

## Forecasting Cross-Sectional Time Series : A Data Mining Approach Using Enterprise Miner™ Software

John Brocklebank , Taiyeong Lee , and Michael Leonard  
SAS Institute Inc., Cary, NC

### ABSTRACT

The practice of developing predictive models on large volumes of data is classified as data mining. Cross-sectional time series manifests itself in cases where data for different departments, locations, part or SKU numbers are collected over time. The target or response measures may include product counts, sales, and a variety of inputs or independent pieces of information that are associated with the target. The inputs are classified as either deterministic (for example month or year where the future values are known) or nondeterministic (for example interest rates or price, where future values are not known and must be forecasted or estimated from the data). The cross-sectional nature comes from the fact that you have separate time series (departments for instance) and you wish to provide a modeling method that estimates correlations across separate strata and the combined series.

This paper examines the robust nature of a future release of Enterprise Miner for handling large volumes of cross-sectional time series data.

### INTRODUCTION

Data mining techniques have seen a tremendous increase in their applicability recently. The biggest challenges have been related to the creation of a meaningful data model and the iterative process of exercising the data warehouse transformation mechanisms. The most common analytic methods include decision trees, regression, neural networks and segmentation methods. Enterprise Miner™'s ability to handle "dirty data" has also assisted in the development of meaningful and optimal results.

Time dependent data is very common in practice and usually manifests itself with several related components like different products, locations, SKU number etc. Economists and statisticians have faced challenges associated with processing vast amounts of time series data long before the concept of data mining became in vogue. A wide class of analytic methods support techniques to handle these problems ranging from simple forecasting methods like exponential smoothing to very advanced multivariate methods like Kalman Filtering and Statespace modeling. Data mining without forecasting is tantamount to knowledge discovery without SAS software. This concept has also been reinforced by Thearling (1998) who documents the need to integrate forecasting methods into the data mining selection of offerings.

### DATA MODEL AND ADAPTIVE STRUCTURE

#### VARIABLE DEFINITIONS

1. **Target** : This represents the variable to be predicted. Typically the target variable is a univariate time series and can often be better characterized by other factors or variables.
2. **Inputs** : Input variables consists of two types of variables, deterministic variables such as an intercept, seasonal dummies, trend variables, and class variables. Class variables are automatically transformed into dummy variables. The other type of variable is non-deterministic. Non-deterministic variables are univariate time series variables, which need to be extended using a forecasting method.
3. **Time id** : This variable specifies time period for each observation.
4. **Cross section id** : This variable specifies a cross section variable such as department, location, SKU number etc.

#### DATA MINING PROCEDURES

Two procedures, PROC DMFOR and PROC AUTOREG will be integrated into a time series modeling node in Enterprise Miner™ in a future release . PROC DMFOR automatically creates dummy variables for the class variables, estimates embedded missing values, and provides extensions for non-deterministic variables. PROC AUTOREG fits regression models within each cross section with time series errors.

#### METHODOLOGY

##### Step 1. Simple forecasting methods :

The simple forecasting methods provide exponential smoothing models with optimized smoothing weights and automatic model selection. This method estimates missing values and also extends non-deterministic variables used in the forecasting process. The following smoothing models are considered and measured for optimality.

- 1) Simple exponential smoothing
- 2) Double exponential smoothing
- 3) Linear exponential smoothing
- 4) Damped-trend linear exponential smoothing
- 5) Seasonal exponential smoothing
- 6) Winters method ( Additive version)
- 7) Winters method ( Multiplicative version)

##### Step 2. Linear regression with time series errors :

This technique estimates and forecasts linear regression models for time series data when the errors are autocorrelated and/or heteroscedastic. Stepwise autoregression is used to automatically select the autoregression error model.

- The parameter estimation methods include ;
  - 1) Yule-Walker estimates
  - 2) Unconditional least squares estimates

- 3) Maximum likelihood estimates
- 4) Iterative Yule-Walker estimates
- This provides a separate model for each cross section.

1994	1486.70	5593.60	2226.30	AA
1995	.	4989.18	2675.10	AA
1996	.	5045.91	3123.89	AA
1975	26.63	290.60	162.00	BB
1976	23.39	291.10	174.00	BB
.	.	.	.	.
.	.	.	.	.
1994	49.34	474.50	468.00	BB
1995	.	476.56	472.35	BB
1996	.	476.56	491.42	BB
1975	24.43	138.00	100.20	CC
1976	23.21	200.10	125.00	CC
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

**EXAMPLE 1. NO TIME SERIES INPUT VARIABLES.**

- 1) **Data** : A cosmetic company collected 2 years of monthly sales data ( JAN1996 to DEC1997) for each SKU. The variables represent monthly sales, SKU and dates. The SKU represents 5 shades of eye-shadows. They want to forecast them individually and jointly.
- 2) **Individual Analysis** : For all SKUs, the best forecasting model is the additive version of Winters method and was selected based on mean absolute percent error (MAPE).
- 3) **Joint Analysis** : The company also wants to have a model for all the SKUs. They would like to forecast monthly sales of eye-shadows combined. Monthly sales were also summed across SKUs (cross sectional units). The additive version of Winters method was also selected based on mean absolute percent error (MAPE). Table 1 shows MAPEs for all the final models.

Table 1. Mean Absolute Percent Error

SKU	1	2	3	4	5	joint
MAPE	8.35	8.92	6.42	8.00	6.78	5.84

Table 2. Predicted Values and 95 % Confidence Intervals

SKU	LCL	Pred.(Jan98)	UCL	LCL	Pred.(Feb98)	UCL
1	25714	39147	44076	18287	27483	36678
2	1288	2263	3237	1850	2824	3799
3	48334	62982	77630	51160	65831	80501
4	3583	4560	5537	3224	4204	5183
5	14791	19244	23696	17148	21819	26491
Joint	98477	122462	146447	96622	120678	144734

\* LCL(UCL) : Lower (Upper) 95 % confidence limits

**EXAMPLE 2.TIME SERIES INPUT VARIABLES**

- 1) **Data** : A partial history of a gross investment data set is displayed below. The data set contains a cross section variable which has four companies (AA, BB, CC, and DD), a target variable( Y-gross investment ), and two input variables (X1-lagged value of the firm and X2-lagged capital stock of the firm). The records from 1975 through 1994 are used for modeling. The data table shows an extended data set including the original data set. That is, for X1 and X2, the observations of 1995 and 1996 are extended using simple forecasting methods.

Year	Y	X1	X2	Company
1975	317.60	3078.50	2.80	AA
1976	391.80	4661.70	52.60	AA
1977	410.60	5387.10	156.90	AA
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
1993	1304.40	6241.70	1777.30	AA

- 2) **Analysis** : To obtain a forecasting model in each cross section, a stepwise autoregression error model is used, The stepwise autoregression method initially fits a high-order model (13 by the default) and then sequentially removes autoregressive parameters until all remaining autoregressive parameters are statistically significant.
  - a) Selection of significant autoregressive lags : The initial order of AR model is 13 and the significance level 0.1 is used as the elimination criterion. Table 3 shows the selected autoregressive error terms. An autocorrelation correction is not required for BB and CC companies.

Table 3. The selected significant autoregressive lags

	Significant AR Lags			
AA	Lag	Coefficient	Std Error	t Ratio
	5	0.52	0.2137	2.428
BB	NONE			
CC	NONE			
DD	Lag	Coefficient	Std Error	t Ratio
	2	0.39	0.1925	1.999
	3	0.49	0.1925	2.533

- b) After the error autocorrelation structure is determined, maximum likelihood estimates are used to construct a forecasting model in this example. Three additional estimation methods are available, which are described in step 2 of the methodology section.
- c) Improvements in R-squares and mean squares errors after adjusting for autocorrelation in error are shown in Table 4.

Table 4. Model MSE and R-squares

Company ⇒	AA	BB	CC	DD
MSE w/o error correction	8423	82.79	88.67	104.3
MSE w. error correction	2970	*	*	39.4
R-sq w/o error correction	0.921	0.666	0.764	0.903
R-sq w. error correction	0.976	*	*	0.941

\* Autocorrection corrections are not needed.

- d) Table 5. shows predicted values of two future years, 1995 and 1996, and their 95 % confidence intervals from the estimated models.

Table 5. Predicted Values and 95 % Confidence Intervals

for Years, 1995 and 1996.

Email: sasml@wnt.sas.com

Company	LCL	Pred. (1995)	UCL	LCL	Pred. (1996)	UCL
AA	1347	1485	1623	1541	1687	1834
BB	45.30	66.99	88.67	46.63	68.55	90.47
CC	55.81	78.37	100.94	58.24	81.09	103.9
DD	57.61	73.07	88.53	73.28	88.90	104.52

\*\* LCL(UCL) : Lower (Upper) 95 % confidence limits.

- e) The prediction intervals for 1995 and 1996 do not take into account the variability associated with predicting futures for the input variables X1 and X2.

## CONCLUSION

These examples demonstrates how Enterprise Miner™ will handle time series data sets with cross sections. The forecasting node in Enterprise Miner™ generates dummy variables and extends future observations for non-deterministic input variables for the purposes of forecasting. The forecasting node provides a predictive modeling mechanism based on linear regression models with autocorrelated errors or heteroscedastic errors. You can also perform several kinds of exponential smoothing technique for a single time series. Finally, you can obtain predicted values and their corresponding confidence intervals to assist users in making business decisions. The forecasting node will be provided in a future release of Enterprise Miner™.

## REFERENCES

Thearling Kurt (1998) : Some thoughts on the current state of data mining software applications, A workshop paper held in conjunction with KDD'98.  
 SAS/ETS User's Guide Version 6, Second Edition.

## CONTACT INFORMATION

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

Author Name : John C. Brocklebank, Ph.D.  
 Company : SAS Institute Inc.  
 Address : Business Solutions Division  
 City state ZIP : Cary, NC 27513  
 Work Phone: 919-677-8001 ext. 7360  
 Fax: 919-677-4444  
 Email: sascjb@wnt.sas.com

Author Name : Taiyeong Lee, Ph.D.  
 Company : SAS Institute Inc.  
 Address : Business Solutions Division  
 City state ZIP : Cary, NC 27513  
 Work Phone: 919-677-8001 ext. 2186  
 Fax: 919-677-4444  
 Email: tailee@wnt.sas.com

Author Name : Michael Leonard  
 Company : SAS Institute Inc.  
 Address : SAS/ETS  
 City state ZIP : Cary, NC 27513  
 Work Phone: 919-677-8001 ext. 6967  
 Fax: 919-677-4444