

Paper 085-2013

Using Data Mining in Forecasting Problems

Timothy D. Rey, The Dow Chemical Company; Chip Wells, SAS Institute Inc.;

Justin Kauhl, Tata Consultancy Services

Abstract: In today's ever-changing economic environment, there is ample opportunity to leverage the numerous sources of time series data now readily available to the savvy business decision maker. This time series data can be used for business gain if the data is converted to information and then into knowledge. Data mining processes, methods and technology oriented to transactional-type data (data not having a time series framework) have grown immensely in the last quarter century. There is significant value in the interdisciplinary notion of data mining for forecasting when used to solve time series problems. The intention of this talk is to describe how to get the most value out of the myriad of available time series data by utilizing data mining techniques specifically oriented to data collected over time; methodologies and examples will be presented.

Introduction, Value Proposition and Prerequisites

Big data means different things to different people. In the context of forecasting, the savvy decision maker needs to find ways to derive value from big data. Data mining for forecasting offers the opportunity to leverage the numerous sources of time series data, internal and external, now readily available to the business decision maker, into actionable strategies that can directly impact profitability. Deciding what to make, when to make it, and for whom is a complex process. Understanding what factors drive demand, and how these factors (e.g. raw materials, logistics, labor, etc.) interact with production processes or demand, and change over time, are keys to deriving value in this context.

Traditional data mining processes, methods and technology oriented to static type data (data not having a time series framework), has grown immensely in the last quarter century (Fayyad, ET. Al. (1996), Cabena, ET. Al. (1998), Berry (2000), Pyle (2003), Duling, Thompson (2005), Rey, Kalos (2005), Kurgan and Musilek (2006), Han, Kamber (2012)). These references speak to the process as well as the myriad of methods aimed at building prediction models on data that does not have a time series framework. The idea motivating this paper is that there is significant value in the interdisciplinary notion of data mining for forecasting. That is, the use of time-series based methods to mine data collected over time.

This value comes in many forms. Obviously being more accurate when it comes to deciding what to make when and for whom can help immensely from a inventory cost reduction as well as a revenue optimization view point, not to mention customer loyalty. But, there is also value in capturing a subject matter expert's knowledge of the company's market dynamics. Doing so in terms of mathematical models helps to institutionalize corporate knowledge. When done properly, the ensuing equations actually become intellectual property that can be leveraged across the company. This is true even if the data sources are public, since it is how the data is used that creates IP, and that is in fact proprietary.

There are three prerequisites to consider in the successful implementation of a data mining for time series approach; understanding the usefulness of forecasts at different time horizons, differentiating planning and forecasting and, finally, getting all stakeholders on the same page in forecast implementation.

Defining the Need

One primary difference between traditional and time series data mining is that, in the latter, the time horizon of the prediction plays a key role. For reference purposes, short ranged forecasts are defined herein as one to three years, medium range forecasts are defined as 3 to 5 years and long term forecasts are defined as greater than 5 years. The authors agree that anything greater than 10 years should be considered a scenario rather than a forecast. Finance groups generally control the "planning" roll up process for corporations and deliver "the" number that the company

plans against and reports to Wall Street. Strategy groups are always in need for medium to long range forecasts for strategic planning. Executive Sales and Operations Planning (ESOP) processes demand medium range forecasts for resource and asset planning. Marketing and Sales organizations always need short to medium range forecasts for planning purposes. New Business Development incorporates medium to long range forecasts in the NPV process for evaluating new business opportunities. Business managers themselves rely heavily on short and medium term forecasts for their own businesses data but also need to know the same about the market. The optimal time horizon is a key driver in formulating and implementing the data transformation steps outlined in following sections. Since every penny a Purchasing organization can save a company goes straight to the bottom line, it behooves a company's purchasing organization to develop and support high quality forecasts for costs associated with raw materials, logistics, materials and supplies, as well as services.

However, regardless of the needs and aims of various stakeholder groups, differentiating a "planning" process from a "forecasting" process is critical. Companies do in fact need to have a "plan" that is aspired to. Business leaders do in fact have to be responsible for the plan. But, to claim that this plan is in fact a "forecast" can be disastrous. Plans are what we "feel we can do," while forecasts are mathematical estimates of what is most likely. These are not the same, but both should be maintained. In fact, the accuracy of both should be tracked over a long period of time. When reported to Wall Street, accuracy is more important than precision. Being closer to the wrong number does not help.

Given that so many groups within an organization have similar forecasting needs, a best practice is to move towards a "one number" framework for the whole company. If Finance, Strategy, Marketing/Sales, Business ESOP, NBD, Supply Chain and Purchasing are not using the "same numbers," tremendous waste can result. This waste can take the form of rework and/or miss-management given an organization is not totally lined up to the same numbers. This then calls for a more centralized approach to deliver forecasts for a corporation which is balanced with input from the business planning function. Chase (2009) presents this corporate framework for centralized forecasting in his book called Demand Driven Forecasting.

Big Data in Data Mining for Forecasting

Over the last 15 years or so, there has been an explosion in the amount of external time series based data available to businesses, Figure 1. To name a few, Global Insights, Euromonitor, CMAI, Bloomberg, Nielsen, Moody's Economy.com, Economagic, etc., not to mention government sources like www.census.gov, www.stastics.gov.uk/statbase, IQSS data base, research.stlouisfed.org, imf.org, stat.wto.org, www2.lib.udel.edu, sunsite.berkeley.edu, etc. All provide some sort of time series data – that is, data collected over time inclusive of a time stamp. Many of these services are for a fee; some are free. Global Insights (ihs.com) alone contains over 30,000,000 time series.



Figure 1. Examples of externally available time series data

This wealth of additional information actually changes the way a company should approach the time-series forecasting problem in that new methods are necessary to determine which of the potentially thousands of useful time series variables should be considered in the exogenous variable forecasting problem. Business managers do not have the time to “scan” and plot all of these series for use in decision making.

Many of these external sources do offer data bases for historical time series data but do not offer forecasts of these variables. Leading or forecasted values of model exogenous variables are necessary to create forecasts for the dependent or target variable. Some services, such as Global Insights, CMAI and others, do offer lead forecasts.

Concerning internal data, IT Systems for collecting and managing data, such as SAP and others, have truly opened the door for businesses to get a handle on detailed historical data for revenue, volume, price, costs and could even include the whole product income statement. That is, the system architecture is actually designed to save historical data. Twenty five years ago IT managers worried about storage limitations and thus would “design out of the system” any useful historical detail for forecasting purposes. With the cost of storage being so cheap now, IT architectural designs now in fact include “saving” various prorated levels of detail so that companies can take full advantage of this wealth of information.

Time Series vs. Transactional Modeling

A couple of important distinctions about time series modeling are important at this point. First, the one thing that differentiates time series data from simple static data is that the time series data can be related to “itself” over time. This is called serial correlation. If simple regression or correlation techniques are used to try and relate one time series variable to another, thus ignoring possible serial correlation, the business person can be misled. So, rigorous statistical handling of this serial correlation is important.

The second distinction is that there are two main classes of statistical forecasting approaches to consider. First there are “univariate” forecasting approaches. In this case, only the variable to be forecast (the “Y” or dependent variable) is considered in the modeling exercise. Historical trends, cycles and seasonality of the Y itself are the only structures considered when building the forecasting model. There is no need for data mining in this context. In the second approach – where the plethora of various time series data sources comes in – various “X’s” or independent (exogenous) variables are used to help forecast the Y or dependent variable of interest. This approach is considered exogenous variable forecast model building. Businesses typically consider this value added; now we are trying to understand the “drivers” or “leading indicators”. The exogenous variable approach leads to the need for data mining for forecasting problems.

Use of External Data

Though univariate or ‘Y only’ forecasts are often times very useful, and can be quite accurate in the short run, there are two things that they cannot do as well as the “multivariate” forecasts. First and foremost is providing an understanding of “the drivers” of the forecast. Business managers always want to know what “variables” (and in this case means what other time-series) “drive” the series they are trying to forecast. ‘Y only’ forecasts do not accommodate these drivers. Secondly, when using these drivers, the exogenous variable models can often forecast further and more accurately than the univariate forecasting models.

The recent 2008/2009 recession is evidence of a situation where the use of proper X’s in an exogenous variable “leading indicator” framework would have given some companies more warning of the dilemma ahead. Univariate forecasts were not able to capture this phenomena as well as exogenous variable forecasts.

The external data bases introduced above not only offer the “Y’s” that businesses are trying to model (like that in NAICS or ISIC data bases), but also provide potential “X’s” (hypothesized drivers) for the multivariate (in X) forecasting problem. Ellis (2005) in “Ahead of the Curve” does a nice job of laying out the structure to use for determining what “mega level” X variables to consider in a multivariate in X forecasting problem. Ellis provides a thought process that, when complimented with the Data Mining for Forecasting process proposed herein, will help the business forecaster do a better job identifying key drivers as well as building useful forecasting models.

The use of exogenous variable forecasting not only manifests itself in potentially more accurate values for price, demand, costs, etc. in the future, but it also provides a basis for understanding the timing of changes in economic activity. Achuthan and Banerji (2004), in "Beating the Business Cycle," along with Banerji (1999), present a compelling approach for determining potential X's to consider as leading indicators in forecasting models. Evans ET. al., (2002) as well as (www.nber.org and www.conference-board.org) have developed frameworks for indicating large turns in economic activity for large regional economies as well as specific industries. In doing so, they have identified key drivers as well. In the end, much of this work speaks to the concept that, if studied over a long enough time frame, many of the structural relations between Y's and X's do not actually change. This offers solace to the business decision maker and forecaster willing to learn how to use data mining techniques for forecasting in order to mine the time-series relationships in the data.

Many large companies have decided to include external data, such as that found in Global Insights as mentioned above, as part of their overall data architecture. Small internal computer systems are built to automatically move data from the external source to an internal data base. This accompanied with tools like SAS Institute Inc. Data Surveyor for SAP, allows bringing both the external Y and X data alongside the internal. Often times the internal Y data is still in transactional form. Once properly processed, or aggregated, e.g. by simply summing over a consistent time interval like month and concatenated to a monthly time stamp, this time stamped data becomes time-series data. This data base would now have the proper time stamp, include both internal and external Y and X data and be all in one place. This time-series data base is now the starting point for the data mining for forecasting multivariate modeling process.

Process and Methods for Data Mining for Forecasting

Various authors have defined the difference between "data mining" and classical statistical inference; Hand (1998), Glymour, et. Al. (1997), Kantardzic (2011), are notable examples. In a classical statistical framework, the scientific method (Cohen (1934)) drives the approach. First, there is a particular research objective sought after. These objectives are often driven by first principles or the physics of the problem. This objective is then specified in the form of a hypothesis; from there a particular statistical "model" is proposed, which then is reflected in a particular experimental design. These experimental designs make the ensuing analysis much easier in that the X's are independent, or orthogonal to one another. This orthogonality leads to perfect separation of the effects of the "drivers" there in. So, the data is then collected, the model is fit and all previously specified hypotheses are tested using specific statistical approaches. Thus, very clean and specific cause and effect models can be built.

In contrast, in many business settings a set of "data" often times contains many Y's and X's, but have no particular modeling objective or hypothesis for being collected in the first place. This lack of an original objective often leads to the data having irrelevant and redundant candidate explanatory variables. Redundancy of explanatory variables is also known as "multicollienarity" – that is, the X's are actually related to one another. This makes building "causes and effect" models much more difficult. Data mining practitioners will "mine" this type of data in the sense that various statistical and machine learning methods are applied to the data looking for specific X's that might "predict" the Y with a certain level of accuracy. Data mining on static data is then the process of determining what set of X's best predicts the Y(s). This is a different approach than classical statistical inference using the scientific method. Building adequate "prediction" models does not necessarily mean an adequate "cause and effect" model was built.

Considering time-series data, a similar framework can be understood. The Scientific Method in time series problems are driven by the "economics" or "physics" of the problem. Various "structural forms" may be hypothesized. Often times there is a small and limited set of X's which are then used to build multivariate times series forecasting models or small sets of linear models that are solved as a "set of simultaneous equations." Data mining for forecasting is a similar process to the "static" data mining process. That is, given a set of Y's and X's in a time series data base, what X's do the best job of forecasting the Y's. In an industrial setting, unlike traditional data mining, a "data set" is not normally readily available for doing this data mining for forecasting exercise. There are particular approaches that in some sense follow the scientific method discussed earlier. The main difference herein will be that time-series data cannot be laid out in a "designed experiment" fashion.

With regard to process, various authors have reported on the process for data mining static data. A paper by Azevedo and Santos (2008) compared the KDD process, SAS Institute's SEMMA process (Sample, Explore, Modify,

Model, Assess) and the CRISP data mining process. Rey and Kalos (2005) review the Data Mining and Modeling process used at The Dow Chemical Company. A common theme in all of these processes is that there are many X's and thus some methodology is necessary to reduce the number of X's provided as input to the particular modeling method of choice. This reduction is often referred to as Variable or Feature selection. Many researchers have studied and proposed numerous approaches for variable selection on static data (Koller (1996), Guyon (2003), etc.). One of the expositions of this article is an evolving area of research in variable selection for time-series type data.

The process for developing time series forecasting models with exogenous variables, Figure 2, starts with understanding the strategic objectives of the business leadership sponsoring the project.

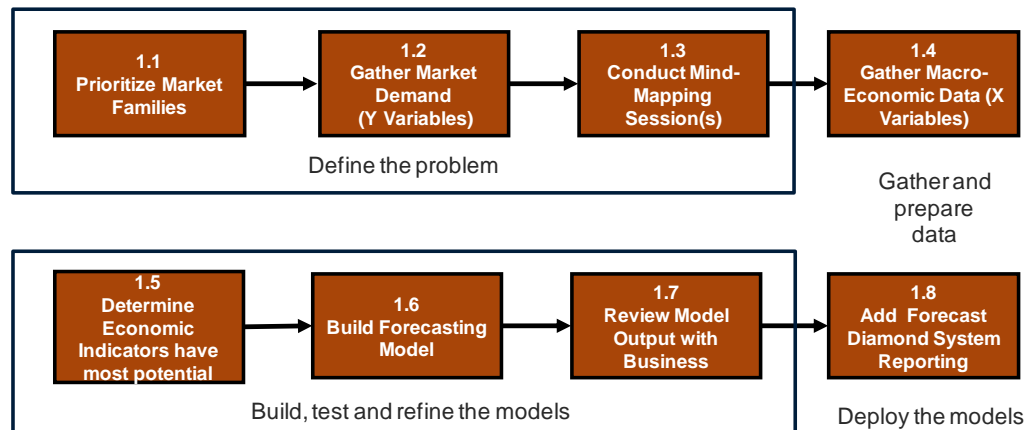


Figure 2 The Process for Developing Time Series Forecasting Models with Exogenous Variables

This understanding is often secured via a written charter so as to document key objectives, scope, ownership, decisions, value, deliverables, timing and costs. Understanding the system under study with the aid of the business subject matter experts provides the proper environment for focusing on and solving the right problem. Determining from here what data helps describe the system previously defined can take some time. In the end, it has been shown that the most time consuming step in any data mining prediction or forecasting problem is in fact the data processing step where data is defined, extracted, cleaned, harmonized and prepared for modeling. In the case of time series data, there is often a need to harmonize the data to the same time frequency as the forecasting problem at hand. Next, there is often a need to treat missing data properly. This missing data imputation may be in the form of forecasting forward, back casting or simply filling in missing data points with various algorithms. Often times the time series data base may have hundreds if not thousands of hypothesized X's in it. So, just as in Data Mining for static data, a specific Feature or Variable Selection step is needed.

Time Series Based Variable Reduction and Selection

As mentioned throughout the discussion, often times a time series data base may have hundreds if not thousands of hypothesized X's for a given set of Y's. As referenced in the process discussion above, in both the static data mining process and in data mining for forecasting, a step of the process exists to help reduce the number of variables for consideration in modeling. We will mention some of the traditional static Feature Selection approaches, adapted to time series data, as well as introduce various new time series specific Variable Reduction and Variable Selection approaches. There are two aspects of this phase of analysis. First, there is a variable reduction step and then there is a variable selection step. In variable reduction, various approaches are used to reduce the number of X's being considered while at the same time "explaining" or "characterizing" the key sources of variability in the X's regardless of the Y's. This variable reduction step is considered a non-supervised method in that the Y's are not being considered. The second step is the variable selection phase. This step contains supervised methodologies since the Y's are now being considered.

One of the key problems with using static Variable Reduction and Variable Selection approaches on time series data is that, in order to overcome the method not being time series based, the modeler has to include lags of the X's in

the problem. This dramatically increases the size of the variable selection problem. For example, if there are 1000 X's in the problem, and the data interval is quarterly, the modeler would have to add a minimum of 4 lags for each X in the Variable Reduction or Selection phase of the project. This is because the Y and X are potentially correlated at any or all positive lags of Y on X, roughly 1 to 4, and thus now there are 4000 'variables' to be processed.

In the traditional data mining literature on static or "non" time series data, many researchers have proposed numerous approaches for variable reduction and selection on static data (again Koller (1996), Guyon (2003), etc.). These static variable selection methods include but are not limited to: Decision Trees, Stepwise Regression, Genetic Programming, etc. The serially correlated structure and the potential for correlation between candidate X's and Y's at various lags makes implementing traditional data mining techniques on time series data problematic. New techniques are necessary for dimension reduction and model specification that accommodate the unique time series aspects of the problem. Below we present three methods that have been found to be effective for mining time series.

1) A Similarity analysis approach can be used for both variable reduction and variable selection. Leonard and Lee (2008) introduce via PROC SIMILARITY in SAS Econometrics and Time Series (ETS), an approach for analyzing and measuring the similarity of multiple time series variables. Unlike traditional time series modeling for relating Y (target) to an X (input) similarity analysis leverages the fact that the data is ordered. A similarity measure can take various forms but essentially is a metric that measures the distance between the X and Y sequences keeping in mind ordering. Similarity can be used simply to get the similarity between the Y's and the X's but it can also be used as input to a Variable Clustering (PROC VARCLUS in SAS STAT) algorithm to then get clusters of X's to help reduce redundant information in the X's and thus reduce the number of X's.

2) A co-integration approach for variable selection is a supervised approach. Engle and Granger (2001) discuss a co-integration test. Co-integration is a test of the Economic theory that two variables move together in the long run. The traditional approach to measuring the relationship between Y and X would be to make each series stationary (generally by taking first differences) and then see if they are related using a regression approach. This differencing may result in a loss of information about the long run relationship. Differencing has been shown to be a harsh method for rendering a series stationary. Thus co-integration takes a different tack. First, the simple OLS regression model (called the co-integrating regression), the residual are obtained use the exogenous variables as the Y and the dependent variable as the X. Then, a test statistic is used to see if the residuals of the model are stationary. This test can be either the Dickey-Fuller Test or the Durbin Watson Test. In the implementation examples below the Dickey-Fuller test is used.

3) A cross correlation approach for variable selection is also a supervised approach. A common approach used in time series modeling for understanding the relationship between Y's and X's is called the Cross Correlation Function (CCF). A CCF is simply the bar chart of simple Pearson Product moment correlations for each of the lags under study.

We have automated all of these methods using SAS EG and combine them into one table for review by a forecaster. The forecaster then reduces the total number of variables to be considered as input to the modeling tool by assessing an overall measure as well as the three inputs described above.

Code Description for Variable Reduction and Selection

The code used for variable reduction and selection had to fulfill several requirements before it would be deemed acceptable in the context of this project. First, and most importantly, it had to incorporate the statistical processes used for variable reduction and selection previously outlined, second it needed to be able to process "X's" or independent (exogenous) variables in large quantities, and lastly it needed to perform the first two tasks while still remaining computationally efficient. The input for this code would be the target Y variable and a pre-determined (via the process mentioned above) set of several hundred to several thousand independent variables. The goal is to produce a prioritized reduced set of meaningful exogenous variables with associated statistics portraying how strongly they are related to the Y.

The framework of the code relies heavily on the SAS macro language. The design includes a collection of independent macro functions. These functions perform alone or in tandem with others to compose a code "node" or

collection of functions that perform a single logical task within the master code flow. The core of these code nodes can perform their task on their own or in conjunction with the others to provide a customized analysis. Within the master code flow there are four core code nodes which handle the bulk of the computations that provide the statistical backbone to variable reduction and selection process and three peripheral nodes that provide error handling and variable cleaning prior to core node processing.

In the bulk of cases when this code is used the input is structured fairly simply. The input is comprised of two files, one file which contains the dependent, Y, variables and one which contains all of the independent, X, variables wherein one column is equivalent to one variable; each with a commonly formatted Date column. While this format of input allows for multiple dependent variables per run, in practice only one dependent variable is contained in each dependent variable file. The code will be extended to handle multiple dependent variables in a future iteration.

When working with large numbers of independent variables, often it is quite an intensive process to manually check each variable to see if the quality of data is sufficient for the analysis. Prior to the statistical core of this process the data needs to be cleaned and brought to a point where it will not cause errors downstream. In practice, the only assumption that could safely be made about the input data series was that they would have the correct time frequency. Everything else, even whether or not the series would have any data points, was an unknown upon input to the system.

Thus, some preprocessing and filtering on the data prior to the analysis was required. Specifically, if independent variables had less than the equivalent of one year's worth of data points plus one year of overlap with the dependent variable, they were dropped. Similarly, if they had a standard deviation of 0, or no variability, they were also dropped. A last task in the simple preprocessing stage was to mark the date of the last data point for each variable. No action is taken on this last point but it is a helpful supplemental piece of information at the end of the process.

At this point all variables that cannot contribute to the analysis are already filtered out. Here it becomes necessary to ensure that the variables that remain are comparable to each other in some meaningful way. As the assumption is that the variables provided could be in any state it is completely possible that they all exist in differing states of differencing. This means that differencing needs to be performed on these series individually so that they all become stationary. To accomplish this, some SAS HPF (High Performance Forecasting) processes are implemented.

```
/* inset = dataset which contains the series to be analyzed*/
/* names = the individual names of the series, space delimited */
/* data_freq = data point frequency */
/* time_idname = name of the date field */
%macro outest(inset, names);
    /* Clear old catalogs out of the directory */
    proc datasets lib=dir mt=catalog nolist;
        delete hpfscore hpfscatalog;
    run;

    proc hpfdiagnose data=dir.&inset. outest=dir.tempest
        modelrepository=dir.hpfscatalog basename=hpfsmodel;
        .....
        arimax;
    run;

    proc hpfbengine data=dir.&inset. inest=dir.tempest
        modelrepository=dir.hpfscatalog
        outest=dir.outest;
        .....
    run;

%mend outest;
```

The results of this code provide an “outest” file from which the differencing factors can be extracted. These can then be applied to the series individually to produce stationary series.

Once these preprocessing steps are completed the bulk of the analysis can proceed. As mentioned previously, there are four main components to this process. They are Similarity, Clustering, Cross-Correlation, and Co integration. All four of these are performed within the master code flow and then the results are concatenated into the final modelers report.

The first step Similarity is accomplished in two separate processes. The first of these processes is Independent-Dependent similarity, X-Y Similarity, and the second as Dependent-Dependent, X-X Similarity. Both are accomplished quite simply by using in Proc Similarity.

```
/* x_var = input independent dataset*/
/* y_var = input dependent dataset*/
/* xnames = list of independent variables to be used, space delimited */
/* ynames = list of dependent variables to be used, space delimited */
/* data_freq = data point frequency */
/* time_idname = name of the date field */
proc similarity data=dir.X_Y_Merged outsum=dir.xysim;
    id &Time_idname interval=&Data_Freq;
    input &xnames / scale=standard normalize=standard;
    target &ynames / measure=mabsdev normalize=standard
    expand=(localabs=0) compress=(localabs=12);
run;
```

The code for X-X Similarity contains almost exactly the same structures except that it less restrictive on the expansion. Clustering of X variables also occurs here as the output of similarity is the input for the clustering step. The clustering step is performed by Proc Varclus, which runs almost entirely in its native form. The one exception to this is redefining of the default maximum cluster number to some number defined by the user at runtime.

```
/*Similarity Step */
proc similarity data=dir.TS_&X_VAR outsum=dir.SIMMATRIX ;
    id &Time_idname interval=&Data_Freq;
    target &xnames /
    normalize=absolute
    measure=mabsdev
    /*The COMPRESS=(LOCALABS=12) option limits local absolute compression to
12. */
    /*The EXPAND=(LOCALABS=12) option limits local absolute expansion to 12.
*/
    expand=(localabs=12)
    compress=(localabs=12);
run;

/*Clustering Step*/
/*Cluster_number = user defined max cluster value */
PROC varclus data=dir.simmatrix maxc=&Cluster_number outstat=dir.outstat_full
noprint;
    var &xnames;
RUN;
```

The third step, Cross-Correlation is accomplished by using Proc Arima with the “crosscorr” option. After the Proc Arima step is run any correlations that reported an absolute correlation less than 2 standard deviations away from the

standard error was dropped. Also any correlation in the negative direction was dropped as well. Then through an iterative process only the top three lags are preserved for each X-Y pairing and reported on. Note that this analysis is not done on pre-whitened data since at this point we are not trying to diagnose the transfer function model we are only seeking key, large cross correlations. It has been the authors experience that for the main correlation and lag, there is no real difference between doing the analysis on raw data or pre-whitened data when use in this context.

```
/* Cross-Correlation */
proc arima data=&inset;
    identify var=&ynames crosscorr=( &xnames )
        nlag=10 outcov=&outset noprint;
run;
```

The final core code node deals with the analysis of co integration. The goal of this step is to gain an associated p value from the Dickey-Fuller test for each X-Y pair. To start this node joins the two variable files together and then stacks them so that each independent variable is joined to the dependent by Date. Once this table is created Proc Regression is run to build a residuals table for the rest of the analysis.

```
PROC REG DATA=stacked_all NOPRINT PLOTS(ONLY)=NONE;
    BY Y_Var X_Var;
    Linear_Regression_Model: MODEL X = Y / SELECTION=NONE;
    OUTPUT OUT=RESIDS
        .....
RUN;
```

This table is then used to derive the p-value. SAS does provide a tool to get the value called “%dfctest”. While this tool sufficiently calculates the required statistic it has the shortcoming of being able to only analyze one X-Y pair at a time. So, to handle an analysis of these subset sizes, the original large residual dataset needed to be broken off, analyzed, and then spliced back together. This causes a very large number of I/O operations which slows the process down considerably. To circumvent this, some changes were applied to the dfctest macro to allow it to process the file as a whole. Our “dfctest_batch” macro is the end result of this modification. This modification resulted in a significant reduction in processing time for this big data problem in this context.

The end result of these processes is a SAS program flow that automated a large amount of the processing that used to be relegated to the modeler. This entire process is fairly efficient: the user can expect to see the results of a single dependent variable compared against several thousand independents in a few minutes at most even on a fairly modest machine. The final analysis output, Table 1, derived by concatenating the results of these different analysis, can be used to greatly speed up the model building process.

NAME OF FORMER VARIABLE	Cluster	RSQUARED	Overall_fit_metr1	CCF_impact_factor	Coimt	XYsimilarity
SA_XNET_NetExports_Q	2	0.75353	1.06783	0.41089	0.82275	1.2136
US_IndProdTextileCarpet_Q	3	0.88058	1.05689	0.96827	0.30134	1.41755
Ratio_NatGas_Crude_Q	10	0.76041	1.00322	0.35933	0.01225	1.13092
EMEA_D17_Purchases_USR_A	10	0.93457	0.95842	0.10296	0.03182	0.95809
EMEA_D341T2_CapEx_USR_A	2	0.90305	0.93874	0.12241	0.24041	1.2694
GB_UKIP3825_IP_Oequip_Q	14	0.76552	0.91922	0.91594	0.30694	1.29469
DE_RSH_RetailSales_M	23	0.92545	0.91316	1.04961	0.21038	1.35571
EMEA_D17_TotalSales_USR_A	10	0.93171	0.89764	0.03428	0.06698	0.98287
EMEA_D17_AppCons_USR_A	10	0.94363	0.8833	0.06878	0.00804	0.69742
GB_UKIPTOT_ProdIDX_Total_Q	14	0.92461	0.87273	1.0093	0.27083	1.38147
EMEA_D17_ValueAdded_USR_A	10	0.91154	0.85845	0	0.01527	1.02758
US_DemandPaintsCoatings_Q	3	0.64531	0.83516	0.81584	0.29373	1.34716
US_PPL_Semiconductors_Q	23	0.93715	0.83511	0.06393	0.0143	0.64068
ECRI_UKCSI_Growth_M	2	0.9333	0.72088	0.11517	0.13984	1.13519
GB_UKIP3831A_IP_Emach_Q	23	0.9621	0.68237	1.01475	0.33737	1.31429
ECRI_FRCMI_Level_M	10	0.93054	0.67723	0.96632	0.23854	1.42202
GB_UKIP3832_IP_Cequip_Q	14	0.42833	0.67295	1.16686	0.15831	1.31584
EMEA_D17_OpProfit_USR_A	10	0.86821	0.67029	0	0.10978	0.85769
DE_CRUDE_GAS_IP_Q	23	0.90158	0.6047	0.23504	0.24053	1.28992
EMEA_D341T2_OpProfit_USR_A	18	0.88417	0.6027	0.12454	0.23764	1.27323
EMEA_D36_ValueAdded_USR_A	10	0.96373	0.6014	0.1382	0.1756	1.36848
US_DemandElecWireCable_Q	10	0.80881	0.58226	0.83917	0.31734	1.29785
EMEA_A_Purchases_USR_A	1	0.97916	0.56493	0.81553	0.26345	1.31931
EMEA_D343_ValueAdded_USR_A	2	0.94043	0.56254	0.12296	0.27745	1.31606
EMEA_D34_OpProfit_USR_A	18	0.84732	0.55947	0.18405	0.41904	1.25575
EMEA_D34_CapEx_USR_A	2	0.97543	0.54202	0.12235	0.19853	1.24471
ECRI_UKCEI_Growth_M	2	0.92079	0.53747	0.13727	0.27509	1.01952
EMEA_A_TotalSales_USR_A	1	0.96984	0.53109	0.63316	0.2083	1.34253
EMEA_A_AppCons_USR_A	1	0.97568	0.52896	0.87562	0.2914	1.31233
WO_D17_OpProfit_USR_A	15	0.7888	0.5221	0.12699	0.07719	1.04882
US_DemandHydCement_Q	14	0.90516	0.503	0.99644	0.21543	1.31523
US_PPL_Coal_Q	8	0.9042	0.49311	1.05682	0.27391	1.46314
EMEA_Infrastr_USR_A	1	0.96235	0.48859	0.2685	0.73586	1.44624
EMEA_D34_ValueAdded_USR_A	2	0.95011	0.48524	0.11406	0.34027	1.2608
ECRI_FRLMI_Level_M	22	0.93451	0.47742	1.07022	0.15624	1.46193
SA_CR_Priv_Consumption_Q	4	0.96182	0.45843	0.33415	0.29073	1.1502
EMEA_GovtCons_USR_Q	21	0.92775	0.45494	0.1373	0.38077	1.40476
EMEA_D341T2_ValueAdded_USR_A	2	0.95336	0.44433	0.11341	0.28852	1.27862
EMEA_D343_OpProfit_USR_A	7	0.78424	0.44147	0.1242	0.25144	1.30967

Table 1. Final Results and Prioritization of 3 methods for Variable Reduction and Selection

In these big data, time series variable reduction and variable selection problems, we combine (Figure 3) all three, along with the prioritized list of variables the business SME's suggest, into one research data base for studying the problem using a forecasting technology like SAS Forecast Studio. We often strive for a 25-50 to 1 reduction in large problems (going from 15,000 X's to 200-300 hundred).

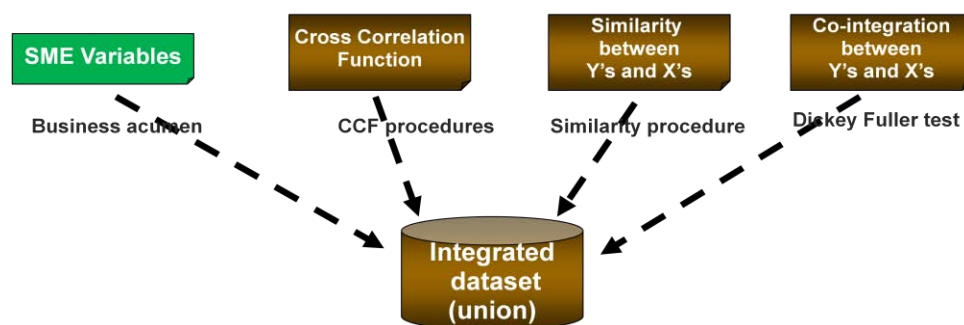


Figure 3. Combining various variable selection methods

Next, various forms of time series models are developed; but, just as in the Data Mining case for static data, there are some specific methods used to guard against over fitting which helps provide a robust final model. This includes, but is not limited to, dividing the data into three parts: model, hold out and out of sample. This is analogous to Training, Validating and Testing data sets in the static data mining space. Various statistical measures are then used to choose the final model. Once the model is chosen it is deployed using various technologies suited to the end user.

First and foremost, the reason for integrating Data Mining and Forecasting is simply to provide the highest quality forecasts as possible. The unique advantage to this approach lies in having access to literally thousands of potential X's and now a process and technology that enables doing the data mining on time series type data in an efficient and

effective manner. In the end, the business receives a solution with the best explanatory forecasting model as possible. With the tools now available through various SAS technologies, this is something that can be done in an expedient and cost efficient manner.

Now that models of this nature are easier to build, they can then be used in other applications beyond forecasting itself inclusive of Scenario Analysis, Optimization problems, as well as Simulation problems (linear systems of equations as well as non-linear system dynamics). So, all in all, the business decision maker will be prepared to make better decisions with these forecasting processes, methods and technologies.

Manufacturing Data Mining for Forecasting Example

Let's introduce a real example. Dow was interested in developing an approach for demand sensing that would provide:

- Cost Reduction
 - Reduction in resource expenses for data collection and presentation
 - Consistent automated source of data for leading indicator trends
- Provide Agility in the Market
 - Shifting to external and future looks from internal history.
 - Broader dissemination of key leading indicator data.
 - Better timing on market trends... faster price responses, better resource planning
 - reducing allocation/force major/share loss on the up side
 - reducing inventory carrying costs and asset costs on the down side
- Improved Accuracy
 - Accuracy of timing and estimates for forecast models
- Visualization – understanding leading indicator relationships

We were interested in better forecasting models for Volume (Demand), Net Sales, Standard Margin, Inventory Costs, Asset Utilization, and EBIT. This was to be done for all businesses and all geographies. Similar to many large corporations, Dow has a complex business/product hierarchy (Figure 4). This hierarchy starts at the top, total Dow, and then moves down through Division's, Business Groups, Global Business Units, Value Centers, Performance Centers, etc... As is the case in most large corporations, it is always changing and is overlaid with Geography. Even lower levels of the hierarchy exist when specific products are considered.

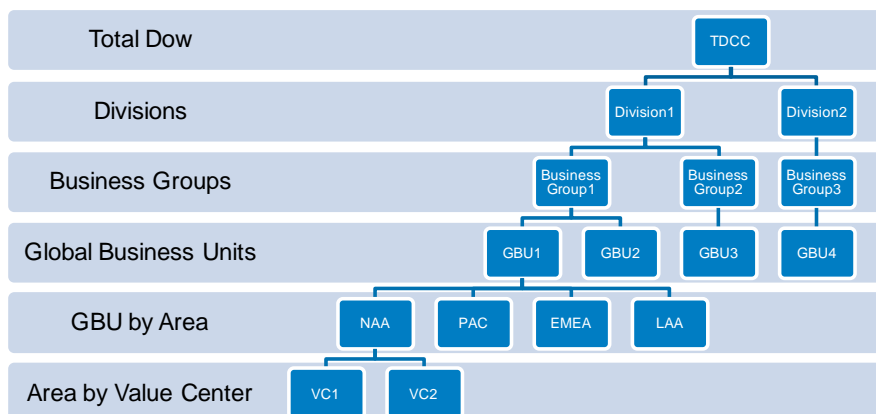


Figure 4. Dow Business Hierarchy

Dow operates in the vast majority of the 16 global market segments as defined in the ISIC market segment structure, some of which are: Agriculture, Hunting and Forestry, Mining and Quarrying, Manufacturing, Electricity, Gas and Water supply, Construction, Wholesale and Retail trade, Hotels and Restaurants, Transport, Storage and

Communications, Health and social work, etc. (Figure 5). This includes commodities, differentiated commodities and specialty products and thus makes the mix even more complex. The value chains Dow is involved in are very deep and complex often times connecting the earliest stages of hydrocarbons extraction and production all the way to the consumer on the street.

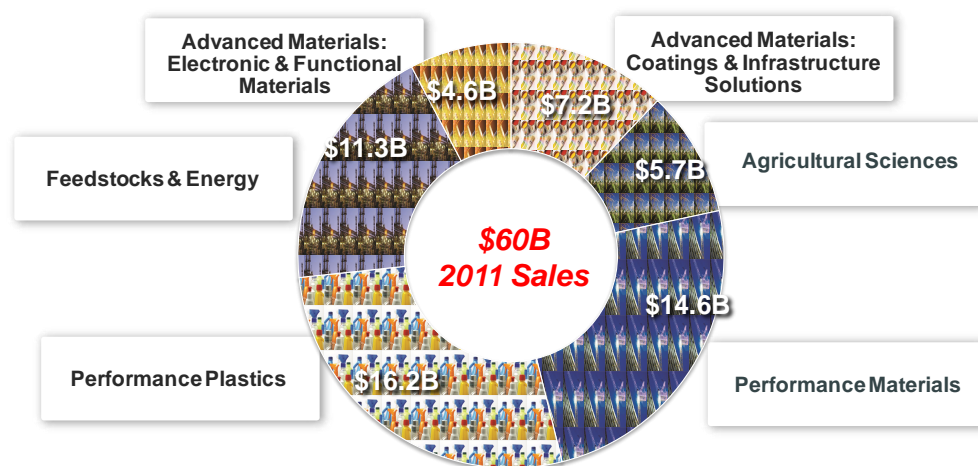


Figure 5. Dow Operating Segments

Before embarking on the project there were a few “industrial” and economic considerations to attack. First, simply multiplying out the number of models, we see that we would have around 7000 exogenous models to build, so, we focused on the top GBU by Area combinations in each Division restricting our initial effort to covering 80% of net sales. Next, we fully realize that the target variables of interest (Volume, Asset Utilization, Net Sales, Standard Margin, Inventory Costs and EBIT) are generally related to one another. Thus Volume is a function of Volume “drivers” (V_x), represented by $f(V_x)$, and AU is a function of Volume and AU “drivers” $f(AU_x)$, Inventory is a function of Volume and INV “drivers” $f(INV_x)$, Net Sales is driven by Volume, various costs (x_{costs}) and NS “drivers” $f(NS_x)$, Standard Margin is driven by Net sales and Standard Margin “drivers” $f(SM_x)$, and finally EBIT is driven by Standard Margin and EBIT “drivers” $f(EBIT_x)$.

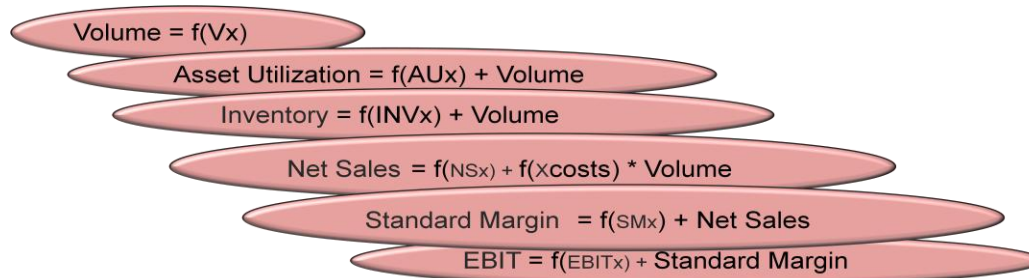


Figure 6. Daisy Chain Approach

The problem, if done only at one level of the hierarchy, fits into a Multivariate in Y approach that could be solved using a VARMAX (Vector Auto Regressive Moving Average with Exogenous variables) system. The complexity here is that we needed to solve the problem across the hierarchy shown above. We proposed that we could mimic the VARMAX structure by building the models in a “daisy chain” fashion shown in Figure 6 above. As a baseline, we thus compared a traditional VARMAX approach to the daisy chain approach at the total Dow level. We also did a traditional Univariate model as well as a traditional ARIMAX model for each Y. The Reconciled* column in Table 2 below was the daisy chain approach used in the hierarchy (implemented via SAS Forecast Studio) and then reconciled up. Given the results in the table below, we were thus confident we could use the daisy chain approach across the hierarchy and get similar benefit to the VARMAX approach. All of the above was accomplished with various SAS forecasting platforms.

Y variable	Univariate (no X's)	ARIMAX	VARMAX	Reconciled*
Volume	29.23	6.29	2.67	N/A
LOD	9.36	14.02	5.40	22.83
Inventory	12.51	1.29	1.42	4.59
Net Sales	12.98	2.94	3.21	1.47
Standard Margin	28.06	6.28	3.77	7.56
EBIT	48.81	29.85	9.18	12.03

Table 2. Model approach comparison results

Following the data mining for forecasting process described above, leads to conducting dozens of mind mapping sessions to have the businesses propose various sets of “drivers” for the numerous GBU and VC by geographic area combinations. This leads to using thousands (over 15,000 in this case!) potential exogenous variables of interest for the 7000 models in the hierarchy. This is truly a big data, thus large scale forecasting problem. A lot of automation, using several of the SAS tools, was necessary for first setting up initial SAS FS research projects as well as automatically building initial univariate and daisy chain models. As described earlier, SAS Enterprise Guide (EG) code was leveraged for the time series variable reduction and variable selection necessary to reduced the X's to a reasonable size. We also built automatically generated pre-whitening analysis for the reduced set of X's for the initial models in case the modelers wanted to build their own competitive models to those proposed by SAS FS. This was also accompanied with SAS EG code. Traditional hold out and out of sample methods were used for testing the quality and robustness of the models being proposed. A small example of the quality of some of the initial models is given in Table 3 below. Models with hold out SMAPES (Symmetric Mean Absolute Percent Error) greater than 15% are reworked as appropriate.

GBU Code	GBU	Area	Volume	NS
0000009807	POLYURETHANE SYSTEMS	EMEA	10.67%	17.10%
0000006325	EPOXY	NAA	5.37%	6.59%
0000016985	POLYGLYCOLS SURFACTANTS & FLUIDS	NAA	5.52%	6.19%
0000016982	POLYOLS	EMEA	11.70%	16.82%
0000006325	EPOXY	EMEA	8.30%	7.43%
0000006274	ISOCYANATES	EMEA	10.47%	14.12%
0000009820	SOLVENTS & INTERMEDIATES	NAA	9.22%	8.79%
0000006325	EPOXY	PAC	14.53%	3.29%
0000013049	AMINES	NAA	2.85%	2.89%
0000016982	POLYOLS	LAA	8.82%	6.61%
0000016982	POLYOLS	NAA	8.31%	16.82%
0000016983	PLASTIC ADDITIVES	NAA	8.71%	10.03%
0000013749	PERFORMANCE SOLUTIONS	EMEA	10.11%	4.48%
0000016983	PLASTIC ADDITIVES	EMEA	4.75%	13.09%
0000009791	GLYCOL ETHERS	PAC	10.31%	8.65%
0000006274	ISOCYANATES	NAA	13.67%	7.91%
0000009808	PO/PG	NAA	2.85%	15.80%
0000016982	POLYOLS	PAC	10.39%	8.54%
0000016985	POLYGLYCOLS SURFACTANTS & FLUIDS	EMEA	4.10%	3.17%
0000009791	GLYCOL ETHERS	EMEA	3.93%	5.40%
0000009807	POLYURETHANE SYSTEMS	NAA	8.55%	6.56%
0000009808	PO/PG	EMEA	3.66%	7.85%
0000006274	ISOCYANATES	PAC	9.20%	9.06%
0000006274	ISOCYANATES	LAA	10.80%	22.51%
0000009820	SOLVENTS & INTERMEDIATES	PAC	10.29%	33.96%
0000009791	GLYCOL ETHERS	NAA	5.06%	6.12%
0000013749	PERFORMANCE SOLUTIONS	NAA	9.68%	4.04%
0000009798	CHLORINATED ORGANICS BUS	NAA	5.06%	7.37%
0000013748	DA ADHESIVES	EMEA	5.90%	8.12%
0000009808	PO/PG	PAC	8.92%	7.23%
0000013049	AMINES	EMEA	5.59%	13.90%

Table 3. Example of Model quality across the GBU by Area level of the Hierarchy

Individual models are presented back to the businesses for approval in graphical form (Figure 7). Drivers are presented in a simple format, as in Figure 8, for consumption by the business.

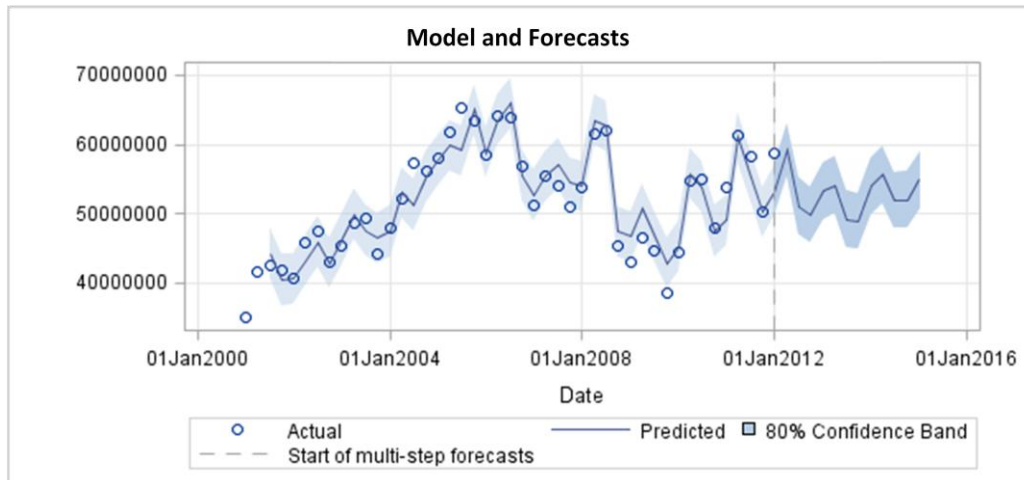


Figure 7. Visual rendition of model quality

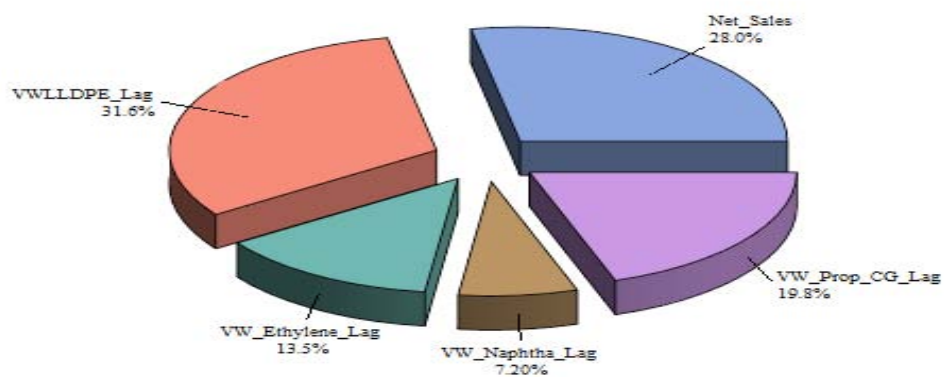


Figure 8. Visual rendition of Model Driver Contributions

Lastly, concerning in use model visualization, the business can gain access to these forecasts in a corporate wide business intelligence delivery system where they can see the history, model, forecast and drivers (Display 1).



Display 1. Real Time Model Visualization

Summary

Big data mandates big judgment. Big judgment has to have short “ask to answer” cycles. In the case of time series data for forecasting, there is certainly the potential for problems with big data given services like IHS Global Insight that provide access to over 30,000,000 time series. These opportunities call for the use of data mining for forecasting approaches which leads us to using special techniques for variable reduction and selection on time series data. These large problems can be complicated by complex hierarchies and special issues driven by known financial structures. Dow has overcome a very “big data” like forecasting problem in a project for corporate demand sensing. Over 7000 models were built drive by over 15,000 initial X’s. Model errors as low as 2 to 5 percent have been obtained on the upper level of the organization structure. Model results are extracted from SAS systems and moved to the corporate business intelligence platform to be consumed by the business decision maker in a visual manner

References

1. Achuthan, L. and Banerji, A. (2004) Beating the Business Cycle, Doubleday.
2. Antunes, C. And Oliveira, A. (2001) “Temporal Data Mining: An Overview, KDD Workshop on Temporal Data Mining.
3. Azevedo, A. and Santos, M. (2008) “KDD, SEMMA and CRISP-DM: A parallel overview”, Proceedings of the IADIS.
4. Banerji, A. (1999) “The Lead Profile and Other Nonparametric to Evaluate Survey Series as Leading Indicators”, 24th CIRET conference.
5. Berry, M. (2000) Data Mining Techniques and Algorithms, John Wiley and Sons.
6. Cabena, P, Hadjinian, P, Stadler, R, Verhees, J and Zanasi, A (1998) Discovering Data Mining: From Concept to Implementation, Prentice Hall.
7. Chase, C. (2009) Demand-driven forecasting: a structured approach to forecasting, SAS Institute, Inc..

8. Cohen, M. and Nagel, E. (1934) An Introduction to Logic and Scientific Method, Oxford, England: Harcourt, Brace xii.
9. CRISP -DM 1.0 (2000) SPSS, Incorporated.
10. SAS Institute Inc. (2003) Data Mining Using SAS®, Enterprise Miner™: A Case Study Approach, Second Edition. Cary, NC: SAS Institute Inc.
11. Duling, D. and Thompson, W. (2005) "What's New in SAS® Enterprise Miner™ 5.2", SUGI-31, Paper 082-31.
12. Ellis, J. (2005) Ahead of the Curve: A common sense guide to forecasting business and market cycles, Harvard Business School Press.
13. Engle, R. and Granger W. (1992) Long-Run Economic Relationships: Readings in Cointegration, Oxford University Press.
14. Evans, C., Liu, C. T., and Pham-Kanter, G. (2002) "The 2001 recession and the Chicago Fed National Activity Index: Identifying business cycle turning points", Federal Reserve Bank of Chicago,.
15. Fayyad, U, Piatetsky-Shapiro, G, Smyth, P and Uthurusamy, R (eds.) (1996a) "Advances in Knowledge Discovery and Data Mining," AAAI Press.
16. Glymour, C., Madigan, D., Pregibon, Smyth, P. (1997) "Statistical Themes and lesson for Data Mining", Data Mining and Knowledge Discovery 1, 11–28, Kluwer Academic Publishers.
17. Guyon, I. (2003) "An introduction to variable and feature selection", The Journal of Machine Learning Research, Vol 3 Issue 7-8 pages 1157-1182.
18. Han, J. and Kamber, M. and Pie, J. (2012) Data Mining: concepts and techniques, Elsevier, Inc..
19. Hand, D. (1998) "Data Mining: Statistics and More", American Statistician, Vol. 52, No. 2.
20. Kantardzic, M. (2011) Data Mining: Concepts, Models, Methods, and Algorithms, Wiley.
21. Koller, D. and Sahami, M. (1996) "Towards Optimal Feature Selection", International Conference on Machine Learning, Volume: 1996, Issue: May, Publisher: Citeseer, Pages: 284-292.
22. Kurgan, L. and Musilek, P. (2006) "A Survey of Knowledge Discover and Data Mining process models," *The Knowledge Engineering Review*, Vol. 21, No. 1, pp. 1-24.
23. Lee, T. and S. Schubert (2011) "Time Series Data Mining with SAS® Enterprise Miner? Paper 160-2011, SAS Institute Inc., Cary, NC.
24. Lee, T., ET. All, (2008) "Two-Stage Variable Clustering for Large Data Sets," SAS Institute Inc., Cary, NC, SAS Global Forum, Paper 320-2008.
25. Leonard, M., Lee, T, Sloan, J. and Elsheimer, B. (2008) "An Introduction to Similarity Analysis Using SAS," SAS Institute White Paper.
26. Leonard, M. And Wolfe, B. (2002) "Mining Transactional and Time Series Data International Symposium of Forecasting.
27. Mitsa, T. (2010) "Temporal Data Mining", Taylor and Francios Group, LLC.
28. Pankratz, A., (1991) Forecasting with Dynamic Regression Models, Wiley.

29. Pyle, D. (2003) Business Modeling and Data Mining, Elsevier Science, 2003.
30. Rey, T. and Kalos, A. (2005) "Data Mining in the Chemical Industry," *Proceedings of the eleventh ACM SIGKDD*.

ACKNOWLEDGEMENTS

Thanks to the all of the Dow, CMU IHBI and TCS team members that helped solve this very large, time series big data problem for Dow.

DISCLAIMER: The contents of this paper are the work of the author(s) and do not necessarily represent the opinions, recommendations, or practices of Dow.

CONTACT INFORMATION

Comments, questions, and additions are welcomed.

Contact the author at:

Tim Rey, Director
Advanced Analytics
WHDC, Bldg 2040
The Dow Chemical Company
Midland, Mi 48674
Phone: (9089) 636-9283
Email: tdrey@dow.com

TRADEMARKS

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