Quality Engineering, 27:477–487, 2015

ISSN: 0898-2112 print / 1532-4222 online
DOI: 10.1080/08982112.2015.1065322

Taylor & Francis
Taylor & Francis Group

A Practical Approach to Data Mining: I Have All These Data; Now What Should I Do?

Ronald D. Snee
Snee Associates, LLC, Newark, Delaware

ABSTRACT Information technology increases our ability to use data to develop and improve processes. Professionals are being asked to make sense out of large volumes of data. Today's literature provides little guidance on how to approach such problems. Addressing this void, this article places a keen focus on the pedigree of the data: process that generated the data, measurement, and data collection process including sampling schemes used. The importance of using subject matter knowledge and recognition of the sequential nature of problem solving is also emphasized. A guiding framework for the execution of data-rich projects is presented and illustrated with case studies.

KEYWORDS analytics, data mining, data pedigree, model validation, project strategy

SO, NOW WHAT DO I DO?

You are a professional with statistical knowledge and skill (statistician, quality engineer, etc.) You are approached by your boss or another of the "higher-ups" who tells you, "Our process is losing money. We have lots of data and statistical software. Find out why our process is losing money and fix the problem!" In another instance the boss might say to you, "Variation in quality of product X is too high. We have lots of data and statistical software. You have some statistical background. Find out what the source of the problem is."

Statisticians and other quality professionals have learned that it takes much more than a large data set and computer software to effectively solve problems. Unfortunately, much of this experience has not yet found its way into the literature. Many professionals are finding themselves being asked to make sense out of large volumes of data often with little guidance except that there must be some software that can help you. The question often arises, "I have all these data; now what should I do?" The answer does not seem to be in textbooks: no roadmaps, no guides, no tips and traps. This article attempts to address this void.

So where do you begin? What is your strategy of responding to this type of request? The popular literature tell us to use "analytics." So what exactly *is* analytics, you ask? There are many definitions. One useful definition is: "Extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions" (Davenport and Harris 2007). I add to this definition the following phrases: "solving problems by developing process and product understanding using analysis and

Address correspondence to Ronald D. Snee, Snee Associates, LLC, 10 Creek Crossing, Newark, DE 19711. E-mail: Ron@SneeAssociates.com

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lqen.

modeling of data," "turning data into actionable information you can use," and "learning from data."

If you refer to the popular media, you get the impression that all you need to solve problems using data is to get a large amount of data, torture the data using some form of one or more statistically-based algorithms, and the results will provide the answer. If only it was so easy—in reality it is not!

This article begins by introducing the "Building Blocks of Analytics," which begins with the problem to be solved and the strategy for solving the problem and concludes with assessing the need for more data to be collected. The importance of addressing the practical, graphical, and analytical aspects of data analysis and problem solving is also considered. This roadmap provides a framework for how to think about problem solving using data. This is followed by a discussion of some critical considerations including the use of subject matter knowledge, the need to assess and understand the pedigree of the data, the impact of multicollinearity on the estimation of the effects of the variables (Xs), the data collection process, and how knowledge of the structure of the data can help lead to a useful analysis.

Next, two case studies are introduced to illustrate the issues discussed at the beginning of the article, including assessing the data pedigree, evaluating the measurement system, developing a strategy and associated tactics for the project implementation and data analysis and multicollinearity. The article concludes with a discussion of model validation, and critical success factors.

Building Blocks of Analytics

I have found it helpful to refer to the "Building Blocks of Analytics" shown in Figure 1. Here we see that there is a lot more involved than Data + Analysis = Success. It is critical to think carefully about the problem, then collect data that answers key questions that are central to solving the problem and do exploratory analysis typically involving some variety of graphics and a variety of models—both qualitative and quantitative (Snee 2002).

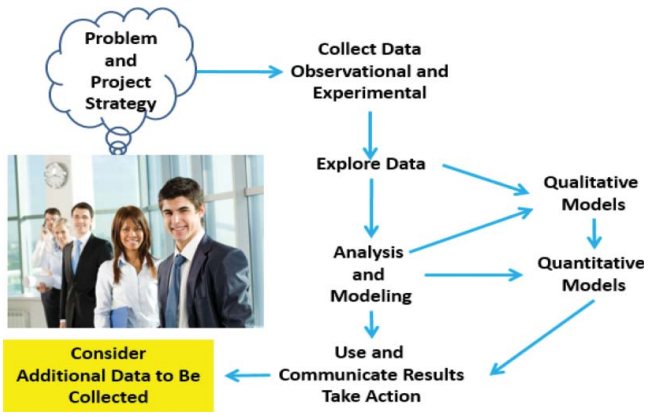

FIGURE 1 Building blocks of analytics.

Quantitative models are those models that we typically think of when fitting equations to data. We hypothesize a model form (an equation) and the variables (Xs) are known and measured. After the model has been fit to the data, we know the direction (+, 0, −) and magnitude of each effect, and the model (equation) can be used to calculate predicted values of the response (Y) for given values of the Xs. Such a model might appear as follows: Yield = 50 + 10(Temperature) −3(Pressure). We can calculate (predict) yield given values for temperature and pressure.

In the case of qualitative models, the predictor variables (Xs) are known, as well as the direction of the effects and sometimes the relative magnitude of the effects. However, the particular model form is unknown. But qualitative models can be very useful and are used extensively in many different areas. For example, it is widely known that weight gain is a positive function of caloric intake (calories up → weight up) and is negatively influenced by the amount of exercise one does (exercise up → weight down). In almost all cases, we do not have a specific model form (an equation) relating weight gain to caloric intake and exercise, but many people use a similar mental model using these two variables to control their own weight.

Another commonly used model is for residential home value as a function of three variables: location, location, and location! Home value assessments using this qualitative model would be based solely on location but without using any quantitative model. In reality, other factors such as square footage, number of rooms, and house age would play a role in the decision process. Quantitative models could of course incorporate these additional variables, arguably more effectively than the qualitative model.

The building blocks have been around for a while but perhaps not in this assembly. Of particular note is the "project strategy" at the beginning and the "collection of additional data" at the end. Projects need planning up front and are rarely complete with a single data set. Problem solving is a sequential process frequently requiring the collection and analysis of several data sets in combination and in series.

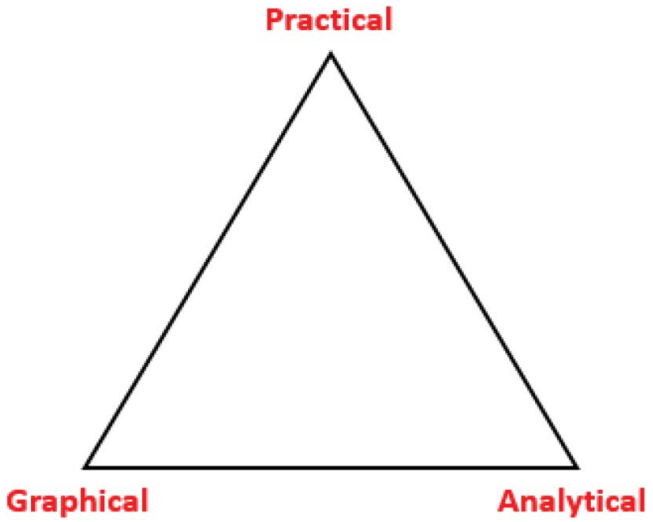

FIGURE 2 **Critical aspects of data-based problem solving.**

As we work our way through a data-based problem solving project and integrate the building blocks (Figure 1) into a solution, there are three aspects that we need to think about: the practical, graphical, and analytical, as shown in Figure 2 (Ross 1999). I consider this to be a "best practice" which I use, teach in my industrial workshops and university graduate courses and strongly recommend to others. The "Practical" refers to the context and subject matter knowledge that surrounds the issue or problem being investigated: the objectives of the study, the process involved, the pedigree of the data, various sources of variation to be studied, and associated subject matter knowledge. "Graphical" refers to the use of graphical techniques to explore, analyze, and communicate the data and results. "Analytical" refers to the statistical and other types of modeling procedures used. It is critical to note that one should always begin with the "Practical" considerations.

It is considered a best practice to visit each of these phases during an analytics study. These three phases interact with each other and are typically visited several times during a study. When the data are small in number the initial sequence is typically *Practical*, *Graphical*, and then *Analytical*. For larger data sets (which may be large in sample size, number of variables, or both), the sequence is often *Practical*, followed by *Analytical* and then *Graphical*. In this case, the analytical methods are used to reduce the dimension (number of variables) of the data set or the effective sample size or both. The analytical techniques used to reduce the dimensionality (complexity) of the data set including data smoothing, averaging and sampling. Graphical procedures are less effective when the number of points plotted gets large. The large number of points make it difficult to see trends and patterns unless the effects are small in number and very large.

It is also important to recall that the assembly of the building blocks (Figure 1) and the critical aspects of data-based problem solving (Figure 2) are constructed using the principles of Statistical Engineering (Hoerl and Snee 2010, 2012; *Quality Engineering* 2012). A variety of concepts, methods, and tools are linked and sequenced to form a methodology for executing data-based problem solving.

Assessing Data Pedigree

At the beginning of any study, there are some general topics that need to be considered. First, we must understand not only the objectives of the study, but also how the process that generated the data works and how the associated product or service is produced. This comment applies to service and manufacturing processes alike. Next, we must understand the pedigree of the data: how the data were collected, the measurement process and science, and the engineering and structure of the process. Assessing the pedigree of the data can help us avoid accepting poor quality data at face value and performing the wrong analysis of the data. Understanding the data pedigree results in a deeper knowledge of the data (Snee and Hoerl 2012; Snee, DeVeaux, and Hoerl 2014).

When assessing the pedigree of the data, two critical questions to keep in mind are the following ones: Do I really understand how the data were collected? Can I trace back and identify the origin of each data point?

Such an assessment is enhanced by the use of data graphics and visual analysis tools. The creation of process diagrams (schematics) are almost always helpful in assessing data pedigree and understanding the problem. The data pedigree should be assessed before, during, and after the analysis, as follows:

- **BEFORE**: Understand the process, sampling procedure, data collection, analysis preparation, and measurement system;
- **DURING**: Constantly check the data and results with the "does this make sense" test aided with extensive use of graphical displays;
- **AFTER**: Evaluate the results to make sure the results and conclusions make sense regarding what is known about the problem being investigated.

It is also important to check the assumptions of the data collection process.

- Are the data from an observational study collected without a well-structured and defined protocol or are the data from a statistically planned experiment or survey? Observational data can be of low, quality for a number of reasons as discussed by Hoerl and Snee (2012, p. 264–265).
- Is the randomization process used understood—is there any evidence of "split-plotting" or restricted randomization?
- Has the possibility of within-experiment nonhomogeneity been evaluated?
- Have equipment warm-up and wear effects been considered? Are they ignored or unknown?
- Was there a protocol for data collection including sampling, and was it followed?

When thinking about data quality per se, it is helpful to look for data that are clearly wrong (e.g., grossly atypical values, pregnant males), results and trends that do not make sense given the technical background of the problem, and missing information and data critical to a useful analysis and making sound conclusions (DeVeaux and Hand 2005).

Assessing Multicollinearity

Multicollinearity is a characteristic of the predictor variables that can be a big problem when the correlation between two or more predictor variables (Xs) in the model is high. When multicollinearity is strong, the prediction model can give erroneous results. For example, coefficients may be too large and/or have the wrong sign. The model may produce inaccurate predictions outside the region of the data (this is a general issue with extrapolation but aggravated by multicollinearity). Best results are obtained when the predictors (Xs) are independent (zero correlation). It is helpful to recall that designed experiments are constructed so that the correlations among the Xs are typically near zero or equal to zero.

A discussion of multicollinearity and how to deal with it can be found in Montgomery, Peck, and Vining (2001, Chapter 10) who point out that there are three ways to deal with multicollinearity, as follows:

- Collecting additional data to reduce the effects of multicollinearity;
- Revising the model form by deleting and combining variables producing the multicollinearity and standardizing variables when estimating curvilinear effects;
- Using ridge regression and other biased estimation procedures to estimate the model coefficients.

These approaches are all effective depending on the specific situation and are often used in combination with each other. Principal component analysis is also effective in detecting the source of the multicollinearity and identifying possible ways to reduce the multicollinearity (Montgomery, Peck, and Vining 2001, pp. 371–375).

Observations on Current Practice

Reflecting on current practice, one often sees data being collected without controls or careful administration of the data collection process. Such data are often referred to as "observational" data as opposed to data collected without the careful planning and administration associated with designed experiments and surveys. Observational data typically contain erroneous results, mistakes in data values, and may be missing variables that are critical to problem solution (Hoerl and Snee 2012, Chapter 6). Data mining as often practiced tends to assume that all data are good data and more data is better. The fact that data reside in electronic files says nothing about the quality of the data.

Knowing how the data were collected is also critical to performing the correct analysis of the data. Understanding the data structure enables us to more easily identify potential sources of variation. The model form that best fits the structure and situation becomes more apparent (e.g., crossed vs. nested factors, split plots; Montgomery 2013). Conducting data analysis without understanding the associated processes, including sampling and testing procedures, greatly increases the risk of erroneous results.

Regarding Data Pedigree, we should "trust but verify." While data quality is an issue with all sources of data, it is an often overlooked issue with large data sets constructed by combining datasets from different sources. Multiple sources of data require careful thought as to data pedigree and how to fit the data bases together to produce useful results.

Different data sets often come from different organizations. This may result in political issues as different organizations usually have different agendas and different objectives for the analysis and use of the data. For example, the vehicle emissions standard case study

discussed below involved data collected by the state, local, and Federal governments. The standards were of interest to the auto industry, gasoline suppliers, and suppliers of the gasoline companies. The result is at least six different agendas, each with a keen interest in the associated modelling and data analysis. Careful thought must also be given to the model form needed to answer the question. Different models can often get one to the same place, or to different places.

Subject matter knowledge is another important consideration that does not appear to be given proper attention in the current literature. Assessment of subject matter knowledge is part of practical data analysis consideration (see Figure 2 and associated discussion). Box and Hunter (1978, p. 291) point out that "Data have no meaning in themselves; they only have meaning within the context of a conceptual model of the phenomenon under study." Subject matter knowledge—the science and engineering underlying the process, product, and data—provides the context for the analytical work from beginning to end. This includes problem definition, data analysis, and modelling including selection of variables and appropriate scales (e.g., log, inverse, and square root) and selection of model form (e.g., linear, curvilinear, multiplicative) and interpretation and implementation of results.

Case Study: Assessment of Data Pedigree for Carbon Monoxide Vehicle Emissions

A study of ambient carbon monoxide (CO) levels and the development of vehicle emission standards reported by Pierrard, Snee, and Zelson (1974) focused in the initial stages on the pedigree of the available (CO) data. The study involved the assembly of data from eight different state and federal agencies involving traffic, meteorology, and vehicle emissions in addition to the CO data. The total data set involved over 150 sampling station-years for hourly data (8,760 hours/yr) resulting in over one million records. The goal of the study was to assess the vehicle emission standard needed to meet the air quality standard. Part of the strategy used was to analyze the ambient CO data.

The ambient air quality standard for CO is 9 ppm (8-hour average) not to be exceeded more than once per year is a human health-based standard. In practice the second highest 8-hour average CO during a year is used to assess the air quality in the vicinity of the sampler. This 2nd highest value is a highly variable statistic. For example, the 2nd highest 8-hour average for the Denver, Colorado, carbon monoxide sampling station in 1971 was 35 ppm with a maximum 8-hour average of 39 ppm, well above the standard.

Hourly data used to compute the 2nd highest value were evaluated. A plot of the hourly CO values for the period in questions showed 10 consecutive hourly readings of 39 ppm, with 4 out of the next 6 hourly readings at 39 ppm and the remaining 2 readings at 36 ppm (Figure 3). This small amount of variation over a 16-hour period is not typical of variation in hourly CO readings and does not represent an accurate characterization of the air quality in the area of the air sampler (Snee and Pierrard 1977). It is highly probable that these results are due to equipment malfunction. A similar problem was found in the CO data from Cincinnati, Ohio, in 1968 (Figure 4).

The investigation then centered on finding a measure of air quality that was less affected by atypical values. It was known that CO data over a year follows a lognormal distribution. The data in hand were found to be lognormally distributed and thus the 2nd highest value could be predicted using the log-normal model. Furthermore, the annual average was found to be a good predictor of the percentage of time that the standard was exceeded. It was a much better estimate than the 2nd highest value, since it was less affected by variations in CO measurement systems, meteorological conditions, and traffic volumes.

Empirical evidence from the data collected at six major cities over a 10 year period and at several sampling stations in New Jersey validated the relationships found. These findings produced the recommendation to use the annual average to make predictions and calculate emission reductions required for areas to meet the CO air quality standard which was designed to protect human health. The 2nd highest value would continue to be used to compare air quality levels at a given site to the Air Quality Standard as the standard is a Federal regulation not easily changed and is still in use today.

Assess Measurement Process

Particular attention should be paid to the measurement process. In my experience, as much as 50 percent of the measurement systems I have encountered have been in need of improvement. Measurement system analysis is a critical step in Lean Six Sigma projects (Snee

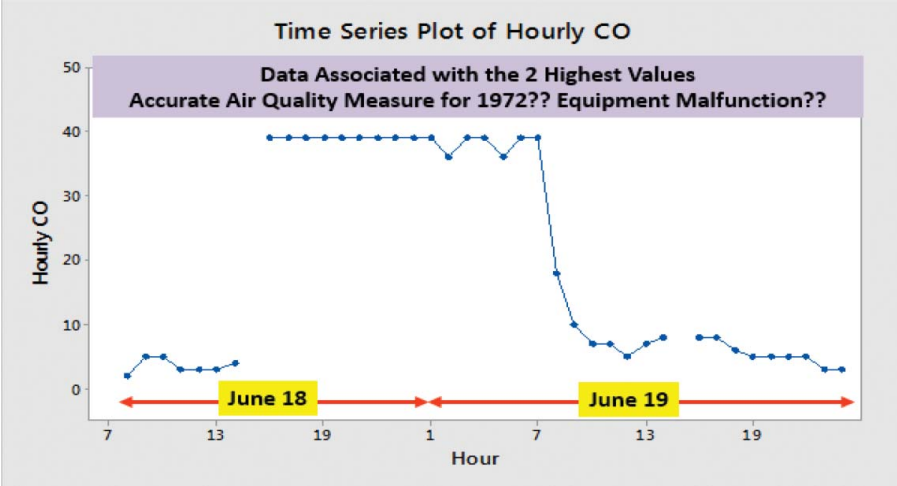

FIGURE 3 Time plot of hourly CO readings in Denver, June 18–19, 1971.

and Hoerl 2003). Common sources of measurement issues include: operator/analyst errors, equipment failures, and poor sampling methods and protocols. Operator technique differences are a common occurrence. Operator fatigue can result in using shortcuts resulting in data that are recorded incorrectly, for example, transposed digits and test randomization not used. Measurement instrument failures such as the CO measurement issues discussed above are a source of measurement problems. Sampling protocols not available, or not followed, and poor sampling techniques are often encountered.

The following example shows what can happen. Poor product quality was being experienced: the defect level was too high. An improvement project was chartered using the Six Sigma methodology (Snee and Hoerl 2003). During the Measure phase of the project, it was discovered that the measurement instrument had not been calibrated for two years. After calibration the product problems completely disappeared (zero defects) saving $157,000 per year in scrap. Case closed.

Case Study: Creating and Executing a Project Plan for Pharmaceutical Contamination Issue

This case study illustrates how a problem and its associated large data set can be approached using the strategy described in this article.

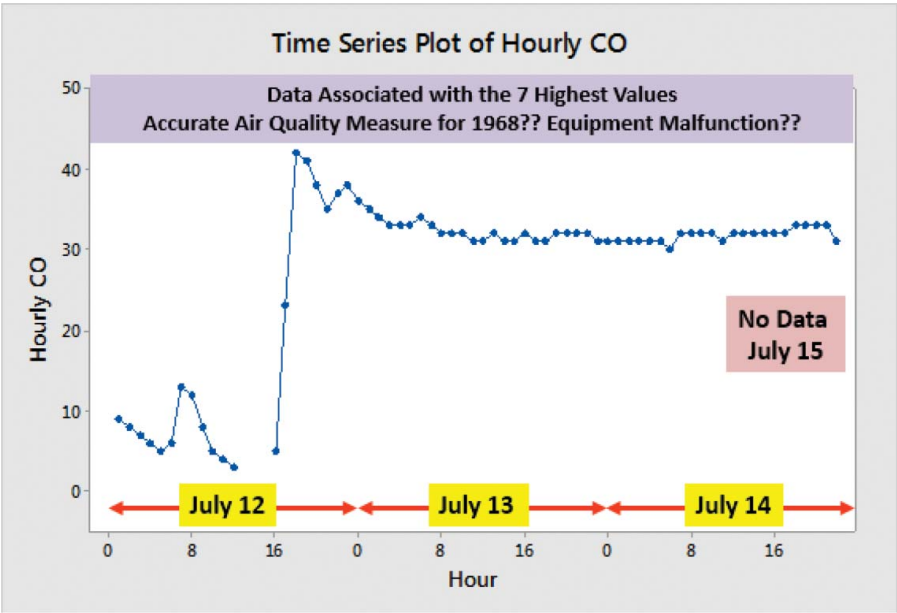

FIGURE 4 Time plot of hourly CO readings in Cincinnati, July 12–14, 1968.

Problem: A blockbuster drug is exhibiting high impurity levels which is significantly affecting the corporate bottom-line.

Problem Structure: Product impurity had an upper specification limit of 0.2 percent. The goal is to find the variables causing the high impurity levels and to put fixes in place so that the problem does not happen again.

Problem Context: Product is produced at five sites around the world. Data on approximately 50 variables (Xs) are available for 146 batches produced over the last 4 years. The data are considered to be "observational" as the data were measurements made on the product and process as the process was operated over the four periods. Manufacturing, R&D, Quality Assurance, Regulatory Affairs, and Quality Control have a keen interest in solving the problem.

The overall project strategy at this point (strategies often evolve over time) is to analyze the available data from all five sites to identify causes of the problem. This involves all sites in the problem solution. If the process changes associated with the causes identified do not fix the problem, then additional studies will need to be done, perhaps using designed experiments.

The following *Analysis Strategy* was developed given the problem definition, structure, and context detailed above:

- Collect the data from the five locations;
- Integrate into a single database;
- Analyze the retrospective data to identify the variables affecting the impurity;
- Use the results to identify possible fixes;
- Implement proposed fixes and test the effectiveness of the fixes;
- Determine what additional studies, if any, are needed.

The *Analysis Tactics* included using regression modeling and statistical process control techniques. Control charting was used to evaluate the stability of the process over time and process capability indices (Pp and Ppk) were used to compare impurity levels to the specification.

Regression analysis was used to identify variables (Xs) having a significant effect on product impurity. In general, the regression modeling approach was to start with a large number of variables in the model and then identify the important variables to keep in the model by evaluating the Student's *t*-statistics and associated *p*-values. Principal Component Analysis and regression coefficient variance inflation factors (VIF) were used to identify sources of multicollinearity among the predictor variables. A variety of models were evaluated, and a residuals analysis was conducted for each of the candidate models.

The regression, control charting, and process capability analyses lead to the following findings regarding product impurity:

- Process was not capable of meeting specification requirements and exhibited instances of instability;
- Using regression analysis statistically significant effects were found for:
- Time between Curing and Coating;
- Duration of Coating Process;
- Active Pharmaceutical Ingredient (API) Lot;
- Product Strength;
- Interaction between Manufacturing Location and API Lot;
- Lot of Excipient (inactive substances such as lubricants, fillers and diluents used in making a pharmaceutical tablet);
- Storage Condition;
- The regression model had an Adjusted R-Square = 84 percent indicating that the model provided a reasonable fit to the data. In my experience an Adjusted R-Square of > 80 percent is considered to be good for observational data assembled from historical production records.

When these effects were taken into account the projected process capability (long-term) was projected to be Ppk = 1.30, which is close to the goal of Ppk = 1.33 (process sigma = 4.0) and well above the baseline of Ppk = 0.50. This result indicates that, if the variation due to the significant variables can be controlled, the process will be able to meet specifications.

However, more studies are needed. The results discussed above are based on data from four plants, not five. The fifth plant did not measure "time between coating and curing." Since this variable was found to have a significant effect, the data from the 5th plant could not be included. Another study should be conducted to assess the importance of this variable across all five plants.

There were also significant effects due to API lots and excipient lots. Variables that characterized the differences between the API lots were identified and found to have significant effects. However, no such set of variables could be found for the excipient lots. The variables available were found to be so highly correlated (multicollinearity) that no useful predictors could be found. Further data collection including the identification of additional variables to characterize excipient lot quality will be needed to adequately understand the effect of the differences among the different excipient lots.

Analysis of the regression model residuals uncovered the fact that the variability of the impurity level was significantly larger in two of the plants compared to the other two plants. This indicates that additional studies will be needed to understand the source of the higher variation in the two plants. Extensive plotting of the residuals versus variables (Xs) in and out of the model did not identify any additional useful predictor variables or different model forms.

These findings illustrate the fact that it is rare that a single data set solves the problem. Problem solving is sequential with the learnings from one data set being used to improve the process and suggest other studies and data that needs to be collected to further improve the process.

Model Validation: Is the Model Useful?

One cannot leave the discussion of analytics without addressing the subject of model validation (Snee 1977; Montgomery, Peck, and Vining 2001, Chapter 15). Confirmation experiments are routinely run to test the predictability of models developed from designed experiments. It seems reasonable that we should do the same for models developed from observational data obtained without a designed experiment.

The analysis of the carbon monoxide (CO) data discussed earlier provides a good example of model validation. The CAMP CO data were found to be described on an annual basis by a lognormal distribution with a coefficient of variation of approximately 50 percent. Analysis of a larger data set collected in the State of New Jersey produced the same results: a lognormal model with a coefficient of variation of approximately 50 percent (Snee and Pierrard 1977) thus validating the model developed for the the CAMP CO data.

Models can be assessed in a number of different ways (Snee 1977) including the following ones:

- Comparison of model coefficients and predictions with scientific and engineering theory;
- Collection of new data to check model predictions;
- Comparison of results with theoretical models and simulated data;
- Reserving a portion of the existing data to measure the prediction accuracy of the model.

Practical considerations should also enter into model validation assessments. Box and Norman (1987, p. 424) pointed out that "all models are wrong, but some are useful." Juran and Godfrey (1999, Chapter 21, page 21.4) defined "quality as fitness for use." Combining the guidance of Box and Norman and Juran and Godfrey suggest that *a good model is one that is useful for its intended use.*

Data splitting is a widely used approach for validating a model. The data (N observations) are split into two sets called the estimation set (N1) and the prediction set (N2) with $N = N1 + N2$. The estimation set is used to build the model, and the prediction set is used to assess how well the model does in making predictions. It provides an independent set of data. Data-splitting is also referred to as cross-validation (Montgomery, Peck, and Vining 2001).

Having an independent set of data to measure the prediction accuracy of the model simulates how the prediction accuracy of the model will be done in use; the assessment will be made on new data, data not used to build the model. Testing the model on an independent set of data provides a check on the whole model construction process; model formulation, statistical analysis, subject matter considerations, and statistical assumptions.

When the data are time ordered, it seems reasonable to pick a point in time and to divide the data into the estimation set and prediction set. For example, given four years of data, one might develop the model using data from years 1–3 and use the year 4 data to check the model prediction accuracy. Some prefer to randomly split the data into two sets assigning the data to the two sets at random. Another data splitting approach is to use the DUPLEX algorithm, which divides the data set into two sets which cover approximately the same

region and have the same statistical properties (Snee 1977). Additional details on the technical mechanics of data-splitting are discussed by Montgomery, Peck, and Vining (2001) and Snee (1977).

CASE STUDY: MODEL VALIDATION FOR YIELD PREDICTIONS USING DATA-SPLITTING

Yield data from a three step process will be used to illustrate the data-splitting approach. Fifty-one weeks of yield data were available. Six predictor variables were included in the linear-effects additive model: Moisture and temperature are measured in each of the three steps in the process. The model was developed from the first 30 weeks of data and then validated by using the resultant model to predict the next 21 weeks of data. Figure 5 shows a control chart for the prediction residuals in which the control limits are three-standard deviation limits based on the residual standard deviation for the model developed from the estimation data. These results show that the prediction residuals are stable and predictable, indicating that the model is predicting with the precision obtained with the estimation data.

Model prediction performance as assessed by the model prediction residuals being within the desired limits and showing no patterns or trends is arguably the best way to demonstrate model validity. But it is not the only way, and we need to be careful not to base our model validation assessment completely on the behavior of the prediction residuals. As noted previously, other model validation checks include comparison of model coefficients and predictions with scientific and engineering theory, comparison of results with theoretical models, and simulated data. Also, model validation should be a continuing process done throughout the time the model is in use.

CASE STUDY: MODEL VALIDATION FOR ELECTRICITY USAGE MODEL

Figure 6 shows the prediction residuals for a production plant electricity usage model. Note the shift in the residuals indicating the model is no longer adequate and needs to be updated. Such shifts in prediction accuracy are frequently due to new variables becoming important or variables currently in the model varying outside the range of the data used to develop the model.

Some Critical Success Factors

As in other endeavors, it is important to have a strategy at the start to guide the effort. One needs a plan which includes a goal and a view of what the end looks like. Beginning with an end in mind will increase the probability that the analytics project is successful. As noted earlier, there are two strategies involved, i.e., how the overall project will be executed, and how the statistical analysis and modelling will be done.

The project execution strategy answers questions such as the following ones:

- What is the goals and objectives for the project?
- What is the timeline and critical milestones?

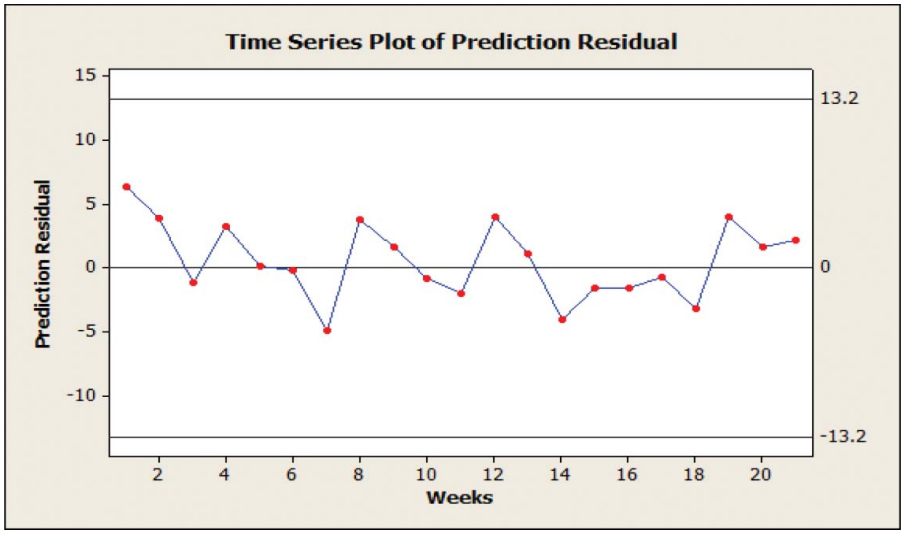

FIGURE 5 **Process yield prediction performance. Prediction residuals are within limits. Model is predicting as expected.**

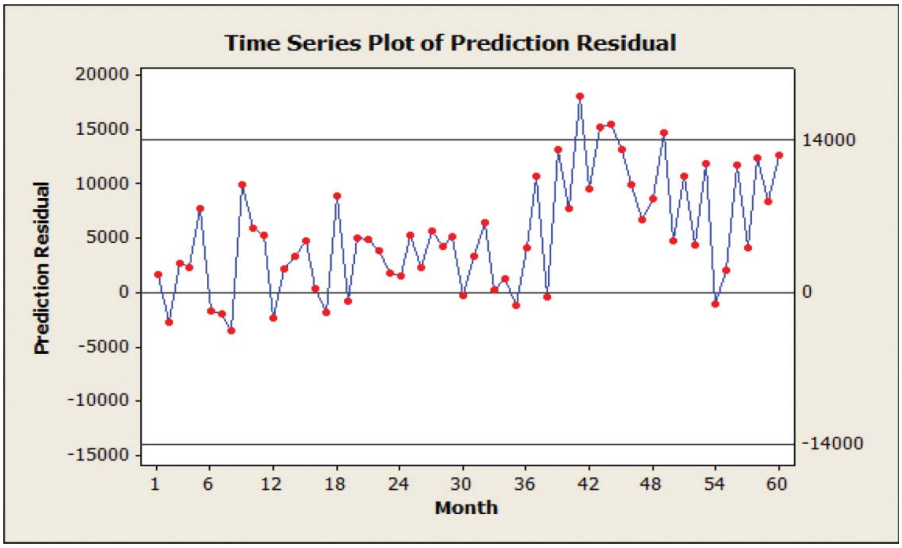

FIGURE 6 **Electricity usage prediction performance. Prediction residuals are out of limits. Model needs to be reevaluated.**

- What are the sources of data?
- What resources are needed: project team members, equipment, funding?
- How will the results be used?

The data analysis and modelling strategy answers questions such as the following ones:

- What specific data sets will be used and how will these data sets be integrated?
- How will the data pedigree be assessed?
- What are the goals for the analysis: Develop a prediction equation? Identify critical variables? Assess process capability and stability? And so on.
- What statistical tools will be used and how will the tools be linked and sequenced?
- How will the results of the statistical analysis be confirmed and validated?
- How will the results be displayed and communicated?

The aspects of strategy are presented as examples of some of the activities a strategy might include. Each situation is different. The strategies are constructed to fit the problem in hand. The guiding question is always, "What do we have to do to have a successful project?"

In addition, there are some other critical success factors that need to be considered including the following ones:

- Do not wait for your data to be "perfect" or you will never get started. But think sequential—Few problems are solved with a single data set;
- Get results quickly and look for quick fixes—Results encourage you and your customers to maintain the investment;
- Do not assume that the data keepers will be cooperative;
- Pay attention to the communication and implementation of results;
- Consider within the business context how the models will be used.

Another success factor is to make "The Data" your #1 priority. Experience shows that 60–80 percent of time is spent on acquiring the data and getting it ready for analysis. Hence, data are at the beginning of the critical path. You also need to make sure that your data can support your project. And of course you need to understand the pedigree of your data.

Make sure that you take full advantage of available subject matter expertise. It will be useful in the planning of the project, selection of the data, interpretation of the results, and aid in the implementation of the solution.

The Glass Is Half-Full: Analytics Brings Opportunity for Success and Failure

The focus on analytics and the associated tools and the availability of large amounts of data offer a unique opportunity to solve important problems that were previously intractable. But we should not lose sight of the fact that the fundamentals of good science, analytical modeling, and interpretation still apply. Ignoring the fundamentals increases the probability that wrong conclusions are reached and inappropriate actions taken.

It is emphasized that that there are three critical aspects of data analysis: the Practical, the Graphical and the Analytical. Particular attention should be paid to the pedigree of the data, including the process that generated the data, the measurement processes, and data collection processes including the sampling schemes used.

Problem solving is sequential in nature, and it is rare that a problem is solved with the analysis of a single set of data. Over time the problem is solved by iterating between conjectures/theories, data collection, data analysis, and the development of new theories. The "Analytics Building Blocks" and "Critical Aspects of Data Analysis" frameworks provide guidance on how to manage and execute on data-rich projects.

The likelihood of success is increased when the following fundamentals are in place: there is a strategy and plan for the project, the pedigree of the data is known and understood, project plans include sequential approaches, and subject matter knowledge is made use of at every opportunity.

This article has been adapted from "Using Analytics: Discovering Variable Relationships that Produce Product and Process Understanding," which appeared in the ASQ Statistics Division Newsletter, June 2014.

ABOUT THE AUTHOR

Ronald D. Snee, Ph.D. is Founder and President of Snee Associates, LLC, a firm dedicated to the successful implementation of process and organizational improvement initiatives. He was employed at the DuPont Company for 24 years prior to pursuing a consulting career. Snee also serves as Adjunct Professor in the Temple University School of Pharmacy and Rutgers Pharmaceutical Engineering program. He received his B.A. from Washington and Jefferson College, and M.S. and Ph.D. degrees from Rutgers University. He is an academician in the International Academy for Quality and Fellow of the American Society of Quality, American Statistical Association, and American Association for the Advancement of Science. Snee's work has been awarded ASQ's Shewhart and Grant Medals, ASA's Deming Lecture, and Dixon Statistical Consulting Excellence Awards. He has authored five books and more than 260 articles in professional journals.

REFERENCES

Box, G. E. P., Norman, R. D. (1987). *Empirical Model-Building and Response Surfaces*. New York, NY: John Wiley and Sons.

Box, G. E. P., Hunter, W. G., J. S. (1978). *Statistics for Experimenters*. New York: John Wiley and Sons.

Davenport, T. H, Harris, J. G. (2007). *Competing on Analytics*. Boston, MA: Harvard Business Press.

DeVeaux, R. D., Hand, D. J. (2005). How to lie with bad data. *Statistical Science*, 20:231–238.

Hoerl, R. W., Snee, R. D. (2010). Statistical thinking and methods in quality improvement: A look to the future. *Quality Engineering* 22(3), July-September (2010):119–139.

Hoerl, R. W., Snee, R. D. (2012). *Statistical thinking: Improving business performance*, 2nd Edition. Hoboken, NJ: John Wiley and Sons.

Juran, J. M., Godfrey, A. B. (1999). *Juran's quality handbook*, 5th Edition. New York, NY: McGraw-Hill.

Montgomery, D. C. (2013). *Design and analysis of experiments*, 8th Edition. Hoboken, NJ: John Wiley and Sons.

Montgomery, D. C., Peck, E. A., Vining, C. G. (2001). *Introduction to linear regression analysis*, 3rd Edition. New York, NY: John Wiley and Sons.

Pierrard, J. M., Snee, R. D., Zelson, J. (1974). A new approach to setting vehicle emission standards. *Air Pollution Control Association Journal* 24(9):841–848.

Quality Engineering. (2012). Special Issue on Statistical Engineering, ed. C. M. Anderson-Cook, and L. Lu. *Quality Engineering* 24(2):107–359.

Ross, W. J. (1999). Aspects of data analysis: Practical, graphical, analytical. *Personal communication*.

Snee, R. D. (1977). Model validation: Methods and examples. *Technometrics* 19:415–428.

Snee, R. D. (2002). Develop useful models—finding the critical few xs allows you to better control and optimize a process. *Quality Progress*, 35:94–96.

Snee, R. D., Hoerl, R. W. (2003). *Leading six sigma—a step by step guide based on experience with general electric and other six sigma companies*. New York, NY: FT Prentice Hall.

Snee, R. D., Hoerl, R. W. (2012). Inquiry on pedigree—do you know the quality and origin of your data? *Quality Progress* December 2012:66–66.

Snee, R. D., Pierrard, J. M. (1977). The annual average: An alternative to the second highest value as a measure of air quality. *Air Pollution Control Association Journal* 27(2):131–133.

Snee, R. D., DeVeaux, R. D., Hoerl, R. W. (2014). Follow the fundamentals—four data analysis basics will help you do big data projects the right way. *Quality Progress*, 47:24–28.