largest values on both criteria. It is possible that the largest G^2 and LogWorth values are obtained for different variables. The default criterion in JMP 6 is to base the split on the LogWorth values. However, the user can change that criterion under the red arrow in the analysis window.

2.10. Lock Columns

The team realizes that the variable “press” shows the largest value for LogWorth, and so, algorithmically, it is the best candidate for the first split. But the team agrees that this variable will not give them information about underlying root causes. All presses must be utilized in production. The question is, “What underlying process behavior is affecting banding, and perhaps also affecting the fact that some presses do better than others?”

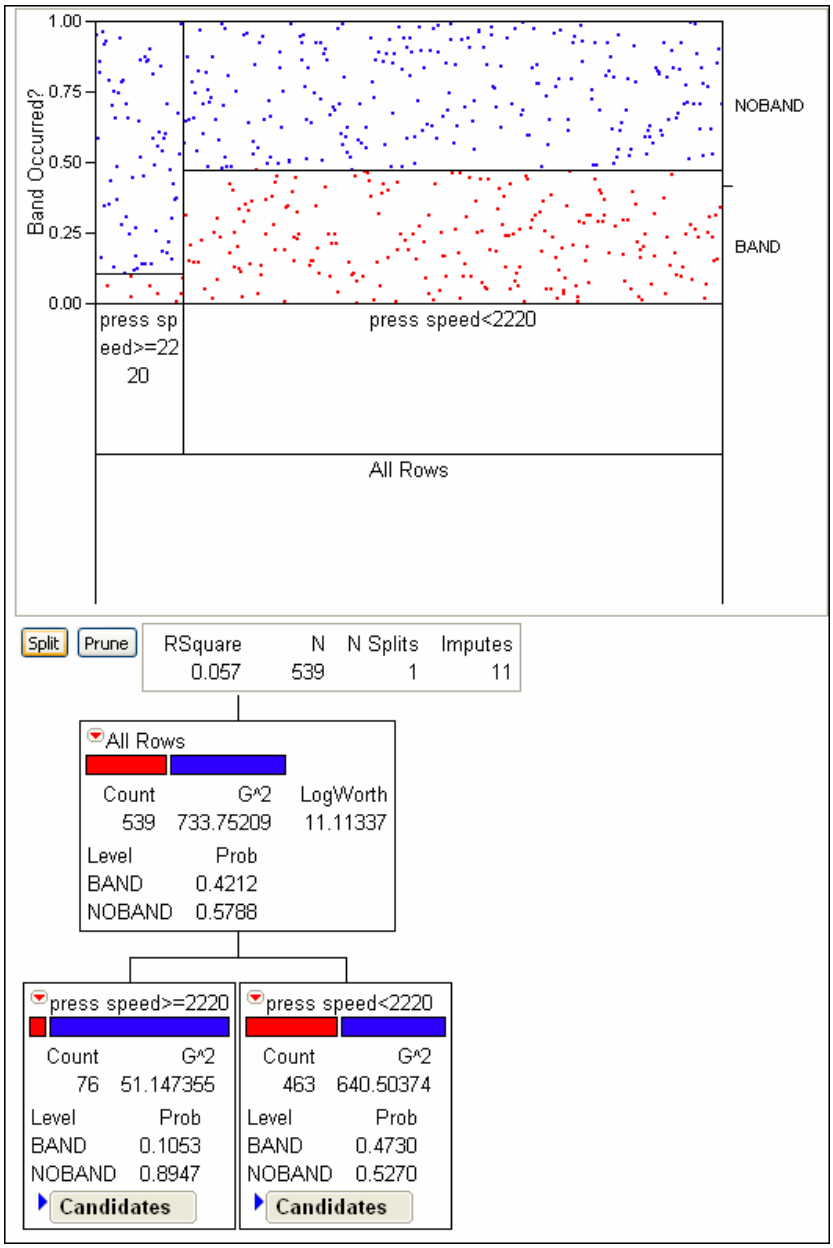
In other words, “press” is not a desirable variable for splitting. The fact that a split would occur on “press” tells the team that certain presses are more affected by banding than others, but that, in itself, does not help the team improve the process.

When splits occur on variables that are only tangentially useful for planning process improvements, one can force attention to more useful predictors by excluding the tangential variables from the partitioning algorithm.

That is done by selecting the Lock Columns option in the Partition platform menu. In our current example, the team first 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 under the red arrow, and then select “press” from the column list.)

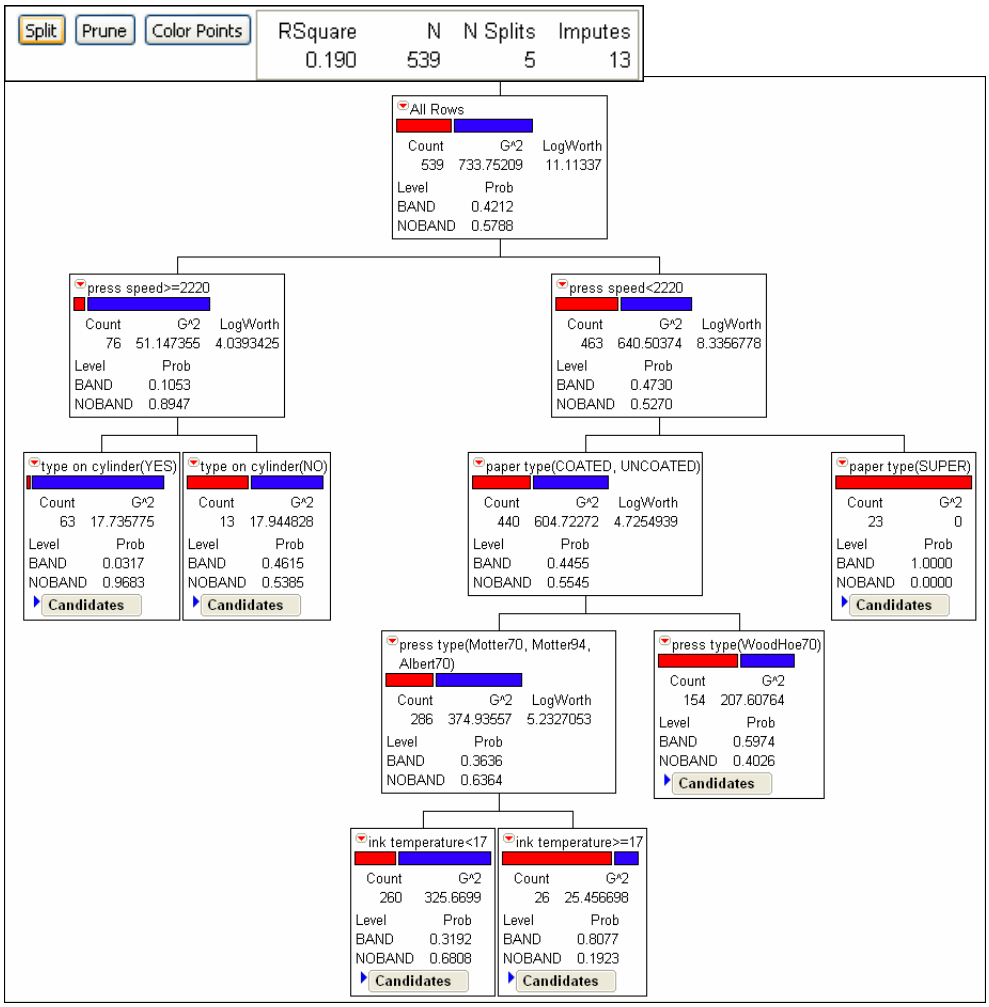
2.11. The Analysis Continues

With the “press” variable locked, the first split occurs on the variable “press speed” (see graph). Knowledge about the effect of “press speed” is useful in terms of process improvement actions. For “press speed >=2220”, the team sees that 10.5% of runs have banding, while for “press speed < 2220”, 47.3% have banding.

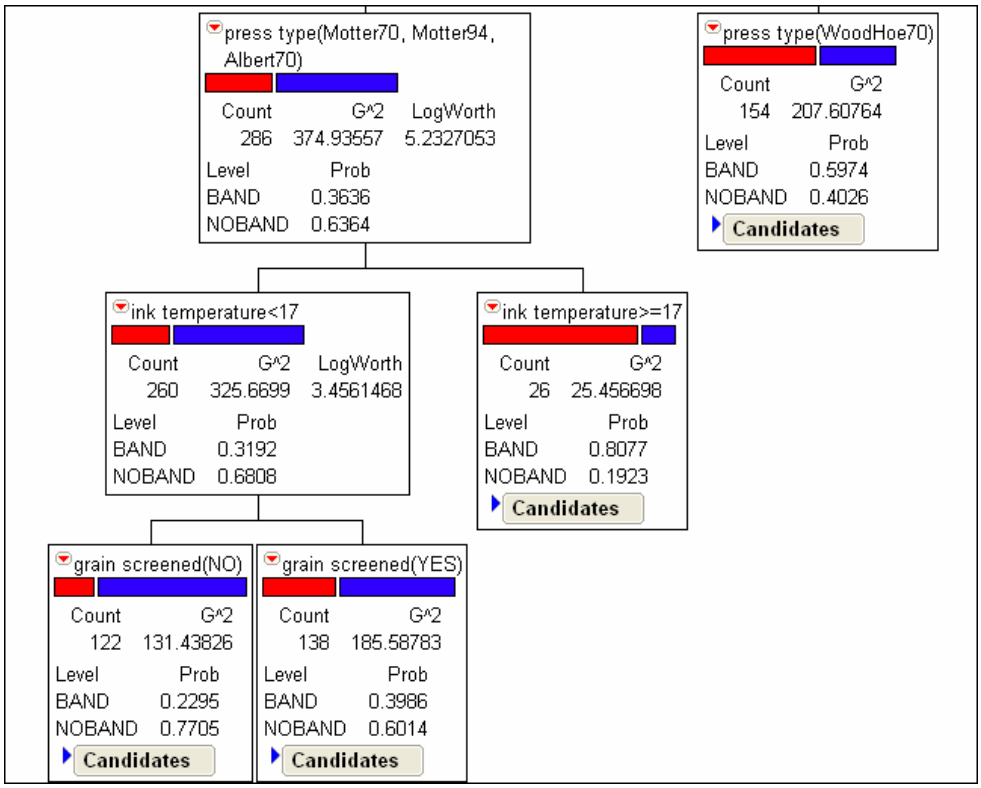


The next graphic shows the tree after five splits. The counts for each of the splits and the observed proportions for each node are displayed. The proportions, “Prob”, are obtained by selecting the appropriate Display Option from the main menu.

Note that a split on “press type” appears at one point. The team might consider locking this variable from the analysis, for reasons similar to those for which “press” was locked out. However, the team suspects that optimal process settings may depend on “press type”. Knowledge of the process should guide decisions of this kind.



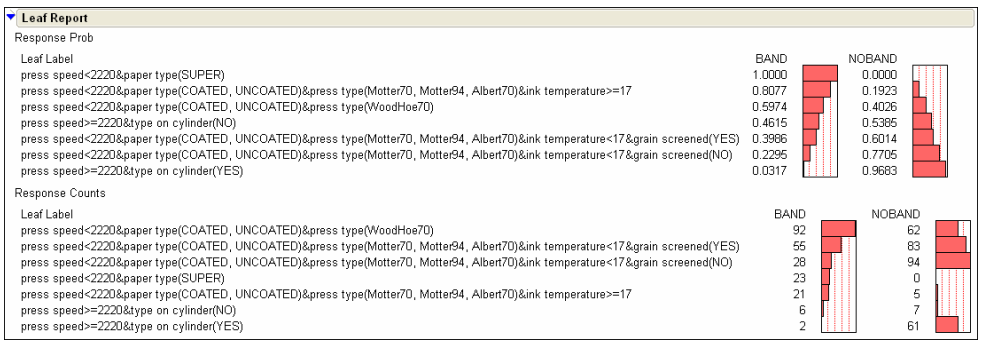
The following graphic shows that a sixth split selects “grain screened” as the partitioning variable in the “ink temperature < 17” branch.



Note that splits have occurred both on nominal and continuous predictors. Note also that, by the sixth split, 26 values have been imputed. This means that 26 of the rows involved in the splits had missing values on some of the split variables. At each split where values are missing, the corresponding rows are randomly assigned to the resulting nodes in a fashion consistent with that variable's population representation.

2.12. The Leaf Report and Predicted Probabilities

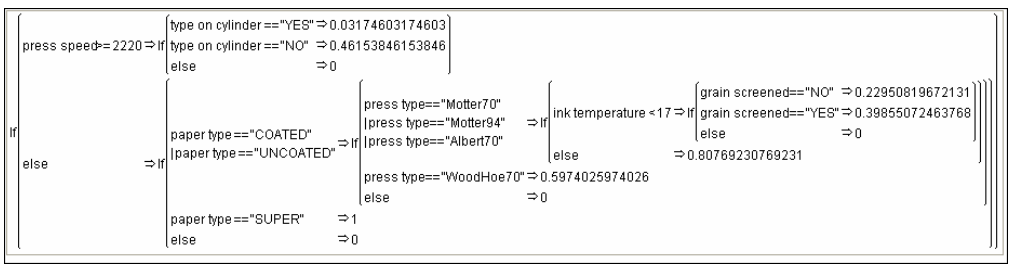
As trees get large, they become visually intractable. JMP provides a leaf report, which gives the rule set and a display of the terminal nodes' discriminatory ability. The leaf report for the team's six split model is shown in the following figure. The leaves have been sorted according to the occurrence of "BAND". This is done by right-clicking in the display in the area of the bar graph and choosing the appropriate options from the menu that appears.



Formulas for the predicted probabilities, leaf numbers, and leaf labels (rule set) can be saved to columns in the JMP data table. The predicted probabilities and leaf labels are shown in the following figure. Note that the leaf labels are long and have been truncated in this figure.

	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 formula for the predicted probabilities, which was saved to the data table, is shown in the following figure. The formula simply follows the splits to the terminal nodes and then assigns the proportion of banding that was observed to a job that falls in that terminal node.



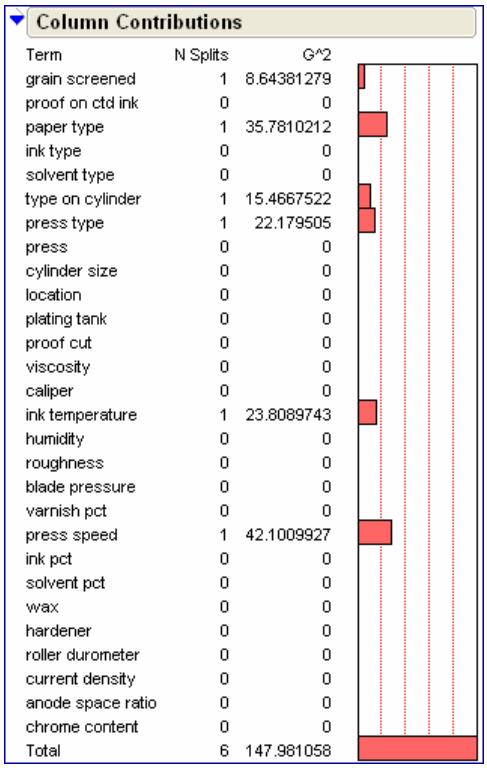
2.13. Model Assessment

In the Partition platform, the user controls splitting. At each split, JMP provides the best splitting variable and grouping of levels of that variable. How does the user evaluate the model defined by a particular selection of terminal nodes?

JMP provides several tools that are of value. These include R^2, a column contributions analysis, and, in the case of a classification (rather than regression) tree, lift curves and receiver operating characteristic (ROC) curves.

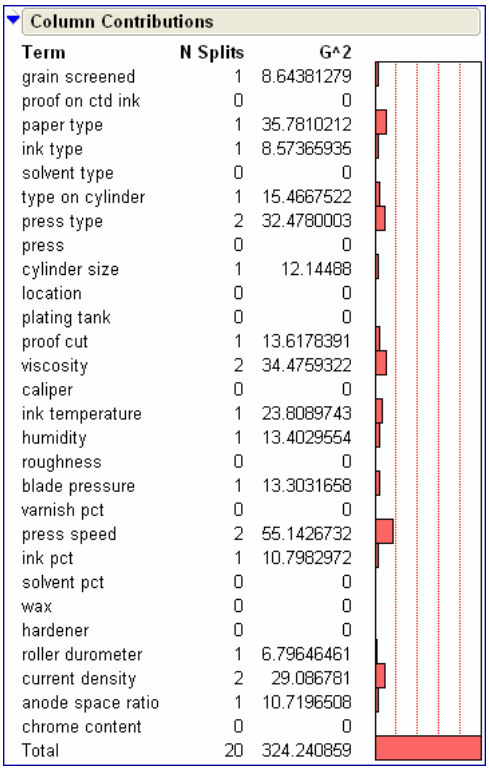
2.14. The Column Contributions Plot

JMP provides a column contributions plot to help determine the influence of the variables on the response. The column contributions plot on the left in the following figure is for our six-split model. The plot on the right is for the model that is obtained after 20 splits.



Column Contributions

Term	N Splits	G^2
grain screened	1	8.64381279
proof on ctd ink	0	0
paper type	1	35.7810212
ink type	0	0
solvent type	0	0
type on cylinder	1	15.4667522
press type	1	22.179505
press	0	0
cylinder size	0	0
location	0	0
plating tank	0	0
proof cut	0	0
viscosity	0	0
caliper	0	0
ink temperature	1	23.8089743
humidity	0	0
roughness	0	0
blade pressure	0	0
varnish pct	0	0
press speed	1	42.1009927
ink pct	0	0
solvent pct	0	0
wax	0	0
hardener	0	0
roller durometer	0	0
current density	0	0
anode space ratio	0	0
chrome content	0	0
Total	6	147.981058



Column Contributions

Term	N Splits	G^2
grain screened	1	8.64381279
proof on ctd ink	0	0
paper type	1	35.7810212
ink type	1	8.57365935
solvent type	0	0
type on cylinder	1	15.4667522
press type	2	32.4780003
press	0	0
cylinder size	1	12.14488
location	0	0
plating tank	0	0
proof cut	1	13.6178391
viscosity	2	34.4759322
caliper	0	0
ink temperature	1	23.8089743
humidity	1	13.4029554
roughness	0	0
blade pressure	1	13.3031658
varnish pct	0	0
press speed	2	55.1426732
ink pct	1	10.7982972
solvent pct	0	0
wax	0	0
hardener	0	0
roller durometer	1	6.79646461
current density	2	29.086781
anode space ratio	1	10.7196508
chrome content	0	0
Total	20	324.240859

Note that, at some point, we begin to split on variables that seem to contribute little in terms of discrimination.

2.15. Lift and ROC Curves

We return to the project team's analysis based on six splits. The model's ability to correctly classify jobs as affected or not affected by banding can be assessed by using a lift curve and/or a receiver operating characteristic (ROC) curve.

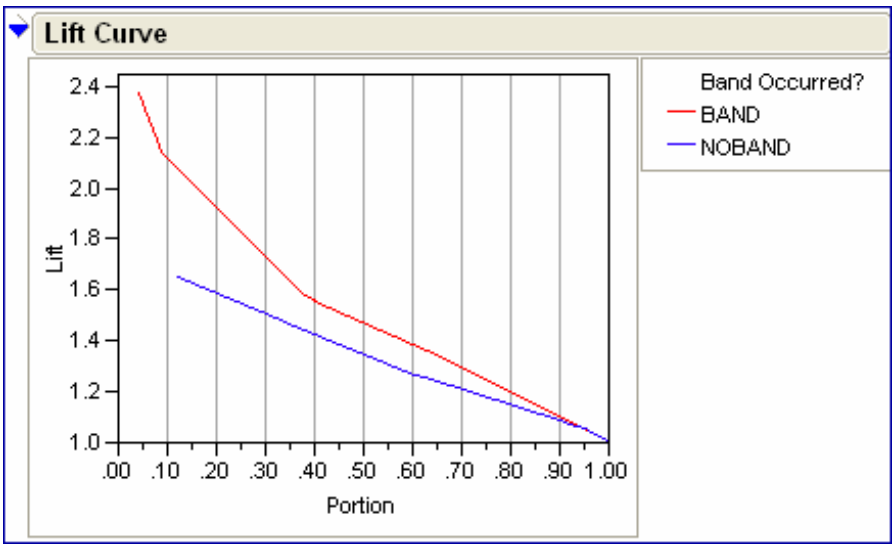
To understand the lift curve, think of the predicted probabilities of "BAND" as being sorted in descending order. Each value of the probability of "BAND" is thought of as a cut point for the decision to classify a record as "BAND". So, each predicted probability defines a percentile, or *portion*, of the data that would be classified as "BAND".

For each predicted probability (or *p*), JMP calculates the portion of the data that has predicted probabilities greater than or equal to *p*. Then, thinking of *p* as a cut point for the decision to classify runs as "BAND", JMP calculates the correct classification rate for those runs with predicted probability greater than or equal to *p*. This rate is divided by the proportion of "BAND" in the full population.

For example, the highest predicted probability in our example is 1.00. The number of jobs that fall in this node is 23 (see the preceding leaf report). This cut point defines the top 23/539 = .043 portion (or 4.3 percentile) of the population.

Because all 23 of these jobs have banding, and so are correctly classified, the correct classification rate is 1.00. In the population, there are 227 jobs that have “BAND” and 539 non-missing records for “BAND”, so the proportion of “BAND” in the population is 227/539 = .421.

So, the lift obtained at the .043 population portion is 1.00/.421 = 2.37. This is plotted on the lift curve as the value at portion .043 (see the top curve in the following graph). Note that, at a portion of .10, the lift value is about 2.10. This means that, if we use the six-split model and classify the runs with predicted probabilities in the top 10% as “BAND”, then we are correctly identifying 2.1 times more jobs than would be identified by chance alone.



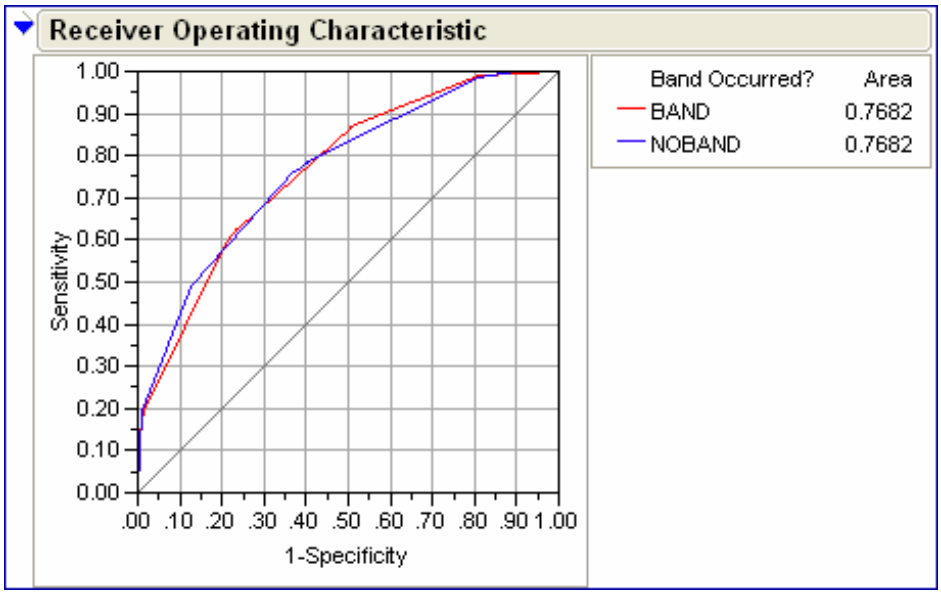
Intuitively, the lift curve measures the superiority of our model to random classification.

(In the construction of the lift curve, we note that a lift value is calculated for each of the predicted probabilities (see the leaf report). The lift values between the predicted probabilities are determined by linear interpolation.)

The receiver operating characteristic (ROC) curve is also based on the idea of treating the probabilities as cut points for a classification scheme.

For a given cut point, the ROC curve plots the proportion of correct classifications (hit rate or true positive ratio) on the Y axis and the proportion of incorrect classifications (false alarm rate or false positive ratio) on the X axis.

An ideal model has a hit rate of 1 and a false alarm rate of 0. The closer the curve is to the left and upper boundaries of the graph, the better the model. The area under the ROC curve measures the sorting efficiency of the model. A value of .5 indicates that the model is equivalent to chance classification, while a value of 1.0 indicates that the model is classifying perfectly.



2.16. When Do You Stop Splitting?

JMP enables users to specify a minimum node size. Splitting on a node ends when that size is reached. However, splitting until nodes can no longer be split because of the minimum size requirement is not wise, as this may result in modeling of noise, rather than structure.

Recall that partition analysis can be used for data exploration as well as for model building.

If the goal is data exploration, splitting can continue until little additional predictive ability is gained by further splitting. This can be assessed by comparing R^2 values, column contributions, lift curves, or ROC curves.

If the goal of the analysis is predictive modeling, it is strongly recommended that the data be separated into a training set and an evaluation set. Model development should take place on the training set. Here, the user can select a number of candidate models based on criteria such as a minimum change in R^2, column contributions, or lift curves.

Then, these models can be evaluated on the evaluation set and a best model chosen. The evaluation set helps guard against both underfitting and overfitting.

2.17. More Features of the Partition Platform

JMP has many features that facilitate use of the Partition platform. Suppose that you have split six times and have produced a lift curve and a leaf report. When you split once more, the lift curve and leaf report update automatically; there is no need to regenerate them.

As we have seen, formulas can be saved to columns. In this form, they can be applied to new records or copied and pasted into a new data table that contains new records. Also, the row state data type in JMP allows you to easily track development and evaluation samples.

3. Custom Design

3.1. The Improve Phase

Our Six Sigma team is content with its six-split model. The team is ready to address root causes. The following variables were identified by the partition analysis:

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

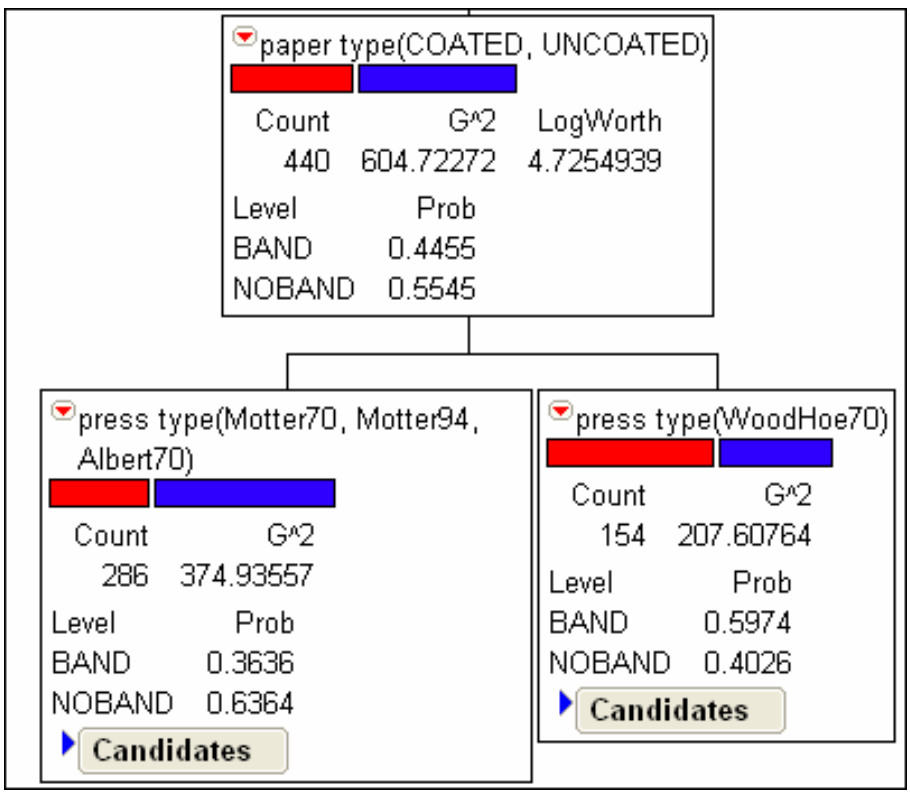
Although the partition analysis suggests an association of these predictors with banding, the team realizes that association is not causality. The team decides to run a designed experiment to determine if these factors and their interactions have a causal effect on banding.

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, and often prohibitive, number of runs to detect factor-level differences.

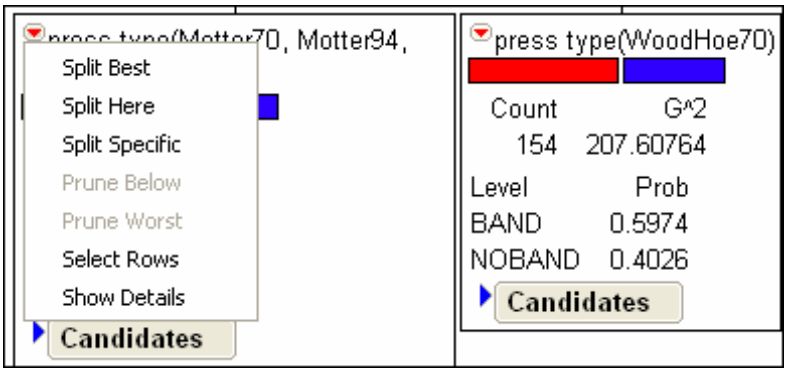
3.2. Partitioning Helps Determine Factor-Level Settings

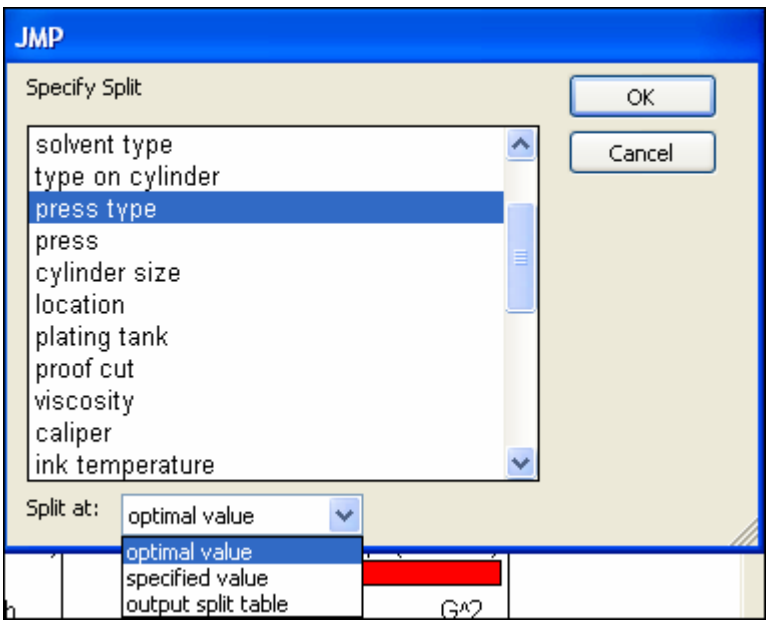
A second challenge to the team is to determine factor-level settings. Here, the Partition platform continues to be of value. For continuous factors, the partition analysis provides a guide to low and high levels based on the cut point that defines the split. For example, the first split is on “press speed” and this is based on the cut point where speed is 2220. There are no further splits on “press speed”. So, it makes sense to choose factor level settings for “press speed” that are aggressive, with a high level above 2220 and a low level below 2220.

The Partition platform also helps in more complex situations. We illustrate with the factor “press type”. Note that there are four press types, and that in our six-split partition model, these are split into two nodes as shown.

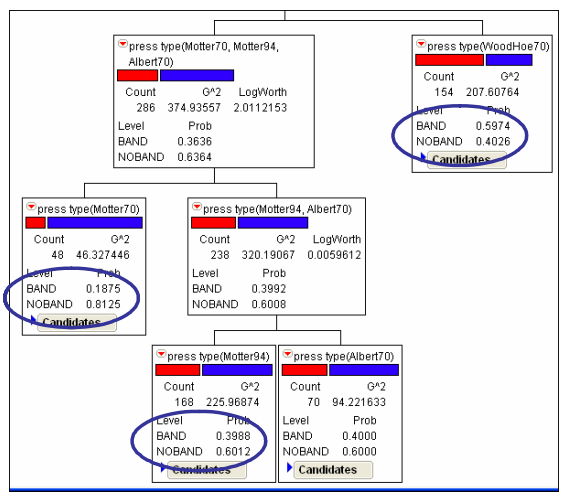


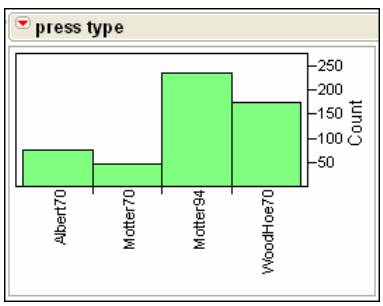
Which press types should be included in the experiment? At each node, the red arrow contains options relating to further splits at that node. At the “press type (Motter70, Motter94, Albert70)” node, the team chooses Prune Below to undo splitting beyond this node. Now the team chooses Split Specific to choose a further split on “press type” at the optimal split value.





The results of this split show that Motter 70 presses and Motter 94 and Albert 70 presses are different in their effects on “BAND” at this point in the tree (see the following partition graph). 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”.

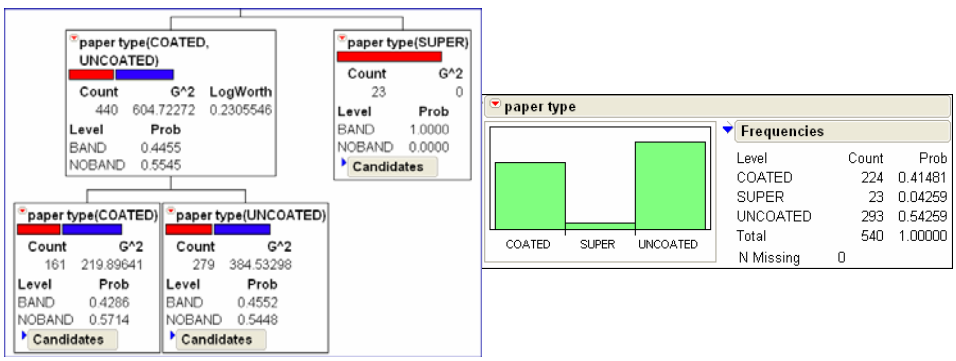




Given this information and the bar chart above, the team decides on three levels for “press type”:

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

The team now turns its attention to “paper type”. “SUPER” paper is rarely used (see the following figure), but it is always affected by banding. However, the team learns that its use is being phased out. A split of the “paper type (UNCOATED, COATED)” node indicates that both paper types seem to be affected at about the same rates (see the following figure). Based on all of this information, the team decides to hold “paper type” constant at “UNCOATED” during the experiment.



3.3. The Randomization Scheme

Once factor levels have been determined for all factors, the team must determine the randomization scheme for the experiment. Since “paper type” has been eliminated as an experimental factor, the experimental factors consist of:

- 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. 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 following factors are *difficult to change*:

- Press type.
- Type on cylinder.
- Grain screened.

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

- Press speed.
- Ink temperature.

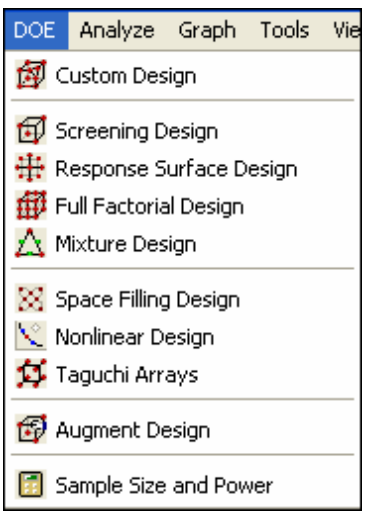
Note that the team faces a fairly complex design problem:

- There is a combination of continuous and categorical factors.
- There is one multiple-level categorical factor (“press type”).
- There are both hard- and easy-to-change factors.
- Two-way interactions among the factors must be estimated.

This last requirement follows from the partition analysis, which suggests the existence of at least two-way interactions among the factors.

3.4. The Custom Design Platform

JMP 6 provides the design of experiments (DOE) options shown below. 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 categorical or continuous factors. JMP also provides Response Surface, Mixture, and other design platforms.



A new feature in JMP 6 is a greatly enhanced Custom Design platform. This is a highly flexible structure for designing both simple and complex experiments. It accommodates:

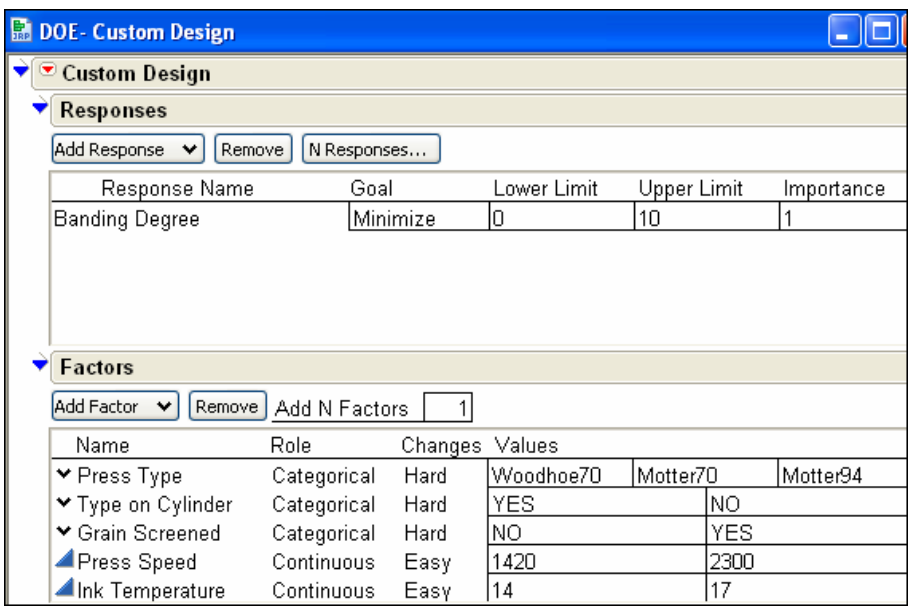
- Continuous and categorical factors with arbitrary numbers of levels.
- Hard- and easy-to-change factors.
- Mixture factors.
- Inequality constraints on factors.
- Covariates and uncontrollable variables.
- User-specified lists of interactions and polynomial terms to be estimated.

By default, Custom Design generates D-optimal designs unless a response surface design is requested; for response surface designs the I-optimality criterion is used. This is an option that can be set by the user, though.

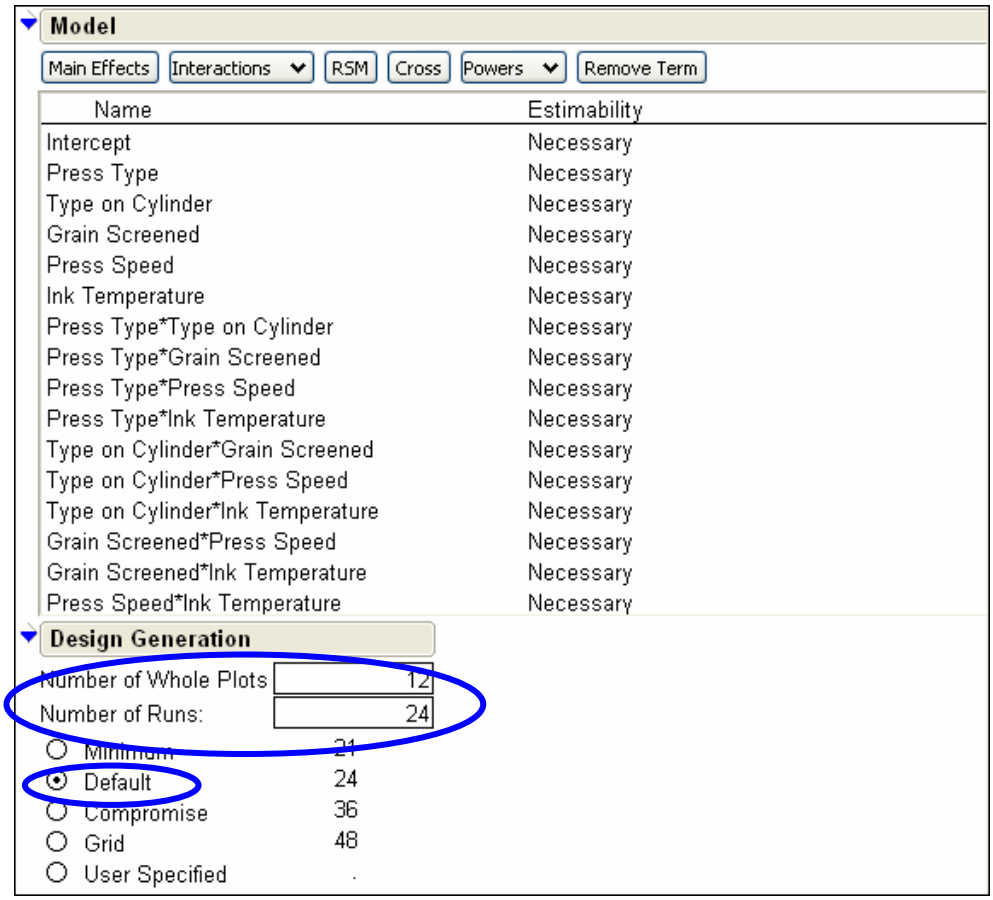
3.5. The Press Banding Team and Custom Design

Given the challenging nature of the design that the Six Sigma team must construct, the team uses the JMP Custom Design platform to facilitate the design process.

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



As shown in the following figure, the team decides on a design that estimates all two-way interactions, as indicated by the partition analysis. Note that the default design will require 24 runs, of which 12 will require changes to the press setup.

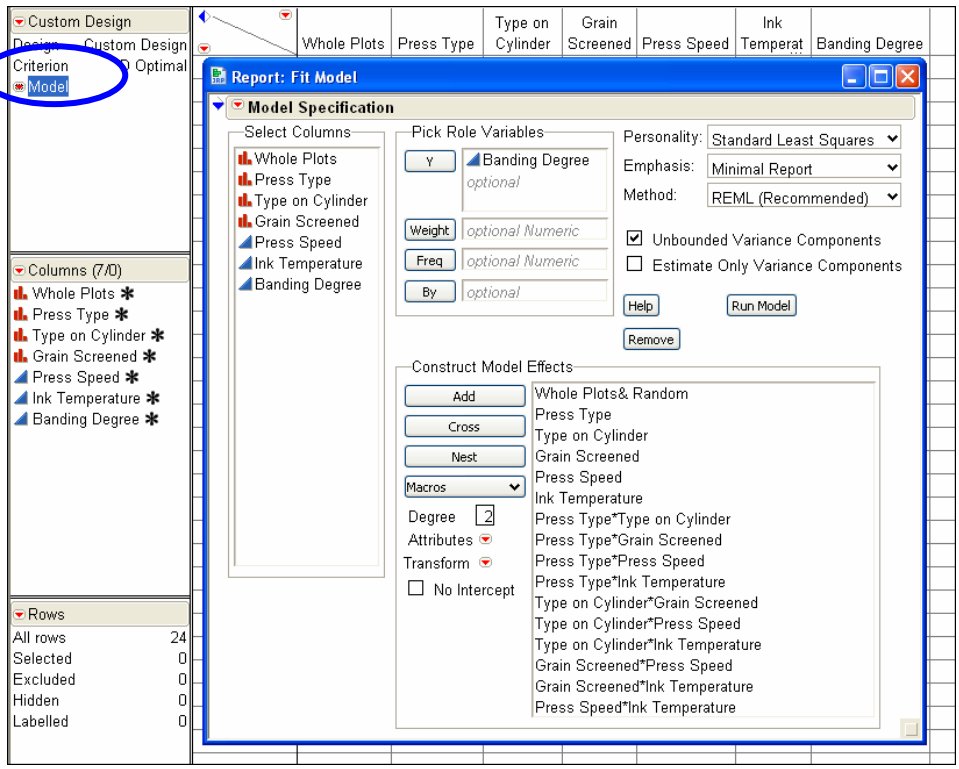


The team agrees that the default design is appropriate. This properly randomized design, generated by JMP, is shown in the following figure.

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	.

JMP conveniently saves the model that will be used to analyze the experiment to the data table. When the team has entered responses from the experiment, the team will simply run this model.



3.6. A Real Application Leads to Success

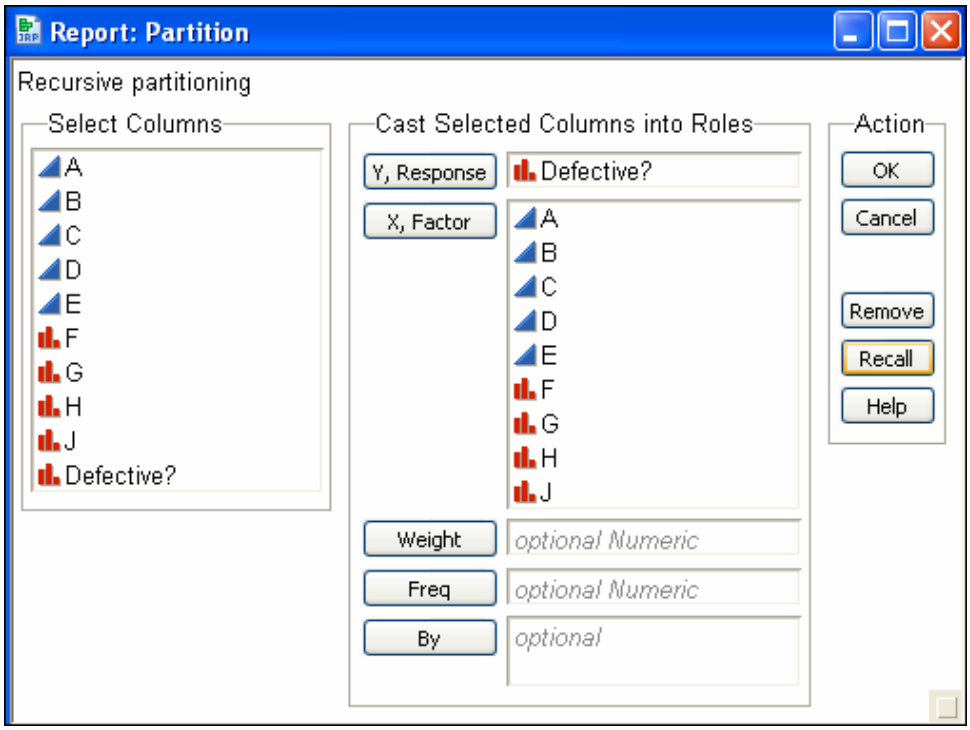
Our fictional Six Sigma team will run its experiment and analyze the results. This will undoubtedly provide valuable knowledge about the root causes of banding.

At this point, we will describe a real application where partitioning, followed by a designed experiment that was informed by the partition analysis, led to a large success. Because the process and data are highly proprietary, we will be able to describe this example only at a summary level.

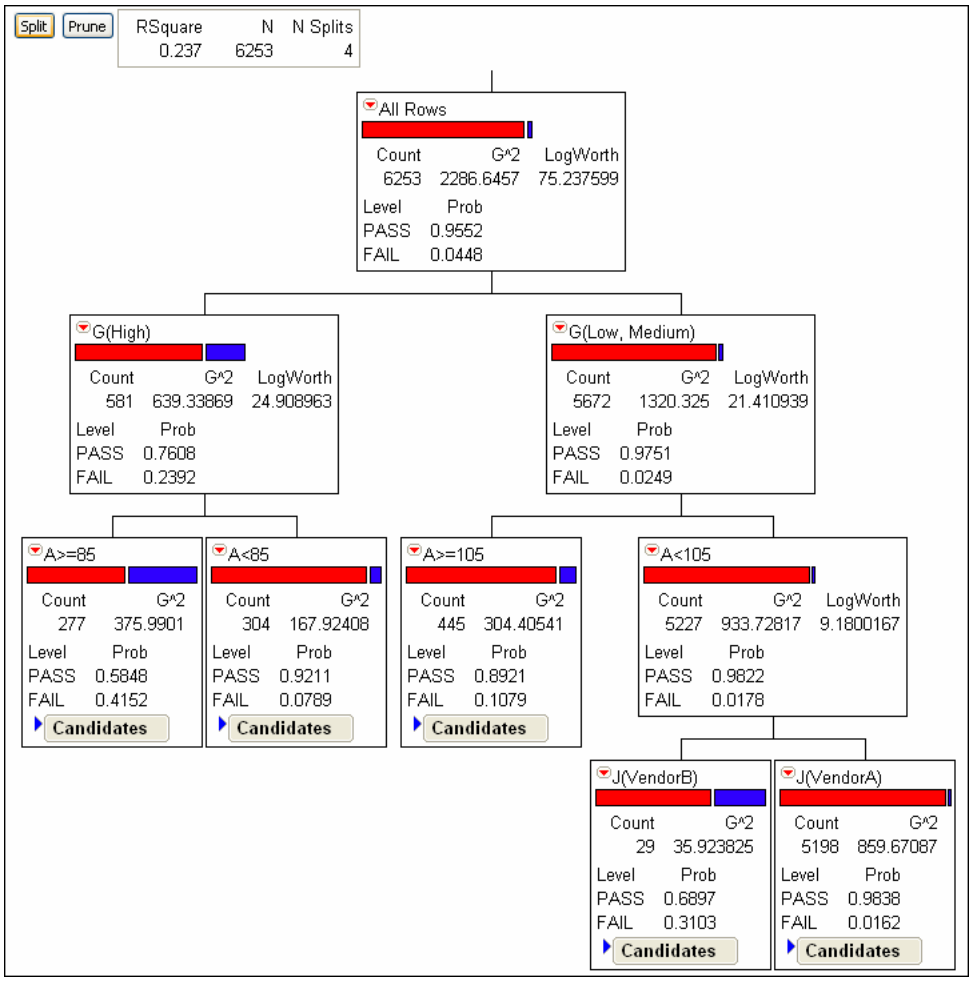
A Six Sigma team was addressing the occurrence of a product defect. Although the occurrence rate was small (4.5%), occurrence costs exceeded $10,000 per incident.

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.

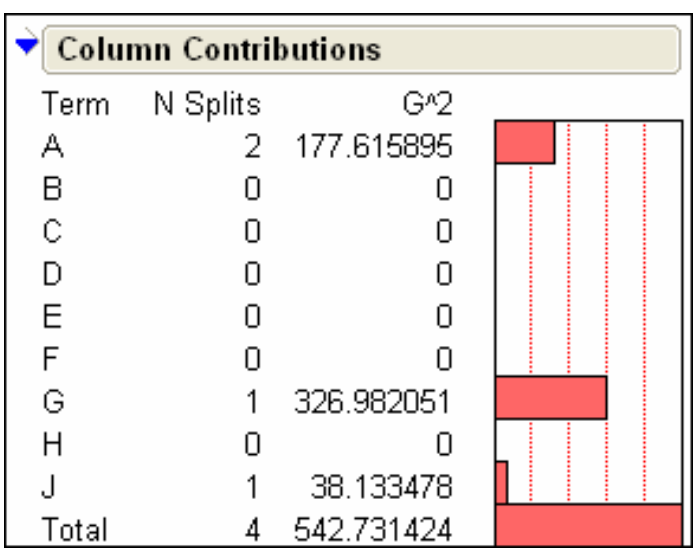
The database contained 6253 records. Nine process and raw material factors (five continuous, four categorical), were used as inputs to the partition analysis. In the following figure, which shows the partition model menu, the factors are generically named to preserve confidentiality.



The partition analysis, based on four splits, is shown below. Three factors are involved: "G", "A", and "J".



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 2^3 factorial experiment with factors “G”, “A”, and “J”. The experiment led to root cause identification and elimination of the defect.

4. Summary

This paper has discussed the use of partition analysis in supporting variable selection for design of experiments, and has illustrated the use of the Custom Design platform in JMP to design a complex experiment. Although both partition examples were classification tree models, regression trees can be used in a similar fashion. We have found the partition/design of experiments pairing extremely valuable in our Six Sigma project work and training.

As we have seen, partitioning overcomes some of the shortcomings of multiple linear regression and logistic regression (traditional regression). Multiple linear regression modeling works well when the predictors and the response are linearly related; however, relationships are not always linear. Traditional regression can be adversely affected by outliers and unruly distributions, both for the predictors and response. And traditional regression does not deal well with categorical predictors that have many levels (for example, Part Number, Distribution Center, Sales Region).

Partition methods assist in data exploration, help with variable reduction, and inform variable recoding (grouping levels of categorical variables into fewer categories). They also often allow the building of better models than would be possible with traditional regression methods. We find that partitioning is intuitive and easily understood by Six Sigma project team members. In manufacturing situations where historical data is available, we have found that partitioning helps teams understand relationships and suggests experimental factors. Combined with design of experiments, it can greatly enhance project success.

We introduce the Partition platform in our Green Belt and Black Belt training. It is a valuable exploratory tool for both transactional and manufacturing projects. Although we initially introduced partitioning in our transactional training, it soon became clear that manufacturing Green and Black Belts would benefit as well. It was then that we first appreciated the value of the Partition platform as a tool for variable reduction prior to designing an experiment.

We are in the process of integrating the enhanced JMP 6 Custom Design platform in our Green Belt and Black Belt training. The Custom Design platform in JMP 6 is a great improvement over the platform in previous versions. With so many industrial experiments dealing with hard- and easy-to-change variables, as well as constraints on the experimental region, we view the new Custom Design platform as an extremely useful and convenient tool for our trainees.

JMP Headquarters
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
USA
Tel: +1 919.677.8000
Fax: +1 919.677.4444
jmpsales@jmp.com
www.jmp.com

JMP Europe
SAS Institute
Henley Road
Medmenham
Marlow
SL7 2EB
United Kingdom
Tel: +44 (0)1628 486 933
Fax: +44 (0)1628 483 203
jmpsaleseur@jmp.com
www.jmp.com

JMP Japan
SAS Japan Head Office
Inui Bldg. Kachidoki
1-13-1 Kachidoki
Chuo-ku Tokyo 104-0054
Japan
Tel: +81 3 3533 3887
Fax: +81 3 3533 1600
jmpjapan@jmp.com
www.jmp.com/japan

JMP China
SAS China
25/F POS Plaza
1600 Century Avenue
Pudong New District
Shanghai 200122
PRC
Tel: + 86 21 6876 5353
Fax: + 86 21 6876 9010
jmpsalesprc@jmp.com
www.jmp.com