

Paper 340-2012

## Variable Selection for Multivariate Cointegrated Time Series Prediction with PROC VARCLUS in SAS® Enterprise Miner™ 7.1

Akkrapol Sa-ngasoongsong and Satish T.S. Bukkapatnam  
Oklahoma State University, Stillwater, OK, USA

### ABSTRACT

Vector Error Correction Model (VECM) has recently become a popular tool for economic analysis and forecasting for multivariate cointegrated time series. However, one problem of this type of model is over-parameterization issue. Traditional method to address this problem is to impose weak exogeneity assumption on variables. Assuming unknown structural relationship among variables, imposing this assumption alone may not be sufficient, especially in case of large number of hypothesized variables. This paper presents a variable selection method for multivariate cointegrated time series prediction using variable clustering procedure (PROC VARCLUS) in SAS® Enterprise Miner™ 7.1. The empirical results show that long-run equilibrium relationship among variables selected by variable clustering procedure can be reasonably identified. The comparison of forecasting performance with other classical time series techniques demonstrates a significant improvement on a prediction accuracy.

### INTRODUCTION

The automobile industry is one of the largest industries within U.S. manufacturing sector. For automobile manufacturers, increasing competition amongst manufacturers has been accelerated since U.S. economic crisis in 2008 when the U.S. Big Three<sup>1</sup> automakers requested government aid to relieve their financial problems [1]. To revive the companies and also remain competitive in market, this has forced automobile companies to devise various strategies to help them overcome competition in the industry. Not only optimal positioning of new products is required, but also controlling activities, such as demand planning, are necessary, to effectively manage resources and maximize revenue. Considering dynamic environment, demand planning is challenging and important. Often times, errors in demand planning have led to enormous costs and loss of revenues. Hence, accurate demand or sales forecasting is vital to a successful strategic planning.

Automobile sales forecasting has received significant attention in the literature since 1970 [2, 3]. Many of the automobile sales forecasting models proposed [4-6] are econometric approaches imposing a certain structure of economic theory on the data. Also, most of the prior forecasting techniques assume short-term forecasting horizons. They are inefficient for effective demand planning which is required long-term prediction. Very few efforts have been made to address forecasting problem for long-term prediction [7, 8]. In the area of forecasting research in economics, some developments in multivariate time series techniques [9, 10] have been specifically designed to quantify long-run impact of related variables to variables of interest. These models are Vector autoregressive (VAR) and Vector error correction model (VECM)[11, 12]. They have been broadly recognized as a powerful theory-driven model that can be used to describe long-run dynamic behavior of multivariate time series. Especially, for multivariate cointegrated nonstationary time series, VECM has theoretically been proven to provide an identification of long-run equilibrium interrelationships among variables in the system.

However, there is some disadvantage of implementing these types of model in practice. A well known problem of VAR and VECM is an over-parameterization issue [13, 14] which is a prohibitively large number of parameters to be estimated. One way to address this problem is imposing theory-based weak exogeneity assumption on variables. The number of equations in the model can be reduced if variables are treated as weakly exogenous in the model. However, imposing the test alone may not be sufficient in case of large scale datasets. Recently, in the field of data mining, significant efforts [15, 16] have been made to address the issue of excessive number of correlated factors. Many dimensional reduction and variable selection techniques have been proposed to solve the problem. In case of dimensional reduction techniques, although they are useful to retain significant portion of explained variance with reduced number of factors, interpretations of the results are no longer straightforward because the components from dimensional reduction techniques are a combination of all of the original variables. It is not easily explicable, and may not be applicable in the context of econometrics. Considering the disadvantage of dimensional reduction techniques, variable selection methods have gained significantly more attention for economists in the sense that the results from variable selection methods can be used and explained directly. This paper proposes a utilization of variable clustering technique (PROC VARCLUS) in SAS® Enterprise Miner™ 7.1 to solve an over-parameterization issue in VAR and VECM models (PROC VARMAX) for automobile sales forecasting. The organization of this paper is as follows: Data section provides brief description of each hypothesized variable for automobile sales prediction. In methodology

---

<sup>1</sup> The U.S. Big Three automakers consist of General Motors, Ford and Chrysler.

section, three-stage methodology for multivariate cointegrated time series, including variable clustering procedure, is presented in details. The next section is implementation details and results section. This section provides empirical results on the methodology described in a previous section. Conclusion is presented in the last section of this paper.

## DATA

The main time series in this paper is the number of monthly retail sales in U.S. (Motor vehicle and parts dealers) during a period of 1992-2011. The dataset consists of automobile sales and eleven economic and related indicators<sup>2</sup>. These variables are hypothesized to have relationship with sales. The details of variables are summarized as shown in Table 1. This paper will investigate and also test hypotheses of the causal relationships and cointegration among sales and selected variables using variable clustering technique (PROC VARCLUS) in SAS® Enterprise Miner™ 7.1.

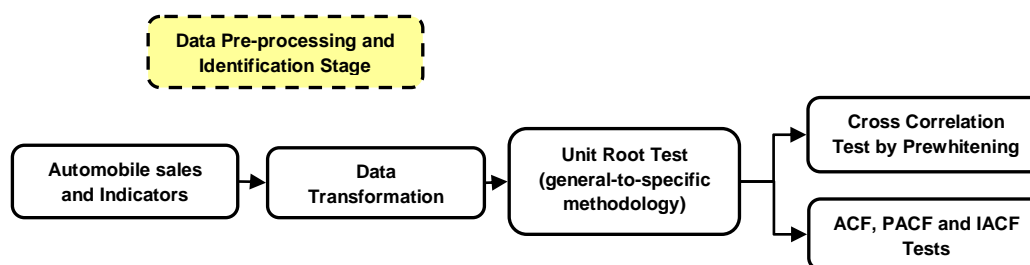
Variables	Source	Description
Sales	BLS*	A monthly retail sales of motor vehicles and part accessories
Current Personal Finance	UMICH**	Current financial situation compared to a year ago
Expected Personal Finance	UMICH**	Expected change in financial situation
Business Condition in 12 Months	UMICH**	Business conditions expected during the next 12 months
Business Condition in 5 Years	UMICH**	Business conditions expected during the next 5 years
Buying Conditions (1)	UMICH**	Buying conditions for vehicles
Buying Conditions (2)	UMICH**	Buying conditions for large household goods
Housing Starts	HUD***	Total new privately owned housing units started
Consumer Price Index (CPI)	BLS*	A monthly data on changes in the prices paid by urban consumers
Unemployment Rate	BLS*	A national unemployment rate (16 years or over)
Employment-Population Ratio	BLS*	A proportion of the country's working-age population that is employed
Gas Prices	EIA****	A monthly U.S. city average retail price (all types of gasoline)

\* Bureau of Labor Statistics, \*\* University of Michigan, \*\*\* Department of Housing and Urban Development, \*\*\*\* U.S. Energy Information Administration

**Table 1. Summary of Variables**

## METHODOLOGY

The methodology in this paper is consisting of the following three stages. The first stage is data pre-processing and identification. This stage of methodology is used to pre-process the data and also identify characteristic of each variable in the dataset for subsequent analysis. In the second stage of variable and model structure selection, the results from the 1<sup>st</sup> stage of methodology and variable clustering technique (PROC VARCLUS) will be used for variable selection. This procedure is used to avoid over-parameterization problem in model parameterization stage (3<sup>rd</sup> stage) by selecting variables in each group (cluster). Then, weak exogeneity assumption will be imposed on selected variables, and model structure (VAR or VECM) will be chosen based on cointegration and granger causality tests. The final stage of methodology is model parameterization and forecasting. The model and variables selected from second step will be validated using out-of-sample data. The comparison of model prediction accuracy with other time series techniques will be done to test forecasting performance of the model. The details of each stage of methodology are as follows:



**Figure 1: Data Pre-processing and Identification Stage**

<sup>2</sup> Five variables from University of Michigan (excluding Buying Condition (1)) are main components of index of consumer sentiment (ICS).

1.) Data Pre-processing and Identification: In the first stage, it is used to prepare data for an analysis, and also identify characteristic of each variable. The procedure in this stage begins with normalization of each variable to attain a comparability among variables. Since same order of integration is required for all endogenous variables in VECM model, stationary condition and integration order of each variable will be identified using unit root test. Cross correlation test by prewhitening technique serves as a useful tool to reveal the relation of between sales and other factors in the dataset. Autocorrelation (ACF), Partial Autocorrelation (PACF) and Inverse Autocorrelation (IACF) serve as useful tools to identify significant lags order to model a filter (Autoregressive Integrated Moving Average, ARIMA) for the prewhitening procedure. Figure 1 presents techniques and hypothesis tests required for the first stage of methodology. The details of each technique and hypothesis test in the 1<sup>st</sup> stage of methodology are as follows:

Data Transformation: Since variables have different range, this can create significant numerical errors when we compare the effect of each variable after modeling. A normalization technique can be used to weigh all variables in the same way. Assuming that variables in the dataset follow Gaussian distribution, the z-Transformation can be obtained by

$$Y_{it} = \frac{X_{it} - u(X_i)}{\sigma(X_i)} \sim N(0,1) \quad (1)$$

where  $Y_{it}$  = z – Transformation of variable  $X_i$  at a particular time  $t, t \in T$  and  $i = \{1, \dots, 11\}$

$u(X_i)$  is the mean and  $\sigma(X_i)$  is the standard deviation of  $X_i$

This process can be performed in SAS using PROC STANDARD [See Appendix A] where mean and variance are specified as shown in Equation (1).

Unit Root Test: To identify the nonstationary condition of variables, SAS offers multiple unit root tests, such as, Phillips-Perron test, a random-walk with drift test, augmented Dickey-Fuller test (ADF) etc, in ARIMA procedure (PROC ARIMA) [See Appendix A]. In this paper, the ADF test is selected to identify nonstationary condition of variables due to its statistical power. There are three main versions of the test as shown in the Equations 2-4.

$$\Delta y_t = \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \dots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t \quad (2)$$

$$\Delta y_t = \alpha + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \dots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t \quad (3)$$

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \dots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t \quad (4)$$

The null hypothesis is  $\gamma = 0$  against the alternative hypothesis of  $\gamma < 0$ . Equation (2) is the test for a unit root with zero mean. Equation (3) is the test for a unit root with drift and Equation (4) is a unit root test with drift and deterministic time trend. The appropriate lag length for ADF test is selected using the general-to-specific methodology [17]. The idea is to select optimal lag ( $p^*$ ) that is significantly difference from zero for the  $p$ th autoregressive process shown in Equations 1-3. The methodology starts with a lag length of  $p^*$  and then pares down the model by the usual  $t$ -test. If the  $t$ -statistic on lag  $p^*$  is insignificant at some specified critical value, then re-estimate the regression using a lag length of  $p^*-1$ . Repeat the process until the last lag is significantly different from zero. In this autoregressive case, this procedure yields the true lag length with an asymptotic probability of unity, provided the initial choice of lag length including the true length. An alternative approach is to examine the information criteria such as the Akaike information criterion (AIC), Bayesian information criteria (BIC) or the Hannan-Quinn information criterion (HIC).

Cross-Correlation by Prewhitening: One way to identify whether lag(s) of one variable have an effect to the current period of another variable is to find the cross-correlation. Consider the following generalization of the transfer function model:

$$y_t = a_0 + A(L)y_{t-1} + C(L)z_t + B(L)\varepsilon_t \quad (5)$$

where  $A(L), B(L),$  and  $C(L)$  are polynomials in the lag operator  $L$ .  $\{z_t\}$  and  $\{\varepsilon_t\}$  are uncorrelated white noise processes.

The cross-correlation between  $y_t$  and  $z_{t-i}$  is defined to be

$$\rho_{yz}(i) = \frac{cov(y_t, z_{t-i})}{\sigma_y \sigma_z} \quad (6)$$

where  $\sigma_y$  and  $\sigma_z$  = the standard deviation of  $y_t$  and  $z_t$ , respectively

Practically,  $z_t$  is rarely white noise process. To obtain the pattern of the coefficients in  $C(L)$ , the appropriate methodology is to filter the  $\{y_t\}$  sequence with the estimated polynomial  $D(L)/E(L)$  where  $D(L)$  and  $E(L)$  are defined as

$$D(L)z_t = E(L)\varepsilon_{zt} \quad (7)$$

This method is called prewhitening. The filtered value of  $y_t$ ,  $D(L)y_t/E(L)$ , is defined as

$$\frac{D(L)y_t}{E(L)} = \frac{D(L)a_0}{E(L)} + \frac{D(L)A(L)y_{t-1}}{E(L)} + \frac{C(L)D(L)z_t}{E(L)} + \frac{B(L)D(L)\varepsilon_t}{E(L)} \quad (8)$$

Given that  $D(L)y_t/E(L) = y_{ft}$ ,  $D(L)y_{t-1}/E(L) = y_{ft-1}$ , and  $D(L)z_t/E(L) = \varepsilon_{zt}$ , so (8) is equivalent to

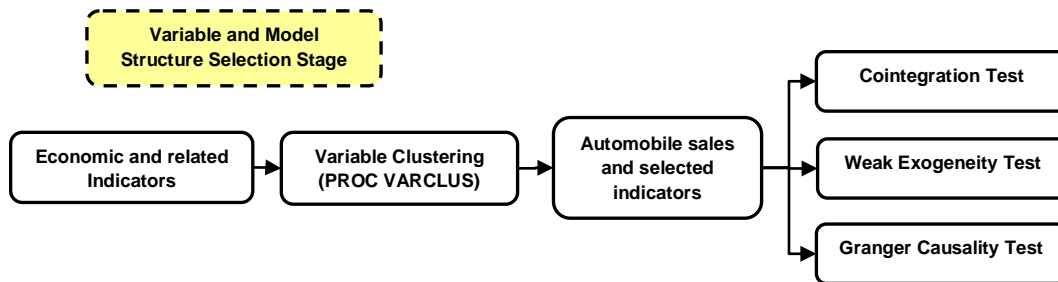
$$y_{ft} = \frac{D(L)a_0}{E(L)} + A(L)y_{ft-1} + C(L)\varepsilon_{zt} + \frac{B(L)D(L)\varepsilon_t}{E(L)} \quad (9)$$

SAS can perform this prewhitening process automatically using PROC ARIMA [See Appendix A]. The results of this cross-correlation using prewhitening technique is used to reveal the relation of between sales and other factors in the dataset.

ACF, PACF and IACF Tests: Box and Jenkins [18] have described the sample autocorrelation (ACF), partial autocorrelation (PACF) and inverse autocorrelation (IACF) as useful tools in identifying and estimating time series models. These three tools can be performed automatically in PROC ARIMA [See Appendix A]. The results of these tests will be used to identify the filter (ARIMA model) for cross-correlation by prewhitening.

## 2.) Variable and Model Structure Selection Stage

The second stage of methodology is used to select variables and model structure for subsequent analysis. As discussed in the introduction section, one problem of VAR and VECM models is over-parameterization issue. Large number of parameters to be estimated is a well known problem for this type of model. One way to address this problem is to test and impose weak exogeneity assumption [19]. However, imposing the test alone may not be sufficient in case of large number of hypothesized variables. This paper utilizes the enhanced method for variable selection in SAS® Enterprise Miner™ 7.1. The results of variable selection (VARCLUS Procedure) will be used to select the model structure. Three hypothesis tests (Cointegration, Weak Exogeneity and Granger Causality Tests) will be tested. These three tests can be performed using SAS VARMAX procedure. The details of SAS codes for each test can be found in Appendix A. Figure 2 presents a framework for the second stage of methodology. The details of variable selection procedure and hypothesis tests are as follows



**Figure 2: Variable and Model Structure Selection Stage**

VARCLUS Procedure: The algorithm of VARCLUS procedure in SAS begins with dividing numeric variables into disjoint hierarchical clusters via a type of oblique principal components analysis. Initially, all variables is in a single cluster. It then splits the cluster by finding the first two principal component (PCs), doing an oblique rotation of the PCs and then assigning each variable to the PC with which it is most strongly correlated. The process will be continued until default criteria is reached (Eigen value greater than 1 for second PC). Domain knowledge is required to select a representative variable in each group (cluster). Alternatively, SAS provides the ratio for each variable in each cluster This ration is calculated as follows:

$$1 - R_{ratio}^2 = \frac{1 - R_{own\ cluster}^2}{1 - R_{next\ closet\ cluster}^2} \quad (10)$$

A small value of this ratio indicates that a variable has strong correlation with variables in its own cluster buy weak correlation with variables in other clusters. Therefore, small values of this ratio is desirable to select representative variable out of each cluster.

Weak Exogeneity Test: As mentioned earlier, imposing weak exogeneity test can be used to avoid over-parameterization problem. The test of weak exogeneity will identify the weak exogeneity effect of each variable to the others. The number of equations in the model can be reduced if the variables are treated as weakly exogenous. To test which variables should be treated as endogenous in the equation, and which ones as exogenous, the  $k$ -vector of  $I(1)$  random variables  $y_t$  is initially partitioned into the  $k_1$ -vector  $y_{1t}$  and the  $k_2$ -vector  $y_{2t}$ , where  $y_t = (y'_{1t}, y'_{2t})$  and

$k = k_1 + k_2$ . From the VECM model (See 3<sup>rd</sup> stage of Methodology), the parameters can similarly be decomposed as  $\delta = (\delta'_1, \delta'_2)$ ,  $\alpha = (\alpha'_1, \alpha'_2)$ ,  $\Phi_i^* = (\Phi_{1i}^*, \Phi_{2i}^*)$  and the variance-covariance matrix as

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \tag{11}$$

The conditional model for  $y_{1t}$  given  $y_{2t}$  is

$$\Delta y_{1t} = \omega \Delta y_{2t} + (\alpha_1 - \omega \alpha_2) \beta' y_{t-1} + \sum_{i=1}^{p-1} (\Phi_{1i}^* - \omega \Phi_{2i}^*) \Delta y_{t-i} + (\delta_1 - \omega \delta_2) + (\varepsilon_{1t} - \omega \varepsilon_{2t}) \tag{12}$$

and the marginal model for  $y_{2t}$  is

$$\Delta y_{2t} = \alpha_2 \beta' y_{t-1} + \sum_{i=1}^{p-1} \Phi_{2i}^* \Delta y_{t-i} + \delta_2 + \varepsilon_{2t} \tag{13}$$

where  $\omega = \Sigma_{12} \Sigma_{22}^{-1}$

The test of weak exogeneity of  $y_{2t}$  determines whether  $\alpha_2 = 0$ .

**Cointegration Test:** Engle and Granger [11] show that if, a linear combination of nonstationary time series is stationary, the time series are cointegrated. Cointegrated processes are processes that are random in the short term but tend to move together in the long term. Vector Error Correction Model (VECM) should be considered if variables are cointegrated. For a test of cointegration, the Johansen's reduced rank methodology [20] is employed. Two test statistics are suggested to test the null hypothesis that there are at most  $r$  cointegrating vectors ( $H_0: \lambda_i = 0$  for  $i = r + 1, \dots, k$ ). The trace and maximum eigenvalue statistics are as follows:

$$\lambda_{trace} = -T \sum_{i=r+1}^k \log(1 - \lambda_i) \tag{14}$$

$$\lambda_{max} = -T \log(1 - \lambda_{r+1}) \tag{15}$$

where  $\lambda$  is the eigenvalue in the Johansen's reduced rank regression model.

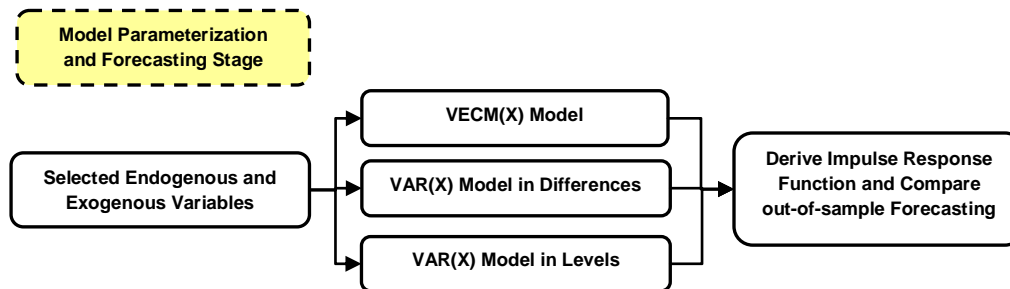
**Granger Causality Test:** Let  $y_t$  be arranged and partitioned in subgroups  $y_{1t}$  and  $y_{2t}$  with dimensions  $k_1$  and  $k_2$ , respectively ( $k = k_1 + k_2$ ): that is,  $y_t = (y'_{1t}, y'_{2t})'$  with the corresponding white noise process  $\varepsilon_t = (\varepsilon'_{1t}, \varepsilon'_{2t})'$ . Considering the bivariate VAR( $p$ ) model with partitioned coefficients  $\Phi_{ij}(B)$  for  $i, j = 1, 2$  as follows

$$\begin{bmatrix} \Phi_{11}(B) & \Phi_{12}(B) \\ \Phi_{21}(B) & \Phi_{22}(B) \end{bmatrix} \begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} \tag{16}$$

The variable  $y_{1t}$  is said to cause (Granger)  $y_{2t}$ , but  $y_{2t}$  do not cause (Granger)  $y_{1t}$  if  $\Phi_{12}(B) = 0$ . This model structure implies that if  $\Phi_{12}(B) = 0$ ,  $y_{1t}$  is influenced only by its own past values and not by the past of  $y_{2t}$ . Consider testing  $H_0: C\beta = c$ , where  $C$  is a  $s \times (k^2p + k)$  matrix of rank  $s$  and  $c$  is an  $s$ -dimensional vector where  $s = k_1 k_2 p$ . Assume that  $\sqrt{T}(\hat{\beta} - \beta) \rightarrow N(0, \Gamma_p^{-1} \otimes \Sigma)$ . The Wald statistic can be obtained from

$$\sqrt{T}(C\hat{\beta} - c)[C(\hat{\Gamma}_p^{-1} \otimes \hat{\Sigma})C']^{-1}(C\hat{\beta} - c) \rightarrow \chi^2(s) \tag{17}$$

### 3.) Model Parameterization and Forecasting Stage



**Figure 3: Model Parameterization and Forecasting Stage**

As discussed at the beginning of methodology section, we have used VAR & VECM classes of multivariate linear models to model sales and related indicators. From the model structure and variables selected in the second

stage, the third stage of methodology is to estimate and validate the model using out-of-sample data. Figure 3 presents a framework for this stage. The details of VAR and VECM are as follows:

VAR and VECM Models: A  $p$ th-order vector autoregression, denoted as  $VAR(p)$ , can be written as

$$y_t = c + \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \dots + \Phi_p y_{t-p} + \varepsilon_t \quad (18)$$

where  $c$  denotes an  $(n \times 1)$  vector of constants and  $\Phi_j$  denotes an  $(n \times n)$  matrix of autoregressive coefficients for  $j = 1, 2, \dots, p$ . The  $(n \times 1)$  vector  $\varepsilon_t$  is a vector with  $\Omega$   $(n \times n)$  symmetric positive definite matrix. For stationary assumption of VAR models, the stationary condition is satisfied if all roots of  $|\Phi(z)| = 0$  lie outside the unit circle. If the stationary condition is not satisfied, a nonstationary model (a differenced model or an error correction model) might be more appropriate. The vector error correction model with the cointegration rank  $r (\leq k)$ , denoted as  $VECM(p)$ , can be written as

$$\Delta y_t = \delta + \Pi y_{t-1} + \sum_{i=1}^{p-1} \Phi_i^* \Delta y_{t-i} + \varepsilon_t \quad (19)$$

where  $\Delta$  is differencing operator, such that  $\Delta y_t = y_t - y_{t-1}$ ;  $\Pi = \alpha\beta'$ , where  $\alpha$  and  $\beta$  are  $k \times r$  matrices;  $\Phi_i^*$  is a  $k \times k$  matrix. The cointegrating vector  $\beta$ , is also called the long-run parameter, and  $\alpha$  is the adjustment coefficient.

## IMPLEMENTATION DETAILS AND RESULTS

Data Pre-processing and Identification: The results of unit root test on transformed variables are shown in Table 2. All variables are stationary after first differencing (1<sup>st</sup> order integration,  $I(1)$ ). The lags order of ARIMA model of each variables have been selected, based on ACF, PACF, IACF and residual diagnostic. These ARIMA models are used as filter for prewhitening process prior to identifying the cross-correlation. The results on cross-correlation show that only six variables have significant cross-correlation with sales. These variables are Housing Starts, CPI, Unemployment Rate, Gas Prices, Buying Conditions (1) and Current Personal Finance. Significant lags of each variable are presented in Table 2.

Variables	Transformed Variables	Stationary Test	Filtering Model	Cross Correlation by Prewhitening (Significant Lags)
Sales	Stnd_Sales	I(1)	ARIMA(2,1,0)	-
Housing Starts	Stnd_HS	I(1)	ARIMA(3,1,1)	-1
Consumer Price Index (CPI)	Stnd_CPI	I(1)	ARIMA(3,1,0)	3,4
Unemployment Rate	Stnd_UP	I(1)	ARIMA(1,1,1)	-1
Employment-Population Ratio	Stnd_EP	I(1)	ARIMA(3,1,0)	-
Gas Prices	Stnd_GP	I(1)	ARIMA(3,1,0)	1,2
Buying Conditions (1)	Stnd_BC	I(1)	ARIMA(1,1,1)	0
Current Personal Finance	Stnd_C1	I(1)	ARIMA(2,1,0)	-1,0,4
Expected Personal Finance	Stnd_C2	I(1)	ARIMA(4,1,0)	-
Business Condition in 12 Months	Stnd_C3	I(1)	ARIMA(4,1,1)	-
Business Condition in 5 Years	Stnd_C4	I(1)	ARIMA(1,1,1)	-
Buying Conditions (2)	Stnd_C5	I(1)	ARIMA(3,1,0)	-

**Table 2: Data Pre-Processing and Identification**

Variable and Model Structure Selection: In the second stage, PROC VARCLUS was used to cluster eleven variables to a certain number of cluster based on clustering algorithm. The results show that optimal number of clusters is three. The proportion of variance explained by these three clusters is 84.19% as shown in Table 3. Figure 4 shows variables in each of the three clusters. In order to select variables for subsequent analysis, the cross-correlation and  $1 - R_{ratio}^2$  criteria are considered together. From six variables in the first cluster (CLUS1), only two (Buying Condition (1) and Current Personal Finance) have significant cross-correlation with sales. The  $1 - R_{ratio}^2$  criteria of these variables are not significantly different. Hence, both variables are selected to represent all variables in CLUS1. For CLUS2, both variables in CLUS 2 also have significant cross-correlation with sales. The  $1 - R_{ratio}^2$  of these variables is quite low (0.0724 and 0.0904), and not significantly different. Both variables are selected to represent variables in CLUS2. In CLUS3, Housing Starts and Unemployment Rate have significant cross-correlation with sales, however, the  $1 - R_{ratio}^2$  of Unemployment Rate is significantly lower than that of Housing Starts. Therefore, only Unemployment rate is selected to represent variables in CLUS3.

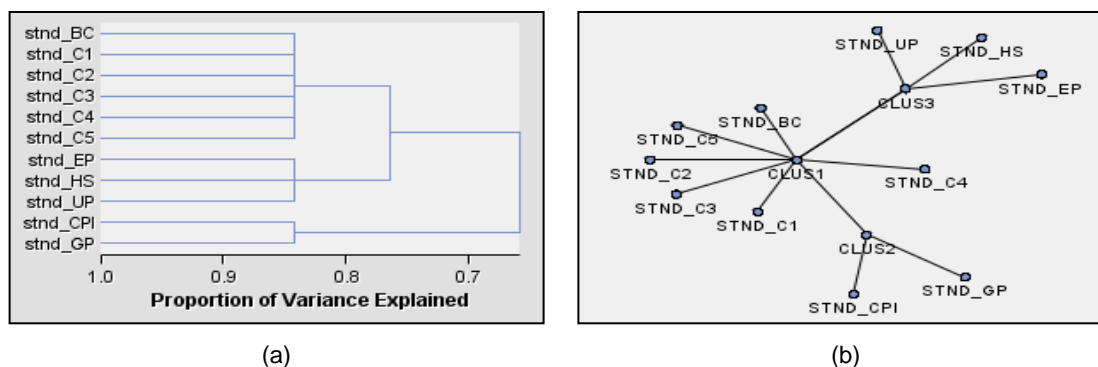


Figure 4: (a) Proportion of Variance Explained by Number of Clusters and (b) Cluster Plot

Cluster Summary					
Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	6	6	4.740105	0.7900	0.5859
2	2	2	1.893345	0.9467	0.1067
3	3	3	2.626988	0.8757	0.3396
Total variation explained = 9.260439    Proportion = 0.8419					

Cluster	Variable	R-squared with		$1 - R_{ratio}^2$
		Own Cluster	Next Closest	
Cluster 1	Stnd_BC	0.5474	0.2427	0.5976
	Stnd_C1	0.8441	0.7261	0.5963
	Stnd_C2	0.8431	0.4793	0.3014
	Stnd_C3	0.8878	0.3790	0.1807
	Stnd_C4	0.8166	0.4116	0.3117
	Stnd_C5	0.8011	0.5363	0.4289
Cluster 2	Stnd_CPI	0.9467	0.2632	0.0724
	Stnd_GP	0.9467	0.4101	0.0904
Cluster 3	Stnd_EP	0.9237	0.5345	0.1638
	Stnd_HS	0.7594	0.4628	0.4479
	Stnd_UP	0.9438	0.4542	0.1029

Number of Clusters	Total Variation Explained by Clusters	Proportion of Variation Explained by Clusters	Minimum Proportion Explained by a Cluster	Maximum Second Eigenvalue in a Cluster	Minimum R-squared for a Variable	Maximum $1 - R_{ratio}^2$ for a Variable
1	7.249686	0.6591	0.6591	1.438778	0.3806	
2	8.412705	0.7648	0.7244	1.090574	0.4284	0.7548
3	9.260439	0.8419	0.7900	0.585862	0.5474	0.5976

Table 3: Cluster Summary

The next procedure in this stage of methodology is to impose weak exogeneity assumption on selected variables. Each candidate exogenous variable was tested with unrestricted cointegration rank. The results are shown in Table 4(a). The null hypothesis of weak exogeneity cannot be rejected for CPI. In contrast, weak exogeneity for Sales, Gas Prices, Unemployment Rate, Buying Condition (1) and Current Personal Finance is strongly rejected at less than 1% significance level. Retesting weak exogeneity with only rejecting weakly exogenous variables confirms that five variables (Table 4(b)) are not weakly exogenous for each other variable. The cointegration test on selected five variables (Table 4(b)) indicates that there is a potential of two cointegrating vectors among five variables. The underlying processes of these variables are random in the short term but tend to move together in long term horizon. The results of the cointegration test are presented in Table 5. Granger Causality test variables and results are shown in Table 6. Each causality test is testing the null hypothesis that variables in Group 1 cause variables in Group 2, but variables in Group 2 do not cause variables in Group 1. For Example, test 1 tests the null hypothesis that sales

causes the other variables (Gas Prices, Unemployment Rate, Buying Condition (1) and Current Personal Finance), but other variables do not cause sales. The results show that the null hypothesis of all tests (1 to 5) is strongly rejected. These test results confirm that selected variables can be used as endogenous variables in the model. For model structure selection, VECM model with one exogenous variable (CPI) is selected due to a potential of cointegrating vectors among selected variables.

Testing Weak Exogeneity of Each Variables			
Variable	DF	Chi-Square	Pr > ChiSq
Stnd_Sales	5	34.24	<.0001
Stnd_GP	5	19.51	0.0015
Stnd_CPI	5	6.08	0.2984
Stnd_UP	5	51.89	<.0001
Stnd_C1	5	31.84	<.0001
Stnd_BC	5	27.29	<.0001

(a)

Testing Weak Exogeneity of Each Variables			
Variable	DF	Chi-Square	Pr > ChiSq
Stnd_Sales	4	37.62	<.0001
Stnd_GP	4	14.13	0.0069
Stnd_UP	4	35.41	<.0001
Stnd_C1	4	28.47	<.0001
Stnd_BC	4	26.86	<.0001

(b)

Table 4: (a) The weak exogeneity on selected five variables and sales and (b) The retesting weak exogeneity on sales and four variables (CPI was excluded)

Cointegration Rank Test Using Trace						
H0: Rank=r	H1: Rank>r	Eigenvalue	Trace	1% Critical Value	Drift in ECM	Drift in Process
0	0	0.2665	142.5700	76.37	Constant	Linear
1	1	0.1748	71.8948	53.91		
2	2	0.0904	28.0781	34.87		
3	3	0.0280	6.4691	19.69		
4	4	0.0000	0.0004	6.64		

Table 5: Cointegration Rank Test using Trace on selected variables.

Granger-Causality Wald Test								
Test	Group 1 Variables	Group 2 Variables				DF	Chi-Square	Pr > ChiSq
1	Stnd_Sales	Stnd_GP	Stnd_UP	Stnd_C1	Stnd_BC	16	50.35	<.0001
2	Stnd_GP	Stnd_Sales	Stnd_UP	Stnd_C1	Stnd_BC	16	30.85	0.0141
3	Stnd_UP	Stnd_GP	Stnd_Sales	Stnd_C1	Stnd_BC	16	64.76	<.0001
4	Stnd_C1	Stnd_GP	Stnd_UP	Stnd_Sales	Stnd_BC	16	74.76	<.0001
5	Stnd_BC	Stnd_GP	Stnd_UP	Stnd_C1	Stnd_Sales	16	75.90	<.0001

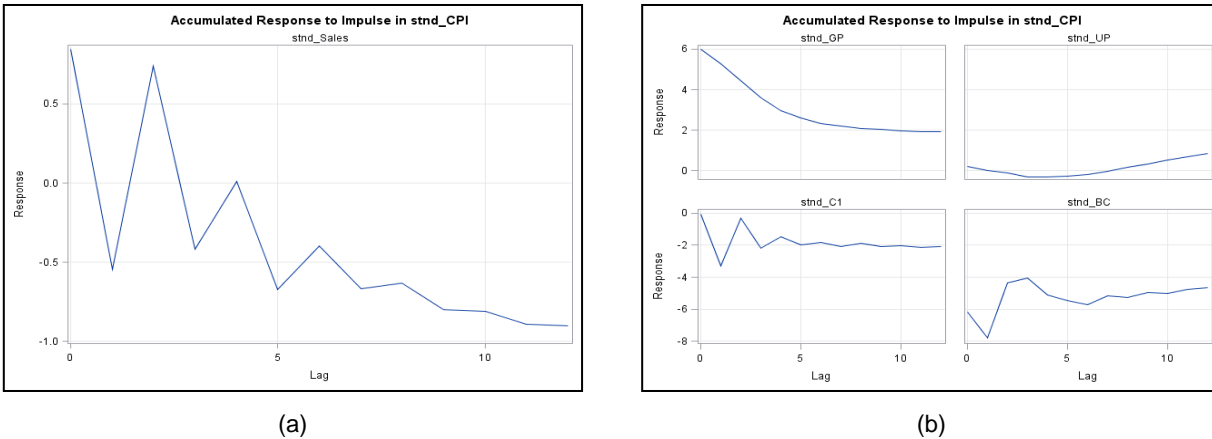
Table 6: Granger-Causality Wald Tests of selected variables

Forecasting Performance Comparison	
Model	RMSE (12-step ahead prediction)
ARIMA(2,1,0)	0.6102
ARIMAX	0.3682
VARX(4,4)	0.3207
VECMX(4,4)	0.2422

Table7: Model Comparison

Model Parameterization and Forecasting Stage: From the results in second stage, VECMX model with 5 endogenous (Sales, Gas Prices, Unemployment rate, Current Personal Finance and Buying Condition (1)) and 1 exogenous (CPI) variables is parameterized, using maximum likelihood estimation method. Cointegration rank for the model is set to two. Out-of-sample data are randomly selected to validate the VECMX model. Three classical time series models are selected to compare with VECMX model. They are ARIMA, ARMAX and VARX models as shown in Table 7. Considering on forecasting performance of sales prediction, VECMX model can improve a prediction accuracy by 60%, 34% and 24%, compared to ARIMA, ARIMAX and VARX models in terms of RMSE. These RMSEs values are quantified on 12-step ahead prediction of sales.





**Figure 5: (a) Accumulated Response of Sales to Impulse in CPI, (b) Accumulated Response of Sales to Impulse of Gas Prices, Unemployment Rate, Current Personal Finance and Buying Condition (1) variables.**

Figure 5(a) presents an accumulated response of sales on one unit change in CPI (Stnd\_CPI). At a one-period delay, automobile sales declines sharply, then the response approximately returns to its initial value at a period of two. The system shows an oscillating decay pattern with a downward sloping trend till a twelve-period delay. This may indicate that one unit increase in CPI tends to have negative impact on the automobile sales. Figure (b) shows accumulated responses of sales on one unit change in Gas Prices, Unemployment Rate, Current Personal Finance and Buying Condition (1) variables. As shown in Figure 5(b), Gas Prices tend to have temporarily negative impact on sales. The accumulated response tend to converge to some value after ten-period delay. One unit increase in Current Personal Finance and Buying Condition (1) variables also tend to have temporary effect on Sales.

## CONCLUSION

The advantage of VECM approach to model an automobile sales is that it provides a clear and quantifiable method to the long-run effect of selected variables. However, one well-known problem of this type of model is an over-parameterization. This paper utilizes variable clustering technique in SAS® Enterprise Miner™ 7.1 incorporating with traditional method to address this issue. The empirical results show that number of variables and parameters in the model can reduce dramatically using this technique. From eleven hypothesized variables, only five variables are selected as endogenous variables in VECM model (Sales, Gas Prices, Unemployment rate, Buying Condition for Vehicles and Current Personal Finance variables). Based on weak exogeneity test results, Consumer Price Index (CPI) is selected as exogenous variable in the model. The empirical results show that VECM model with selected endogenous and exogenous variables can significantly improve a prediction accuracy of automobile sales for long term prediction.

## APPENDIX A: SAS CODES

```
libname sas 'H:\';
run;

DATA WORK.sas;
  SET sas.sas;
  stnd_Sales = sales; LABEL stnd_Sales="Standardized Sales: mean = 0 standard deviation = 1";
  stnd_HS = HS; LABEL stnd_HS="Standardized HS: mean = 0 standard deviation = 1";
  stnd_CPI = CPI; LABEL stnd_CPI="Standardized CPI: mean = 0 standard deviation = 1";
  stnd_UP = UP; LABEL stnd_UP="Standardized UP: mean = 0 standard deviation = 1";
  stnd_EP = EP; LABEL stnd_EP="Standardized EP: mean = 0 standard deviation = 1";
  stnd_GP = GP; LABEL stnd_GP="Standardized GP: mean = 0 standard deviation = 1";
  stnd_BC = BC; LABEL stnd_BC="Standardized BC: mean = 0 standard deviation = 1";
  stnd_C1 = C1; LABEL stnd_C1="Standardized C1: mean = 0 standard deviation = 1";
  stnd_C2 = C2; LABEL stnd_C2="Standardized C2: mean = 0 standard deviation = 1";
  stnd_C3 = C3; LABEL stnd_C3="Standardized C3: mean = 0 standard deviation = 1";
  stnd_C4 = C4; LABEL stnd_C4="Standardized C4: mean = 0 standard deviation = 1";
  stnd_C5 = C5; LABEL stnd_C5="Standardized C5: mean = 0 standard deviation = 1";
RUN;

/*-- DATA NORMALIZATION --*/
PROC STANDARD
DATA=work.sas
OUT=WORK.sas
MEAN=0
```

```

        STD=1
        ;
        VAR stnd_sales stnd_HS stnd_CPI stnd_UP stnd_EP stnd_GP stnd_BC stnd_C1 stnd_C2 stnd_C3 stnd_C4 stnd_C5;
RUN;

*-- Augmented Dickey-Fuller Unit Root Tests, Autocorrelation(ACF), Partial Autocorrelation(PACF) and Inverse
Autocorrelation(IACF) --*/
/*-- Lag Length Selection has been tested using the general-to-specific methodology, and AIC, BIC criteria--*/
PROC ARIMA data=work.sas;
    identify var=stnd_sales stationarity=(adf=(4));
    identify var=stnd_HS stationarity=(adf=(4));
    identify var=stnd_CPI stationarity=(adf=(1));
    identify var=stnd_UP stationarity=(adf=(6));
    identify var=stnd_EP stationarity=(adf=(5));
    identify var=stnd_GP stationarity=(adf=(6));
    identify var=stnd_BC stationarity=(adf=(6));
    identify var=stnd_C1 stationarity=(adf=(6));
    identify var=stnd_C2 stationarity=(adf=(6));
    identify var=stnd_C3 stationarity=(adf=(6));
    identify var=stnd_C4 stationarity=(adf=(6));
    identify var=stnd_C5 stationarity=(adf=(6));
RUN;

/*--Cross-Correlation by prewhitening, selected lags order for each variable are presented in Table2 --*/
PROC ARIMA data=work.sas; /*
    identify var=stnd_HS(1);
    estimate p=2 q=1; /* Lags order are selected based on ACF, PACF, IACF and Residual diagnostic */
    identify var=stnd_sales(1) crosscorr=stnd_HS(1);
run;

/*-- Weak Exogeneity Test --*/
PROC VARMAX data=work.sas;
    model stnd_sales stnd_gp stnd_cpi stnd_up stnd_c1 stnd_bc/ p=4 ecm=(rank=5 normalize=stnd_sales);
    cointeg rank=5 exogeneity;
RUN;

/*-- Weak Exogeneity Test (Retest) --*/
PROC VARMAX data=work.sas;
    model stnd_sales stnd_gp stnd_up stnd_c1 stnd_bc/ p=4 ecm=(rank=4 normalize=stnd_sales);
    cointeg rank=4 exogeneity;
RUN;

/*-- Cointegration Rank Test --*/
PROC VARMAX data=work.sas;
    model stnd_sales stnd_gp stnd_up stnd_c1 stnd_bc/ p=4 cointtest=(johansen=(normalize=stnd_sales));
    cointeg rank=4 exogeneity;
RUN;

/*-- Granger Causality Test --*/
PROC VARMAX data=work.sas;
    model stnd_sales stnd_gp stnd_cpi stnd_up stnd_c1 stnd_bc/ p=4 ;
    causal group1=(stnd_sales) group2=(stnd_gp stnd_cpi stnd_up stnd_c1 stnd_bc);
    causal group1=(stnd_gp) group2=(stnd_sales stnd_cpi stnd_up stnd_c1 stnd_bc);
    causal group1=(stnd_cpi) group2=(stnd_sales stnd_gp stnd_up stnd_c1 stnd_bc);
    causal group1=(stnd_up) group2=(stnd_sales stnd_cpi stnd_gp stnd_c1 stnd_bc);
    causal group1=(stnd_c1) group2=(stnd_sales stnd_cpi stnd_up stnd_gp stnd_bc);
    causal group1=(stnd_bc) group2=(stnd_sales stnd_cpi stnd_up stnd_c1 stnd_gp);
RUN;

/*-- VECMX(4,4) and Impulse Response Function --*/
PROC VARMAX data=work.sas plot=impulse;
    model stnd_gp stnd_up stnd_c1 stnd_bc stnd_sales= stnd_cpi /p=4 noint xlag=4 lagmax =12 ecm=(rank=2
    normalize=stnd_sales) print=(impulse=(all) impulsx=(all));
    output lead=12;
run;

/*-- Model Comparison-ARIMA(2,1,0) --*/
PROC ARIMA data=work.sas;
    identify var=sales(1);
    estimate p=2;
    forecast lead=12;
run;

/*-- Model Comparison-ARIMAX --*/
PROC VARMAX data=work.sas;
    model stnd_sales= stnd_cpi /p=2 noint xlag=4 ;
    output lead=12;
run;

```

```

/*-- Model Comparison-VARX(4,4) --*/
PROC VARMAX data=work.sas;
model stnd_gp stnd_up stnd_cl stnd_bc stnd_sales= stnd_cpi /p=4 noint xlag=4 lagmax =12 ;
output lead=12;
run;

```

## REFERENCES

1. Freedman, J., *The U.S. Auto Industry: American Carmakers and the Economic Crisis*. 2011, New York: The Rosen Publishing Group, Inc.
2. James, B., *Forecasting automobile demand using disaggregate choice models*. Transportation Research Part B: Methodological, 1985. **19**(4): p. 315-329.
3. Mannering, F.L. and K. Train, *Recent directions in automobile demand modeling*. Transportation Research Part B: Methodological, 1985. **19**(4): p. 265-274.
4. Train, K., *Qualitative Choice Analysis: Theory, Econometrics, and an Application to Automobile Demand*. 1986, Boston: MIT Press series in transportation studies.
5. Suits, D.B., *Forecasting and Analysis with an Econometric Model*. The American Economic Review, 1962. **52**(1): p. 104-132.
6. Hess, A.C., *A Comparison of Automobile Demand Equations*. Econometrica, 1977. **45**(3): p. 683-701.
7. Hyndman, R.J. and F. Shu, *Density Forecasting for Long-Term Peak Electricity Demand*. Power Systems, IEEE Transactions on, 2010. **25**(2): p. 1142-1153.
8. Dong, R. and W. Pedrycz, *A granular time series approach to long-term forecasting and trend forecasting*. Physica A: Statistical Mechanics and its Applications, 2008. **387**(13): p. 3253-3270.
9. Johansen, S., *Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models*. Econometrica, 1991. **59**(6): p. 1551-1580.
10. Lütkepohl, H., *Comparison of Criteria for Estimating the Order of a Vector Autoregressive Process*. Journal of Time Series Analysis, 1985. **6**(1): p. 35-52.
11. Engle, R.F. and C.W.J. Granger, *Co-Integration and Error Correction: Representation, Estimation, and Testing*. Econometrica, 1987. **55**(2): p. 251-276.
12. Kim, K.H., *US inflation and the dollar exchange rate: a vector error correction model*. Applied Economics, 1998. **30**(5): p. 613-619.
13. Gary L, S., *Multiple cointegrating vectors, error correction, and forecasting with Litterman's model*. International Journal of Forecasting, 1995. **11**(4): p. 557-567.
14. Kim, J.H. and T. Ngo, *Modelling and forecasting monthly airline passenger flows among three major Australian cities*. Tourism Economics, 2001. **7**(4): p. 397-412.
15. Bi, J., et al., *Dimensionality Reduction via Sparse Support Vector Machines*. Journal of Machine Learning Research, 2003. **3**: p. 1229-1243.
16. Hartmann, W., *Dimension Reduction vs. Variable Selection: Applied Parallel Computing. State of the Art in Scientific Computing*. J. Dongarra, K. Madsen, and J. Wasniewski, Editors. 2006, Springer Berlin / Heidelberg. p. 931-938.
17. Enders, W., *Applied Econometric Time Series: 3rd Edition*. 2010: John Wiley & Sons, Inc.
18. Box, G.E.P., G.M. Jenkins, and G.C. Reinsel, *Time Series Analysis: Forecasting and Control (Wiley Series in Probability and Statistics)*. 2008.
19. Søren, J., *Testing weak exogeneity and the order of cointegration in UK money demand data*. Journal of Policy Modeling, 1992. **14**(3): p. 313-334.
20. Johansen, S. and K. Juselius, *Maximum Likelihood Estimation and Inference on Cointegration - with Applications to the Demand for Money*. Oxford Bulletin of Economics and Statistics, 1990. **52**(2): p. 169-210.

## CONTACT INFORMATION

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

Name: Akkarapol Sa-ngasoongsong  
 Enterprise: Oklahoma State University  
 Address: 245N. University Place#301  
 City, State ZIP: Stillwater, Oklahoma, 74075, USA  
 E-mail: [akkarap@okstate.edu](mailto:akkarap@okstate.edu)

Name: Satish T.S. Bukkapatnam  
 Enterprise: Oklahoma State University  
 Address: 322 Engineering North  
 City, State ZIP: Stillwater, Oklahoma, 74078, USA  
 E-mail: [satish.t.bukkapatnam@okstate.edu](mailto:satish.t.bukkapatnam@okstate.edu)

Akkarapol Sa-ngasoongsong is a PhD student in School of Industrial Engineering and Management at Oklahoma State University. He has three years of professional experience as R&D Engineer. He has earned two SAS certifications: SAS Certified Base Programmer for SAS® 9 and Certified Predictive Modeler using SAS® Enterprise Miner™ 6.1. He was 1<sup>st</sup> award winner of the M2010 conference's Data Mining Shootout.

Satish T.S. Bukkapatnam is AT&T professor in School of Industrial Engineering and Management at Oklahoma State University. His research is published in numerous academic journals. He was a recipient of Alpha Pi Mu/ Omega Rho Outstanding Teacher of the Year in Industrial Systems Engineering, and Outstanding Young Manufacturing Engineer Award from the Society of Manufacturing Engineers.

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.