Bridging the Gap between Theory and Practice in Basic Statistical Process Monitoring

4th Stu Hunter Research Conference Paper
2/8/2016 Draft

William H. Woodall
Department of Statistics
Virginia Tech
Blacksburg, VA 24061-0439

ABSTRACT Some issues are discussed relative to the gap between theory and practice in the area of statistical process monitoring (SPM). Among other issues, it is argued that the collection and use of baseline data in Phase I needs a greater emphasis. Also, the use of sample ranges in practice to estimate process standard deviations deserves reconsideration. A discussion is given on the role of modeling in SPM. Then some work on profile monitoring and the effect of estimation error on Phase II chart performance is summarized. Finally, some ways that researchers could influence practice more effectively are discussed along with how SPM research could become more useful to practitioners.

KEYWORDS average run length, control chart, false alarm rate, SPC, SPM, statistical process control.

INTRODUCTION

In this paper I discuss some issues regarding the relationship between theory and application in the area of statistical process monitoring (SPM) and make some recommendations. In some sense, this paper is a follow-up to the paper by Woodall (2000) on controversies and contradictions in statistical process monitoring and the associated discussion. Many of my ideas on current directions in SPM are contained in the recent paper by Woodall and Montgomery (2014) and the overview paper on
Phase I issues and approaches by Jones-Farmer et al. (2014), so repetition of the content of these two papers is avoided to the extent possible.

Because it better reflects the application of the methods, I believe that the use of control charts and other monitoring methods should be referred to as “statistical process monitoring”, not “statistical process control (SPC).” The use of the word “control” implies that control actions based on an adjustment variable are part of the practice and theoretical framework in the sense of Box and Narasimhan (2010). This is most often not the case. As Steiner and MacKay (2000) stated “… the word “control” in SPC causes endless confusion.” In addition, many people associate the term “SPC” solely with the simplest monitoring methods such as the $X$-bar and $R$-charts. It seems that the use of “statistical process monitoring” is becoming more and more common as evidenced by its use by two previous Stu Hunter Research Conference speakers, Capizzi (2015) and De Ketelaere (2015). Dr. Donald J. Wheeler, e.g. in Wheeler (2011c) and elsewhere, referred to control charts as “process behavior charts” in a similar spirit.

Some of the viewpoints and recommendations given in this paper conflict with recommendations of Dr. Wheeler. Dr. Wheeler trains many practitioners and has a large impact on practice. He has published a number of articles in Quality Digest, many of which can be found on-line at www.spcpress.com/djw_columns.php. Quality Digest is a web-based magazine with over 40,000 subscribers, thus reaching an audience of practitioners one to two orders of magnitude larger than Quality Engineering or the Journal of Quality Technology. Most of Dr. Wheeler’s work does not appear in refereed journals. This gives him considerable freedom to reach practitioners without necessarily citing related work or giving what some researchers would consider to be both sides of the story, not unlike what was the case with Genichi Taguchi in the area of designed experimentation in the 1980s. Wheeler (2014b) claimed that probability models and statistical theory do not apply to process monitoring. Taguchi et al. (2001) claimed that the theory of statistics did not apply to his methods either, writing, for example, in the book on the Mahalanobis-Taguchi System,

“The core of statistics is distribution. In quality engineering, however, distribution is not considered. Instead the SN (signal-to-noise) ratio is calculated to evaluate the magnitude of diagnosis error, which can be calculated without considering distribution.”
While many of Dr. Wheeler’s ideas are quite good, some are not. I argue in this paper that theory can play an important role in assessing the performance of existing methods and in the development of new methods. As Deming (1993, p. 106), Dr. Wheeler’s mentor, wrote,

“Without theory, experience has no meaning. Without theory, one has no questions to ask. Hence without theory, there is no learning.”

The gap between theory and practice in industrial statistics was the topic of Steinberg (2015) and Jarrett (2016) in papers presented at the 3rd and 4th Stu Hunter Research Conference, respectively. My discussion is limited to the area of SPM. There are gaps between theory and practice in every field. New methods based on theory must be convincingly shown to be effective in practice and adequately communicated before they can be accepted by practitioners. In addition, there is a considerable amount of inertia that needs to be overcome for a new method to be accepted or for an old method to be discarded. Many excellent points regarding the relationships between theory and practice were made by the discussants of Woodall (2000).

It is helpful to identify the group of people referred to here as “practitioners”. There is considerable variation among practitioners with respect to statistical background. One level includes operators and Six Sigma green belts who have had limited statistical training. Another level includes process engineers, quality engineers, Six Sigma black belts and master black belts. There are practitioners with master’s degrees in statistics or industrial engineering. There are some practitioners who have Ph.D.s in statistics or industrial engineering. In many cases those with the most statistical knowledge train those with lesser knowledge. Another group of practitioners consists of researchers in many other fields, such as engineering, public health surveillance, network monitoring, and healthcare, who implement SPM in their applications.

My focus is on the basic control charting methods, e.g., the $X$-bar chart, the $R$-chart, the $p$-chart, etc. Prior SPM papers presented at the Stu Hunter Research Conference, by Ferrer (2014), Capizzi (2015) and De Ketelaere (2015), have dealt primarily with more sophisticated methods for applications involving what could be an extensive amount of multivariate data.
TWO SIGNIFICANT GAPS

In this section I discuss two gaps between theory and practice in SPM. I would expect that readers and the discussants could identify others.

Importance of Phase I

There remains a need for greater emphasis on the important practical aspects of Phase I, i.e., the collection and analysis of baseline data. Many authors of SPM research papers, including this one upon occasion, have simply assumed the existence of accurate estimates of the in-control parameter(s) from a set of in-control data and then moved directly into a discussion of the on-going monitoring of Phase II. Although the majority of SPM research papers are on Phase II topics, much or more can be learned about process performance in Phase I. As Woodall (2000) mentioned, it is not possible to capture the exploratory and investigative nature of Phase I in a mathematical model. Thus many researchers continue to mention Phase I only in passing as if it were unimportant and uninteresting. Any discussion that is devoted to Phase I is frequently oversimplified. Research on Phase I methods is concentrated on change-point and outlier detection methods.

We should emphasize the importance of Phase I to practitioners and encourage good practice. The reader is referred to Jones-Farmer et al. (2014) for a comprehensive discussion of Phase I issues, goals, and methods. In practice, one could argue, along with Thyregod and Iwersen (2000) and others, that Phase I analysis and actions can be much more important than Phase II monitoring. New researchers reading our statistical literature would likely come away believing that Phase I involves only collecting some data over time and estimating parameters when, in fact, it includes the selection of quality characteristics, assessing the measurement process, determining the sampling strategy based on rational subgrouping ideas, assessing short-term and long-term stability, understanding process variation, implementing process improvement, collecting additional data as needed, and deciding how to best monitor in Phase II. Some researchers equate the use of “rational subgroups” to the use of sample sizes greater than one, which misses the point entirely.

Even without any type of formal decision rules, the plotting of Phase I process data on a run chart can be very informative. For good advice on approaches to reduce process variation, I recommend Steiner and MacKay (2005).
To check for stability of the process over time in Phase I, nonparametric methods, such as those discussed by Capizzi (2015), are applicable since one cannot reliably assume the model form of any underlying distribution(s). Nonparametric methods seem much less appropriate in Phase II, however, since information will be available from Phase I about the form of a suitable underlying distribution. An adequate model is necessary for capability analysis and can lead to a reasonable model for Phase II.

A primary goal in Phase II is to check for the continued stability of the process. As discussed by Steiner and MacKay (2000), the often stated purpose of tracking down assignable causes associated with Phase II signals in order to reduce overall variation is rarely achieved. Greater opportunities for variation reduction are often obtained off-line or in Phase I.

It is assumed in the research literature that Phase II monitoring is done on an observation-by-observation or a sample-by-sample basis. In many cases in practice data are collected at one level of frequency and the control chart examined at another. For example, data could be collected daily, but the chart examined weekly (or even monthly or quarterly). This could be due to time constraints on those responsible for the process. This sequential retrospective approach in Phase II has not been studied.

**Use of Sample Ranges to Assess Variation**

Another major gap between theory and practice regards the use of sample ranges to estimate standard deviations. Researchers advocate moving away from the use of ranges, but practice has not changed very much, if at all. The use of sample ranges to estimate the process standard deviation is used, for example, as the default option in Minitab 17 and JMP for the design of $X$-bar and $R$-charts. This practice was criticized by Nair (2013), who held it up as an example to illustrate how little SPM practice has progressed. Vardeman (1999), Mahmoud et al. (2010), and many others have recommended, with little success, that other estimators of the standard deviation be used based on the resulting improvements in estimation efficiency.

Sample variances and standard deviations are used in every other area of statistics, so why is the situation different in SPM? The primary reason is likely tradition. When Walter Shewhart and others proposed control charting methods in the early part of the 20th century, there were obviously no calculators or computers. It was far more computationally convenient to use sample ranges. The practice,
however, has continued even though the computational advantage of the range is no longer an issue.

It would be a mistake to think that all statisticians agree regarding the use of ranges. Wheeler (2014a), for example, strongly recommended the continued use of sample ranges for three reasons. First, he claimed statistical robustness of the range estimator compared to the use of the pooled standard deviation. Second, he stated that the range and standard deviation are practically equivalent for small subgroup sizes. Third, he argued that the range is a more intuitive measure of variation for many practitioners and much easier to teach. He augmented his arguments by an analysis of case study data where the use of ranges gives more statistically significant effects than with the use of the pooled standard deviation. Montgomery (2013, p. 117) also stated that, with respect to standard deviation estimation for sample sizes of six or less, “the range works very well and is entirely satisfactory.”

I will take each of these justifications one-by-one, but first it is important to realize that the outcome with the use of case study data can never be used to support the use of one estimator over another because the correct decision is unknown. Identifying a larger number of significant effects, as Wheeler (2014a) did with his case study data, is not better if the additional effects have been identified incorrectly. Thus, Wheeler’s (2014b) case study analysis does not support his claim.

It is true that the use of the average sample range provides a more robust estimator of the standard deviation in the presence of outliers than the use of the very non-robust pooled standard deviation. Schoonhoven et al. (2011) showed that the pooled standard deviation was the least robust estimator for estimating the standard deviation among many considered. Schoonhoven and Does (2013) showed, however, that the use of the range-based estimator is far from robust. They showed that the use of the unbiased estimator based on the average standard deviation has slightly better robustness. If one is concerned about outliers, they proposed a much more robust estimator based on the interquartile range. Nazir et al. (2014) recommended a robust Phase I estimator in order to identify outliers with the follow-up use of a more efficient estimator based on the standard deviations to determine the control limits for Phase II. Certainly there is no reason to use the same estimators in Phase II as used in Phase I.

As a more important issue, a standard recommendation is that the range or standard deviation chart be dealt with prior to considering the $X$-bar chart.
Montgomery (2013, p. 252), for example, wrote, “In interpreting patterns on the X-bar chart, we must first determine whether or not the R-chart is in control. … Never attempt to interpret the X-bar chart when the R-chart indicates an out-of-control process.” This standard practice helps to address the outlier problem since outliers would tend to result in out-of-control signals on the control chart for dispersion. One could then consider deleting any outliers attributable to assignable causes before constructing the X-bar chart. Note, however, that Wheeler (2014b) advises against removing outliers from the data before computing control limits.

Given that the chart for variation should be dealt with prior to the chart for location, why don’t we refer to the use of “R and X-bar charts” instead of vice-versa? Why isn’t the chart for variation shown above the X-bar chart instead of below it in software output, such as with Minitab 17, so that practitioners would see it first? I think this change in software output would be helpful.

Let us assume now that there are \( m \) Phase I samples each of size \( n \) (\( n > 1 \)). The most common estimators of the standard deviation are the unbiased estimators \( \overline{R}/d_2 \) and \( \overline{S}/c_4 \), where \( \overline{R} \) and \( \overline{S} \) are the average sample range and average sample standard deviation, respectively. The control chart constants \( d_2 \) and \( c_4 \) depend on the sample size \( n \). These constants are tabled in many references, including Montgomery (2013, p. 720). The usual comparison of the mean-squared error (MSE) involves only these two estimators. Note that the MSE is the expected squared deviation of the estimator from the true standard deviation, with lower values thus being better.

The relative efficiency values with respect to the estimator \( \overline{R}/d_2 \), i.e., the ratios of their mean squared errors, are given in Table 1 for \( m = 1 \) for three estimators. With \( m = 1 \), we have \( \overline{R}/d_2 = R/d_2 \) and \( \overline{S}/c_4 = S/c_4 \). If the relative efficiency is below one, then the estimator in the corresponding column tends to be more accurate than \( \overline{R}/d_2 \). The relative efficiency values given by Montgomery (2013, p.117) in the second column were said incorrectly to correspond to the standard deviation \( S \), whereas they correspond to the values for the unbiased estimator \( S/c_4 \). With the relative efficiencies for the unbiased estimator in column two, it is typically argued that the loss of efficiency is not large enough to be of concern when the sample size is low. It can be seen from the third and fourth columns, however, that if one uses a biased estimator such as the standard deviation, \( S \), or the multiple of \( S \) with the minimum MSE, \( c_4 S \), then there is a much weaker justification for the unbiased range-based estimator.
Table 1 Relative efficiencies of standard deviation estimators relative to the use of the range estimator \( R/d_2 \)

<table>
<thead>
<tr>
<th>Sample size ( n )</th>
<th>( S/c_4 )</th>
<th>( S )</th>
<th>( c_4S )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.000</td>
<td>0.7082</td>
<td>0.6367</td>
</tr>
<tr>
<td>3</td>
<td>0.992</td>
<td>0.8260</td>
<td>0.7790</td>
</tr>
<tr>
<td>4</td>
<td>0.975</td>
<td>0.8617</td>
<td>0.8278</td>
</tr>
<tr>
<td>5</td>
<td>0.955</td>
<td>0.8696</td>
<td>0.8435</td>
</tr>
<tr>
<td>6</td>
<td>0.930</td>
<td>0.8658</td>
<td>0.8448</td>
</tr>
<tr>
<td>10</td>
<td>0.850</td>
<td>0.8152</td>
<td>0.8040</td>
</tr>
</tbody>
</table>

Certainly the range is a simpler statistic than the standard deviation, but should this simplicity override the efficiency advantages of other estimators? Wheeler (2014a) claimed that discussion of the standard deviation in his classes often turns into “gobbledygook”, but is this statistic really too hard to teach to most practitioners? In my academic experience, engineering and other students are capable of understanding the sample standard deviation.

If the standard deviations are used in Phase I, I recommend using \( S\text{-bar}/c_4 \), not an estimator based on the pooled standard deviation. The pooled standard deviation will be most affected by outliers. One should note that most of the studies showing the relative efficiencies of estimators have been based on the assumption of normality without the presence of outliers. Use of the pooled standard deviation is an option in Minitab 17 and the default estimator if one plots only an \( X\text{-bar} \) chart.

Of course a practitioner could consider the evidence in favor of using standard deviations instead of ranges and still reject the idea, but any arguments on behalf of using the ranges should be legitimate ones. In any case, practitioners might reasonably think that the use of ranges has worked well for decades, so why not continue to use them? For this or any other change to be implemented in practice, the advantages of the change must be very convincing and easily accomplished. Some researchers, too, might decide that there are more important issues than this one. I recommend, however, the use of estimators based on standard deviations, but not the use of the pooled standard deviation in Phase I due to its lack of robustness to outliers.

Note that in the frequent case of \( n=1 \), Rigdon et al. (1994) showed that the moving range chart was not effective for detecting changes in process variability. Nelson
(1982) recommended that the moving ranges not even be plotted, stating, “Some delicacy of interpretation would be required since the moving ranges are correlated. Furthermore, the chart of the individual values actually contains all of the information available.” I believe the moving range chart should be deemphasized.

**USE OF PROBABILITY DISTRIBUTIONS IN SPM**

**Use of Models**

We are all likely familiar with George Box’s statement, “All models are wrong, but some models are useful.” In SPM research it is common to assume an underlying probability model such as the normal, exponential, binomial, or Poisson. These models are hypothesized in Phase I and assumed in Phase II. In applications, many of the control limit formulas are based on an assumed underlying probability distribution.

Wheeler (2014b) recommended against using probability models, advocating the use of generic 3-sigma limits which he claims provide

“… a reasonably conservative analysis with virtually every type of homogeneous data set. Thus, no probability model has to be specified. No alpha level is required. No critical values are needed. With this conservative, one-size-fits-all approach any signals found are almost certain to be real, and this allows us to reliably characterize the process behavior without going through the rigamarole (sic) associated with statistical inferences.”

Wheeler (1996, 2011a, 2011b) advised plotting count data with individuals and moving range charts, i.e., $XmR$ charts. Wheeler (1996, p. 50) stated,

“… deciding which probability model is appropriate requires judgement that most students of statistics do not possess.”

He claims that use of the $XmR$ charts will lead to limits close to those based on the probability distribution if the distributional assumptions hold and will be correct if the distributional assumptions are incorrect. One issue with this approach for count data is that the $X$-chart limits will reflect within-sample and between-sample variation. Between-sample variation, which could reflect the presence of assignable causes, will be automatically included in the common cause variation. The use of
*XmR* charts with binomial data can be useful when samples sizes are very large, but they should be used cautiously as an alternative to the *p*-chart.

The use of probability limits is tied to the assumption of a distribution. With the probability limit approach, one would assume a particular distribution and determine the control limits such that the probability of a false alarm is a specified value. Researchers commonly take this approach, or instead control the in-control average run length (ARL), in Phase II studies when two or more monitoring methods are compared.

Deming (1993, p. 181) argued against the use of probability limits, writing,

“It is wrong (misuse of the meaning of a control chart) to suppose that there is some ascertainable probability that either of these false signals will occur. We can only say that the risk to incur either false signal is very small. (Some textbooks on the statistical control of quality lead the reader astray on this point.)”

Wheeler (2015) argued that Shewhart (1931) based the 3-sigma control limits on the Chebyshev inequality without the assumption of a model. Shewhart (1931, p. 276) stated, in fact, that we don’t know the distribution well enough to set up probability limits. Later, however, Shewhart (1939, p. 36) still found it informative to write,

“The control limits as most often used in my own work have been set so that after a state of statistical control has been reached, one will look for assignable causes when they are not present not more than approximately three times in 1000 samples, when the distribution of the statistic used in the criterion is normal.”

Wheeler (2015) argued that any model fit in practice will be adversely affected by the extreme values in the data, which will then affect the control limits. He stated that the control chart should be used to identify outliers. In his view the symmetric 3-sigma limits are sufficiently conservative to work with all types of probability models and are robust to extreme values in the data. Wheeler (2015) showed that the false alarm rate with 3-sigma limits varies by a factor of six (0.003 for the normal distribution to 0.018 for the exponential distribution), but remains low enough in his opinion to not be a concern.
In his argument against probability limits, Wheeler (2015) stated,

“I do not wish to engage anyone in a debate, nor do I wish to raise anyone’s blood pressure.”

Nevertheless, there are some issues with his approach. With time-between-event data, for example, an exponential or Weibull distribution is frequently reasonable. With such skewed distributions, the lower 3-sigma limit will be negative. Thus, there won’t be a lower control limit. Using symmetric control limits for a distribution that is not symmetric can lead to such problems. This is an issue with time-between-event data because the primary interest is in detecting decreases in the mean time between adverse events. One could add supplementary runs rules to detect deceases in the mean, but without some theory we do not know which rules are best to use.

As another issue, without models it seems impossible to extend process monitoring to applications such as spatiotemporal monitoring in public health surveillance or profile monitoring. In addition, applications such as the risk-adjusted monitoring of surgical quality reviewed by Woodall et al. (2015b) cannot be done without a risk model. Restricting ourselves to Shewhart charts with 3-sigma control limits means that we cannot deal with some important applications.

Use of Statistical Inference

Wheeler (2014b) stated his view that statistical inference applies to experimental data whereas process monitoring deals with observational data. This perspective that statistical inference applies only to experimental data would imply that statistical inference is very limited and could not be applied, for example, with most time series data or in fields such as econometrics. Only a small percentage of data are experimental data.

Wheeler (2014b) further stated,

“There is a time and a place for the techniques of statistical inference. There is a time and a place for the use of SPC. To use both effectively you need to understand how they differ. You also need to avoid those who try to merge the two into one ‘unified’ approach. Alpha levels, critical values, distributional assumptions, tests for lack of fit, and the like all belong to the world of experimental studies and statistical inference."
From the very beginning the elements of statistical inference had no place in SPC. Today they still have no place in SPC. And they shall never have any place in SPC, world without end, amen. Understanding the differences between statistics and SPC can bring clarity. Seeking to unify statistics and SPC will simply create confusion.”

If one accepts Wheeler’s (2014b) view that probability models and inference should not be used in SPM, then virtually all of the research on process monitoring would be irrelevant. His approach seems to be that all models are wrong, so no model is useful. His views have changed since he wrote in Wheeler (2000),

“I agree fully with Woodall about the need for theoretical evaluations of techniques. We cannot make progress without the use of mathematical models and theoretical evaluations.”

I believe that if a method proves useful in practice, then there will be a reasonable model which provides theoretical justification for the method. On the other hand, if the method has poor performance under realistic models, then it is unlikely to be useful in practice.

**Choice of Methods**

Researchers assume particular models in order to compare the relative performance of competing methods. If one assumes that the change of interest is a sustained step shift in a parameter, then cumulative sum (CUSUM) and exponentially weighted moving average (EWMA) methods are frequently recommended. Wheeler (1995) preferred a Shewhart chart with Western Electric runs rules to these two types of charts based on the ARL comparisons of Champ and Woodall (1987). I agree with Box and Ramírez (1992) who wrote,

“One of the many virtues of the Shewhart chart is that it is a direct plot of the actual data and so can expose types of deviations from statistical stability of a totally unexpected kind.”

If one has particular out-of-control patterns in mind, however, one can tailor a monitoring method such as a CUSUM chart to have better performance. In general, CUSUM charts should supplement the use of Shewhart charts. Shewhart charts are
particularly useful in Phase I while CUSUM and EWMA charts are more useful in Phase II.

If a CUSUM chart is used, I strongly recommend the use of the tabular form. As discussed by Montgomery (2013, pp. 429-431), the V-mask alternative, which is the default in JMP, is inconvenient to use and often designed incorrectly. For example, JMP and SAS use the discredited Johnson (1961) design approach using error probabilities \( \alpha \) and \( \beta \). See Woodall and Adams (1993). The V-mask and tabular forms can be made equivalent, but only if a point is plotted at (0, 0) for the V-mask. This point at the origin is not included by Montgomery (2013, p. 429), Minitab, JMP, or anyone else to my knowledge, but without it the CUSUM chart cannot signal at the first observation no matter how extreme it may be.

If runs rules are used with an X-bar chart, then the revision proposed by Antzoulakos and Rakitzis (2008) leads to uniformly better ARL performance in detecting sustained step shifts in the mean. With their intuitive approach, a runs rules such as the one which signals if 2-of-3 consecutive points are beyond the same 2-sigma limits would not signal unless the last two points were on the same side of the center line.

In summary, I believe that theory and practice should complement each other. As pointed out by Woodall and Faltin (1996) in their discussion of the role of theory, Shewhart (1939, p. 151) supported the use of theory to generate new techniques, writing,

> “Throughout this monograph care has been taken to keep in the foreground the distinction between the distribution theory of formal mathematical statistics and the use of such theory in statistical techniques designed to serve some practical end. Distribution theory rests upon the framework of mathematics, whereas the validity of statistical techniques can only be determined empirically. Because of the repetitive character of the mass production process, it is admirably suited as a proving ground wherein the try out the usefulness of proposed techniques. The technique involved in the operation of statistical control has been thoroughly tested and not found wanting, whereas the formal mathematical theory of
distribution constitutes a generating plant for new techniques to be tried.”

TWO EXAMPLES ILLUSTRATING USEFULNESS OF THEORY

Models can be useful to examine the performance of proposed methods under idealized conditions. If a method fares poorly in idealized cases, it would likely fare poorly in applications. In the case of monitoring the occurrence rate of rare events, Wheeler (2011b) proposed plotting the number of conforming items between non-conforming items on an $X$-chart with an upper control limit to detect process improvement and plotting rates calculated as the reciprocal of these counts on an $X$-chart with an upper control limit to detect process deterioration. He used a case study data set to illustrate his methods, but the statistical performance of this approach was not investigated. Plots for his case study data are given in Figures 1 and 2.

![Figure 1 Wheeler’s X-chart of geometric random variables](image-url)

*Figure 1 Wheeler’s $X$-chart of geometric random variables*
Our simulation study results (not shown here) show that roughly 12% of the points plotted on the rate chart will fall outside the upper control limit when the data are generated from a stable Bernoulli process. This percentage holds for a wide range of probabilities of a nonconforming item. For Wheeler’s case study with 75 values, one would expect nine false alarms, i.e., more than the number Wheeler (2011b) interpreted in his case study as meaningful signals. More work is in progress, but theory can and should be used to test methods such as this one.

It is important to note limitations of proposed methods and to make performance comparisons to existing methods. Woodall and Driscoll (2015a) reviewed the extensive literature on methods for monitoring the rate of a rare event.

As another example, consider the effect of autocorrelation on control chart performance. This is an increasingly important issue as data are collected more frequently over time and closer spatially. Wheeler (1995, p. 288) stated that autocorrelation will not have an appreciable effect on the control limit unless the sample lag 1 autocorrelation coefficient is greater than 0.7. The simulation results of Maragah and Woodall (1992) tell a different story, however, especially if runs rules are used. An X-chart for data with a sample autocorrelation coefficient of 0.47 is shown in Figure 3. These data were randomly generated from an AR(1) model with a lag 1 autocorrelation coefficient of 0.5. If one uses the Western Electric runs rules, as Wheeler (1995) recommended, there are many false alarms, with ten in Figure 3. Psarakis and Papaleonida (2007) provided a review of the literature on process
monitoring with autocorrelated data with methods designed to account for such issues.

**Figure 3 X-Chart with Autocorrelated Data.**

It would be beneficial for many practitioners to know more about autocorrelation and basic time series modeling. The effect of positive autocorrelation is to increase the number of signals, which can then be misinterpreted as isolated shocks to the process. The root cause of autocorrelation should be removed if possible in order to reduce process variation. The next option would be to use engineering process control, as discussed by Box and Narasimhan (2010) and others, to keep the process variable on target. If these two options are not feasible, then one can monitor the process using a time series approach. For monitoring I recommend the approach of Alwan and Roberts (1988) who recommended plotting the fitted values on a chart with engineering-based limits and plotting the one-step-ahead residuals on an X-chart to detect unusual shocks to the process. Many researchers studying methods for SPM with autocorrelated data consider only the residuals chart. The other chart is needed in practice, however, because the process could wander unacceptably far from target without a signal from the residuals chart.
TWO RECENT DEVELOPMENTS

Profile Monitoring

In an increasing number of applications, quality is best assessed using a function, i.e., a relationship between a response variable and one or more explanatory variables. The monitoring of these functions over time is referred to as “profile monitoring”. Almost all of the work on profile monitoring has occurred since 2000. Noorossana et al. (2011) provided an overview of many profile monitoring methods. Many of the proposed methods involve fitting a parametric model and using the vectors of estimated parameters as input to a multivariate control chart.

Nair (2013) discounted the work on profile monitoring, primarily because in his view the methods were not adequately tied to root cause analysis in applications. Root cause analysis can often be challenging in process monitoring applications. A considerable amount of the work on profile monitoring, however, is based on applications. The origin of the approach stemmed from the monitoring of calibration lines in clinical chemistry and at the National Bureau of Standards. Much of the work remains tied to applications. The work of Bianca Colosimo on monitoring the shape of manufactured parts, for example, has been done in collaboration with mechanical engineers. See, for example, Colosimo (2011) and Colosimo et al. (2014). As another example, Jensen et al. (2015) recently provided an industrial case study application involving the analysis of profile data.

In order to address root cause analysis, research needs to be application-based. Many researchers on profile monitoring take a more hypothetical approach in which they assume a specific model form and investigate the performance of various methods that are designed to detect specified changes in their model. I think this type of work can be useful, although it is always helpful to illustrate the methods using application data with a discussion of the practical aspects and interpretation. A large portion of the profile monitoring literature deals with Phase I issues and includes an analysis of application data.

Effect of Estimation Error

In Phase I we estimate the parameter values used to determine the Phase II control limits. I believe that the effect of the estimation error on the performance of control charts in Phase II has tended to be seriously underestimated by many researchers,
including myself. Much of the research has considered the marginal run length performance of control charts averaged over the distribution of the parameter estimators. This provides average performance metrics that would apply across many practitioners or in repeated applications for a single practitioner. I believe that the study of conditional performance is much more valuable. If a practitioner desires a monitoring procedure with an in-control average run length of 100 in a particular application, for example, then he/she would be interested in the variation about this value for a given amount of Phase I data.

This issue has been brought up by Grant (1946, p. 119), Wheeler (2000), and others, but the extent of the problem is perhaps greater than they realized. Albers and Kallenberg (2004) recognized the extent of the estimation problem, but their paper has been overlooked for the most part. Recent research has demonstrated that it is not possible for a practitioner to be confident of having an in-control ARL within 20% of the desired value even with an unrealistically large amount of Phase I data. See, for example, Saleh et al. (2015a). The phenomenon is illustrated in Figure 4 where the conditional in-control ARLs of the X-bar chart with 3-sigma limits are given for \( n = 5 \) and \( m = 1,000 \) with the standard deviation based on R-bar/d_2. Even with this unrealistically large amount of Phase I data, the variation in the in-control ARL values obtained could be considered to be substantial.

![Figure 4 Histogram of in-control ARL values \( n = 5 \) and \( m = 1000 \). (based on 10,000 simulated X – charts)](image)
The comparisons of various monitoring methods in the research literature are typically set up under the assumption that the in-control parameter values are known. The in-control ARLs are then set to be equal by adjusting the control limits and out-of-control performance is compared. The most commonly assumed scenario is an assumed sustained step shift in the value of the parameter. The fact that in practice we must estimate parameters with a limited amount of data results in a wider gap between this aspect of theory and practice. In theory we can control the in-control ARL, but in practice we cannot.

Historically, a fundamental component of SPM research has been the control of the false alarm rate, the average number of samples between false alarms, or some other metric that reflects the occurrence of false alarms. False alarms can lead to unneeded distraction, wasted time and effort, and increased variability. An excessive number of false alarms can lead practitioners to ignore signals from the monitoring method altogether. This is always the case, but can be an even more important issue when many variables are being monitored. An important question concerns how precisely do practitioners need to know a false alarm rate. If an in-control ARL of 200 is desired, is having an ARL within (150, 250) acceptable? What about (100, 300)?

Since an excessive number of false alarms is a primary issue, a reasonable approach is to control the percentage of in-control ARLs that are below a given value. This approach was recommended by Jones and Steiner (2012) and Gandy and Kvaløy (2013) and has been successfully implemented by Saleh et al. (2015b) and others.

**NARROWING THE GAP BETWEEN THEORY AND PRACTICE**

Academic researchers are primarily rewarded for obtaining grants and writing papers. The greatest rewards are for publishing in the more mathematical journals, i.e., the ones practitioners are least likely to read. In academia the *Journal of the American Statistical Association* and *Annals of Statistics* are ranked well above *Technometrics*, which is ranked above the *Journal of Quality Technology* and *Quality Engineering*, and down the line.

The number of papers published on SPM topics seems to be growing exponentially. It is impossible to keep fully abreast of advances in the literature. Almost all statistical journals publish papers on SPM and the number of journals is
growing. Woodall and Montgomery (2014) cited nearly fifty review papers on SPM that have appeared in just the last ten years. As more and more researchers from around the world are required to publish in ISI Web of Science journals, the number of SPM papers is likely to increase at an increasing rate.

Readers may be most familiar with *Quality Engineering*, *Quality and Reliability Engineering International*, *Journal of Quality Technology*, and *Technometrics*, but a considerable amount of work on process monitoring and change-detection has appeared in the more mathematical journals such as *Annals of Statistics*. See, for example, Poor and Hadjiliadis (2009), Shiryayev (2010) and Tartakovsky, Nikiforov, and Basseville (2015). The more mathematical research has an emphasis on asymptotic performance and optimality results. In addition, there is more of a focus on the Shiryaev-Roberts approach.

Academic researchers are indirectly rewarded based on citation counts, but these result to a large extent from academic researchers citing other academic researchers. Certainly researchers are pleased when their methods are incorporated into software such as Minitab or JMP, but this has little to no effect on any evaluations of their academic performance. Most academics do not do training for industry. For the most part this direct influencing of practice is either done in-house or by consultants. Much of the effect of academics may come from influencing their students who then take industrial jobs and influence others.

The popular textbook by Montgomery (2013) provides discussion of many of the newer techniques and ideas. This textbook reaches many university students and, perhaps to a lesser extent, practitioners. The book by Wheeler and Chambers (1992) seems more widely used for industrial training. Journals and magazines can have a direct impact on practice as long as they are easily accessible and written at an appropriate level. In this respect *Quality Digest* seems to have the upper hand over the quality-related academic journals since all issues are freely available on-line and articles on basic process monitoring methods are written at a level understandable to practitioners. In part because it is not a refereed publication, academic statisticians are generally not rewarded for publishing in outlets such as *Quality Digest*.

There are two main issues. How can we influence practice to better reflect what is supported by important advances in theory? How can we improve SPM research efforts so that they have a more meaningful connection to practice?
Transferring Good Ideas to SPM Practice

The following are some ideas on how we can improve practice by incorporating methods that have been shown to be effective:

1. We can write articles in *Quality Progress* and *Quality Digest*. Note that *Quality Progress* publishes “Statistics Roundtable”, where many in our community have already made important contributions.
2. We can provide input to Minitab, JMP and other software companies if we believe an improvement or addition is needed.
3. Software providers could be more assertive in offering advice on good practice to practitioners.
4. We can write papers to be more understandable. We should try to address the bigger picture or cite relevant papers that provide perspective to the more specific technical issues at hand.
5. We should be careful not to oversell our methods. Replacing all Shewhart charts with CUSUM charts, for example, would be a major step backward, not forward. In addition, one should only monitor when it adds value. SPC was frequently oversold in the 1980s.

Improving SPM Research

The following are some ideas and comments on improving SPM research:

1. We need a forum or an avenue through which practitioners could share more freely the process monitoring challenges they face. Greater interaction between practitioners and researchers would lead to better research.
2. Senior researchers could perhaps be more willing to share promising ideas and directions more widely with newer researchers. With the challenge of monitoring many streams of data, for example, I believe the use of false discovery rate methods have promise and particularly like the approach of Gandy and Lau (2013). Most of the research on process monitoring has been on increasing the sensitivity to small shifts in a quality characteristic. As more and more data streams are monitored, decreasing the sensitivity becomes important. As another research area, I believe our research community has much to offer in the area of social network monitoring. See Woodall et al. (2015c).
3. As research progresses on a particular method or idea, there is a tendency for researchers to tweak assumptions, add more bells and whistles, or consider cases that are more and more specialized. This process can quickly reach the point of diminishing returns, but is limited only by what remains publishable. How can this pattern be changed? Should it be changed? It is important to keep in mind that what seems to be work on a narrow topic can sometimes lead to a breakthrough idea.

4. We should write our papers with practitioners in mind, whenever possible, as opposed to solely gearing our papers for other researchers. Some authors seem to believe that the more complicated they can make a paper or a proposed method, the better it is. We should write from a reader perspective. Papers containing unnecessarily complicated approaches with incomplete or poor explanations can get published, but will tend to have a limited, or even negative, impact.

6. As referees, we need to require papers to be more relevant and understandable.

7. Journal editors should require discussion of practical issues in papers. They also have to work against the natural tendency of journals to become more mathematical over time. Inviting overview papers with industrial statisticians as discussants can be helpful. Practitioners should continue to be asked to referee papers. As Hoerl (2000) stated, our research community “breathes its own exhaust” without practitioner input.

8. It is helpful if authors provide software for their proposed methods and make any case study data available when allowed.

9. We should recognize that the assumption of known in-control parameter values is an exceptionally strong one. It is not helpful to address a new scenario from a Phase II perspective without discussing Phase I issues. The conditional performance of Phase II methods based on Phase I results is much more informative than the use of marginal performance.

10. The performance of any new method should be compared to that of the strongest competitor. Generally steady-state performance comparisons with the conditional expected delay metric are more convincing than zero-state performance comparisons. See, for example, Kenett and Pollak (2012). Steady state comparisons are based on the assumption of a delayed process change, whereas zero-state comparisons are based on a less realistic assumption that any process change occurs at the time monitoring begins or when the monitoring statistic is at its initial value. Some methods, such as the synthetic control charts or the sets method in public health surveillance (Sego et al., 2008), have good zero-state performance, but fare poorly for delayed process changes.
Perhaps in some cases we need to be more outspoken in order to challenge the usefulness of particular approaches. Some refer to work showing weaknesses or disadvantages in existing methods as “negative research”, a term with an unfortunate negative connotation. Knowing what not to do or what approach not to take, however, can be of significant value. I consider, for example, the economic design of control charts to be of little to no practical benefit (Woodall (1986)) and the use of synthetic control charts to involve an unnecessary loss of information (Davis and Woodall (2002) and Knoth (2015)). Davis and Woodall (1988) and others showed that the use of the very common trend rule which signals for a specified number of consecutively increasing or decreasing values on a Shewhart X-bar chart has a greater effect on increasing the false alarm rate than improving the detection of a sustained shift or a trend in the underlying mean of the process. As another example, Woodall et al. (1997) questioned the need for control charts based on fuzzy logic.

**CONCLUSION**

Despite the quotes from W. A. Shewhart given in my paper, it seems past time for us to move beyond justifying what we do, or criticizing the work of others, based on appeals to authority. The manufacturing environment has changed in many ways over the past half-century. Data are now collected in a much higher volume and at a much higher frequency, requiring some modifications to the traditional methods. No one knows how Shewhart’s thinking might have changed to reflect, for example, our current measurement systems, sensor technology and computing capability. Researchers in statistical inference don’t let the views of Sir R. A. Fisher determine the legitimacy of their approaches. The writings of Fisher would lend no support to Bayesian inference, for example, which is now the dominant approach to statistical inference. It seems that in the quality area and in SPM, we reply on the writings of those who founded our field to a much greater extent than in other areas of statistics.

I appreciate the opportunity to present my thoughts at the 4th Stu Hunter Research Conference. There is an incredibly large effort devoted to process monitoring research. Process monitoring itself is receiving increased interest with many challenging problems to be solved in medicine, network monitoring, and other areas. It benefits all of us to align the needs of practitioners with the work of researchers to the fullest extent possible.
ACKNOWLEDGEMENT

The author’s work was supported in part by National Science Foundation Grant CMMI-1436365. The author appreciates the helpful comments from a number of Virginia Tech graduate students and Willis Jensen of W. L. Gore.

REFERENCES


Nair, V. N. 2013. Deming Lecture, Joint Statistical Meetings, Montreal, Canada.


Wheeler, D. J. 1995. *Advanced topics in statistical process control – the power of Shewhart’s charts*. SPC Press, Knoxville TN.


Woodall, W. H., M. Zhao, K. Paynabar, R. Sparks, and J.D. Wilson. 2015c. An overview and perspective on social network monitoring. Submitted for publication.