Reprinted from
PHARMACEUTICAL ENGINEERING®
The Official Magazine of ISPE
March/April 2007, Vol. 27 No. 2

Lean Data Analysis: Simplifying the Analysis and Presentation of Data for Manufacturing Process Improvement

by Malcolm Moore

This article introduces a data analysis maturity model that maps various tools and methodologies aimed at predicting, analyzing, improving, or controlling the drivers of product quality to the extent to which these techniques may help reduce defects. By mapping tools currently deployed in a particular manufacturing facility to the maturity model, it is possible to define a cost-effective road map for various initiatives aimed at improving product quality through increased process understanding. Pragmatic data analysis and reporting approaches are introduced to aid process understanding for mainstream users and the deployment of that understanding in manufacturing to increase product performance.

Introduction

To achieve increased process understanding via Six Sigma, Process Analytical Technology (PAT), or other methodologies requires adoption of at least three types of technology:

1. measurement technology to gauge process and material inputs and intermediate product
2. data integration and cleansing technology to bring together disparate sources of data – including process, material, intermediate, and final product data sources – in a timely and effective manner
3. data analysis and reporting technology to bring understanding from integrated data collected in the context of a problem or improvement opportunity

Emphasis on measurement technology alone will increase the extent to which process and materials are measured, and will drive up costs and data volumes. The lack of effective data integration and data analysis methods for all consumers of the data will limit the growth in process understanding and the ability of manufacturing to exploit this understanding.

Figure 1 presents a high-level process model of data integration and data analysis in manufacturing. The components represented in blue depict the IT function of integrating disparate data sources, including databases, electronic and paper sources, then cleansing and transforming data to an analysis-ready state with a data model that is easily maintained and extended as the number and type of data sources grow.

The need for a data integration solution – and the level of sophistication required of it – will depend upon the extent to which inputs are measured. In newer production lines, it may be common to measure hundreds of input variables via NIR spectroscopy, and other inline,

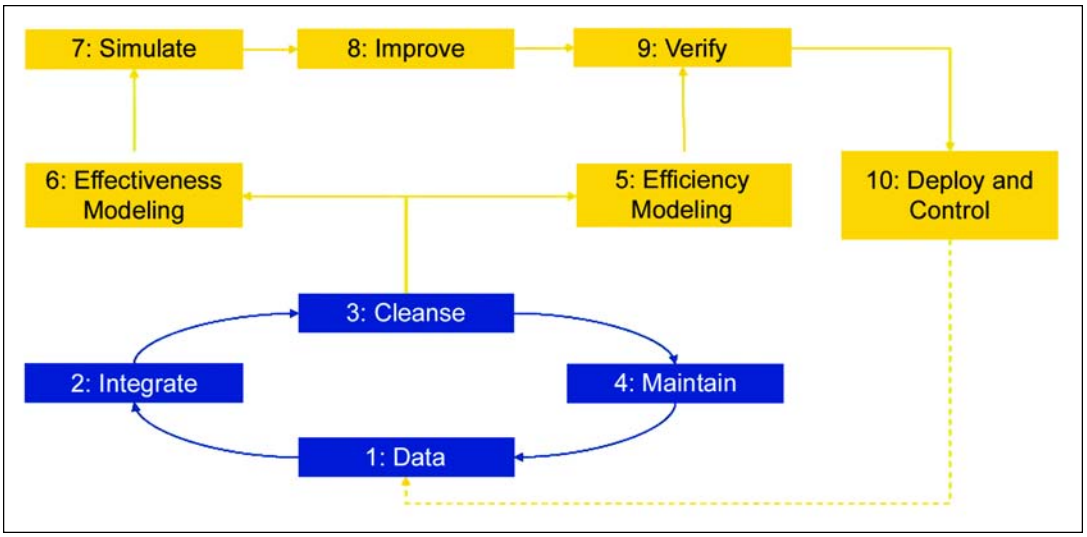

Figure 1. PAT Data Integration, Modeling, Improvement, and Control Process.

at-line, online, or offline methods, requiring a data integration solution. However, older production facilities may focus on offline laboratory testing of intermediate and end-of-line products and use measurement technologies for process inputs on an as-needed basis.

The process steps represented in orange symbolize some ways that business users might analyze their data. At least two different approaches to modeling the relationships in cleansed or analysis ready data are available and the terms *efficiency* and *effectiveness* modeling are introduced to distinguish the two approaches. Efficiency modeling is used to classify models that use multivariate relationships to predict manufacturing problems. Such models do not necessarily result in model simplification or reduction of the number of dimensions that need to be measured, nor do they greatly increase understanding of how the key inputs drive variation in product quality. Effectiveness modeling, on the other hand, is used to describe approaches that identify the critical few inputs and define empirical transfer functions that describe how these key inputs operate together to drive manufacturing problems or issues, increasing our understanding of how those inputs affect variations in quality. These two modeling approaches are described in more detail below and are illustrated by case studies.

This article focuses on pragmatic approaches to data analysis and reporting that work regardless of the extent to which inputs are measured. It introduces ways of simplifying data analysis and reporting approaches associated with PAT, Six Sigma, and related methodologies and proposes a way to define a road map for the adoption of manufacturing improvement technologies relative to the current level of measurement maturity. Mapping of a broad set of tools to a data analysis maturity model are presented along with examples of various data analysis approaches, including a set of pragmatic analysis techniques that are simple to apply and understand at all levels of an organization.

PAT Data Analysis Methods

Modeling Approaches

Statistical modeling approaches to PAT are classified in two ways: models for increasing the efficiency of manufacturing – reducing waste; and models for increasing effectiveness of manufacturing – enhancing process understanding and utilizing it to improve manufacturing performance.

Efficiency models consist of classification modeling techniques, such as discriminant analysis, cluster analysis, and decision trees, along with predictive modeling techniques such as Partial Least Squares (PLS) and Principal Component Regression (PCR). These techniques exploit the multivariate relationships among a large number of measured inputs to predict product performance or batch failures ahead of time. Compared with effectiveness modeling methods,

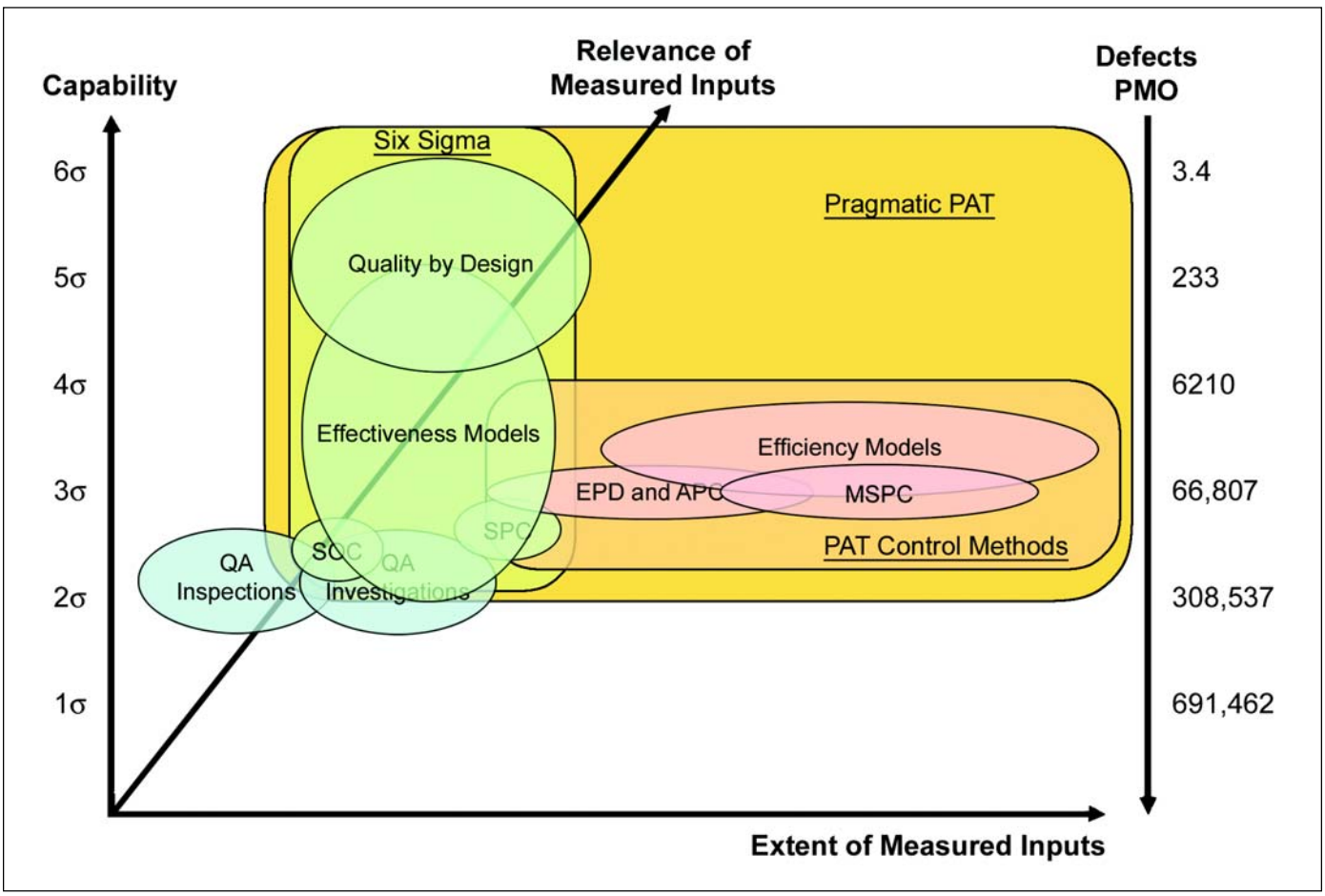

Figure 2. Mapping of data analysis technology to process capability and dependence on extent and relevance of measured inputs.

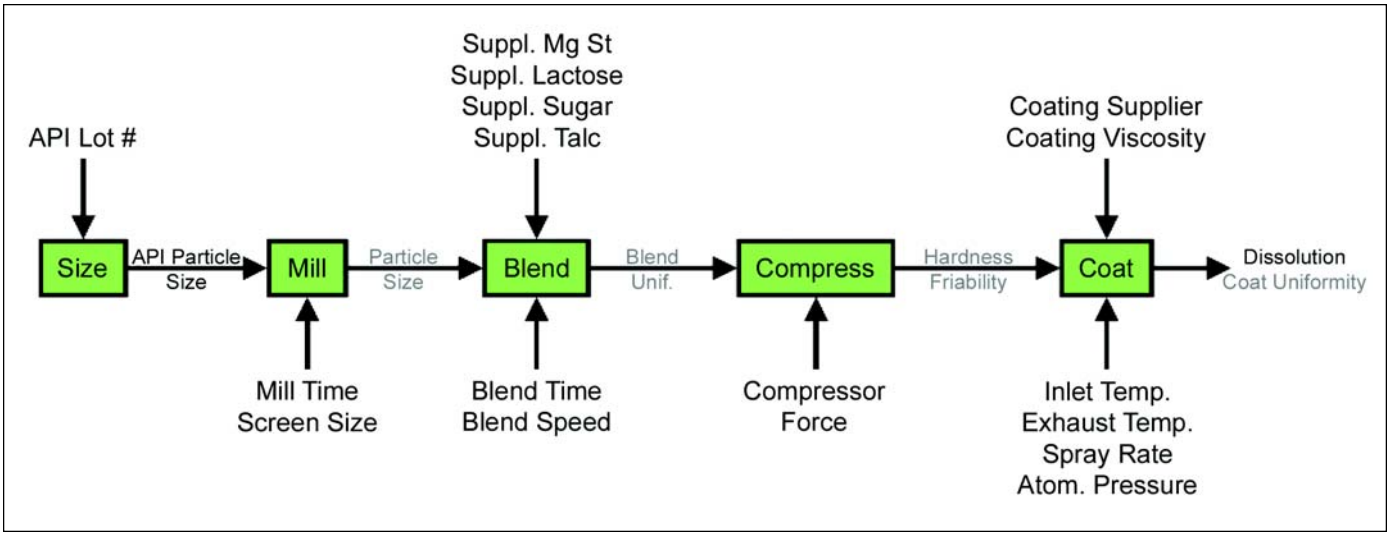

Figure 3. Key processes and inputs associated with excessive variation in 60 minute dissolution.

efficiency models require a large number of measured inputs – in fact, the more the better – and tend to be used for "black box" batch classification or prediction of likely product performance. In other words, they make good predictions, but they do not necessarily deliver fundamental changes in process understanding.

Effectiveness models consist of variable reduction or exploratory data analysis methods, such as data mining, correlation analysis, process mapping, cause-and-effect analysis, Quality Function Deployment (QFD), Failure Mode Effects Analysis (FMEA) to identify the critical few inputs that are investigated in more detail via Design of Experiments (DOE), multiple regression, and generalizations of multiple regression for non-normally distributed measures of product quality. With care, these latter techniques develop empirical models that approximate the causal relationships between the critical few inputs and product quality.

Examples of some of these different modeling approaches are provided in the case studies below.

		Actual			
	Count Total % Col % Row %	Too Low	Good	Too High	
Predicted	Too Low	1 1.39 100.00 50.00	1 1.39 1.54 50.00	0 0.00 0.00 0.00	2 2.78
	Good	0 0.00 0.00 0.00	63 87.50 96.92 98.44	1 1.39 16.67 1.56	64 88.89
	Too High	0 0.00 0.00 0.00	1 1.39 1.54 16.67	5 6.94 83.33 83.33	6 8.33
	Total	1 1.39	65 90.28	6 8.33	72

Table A. Predicted by actual batch classification.

Maturity Model

Figure 2 may be useful when considering the best mix of data analysis methodologies to increase process understanding for a particular manufacturing facility. It may help establish a baseline for your manufacturing facility with regard to product quality performance, define goals for a proposed PAT investment, and help define a road map for getting to those performance goals.

This maturity model maps data analysis methodologies against sigma capability. Sigma is the measure of variability in the product quality measure, usually calculated by assuming the product quality measure is normally distributed. A 2 sigma process is one where the mean ± two standard deviations coincide with the specification limits of the product quality measure. In this case, approximately 5% of batches would not meet the required quality specification (approximately 2.5% in each tail of the distribution). Defects Per Million Opportunities (PMO) is calculated after assuming a shift of 1.5 sigma in the mean of the product quality measure. Hence, a 2 sigma process encountering a 1.5 sigma shift in the mean from target would result in 308,537 defects PMO. Thus, a high sigma capability value such as 5 or 6 is required to ensure little or no defects after allowing for a shift in the process mean.

Most mature manufacturing facilities deploy a combination of QA inspections, Statistical Quality Control (SQC) – control charts applied to product quality measures, and QA investigations in an attempt to trace the cause of batch exceptions. Such approaches generally achieve sigma capability of up to 2.5. The introduction of SPC, where control charts are applied to intermediate product measurements, may get performance up to the region 3 sigma.

More sophisticated control methods, such as End-Point Detection (EPD) and Advanced Process Control (APC) can be deployed to reduce variation in intermediate product and help reduce variation in final product to 3 sigma or thereabouts. Utilization of inline measurement tools in conjunction with EPD to achieve a specified moisture content in

drying operations or specified blend uniformity in blending operations, or APC to have short loop controls to make "corrective" adjustments during unit process operations are appealing control strategies. However, to have real impact, these approaches require detailed understanding of the extent to which changes to the mean, within-batch and between-batch variability of factors, such as moisture content, affect the sigma capability of final product quality. Used in isolation, EPD and APC are unlikely to get final product capability much beyond 3 sigma. Multivariate Statistical Process Control (MSPC) is an extension of SPC that exploits the correlations between measured inputs to give greater resolution in detecting problems ahead of time and reduce the extent of false alarms. Like EPD and APC, MSPC is most effective when deployed in conjunction with effectiveness modeling strategies. Early indications of process drift, detected by a control mechanism when investigated via effectiveness modeling methods, may result in increased process understanding that may be used to revise the control strategy to sustain the gain. Used in isolation, neither MSPC nor any other control method is likely to get final product quality appreciably beyond 3 sigma capability.

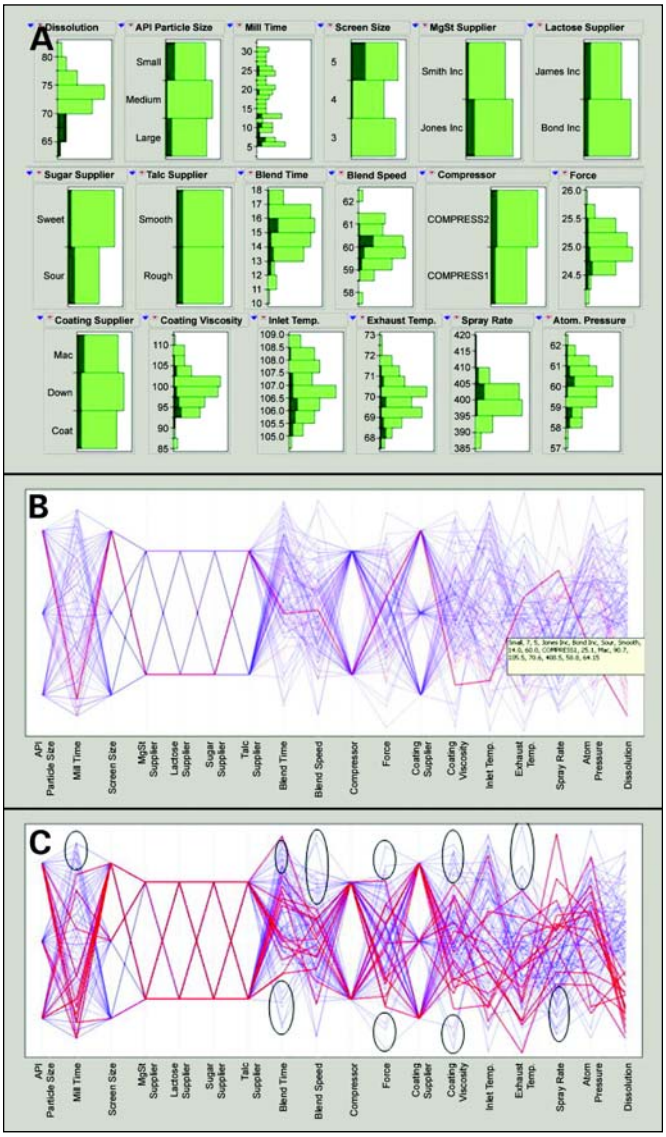

Figure 4. Visual exploration of relationships.
a. Distributions of 60 minute dissolution and key inputs with dissolution failures identified in dark green.
b. Parallel of key inputs with the values of the processing conditions identified for a failing lot.
c. Parallel plot with settings of some key inputs worthy of further investigation.

Efficiency models such as PLS and PCR work best when tens or hundreds of measured inputs are available for each production batch. They exploit the correlations between the inputs to produce reliable classification or prediction models of product quality. These techniques need to be used in conjunction with effectiveness modeling methods if the goal is to increase sigma capability of product quality through increased process understanding. Efficiency models provide an effective basis for prioritizing the input variables for inclusion in effectiveness modeling activities. Used in isolation, efficiency models are likely to get capability in the region of 3 to 4 sigma.

The focus of effectiveness modeling is to identify the critical few inputs and to develop empirical models of the effects of these on product quality that approximate the causal relationships between inputs and product quality. This new knowledge is then deployed to reduce variation in final product quality and achieve performance requirements through revised process and material specifications and controls. Effectiveness modeling approaches applied to the process development of new products is otherwise called Quality by Design (QbD). For new products, Quality by Design is a good way of achieving six sigma quality performance. This approach ensures a high level of process understanding along with cost-effective control strategies in manufacturing that are based on measurement and control of the critical few relevant inputs. Measuring and controlling everything that can be measured increases production costs and cycle times unnecessarily. Compared to process development of new products in R&D, effectiveness modeling of mature manufacturing processes requires a little more care due to the inherent correlations in measured inputs and the limited range over which inputs are varied. With appropriate consideration of these constraints, effectiveness modeling can deliver increased process understanding and higher levels of sigma capability, along with specifications and controls concerning the relevant few process and material inputs.

Efficiency modeling along with advanced control strategies, such as EPD, APC, and MSPC, are classified as PAT Control Methods. Effectiveness modeling and QbD are classified as Six Sigma Methods. Lean Data Analysis or Pragmatic PAT as illustrated in the case studies is an appropriate blend of simplified Six Sigma Methods and PAT Control Methods. The best blend for a particular manufacturing facility depends on where the facility is positioned within the matrix.

Various factors need to be considered when determining the best mix of data analysis methodologies to enable increased process understanding for a particular manufactur-

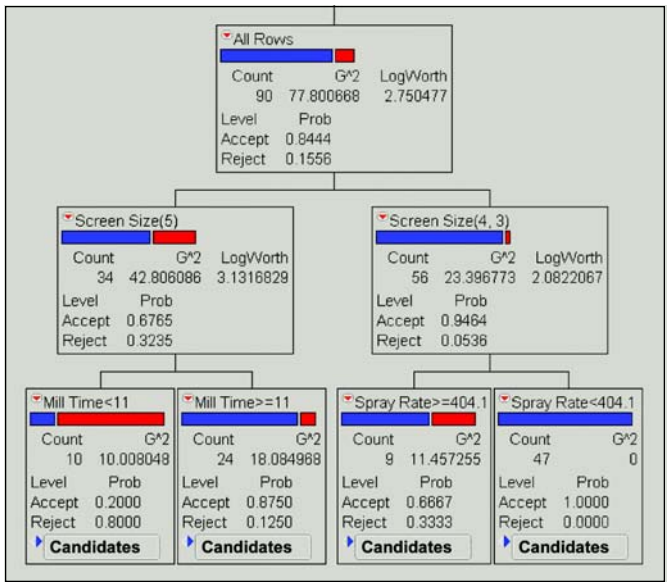

Figure 5. Recursive partitioning decision tree.

ing facility, including product maturity, sigma capability, and the extent and relevance of measured and unmeasured inputs. A mature manufacturing facility is unlikely to have extensive inline, at-line, or online measurement tools in place; therefore, greater emphasis on effectiveness modeling and Six Sigma approaches will be appropriate. For newer manufacturing processes with an extensive number of measured inputs, there may be a greater mix of PAT Control Methods although appropriate use of effectiveness modeling methods also will be required to ensure fundamental understanding of the process. After positioning a particular manufacturing facility within the maturity matrix, it is possible to map out short- and long-term goals of a quality improvement or PAT investment and define a high-level road map for achieving those goals.

Case Studies

These are fictional case studies based on simulated data, copies of which are available on request from the author. The scenarios around which the data have been simulated are fairly typical of the data sparse situation of mature manufacturing and data rich position of some manufacturing facilities of new products. These simulated situations are not based on any particular case, but they do try to reflect the realities of the two situations and by so doing provide data analysis examples that are easier to apply and understand for the mainstream.

Case Study 1: Mature Manufacturing with Few Measured Inputs

This case study concerns a manufacturing facility that has been producing an established product in the form of solid doses at various concentrations for several years. Current measurement systems are based on storing finished material, while offline QA tests are performed to assure the finished product meets the performance specification.

The case study focuses on investigating the process for tablets produced at a single concentration. The key performance metric is 60-minute mean dissolution, which must be no less than 70%. Historically, 16% of production batches fail to meet the 60-minute dissolution requirement and QA investigations into these lot failures rarely find an assignable cause.

In this data-sparse scenario, the manufacturing team was commissioned to investigate the process and dramatically improve sigma capability. The team adopted a variety of effectiveness modeling techniques, starting with process mapping, which was used to identify the key process steps and to identify the set of inputs that were most relevant to the problem and easy to collect information about retrospectively. The results of this process-mapping exercise are documented in Figure 3; the set of inputs that might have an impact on 60-minute mean dissolution and are easily collected retrospectively are identified in black type. Inputs occurring above a process step represent material properties; inputs occurring below a process step represent process parameters.

Data on the inputs identified in black type along with mean dissolution were collated for the last two years of production batches, which resulted in a data set consisting of 90 rows and 19 columns.

Exploratory data mining methods as indicated in Figure 4 were deployed to help determine the inputs most strongly associated with dissolution failures. Part (a) of Figure 4 shows simple histograms of each variable with the failing batches identified in dark green. This shows a particularly strong relationship between screen size in the milling step and batch failure with a larger screen size resulting in a greater proportion of failures – presumably a larger screen size results in larger API particle size and these larger particles take longer to dissolve. Spray rate in the coating

V9	V15	V18	V21	% at stage 3-4
0.25	0.5	0.9	0.7	29.24
0.25	0.35	0.9	0.95	27.10
0.25	0.5	0.65	0.95	17.88
0.25	0.2	0.65	1.2	21.93
0.2	0.2	0.9	0.7	33.24
0.3	0.2	0.4	0.95	22.47
0.2	0.2	0.4	1.2	19.46
0.3	0.2	0.9	0.7	34.69
0.3	0.5	0.9	1.2	17.83
0.2	0.5	0.9	1.2	14.61
0.3	0.35	0.65	0.7	25.96
0.25	0.2	0.4	0.7	23.77
0.3	0.5	0.4	0.7	20.36
0.2	0.5	0.4	0.7	19.52
0.25	0.35	0.4	1.2	14.80
0.2	0.35	0.65	0.95	20.56

Table B. DOE worksheet.

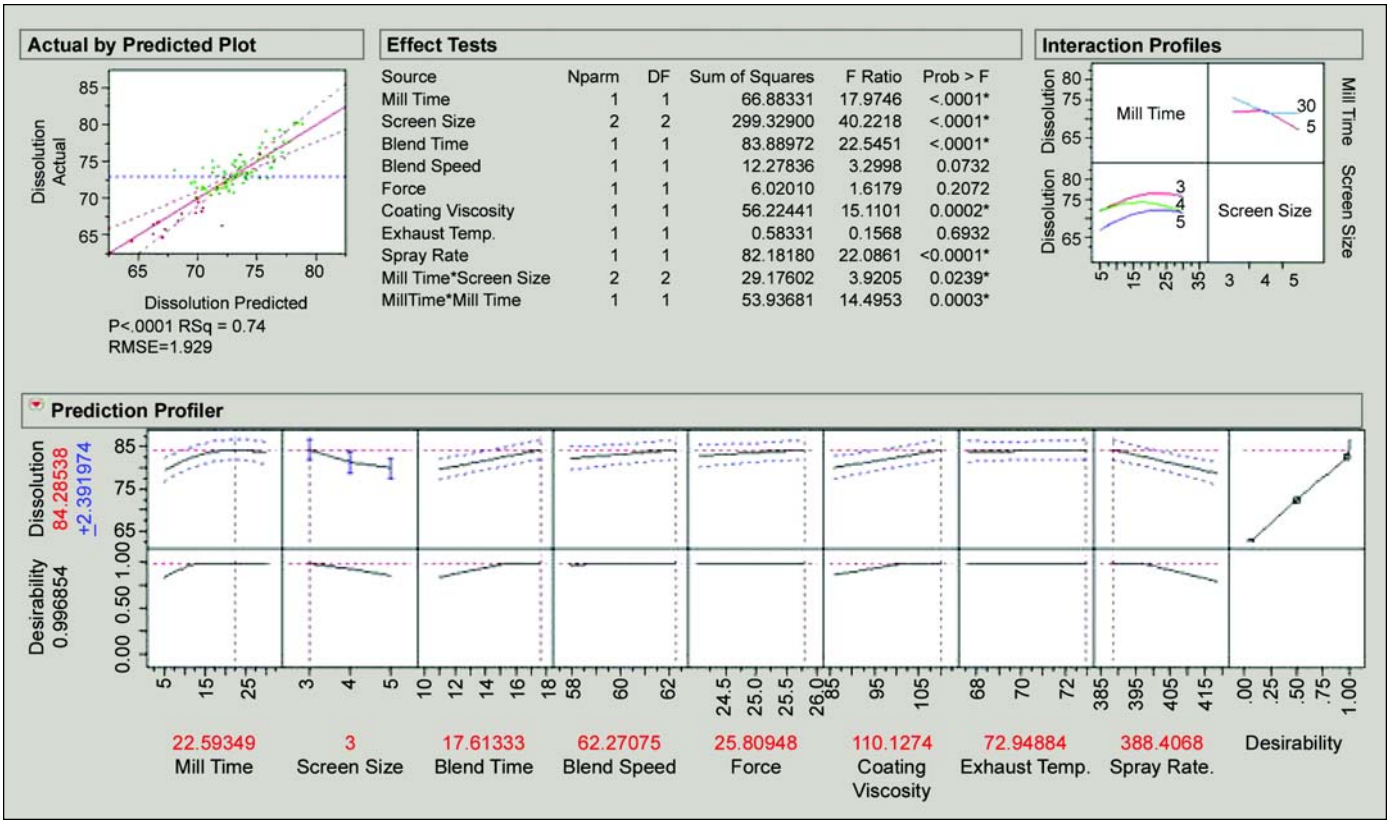

Figure 6. Multiple regression analysis summary.

step also has a strong association with batch failures; in particular, lower spray rates have no batch failures. Part (b) of Figure 4 shows the multivariate relationship of the 18 variables in a single graph called a parallel coordinates plot. There are 90 lines on this graph – one line for each production batch – with the path of each line representing the processing conditions of each batch along with the resulting dissolution test result for that batch. By design, there is no scale on the y-axis; instead, a plotting range for each variable is selected to show the span of data values for that variable. The failing batches are identified in red and the passing batches in blue. The values of each variable for one of the failing batches is illustrated in bold red, showing API with a small particle size, processed with a mill time of seven minutes, screen size of five, and so on. Part (c) of Figure 4 shows the parallel coordinates plot with all failing batches identified in bold red; this version of the graph enables the identification of processing conditions associated with passing or failing batches. Some processing conditions associated with passing batches are circled, e.g., high mill time, low and high blend time (with one exception), high blend speed, low and high force, low and high coating viscosity, high exhaust temperature, and low spray rate appear to be more favorable processing conditions. Potential interactive effects of two or more inputs on dissolution can be investigated on both graph types by coloring the points according to different rules. For example, to investigate the size of the interactive effect of blend time and blend speed on mean dissolution, a cut point would be defined for each input (giving high and low values of each input) and then color the points differently for the four combinations. We would then look to see if there is an appreciable change in mean dissolution across the four combinations of the two variables. One potential draw-back of the parallel coordinates plot is that it is not as effective at exploring the effects of categorical variables such as API Particle Size, Screen Size, and Coating Supplier, due to the inability to display the proportion of failing/passing batches processed at each level of a categorical variable. Nonetheless, it is a good visual data mining tool that helps identify key continuous variables for further investigation.

Another useful exploratory data mining method is recursive partitioning. This method repeatedly partitions data according to a relationship between the input variables and an output variable, creating a tree of partitions. It finds the critical input variables and a set of cuts or groupings of each that best predict the variation in batch failures. Variations of this technique are many and include: decision trees, CART™, CHAID™, C4.5, C5, and others.

Figure 5 shows the resulting decision tree using recursive partitioning to explore the main drivers of batch failures. The right-hand branch of the decision tree shows that 47 of the 90 batches were processed using a screen size of four or three in conjunction with a spray rate less than 404. All 47 batches passed the dissolution test. At the other extreme, the left-hand branch shows that 10 batches were processed using a screen size of five and a mill time of less than 11. Eight of these batches failed the dissolution test.

These exploratory data mining methods have collectively

identified a subset of inputs – Mill Time, Screen Size, Blend Time, Blend Speed, Force, Coating Viscosity, Exhaust Temperature, and Spray Rate – worthy of further investigation. The methods have several advantages over conventional statistical approaches, including:

1. ease interpretation and communication, enabling everyone to gain insight into the potential key relationships in data

2. inform the mainstream about the principles of statistical thinking, particularly those of modeling variation in process outputs and identifying the key drivers of process variation

The effects of this subset of input variables upon 60 minute mean dissolution were investigated in more detail using multiple regression in Figure 6. The graph at the top shows that the model predicts actual values of dissolution reasonably well, and the effects tests summary shows that all, but blend speed, force, and exhaust temperature significantly contribute to variation in 60 minute mean dissolution at the 5% level. Further mill time and screen size have an interactive effect and mill time has a quadratic effect on 60 minute mean dissolution as illustrated in the interaction profile. The prediction profiler at the bottom of Figure 6 shows the direction and strength of the effects of each factor on 60 minute mean dissolution with the optimum settings of each input given in red. Since blend speed, force, and exhaust temperature do not significantly affect 60 minute mean dissolution at the 5% level, any value of these three inputs within the observed range are acceptable. The anomaly of one failing batch with a high blend time in Figure 4(c) was due to a low mill time (10 minutes) and screen size of five.

To investigate process robustness against the proposed new set points, the simulation illustrated in Figure 7 was performed. Using the multiple regression model as the transfer function between the key process inputs and 60 minute mean dissolution, 1000 simulations were performed with mean settings close to the best setting of the inputs with tolerances as indicated in Figure 7. The target and tolerance used for blend speed, force, and exhaust temperature was the same as currently used in manufacturing since 60 minute mean dissolution was robust to this level of variation in these three inputs. The target and tolerance of mill time, screen size, blend time, coating viscosity, and spray rate were adjusted per the knowledge gained via multiple regression to ensure acceptable distribution of 60 minute mean dissolution relative to the lower specification limit of 70%. The simulation confirms expectations of consistent product performance with a predicted Cpk of 1.4 (equivalent to a sigma level of 5.6). The proposed solution is wholly within the bounds of the currently validated process.

Case Study 2: New Production Facility with Many Measured Inputs

This case study concerns a relatively new manufacturing facility that has been producing commercial batches of an inhaler product for a couple of years. Extensive inline measurement systems were designed into the facility, resulting in a data-rich environment of 520 measured inputs. The first 30 inputs are processing parameters of the milling, blending, and packaging steps; variables 31-100 are properties of ingredient 1; variables 101 to 170 are properties of ingredient 2; and the remaining variables are properties of ingredient 3 (active ingredient).

The key performance metric is a percentage of a given dose reaching stage 3-4 of a cascade impactor, which must be between 15% and 25%. Since the start of commercial production, 240 batches have been manufactured, approximately

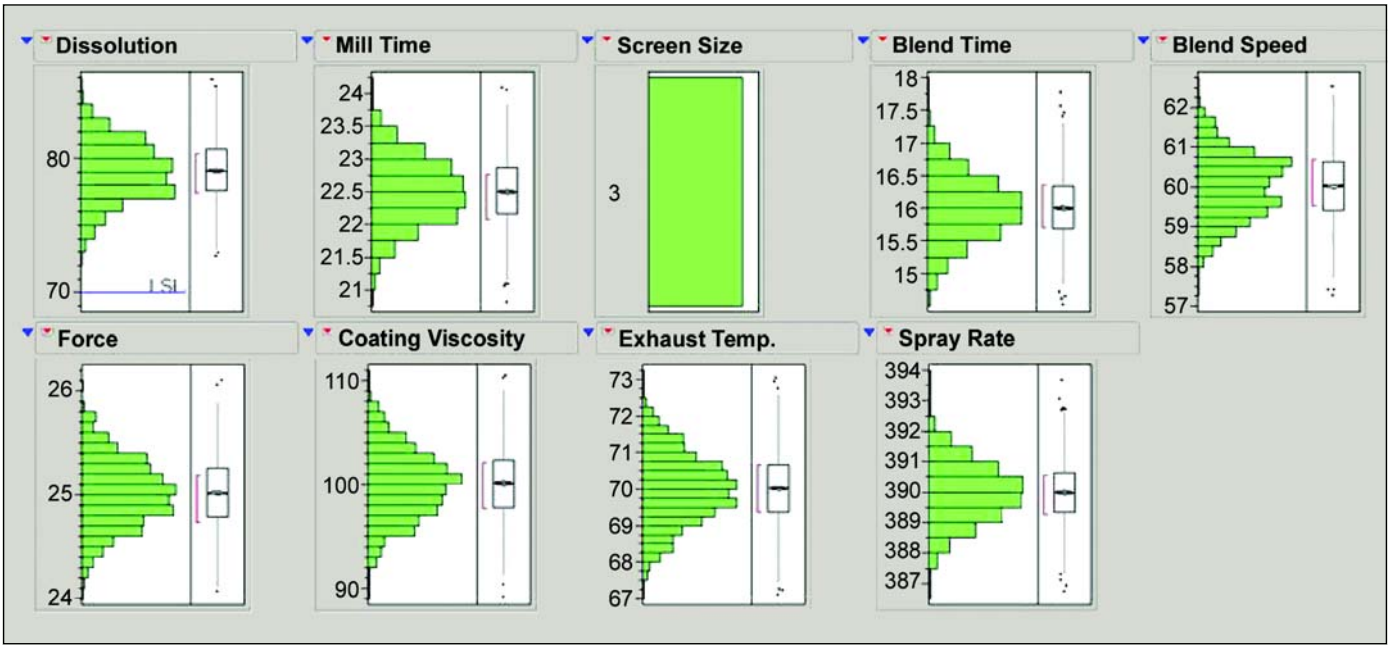

Figure 7. Simulation study to investigate process robustness.

14% of which have failed to meet the performance requirement of the cascade impactor test. QA investigations into these batch failures have been unable to identify any obvious assignable causes.

In this data-rich scenario, the manufacturing team commissioned with investigating the process and dramatically improving sigma capability adopted a variety of effectiveness and efficiency modeling techniques. Figure 8 illustrates the results of recursive partitioning to help determine the inputs most strongly associated with the percentage of a dose reaching stage 3-4.

The decision tree shows how the distribution of % at stage

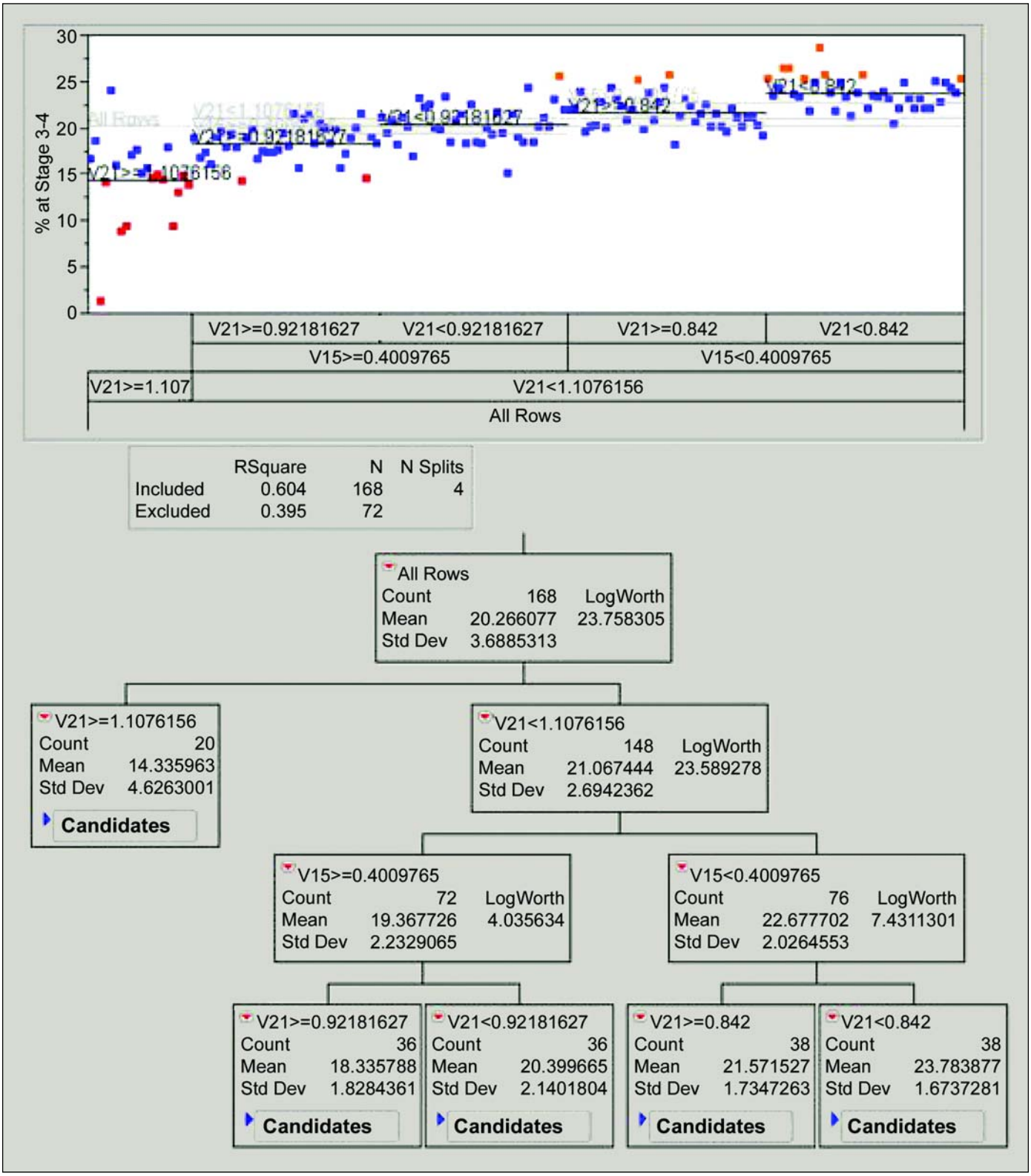

Figure 8. Recursive partitioning decision tree identifies inputs most strongly associated with variation in % at stage 3-4.

3-4 changes according to splits derived from the levels of two process variables - V21 and V15. The left-hand branch of the decision tree shows that when V21 >= 1.1, the distribution of % at stage 3-4 has a mean of 14.3 and standard deviation of 4.6. The graph at the top of Figure 8 shows a greater proportion of rejected batches (red points) than passing batches (blue points) in this subgroup. The middle branch of the decision tree defined by V15 >= 0.4 and V21 < 0.9 yields a distribution of % at stage 3-4 with a mean of 20.4 and standard deviation of 2.1. Just one of the 36 batches processed this way results in a rejected lot.

The decision tree model was built by excluding 72 (30%) of the 240 batches, these excluded batches were used to validate the decision tree. The partitions defined by the levels of V15 and V21 in the decision tree explain 40% of the variation in % at stage 3-4 of the 72 batches that were excluded from building the model. Nine of the 72 model validation batches met the criterion of V15 >= 0.4 and V21 < 0.9 with all nine of these batches passing the compliance test. Thus, V15 and V21 are verified as being strongly associated with some of the excessive variation in % at stage 3-4 and are potential drivers of resultant batch failures.

Recursive partitioning is a good visualization tool to help all consumers of process data see and communicate understanding about the dominant drivers of product variation; however, the method requires a large number of batches to reliably build decision trees with a greater number of branches (possibly utilizing other input variables to define the additional branches). Nonetheless, it is a great tool for aiding understanding and communication about the potentially dominant drivers of a problem.

To determine if there are additional input variables that may enable us to further reduce variation in % at stage 3-4, a PLS analysis was performed using cross validation. The coefficients from the resulting model are illustrated in Figure 9. The area of each rectangle of the Tree Map is in proportion to the size of the PLS model coefficient for the corresponding input variable. Blue rectangles stand for negative coefficients and red for positive coefficients. Two dominant factors, in addition to V15 and V21, are identified as V9 and V18. The sign of these four model coefficients tell us that increasing the values of V9 and V18, and reducing the values of V15 and V21 will result in higher values of % at stage 3-4.

Table A compares the observed vs. predicted result of the batch acceptance test, based on 72 batches that were excluded from the model fitting. The PLS model predicts batch performance of the 72 batches excluded from the model with three misclassifications. However, as a prediction model of batch performance, the center branch of the decision tree in Figure 8 works just as well as a predictive model of batch failures in helping to reduce future occurrences of batch failures. With nine of the 72 model validation batches meeting the criterion of V15 >= 0.4 and V21 < 0.9 and with all nine batches passing the compliance test, the recursive partitioning decision tree appears to be a simpler and sufficient predictor of batch failures.

To investigate in greater detail the effects of the input

Figure 9. Tree map of PLS model coefficients.

factors V9, V15, V18, and V21 on % at stage 3-4, a D-optimal DOE with a full quadratic model was performed. The resulting DOE worksheet is presented in Table B.

Summary results for the DOE analysis are presented in Figure 10, which shows significant linear effects of all four factors and a significant interaction between V15 and V21 at the 5% level. The direction of the relationship between % at stage 3-4 and each of the four inputs is in agreement with the sign of the coefficients from the PLS model. Multiple optimum solutions that get the mean of % at stage 3-4 on target exist, one of which is to operate close to V9=0.25, V15 = 0.4, V18 = 0.9, and V21 = 1.2. To explore the viability of this solution, the regression model was used to simulate the propagation of variation from the four inputs when set at the above values with a tolerance as indicated in the bottom part of Figure 10, and random batch to batch variability defined by a standard deviation of 0.5 (more than twice the standard deviation of the residuals in the fitted regression model). This predicts a distribution of % at stage 3-4 wholly within the required range (Figure 10) with a predicted sigma quality level of 4.8. In practice, before accepting this solution, it would be necessary to validate the model and predicted behavior with model validation batches performed at or within the proposed tolerance of the four process settings.

Summary

The blend of three key technology enablers – measurement, data integration, and data analysis systems – required to improve product quality through increased process understanding, depends upon the circumstances of the particular manufacturing facility.

Mature manufacturing facilities are unlikely to have extensive inline, at-line, or online measurement systems in place for tracking process inputs. Thus, the adoption of effectiveness modeling is a way to improve product quality through increased process understanding. The focus is to identify the critical few inputs and to develop empirical models of the effects of these on product quality that approximate the causal relationships between inputs and product quality. These models are then deployed to reduce variation

in final product quality and achieve performance requirements through improved process and material specifications and controls. A subset of some visual and statistical effectiveness modeling techniques in the context of mature manufacturing was illustrated in Case Study 1.

Manufacturing facilities for newer products are more likely to have extensive inline, at-line, or online measurement systems for tracking process inputs. The path to improved product quality through increased process understanding is a combination of efficiency and effectiveness modeling. Efficiency modeling methods are deployed to predict product performance, define some temporary controls to reduce batch failures, while effectiveness studies are conducted. The efficiency models also help identify and prioritize the inputs to be investigated in detail through effectiveness modeling techniques. The combined use of efficiency and effectiveness models may help reduce the number of process inputs that are routinely measured to the critical few if this helps accelerate cycle time or reduce other risks. A subset of some efficiency and effectiveness techniques in the context of a data-rich measurement environment was illustrated in Case Study 2.

Quality by Design is effectiveness modeling applied in process R&D, where it is possible to explore wider ranges of process inputs. The goal is to design a robust process that identifies the critical few inputs and tolerances for each key input that must be maintained in manufacturing. From a measurement systems viewpoint, the goal is to define the few inputs that must be measured or controlled in manufacturing and to achieve this knowledge through a high level of process understanding.

Simplifying data analysis and reporting is critical if more people in process development and manufacturing are to interpret and communicate around models that enhance process understanding. This article has introduced visual modeling methods that are easy to deploy for mainstream

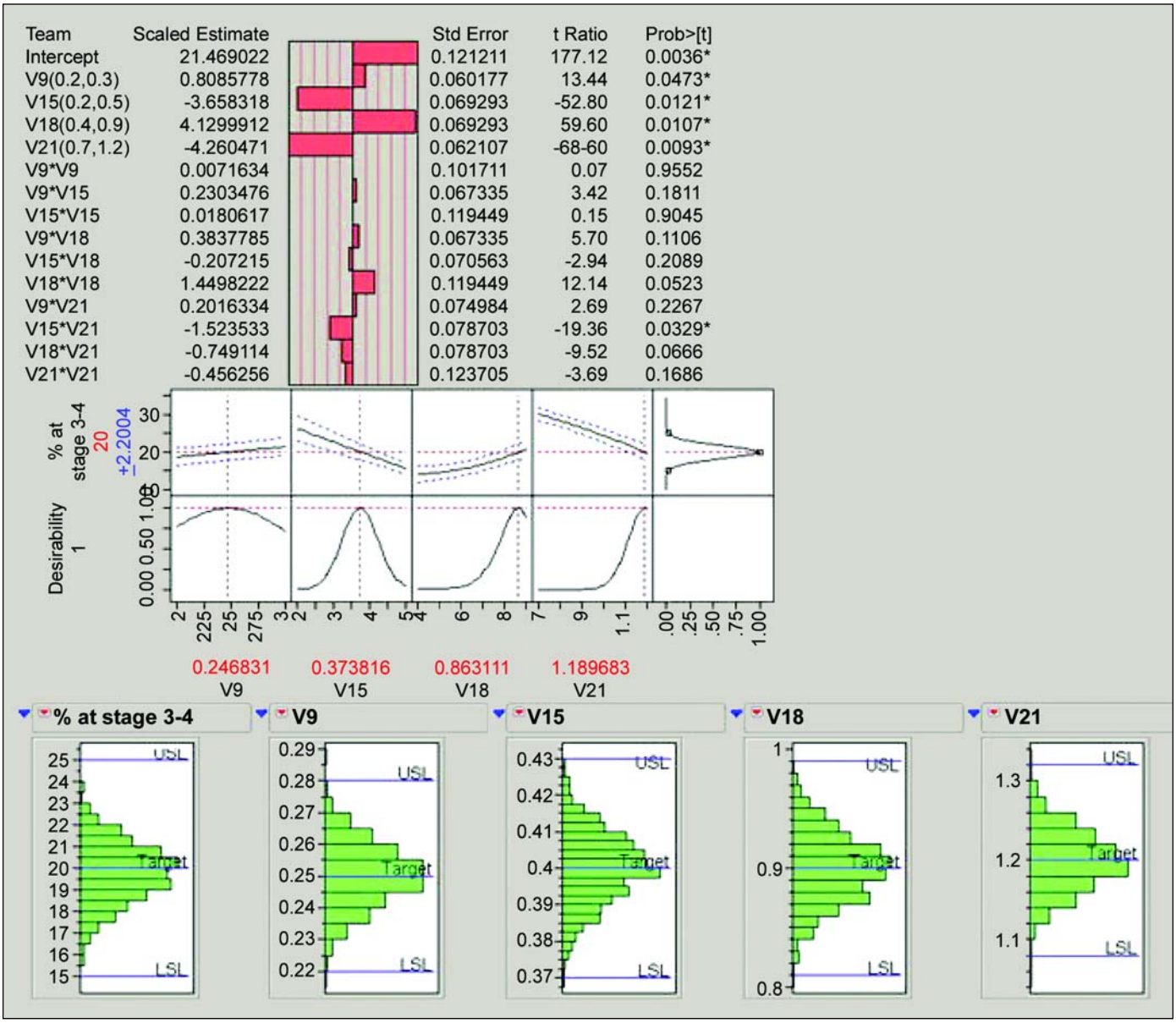

Figure 10. DOE summary analysis.

users and help them apply the principles of statistical thinking, particularly those of modeling variation in process outputs and identifying the key drivers of process variation.

About the Author

Dr. Malcolm Moore has worked with clients to integrate statistical methods and software into R&D, quality improvement, defect reduction, cycle time reduction, and corporate Six Sigma consulting activities for a variety of industries, including pharmaceutical and semiconductors. Prior to joining SAS UK, he worked at Light Pharma, BBN, Astra Zeneca, and lectured in medical statistics at Newcastle University. He is an expert in design of experiments, and received his PhD in design of non-linear experiments at London University. He can be contacted by e-mail at: malcolm.moore@jmp.com.

SAS Institute, Whittington House, Henley Road, Medmenham, Marlow SL7 2EB, United Kingdom.