Paper 318-2012

Monitoring the Variation in Your Multivariate Process: An Introduction to the MVP Procedures

J. Blair Christian and Bucky Ransdell, SAS Institute Inc., Cary, NC

ABSTRACT

Complex processes in modern manufacturing and business environments can generate hundreds and even thousands of process measurements that vary over time. Early detection of process instability is critical for avoiding costly failures and minimizing risk. When the process measurements are correlated, multivariate statistical process monitoring methods are appropriate. Three new procedures in SAS/QC® 12.1, the MVPMODEL, MVPMONITOR, and MVPDIAGNOSE procedures, implement methods that are based on a principal components approach to process monitoring, which was developed in the field of chemometrics. They provide T^2 and SPE charts, which are multivariate summaries of process variation. An example from social media sentiment analysis illustrates how the procedures work together and demonstrates the power of the methods for discovering and diagnosing unusual variation.

INTRODUCTION

Multivariate process monitoring based on principal components is an effective approach for dealing with hundreds or thousands of correlated process measurements. This approach was introduced during the 1990s for applications in the chemical process industries (Kourti and MacGregor 1995, 1996).

In manufacturing applications, the goal of statistical process monitoring—more commonly referred to as statistical process control (SPC)—is to create a stable, predictable process by identifying and removing special causes of variation. In a business environment on the other hand, this form of SPC is seldom feasible because it is often impossible to eliminate the special causes. Nevertheless, process monitoring can be used to better understand process variability and to detect problems early on, thereby minimizing risk and reducing costs. An example of this type is early detection of sentiment change in social media, which is used in this paper to illustrate the use of new SAS/QC procedures for multivariate process analysis.

The new MVPMODEL, MVPMONITOR, and MVPDIAGNOSE procedures in SAS/QC software enable you to use multivariate measurements that are collected over time to monitor a manufacturing or business process for unusual variation. You use these procedures in the following order:

1. You use the MVPMODEL procedure to create a principal components model that uses a small number of components to characterize the variation in the data. It builds a principal components model and saves the loadings and scores in output data sets.
2. With this model, you use the loadings and scores data sets as inputs to the MVPMONITOR procedure to create multivariate control charts of the T^2 and SPE (squared prediction error) statistics to find unusual variation.
3. If you discover unusual variation, you can use the MVPDIAGNOSE procedure to help diagnose and interpret the variation.

This paper refers to these three procedures collectively as the MVP procedures.

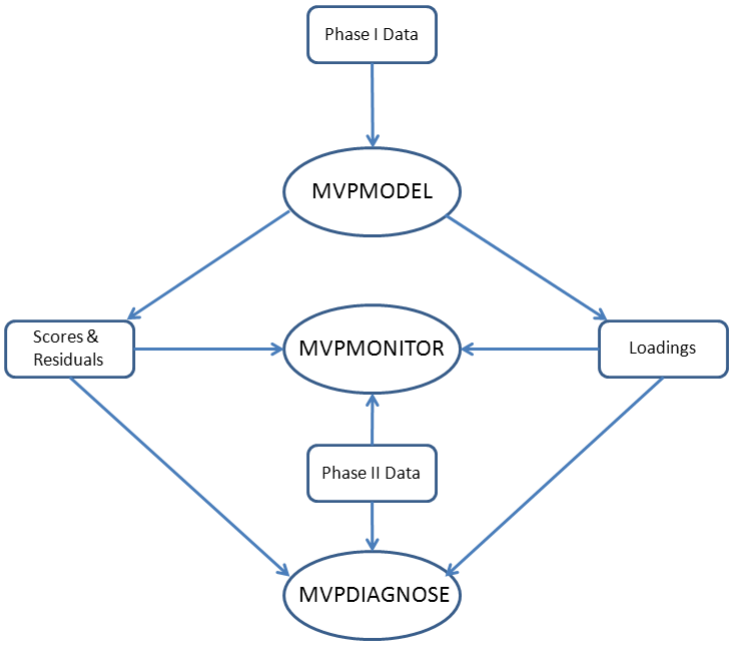
Statistical process control often occurs in two phases:

1. In Phase I, it is not assumed that the process is stable. The goal is to identify (and remove) special causes of variation by constructing a control chart from an initial set of data. You use all three MVP procedures in a Phase I analysis.
2. In Phase II, the goal is to monitor a stable process by constructing a control chart for new data with control limits derived from a previously established model for stable variation. You use model loadings saved by the MVPMODEL procedure as input to the MVPMONITOR procedure.

Figure 1 shows how the MVP procedures work together. The time of measurement is displayed on the control chart although it is not used in the creation of the model.

Although other SAS® procedures can perform a principal components analysis, the MVPMODEL procedure provides convenient features for working with the MVPMONITOR and MVPDIAGNOSE procedures. The MVPMODEL procedure also supports cross validation for selecting the number of components.

Figure 1 Multivariate Process Monitoring with the MVP Procedures



For more information about computational details, examples, syntax, and discussions for the MVP procedures, see the *SAS/QC User's Guide*.

NOTE: The MVP procedures are experimental in SAS 9.3. The examples in this paper reflect the production versions of these procedures, which will be included with SAS/QC 12.1.

SOCIAL MEDIA SENTIMENT DATA

This paper analyzes social media sentiment data as an example of a business process analysis. Figure 2 shows a partial listing of the data set InitialData, which provides the measures of sentiment.

Figure 2 First Five Observations of InitialData

DATE	TWITTER_ POS	TWITTER_ NEU	TWITTER_ NEG	BLOG_POS
02JAN11	0.07927	0.30550	0.32370	0.07493
03JAN11	0.36758	0.18495	0.24483	0.04450
04JAN11	-0.12600	-0.01648	0.18132	0.42822
05JAN11	0.10815	0.05947	0.16864	0.39896
06JAN11	0.18019	0.30424	0.41134	0.02784

BLOG_NEU	BLOG_NEG	NEWS_POS	NEWS_NEU	NEWS_NEG
0.89294	-0.10580	0.21893	0.75615	0.30278
0.26116	0.35238	0.32944	0.12125	0.31837
0.43767	-0.68026	-0.27414	0.30736	0.68433
-0.09716	0.62414	0.29168	-0.34969	-0.63978
0.82759	0.31462	-0.19913	0.02810	0.75023

Nine variables contain measurements of three levels of sentiment from each of three sources. The first part of each variable name identifies the source as one of the following types of social media documents: *TWITTER* for tweets, *BLOG* for blog posts, and *NEWS* for news articles. The last part of the variable name corresponds to the type of sentiment: *POS* for positive, *NEU* for neutral, and *NEG* for negative sentiment. The Date variable indicates the day during which each source was collected. The data were collected over 200 days during the first half of 2011.

BUILD A PRINCIPAL COMPONENTS MODEL WITH PROC MVPMODEL

In the first step of a Phase I analysis, you build a principal components model. This first step often consists of three substeps:

1. building a preliminary model using either all possible components or cross validation
2. examining the output of the preliminary model to determine whether the number of components is adequate
3. creating another model if necessary with a new number of components, and saving the model information in an output data set

The following statements use the MVPMODEL procedure to carry out a preliminary principal components analysis for the data in InitialData:

```
ods graphics on;
proc mvpmodel data=InitialData outloadings=MvpOutloadings;
   var TWITTER_POS TWITTER_NEU TWITTER_NEG BLOG_POS BLOG_NEU BLOG_NEG
      NEWS_POS NEWS_NEU NEWS_NEG ;
run;
```

The DATA= option specifies the input data set, which contains the process measurement variables. The OUTLOADINGS= option specifies the output data set, which contains the principal component loadings. The VAR statement specifies the process measurement variables to be analyzed. The ODS GRAPHICS ON statement enables ODS Graphics, which produces plots for interpreting the model.

The procedure first outputs a summary of the model and the data, shown in Figure 3.

Figure 3 Summary of Model and Data Information

The MVPMODEL Procedure

Data Set	WORK.INITIALDATA
Number of Variables	9
Missing Value Handling	Exclude
Number of Observations Read	200
Number of Observations Used	200
Number of Principal Components	9

By default, the MVPMODEL procedure produces a model with the same number of principal components as number of variables. The loadings are shown in Figure 4.

Figure 4 Listing of Output Data Set MvpOutloadings

_PC_	TWITTER_ POS	TWITTER_ NEU	TWITTER_ NEG	BLOG_POS	BLOG_NEU
0	5.27629	0.85906	0.69358	0.54987	0.51044
1	0.35538	0.36967	0.34234	0.35371	0.35355
2	-0.03100	-0.08450	0.42655	-0.23443	-0.22495
3	-0.36110	-0.13214	-0.19858	-0.19191	-0.15681
4	0.15872	-0.39936	0.06438	0.22425	-0.41339
5	-0.43418	0.12377	0.20643	-0.52173	0.40981
6	0.00066	0.05213	-0.26918	-0.14974	-0.34667
7	0.16880	0.53958	-0.61454	-0.10594	-0.04169
8	0.64731	-0.42993	-0.15926	-0.50971	0.31158
9	-0.28485	-0.43039	-0.37307	0.40885	0.49192

BLOG_NEG	NEWS_POS	NEWS_NEU	NEWS_NEG	_NOBS_
0.41142	0.32727	0.27460	0.09747	200
0.30004	0.32930	0.29797	0.28721	200
0.68078	-0.25216	-0.38160	0.17812	200
-0.03043	0.34421	0.13117	0.78402	200
0.11023	-0.33177	0.68329	0.02993	200
0.09946	-0.13902	0.49773	-0.19095	200
0.43301	0.61523	0.14332	-0.44236	200
0.24964	-0.43883	0.08293	0.16724	200
0.00842	0.07598	0.00076	0.08973	200
0.41672	-0.05700	-0.07973	-0.04522	200

The loadings contain the model information. Observation 0 for the variable *_PC_* in Figure 4 contains the eigenvalues.

After outputting the model information, the procedure outputs the correlation matrix of the variables, shown in Figure 5.

Figure 5 Correlation Matrix

Correlation Matrix

	TWITTER_ POS	TWITTER_ NEU	TWITTER_ NEG	BLOG_POS	BLOG_NEU	BLOG_NEG
TWITTER_POS	1.0000	0.6316	0.5882	0.7449	0.6206	0.5435
TWITTER_NEU	0.6316	1.0000	0.5740	0.6634	0.7647	0.5556
TWITTER_NEG	0.5882	0.5740	1.0000	0.5777	0.6216	0.6963
BLOG_POS	0.7449	0.6634	0.5777	1.0000	0.5647	0.3941
BLOG_NEU	0.6206	0.7647	0.6216	0.5647	1.0000	0.3828
BLOG_NEG	0.5435	0.5556	0.6963	0.3941	0.3828	1.0000
NEWS_POS	0.5308	0.6223	0.4475	0.5801	0.5939	0.4110
NEWS_NEU	0.4923	0.4992	0.4273	0.5522	0.5388	0.3415
NEWS_NEG	0.4087	0.4681	0.4701	0.4574	0.4357	0.4677

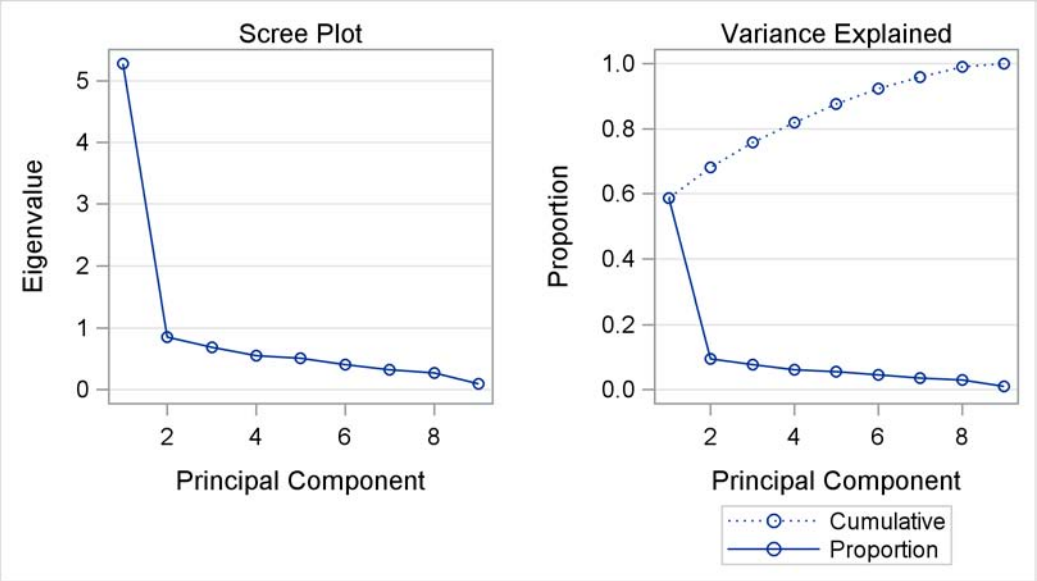
Correlation Matrix

NEWS_POS	NEWS_NEU	NEWS_NEG
0.5308	0.4923	0.4087
0.6223	0.4992	0.4681
0.4475	0.4273	0.4701
0.5801	0.5522	0.4574
0.5939	0.5388	0.4357
0.4110	0.3415	0.4677
1.0000	0.4966	0.5218
0.4966	1.0000	0.4060
0.5218	0.4060	1.0000

The correlations between Twitter sentiments are all above 0.57, while the correlations between the news sentiments are all below 0.53. Also, all of the correlations between Twitter and blog sentiments are greater than both the correlations between Twitter and news and the correlations between blog and news sentiments. This is not surprising since these variables are all related to sentiment volume, so they tend to vary together.

Next, the MVPMODEL procedure produces a scree plot and a variance-explained plot, shown in Figure 6.

Figure 6 Scree Plot and Variance-Explained Plot



DETERMINE THE APPROPRIATE NUMBER OF COMPONENTS

The scree plot shows the eigenvalues for each principal component. Traditionally, the location of the "knee" has been recommended as an aid in selecting the number of principal components for the model (Mardia, Kent, and Bibby 1979). The variance-explained plot shows both the proportion of variance and the cumulative variance that is explained by the principal components. This information is also included in the table of eigenvalues, shown in Figure 7.

Figure 7 Eigenvalue and Variance Information

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	5.27628765	4.41723228	0.5863	0.5863
2	0.85905537	0.16547263	0.0955	0.6817
3	0.69358275	0.14370820	0.0771	0.7588
4	0.54987455	0.03942995	0.0611	0.8199
5	0.51044459	0.09902904	0.0567	0.8766
6	0.41141556	0.08414831	0.0457	0.9223
7	0.32726725	0.05266537	0.0364	0.9587
8	0.27460188	0.17713148	0.0305	0.9892
9	0.09747040		0.0108	1.0000

The eigenvalues are the variances of the principal components, and the proportions reflect the relative amount of variance explained by each component (Jackson 1991).

In general, the goal of a principal components analysis is to create a parsimonious model. By examining the "knee" in the plots in Figure 6, you can see that the amount of variance explained by Components 3 and higher is small. Although the first two components explain only 68% of the variance, in this case a two-component model is adequate.

CREATE A MODEL WITH A SPECIFIED NUMBER OF COMPONENTS

To build a model with a smaller number of principal components, you use the NCOMP= option. The following statements build a model with two principal components:

```
proc mvpmodel data=InitialData
   ncomp=2
   plots=(loadings)
   outloadings=MvpOutloadings
   out=MvpOut;
   var TWITTER_POS TWITTER_NEU TWITTER_NEG BLOG_POS BLOG_NEU BLOG_NEG
      NEWS_POS NEWS_NEU NEWS_NEG ;
run;
```

The PLOTS=(LOADINGS) option requests a loadings plot in addition to the default scree and variance-explained plots. A loadings plot is a scatter plot of the variable loadings for a pair of principal components; it helps you understand the relationships among the variables. The OUTLOADINGS= option saves the model information in the MvpOutloadings data set, which is used as input to the MVPMONITOR procedure. The OUT= option produces an output data set with principal component scores and residuals, which is partially listed in Figure 8.

Figure 8 Partial Listing of Output Data Set MvpOut

DATE	TWITTER_ POS	TWITTER_ NEU	TWITTER_ NEG	BLOG_POS	BLOG_NEU	BLOG_NEG	NEWS_POS
02JAN11	0.07927	0.30550	0.32370	0.07493	0.89294	−0.10580	0.21893
03JAN11	0.36758	0.18495	0.24483	0.04450	0.26116	0.35238	0.32944
04JAN11	−0.12600	−0.01648	0.18132	0.42822	0.43767	−0.68026	−0.27414
05JAN11	0.10815	0.05947	0.16864	0.39896	−0.09716	0.62414	0.29168
06JAN11	0.18019	0.30424	0.41134	0.02784	0.82759	0.31462	−0.19913

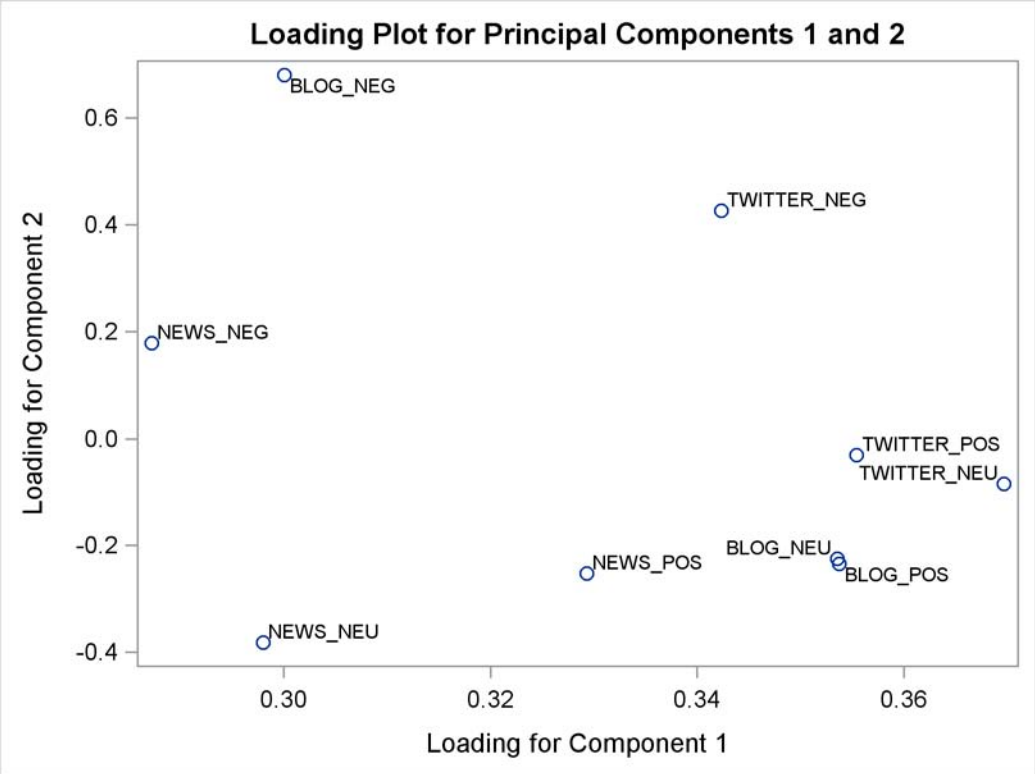
NEWS_NEU	NEWS_NEG	Prin1	Prin2	_NOBS_	R_TWITTER_ POS	R_TWITTER_ NEU	R_TWITTER_ NEG
0.75615	0.30278	1.99377	−0.55730	200	−0.50551	0.13278	0.64243
0.12125	0.31837	1.65882	0.59848	200	0.54576	−0.01906	0.00934
0.30736	0.68433	0.52745	−0.75733	200	−0.62885	−0.33924	0.76983
−0.34969	−0.63978	0.55310	0.93157	200	0.14242	0.02920	−0.00034
0.02810	0.75023	2.00454	0.94434	200	−0.14901	0.25179	0.28132

R_BLOG_ POS	R_BLOG_ NEU	R_BLOG_ NEG	R_NEWS_ POS	R_NEWS_ NEU	R_NEWS_ NEG
−0.66745	0.76478	−0.35606	−0.46662	0.39377	0.06783
−0.33429	−0.06047	−0.14776	0.11767	−0.13677	−0.01444
0.45774	0.37082	−0.90162	−0.84870	−0.00295	1.19501
0.79047	−0.27729	0.48787	0.50341	−0.47489	−1.43972
−0.40631	0.97424	−0.56068	−0.78211	−0.26498	0.58346

The variables Prin1 and Prin2 provide the scores for the principal components. Variables R_TWITTER_POS through R_NEWS_NEG provide the residuals for each of the original variables.

Figure 9 displays the loading plot. Loadings are the variable coefficients in the eigenvectors (linear combination of variables) that define the principal components. The loadings explain how variables contribute to the linear combination.

Figure 9 Loading Plot for Principal Components 1 and 2



In Figure 9, the loadings for the first principal component are all positive and are all similar in value. The first component appears to capture an average of the variables. The second principal component appears to be a contrast between negative and nonnegative sentiment. See Jackson (1991) for more information about the details and interpretation of principal components loadings and scores.

USE THE MVPMONITOR PROCEDURE TO MONITOR THE PROCESS

In the second step of a Phase I analysis, you monitor the process for unusual variation. The following statements produce the multivariate control charts for T^2 and SPE statistics:

```
proc mvpmonitor history=MvpOut loadings=MvpOutloadings;
   time Date;
   tsquarechart / npanelpos=100;
   specchart / npanelpos=100;
run;
```

In a Phase I analysis, you specify the input data set with the HISTORY= option. The LOADINGS= option specifies the data set that contains the loadings for the model, which were generated by the MVPMODEL procedure in the previous section.

The TIME statement specifies a variable that provides the chronological ordering of the observations. The TSQUARECHART statement requests a T^2 chart, and the SPECHART statement requests an SPE chart. The NPANELPOS= option used in both chart statements specifies that 100 points be plotted on each panel of the chart.

Figure 10 shows the T^2 chart, which is displayed in two panels.

Figure 10 Multivariate Control Chart for the T^2 Statistics

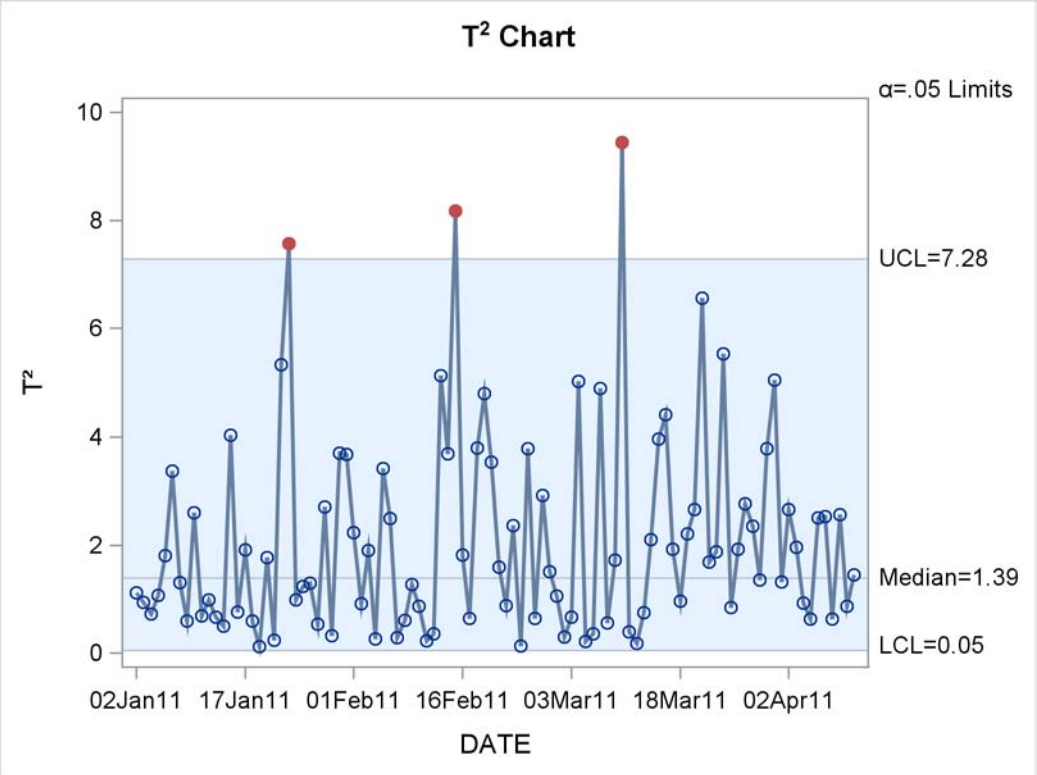
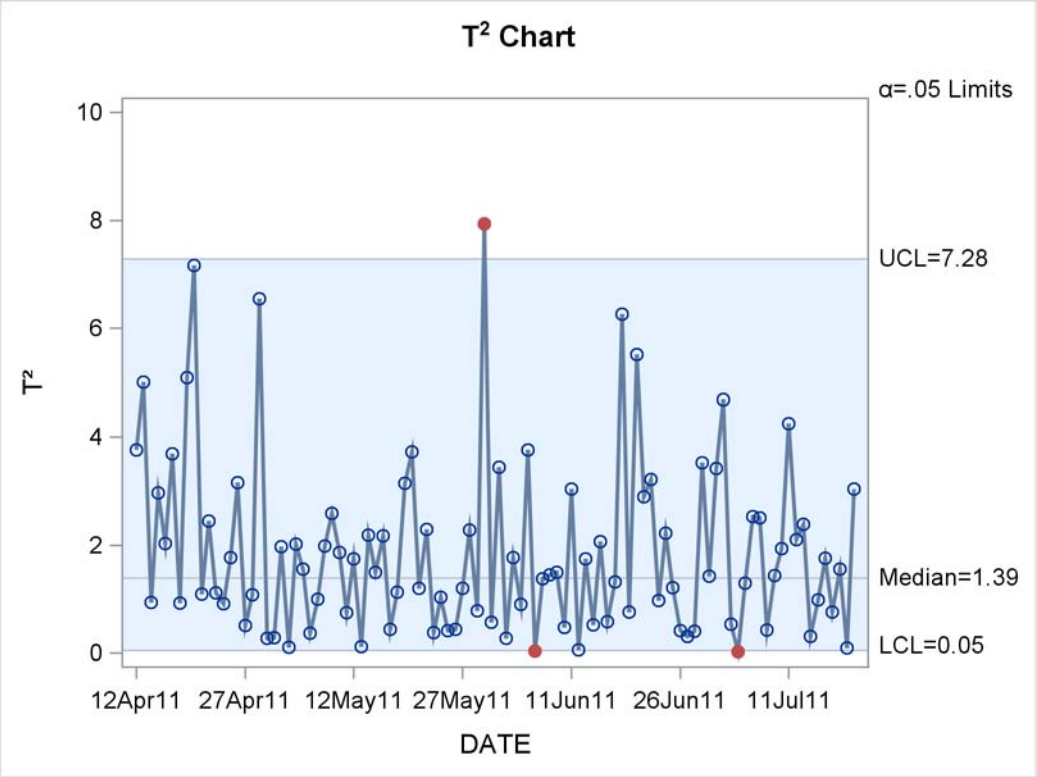


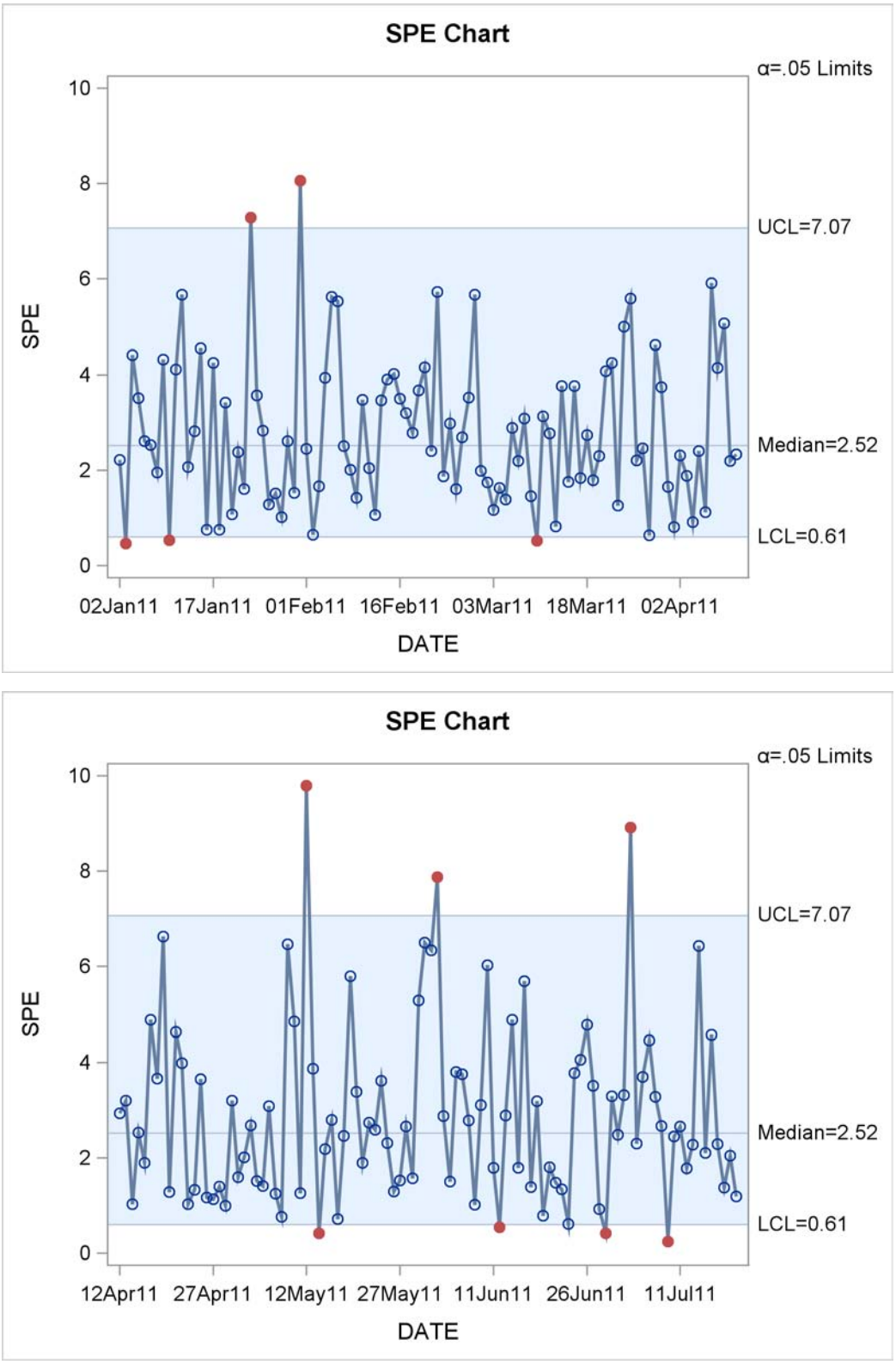
Figure 10 *continued*



The T^2 chart shows the variation in the model plane that is defined by the two principal components. Four points exceed the upper control limit, which is not unusual given 200 total points and control limits that correspond to α=0.05. The two points below the lower limit represent measurements that are very close to the origin in the model plane and do not indicate unusual variation.

The SPE chart (shown in two panels in Figure 11) displays variation that is away from the model plane. Five points exceed the upper control limit. As with the T^2 chart, these do not represent unusual variation. Likewise, the seven points below the lower limit are not a cause for concern.

Figure 11 Multivariate Control Chart for the SPE Statistics



The T^2 and SPE charts show that the variation is stable in the initial time period. The model used to construct these charts can also be used to monitor new data as described in the next section.

MONITOR NEW DATA

You can use your previously constructed model to analyze data collected after you have finished a Phase I analysis. This is called a Phase II analysis.

Figure 12 shows a partial listing of the data set NewData, which contains 37 days' worth of additional sentiment data collected after construction of the principal components model.

Figure 12 First Five Observations of NewData

DATE	TWITTER_ POS	TWITTER_ NEU	TWITTER_ NEG	BLOG_POS
21JUL11	−0.16577	−0.04058	−0.26328	0.07264
22JUL11	0.10000	0.09139	−0.40487	0.08667
23JUL11	−0.11996	0.22553	0.05657	0.24319
24JUL11	0.05238	0.25206	0.39255	0.07407
25JUL11	0.22180	0.12058	0.01496	0.65308

BLOG_NEU	BLOG_NEG	NEWS_POS	NEWS_NEU	NEWS_NEG
−0.08641	−0.53910	−0.42225	−0.62045	0.26873
0.13673	−0.30414	−0.38777	0.17440	−0.31861
0.61805	−0.33558	−0.14428	−0.00481	−0.75728
0.35110	0.18153	0.17487	−0.14056	0.70672
0.60317	−0.18577	−0.09454	0.66571	−0.39414

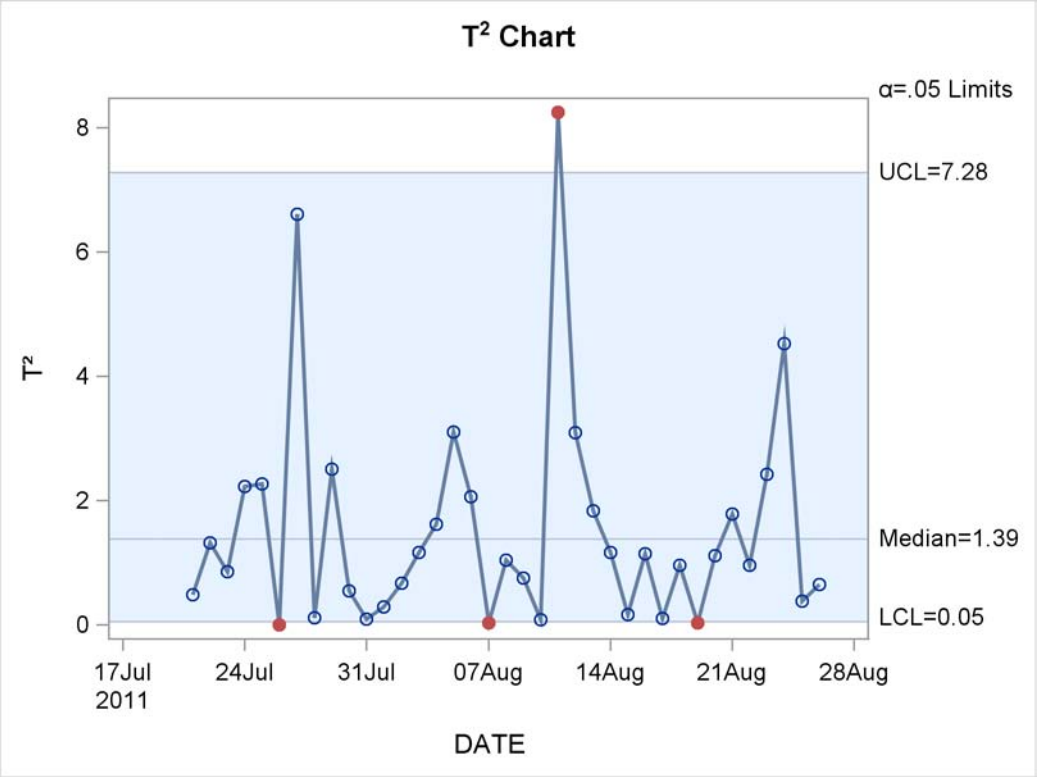
The following statements apply the model to the new data:

```
proc mvpmonitor data=NewData loadings=MvpOutloadings;
   time date;
   tsquarechart / outtable=T2Tab;
   specchart / outtable=SpeTab;
run;
```

The model information is contained in the principal component loadings in the MvpOutloadings data set that was previously created by the MVPMODEL procedure. In a Phase II analysis, you specify the input data with the DATA= option. You can specify the OUTTABLE= option in the TSQUARECHART and SPECHART statements to save a summary of the chart in an output data set for subsequent use by the MVPDIAGNOSE procedure.

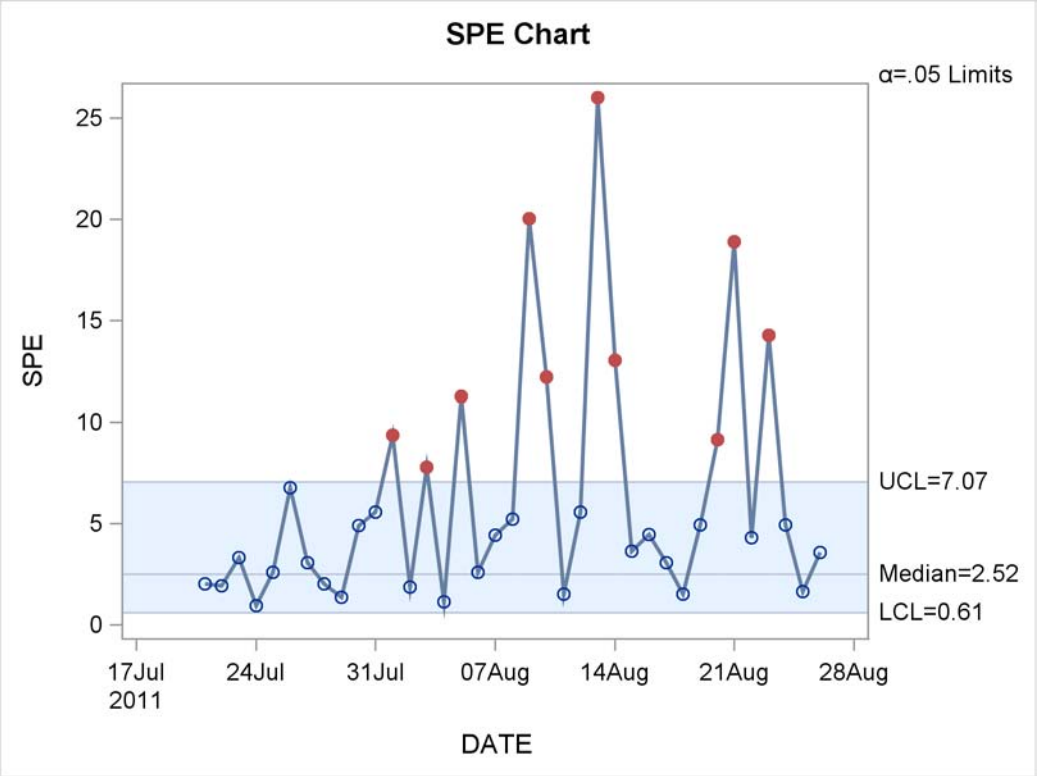
The T^2 chart is shown in Figure 13.

Figure 13 T^2 Chart Created Using NewData



Based on this chart, the variation appears to be stable, but the SPE chart shown in Figure 14 tells a different story.

Figure 14 SPE Chart Created Using NewData



In the SPE chart, there are many points with unusual variation, and the unusual variation in the SPE statistic increases between August 1, 2011, and August 14, 2011. This shows that the process is departing from the model plane. In other words, the process has shifted in a way that is not explained by the model.

DIAGNOSING UNUSUAL VARIATION

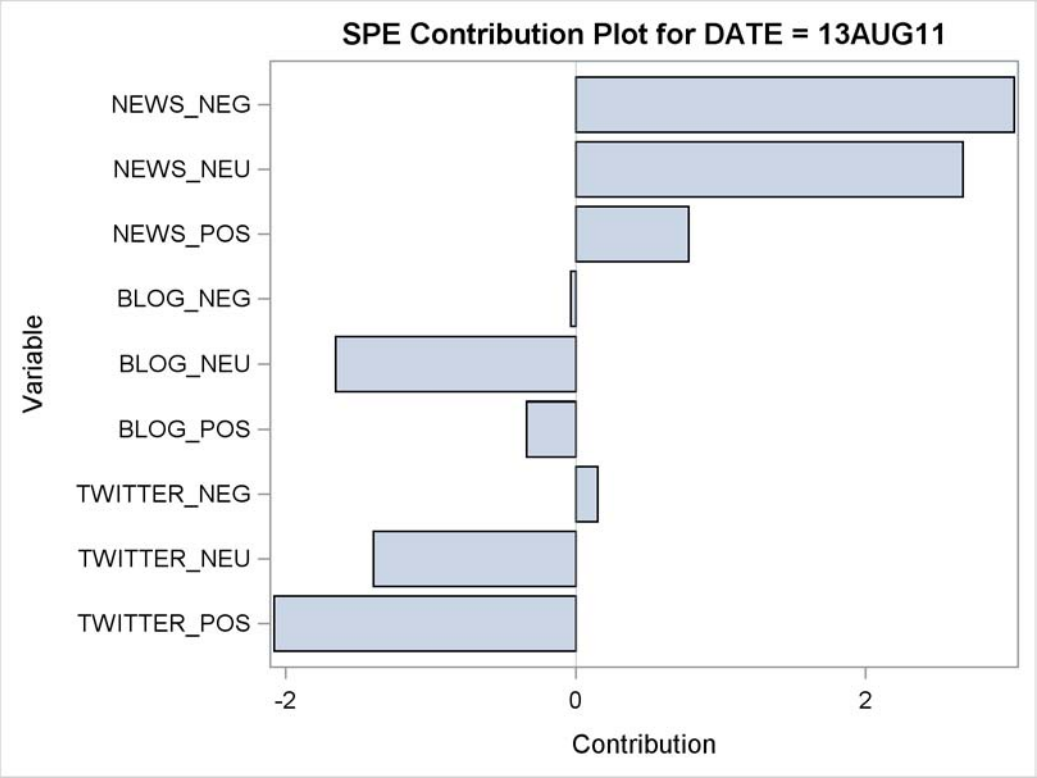
You can use the MVPDIAGNOSE procedure to interpret unusual variation in T^2 and SPE charts. The following statements create a contribution plot that provides an interpretation for the large SPE statistic on August 13.

```
proc mvpdiagnose table=SpeTab;
   where time='13AUG11'd;
   time Date;
   contributionplot;
run;
```

The TABLE= option specifies the SpeTab data set that was produced by the MVPMONITOR procedure. The WHERE statement selects the August 13 observation from the SpeTab data set, and the CONTRIBUTIONPLOT statement produces a contribution plot for that date, which is shown in Figure 15.

A contribution plot shows how the original process measurement variables contribute to an unusual point on a multivariate control chart. In an SPE contribution plot, the contributions are simply the residuals of the variables.

Figure 15 Contribution Plot



The contributions of the News variables have large positive values, while the other variables, Blog and Twitter, have small positive values or negative values. In this example, the data have changed in a fundamental way. Further analysis reveals that this change is in the correlation structure of the data. This change would not be detected by using univariate charts, which do not take correlation into account.

CONCLUSION

The sentiment analysis example illustrates how the MVP procedures work together to discover and diagnose unusual variation in complex processes that might use hundreds or thousands of process measurement variables. The key to the statistical approach implemented by the procedures is a principal components model, which captures the variation in a small number of components. The advantage of this approach over making individual Shewhart charts for each of the variables is that this approach can uncover changes in the relationships of the variables as explained by the model. A relatively small number of multivariate displays (loading plots, score plots, T^2 charts, SPE charts, and contribution plots) yields a wealth of information about the variation in the process.

REFERENCES

Jackson, J. E. (1991), *A User's Guide to Principal Components*, New York: John Wiley & Sons.

Kourti, T. and MacGregor, J. F. (1995), "Process Analysis, Monitoring and Diagnosis, Using Multivariate Projection Methods," *Chemometrics and Intelligent Laboratory Systems*, 28, 3–21.

Kourti, T. and MacGregor, J. F. (1996), "Multivariate SPC Methods for Process and Product Monitoring," *Journal of Quality Technology*, 28, 409–428.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), *Multivariate Analysis*, London: Academic Press.

ACKNOWLEDGMENTS

The authors are grateful to Bob Rodriguez of the Advanced Analytics Division at SAS Institute Inc. for his valuable assistance in the preparation of this manuscript. They are also grateful to Anne Baxter for her editorial contributions.

CONTACT INFORMATION

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

J. Blair Christian
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
(919) 531-8816
blair.christian@sas.com

Bucky Ransdell
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
(919) 531-7928
bucky.ransdell@sas.com