

# Effects of PROC EXPAND Data Interpolation on Time Series Modeling When the Data are Volatile or Complex

Keiko I. Powers, Ph.D., J. D. Power and Associates, Westlake Village, CA

## ABSTRACT

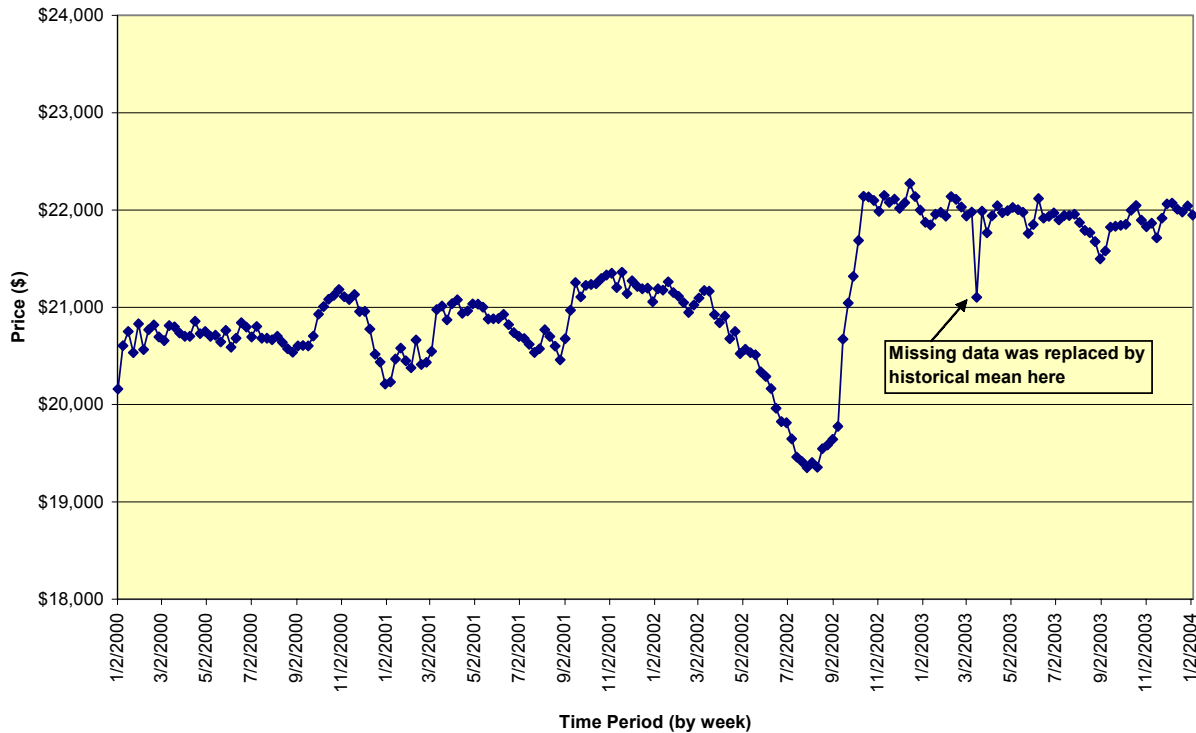
Discrete time series modeling, such as ARIMA, transfer function modeling, etc, assumes that the data are based on the equally spaced intervals of time and no missing data. In the case of missing data, data imputation has to be performed first before time series modeling, and PROC EXPAND of SAS/ETS<sup>®</sup> allows interpolation of time series data. Using the automobile point-of-sales data, the present study investigates if and how PROC EXPAND data interpolation affects time series modeling, when the data are volatile (e.g., vehicle sales price) or complex (e.g., sales volumes, which exhibit seasonality with respect to week, month, and year). Just as with the study by Powers (2004) that investigated another PROC EXPAND function, i.e., frequency conversion, it is important to know the existence and nature of effects of missing data imputation when developing and interpreting time series models. .

## INTRODUCTION

SAS/ETS, PROC EXPAND offers various data-management functions specifically for time series data that are useful for time series modelers. Knowing how these functions perform in various settings allow us to perform PROC EXPAND data management steps more efficiently. Powers (2004) investigated the frequency conversion feature of PROC EXPAND and recommended that the high-to-low frequency conversion should be handled with caution when the time series data contain seasonal components. Another key feature of PROC EXPAND is missing data interpolation. Because commonly-used time series approaches, such as the Box-Jenkins ARIMA or autoregression model, require data with no missing values, performing missing data interpolation with a statistically sound method is particularly important.

In the general statistics domain, there is much literature on advanced approaches to missing data (e.g., Johnson, and Davis, 1998; Little and Rubin, 2002; Marsh, 2000; Yuan, 2000). There are new SAS procedures, PROC MI and PROC MIANALYZE, that handles some of the complex approaches (Yuan, 2000). When time series data are involved, special considerations should be given for missing data imputation, due to its time-related dependency. The illustration below highlights its unique property. In this example that shows the weekly time series of the average vehicle price, we have a missing data point, and missing data imputation was performed by a simple mean replacement using the historical mean. It is clear that this procedure is introducing an outlier data point. Since the data display a somewhat step-wise pattern, it would be more appropriate to use a local mean based on adjacent values rather than the global historical mean. Obviously, when the time series data are less volatile and do not display an upward or downward trend over time, missing data replacement by the historical mean would suffice, but special caution is needed when dealing with volatile or complex time series data. The purpose of the present study is to investigate how PROC EXPAND missing data interpolation performs with volatile or complex time series data. The analysis results help us better understand how to handle missing data for various time series modeling approaches.

### Illustration of how historical mean replacement could create a problem in time series data



## DATA AND PROCEDURE

Power Information Network (PIN), a division of J. D. Power and Associates, has been collecting and archiving point-of-sale data from automobile retailers in the United States for over a decade. The sales data are electronically sent to the PIN computer system daily. The data consist of new and used vehicles purchased by consumers and cover 26 key markets (e.g., New York, Florida, Detroit/Chicago area, Texas, California, etc) with over 6,000 retailers currently providing data to PIN. The total cumulative number of transactions now exceeds 17 million records. The transaction data can be aggregated and summarized to produce historical data, which allow us to analyze key market indicators to monitor over time the automobile market in the USA. These indicators include various features, such as vehicle price, profit margin, consumer incentives, or days to turn (i.e., how many days it takes to sell).

For the present study, effects of missing data interpolation on univariate time series modeling were investigated by simulating missing data situations with different numbers of consecutive missing data points introduced to originally complete time series variables. The number of missing observations was increased up to 10 percent of the data; for example, the weekly time series variables had 210 observations, and therefore scenarios with up to 20 missing observations were simulated.

Three time series variables were studied with the following setups:

- (1) Days to Turn (DTT) time series data: displaying a volatile pattern
  - a. Time series frequency: weekly
  - b. Missing data: at two time periods – stable period and volatile period
  - c. Missing data duration: # of consecutive weeks= 4, 8, 12, 16, 20
  - d. Imputation method: (1) join (linear) and (2) spline (cubic)

- (2) Sales Unit (SALE-W) time series data: displaying a complex seasonality pattern
  - a. Time series frequency: weekly
  - b. Missing data: at one time period starting on January 2002
  - c. Missing data duration: # of consecutive weeks = 4, 8, 12, 16, 20
  - d. Imputation method: (1) join (linear) and (2) spline (cubic)
  
- (3) Sales Unit (SALE-M) time series data: displaying a complex seasonality pattern
  - a. Time series frequency: monthly
  - b. Missing data: at one time period starting on January 2002
  - c. Missing data duration: # of consecutive months = 1, 2, 3, 4, 5
  - d. Imputation method: (1) join (linear) and (2) spline (cubic)

Days to Turn (DTT) refers to the number of days it takes an automobile dealer to sell a vehicle, and Figure 1 shows the weekly mean DTT for a mid-size car from January 2000 through December 2003. As can be seen in Figure 1, the data are fairly volatile, displaying three clear 'hills'. The pattern shows considerable variations from time to time due to an introduction of the new model year. The missing data simulation is conducted with two starting points for this variable – from May 2000 and from September 2001 – in order to assess possible differential effects of missing imputation between when the data are stable versus volatile. Sales Unit data refer to the number of vehicles sold by the PIN participating dealers. Two time-series variables, one weekly (SALE-W) and the other monthly (SALE-M) (see Figures 2 and 3, respectively), are prepared, both of which clearly show the existence of seasonality. For each setup, the missing data imputation pattern is first investigated by visual inspection of time series plots, and the effects of missing data imputation on ARIMA model specifications are compared for various setups stated above. It should be noted that the number of data points for the monthly SALE variable is 48, which is less than the number of data points recommended for univariate time series analysis. This setup is included to illustrate the impact of missing data when the number of data points is not ideal.

**Figure 1. Weekly Time Series of Days to Turn**

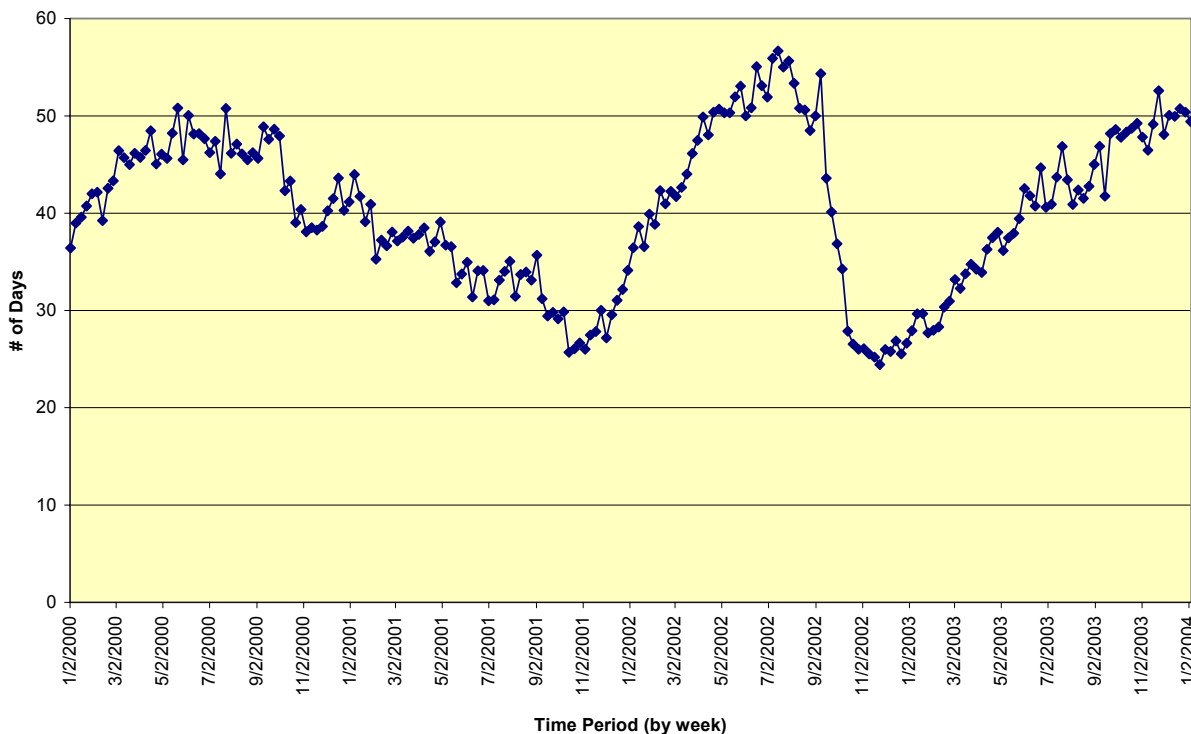


Figure 2. Weekly Time Series of # of Units Sold

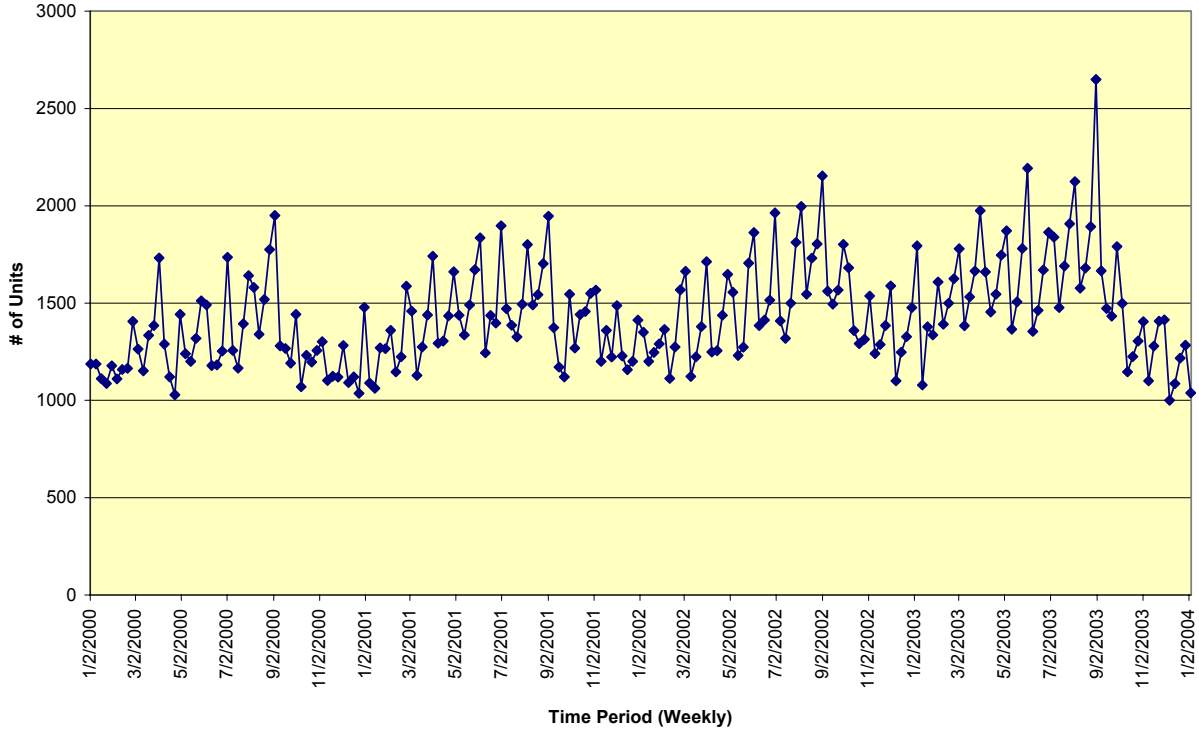
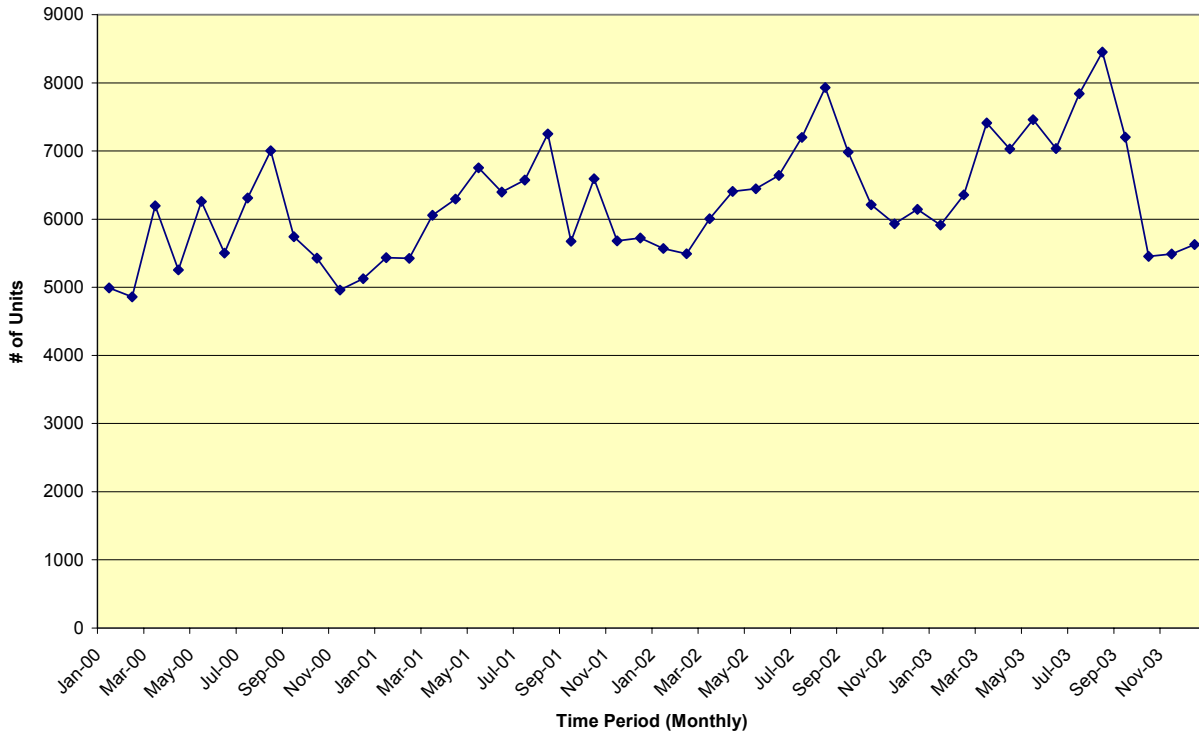


Figure 3. Monthly Time Series of # of Units Sold



For performing missing data interpolation, the SAS codes below were used:

```
/* Days to Turn: missing data interpolation with method = spline */
proc expand data=d2 out=out1 from=week;
id enddat;
convert daystotu;
run;

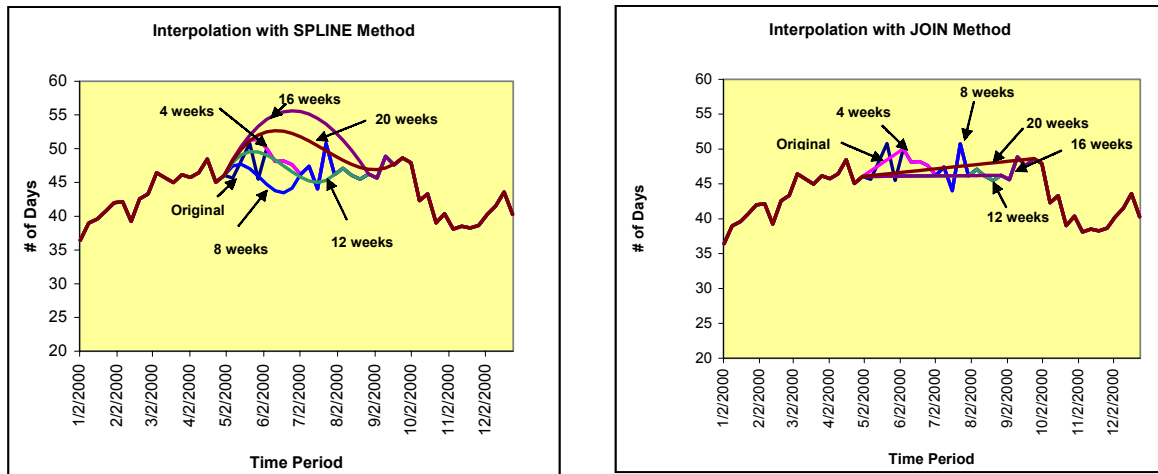
/* Days to Turn: missing data interpolation with method = join */
proc expand data=d2 out=out1 from=week;
id enddat;
convert daystotu / method=join;
run;
```

The default method, SPLINE, fits a third-degree polynomial function for data interpolation, whereas the JOIN method is based on a linear function. These derived variables are then modeled with PROC ARIMA, and the resultant ARIMA specifications are compared against those based on the corresponding original variables.

## RESULTS AND DISCUSSION

First, to visually inspect the performance of PROC EXPAND derived variables, these variables were plotted with the original variables. Figure 4. shows missing data interpolation results for Days to Turn variable, using the SPLINE and JOIN methods, when the numbers of missing data points are 4, 8, 12, 16, 20 weeks. The starting time point is May 2000 for these charts, or from the section of the time series where the data dynamics is fairly stable.

Figure 4. Visual Inspection of Missing Data Interpolation of Days To Turn Variable with SPLINE and JOIN Methods



In general, observed differences in the interpolated values are small between the two methods when the time series variable exhibits a smoother pattern. Compared to Days to Turn, the weekly variable for sales unit (SALE-W) showed a substantially different interpolation results between SPLINE and JOIN methods. The SPLINE method tends to exaggerate the upward or downward pattern, and when the time series variable consists of a series of spikes (e.g., weekly series for Sales Unit) the interpolated values could get extremely high or low. For example, the missing interpolation for the number of missing weeks being 12 and 16 resulted in considerably different patterns for this variable.

Next, the performance of PROC EXPAND data interpolation was investigated with Box- Jenkins ARIMA modeling. ARIMA models were developed for each of the three variables (weekly DTT, weekly SALE, and monthly SALE) using the original non-missing data. These model specifications were then applied to the set of the interpolated data to test if the original model specification will be recovered. The goodness of fit criteria (i.e., white noise tests with Q statistics, significance tests for parameter estimates) were used for this purpose. For example, the ARIMA model based on the original non-missing DTT was ARIMA (1,1,0). The same model specification was examined with data having varying numbers of missing data points from 4 to 20.

The analysis results in Table 1 are based on the SPINE method. Since the ARIMA results were similar between the SPINE and JOIN methods, only the outcomes with the former approach are presented. The summary table indicates that the model specification for DTT was not affected by the interpolation procedure even when the number of missing data points was as high as 12, whether the missing data time period was for May 2000 or for September 2001. On the other hand, the two SALE variables displayed quite different patterns between the weekly series and monthly series. While the monthly series did not get affected by the missing data points, the weekly series did very poorly even when the number of missing data points was only 4 (or less than 2% of the 210 data points in the weekly SALE series). The poor results by the weekly SALE variable might be due to the spiky pattern of this variable. On the other hand, the short time series, SALE-M, with 48 observations, did not display any negative effects from the data interpolation despite its small number of observations.

**Table 1. ARIMA Results Based on Different Number of Consecutive Missing Data Points**

**(1) Days to Turn (DTT) weekly time series variable**

# of consecutive missings	ARIMA specification	Parameter Estimates	Q statistics on residuals
0: original series - no missing	(1,1,0)	$(1-B) X_t = 0.061 + (1 + 0.244 B) e_{t-1}$	Lag 6: n.s. Lag 12: n.s.
Missing starts at May 2000			
4	(1,1,0)	$(1-B) X_t = 0.061 + (1 + 0.203 B) e_{t-1}$	Lag 6: n.s. Lag 12: n.s.
8	(1,1,0)	$(1-B) X_t = 0.061 + (1 + 0.205 B) e_{t-1}$	Lag 6: n.s. Lag 12: n.s.
12	(1,1,0)	$(1-B) X_t = 0.061 + (1 + 0.160 B) e_{t-1}$	Lag 6: n.s. Lag 12: n.s.
16	(1,1,0)	no estimates due to Q stat results	Lag 6: p<.05 Lag 12: n.s.
20	(1,1,0)	no estimates due to Q stat results	Lag 6: p<.05 Lag 12: n.s.
Missing starts at Sep 2001			
4	(1,1,0)	$(1-B) X_t = 0.061 + (1 + 0.233 B) e_{t-1}$	Lag 6: n.s. Lag 12: n.s.
8	(1,1,0)	$(1-B) X_t = 0.061 + (1 + 0.234 B) e_{t-1}$	Lag 6: n.s. Lag 12: n.s.
12	(1,1,0)	$(1-B) X_t = 0.061 + (1 + 0.209 B) e_{t-1}$	Lag 6: n.s. Lag 12: n.s.
16	(1,1,0)	$(1-B) X_t = 0.061 + (1 + 0.226 B) e_{t-1}$	Lag 6: n.s. Lag 12: n.s.
20	(1,1,0)	no estimates due to Q stat results	Lag 6: p<.01 Lag 12: p<.05

**(2) Sales Unit (SALE-W) weekly time series variable**

# of consecutive missings	ARIMA specification	Parameter Estimates	Q statistics on residuals
0: original series - no missing	(1,0,0)(0,1,0) <sub>4</sub>	$(1-B_4) X_t = 7.585 + (1 - 0.253 B) e_{t-1}$	Lag 6: n.s. Lag 12: n.s.
4	(1,0,0)(0,1,0) <sub>4</sub>	no estimates due to Q stat results	Lag 6: p<.05 Lag 12: p<.01
8	(1,0,0)(0,1,0) <sub>4</sub>	no estimates due to Q stat results	Lag 6: p<.01 Lag 12: p<.01
12	(1,0,0)(0,1,0) <sub>4</sub>	no estimates due to Q stat results	Lag 6: p<.01 Lag 12: p<.01
16	(1,0,0)(0,1,0) <sub>4</sub>	no estimates due to Q stat results	Lag 6: p<.01 Lag 12: p<.01
20	(1,0,0)(0,1,0) <sub>4</sub>	no estimates due to Q stat results	Lag 6: p<.01 Lag 12: p<.01

**(3) Sales Unit (SALE-M) monthly time series variable**

# of consecutive missings	ARIMA specification	Parameter Estimates	Q statistics on residuals
0: original series - no missing	(2,0,0)(0,1,0) <sub>12</sub>	$X_t = 5035 + (1 - 0.453 B - 0.329B_2)(1 - 0.864B_{12}) e_{t-1}$	Lag 6: n.s. Lag 12: n.s.
1	(2,0,0)(0,1,0) <sub>12</sub>	$X_t = 5037 + (1 - 0.454 B - 0.329B_2)(1 - 0.863B_{12}) e_{t-1}$	Lag 6: n.s. Lag 12: n.s.
2	(2,0,0)(0,1,0) <sub>12</sub>	$X_t = 5055 + (1 - 0.444 B - 0.332B_2)(1 - 0.855B_{12}) e_{t-1}$	Lag 6: n.s. Lag 12: n.s.
3	(2,0,0)(0,1,0) <sub>12</sub>	$X_t = 5096 + (1 - 0.459 B - 0.304B_2)(1 - 0.847B_{12}) e_{t-1}$	Lag 6: n.s. Lag 12: n.s.
4	(2,0,0)(0,1,0) <sub>12</sub>	$X_t = 5096 + (1 - 0.462 B - 0.302B_2)(1 - 0.845B_{12}) e_{t-1}$	Lag 6: n.s. Lag 12: n.s.
5	(2,0,0)(0,1,0) <sub>12</sub>	$X_t = 5109 + (1 - 0.472 B - 0.292B_2)(1 - 0.832B_{12}) e_{t-1}$	Lag 6: n.s. Lag 12: n.s.

Overall, the results indicated that the PROC EXPAND interpolated data perform well with ARIMA models for time series with smooth movements, such as DTT, even when the number of consecutive missing data points is fairly high.

On the other hand, when the data are spiky, as has been observed with the weekly SALE variable, the original ARIMA model specification fails to stand even when the number of missing observations is rather small. A possible reason is that the big zigzag movements between adjacent data points, present in the weekly SALE variable, are not captured and recovered well when missing data interpolation is performed. Therefore, when performing missing data interpolation to time series data with volatile or complex patterns, it is very important to first study the data dynamics carefully to make sure that missing data interpolation does not introduce data points that do not fit the overall trend or the cyclic pattern of the time series data.

## CONCLUSIONS

Missing data imputation for time series variables requires special considerations because of the time-related dependency among observations. PROC EXPAND offers options for missing data interpolation, and understanding how it performs with various data patterns is important for applied time series modelers. For example, if there are a few missing observations in the middle of time series data, it is not wise to discard the portion before the missing data period, because many time series approaches require a fairly large number of observations. At the same time, discrete time series approaches assume no missing time periods, and for this reason, it is valuable to have an approach that generates sound missing data imputation outcomes. Overall, the results from this study showed that PROC EXPAND is a valuable tool for this purpose, displaying minor impacts due to consecutive missing data points up to 10 percentages of the total number of observations. This is particularly true when the time series data display a smooth pattern with small adjacent point-to-point movements. On the other hand, when the data are complex and volatile with sharp increases/decreases between adjacent observations, visual inspection of the time series plot is highly recommended.

## REFERENCES

- Johnson, M.J. and Davis, P. (1998). The Effect of Missing Data on Sample Sizes for Repeated Measures Models. Proceedings of the Twenty-Third Annual SAS Users Group International Conference, Paper 231.
- Little R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data, 2nd Ed. New York: John Wiley & Sons, Inc.
- Marsh, L.C. (2000). Correcting for Missing Discrete Responses in Business Surveys. Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference, Paper 269-25.
- Powers, K.I. (2004). Empirical Investigation of Time Series Frequency Conversion with PROC EXPAND. Proceedings of the Western Users of SAS Software: 12th Annual Conference in Pasadena, CA.
- Yuan, Y.C. (2000). Multiple Imputation for Missing Data: Concepts and New Development. Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference, Paper 267-25.

## CONTACT INFORMATION

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

Keiko I. Powers, Ph.D.  
J. D. Power and Associates  
2625 Townsgate Road  
Westlake Village, CA 91361  
Work Phone: 805-418-8114  
Fax: 805-418-8241  
E-mail Address: [Keiko.Powers@jdpa.com](mailto:Keiko.Powers@jdpa.com)

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.