International Journal of Industrial Engineering, 16(3), 206-213, 2009.

Troubleshooting Analyses of Production Data

Duane Steffey[1], Andrew Ostarello[1], Jason Clevenger[2], and Marta Villarraga[3]

[1]Statistical and Data Sciences
Exponent, Inc.
149 Commonwealth Drive
Menlo Park, CA 94025

[2]Mechanical Engineering and Materials Science
Exponent, Inc.
9 Strathmore Road
Natick, MA 01760

[3]Biomechanics
Exponent, Inc.
3401 Market Street, Suite 300
Philadelphia, PA 19104

Corresponding author's e-mail: {Duane Steffey, dsteffey@exponent.com}

Production data present substantial challenges for statistical analysis, but they may hold information of great value in resolving persistent manufacturing deficiencies. Because these data are not generated in controlled experiments, frequently key factors are confounded or nested, observations are unbalanced across factor levels, and a substantial number of values may be missing. On the other hand, these data represent the most realistic characterization of the manufacturing process on a production scale. This paper aims to describe and illustrate, with a case study adapted from an industrial project, how production data can be analyzed to provide clues to the sources of quality problems. Although such observational studies cannot definitively prove the existence of a cause-and-effect mechanism, results of troubleshooting analyses may suggest potential targets for corrective actions, as well as off-line experiments or further measurements and analyses to confirm the root cause of the manufacturing problem. These investigations may be regarded as part of the analysis step in a six sigma DMAIC methodology. They are undertaken to improve the process mean with respect to specification limits and to control process variation, and their effectiveness can be measured in subsequent capability studies.

Significance: This paper illustrates how observational production data may be analyzed statistically to yield clues to the root causes of quality problems in manufacturing, with the ultimate goal of improving process capability in a manner consistent with six sigma DMAIC methodology.

Keywords: Hierarchical linear models, nesting, unbalanced data, variance components, regression trees.

(Received: 2 March 2009; Accepted in revised form: 21 May 2009)

1. INTRODUCTION

Traditional applications of statistics in industrial manufacturing comprise three major areas: experimental design, acceptance sampling, and statistical process control. Designed experiments provide a scientific basis for making cause-and-effect judgments by systematically varying process inputs and subsequently observing the effects on outputs. Fractional factorial designs efficiently screen potential factors affecting the process output, and response surface methods provide a formal mechanism to optimize process inputs. Acceptance sampling plans may be used at the beginning of the manufacturing process to assure the quality of components from a supplier or at the end of production as a final check before shipment to a customer. Tools for statistical process control enable real-time monitoring of intermediate or end-of-line outcomes to provide prompt indication of quality problems as they occur. However, such tools focus exclusively on output measures, not process inputs, and provide no direct guidance for corrective action.

This paper explores the application of statistical methods—specifically, regression trees and hierarchical models—to the analysis of contemporaneous data on production inputs and outputs. Production data present substantial challenges for statistical analysis, but they may hold information of great value in resolving persistent manufacturing deficiencies. Because these data are not generated in controlled experiments, frequently key factors are confounded or nested, observations are unbalanced across factor levels, and a substantial number of values may be missing. On the other hand, these data represent the most realistic characterization of the manufacturing process on a production scale. Identifying

cause-and-effect relationships is a key task in the analysis step of a six sigma DMAIC (Define, Measure, Analyze, Improve, Control) program, and a well conceived and executed troubleshooting analysis can therefore be crucial to improving an existing process.

A stylized case study, adapted from an actual industrial project, is used to describe and illustrate how production data can be analyzed to provide clues to the sources of quality problems. Recent batches of product had experienced unacceptable outcomes in an end-of-line test administered prior to release. Batch production data from multiple sources were integrated and analyzed to assess the relative importance of factors potentially associated with the unacceptable results. In particular, the investigation focused on the effects of using different lots of certain raw materials in the production process.

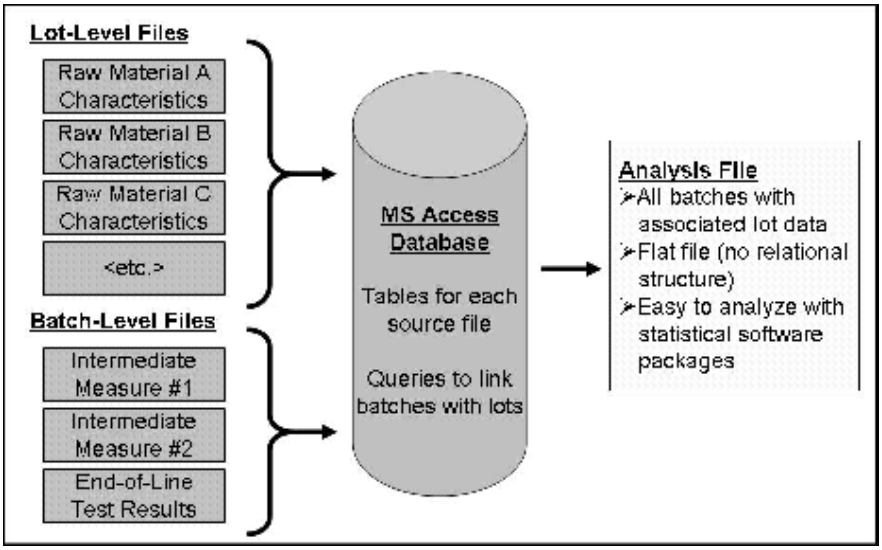


Figure 1. Schematic of Data Processing

2. DATA INTEGRATION AND DESCRIPTION

If a problem is identified in the end-of-the-line product, a troubleshooting study may be conducted to determine whether changes in certain inputs (e.g., raw materials or process characteristics) are associated with the output variables. Such a study requires the integration of data on production inputs, intermediate process measures, and end-of-line outcomes.

2.1 Overview of Data Processing

Troubleshooting studies typically require integration and processing of data files to be put into a form that can be analyzed using statistical software. In this case study, the data files originally received were of two types: batch-level files containing data on process outcomes, and lot-level files identifying the raw material lots used in producing individual batches, as well as measures of physical characteristics of the raw material lots. Data from these files were cleaned to address coding issues and then imported into a Microsoft Access® relational database. The inclusion of a field common to all files, containing unique batch identification numbers (IDs), enabled the creation of "flat" files with no relational structure that were suitable for statistical analysis. Figure 1 outlines this process.

2.2 Description of End-of-Line Outcome

In our investigation, the outcome of interest was the mean of the tested sample units from individual batches. The range of batch mean outcomes in the production dataset was rather large, as shown in Figure 2. The dark grey bars indicate batches for which the mean outcome is below the lower specification limit of 38. Of the 1,000 batch means plotted in Figure 2, 167 or 17 percent are less than 38. With a grand mean of 41.1 and standard deviation of 3.03, the process capability index during the production period was only 0.34, far below the minimum value of 1.25 recommended by Montgomery (2004) for existing processes. For a Six Sigma quality process the capability index is at least 2.00. A troubleshooting analysis seeks to identify production inputs and intermediate process measures that are most strongly associated with batch outcomes. If these associations are subsequently verified as cause-and-effect relationships and controlled, then the process mean will improve with respect to specification limits, process variation will be reduced, and the effectiveness of these analyses will be confirmed in subsequent capability studies

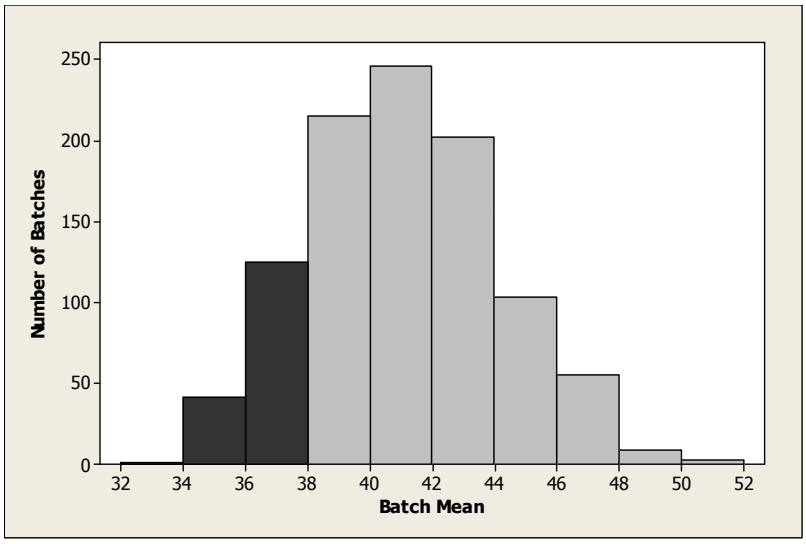


Figure 2. Distribution of Batch Means

3. STATISTICAL METHODS

The application of complementary methods of analysis provides a check of the strength and robustness of conclusions. Two methods were applied in the case study: linear modeling and regression trees.

3.1 Linear Models

Linear models (e.g., Neter, et al., 1996) are versatile statistical tools for studying the relationship between a random outcome (or response) of interest and one or more factors that potentially influence the outcome. The linear models considered in the analysis have the general structure:

$$Y_{ijkl\ldots} = \mu + \alpha_j + \beta_{k(j)} + \gamma_{l(jk)} + \ldots + r_{ijkl\ldots} . \quad (1)$$

In this expression, the batch outcome $Y_{ijkl\ldots}$ denotes the mean response value observed for a sample of units selected for testing from the *i*-th individual batch during production. The grand average μ represents the average value that would be obtained if test results were available for all units from all batches. The subsequent terms constitute adjustments to the average value to account for systematic effects on the outcome that arise from using particular lots of particular raw materials during batch production. For example, α_j represents the effect attributable to the use of the *j*-th lot of Raw Material A, $\beta_{k(j)}$ represents the effect attributable to the use of the *k*-th lot of Raw Material B, and so forth. Finally, the random variation term $r_{ijkl\ldots}$ represents the inherent sampling variability associated with measuring a sample rather than all units in a batch. The error terms for the *n* batches $r_{ijkl\ldots}$, $i = 1,2,\ldots,n$, are assumed to be independent and identically distributed normal random variables with mean zero and variance σ^2.

The linear models considered in this investigation possess two other characteristics of note. First, the raw material lots are modeled as having random, rather than fixed, effects on the outcome, because the lots used in production may be regarded conceptually as being selected from the large number of manufactured lots of each raw material. For example, the effects of the *a* lots of Raw Material A, α_j, $j = 1,2,\ldots,a$, are modeled as independent and identically distributed normal random variables with mean zero and variance σ_α^2. Linear models that decompose the outcome into a sum of random effects are sometimes called variance component models. Essentially, the task is to identify which components of variance (*e.g.*, raw materials) account for most of the observed batch-to-batch variation in mean outcome.

Second, raw material effects are modeled as nested effects. Reflecting the nature of the production process, lots of raw materials that change more frequently during production are nested within lots of raw materials that change less frequently. For example, the parentheses in the subscript for the Raw Material B effect, $\beta_{k(j)}$, indicates that the effect of the *k*-th lot of Raw Material B is nested within the *j*-th lot of Raw Material A. The use of nested models is appropriate here, because all lots of one raw material are not used (or "crossed") with all lots of another raw material.

3.2 Regression Trees

A regression tree (Hastie, et al., 2001) is a computationally intensive method of classifying observations. Regression trees are constructed by repeated binary splits of subsets of the data, beginning with the complete dataset. Each binary split

creates descendent subsets that are "purer", or more homogeneous, than the parent subset. This is done by computationally finding a split that achieves the largest decrease in the average impurity of the descendent subsets.

The nodes created by each binary split are recursively split until one of the following three conditions becomes true:

1. All cases in the node are of the same observed class (*i.e.*, the impurity is equal to zero);
2. The node only contains observations that have identical measurements (*i.e.*, there is no way to split the Remaining observations; or,
3. The node is small, typically 1 to 5 observations.

Once a terminal point has been reached for every node, the tree is pruned upward. This procedure creates a sequence of smaller and smaller trees. The overall impurity of each of these trees can be measured, and the one with the smallest total impurity may be regarded as the "best" regression tree (Brieman, et al., 1984).

In this case study, the goal of regression tree modeling is to define subsets of batches with low mean outcome values. If these batches can be effectively isolated from the other "good" batches, then the characteristics used in defining the batch subsets (*e.g.*, use of particular lots of particular raw materials) would suggest which factors are likely to be most strongly influencing production outcomes.

4. ANALYSIS AND RESULTS

4.1 Linear Modeling Analysis

Table 1 lists the number of distinct vendor lots of selected raw materials used during the batch production period of interest.

Table 1. Number of Distinct Lots of Selected Raw Materials

Raw Material	Number of Lots
A	4
B	6
C	7
D	18
E	42

The list in Table 1 was compiled for 913 batches with complete data on vendor lots of these raw materials. A time series plot of the batch means in order of production (Figure 3) shows that quality was poorest at the beginning and end of the production period.

Results of fitting a nested variance components model to the available data are shown in Table 2. The nesting structure is indicated by the variables in parentheses: B(A) means that factor B is nested within factor A. Because a residual analysis of the initial model indicated significant first-order autocorrelation, the final model summarized in Table 2 also includes an autoregressive term, AR(1), using the preceding batch mean as a covariate. The residuals from the final model showed no evidence of any remaining serial correlation.

Table 2 indicates that, over the entire production period, Raw Material E is most strongly associated with the observed variation in batch mean outcome, followed by Raw Material D (as indicated by the table entries in boldface). Raw Material C is of marginal significance. Other raw materials are not statistically significant predictors, as evidenced by the high p-values (well above 0.05).

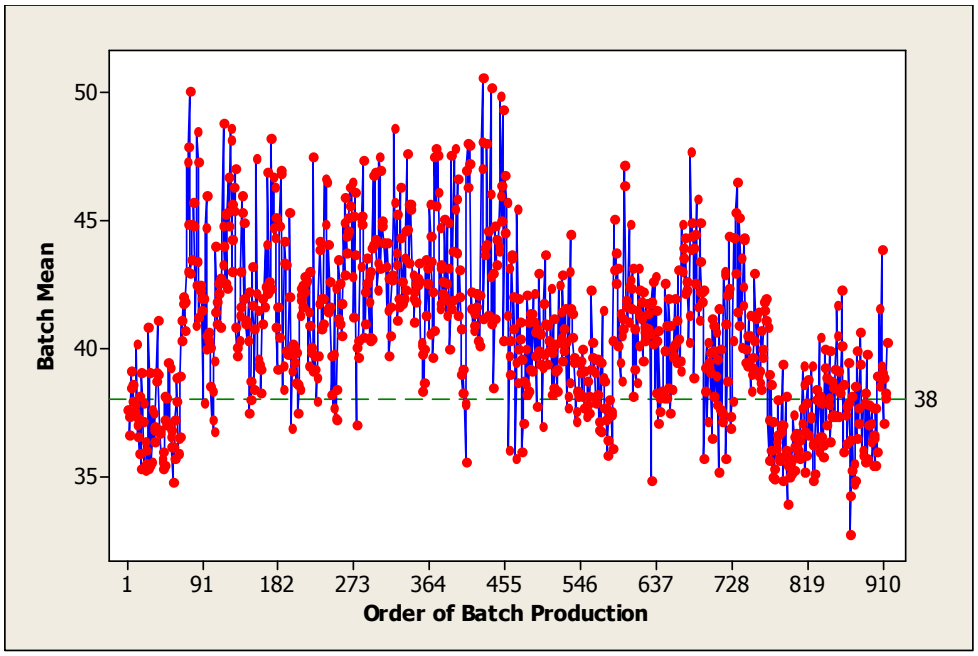


Figure 3. Time Series Plot of Batch Outcomes

Table 2. Analysis of Variance Results for Five-Term Nested Variance Components Model

Source	Degrees of Freedom	Sequential Sum of Squares	Adjusted Sum of Squares	Adjusted Mean Square	F-Statistic	P-Value
AR(1)	**1**	**4,547**	**52**	**52.03**	**13.08**	**< 0.001**
A	3	396	208	69.35	2.25	0.201
B(A)	5	245	213	42.50	2.14	0.150
C(A B)	**9**	**176**	**172**	**19.16**	**2.07**	**0.054**
D(A B C)	**19**	**186**	**286**	**15.04**	**1.99**	**0.018**
E(A B C D)	**47**	**479**	**479**	**10.20**	**2.56**	**< 0.001**
Error	827	3,289	3,289	3.98		
Total	911	9,318				

4.2 Regression Tree Analysis

The statistical software package R (http://www.r-project.org/) was used to construct a regression tree using production data for the 1,000 batches depicted in Figure 2. To satisfy computational constraints, lots of Raw Material E were consolidated into 21 aggregate lots. The resulting tree is shown in Figure 4. The rectangles represent decision nodes and the ovals represent terminal nodes. Observations enter at the top of the tree. At each decision node, observations that meet the split criteria travel down the left branch; observations that do not meet the criteria travel down the right branch. The observations stop when they reach a terminal node. The value in each terminal node is the predicted mean outcome for the observations that end up in that node.

Note that the decision nodes closer to the top of the tree utilize the variables that have the greatest ability to split the data into homogenous groups. In this case, the lot of Raw Material E used in the creation of the batch is the variable used in the first decision node; the regression tree algorithm determined that the lot of this raw material provides the best split of these data into homogenous groups. Decision nodes for Raw Materials D and C exist farther down the tree. Raw Materials A and B are not used in any decision nodes.

The regression tree indicates that the Raw Material E is most significant splitting variable for the batch production dataset, followed by Raw Materials D and C. Of particular interest is the fact that the terminal node with the lowest predicted mean outcome (37) depends only upon the lot of Raw Material E used to make the batch. The terminal nodes with the second lowest predicted mean outcomes (40) depend on combinations of lots from Raw Materials C, D, and E.

This result complements the results from the linear models approach, which also found that Raw Material E explained more variability in batch mean outcome than any other variable. Raw Materials C and D also appear as significant predictors in the regression tree model, but the tree prioritizes Raw Material E over all other inputs. The regression tree indicates that the Raw Material E lot variable is good for distinguishing the batches with low mean outcome from the rest of the batches. On the other hand, the linear models approach indicates that Raw Materials C and D also explain a significant portion of variability across the *entire range* of batch means. Both conclusions provide useful insight into the relationship between the raw materials and the batch mean outcome.

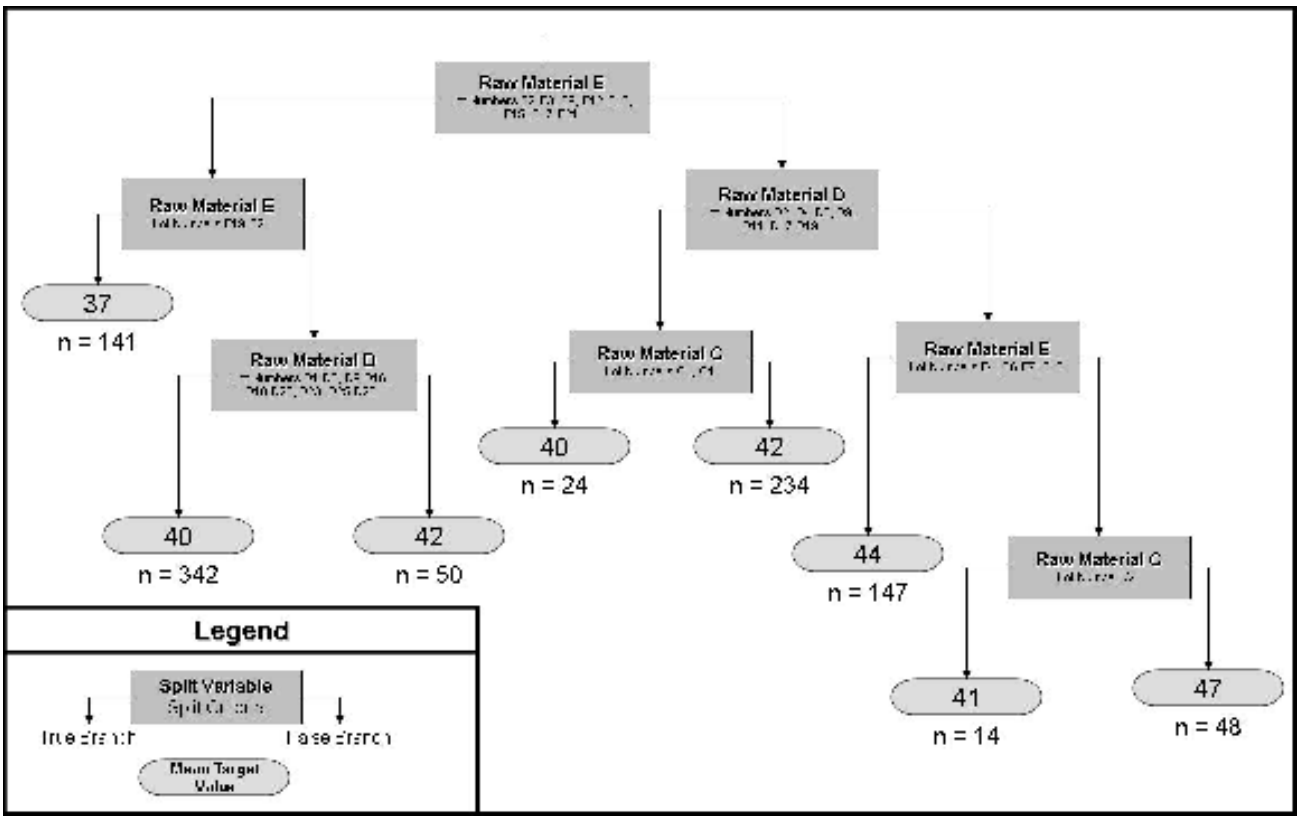


Figure 4. Regression Tree for Batch Production Data

5. DISCUSSION

The linear modeling analysis in the preceding section can be extended with information on physical characteristics of raw material lots. Interest then focuses on whether the variability in batch outcomes attributable to changing lots of a raw material can be explained by lot-level covariates. Hierarchical regression models (Raudenbush and Bryk, 2002) are well suited to addressing such questions and may provide further insight into the physical processes affecting product quality.

Suppose, for example, that W_j denotes a measured characteristic of the j-th lot of Raw Material D and X_{jk} denotes a measured characteristic of the k-th lot of Raw Material E (nested within the Raw Material D lot). Restricting attention to these two raw materials, a hierarchical regression model for the outcome of the i-th batch can be developed from a multilevel specification:

$$\begin{aligned} Y_{ijk} &= \varepsilon_{jk} + r_{ijk} \\ \varepsilon_{jk} &= \delta_{0j} + \delta_{1j} X_{jk} + u_{jk} \\ \delta_{0j} &= \theta_{00} + \theta_{01} W_j + v_{0j} \\ \delta_{1j} &= \theta_{10} \end{aligned} \tag{2}$$

where v_{0j}, u_{jk}, and r_{ijk} are normal random variables with mean zero and respective variances σ_δ^2, σ_ε^2, and σ^2. Combining the equations in (2) yields a hierarchical regression model:

$$Y_{ijk} = \theta_{00} + \theta_{01}W_j + v_{0j} + \theta_{10}X_{jk} + u_{jk} + r_{ijk} \quad (3)$$

Results of the regression tree analysis in the preceding section imply that a screen based on a Raw Material E characteristic could improve quality. When the 141 batches made with lots E19 through E21 (see Figure 4) are removed, only 7 percent, rather than 17 percent, of the remaining batch means fall below the lower specification limit, and the process capability index increases from 0.34 to 0.45. Although this revised index value is still quite low, supplemental investigations established that measurement of the end-of-line outcome was affected by factors unrelated to final product quality. Therefore, subsequent efforts in this case study focused on developing more robust metrics and methods of measuring end-of-line quality.

6. CONCLUSIONS

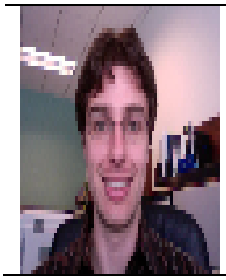
Although observational studies cannot definitively prove the existence of a cause-and-effect mechanism, results of troubleshooting analyses may suggest potential targets for corrective actions, as well as off-line experiments or further measurements and analyses to confirm the root cause of the manufacturing problem. Such analyses may be particularly indicated in cases when traditional experimentation is not feasible because of the amount of required test infrastructure, the perishability of raw materials, or prohibitive costs associated with a temporary stop or reduction in production. Identifying cause-and-effect relationships is a key task in the analysis step of a six sigma DMAIC program, and a well conceived and executed troubleshooting analysis can therefore be a crucial component of this effort. Ultimately, the aim is to improve the process mean with respect to specification limits and to control process variation. Program effectiveness can and should be measured in subsequent capability studies.

7. REFERENCES

1. Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.
2. Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer, New York.
3. Montogomery, D. (2004). *Introduction to Statistical Quality Control*. Wiley, New York.
4. Neter, J., Kutner, M.H., Nachtsheim, C.J. and Wasserman, W. (1996). *Applied Linear Statistical Models*, 4th edition. Irwin, Homewood, IL.
5. R: R Home Page. http://www.r-project.org/
6. Raudenbush, S.W. and Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd edition. Sage, Thousand Oaks, CA.

BIOGRAPHICAL SKETCH

Jason O. Clevenger is a Senior Managing Scientist in the Mechanics and Materials practice of Exponent Failure Analysis Associates. Dr. Clevenger earned his Ph.D. in Physical Chemistry from the Massachusetts Institute of Technology (MIT), and has worked extensively in the field of materials characterization and process engineering for semiconductor, medical device, and pharmaceutical applications.



Andrew Ostarello is the Lead Research Statistician at Scientific Learning Corporation. Prior to that, he was a Scientist in the Statistical and Data Sciences practice at Exponent, Inc. Mr. Ostarello is an experienced statistician and database programmer. He holds an M.S. degree in Statistics from California State University, East Bay.

Duane L. Steffey is a Senior Managing Scientist and Director of the Statistical and Data Sciences practice at Exponent, Inc., an engineering and scientific consulting firm. He was formerly Professor of Statistics at San Diego State University and also held positions at the National Research Council and Westinghouse Electric Corporation. He earned a Ph.D. in Statistics from Carnegie Mellon University. Dr. Steffey specializes in the application of statistical methods in projects involving product development, manufacturing process control, regulatory and safety issues. His recent applied research has concerned the design and analysis of engineering and scientific experiments, sample surveys, and observational studies. Dr. Steffey has evaluated product performance using manufacturing data on process outcomes and field data on incidents and exposure.

Marta Villarraga is a Principal in Exponent's Biomechanics practice in Philadelphia, PA. Dr. Villarraga has a Doctoral Degree in Biomedical Engineering from Tulane University. She specializes in spine biomechanics and in failure analysis of medical devices. She has experience with orthopedic, spinal, reconstructive surgery, and diagnostic medical devices from product liability, intellectual property, regulatory compliance, and product development perspectives. Dr. Villarraga also has experience in evaluating quality control issues as applied to medical devices and pharmaceuticals, with an emphasis in contamination, manufacturing compliance, and finished device evaluations.