Paper SA-08

Are Sales Figures in Line With Expectations? Using PROC ARIMA in SAS[®] to Forecast Company Revenue

Saveth Ho and Brian Van Dorn, Deluxe Corporation, Shoreview, MN

ABSTRACT

The distribution of corporate sales, like many types of financial data, tends to be serially correlated, with revenue in one time period related to revenue in the past. This allows sales revenue to be described by an ARIMA time series model. These models use autoregressive and moving average components that relate current observations to prior observations, and also contain seasonal components that utilize the similarity of values to those obtained at the same time point in prior years. Special indicator variables can account singular events such as holiday weeks, major market campaigns or price increases.

The ARIMA procedure in SAS/ETS[®] allows for the identification and estimation of these time series models, with many diagnostics available for describing the serial correlation and for selecting an appropriate model. In addition, values can then be forecast for subsequent time periods, along with their corresponding confidence intervals. These intervals can then be compared to actual observed values to determine if the values obtained are within an expected range, or instead suggest that an outside factor is at work that may shift the distribution of revenue. The process is suited for statisticians with a working knowledge of time series data, or by those interested in accounting for serial correlation in standard regression models.

This process will be illustrated by modeling weekly business to business corporate revenue experienced by the Deluxe Corporation's Small Business division. The forecasts derived from these models could aid business decisions that rely on revenue by providing a context for actual revenue when compared with expectations.

INTRODUCTION

One of the main tasks of a statistician is to adjudicate whether data trends are simply random variability or instead are evidence of something real. Non-statisticians, when presented with random noise, will nevertheless assess some sort of causation and attribute what happened to outside factors. It takes statistical judgment to exercise the restraint required to avoid conclusions of causation.

In the corporate world, it is difficult to find a better example of this than corporate earnings. Earnings are the collective result of employees work product and of management's decisions, so just about any change from the status quo will be seen as the result of this, that outcomes would have been different had the inputs been changed. But earnings, like other statistical processes, can simply be random. It is thus beneficial to find a criterion to distinguish what could be random from what can be shown to be significant.

While assessing significance is the goal of any statistical model, earnings or revenue have certain properties that lend themselves well to a time series model. The observations are serially correlated, since a company that did well last quarter can be expected to perform similarly in this quarter. (This also means that standard regression models are inadequate since their resulting error terms show varying degrees of correlation with other errors). Observations are also seasonal, with sales and other revenue corresponding to times of the year or month. They can also show a linear or quadratic trend or be influenced by exogenous variables that shift the data.

All of these properties make earnings or revenue a good candidate for modeling using the ARIMA procedure in SAS. There are a number of econometrics and time series procedures in SAS/ETS® software, but the structure of PROC ARIMA lends itself well to compartmentalizing the process of developing a time series model. Specifically, the procedure uses three major components:

- An IDENTIFY statement, which describes the serial correlation of the data along with any relations to external factors, and uses diagnostics to suggest an appropriate model.
- An ESTIMATE statement, which estimates the parameters of the model selected and tests for significance and goodness of fit.
- A FORECAST statement, which uses the model selected to predict subsequent observations and their corresponding confidence intervals.

The confidence intervals from these forecasts are of special significance with respect to corporate revenue. Using these, one can determine whether subsequent results are within an expected range. While results within this range are analogous to a lack of significance and thus do not prove no change occurred, results outside of expected ranges provide strong evidence that some change in the company affected revenue. That is, they can be used as a metric to show a yet to be determined cause of a difference in revenue.

APPLYING PROC ARIMA TO SMALL BUSINESS REVENUE FOR THE DELUXE CORPORATION

OBSERVING THE DATA

This process of modeling corporate revenue can be illustrated by using sales data obtained from the Deluxe Corporation. Deluxe sells business products such as checks, forms, and other supplies to small businesses around the United States and Canada. The landscape for selling these products frequently changes, leading management to wonder if changes in revenue are the result of an unknown factor or is simply business as usual.

The data obtained was U.S. sales revenue by week from 2006 through early 2012 over all product lines and all Deluxe corporate brands. Actual sales numbers were blinded by using a scale factor to express all revenue as a number between \$0.00 and \$1.00. The first step in the modeling process is to use PROC SGPLOT to graph the data. By using the SCATTER and SERIES statements in this procedure, one can produce a sequential graph that connects markers at each time point.

```
proc sgplot data=forecast.overall_week_normalized;
    scatter x=Week y=weekrev;
    series x=Week y=weekrev;
run;
```

The output for this statement can be seen in Figure 1.

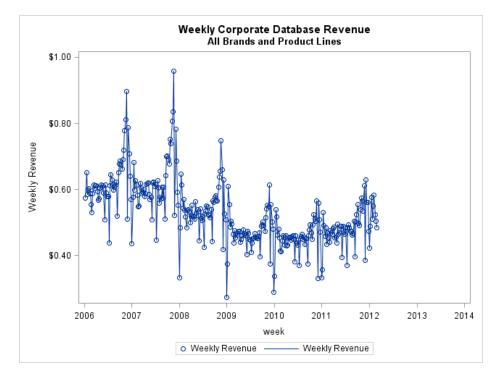


Figure 1. Output of PROC SGPLOT—Corporate Weekly Revenue from 2006 to 2012

A few properties are quickly apparent from this plot. There is a slight linear trend that indicates that revenue may have started to decrease around the time of the 2008 economic recession. Apart from this, each year taken in

isolation appears to show a U shape, with revenues high in January as business stock up on supplies, decrease towards the middle of the year, and then pick back up toward the end of the year as holiday related sales occur. There are also spikes in the data where revenue is much lower than normal. These generally correspond to holiday weeks where four day work weeks result in fewer orders.

THE IDENTIFY STATEMENT

The first step, then, in PROC ARIMA, is to use the IDENTIFY statement to determine candidate models that incorporate these trends. These trends are all accounted for using the CROSSCORR= option:

run;

The variables used in the CROSSCORR=option are

- time, denoting a linear longitudinal effect of time. This variable is represented in the forecast.overall_week_normalized dataset as a counter from 1 to 322, or equivalently the number of weeks since January 1, 2006.
- holiday, an indicator variable corresponding to whether the week was one of six holiday weeks (New Year's Day, Memorial Day, Independence Day, Labor Day, Thanksgiving or Christmas).
- Eleven indicator variables, Mon1 to Mon11, denoting whether the week began during the months of January through November. These variables model the annual seasonality of the data. A variable for the month of December is omitted to avoid collinearity but is included as part of the model's intercept when all of the other indicator variables equal zero. Variables must be listed explicitly since PROC ARIMA does not recognize shorthand notation such as MON1 – MON11.

With respect to the seasonal variables, many seasonal models use trigonometric functions such as $sin(2\pi t/S) + cos(2\pi t/S)$, where t is the time being modeled and S is the period of the model being fit. (Had it been used for this model, S would have equaled 52 since there are 52 weeks in a year). These regressors would be included, and significance tested, as a sine/cosine pair. Other models have used the classic Box-Jenkins approach of modeling $Z_{t} = Y_t - Y_{t \cdot S}$, as the dependent variable, again where S is the period of the model. (Applying this approach would have resulted in replacing var=weekrev with var=weekrev(52) to denote the seasonality of the dependent variable. The time cross-correlation variable could also be removed by using var=weekrev(1 52) to substitute the differences of successive observations for a linear trend variable). However, for Deluxe's revenue data, better results were obtained using the monthly indicator variables.

Three other options are used in the IDENTIFY statement. The NLAG= option determines how far out to plot the autocorrelation function. A lag *m* of a time series observation is the observation that is *m* time periods removed from it. In this case since NLAG=12, observations are compared for autocorrelation to all observations within 12 weeks. The ESACF option uses the Extended Sample Autocorrelation Function to suggest autoregressive moving average (ARMA) models for the residuals after the cross-correlation factors are accounted for. The STATIONARITY= option requests the augmented Dickey-Fuller unit root tests for a specified series of lags. This tests for whether the resulting series is stationary after accounting for the other factors in the model.

Since this PROC ARIMA step uses the PLOTS=ALL option, a number of plots are supplied in the resulting output. (ODS GRAPHICS must be turned on in order to view these plots.) One of these is the "Trend and Correlation Analysis" found in Figure 2 that provides a four plot summary of the overall graph along with the autocorrelation (ACF), partial autocorrelation (PACF), and inverse autocorrelation (IACF) functions.

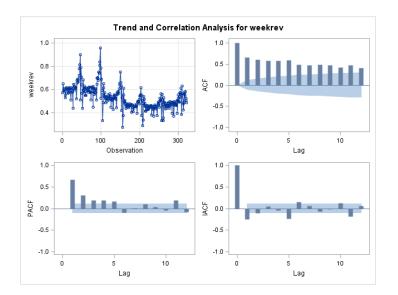


Figure 2. Trend and Correlation Analysis Provided by the IDENTIFY Statement

The overall graph largely replicates the previous output from PROC SGPLOT. The autocorrelation functions reveal a slowly dying ACF as well as a PACF with three and possibly four significant early lags. This suggests that an ARMA model of order 3 could potentially fit.

The ESACF option provides the output in Table 1 to the right. This output is used to suggest the candidate ARIMA (p,d,q) models. The top row suggests that a p+d equal to 3 and a q equal to 3 may provide the best fit. The d term in an ARIMA model, if greater than zero, results in a model where the *differences* of subsequent observations are modeled rather than their actual values. However, if a linear trend of time is included in the model, as is the case here, then the linear term serves the purpose of differencing the observations. Thus differencing is unnecessary and d = 0. This then suggests that an ARMA (3,3) model may provide the best fit.

ARMA(p+d,q) Tentative Order Selection Tests	
ESACF	
p+d	q
3	3
5	4
2	5

Table 1. ESACF Output

The STATIONARITY=(adf=(0 1 2 3 4 5)) option provides the following table (Table 2) to test for stationarity as a function of lags 0 through 5:

Augmented Dickey-Fuller Unit Root Tests							
Туре	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-3.56	0.1943	-1.37	0.1593		
	1	-1.64	0.3750	-0.99	0.2884		
	2	-1.03	0.4674	-0.80	0.3710		
	3	-0.69	0.5290	-0.68	0.4220		
	4	-0.52	0.5663	-0.62	0.4491		
	5	-0.56	0.5578	-0.62	0.4476		
Single Mean	0	-108.51	0.0001	-8.06	<.0001	32.46	0.0010
	1	-57.73	0.0016	-5.38	<.0001	14.46	0.0010
	2	-38.46	0.0016	-4.25	0.0007	9.05	0.0010
	3	-25.32	0.0029	-3.43	0.0111	5.88	0.0143
	4	-17.48	0.0203	-2.85	0.0528	4.09	0.0818
	5	-21.28	0.0078	-3.06	0.0312	4.70	0.0472
Trend	0	-160.64	0.0001	-10.31	<.0001	53.12	0.0010
	1	-99.24	0.0007	-6.98	<.0001	24.40	0.0010
	2	-72.52	0.0007	-5.64	<.0001	15.89	0.0010
	3	-49.74	0.0007	-4.55	0.0015	10.36	0.0010
	4	-34.54	0.0025	-3.74	0.0211	7.03	0.0297
	5	-47.00	0.0007	-4.13	0.0063	8.55	0.0010

 Table 2. Augmented Dickey-Fuller Tests Provided by the

 STATIONARITY= option

In the Dickey-Fuller unit root test, the null hypothesis for each lag is that the characteristic function of the autoregressive model for that lag has a single root on the unit circle, meaning that the process is not stationary. When a model is not stationary its residuals do not have a fixed mean or variance over time and should not be fit with an ARMA model. The alternative hypothesis is that the process is stationary.

Here virtually all of the p-values for the single mean test and the trend test have a p-value below 0.05. This suggests that the null can be rejected at all lags and that the process is stationary.

THE ESTIMATE STATEMENT

With this ARMA(3,3) model suggested, an ESTIMATE statement fits the model and performs diagnostic tests.

run;

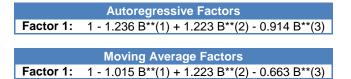
Since the ARIMA procedure is interactive, the ESTIMATE statement can follow the previous RUN statement without re-invoking the procedure. This resulting parameter estimates are given below in Table 3:

Maximum Likelihood Estimation								
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag	Variable	Shift	
MU	0.649	0.0326	19.89	<.0001	0	weekrev	0	
MA1,1	1.014	0.0681	14.91	<.0001	1	weekrev	0	
MA1,2	-1.223	0.0452	-27.06	<.0001	2	weekrev	0	
MA1,3	0.663	0.0692	9.59	<.0001	3	weekrev	0	
AR1,1	1.236	0.0363	34.08	<.0001	1	weekrev	0	
AR1,2	-1.223	0.0356	-34.37	<.0001	2	weekrev	0	
AR1,3	0.914	0.0321	28.46	<.0001	3	weekrev	0	
NUM1	-0.00047	0.000161	-2.93	0.0034	0	Time	0	
NUM2	-0.141	0.00702	-20.15	<.0001	0	holiday	0	
NUM3	-0.0129	0.0130	-0.99	0.3230	0	Mon1	0	
NUM4	-0.0556	0.0153	-3.66	0.0003	0	Mon2	0	
NUM5	-0.0426	0.0176	-2.42	0.0154	0	Mon3	0	
NUM6	-0.0563	0.0185	-3.05	0.0023	0	Mon4	0	
NUM7	-0.0483	0.0190	-2.55	0.0109	0	Mon5	0	
NUM8	-0.0349	0.0191	-1.83	0.0672	0	Mon6	0	
NUM9	-0.0334	0.0192	-1.77	0.0771	0	Mon7	0	
NUM10	-0.0517	0.0184	-2.81	0.0049	0	Mon8	0	
NUM11	-0.0124	0.0176	-0.70	0.4820	0	Mon9	0	
NUM12	-0.00178	0.0155	-0.11	0.9085	0	Mon10	0	
NUM13	0.0620	0.0127	4.88	<.0001	0	Mon11	0	
NUM13	0.0620		4.88		0	Mon11	0	

Table 3. Parameter Estimates from the ESTIMATE Statement

After an intercept term μ , the table starts with three moving average (MA) and three autoregressive (AR) parameters. All p-values for these parameters are significant, suggesting that their inclusion in the model is justified. The parameter estimate for time is negative with a p-value of 0.0034, confirming a linear decrease in overall sales. (Despite a possible curve in the revenue graph given above, an estimate of the quadratic effect for time was not significant and was not included in the model.) The estimate for holiday is highly significant, indicating that the model needs to account for these four day weeks represented by the variable. Without that term in the model, sales estimates for weeks following a holiday, being serially correlated, would be closer to the holiday week estimate even though the holiday condition no longer applies. The indicator variables MON1 through MON11 are not uniformly significant. However, many of them are significant, and a model using monthly indicator variables as seasonal factors needs to either include all of the terms, or include none of them. Since the variability attributed to all eleven variables, as a group, is significant using the Chi-Square test on 11 degrees of freedom, the terms are justifiably included in the model. The first ten parameters, for MON1 through MON10 are negative, due to the fact that the excluded MON12 variable corresponds to December, when sales are much higher than in the first ten months. Since these months are being compared to December their coefficients will therefore be negative. The MON11 parameter is actually positive; indicating that holiday related revenue is actually higher in November than in December.

The components of the ARMA (3,3) model are given above but can be summarized in tables 4 and 5. The B^{**}(n) notation used below is called the *backshift* notation, where B^{**}(n) * $y = y_{t-n}$.



Tables 4 and 5. Autoregressive and Moving AverageCoefficients

The residual correlation diagnostics for the model are given below in Figure 3:

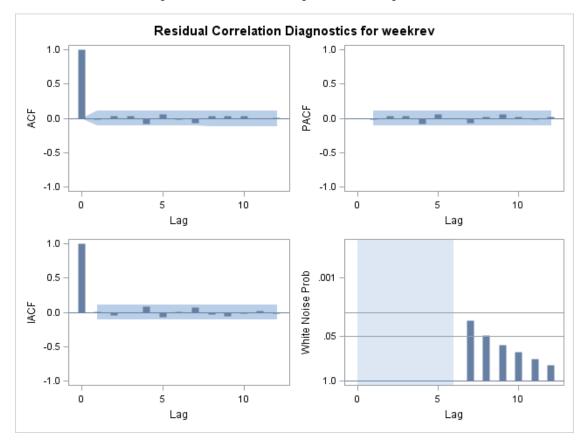


Figure 3. Residual Correlation Diagnostic Plots Provided by the ESTIMATE Statement

The first three charts show the autocorrelation functions for the residuals from the model, after accounting for outside variables and for the ARMA model. All of the lags past lag zero show spikes contained within the shaded error bars, suggesting no further autocorrelation to be explained. The final chart graphs white noise probability as a function of lag. In a pure white noise process, no observation is correlated with any other observation in the process, that is, all observations are independent. This plot graphs the p-values of the null hypothesis for each lag that the

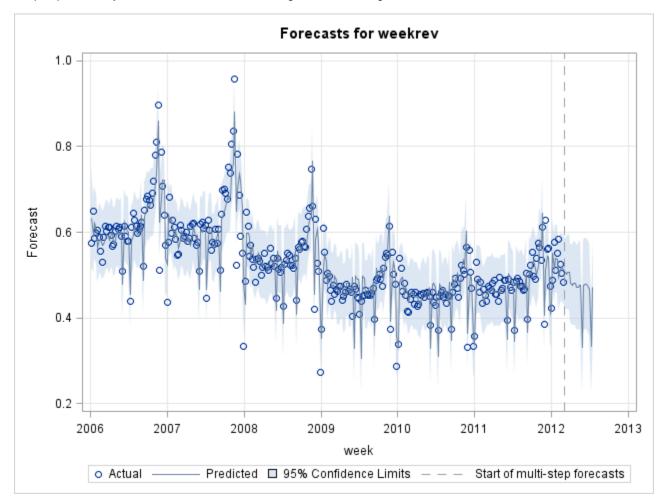
autocorrelations of lag *m* are not zero. The y-axis of this graph decreases from 1 to 0 as the graph ascends, indicating that values plotted above a corresponding p-value line have a low p-value and provide evidence of autocorrelation. In this case, the p-value for lag 7, the first lag to the right of the shaded region, is above the line for 0.05, providing a questionable lack of autocorrelation. However, if the only autocorrelation possibility is at lag 7 the concern is minimal and no other lags appear to be alarming. The model appears to be adequate for describing weekly sales revenue.

THE FORECAST STATEMENT

Once the model is fit, forecasts for future weeks can be obtained. These are provided from the FORECAST statement.

```
forecast lead=20 id=Week interval=week out=Work.arma33;
run;
```

The LEAD= option specifies how long to carry out the forecasts, 20 weeks. The ID= option provides the date variable in the input data set, *Week*, that identifies the time periods associated with the observations. The INTERVAL= option specifies that the observations are weekly. The OUT= option sends the forecasts, as well as standard errors and 95% confidence intervals to the output data set *Work.arma33*.



Graphs provided by the FORECAST statement are given below in Figure 4:

Figure 4. Forecasts and Confidence Intervals Provided by the FORECAST Statement

The dashed vertical line in the forecast plot corresponds to March 2012, where the model ends and the forecasts begin. The model forecasts a slight decrease in sales in the upcoming weeks as December and January fall into the past, with slight upticks being seen due to variability in the monthly indicator variables. Two sharp decreasing spikes can be seen that correspond to the 2012 Memorial Day and Independence Day holidays.

COMPARING FORECASTS WITH FUTURE REVENUE

The above forecasts are accompanied by the shaded zone representing the confidence intervals. Actual earnings, once obtained can be placed into the graphs to determine whether the values were within an expected error range. Once this is done, it becomes apparent that forecast error bars cover sales revenue within a wide range. To illustrate this, the 2012 results of the output *Work.arma33* dataset from the FORECAST statement executed above were merged with their corresponding estimates and actual values from 2011 to provide Table 6.

Week in 2012	Actual Weekly Revenue	FORECAST	STD	L95	U95	RESIDUAL	Estimated Value from 2011	Actual Value from 2011
1/2/2012	0.423	0.418	0.040	0.340	0.497	0.004	0.341	0.357
1/9/2012	0.488	0.550	0.040	0.472	0.629	-0.062	0.484	0.530
1/16/2012	0.577	0.517	0.040	0.438	0.595	0.061	0.490	0.491
1/23/2012	0.511	0.538	0.040	0.460	0.617	-0.028	0.489	0.459
1/30/2012	0.551	0.537	0.040	0.458	0.615	0.014	0.487	0.483
2/6/2012	0.584	0.495	0.040	0.417	0.574	0.088	0.450	0.433
2/13/2012	0.525	0.514	0.040	0.435	0.592	0.011	0.439	0.468
2/20/2012	0.503	0.518	0.040	0.439	0.597	-0.015	0.449	0.453
2/27/2012	0.484	0.513	0.040	0.435	0.592	-0.029	0.457	0.440
3/4/2012		0.508	0.040	0.430	0.587		0.464	0.475
3/11/2012		0.502	0.042	0.419	0.585		0.462	0.468
3/18/2012		0.506	0.044	0.419	0.593		0.466	0.464
3/25/2012		0.506	0.046	0.417	0.595		0.471	0.482
4/1/2012		0.481	0.047	0.390	0.573		0.457	0.487
4/8/2012		0.476	0.048	0.382	0.570		0.462	0.454
4/15/2012		0.480	0.049	0.383	0.577		0.460	0.456
4/22/2012		0.480	0.050	0.382	0.578		0.462	0.438
4/29/2012		0.472	0.051	0.372	0.571		0.457	0.495
5/6/2012		0.471	0.052	0.371	0.572		0.467	0.465
5/13/2012		0.476	0.052	0.373	0.578		0.467	0.470
5/20/2012		0.476	0.053	0.372	0.579		0.469	0.489
5/27/2012		0.331	0.053	0.227	0.435		0.351	0.393
6/3/2012		0.475	0.054	0.370	0.581		0.473	0.488
6/10/2012		0.479	0.054	0.373	0.586		0.502	0.490
6/17/2012		0.479	0.055	0.372	0.586		0.496	0.472
6/24/2012		0.472	0.055	0.365	0.579		0.483	0.466

Table 6. Estimates and Forecasted Revenue Values for 2012

Values provided are actual and forecasted revenue, along with standard errors, lower and upper 95% confidence intervals for the forecasted values, and model residuals. The table provides actual revenue until March 2012, when the modeling concludes and modeled values are replaced by the forecasts. One can tell from the table that for March and April 2012 the forecasted values are generally higher than the estimates or actual values from 2011. This is likely due to the model adjusting to revenue coming in better than expected in January and February 2012, and the resulting serial correlation with future values. Conversely, forecasted values for June are lower than the 2011 estimates, due to higher than expected May 2011 revenue that the model does not see repeating for 2012.

By isolating one row of the table, one can determine what size of increase would be needed to exceed the expectations determined by the model. This is illustrated by comparing the model estimate from early May 2011 against the forecast for early May 6, 2012 in Table 7.

Comparison of Sales Revenue May 9, 2011 vs. May 6, 2012

Date	Estimate/Forecast	Actual Observed Value	Lower Forecast 95% Confidence Interval	Upper Forecast 95% Confidence Interval
09May11	0.467	0.465		
06May12	0.471		0.371	0.572
			Increase of the upper 95% limit over the 2011 estimate	22.5%

Table 7. Comparison of Forecasts with Actual Sales Revenue

The upper confidence limit from the May 2012 estimate is 23% higher than the estimated model value from the year before, mostly due to the variability of the forecast. This means that a real sales revenue increase, even as high as 15% to 20%, would be well within expectations even though it may surprise corporate management. Providing an ARIMA model for previously observed revenue can assist in keeping a perspective in judgment should revenue increases or decreases occur.

It should be noted, however, that while observed values outside of their 95% confidence intervals provide evidence of a real change, the cause of the change is still unclear. Since the model contains only trend, seasonal, and holiday factors, it makes no claims about the influence of factors unrelated to time. Corporate management can hypothesize about causes such as management initiatives, price increases, or changes in the economic climate, but the model itself only informs as to when the change occurred. Outliers are inherently difficult to explain.

Estimates such as the ones provided above can also help guide business decisions. Revenue expectations can allow companies to allocate resources efficiently and to fine-tune budgets. Once the seasonality of this revenue is modeled, marketing campaigns can be coordinated around high revenue periods, or conversely could be scheduled to address expected lows.

Another potential application of the ARIMA models and forecasts involves assessing the effect of isolated events such as natural disasters, economic events, or national elections. The ARIMA procedure allows for indicator variables to assess a one-time impact which confines a revenue change to a brief period, or to act as a trigger for a ramping or decaying temporary effect. The use of these variables is known as *transfer function modeling*, but even if this modeling is not performed, the existing estimates and confidence intervals for the original model can provide a context to determine whether these events provided an anomalous impact on revenue.

CONCLUSION

Many times series and estimation tools exist within SAS, with many effective ways of analyzing a particular time series. PROC ARIMA, however, is especially useful due to its transparency in providing a division of labor of the three main components of Box-Jenkins modeling methodology—identification, estimation, and forecasting. It also provides a straightforward method of introducing cross-correlation variables that attribute exogenous sources of variability into a model. In particular, these variables can be linear trend, holiday indicator, or seasonal variables that provide an alternative to standard methods of accounting for seasonality.

When this approach is applied to corporate revenue, a good fit can be obtained due to this seasonality of revenue. The modeling provides estimates and confidence intervals for forecasts, allowing future observations to be seen in a context reflecting the uncertainty of these estimates. Observations that exceed their confidence intervals can be taken as strong evidence that a real change has occurred to the revenue environment.

REFERENCES

- SAS Institute, Inc. (2004) SAS/ETS® 9.1 User's Guide. Cary, NC: SAS Institute Inc.
- Dickey, David A. and Woodfield, Terry. (2011) Forecasting Using SAS® Software: A Programming Approach: Course Notes. Cary, NC: SAS Institute Inc.
- SAS Institute, Inc. (1996) Forecasting Examples for Business and Economics using SAS®. Cary, NC: SAS Institute Inc.
- Brocklebank, John C. and Dickey, David A. (2003) SAS® for Forecasting Time Series, Second Edition. Cary, NC: SAS Institute Inc.
- Box, George E.P., Jenkins, Gwilym M., and Reinsel, Gregory C. (2008) *Time Series Analysis: Forecasting and Control, Fourth Edition.* Hoboken, NJ: John Wiley & Sons, Inc.

ACKNOWLEDGMENTS

The authors would like to acknowledge Mark Antiel and the Deluxe Research and Decision Science team for the paper concept and may helpful suggestions.

CONTACT INFORMATION

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

Brian Van Dorn Deluxe Corporation 3680 Victoria Street North Shoreview MN 55126 (651) 787-1283 (Work) brian.vandorn@deluxe.com Saveth Ho Deluxe Corporation 3680 Victoria Street North Shoreview, MN 55126 (651) 490-8087 (Work) saveth.ho@deluxe.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 registered trademarks or trademarks of their respective companies