

Variable Selection and Reduction of Time Series Data

Heidi Jensen
Yuanzhi Li
Andrew Worth

with the assistance of

Aaron Warshay

May 6, 2010

Abstract Multivariate time series are used for data mining and forecasting. An essential task is to determine the most relevant variables from a large set of data by variable reduction and/or variable selection to obtain maximum efficiency and accuracy for modeling. This project summarizes many methods of variable selection/reduction and investigates new methods in feature selection: two stage variable clustering and CLeVer. Their performances are compared based on computation time and chosen variables to similarity with VARCLUS, a technique currently in use by the Dow Chemical Company.

Table of Contents

Introduction	1
Literature Review	1
Traditional Data Mining Techniques	2
Similarity Analysis	3
Cointegration	4
K-Means Clustering	4
Two Stage Variable Clustering	5
CLeVer	7
Support Vector Machines	8
Corona	8
Method	8
Data	9
Analysis/Results	9
Test Data Set	9
Economic Data Set	10
Conclusions	12
Project in Phase II	13
Acknowledgements	14
References	15
Appendix A	16
Appendix B	20

Introduction

The Dow Chemical Company is a leading company in science and technology, and supplies more than 3,300 products to customers in 160 countries. Predicting future economic situations is invaluable for a company of this size and breadth. This can be done using data mining and modeling practices. Data mining involves searching for patterns in data using a variety of techniques. Modeling entails solving mathematical problems, usually something that will help make a strategic decision for a company. The Data Mining and Modeling (DMM) department at Dow uses tools from the popular statistical software program SAS, such as JMP and SAS Enterprise Miner for mining data. DMM has successfully completed projects regarding optimization, prediction and forecasting and many others.

Many economic activities can be predicted by economic time series data such as unemployment or consumer spending. However, in determining which of the many economic factors are good predictors of future economic activity can involve investigating hundreds or even thousands of time series variables. Traditional data mining techniques to select or reduce these variables can be extremely time consuming.

Dow is interested in developing a superior method to carry out variable selection and reduction. In particular, methods of *variable reduction* using similarity analysis, principal component analysis (PCA) and cluster analysis as well as methods of *variable selection* using similarity analysis and traditional variable selection methods for non-time series data mining.

This project seeks to determine the best approach for time series variable selection and reduction by examining and comparing traditional and more progressive avenues. The methods described in this project are considered for contexts when there are a great number of variables. Specifically, they can be used for data mining in the realm of forecasting future economic activity based on many different economic time series data.

Literature Review

Many pieces of literature regarding data mining and variable reduction and selection are available. These writings were considered when choosing a method to implement. Among the topics researched were traditional data mining techniques, similarity analysis and cointegration. Although many methods of variable selection and reduction are available, only some of them have known applications for time series data.

It may be important to note in the following descriptions that *variable selection* refers to a supervised process, that is, the target variable is considered in the process of choosing representative variables. *Variable reduction* is an unsupervised process, in other words, the target is not considered in this process.

Traditional Data Mining Techniques

Traditional data mining uses a variety of techniques for variable selection and reduction. Among these are decision trees, simple correlation, partial least squares regression, variable selection nodes, stepwise regression assuming a polynomial function, and genetic programming.

Decision trees are used for classification. For variable selection, rules are made to divide the variables into subsets. This kind of model can be helpful because the rules make the model very easy to understand and explain [1].

Simple correlation is a measure of how closely two items are related to each other. The closeness of the correlation is measured by the correlation coefficient. A perfect correlation will have a correlation coefficient of 1.

Partial least squares regression expands a multiple linear regression model to a multivariate situation. It is less restrictive than many other extensions of this model [2].

There are two kinds of variable selection nodes, R-square and Chi-square. Chi-square is used for a binary target. Inputs are selected by making a tree. R-square can be used for a binary target or an interval scaled target. The R-square between all of the inputs and the target are calculated. A model can then be constructed by using the input with the highest correlation. Then the next most valuable variables in terms of highest correlation with the target can be added in one by one. The process will stop when the inputs have a correlation below a certain threshold [3].

Stepwise regression is similar to the variable selection nodes method in respect to how variables are chosen. In stepwise forward regression, the input variable that is most correlated with the output variable is selected to first create a model. The input variable that has the most correlation with the residuals of this new model is the next one to be chosen. This selection continues until the remaining input variables do not have a significant correlation with the residuals of the most recent model. Stepwise backward regression starts with all of the input variables in the model. Then variables are expunged on the basis of the t value to determine which of the variables is the least significant [4].

Genetic programming searches for the best parameters for a predefined model. The best inputs are chosen by an evolution of possible solutions that converge to the best possible solution [4].

Following the traditional approach, Dow's DMM would use a variety of methods stated above to select or reduce variables and then use the union of the variables chosen from each method. This is very time consuming. Their newer methods of similarity analysis and cointegration are better, which are discussed next.

Similarity Analysis

Similarity measures the distance between two sets of time series data while considering the ordering of the data. Comparisons can occur in an unsupervised approach between two input series, or in a supervised approach with an input and a target series. These approaches can include a comparison to several inputs or targets as well. Dow currently uses both supervised and unsupervised similarity and then uses VARCLUS, a variable clustering function in SAS, to select and or reduce variables.

When comparing two sets of time series data, A and B, a similarity measure is chosen to measure the distance between each element in A and each element in B which is used to create a similarity matrix. Figure 1 shows a direct path through a similarity matrix between an input and a target of time series data. The path goes from the lower left corner, the similarity measure between the first elements of each time series, to the upper right, the similarity measure between the last elements of time series data. A diagonal movement through the matrix is considered a direct route. However, unless the matrix is square, there must be some vertical or horizontal movements which are described as compressions and expansions. The compression and expansion path limits are considered before choosing a path to traverse through the matrix. The goal is to find the path with the least cost.

				(5,5)
			(4,4)	
		(3,3)		
	(2,2)			
(1, 1)				

Figure 1. A similarity matrix where the direct path is from the lower left corner to the upper right, through the labeled entries.

When the best path is chosen, there are many path statistics that can be computed. These describe some of the features of the path taken such as the direct, compression and expansion maps on the path [5].

Cointegration

Cointegration is a technique introduced in the 1980s to treat non-stationary time series. Often, economic variables such as wealth and consumption do not show a typical linear regression, i.e. they do not correspond to a stationary model. In fact, it was shown that using stationary models to analyze non-stationary time series would lead to spurious results: one might find a correlation where it does not exist, producing errors in data forecasting. The discovery that some economic variables are “co-integrated,” that is, the linear combination of two nonstationary time series can be stationary, resolved this issue. Much economic work in the last 30 years has been to develop the theory of co-integrated variables. The task is twofold: first, to find variables that are co-integrated, and second, to use this new analysis to interpret the results [6].

Cointegration creates a regression model that can be tested to see if the residuals of the model are stationary. This can be done using the Dickey-Fuller Test, which will give evidence of whether a model is stationary or not. It identifies whether a unit root is present, which would indicate that the model is non-stationary. This is an improvement from the traditional approach which makes each series stationary to see if they are related, which can result in a loss of information about the long term relationship in the series. Dow currently uses cointegration as a supervised method to select variables.

K-Means Clustering

Clustering variables into several groups is a good way to reduce or select variables. K-means clustering is one of popular algorithms for clustering time series data, attempting to find the groups of data sets which have similar characteristics . The basic principle behind K-means clustering is to partition objects into K clusters so that the distance within a cluster is minimized. Also, the choice of how many initial cluster centers to use impacts on the quality of results. The algorithm for K-means clustering is outlined in Table 1 below.

Table 1. The k-means clustering algorithm [7].

<p>The K-Means Algorithm</p> <p>Step 1. Specify the number of clusters, K.</p> <p>Step 2. Select K initial cluster centers.</p> <p>Step 3. Assign N objects to the nearest cluster center and decide the class memberships.</p> <p>Step 4. Recalculate the K cluster centers by assuming the memberships found above are correct.</p> <p>Step 5. Stop and exit, if none of the N objects changed membership in the last iteration. Otherwise, go to step 3.</p>

For the economic data from the Dow Chemical Company, this could mean clustering the input variables into 20-50 clusters according to the results of similarity method. Then choose K initial cluster centers randomly. Then assign the 1,700 input variables to the nearest random initial centers,

which gives us 25-50 clusters, then re-estimate the cluster centers, reassign variables until these variables will not change the clusters.

Since the choices of initial centers will affect the clustering result, the K-means algorithm can be used in conjunction with methods like PCA similarity factors [8].

Two Stage Variable Clustering

In addition to the variable clustering procedure currently used by Dow and the previously mentioned K-means clustering, there is a proposed method that combines the speed of observational clustering with the accuracy of variable clustering. This method is referred to as two-stage variable clustering. The first stage consists of using fast observational clustering techniques to create global clusters and then variable clustering is performed on each global cluster, creating sub clusters. The principal components of these sub clusters comprise the reduced variable set. The algorithm for two stage variable clustering is shown in Table 2.

Table 2. Two-stage variable clustering algorithm [9].

Stage 1: Variable clustering based on a distance matrix

1. Calculate the correlation matrix of the variables.
2. Apply a hierarchical clustering algorithm to the correlation matrix.
3. Using a predefined cluster number, cluster variables into homogeneous groups.
The cluster number is generally no more than the integer value of $(nvar/100+2)$.
These clusters are called global clusters.

Stage 2: Variable clustering based on latent variables

1. Run PROC VARCLUS with all variables within each global cluster as you would run a single-stage, variable clustering task.
2. For each global cluster, calculate the global cluster components, which are the first principal component of the variables in its cluster.
3. Create a global cluster structure using the global cluster components and the same method as 1 at Stage 2.
4. Form a single tree of variable clusters from 1 and 3.

One of the major reasons for considering this method, is that the unsupervised portion of the variable reduction is the most time consuming step. The clustering procedure currently in use by Dow, PROC VARCLUS, is not very scalable with large data sets, as the memory requirements are proportional to the square of the number of variables. The two stage clustering implemented in this study first uses PROC FASTCLUS, which has memory requirements that are proportional to the number of variables. Table 3 shows the memory formulas for VARCLUS and FASTCLUS. Figure 2

demonstrates the difference in computer memory requirements for the two methods. In this figure, the number of variables is taken to be 1,750 and the desired number of clusters is 50, where the two stage method creates 10 global clusters and 5 sub-clusters within each global cluster. The memory values for the two stage method are calculated assuming that each global cluster has an equal number of elements, which is not necessarily realistic but with large data sets the two stage method will still be faster.

Table 3. Memory requirements in bytes for SAS clustering procedures used [9].

<p>VARCLUS: $v^2 + 2vc + 20v + 15c$</p> <p>FASTCLUS: $4(19v + 12cv + 10c + 2\max(c+1,v))$</p> <p>v = number of variables, c = number of clusters</p>
--

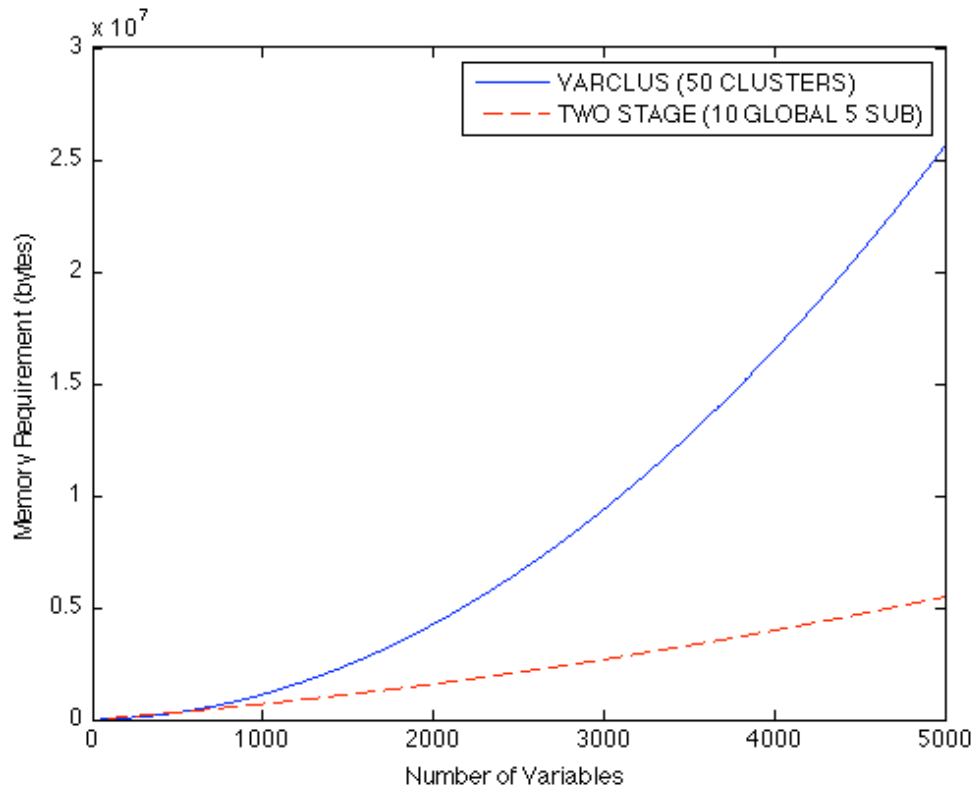


Figure 2. Comparison of memory requirements of PROC VARCLUS and the two stage method.

Speed is not the only concern when reducing a large data set. The reduction technique also needs to select a good set of variables, otherwise nothing is gained by using the reduced set of

variables. The two stage article claims that there is relatively little loss of information, meaning the two stage method selects a similar set of variables compared to the single stage PROC VARCLUS method [9].

CLeVer

CLeVer is a variable reduction method specifically for multivariate time series (MTS). The method was developed for classifying time series variables accurately and efficiently. In the original experiments there were several matrices representing time series data collected from repeated experiments with a certain set of variables. For example, one experiment measured data from various markers on a fifteen people's bodies as they walked. The goal was to find the features that best identified who was walking. In the matrices each column represented a variable and each row was an observation. The matrices were labeled based on the individual or group to which the data belonged, i.e. which person was walking. This method includes three basic steps to select variables. The first step is to calculate the principal components (PC) for each MTS matrix. The second step is to calculate the descriptive common principal components (DCPC) per label and their interconnectedness. This information is utilized to create clusters to aid in variable selection. The CLeVer algorithm is given in Table 4 [10].

Table 4. The CLeVer algorithm [10].

The CLeVer Algorithm

Step 1: Calculate the PCs.

1. Find the correlation matrix, X , of each set of time series data.
2. Calculate singular value decomposition (SVD) of each matrix of X . Then $X = USV^T$. Each row of V^T represents a principal component of X .

Step 2: Calculate the DCPCs.

1. The variance will be contained in the diagonal entries of S .
2. Find p , the number of diagonal entries of S where the variance is less than a prescribed percent of the total variance for each matrix.
3. Use the maximum p for all of the time series matrices and use that number of rows as a cut off point of the loadings. Create the matrix H defined as $H = \sum_i L_i^T L_i$ where L_i is the matrix of the first p PCs in the i th time series matrix.
4. Compute SVD of H .
5. Use the first p rows of V^T . These are the DCPCs.

Step 3: K-means clustering.

1. Do k-means clustering on the DCPCs.
2. The variables closest to the cluster centers will be the selected variables.

The type of data used in the original CLeVer experiment is different from the economic data from the Dow Chemical Company, since it was based on repeated experiments for each person with the same variables. However, since there are many observations in the economic data set, the matrix could be split into two or more matrices, each with a certain number of the observations for all of the variables. It is possible that the same variables will exhibit similar behaviors with respect to time regardless of the start and finish time.

Support Vector Machines

Support vector machines (SVM) use pattern recognition to categorize variables. The machine must be provided with a set of training examples, each belonging to certain categories. The the SVM builds a model that predicts in which category new information belongs [11].

Corona

Corona is a supervised method from the authors of CLeVer that was tested on the same data sets. It is a method of recursive feature elimination that uses SVMs and requires a multivariate time series to be represented as a vectors. The following algorithm in Table 4 outlines the process used in Corona [12].

Table 4. The Corona algorithm [12].

The Corona Algorithm

1. Create a correlation matrix for each multivariate time series.
2. Use the upper triangle of each matrix to vectorize the matrix.
3. Train the SVM on the vectorization.
4. Rank the variables based on weights from the SVM.
5. Remove variables with the lowest ranks until the desired number remain.

Method

Traditionally, data mining can be very time consuming when large data sets with many variables are being examined. There is a need for a more efficient method. After examining approaches suggested in many pieces of literature on variable reduction, variable selection, and data mining in forecasting problems, the methods of similarity, cointegration, CLeVer, k-means clustering and a two-step clustering approach were tested on a large sample multivariate time series data set. The best approach was determined by implementing the methods on a small test data set and a large set of real economic data. The methods were compared based on two criteria: similarity to variables chosen by Dow's current unsupervised method, and the computation time.

Data

Two sets of data were used for the variable reduction process. The first was a smaller data set that included input variables labeled x1 through x40. There was one target variable. Each of the variables was a sequence of time series data.

The second data set was a real world data set containing approximately 10 possible targets and over 1,700 inputs. Each of these has approximately 75 rows of quarterly time series data for various economic indicators and responses such as employment, consumer spending and import and export information. Since the methods used were unsupervised variable reduction methods, the target variables could be ignored.

Analysis/Results

Test Data Set

For the test data we compared six methods: two-stage clustering, Dow's current method, FASTCLUS used on the similarity matrix, VARCLUS used on the data set with no similarity, FASTCLUS used on the data set with no similarity and CLeVer. FASTCLUS and VARCLUS are two built in functions in SAS. Each of these methods were used to select 10 of the 40 possible input variables. The two-stage clustering chose less because one global cluster had less than five elements. The results of all of these methods can be seen in Table A1 of Appendix A. Since this data set is very small, computation time was negligible. Therefore, comparisons of methods on the test data are based only on comparing the reduced variable sets.

Several variables were chosen in almost in every method. For example, all methods except VARCLUS without similarity pick up variable x13. Also variables x31, x6, x5 are selected by most methods. It appears these variables play a strong role in the data set.

Next look at FASTCLUS using similarity and FASTCLUS using the data set. Since the main idea behind FASTCLUS is k-means clustering, the data had to be transposed when we applied FASTCLUS on the data set because FASTCLUS clusters rows. From the result, we can see that four variables x5, x6, x13 and x29 are the same for both methods, but the rest are chosen differently.

Then, looking at Dow VARCLUS selection and FASTCLUS with similarity closely, six of ten variables are the same. This is probably because both methods used similarity matrix, which removed the time-label. When comparing Dow VARCLUS with VARCLUS without similarity, the selection set are quite different, only four variables are the same, which includes x31 and x6.

Furthermore, two-stage selection only gave us eight variables, because one global cluster had less than five elements. The results indicate that four of eight variables are the same as Dow VARCLUS. Since this method uses the similarity matrix instead of data set during the first stage of clustering, we can assume if we use the raw data set, it may give us a quite different selection set.

CLeVer gave a very different selection set from other methods. When CLeVer was implemented, the matrix was split into two matrices, one containing the first half of the observations and the second containing the second half of the observations. When the number variables chosen, and hence the of clusters is small, the algorithm returns the same results consistently. However, when ten variables were chosen, the results varied more. The results in Table A1 shows the most consistently appearing results.

Economic Data Set

Three methods were used to cluster Dow's dataset, each using the similarity matrix that Dow currently uses with different clustering methods, their current clustering using VARCLUS, the new method of two stage clustering, and clustering using FASTCLUS, an observational clustering procedure in SAS. In order to compare each other, all three methods choose 50 clusters, which help reduce the data from 1,752 to 50 variables. These methods are only compared to Dow's unsupervised similarity method since they are also unsupervised methods.

There were several problems that occurred when CLeVer was implemented on the large data set. Splitting the data into two or more matrices of earlier and later observations resulted in a correlation matrix with not-a-number entries. This most likely occurred because some of the variables did not change significantly in the smaller number of observations recorded in the matrices. A second problem was that the program would output a different subset of variables each time it was run on the data. This could be due to the random assignment of cluster centers in the K-means clustering. An attempted remedy was to increase the number of repetitions of the clustering process and choosing the one with the smallest distance between the cluster centers and elements. However, even when 1,000 repetitions were used, the results were still not consistent. Therefore, the results from CLeVer are not included in the remainder of the analysis.

The variables selected from the real data set by the similarity between inputs is shown in Table A2 of the appendix. This list includes one variable selected from each of 50 clusters created from the similarity data. The variables were selected from the clusters based on the best R-square value within that cluster.

Using similarity and VARCLUS, yields 50 clusters. Within each cluster the variable with highest R-square value is chosen as a representative. All of the R-square values are above 0.9 and close to 1, which indicates that this method give a pretty good choice of variables.

Applying two stage clustering on similarity matrix results in ten global clusters in the first stage, each with five elements except clusters 5, 8, 9. Since these global clusters have less than five elements,

only 43 variables are chosen instead of 50. Again the variables were chosen according to their R-square value. In this case, the smallest R-square is about 0.74, and the highest is 1. A list of all of the variables chosen by this method can be found in Table A3 of Appendix A.

Similarity and FASTCLUS is an example of K-means clustering, therefore it chooses variables based on their distance from cluster center. The variable closest to the cluster center is chosen. Even when the smallest distance was chosen, most of them are above 0.5. The reduced variable set for this method is shown in Table A4 of Appendix A.

Table 5. Comparison of run time with Dow's current method of similarity and VARCLUS.

Method	Real Time	CPU Time	% reduction of CPU time
Similarity & VARCLUS	8 min 13.3 sec	8 min 9.43 sec	0%
Similarity & Two Stage	5 min 18.18 sec	5 min 9.98 sec	36.7%
Similarity & FASTCLUS	4 min 25.14 sec	4 min 19.16 sec	47%

The real time and CPU time to run the three methods is summarized in Table 5. Similarity and VARCLUS takes the longest time, and two stage clustering improves the run time about 37%. FASTCLUS improves the time by 47%. Judging strictly by run time, the similarity and FASTCLUS would be the best method among these three.

Table 6. Comparison of the number of common variables with Dow's current method of similarity and VARCLUS.

Method	# of variables in common	% of variables in common
Similarity & VARCLUS	50	100%
Similarity & Two Stage	4	6.7%
Similarity & FASTCLUS	4	6.7%

A comparison of the actual variables chosen using VARCLUS versus the other methods is summarized in Table 6. The methods of two stage clustering and FASTCLUS each only chose 4 variables in common with VARCLUS. This does not necessarily indicate that these are poor methods, simply that they are different. The results could be more fully analyzed by testing the ability of the variables chosen by each method in a model.

Conclusions

The following summarizes the results of variable reduction using similarity and VARCLUS, similarity and two stage clustering, similarity and FASTCLUS and CLeVer.

On the test data set:

- Computing time is negligible.
- Similarity and FASTCLUS chooses the set of variables that is most similar to Dow's current method.
- Similarity and two stage clustering is a close second.
- CLeVer has a much different set of variables.

On the real data set:

- CLeVer needs more work to perform on this data set.
- Similarity and FASTCLUS is the fastest method, reducing the computing time by 47.0%.
- Similarity and two stage clustering reduced the computing time by 36.7%.
- Neither FASTCLUS nor two stage clustering chose variables that were significantly similar to VARCLUS.

Since economic data sets will undoubtedly get larger and larger, FASTCLUS and two stage clustering are excellent ways to cut down computing time on very large data sets. More information is needed about the accuracy of these two methods to make a complete comparison of their merits. CLeVer may prove to be useful on large data sets after some of the problems are eradicated.

Project in Phase Two

Through the literature review, multiple new methods have been proposed. FASCLUS and two-stage clustering are two new methods in addition to Dow's VARCLUS method for unsupervised variable clustering, both of which save computational resources. However, one still cannot say these are best methods. Further examination of the validity of the results in a real modeling situation must be done in order to see if the reduced variable set is a good choice.

Since the dataset is time series data, the time factor must be considered during the data mining procedure. The time label is removed through the use of similarity, which makes it easier to cluster the data set, but generation of the similarity matrix can become time intensive when the number of variables is large. The similarity matrix was created before using FASTCLUS, however it is uncertain that this clustering technique should be used on a similarity matrix.

K-means clustering is used in both CLeVer and two stage clustering, however, K-means seems to only cluster observations, not variables. Therefore it is not used directly in this project. In the paper *Iterative Incremental Clustering of Time Series* [7], there is a suggestion to use the K-means algorithm to cluster time series data directly. This can be done using some background information about the wavelet transform. Wavelets are a technique used to deal with high-dimensional time series data, so this may be a good topic for future study.

Throughout the project, mainly new methods for unsupervised variable reduction were considered. Another future area of study would be more emphasis on supervised methods for variable selection. Dow currently uses similarity and co-integration as supervised methods to select variables. Most supervised methods involve model generation to pick variables, which is not as feasible with the large data sets typically found in economic applications. One potential supervised method is *Corona*, from the paper *A Supervised Feature Subset Selection Technique for Multivariate Time Series* [12]. This and other supervised procedures is where future resources should be devoted.

Acknowledgements

Francsics, Gabor

Dr. Francsics is a mathematics professor at Michigan State University. We are grateful for his advice and support as the faculty manager for this project.

Rey, Timothy

Mr. Rey specializes in research statistics and data mining at the Dow Chemical Company. We would like to thank him for proposing this project and for directing us and focusing our efforts throughout this project.

Speaker, Paul

Dr. Speaker works in research statistics and data mining at the Dow Chemical Company. We would like to thank him for his guidance and advice during the development of this project.

Wu, Peiru

Dr. Wu is a mathematics professor and the director of the Masters of Science in Industrial Mathematics program at Michigan State University. We would like to thank her for her feedback as we prepared the report for this project as well as her commitment to providing opportunities for experience in industry.

References

- [1] Michael J. A. Berry, Gordon Linoff, *Data Mining Techniques: For Marketing Sales, and Customer Support*, John Wiley & Sons, Inc., New York, 1997.
- [2] “Partial Least Squares (PLS)”, *StatSoft: Electronic Statistics Textbook*, <<http://www.statsoft.com/textbook/partial-least-squares/>>.
- [3] Kattamuri S. Sarma, “Variable Selection and Transformation of Variables in SAS® Enterprise Miner™ 5.2,” Ecostat Research Corp., White Plains NY, NorthEast SAS® Users Group, Inc., 2007.
- [4] Spyros Makridakis, Steven C. Wheelwright, Rob J. Hyndman, *Forecasting: Methods and Applications*, Third Edition, John Wiley & Sons, New York, 2008.
- [5] Michael Leonard, Jennifer Sloan, Taiyeong Lee, Bruce Elsheimer, “An Introduction to Similarity Analysis Using SAS®,” SAS Institute Inc., Cary, NC.
- [6] “Time-series Econometrics: Cointegration and Autoregressive Conditional Heteroskedasticity,” *Advanced information on the Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel*, The Royal Swedish Academy of Sciences, October 2003, <http://nobelprize.org/nobel_prizes/economics/laureates/2003/ecoadv.pdf>.
- [7] Jessica Lin, Michail Vlachos, Eamonn Keogh, Dimitrios Gunopulos. “Iterative Incremental Clustering of Time Series,” *Lecture Notes in Computer Science 2992*, Springer-Verlag Berlin Heidelberg, 2004: pp. 106-122.
- [8] Ashish Singhal, Dale E. Seborg, “Clustering multivariate time-series data,” *Journal of Chemometrics*, Vol. 19, August 2005: pp. 427-438.
doi: 10.1002/cem.945
- [9] Taiyeong Lee, David Duling, Song Liu, Dominique Latour, “Two-Stage Variable Clustering for Large Data Sets,” SAS Institute Inc., Cary, NC, SAS Global Forum 2008, Paper 320-2008.
- [10] Kiyoungh Yang, Hyunjin Yoon, Cyrus Shahabi, “CLeVer: A Feature Subset Selection Technique for Multivariate Time Series,” Computer Science Department, University of Southern California, Los Angeles, March 2005, Technical Report 05-845.
- [11] Alain Rakotomamonjy, “Variable Selection Using SVM-based Criteria,” *Journal of Machine Learning Research*, Vol.3, March 2003: pp. 1357-1370.
- [12] Kiyoungh Yang, Hyunjin Yoon, Cyrus Shahabi, “A Supervised Feature Selection Subset Technique for Multivariate Time Series,” Computer Science Department, University of Southern California, Los Angeles, April 2005.

Appendix A

The following pages include tables showing the complete list of variables chosen from the methods examined, including input to input similarity with VARCLUS, two stage clustering and FASTCLUS, and CLeVer.

Table A1. Reduced variable subsets from the test data set by using various methods.

Two stage	Dow VARCLUS	Only FASTCLUS	VARCLUS (no similarity)	FASTCLUS (no similarity)	CLeVer
x12	x6	x31	x40	x20	x18
x13	x10	x2	x4	x30	x26
x24	x8	x17	x31	x3	x20
x27	x22	x29	x21	x29	x22
x29	x31	x5	x16	x5	x1
x31	x13	x6	x6	x6	x36
x5	x5	x7	x33	x38	x21
x6	x38	x4	x22	x16	x4
	x7	x13	x2	x15	x13
	x34	x10	x8	x13	x33

Table A2. Variables found in real data set by using input to input similarity, clustering similar variables using FASTCLUS, and choosing one variable from each cluster based on the R-squared value.

Variable	R ²	Variable	R ²
BOPCRNTAC	0.9724	IFRES	0.9723
CDMVLVR	0.9402	IPSG334	0.9425
CDOOR	0.9804	IPSG3363	0.9613
CKF	0.9515	IPSN31411	0.9145
CMED	0.9402	JPCADJ	0.9474
COSTEMISC	0.9567	JPGFML	0.9523
COSTSBAO	0.9037	JPGFMLGI	0.9247
CSVHOPER	0.9528	JPIFNREE	0.9367
DOMPCCO	0.9211	JPIFNRESOB	0.9796
EDRSPUO	0.8884	JQPCMHFE	0.9541
EEPBS	0.9828	KNRADR	0.9012
GFMLC	0.9562	LIFEEIND	0.9523
GIUSCONPUTMILMANUOTHE_2	0.9715	NETXSV	0.9731
GIUSCONPUTMILTRAN	0.9303	PGO40_45	0.9552
GNP	0.9766	PVDETO	0.9689
GO22X7	0.9517	RMTCM10Y	0.9457
GO355	0.9775	RTXCGFS	0.9357
GO371TATRL	0.9806	TSMNH_USECON_Sum	0.9686
GO40_45	0.9701	TXCORPRW	0.9641
GSLINFRAR	0.9354	WPI141101	0.9047
HU1ESOLD	0.9618	WPIW_M1_PET	0.9670
IDEIND	1.0000	XGK	0.9794
IFNREEIPCSR	0.9398	YGSLTRF	0.9205
IFNRESPPR	0.9522	YPTRFGSL	0.9449
IFNRESXF	0.9788	ZBIVARW	0.9623

Table A3. Variables found in real data set by using input to input similarity with two-stage clustering. Variables without an R-squared value did not have enough members in the global cluster to make subclusters.

Variable	Global Cluster	R ²
GFEXPUNIADJ	1	0.8862
PJ287	1	0.7821
WPI051	1	0.8516
PJ345_7	1	0.8702
GIUSCONPUTMILMANUOTH	1	1.0000
GDPFERAWR	2	0.9930
EDREMISC	2	0.8796
KNIFNREEOR	2	0.9643
EDREIPCC	2	0.9395
KNIFNRESPUOR	2	0.9937
PJ3714	3	0.8900
KHUPS2ADIS	3	0.8637
IFNRESPP	3	0.8183
JPCSVHOPWAS	3	0.8178
CMED	3	0.8814
IFRESO	4	0.9630
CNOR	4	0.9074
COSTSPU	4	0.8708
rsh_nrs_usecon_Sum	4	0.9537
KNEFXNRER	4	0.8922
RESFRBB	5	
RESFRBE	5	
RESFRBT	5	
YPTRFGFFEO	5	

Variable	Global Cluster	R ²
GO361A2	6	0.9286
PJ131A2	6	0.7406
MGINR	6	0.8639
JPCDMVNA	6	0.8063
GIUSCONPUTMILPRIV_	6	0.8304
COSTEFXNREXE	7	0.9506
TITH_USECON_Sum	7	0.9734
COSTETO	7	0.8882
COSTEO	7	0.9654
SRTAFS	7	1.0000
RTXCGFS	8	
TDPASSLOSS	8	
RRDTE	9	
TDINTC	9	
IDEIND	10	0.9943
JPIFNRESOB	10	0.9926
KHUMFG	10	0.8742
JPCDFHEMAVC	10	0.9300
NP16A	10	0.8030

Table A4. Variables found in real data set by using input to input similarity with FASTCLUS. These are the variables with the closest distance to the cluster center.

Variable	Distance
COSTEIPO	0.7296
COSTEMISC	0.7459
COSTETLV	0.6851
COSTSBAOCP	0.7255
CSVTSUOXLSER	0.0000
EDREIPCS	0.8293
EDREIPCT	0.7303
EDRETAC	0.7543
EDRSPU	0.4934
EINFO	0.7539
GASTAXF	0.0000
GFEXPUNIADJ	0.5216
GFMLCKFR	0.8500
GIUSCONPUTMILMANUOTHE_2	0.6973
GIUSCONPUTMILRELI_2	0.7432
GIUSCONPUTMILRESI	0.6089
GSLGISNED	0.6700
IFNRESMFGR	0.6637
IFNRESMI	0.8011
IVACORP	0.7110
JPCDFHEMAVCSW	0.5594
JPCSVHOPDOM	0.7169
JPCXCDMVLV	0.6792
JPGDPEXP85	0.8136
JPIFNRESOTH	0.9225

Variable	Distance
JQPCMHFE	0.6286
KNIFNREEMISCR	0.7424
MTGFARMNA	0.0000
NP	1.0193
NP16A	0.8016
NP65A	0.3005
PDIINVMISC	0.6306
PHU1NMEDNS	0.5348
PHU1OFHEOXRNS	0.0000
PJ284	0.5878
QMGCR	0.5617
RESFRBE	0.7435
RESFRBF	0.0000
RESFRBNBA	0.0000
RMCD3SEC	0.5501
RRDTE	0.2179
RTXCGFS	0.8039
RTXSIGSL	0.7045
SUVGOV	0.6867
TSTH_USECON_Sum	0.5730
W8	0.5567
WPI051	0.4913
WPIW_4	0.6298
YPDADJ	0.5201
YPTRFGFFEO	0.0000

Appendix B

The code used to generate each of the resulting sets of variables for each of the methods of variable selection or reduction are shown in the tables below.

Table B1. The SAS code used for input to input similarity, clustering similar variables, and choosing one variable from each cluster based on the R-square value.

```
proc similarity data=egdcs.DCS_1990_2008 outsum=egdcs.SIMMATRIX;
id date interval=quarter;
target gdp_usecon--NRSJ21X_USECON /sdif=(1,4) normalize=absolute
measure=mabsdev
expand=(localabs=12)
compress=(localabs=12);
run;

proc VARCLUS data=egdcs.simmatrix maxc=50 outstat=egdcs.outstat_full;
var gdp_usecon--NRSJ21X_USECON;
run;
```

Table B2. The SAS code for the two stage variable clustering.

```
/* Set number of global clusters and sub-clusters within each global cluster*/
/* sfromg is letting you know which global cluster you are getting sub clusters
from*/

%let global=10;
%let sub=5;

proc similarity data=_exp0_.DCS_1990_2008 outsum=SIMMATRIX1;
id date interval=quarter;
target gdp_usecon--NRSJ21X_USECON /sdif=(1,4) normalize=absolute
measure=mabsdev
expand=(localabs=12)
compress=(localabs=12);
run;

/* Observational clustering to obtain global clusters*/

proc fastclus data=simmatrix1 maxclusters=&global outstat=fastclusstat
out=fastclusout noprint;
id _input_;
var gdp_usecon--NRSJ21X_USECON;
run;

/* Create clusters as cluster assignment from fastclus */

data clusters;
set fastclusout;
keep cluster;
run;

%macro twostage(num);
/* sfromg means sub cluster from global cluster, so num=number of global clusters*/
%do sfromg = 1 %to &num;
/* Create gcl as output of fastclus omitting distance and status*/

data gc&sfromg;
set fastclusout;
drop Distance _Status_;
run;

/*sort global clusters by cluster and keep only entries from global cluster of
interest*/

proc sort data=gc&sfromg;
by cluster;
where cluster=&sfromg;
run;

/*transpose global cluster*/

proc transpose data=gc&sfromg out=gc&sfromg;
copy _input_ cluster;
id _input_;
run;
```

```

/*merge transposed global cluster with the original cluster assignment*/

data gc&sfromg;
merge gc&sfromg(drop = _input_ cluster) clusters;
run;

/*keep only entries from global cluster of choice */
/*(this creates similarity matrix for global cluster)*/

proc sort data=gc&sfromg;
by cluster;
where cluster=&sfromg;
run;

/*remove cluster variable (all are now the same cluster)*/

data gc&sfromg;
set gc&sfromg;
drop cluster;
run;

/*run varclus on global cluster to create sub clusters */
/* need to work on this. the variable clustering doesnt work
when all elements in global cluster have similarity 0 */

proc varclus data=gc&sfromg maxc=&sub outstat=gc&sfromg._stat noprint;
run;

/*filter output so only has cluster assignment and rsquared for max clusters*/

data gc&sfromg._stat;
set gc&sfromg._stat;
where _ncl_=&sub AND (_type_='RSQUARED' OR _type_='GROUP');
drop _ncl_ _name_;
run;

/*transpose data so sorting can happen and get element with largest rsquared*/

proc transpose data=gc&sfromg._stat out=gc&sfromg._final;
id _type_;
run;

/*sorting by cluster assignment then by rsquared*/

proc sort data=gc&sfromg._final;
by group descending rsquared;
run;

data gc&sfromg._final;
set gc&sfromg._final;
cluster=&sfromg;
run;

```

```

/*keep only first entries (those with the highest rsquared)*/
/*remove this data step to get full cluster structure */

data gc&sfromg._final;
  set gc&sfromg._final;
  by group;
  if first.group;
run;

%end;
%mend;
%twostage(10)

/* append macro doesn't quite work...*/
/* it is still here because if it works, this step would be automated*/
/*
%macro appendmacro(num);

data afinal;
set gc1_final;
run;

%do i=2 %to &num;
proc append BASE=afinal DATA=gc&i._final ;
%end;
%mend;
%appendmacro(11)
*/

data clusterstructure;
set gc1_final
  gc2_final
  gc3_final
  gc4_final
  gc5_final
  gc6_final
  gc7_final
  gc8_final
  gc9_final
  gc10_final
;

run;

filename output "E:\Dow Project\clusters\twostagefinaldrift.xls";
PROC EXPORT
  DATA=twostagefinal
  OUTFILE=output
  DBMS=EXCEL2000 REPLACE;
RUN;

```

Table B3. The Matlab code used to select variables by the CLeVer method.

```

%CLEVer
y=importdata('filename.csv','') %import data columns=variables,rows=observations
Y=y.data;
X{1}=Y(1:38,:); %breaks matrix into earlier/later observations...
X{2}=Y(39:76,:); %...change the numbers to match data used
d=0.8; %threshold
DCPC=0; %initialize
H{1}=0; %initialize
N=2; %the number of MTS data groups

for i=1:N
    X{i}=corrcoef(X{i}); %create correlation matrices
    (nxn)
    z= find(isnan(X{i}));
    X{i}(z)= zeros(size(z));
    [m,n]=size(X{i}); %define size of matrix
    [U,S,V]=svd(X{i}); %calc. SVD
    loading{i}=V'; %define loadings
    variance=diag(S); %find variance
    percentVar=zeros(m,1); %initialize %variance
    tot=zeros(m); %inititalize sum of %variances
    p=zeros(m); %initialize p
    for j=1:m
        percentVar(j)=(variance(j)/sum(variance)); %find %variance
        if tot(i)>d %if the sum of %variance > the threshold
            break
        end
        tot(i)=tot(i)+percentVar(j); %sum of %variance
        p(i)=j; %return cell where variance > 0.8
    end
end
p=max(p); %p=max(p(i)) where variance exceeded
L{1}=loading{1}(1:p,1:n);
H{1}=H{1}+(L{1}'*L{1});
for i=2:N
    L{i}=loading{i}(1:p,1:n); %the first p rows of loading{i}
    H{i}=H{i-1}+(L{i}'*L{i});
end
[U,S,V]=svd(H{N}); %find SVD of H
V=V';
DCPC=V(1:p,:); %DCPCs
K=50; %number of clusters

DCPC=DCPC'; %transpose DCPC so that clustering happens on columns

[idx,cnt,sumd,D]=kmeans(DCPC,K,'replicates',20); %perform K-means clustering
[C,I]=min(D); %get index with min distance from cluster centers
I=sort(I) %return variables selected

```