Improving Processes With Statistical Models

By Malcolm Moore, Robert Anderson, and Phil Kay, SAS

TECHNICAL PRIMER

Table of Contents

Introduction

Do you face scientific, engineering or business challenges that can't be solved by an appeal to expert judgment, simple spreadsheet analysis, or by varying only one thing at a time?

Do you seem to have more problems than time available to solve them?

Are you forced to regularly make decisions from incomplete data or information, limiting your understanding of the real drivers of robust, effective and efficient processes?

Does the need to fix problems with existing processes limit the time you can spend on innovating and developing new processes and products?

JMP® statistical discovery software from SAS helps you gain insights to fix such issues quickly and permanently, giving your organization a competitive edge now and freeing up resources to innovate for future growth.

Finding workable solutions – whether in research, development, manufacturing, sales or marketing – from prior data is sometimes difficult due to the complex, noisy relationships between many variables that are difficult or impossible to easily spot. Learn how JMP users are solving more scientific, engineering and business problems correctly the first time – and in less time – by extracting powerful insights from existing data using proven, simple statistical modeling methods.

The real-world case studies that follow will help you discover best practices to interactively explore the patterns in your data, build useful statistical models of the important patterns of variation, and visually interact with these models to communicate and drive key improvement opportunities.

How can statistical modeling help my business?

Let's see how one manufacturer is using JMP to add value. The organization was unable to consistently meet demand for one of its products. Often it was able to manufacture without issues, yet at other times the organization had significant yield and quality issues. The primary incoming material was a natural product with considerable variation in key properties, so the manufacturer had to make routine adjustments in the process. Operators made these adjustments based on experience, yet overall the company was unable to get the process running at more than 75 percent of its theoretical capacity.

Significant volumes of process data existed, but engineers were unable to make sense of it with existing software. This meant the critical variables to adjust for incoming material variation were not well defined. Demand for the product was increasing, and to facilitate this growth, the company was faced with building a new production facility at significant capital cost to meet market needs.

Using the integrated statistical modeling, profiling and simulation capabilities of JMP, the engineers created a powerful new understanding of the process. They created statistical models involving all the measured inputs, from which they identified the critical ones for consistent operations. This allowed them to define operating envelopes (control tolerances) for the critical inputs to ensure consistent product quality and throughput. With a robust operating envelope, they could reduce the number of process adjustments for incoming material variability. And when an adjustment was necessary to correct for a major change in the input material, they used the statistical model to define which other inputs to adjust and by how much. This yielded a process with predictable product quality and increased throughput.

The resulting increase in capacity saved hundreds of millions of dollars by avoiding the need to build a new production line, and enabled millions in additional revenue via the resulting predictable increase in capacity. Improving the current process enabled enhanced profitability, increased market share and faster business growth, freeing up capital for other projects. The organization is now adopting statistical modeling upstream to reduce variation in input material and further improve production predictability, capacity and quality.

When should you consider statistical modeling?

Many companies are applying statistical modeling routinely to help predictably, efficiently and effectively drive technical and business decisions. As the above example illustrates, the value to your business and personal satisfaction of solving what appear to be insurmountable problems is huge.

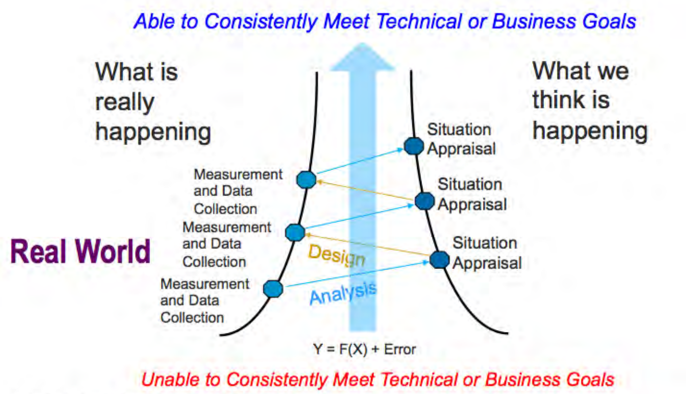
When do you know if statistical modeling might help you? Here are some telltale signs:

- You spend a lot of time dealing with unanticipated problems.
- Your problem-solving efforts yield unpredictable returns – and the same problem or a related one reoccurs.
- You feel you don’t have the complete picture and have to make decisions based on incomplete information.
- You transfer products and processes with limited understanding of how they work.
- You feel stuck in a vicious circle where managing existing problems or processes limits your time for optimization, innovation and new development.

Learning is incremental

Learning is often incremental. Figure 1 illustrates how we typically start with data or a theory, which is analyzed to help assess our situation or theory, which typically leads to more questions that require collection of new data to provide answers. We may iterate many times through this cycle of learning before building a good enough picture of our situation to make a reliable decision. When under time pressure we may omit one or more cycles of learning, with the impact that we may not have closed the gap between the real world and what we think is happening based on our learning to date.

JMP statistical discovery software from SAS helps you extract more knowledge from each cycle of learning, determine the set of inputs, and understand how they work or interact together to create your problem. By extracting more information from each cycle of learning, JMP reduces the number of cycles of learning required to deliver the information you need to make the correct decision, and increases the predictability of getting to those decisions with a limited budget and time.


Able to Consistently Meet Technical or Business Goals
What is really happening
What we think is happening
Situation Appraisal
Measurement and Data Collection
Situation Appraisal
Measurement and Data Collection
Real World
Design
Situation Appraisal
Measurement and Data Collection
Analysis
Y = F(X) + Error
Unable to Consistently Meet Technical or Business Goals

Adapted from Box, Hunter and Hunter

Figure 1: Learning is incremental

Why apply statistical modeling?

Statistical modeling results in many personal and business gains. You can:

- Reduce number of learning cycles. Statistical modeling helps extract more information from complex problems where many factors together are responsible for your opportunity, problem or issue.
- Reduce total problem-solving time. By increasing the knowledge gained from a cycle of learning, statistical modeling helps reduce the total number of learning cycles needed to create the knowledge required to make reliable decisions.

- Increase predictability of improvement and problem-solving activities by extracting more information from your data and reducing the number of learning cycles needed for key decisions. Statistical modeling makes answering your research, development, production, marketing and sales questions more predictable and effective.
- Increase useful understanding. Rather than creating incomplete knowledge from your data – which may result in false solutions – statistical modeling helps extract key relationships from many input variables that may be related to or even interacting with each other, increasing clarity and diminishing confusion.
- Make sustainable decisions within the required time frame, instead of using nonstatistical approaches resulting in poor decisions based on incomplete understanding, which cause problems that must be solved again and again.
- Transfer more complete process knowledge to upstream colleagues.
- Increase time available for new projects to grow future opportunities because it is less likely the same problem comes back to be solved again (and again).

Where to apply statistical modeling

You can apply statistical modeling anytime you have data that can be organized into rows and columns, and tabular or graphical summaries alone are unable to provide the understanding needed for a reliable decision. This is likely to be when:

- No single input is exclusively responsible for output variation, the problem, or the improvement opportunity.
- You need to understand relationships between two or more inputs and one or more outputs.
- Inputs usually vary together, so are correlated and it is difficult to determine which inputs may be responsible for output variation.
- There are holes or gaps in your data table that reduce the amount of data that can be used with nonstatistical approaches.

In product design situations, statistical modeling can improve the understanding of the relationship between attributes and product performance, generating needed product concept and product improvement insight. These are complex multivariate problems where commonly used visualization and modeling methods are often ineffective in extracting nonlinear and higher dimensional effects, leading to increased time to market and reduced product performance.

In development of high-tech products, we need to learn how to make high-margin, complex product designs. Technically it is difficult to realize a product that meets all technical requirements and tolerances, and competitive time pressures result in rushed development. This leads to problems in manufacturing and subsequent redevelopment, which are costly and time consuming. This in turn diverts the resources that should be aligned with product development for new cycles, creating a vicious circle.

In high-tech manufacturing, we are faced with a continual pipeline of new products that need to be made. New technologies may be introduced that require us to develop, qualify and introduce a new process on a regular basis to keep pace with the product design changes that are needed to sustain our competitiveness. The next cycle often starts before yields are optimized with the current cycle. Each new process or technology presents previously unseen problems and challenges, and we are faced with more problems than resources to address them. Engineers struggle to keep their heads above water. And the volumes of data just keep growing.

Sales and marketing must continually attract new customers, help existing customers get more value from products and services, and help customers find new problems or improvement opportunities. There is often an abundance of data, but since not all customer segments behave the same, it is difficult to extract knowledge that can help increase the efficiency and effectiveness of sales and marketing activities.

What is a statistical model?

To benefit from statistical modeling, you obviously require data. It will need to be organized with observations (measured units) in rows and variables (measured values of different attributes) in columns. Some variables may be numeric; others may be character data. Your input variables are referred to as Xs and the other variables are referred to as Ys or outputs.

Given this framework, a statistical model is an empirical model that relates your set of inputs (Xs) to one or more outcomes (Ys). We think of a statistical model as being derived from the relationships in our data to separate the output variation into signal and noise:

$$Y = f(X) + E$$

Where

- **Y** is one or more continuous or categorical responses.
- **X** is one or more continuous or categorical predictors.
- *f*(**X**) describes predictable variation in **Y** (signal).
- **E** describes unpredictable variation in **Y** (noise).

Extracting a useful model – particularly when your Xs are correlated and there are nonlinear and interactive effects of the Xs on our Ys – is getting easier with technology such as JMP.

Case Study 1: Improving an existing process using data exploration and data mining

Three levels of modeling are used:

1. Visual modeling via dynamically linked graphs with zoom and filter to identify major effects.
2. Explanatory modeling using decision trees to isolate joint or interaction effects.
3. Predictive modeling using multiple regression to identify and exploit the way in which many Xs operate together to create variation in our Y.

Example 1

A mature manufacturing process occasionally produces out-of-specification products. Each time this occurs, it requires engineers to troubleshoot and attempt to find the root cause. Key process stages get checked to confirm operation within validated ranges, and no obvious assignable cause emerges. While this effort continues, the process returns to normal without making any changes. Engineers breathe a deep sigh of relief and go back to normal duties – until the next time.

Over the years, many theories as to the cause have been proposed, but none confirmed. Yet continuation of this situation represents significant wasted cost and opportunity:

- Manufacturing and disposal cost of rejected batches costs money.
- Each rejected batch must be reported in an annual product review, and products with a high reject rate are likely to come to the attention of a regulatory agency (or customer).
- In some cases the regulatory agency might remove the license to manufacture due to lack of process understanding.
- In other industries, customers might switch to an alternative supplier able to meet their quality and supply needs.
- Declining market share and reputation.
- Lower profitability affects dividends returned to investors and the availability of capital to reinvest.

Data on the process is abundantly available, and prior approaches to extract meaning from this data include engineering judgment, tabular and graphical summaries, basic statistical comparisons and simple regression. The solutions proposed have not prevented problem reoccurrence. After the last exception, an engineer fresh from a statistical modeling workshop suggested the problem might be due to many factors working together, and that a statistical modeling approach might be helpful. The manager was dubious at first; an engineer himself, he thought his engineers should be able to solve the problem using their engineering knowledge alone. The engineer from the workshop pointed out the obvious: If the problem was easily solved using engineering knowledge alone, why was the problem still unsolved? After some persuasion her manager agreed to allow the engineering group to give statistical modeling a try.

Figure 2 represents a process map of the production process. It is a secondary drug process that involves milling an active pharmaceutical ingredient (API) into a powder of uniform particle size. The milled material is then blended with other ingredients to bulk up and evenly distribute the API. This blended material is then compressed into tablets, which are finally coated to aid shelf life, taste and other properties. At the end of the process, a dissolution test is performed, and if the average dissolution value of several tablets is less than 70 percent dissolved at 120 minutes, the entire batch of tablets is rejected, incurring disposal cost.

The first step was to identify the data that was routinely measured, which is identified in bold type on the process map (Figure 2). The engineer also wanted the variables in gray, but these measurements were not routinely made, so she decided to progress and see what headway she could make with the data that was available to her. Getting her data into a spreadsheet view with rows representing observations and columns representing variables (Figure 3) was straightforward, and she was soon ready to explore and statistically model her data.

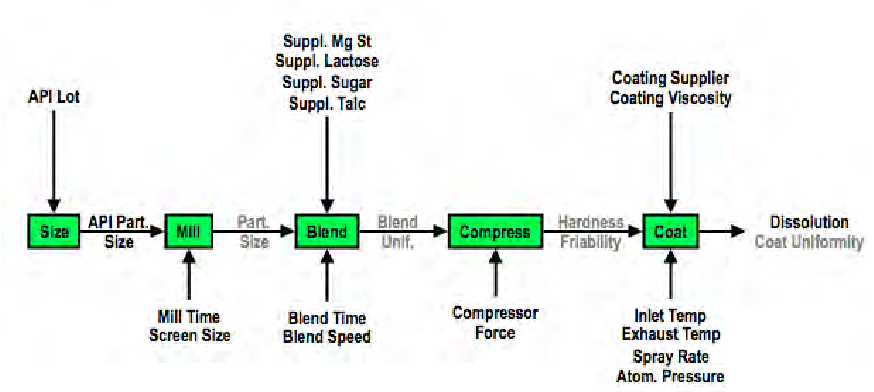
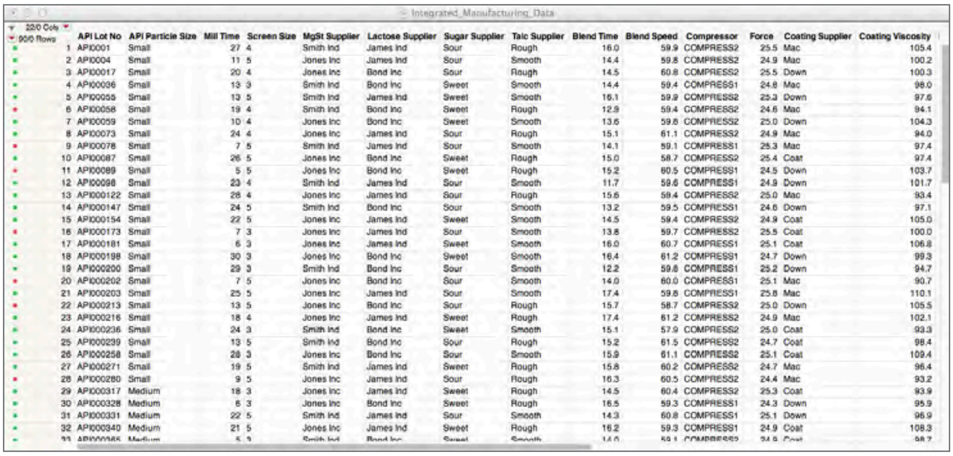


Figure 2: Manufacturing process



	API Lot No	API Particle Size	Mill Time	Screen Size	MgSt Supplier	Lactose Supplier	Sugar Supplier	Talc Supplier	Blend Time	Blend Speed	Compressor	Force	Coating Supplier	Coating Viscosity
1	API0001	Small	27	4	Smith Ind	James Ind	Sour	Rough	16.0	59.9	COMPRESS2	25.5	Mac	105.4
2	API0004	Small	11	5	Jones Inc	James Ind	Sour	Smooth	14.4	59.8	COMPRESS2	24.9	Mac	100.2
3	API00017	Small	20	4	Jones Inc	Bond Inc	Sour	Rough	14.5	60.8	COMPRESS2	25.5	Down	100.3
4	API00036	Small	13	3	Smith Ind	Bond Inc	Sweet	Smooth	14.4	59.4	COMPRESS1	24.8	Mac	98.0
5	API00055	Small	13	5	Smith Ind	James Ind	Sweet	Smooth	16.1	59.9	COMPRESS2	25.3	Down	97.6
6	API00058	Small	19	4	Smith Ind	Bond Inc	Sweet	Rough	12.9	59.4	COMPRESS2	24.6	Mac	94.1
7	API00059	Small	10	4	Jones Inc	Bond Inc	Sweet	Smooth	13.6	59.8	COMPRESS2	25.0	Down	104.3
8	API00073	Small	24	4	Jones Inc	James Ind	Sour	Rough	15.1	61.1	COMPRESS2	24.9	Mac	94.0
9	API00078	Small	7	5	Smith Ind	James Ind	Sour	Smooth	14.1	59.1	COMPRESS1	25.3	Mac	97.4
10	API00087	Small	26	5	Jones Inc	Bond Inc	Sweet	Rough	15.0	58.7	COMPRESS2	25.4	Coat	97.4
11	API00089	Small	5	5	Jones Inc	Bond Inc	Sweet	Rough	15.2	60.5	COMPRESS1	24.5	Down	103.7
12	API00098	Small	23	4	Smith Ind	James Ind	Sour	Smooth	11.7	59.6	COMPRESS1	24.9	Down	101.7
13	API000122	Small	28	4	Jones Inc	James Ind	Sour	Rough	15.6	59.4	COMPRESS2	25.0	Mac	93.4
14	API000147	Small	24	5	Smith Ind	Bond Inc	Sour	Smooth	13.2	59.5	COMPRESS1	24.6	Down	97.1
15	API000154	Small	22	5	Jones Inc	James Ind	Sweet	Smooth	14.5	59.4	COMPRESS2	24.9	Coat	105.0
16	API000173	Small	7	3	Jones Inc	James Ind	Sour	Smooth	13.8	59.7	COMPRESS2	25.5	Coat	100.0
17	API000181	Small	6	3	Jones Inc	James Ind	Sweet	Smooth	16.0	60.7	COMPRESS1	25.1	Coat	106.8
18	API000198	Small	30	3	Jones Inc	Bond Inc	Sweet	Smooth	16.4	61.2	COMPRESS1	24.7	Down	99.3
19	API000200	Small	29	3	Smith Ind	Bond Inc	Sour	Smooth	12.2	59.6	COMPRESS1	25.2	Down	94.7
20	API000202	Small	7	5	Jones Inc	Bond Inc	Sour	Smooth	14.0	60.0	COMPRESS1	25.1	Mac	90.7
21	API000203	Small	25	5	Jones Inc	James Ind	Sour	Smooth	17.4	59.8	COMPRESS1	25.8	Mac	110.1
22	API000213	Small	13	5	Jones Inc	Bond Inc	Sour	Rough	15.7	58.7	COMPRESS2	25.0	Down	105.5
23	API000216	Small	18	4	Jones Inc	James Ind	Sweet	Rough	17.4	61.2	COMPRESS2	24.9	Mac	102.1
24	API000236	Small	24	3	Smith Ind	Bond Inc	Sweet	Smooth	15.1	57.9	COMPRESS2	25.0	Coat	93.3
25	API000239	Small	13	5	Smith Ind	Bond Inc	Sour	Rough	15.2	61.5	COMPRESS2	24.7	Coat	98.4
26	API000258	Small	28	3	Jones Inc	Bond Inc	Sour	Smooth	15.9	61.1	COMPRESS2	25.1	Coat	109.4
27	API000271	Small	19	5	Smith Ind	James Ind	Sweet	Rough	15.8	60.2	COMPRESS2	24.7	Mac	96.4
28	API000280	Small	9	5	Jones Inc	James Ind	Sour	Rough	16.3	60.5	COMPRESS2	24.4	Mac	93.2
29	API000317	Medium	18	3	Jones Inc	James Ind	Sweet	Rough	14.5	60.4	COMPRESS2	25.3	Coat	93.9
30	API000328	Medium	6	3	Jones Inc	Bond Inc	Sweet	Rough	16.5	59.3	COMPRESS1	24.3	Down	95.9
31	API000331	Medium	22	5	Smith Ind	James Ind	Sour	Smooth	14.3	60.8	COMPRESS1	25.1	Down	96.9
32	API000340	Medium	21	5	Jones Inc	James Ind	Sweet	Rough	16.2	59.3	COMPRESS1	24.9	Coat	108.3
33	API000365	Medium	5	3	Smith Ind	Bond Inc	Sweet	Smooth	14.0	59.1	COMPRESS2	24.6	Coat	98.7

Figure 3: Analysis-ready data

One visual modeling approach illustrated in Figure 4 involved plotting the Y (dissolution) against the Xs using side-by-side histograms and dynamic linking to select the failing batches – those with a dissolution below 70 percent dissolved at 120 minutes – and seeing whether the failing batches were clustering at one end of the data range of one or more Xs. Some obvious clustering of defects seems to occur at low values of mill time, high values of screen size and higher values of spray rate.

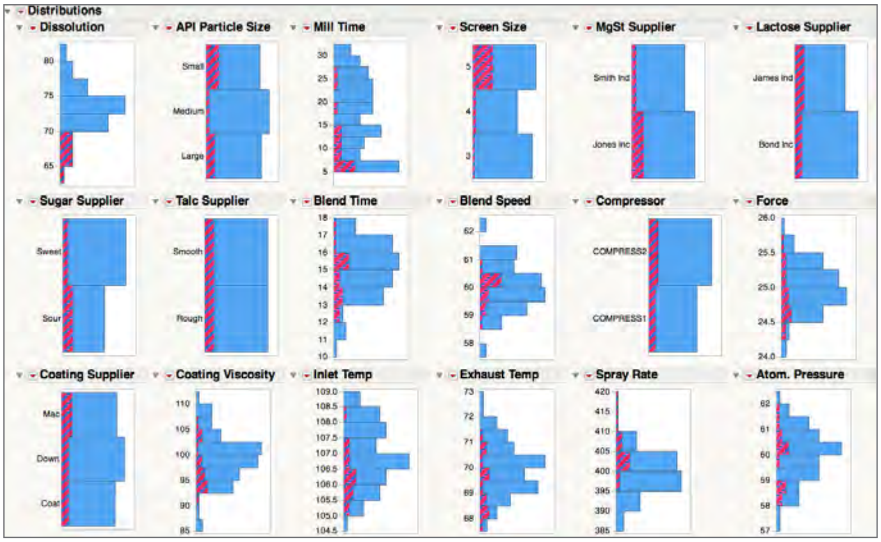
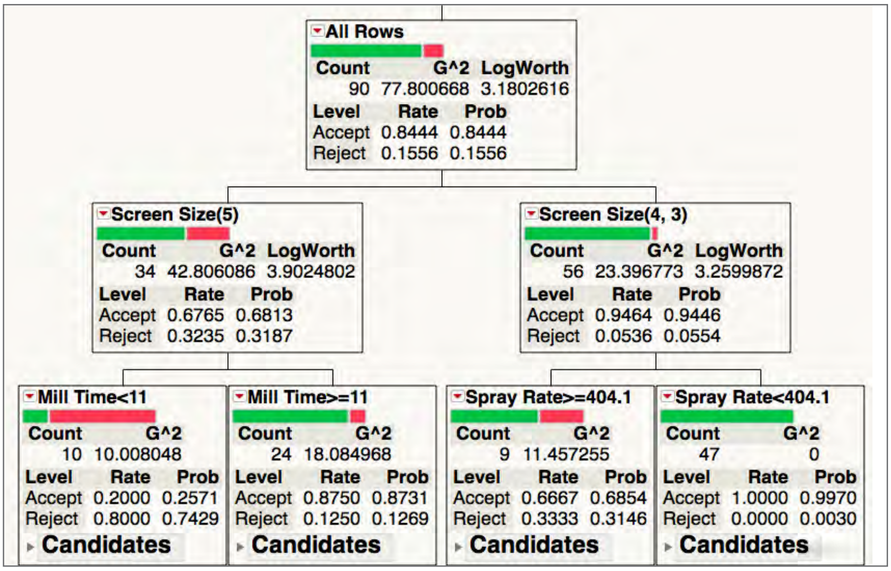


Figure 4: Visual insight

Many of the Xs appear to be related to the dissolution failures, and the engineer wanted to explore how some of the factors may be interacting. A decision tree was constructed using recursive partitioning to identify the top Xs. This can be thought of as an organizational chart of some of your Xs that might be responsible for the variation in a Y. In our particular case, it shows that at the top of the tree we have an overall reject rate of 15.6 percent. With the subset defined by a screen size of 3 or 4 and a spray rate of less than 404.1 (the right-hand branch of the tree), we have a reject rate of 0 percent. Contrast this with a high reject rate of 80 percent in the left-hand branch, which is defined by screen size of 5 and mill time of less than 11 minutes. A decision tree is an organizational chart of the conditional way in which some of our Xs sort the Y from good to bad.



All Rows
Count 90 G^2 77.800668 LogWorth 3.1802616
Level Rate Prob
Accept 0.8444 0.8444
Reject 0.1556 0.1556

Screen Size(5)
Count 34 G^2 42.806086 LogWorth 3.9024802
Level Rate Prob
Accept 0.6765 0.6813
Reject 0.3235 0.3187

Screen Size(4, 3)
Count 56 G^2 23.396773 LogWorth 3.2599872
Level Rate Prob
Accept 0.9464 0.9446
Reject 0.0536 0.0554

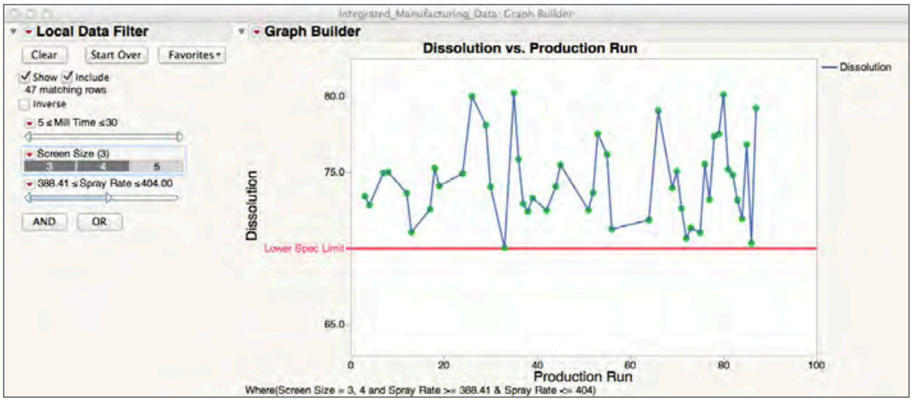
Mill Time<11
Count 10 G^2 10.008048
Level Rate Prob
Accept 0.2000 0.2571
Reject 0.8000 0.7429
Candidates

Mill Time>=11
Count 24 G^2 18.084968
Level Rate Prob
Accept 0.8750 0.8731
Reject 0.1250 0.1269
Candidates

Spray Rate>=404.1
Count 9 G^2 11.457255
Level Rate Prob
Accept 0.6667 0.6854
Reject 0.3333 0.3146
Candidates

Spray Rate<404.1
Count 47 G^2 0
Level Rate Prob
Accept 1.0000 0.9970
Reject 0.0000 0.0030
Candidates

Figure 5: Explanatory insight



Integrated_Manufacturing_Data: Graph Builder

Local Data Filter

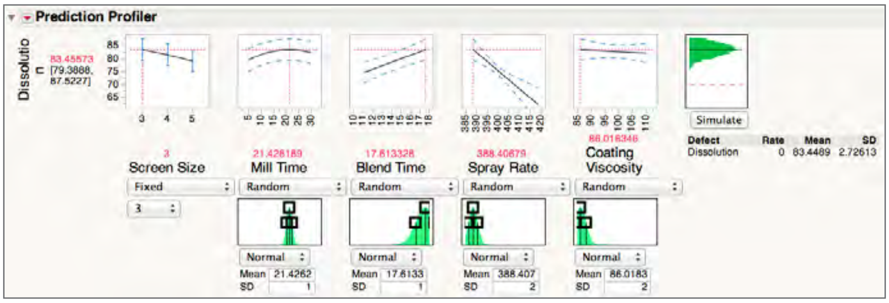
☑ Show ☑ Include
47 matching rows
□ Inverse
5 ≤ Mill Time ≤30
Screen Size (3)
388.41 ≤ Spray Rate ≤404.00

Graph Builder

Where(Screen Size = 3, 4 and Spray Rate >= 388.41 & Spray Rate <= 404)

Figure 6: "Best" subset identified by decision tree

While the solution proposed by the decision tree is an improvement, Figure 6 shows the trend of dissolution for this subset and indicates the process is sometimes likely to operate close to the threshold of failure. It is necessary to shift the mean dissolution to a higher value and preferably shrink the variability around the mean to ensure all future batches stay above the lower specification limit of 70 percent dissolved by 120 minutes.



Factor	Value	Setting	Distribution	Mean	SD
Screen Size	3	Fixed	3		
Mill Time	21.428189	Random	Normal	21.4262	1
Blend Time	17.813328	Random	Normal	17.8133	1
Spray Rate	388.40679	Random	Normal	388.407	2
Coating Viscosity	86.016346	Random	Normal	86.0183	2

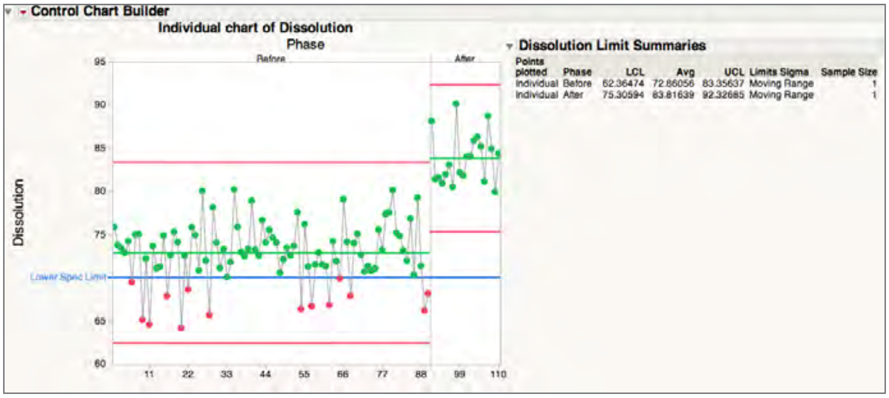
Defect	Rate	Mean	SD
Dissolution	0	83.4489	2.72613

Figure 7: Predictive insight

To gain the level of improvement needed, the engineer used a multiple regression model. Figure 7 shows the critical Xs determined by this method, which are Screen Size, Mill Time, Blend Time, Spray Rate and Coating Viscosity. Further, it shows the nature of the relationship between each X and average dissolution. To get the biggest gain in average dissolution, we need low screen size, a middle-to-high mill time, a high blend time, a low spray rate and low coating viscosity. At the values indicated of 3, 21.4, 17.6, 388 and 86, respectively, the model indicates we can expect an average dissolution of 83.5.

The model can also be queried using Monte Carlo simulation to determine how much variation in the Xs can be tolerated before transmitting undesirable batch-to-batch variation into dissolution. If Mill Time, Blend Time, Spray Rate and Coating Viscosity were unable to be controlled exactly in large-scale manufacturing, and if, for example, the uncontrolled variation is defined by normal distributions with standard deviations of 5, 1, 3 and 5 around the respective means, then Monte Carlo simulation of 5,000 production runs from these input distributions indicate we can expect an output distribution for dissolution with a mean of 83 percent and a standard deviation of 2.7 percent. Further, none of the batches would have a dissolution value below 70 percent dissolved at 120 minutes.

Changing the settings of the five critical inputs to those indicated in Figure 7 resulted in a robust and adequately controlled process, as indicated by the next 20 production batches in Figure 8.



Points plotted	Phase	LCL	Avg	UCL	Limits Sigma	Sample Size
Individual	Before	62.36474	72.86056	83.35637	Moving Range	1
Individual	After	75.30594	83.81639	92.32685	Moving Range	1

Figure 8: Before vs after comparison of dissolution

Summary

Visual and explanatory modeling provided clues as to some of the Xs responsible for the excessive variation in dissolution. Predictive modeling identified the complete set of Xs affecting dissolution and helped the engineers understand the way in which these Xs operated together to drive undesired variation in dissolution. Profiling and simulation capabilities around the model provided the necessary insight to understand what was causing batch-to-batch variation in dissolution and provided the insight needed to fix the problem for good. The solution created significant savings by eliminating the cost of dealing with rejected batches and reduced regulatory risk.

Case Study 2: Accelerating R&D using data exploration and data mining

This case study will introduce additional modeling methods that are helpful when our Xs are strongly related to one another, e.g., X1 and X2 would be described as strongly related if X2 increases (or decreases) when X1 increases. Special modeling methods such as ridge or lasso regression, partial least squares and neural networks may be beneficial in such situations.

Example 2

A drug discovery company typically has several million chemical compounds in its library and wishes to improve efficiency and effectiveness of drug candidate identification and improvement using data from the company's chemical and biological databases. In particular the company is keen to improve the knowledge created to answer questions such as:

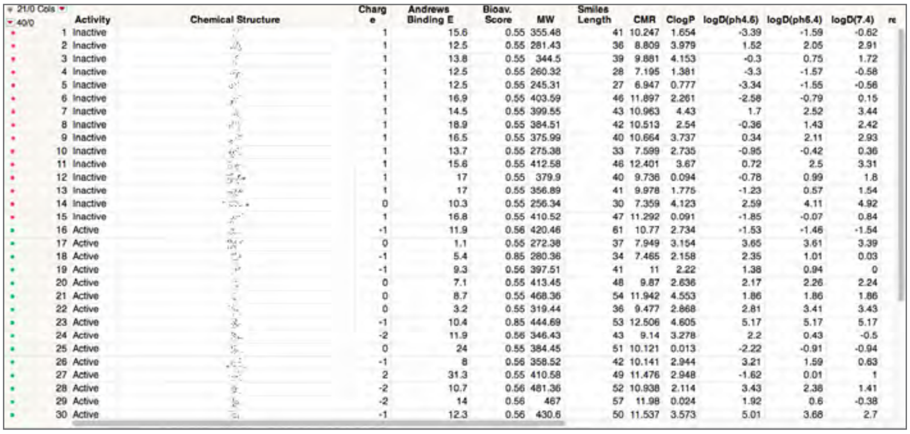
- Which compounds are more likely to be active for a particular disease?
- Which parts of a molecule should be targeted to improve activity or safety, and how?

The chemical database contains various chemical descriptors, e.g., length, shape, charge and molecular weight. The biological database contains activity and safety indicators for various target diseases. To improve the knowledge gained from these databases, we need to devise better models of the relationship between chemical descriptors and biological activity/safety.

Chemists frequently used 2-dimensional and 3-dimensional visual insight methods, multivariate modeling methods that assume linear relationships between chemical descriptors and activity/safety were also used. The relationships between activity and chemical descriptors are complex and it is the way in which many descriptors operate together that influences activity. Two-dimensional and three-dimensional visual insight methods do not scale to providing insight from how a few descriptors influence activity to how many descriptors work together to influence activity. Current multivariate statistical technology was too complex for many researchers and did not cope particularly well when the relationships are interactive and/or nonlinear in nature.

Many researchers were not performing multivariate statistical analysis or were not able to extract nonlinear and interactive effects with multivariate statistical analysis. Therefore the company was not realizing the potential value of multivariate statistical analysis. Decisions were based on partial understanding from visualization methods or less-effective multivariate modeling methods, resulting in a larger number of learning cycles, which affected the speed and predictability of R&D.

Getting his data into a spreadsheet view with rows representing observations and columns representing variables (Figure 9) was straightforward, and the chemist was soon ready to statistically model his data.



21/0 Cols 40/0	Activity	Chemical Structure	Charge	Andrews Binding E	Bioav. Score	MW	Smiles Length	CMR	ClogP	logD(ph4.6)	logD(ph6.4)	logD(7.4)	rc
1	Inactive		1	15.6	0.55	355.48	41	10.247	1.654	-3.39	-1.59	-0.62	
2	Inactive		1	12.5	0.55	281.43	36	8.809	3.979	1.52	2.05	2.91	
3	Inactive		1	13.8	0.55	344.5	39	9.881	4.153	-0.3	0.75	1.72	
4	Inactive		1	12.5	0.55	260.32	28	7.195	1.381	-3.3	-1.57	-0.58	
5	Inactive		1	12.5	0.55	245.31	27	6.947	0.777	-3.34	-1.55	-0.56	
6	Inactive		1	16.9	0.55	403.59	46	11.897	2.261	-2.58	-0.79	0.15	
7	Inactive		1	14.5	0.55	399.55	43	10.963	4.43	1.7	2.52	3.44	
8	Inactive		1	18.9	0.55	384.51	42	10.513	2.54	-0.36	1.43	2.42	
9	Inactive		1	16.5	0.55	375.99	40	10.664	3.737	0.34	2.11	2.93	
10	Inactive		1	13.7	0.55	275.38	33	7.599	2.735	-0.95	-0.42	0.36	
11	Inactive		1	15.6	0.55	412.58	46	12.401	3.67	0.72	2.5	3.31	
12	Inactive		1	17	0.55	379.9	40	9.736	0.094	-0.78	0.99	1.8	
13	Inactive		1	17	0.55	356.89	41	9.978	1.775	-1.23	0.57	1.54	
14	Inactive		0	10.3	0.55	256.34	30	7.359	4.123	2.59	4.11	4.92	
15	Inactive		1	16.8	0.55	410.52	47	11.292	0.091	-1.85	-0.07	0.84	
16	Active		-1	11.9	0.56	420.46	61	10.77	2.734	-1.53	-1.46	-1.54	
17	Active		0	1.1	0.55	272.38	37	7.949	3.154	3.65	3.61	3.39	
18	Active		-1	5.4	0.85	280.36	34	7.465	2.158	2.35	1.01	0.03	
19	Active		-1	9.3	0.56	397.51	41	11	2.22	1.38	0.94	0	
20	Active		0	7.1	0.55	413.45	48	9.87	2.636	2.17	2.26	2.24	
21	Active		0	8.7	0.55	468.36	54	11.942	4.553	1.86	1.86	1.86	
22	Active		0	3.2	0.55	319.44	36	9.477	2.868	2.81	3.41	3.43	
23	Active		-1	10.4	0.85	444.69	53	12.506	4.605	5.17	5.17	5.17	
24	Active		-2	11.9	0.56	346.43	43	9.14	3.278	2.2	0.43	-0.5	
25	Active		0	24	0.55	384.45	51	10.121	0.013	-2.22	-0.91	-0.94	
26	Active		-1	8	0.56	358.52	42	10.141	2.944	3.21	1.59	0.63	
27	Active		2	31.3	0.55	410.58	49	11.476	2.948	-1.62	0.01	1	
28	Active		-2	10.7	0.56	481.36	52	10.938	2.114	3.43	2.38	1.41	
29	Active		-2	14	0.56	467	57	11.98	0.024	1.92	0.6	-0.38	
30	Active		-1	12.3	0.56	430.6	50	11.537	3.573	5.01	3.68	2.7	

Figure 9: Analysis-ready data

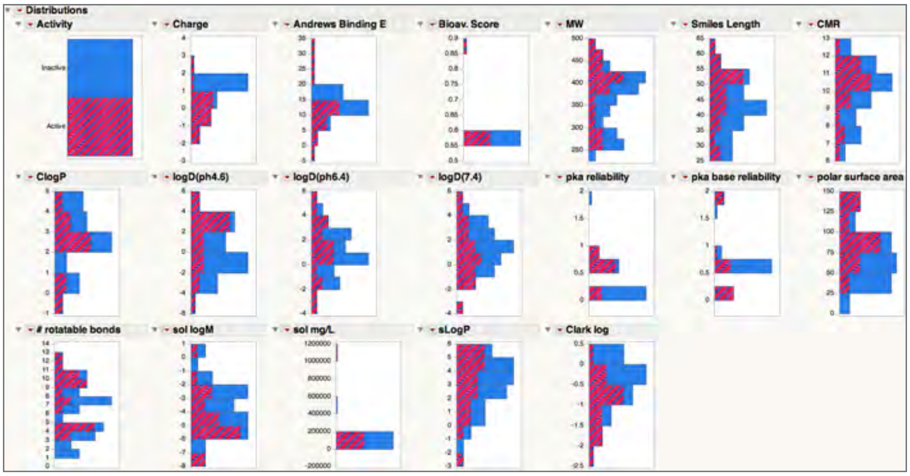


Figure 10: Visual insight

By plotting activity alongside the Xs using side-by-side histograms, as illustrated in Figure 10, and using dynamic linking to select the active chemicals, it was possible to see if the active chemicals clustered at one end of one or more Xs. Some obvious clustering of active chemicals seems to occur at low values of charge, high values of smiles length (bigger molecules), higher log dissolution at pH 4.6, larger polar surface area and lower clark log.

Regular two-dimensional graphs as depicted in Figure 11 enable drill-down to view molecular structure or other images.

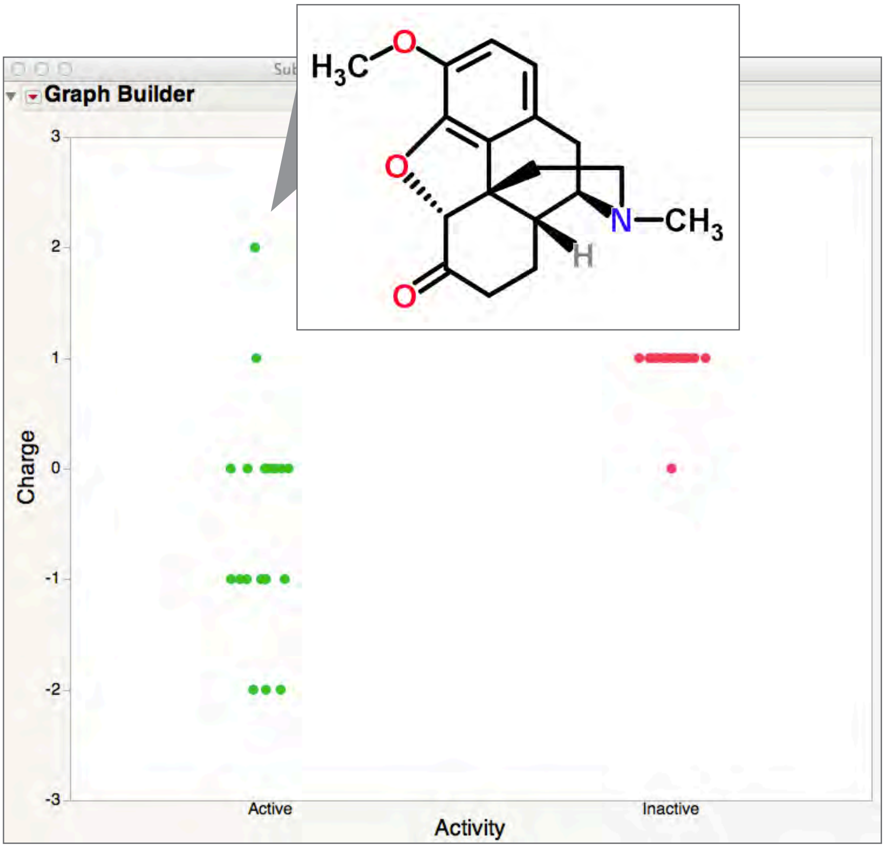


Figure 11: Hover over any data point to generate a pop-up graph showing detailed molecular structure

Various statistical modeling methods were investigated and compared, including lasso regression, logistic regression, partial least squares and neural networks. A neural network was determined as the best model since it was able to predict the activity of other compounds not used in model building more accurately than any of the other methods. Based on examination of the Profiler resulting from the neural network model in Figure 12, it appears that the neural network model was able to identify nonlinear relationships between the Xs and activity.

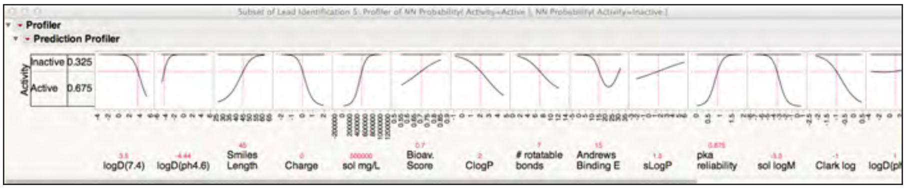


Figure 12: Profiler of neural network model

The neural network model was used to score the chemicals database to predict the probability of each chemical within the database being active for the target disease. The top 5 percent of chemicals – the top 5 percent with the highest prediction probability – were selected for further study. From a chemical understanding perspective, the Profiler suggests that chemicals with a low log dissolution at pH 7.4, a high log dissolution at pH 4.6, a long smiles length (larger molecules), a lower charge and so on are more likely to be active chemicals.

Summary

Visual modeling is effective for low-dimensional problems; however, predictive modeling better provides the insight necessary to solve high-dimensional problems. Predictive modeling delivered a scoring formula to rank or sort the database from most likely to be active to least likely to be active, and the top 5 percent of chemicals were selected for further study.

Statistical modeling drives deeper understanding better, faster and easier, with the resulting benefit of fewer cycles of learning and shorter R&D cycles and a more predictable outcome. Because of the easy-to-use interface of JMP statistical discovery software, more researchers are able to gain a better understanding of their data with statistical modeling.

Conclusion

To keep it brief, just two examples were presented. With access to data on the process or system you wish to investigate, you could similarly create a spreadsheet view of your data and start to model the dependence of your Ys on your Xs.

This paper has attempted to explain:

- What is a statistical model.
- Types of problem that benefit from statistical modeling.
- Simple and effective ways to build statistical models.
- How to extract understanding from statistical models.
- How to present and communicate statistical models and resultant understanding to other stakeholders.
- How to make better decisions, faster.

Statistical modeling might help you or your company to:

- Accelerate innovation.
- Deliver robust products and process that work every time.
- Speed time to market.
- Deliver a competitive edge.
- Improve customer loyalty.
- Increase growth and return.
- Reduce costs.
- Enhance your brand and claim more market share.

Questions? Contact the JMP office nearest you: **jmp.com/contact**.

About SAS and JMP

JMP is a software solution from SAS that was first launched in 1989. John Sall, SAS co-founder and Executive Vice President, is the chief architect of JMP. SAS is the leader in business analytics software and services, and the largest independent vendor in the business intelligence market. Through innovative solutions, SAS helps customers at more than 75,000 sites improve performance and deliver value by making better decisions faster. Since 1976 SAS has been giving customers around the world THE POWER TO KNOW®.

sas®
THE POWER TO KNOW®

SAS Institute Inc. World Headquarters +1 919 677 8000
JMP is a software solution from SAS. To learn more about SAS, visit **sas.com**
For JMP sales in the US and Canada, call 877 594 6567 or go to **jmp.com**