**Paper DK-02**

# SPC Data Visualization of Seasonal and Financial Data Using JMP®

Diane K. Michelson, SAS Institute Inc, Cary, NC
Annie Dudley Zangi, SAS Institute Inc, Cary, NC

## ABSTRACT

JMP® Software offers many types of Statistical Process Control (SPC) charts, including Shewhart, Cusum, and Moving Average charts. SPC chart features in JMP can be accessed through the menus and dialogs, scripting, and now in version 10.0, through a drag and drop interface. The periodic nature of some financial data makes it unsuitable in its original form for detecting anomalies using a SPC chart. One viable method for presenting this data on a SPC chart is to apply time series techniques first, then chart the output. This paper investigates two case studies applying these techniques.

SPC charts work very well under the ideal conditions of data independence and normality. SPC has traditionally been used in manufacturing, where these conditions are often satisfied. However, SPC is beginning to be used more outside of manufacturing, in areas like insurance claims processing, banking, health care, and survey research. In many of these environments, the desired mean may be shifting up or down, or the responses may be cyclic in nature. In this paper, we examine some of the problems with plotting time series data on control charts and suggest remedies.

## BACKGROUND

Statistical Process Control (SPC) is a methodology for monitoring a time series process to detect shifts in either the location or scale of the process. SPC has been used for almost 100 years in manufacturing. In the recent past, it has become more prevalent in other industries, including finance, health care, insurance, and survey research. (see Roberts, 2006).

Data for control charts is collected in subgroups. When the subgroup size is one, an *Individual-Moving Range* (X-MR) chart is used to control the process. An X-MR control scheme consists of two related run charts. An Individual (or X) chart plots the data in time order, along with *control limits*. Changes in the mean are detected when any point is outside of the control limits. The absolute value of successive differences – the moving ranges – are plotted on an MR chart, along with control limits to detect changes in the process spread. Figure 1 is an example of a typical X-MR chart commonly seen in manufacturing. When the subgroup size is greater than one, *Xbar-S* control schemes are used. The Xbar chart plots the mean of the subgroup; the S chart plots the standard deviation of the subgroup. The upper control limit and lower control limit are drawn at a distance of three standard deviations ($3\sigma_\theta$) of the estimate of the parameter from the centerline. That is, we assume the process has been stable for long enough to estimate $\theta$ without error, and draw control limits at $\theta \pm 3\sigma_\theta$. Control limits are calculated based on historic subgroup mean and standard deviation values. (see Wheeler (2010))

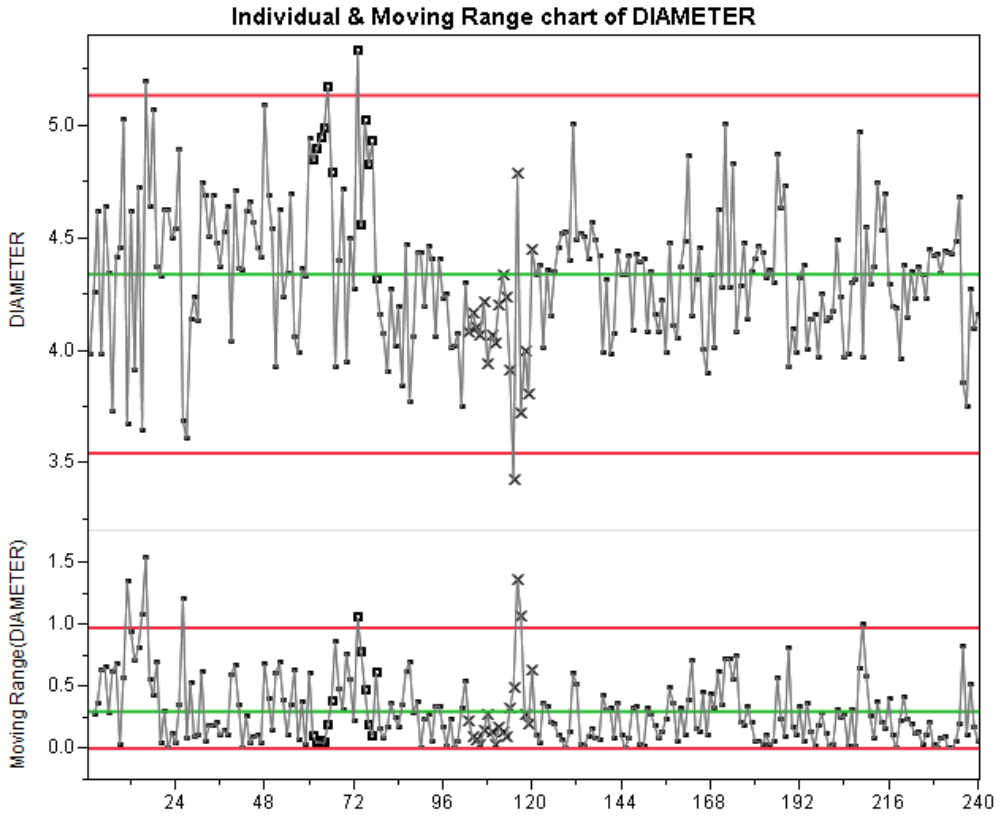


**Figure 1. Typical X-MR chart.**

One of the assumptions in using control charts is that the data come from an independent process. That is, knowing the value of the time series at any point gives no information about future trends of the process, except for the location and scale. In manufacturing, this assumption can often be met by appropriate choice of subgrouping. For example, if data is highly autocorrelated, sampling can be done less frequently. Financial or other business data is often collected daily, weekly or monthly and the sampling plan cannot be changed. Using traditional control charts on positively autocorrelated data causes false signals. (see Michelson, 1994)

One remedy is to model the autocorrelation using *autoregressive-moving average* (ARMA) time series models. The residuals from the model should be uncorrelated and thus are suitable for displaying with a control chart. If a shift occurs in the process mean at time $t_0$, the same shift will occur in the first residual value after $t_0$. The residuals will not see the full shift after time $t_0$, but they will see a dampened shift, leading to longer times until signal for positively correlated data.

In this paper, we give an introduction to ARMA models through two examples. In the first, an equipment company who had successfully used SPC charts in their manufacturing department in the past was interested in monitoring monthly revenue. Often revenue figures are cyclic in nature, dropping off at the beginning of each year and peaking in December. This company, however, produced telescopes and the cycles corresponded directly to astronomy events like eclipses, meteor showers and planet sightings. While they could visually see the trend, weeks with unusual patterns were masked by all the weeks triggering alarms, as traditional control charting methods failed to work.

In another company, control charts were being used on water quality data in a water purification scheme. City water was pumped through a series of filters. At the end of the filtering process, the quality of the water was measured twice per second. The frequency of measurement was important to catching dirty water before it arrived at processing equipment, but the resulting positive serial correlation led to an increase in the false alarm rate of the control chart.

## EXAMPLE 1

### Background

A telescope company who had successfully used SPC charts in their manufacturing department in the past was interested in monitoring revenue. Their revenue figures were cyclic in nature, peaking several weeks before astronomical activities (total eclipses, meteor showers and planetary sightings), then dropping during periods of lower activity. The cyclical nature meant that not only did the data have large swings rather than spikes, but the swings gradually ascended and descended.

**First Attempt at Control Charts**

To graph the data in JMP,

- Select Analyze → Quality and Process → Control Chart Builder
- Drag **Revenue** to the Y drop zone
- Drag **Date** to the X region

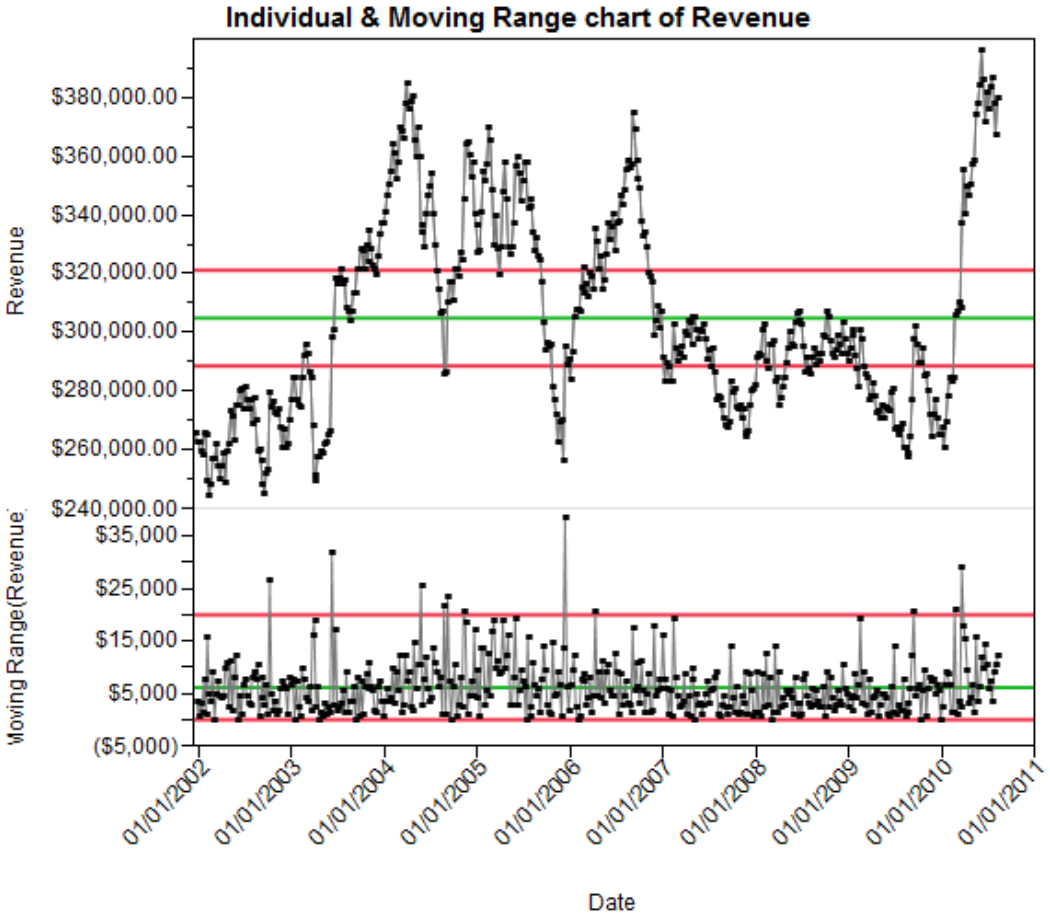


**Figure 2. Individual and Moving Range chart of weekly revenues**

Figure 2 shows ten years of weekly total revenue for the company. The swings were so large that nearly 50% of the data on the Individual chart fell outside the limits. Turning on the tests for points out of control would be meaningless for this chart as so many of the points would signal. The limits, on the other hand, were narrow because the relative change from week to week was typically small, so the moving range chart didn't widen or detect many odd spikes.

While they could visually see the trend, weeks with unusual patterns were masked by all the weeks triggering alarms as traditional control charting methods failed to work.

**Look for Autocorrelation using Lag and Bivariate**

To evaluate the possibility of autocorrelation, look at the relationship between each consecutive point. To do this in JMP:

- Create a new column with the formula, `Lag(:Revenue, 1)`.
- Select Analyze → Fit Y by X, using **Revenue** as the Y and **Lag(Revenue)** as the X.
- In the Bivariate report, select Density Ellipse → 0.95

**Correlation**

| Variable | Mean | Std Dev | Correlation | Signif. Prob | Number |
|---|---|---|---|---|---|
| Lag(Revenue) | 304644 | 34155.55 | 0.971812 | <.0001* | 451 |
| Revenue | 304896.5 | 34289.39 | | | |

**Table 1. Fit Y by X correlation output**

In this output, we see that the correlation is high at 0.97 and significantly different from zero, with a p-value <.0001. From this test, we can reject the hypothesis that the lag of Revenue is uncorrelated with Revenue and conclude there is an autocorrelation effect between consecutive weeks.

### Applying a Time Series Model

JMP offers many options for modeling autocorrelation through the Time Series platform. To try out some modeling options:

- Select Analyze → Modeling → Time Series
- Select **Revenue** as the Y, Time Series and **Date** as the X, Time ID
- Click **OK**
- In the output window, select **ARIMA model group** with the inputs shown in Dialog 1:

Specify ARIMA Model

| ARIMA | | | Seasonal ARIMA | | |
|---|---|---|---|---|---|
| p, Autoregressive Order | 0 | 5 | P, Autoregressive Order | 0 | 0 |
| d, Differencing Order | 0 | 0 | D, Differencing Order | 0 | 0 |
| q, Moving Average Order | 0 | 5 | Q, Moving Average Order | 0 | 0 |
| | | | Periods Per Season | 12 | 12 |

**Dialog 1. Model Group dialog for specifying the ARIMA Model**

The results show that there are a few models that fit very well. The simplest model is the AR(1) model, which has the lowest value of Schwarz's Bayesian Criterion (SBC), as well as a low value of Akaike's Information Criterion (AIC). Next, select output from the AR(1) model and select Save Columns to save the residuals.

| Report | Graph | Model | DF | AIC | SBC |
|---|---|---|---|---|---|
| ☑ | ☑ | AR(1) | 450 | 9423.1170 | 9431.3444 |
| ☐ | ☐ | ARMA(1, 1) | 449 | 9423.4851 | 9435.8262 |
| ☐ | ☐ | AR(2) | 449 | 9423.6594 | 9436.0005 |
| ☐ | ☐ | ARMA(2, 1) | 448 | 9422.0945 | 9438.5492 |
| ☐ | ☐ | ARMA(1, 2) | 448 | 9424.2225 | 9440.6772 |

**Output 1. Time Series Model comparison output**

### Second attempt at Control Charts

Because the significant autocorrelation has been removed from the residuals of the Time Series model, we can use these residuals to identify unusual dates on the Control Chart. To run a Control Chart using the residuals saved in a new data table from the AR(1) model:

- Select Analyze → Quality and Process → Control Chart Builder
- Drag Residual Revenue to the Y-axis drop region

In Figure 3, we see most of the points are now in control and we can investigate the few that went out of control more carefully to find the special causes influencing these dates.

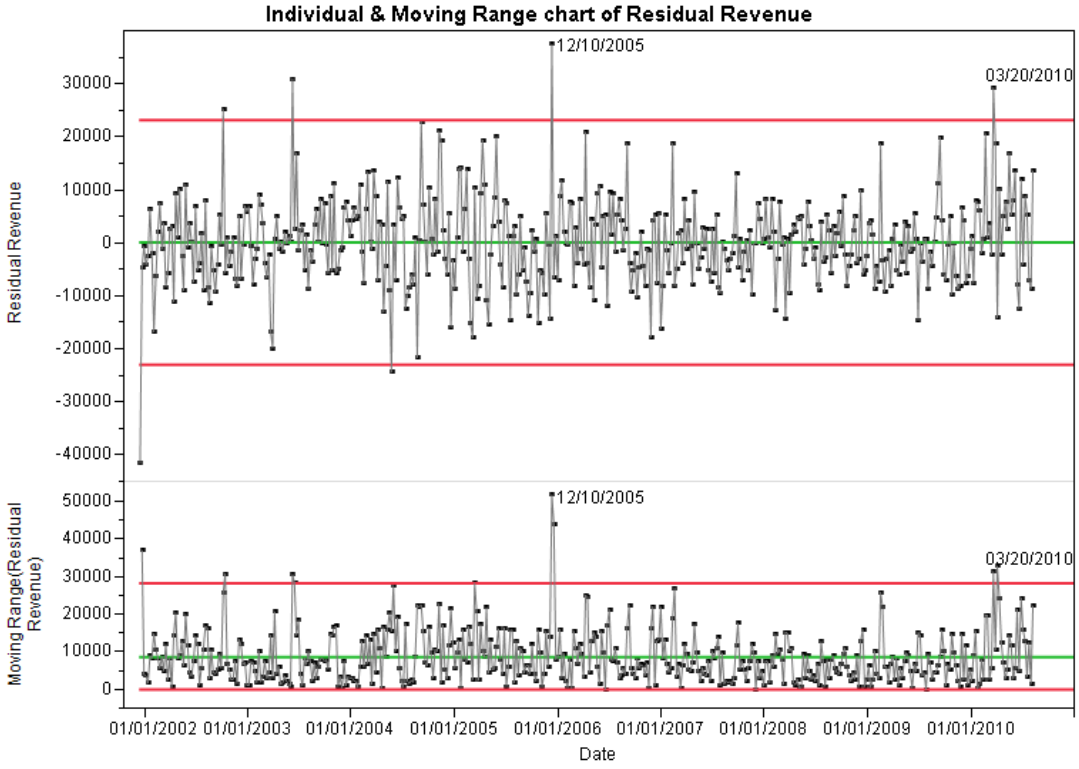


**Figure 3. Individual and Moving Range chart of Revenue Residuals by week**

In fact, we can now see the value on 3/20/2010 flags as an out of control point on the graph of residuals. In preparation of the July 2010 Total Solar Eclipse, the marketing department began running advertisements on 3/01/2010 in popular astronomy magazines. In Figure 3, the positive change in revenue is much more clearly identified than in the control chart in Figure 1.

The high residual on 12/10/2005 has a similar explanation. A marketing campaign targeted toward the Total Solar Eclipse of March, 2006 caused a spike in sales three months prior.

## EXAMPLE 2

### Background

Many industries, including semiconductor and pharmaceutical processing, as well as power generation, rely on water use for many of their steps. City water is fed into a facility through pipes. While the water is safe for human consumption, the particulates and organic compounds in the water can cause all sorts of problems in the manufacturing process. Water is fed through filters and is subjected to chemical and other processing to reduce levels of contaminants to parts-per-billion or parts-per-trillion levels. The resulting water is often called ultrapure water (UPW).

At a certain semiconductor manufacturer, the count of particulates and percent total oxygenated compounds (TOC) were measured before the water was sent from the central utility building (after filtration) to the unit process step. An automated system was set up to signal the engineers if the counts ever went over a specification limit. The process had improved to the point that engineers were interested in moving to a data-based decision system, relying on control limits to tell them when the process had changed, rather than using a specification limit for process control. Data was collected twice per second, due to the need to quickly signal when the particulates or TOC were above the

specification limit, so that the contaminated water would not reach the processing equipment.

Due to the frequency of data collection, both the particulate count data and TOC data time series were highly positively autocorrelated. The engineers knew this positive correlation would cause an increase in false alarms on the Individual chart.

**Time Series Modeling**

One hour's worth of data was collected for modeling purposes. There were 7200 observations of TOC. Due to the proprietary nature of the data, the TOC values in this paper were simulated from the same model used in the application.

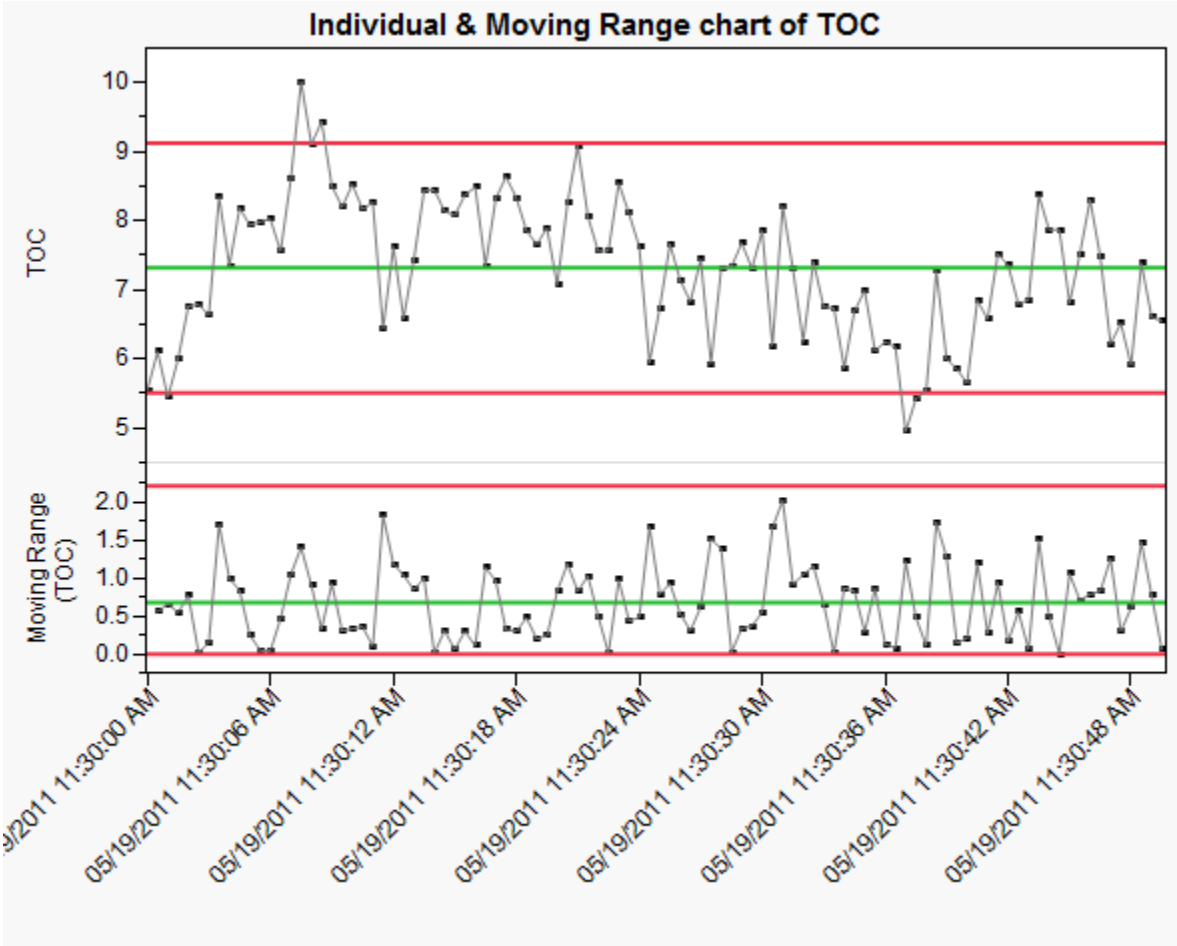


**Figure 4. Individual and Moving Range chart of Total Oxygenated Compounds (percent).**

Figure 4 contains an Individual and Moving Range chart of the first 100 observations taken in 50 seconds. The positive serial correlation has caused two signals on the Individual chart.

Control limits were calculated on the first 1000 observations of the TOC process, using the Control Chart Builder. The limits were copied to a column property that were then used on the control chart for all 7200 observations.

- Select Analyze → Quality and Process → Control Chart Builder.
- Drag **TOC** to the Y zone.
- Drag **Time** to the X zone.

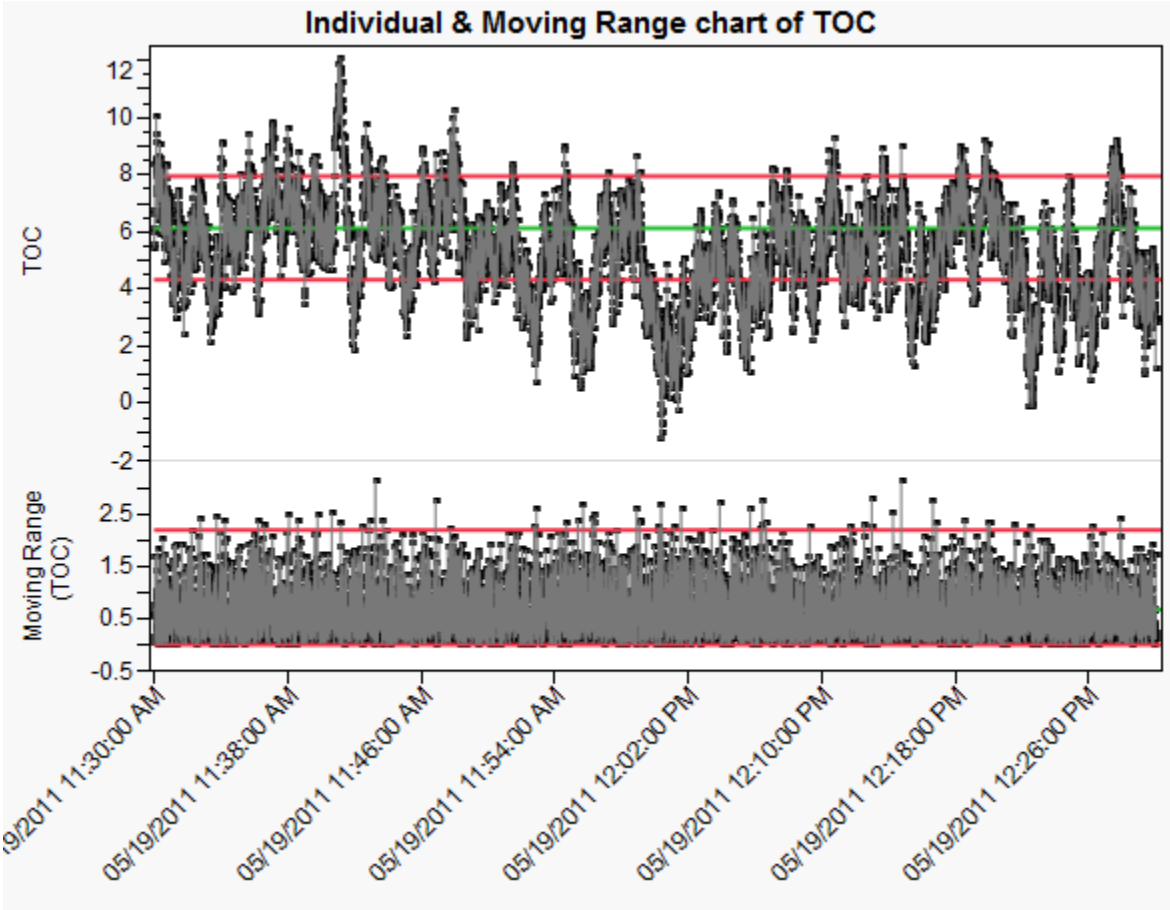


**Figure 5. Individual and Moving Range chart on 7200 data point from TOC process.**

Figure 5 shows that the limits look too tight, and the data wanders around the mean value of just over 6%. The first 1000 data points were used to fit time series models, and the best model was chosen based on the statistics AIC, SBC, as well as model simplicity. An AR(5) process was determined to be the best fit. The prediction formula from this model was saved to the full table and residuals were calculated for each point. These residuals were plotted on a control chart, shown in Figure 6.

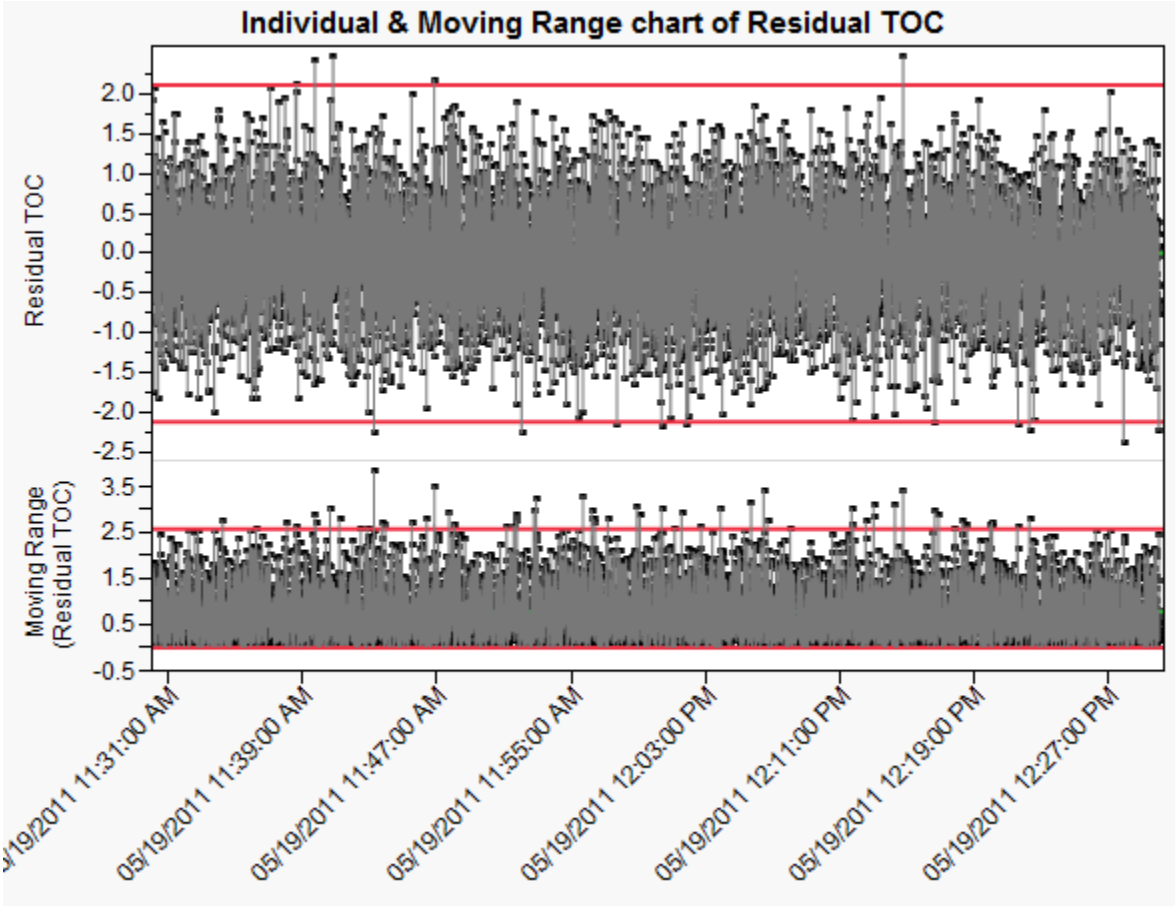


**Figure 6. Individual and Moving Range chart on residuals from AR(5) model on TOC process.**

The reduction in false alarms is seen on the residual chart. In practice, the engineers decided to widen the limits of the residual chart to avoid even these false alarms, due to the frequent sampling of the process. The final system consisted of specification limits on the raw data, and widened limits on a residuals chart. It has been successfully monitoring water quality, signaling quality shifts, for over 10 years.

## STATISTICAL PROPERTIES OF RESIDUAL CHARTS

The use of residual charts for processes with positive autocorrelation has a theoretical justification. Control charts are measured by their *run length*, a random variable that represents the time until a signal. When the process is stable, and consists of just one distribution, the run length should be long. When the process mean has shifted, the run length should be short. It is easily shown that the average run length (*arl*) is 1/*p*, where *p* is the probability of a signal on the chart. We denote the average run length as *arl*(Δ), where Δ represents the size of a mean shift in standard deviation units. For an Individual chart with 3σ limits, assuming the process follows a normal distribution, *p*=0.0027 and *arl*(0)=370. That is, it takes on average 370 runs before this chart falsely signals a mean shift. By contrast, for this chart, *arl*(1)=43.9. That is, it takes on average 43.9 runs before this chart signals a mean shift of 1σ.

Suppose now that $\{X_t\}$ is an autoregressive process of order 1 (AR(1)) with a mean shift of size Δσ at time $t = t_0$. Then

$X_t = \mu_t + \phi(X_{t-1} - \mu_{t-1}) + \varepsilon_t$, where $\mu_t = \begin{cases} 0, & t < t_0 \\ \Delta\sigma, & t \geq t_0 \end{cases}$, and $\varepsilon_t \sim N(0, \sigma^2)$. The residuals from the model are $\{e_t\}$, where $e_t = X_t - \phi X_{t-1}$ for all *t*. Properties of the residuals are given as follows.

$$E[e_t] = \begin{cases} 0, & t < t_0 \\ \Delta\sigma, & t = t_0 \\ (1-\phi)\Delta\sigma, & t > t_0, \end{cases}$$

$$Var[e_t] = (1-\phi^2)\sigma^2,$$

$$Cov[e_t, e_s] = 0, \text{ for all } s \neq t.$$

The entire shift is seen by the residuals at time $t = t_0$. For $t > t_0$, only a fraction of the shift is seen, for $\phi$>0.

Smaller shifts are harder to detect, resulting in higher run lengths for positively correlated data. For example, *arl*(1)=101.9 for an Individual chart on an AR(1) process with $\phi$=0.4, and *arl*(1)=223.3 for an Individual chart on an AR(1) process with $\phi$=0.9. For extremely high correlation the average run length is low, for example, for an AR(1) process with $\phi$=0.98, *arl*(1)=8.56.

The probability of a signal on the first observation after the shift is higher for highly correlated data than it is for independent data. Thus, even though the average run length is large, the probability of a signal on the first observation is greater for correlated data than for independent data. For example, the probability of signal on the first observation after a 1σ mean shift on an Individuals chart on independent data is about 2%. The same probability for a residual chart on an AR(1) process with $\phi$=0.9 is 24% and for $\phi$=0.95 is 58%.

## CONCLUSION

At first glance, statistical process control charts appear wholly unsuitable for seasonal or financial data. As illustrated in the two examples given, modeling techniques applied to the data can sufficiently smooth out the cyclical behavior through the residuals. This smoothed data allows for both applying statistical control charts to the process and further identification and investigation of unusual time periods.

Instead of approximating the expected shape and reacting to changes, by applying Time Series techniques together with statistical process control on the model residuals, the analyst can gain considerable knowledge by both better understanding the financial state itself and identifying which outside indicators or events might affect the state both positively and negatively.

While this is a very effective method, it is computationally intensive and is best suited for a system that is controlled by computers. It is not suitable for paper charts filled out by human operators.

## REFERENCES

Roberts, L. (2006), *SPC for Right-Brain Thinkers: Process Control for Non-Statisticians*: Quality Press, Milwaukee.

Wheeler, D. W., and Chambers, D. S. (2010), *Understanding Statistical Process Control*:, 3rd Ed. SPC Press, Knoxville.

Michelson, Diane K. (1994) *Statistical Process Control for Correlated Data*: unpublished Ph.D. dissertation, Texas A&M University, College Station.

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Diane K. Michelson, Education Division
Annie Dudley Zangi, JMP Group
SAS Institute Inc.
Cary, NC 27513
919-531-9869, 919-531-7624
Di.Michelson@sas.com, Annie.Zangi@jmp.com
www.jmp.com