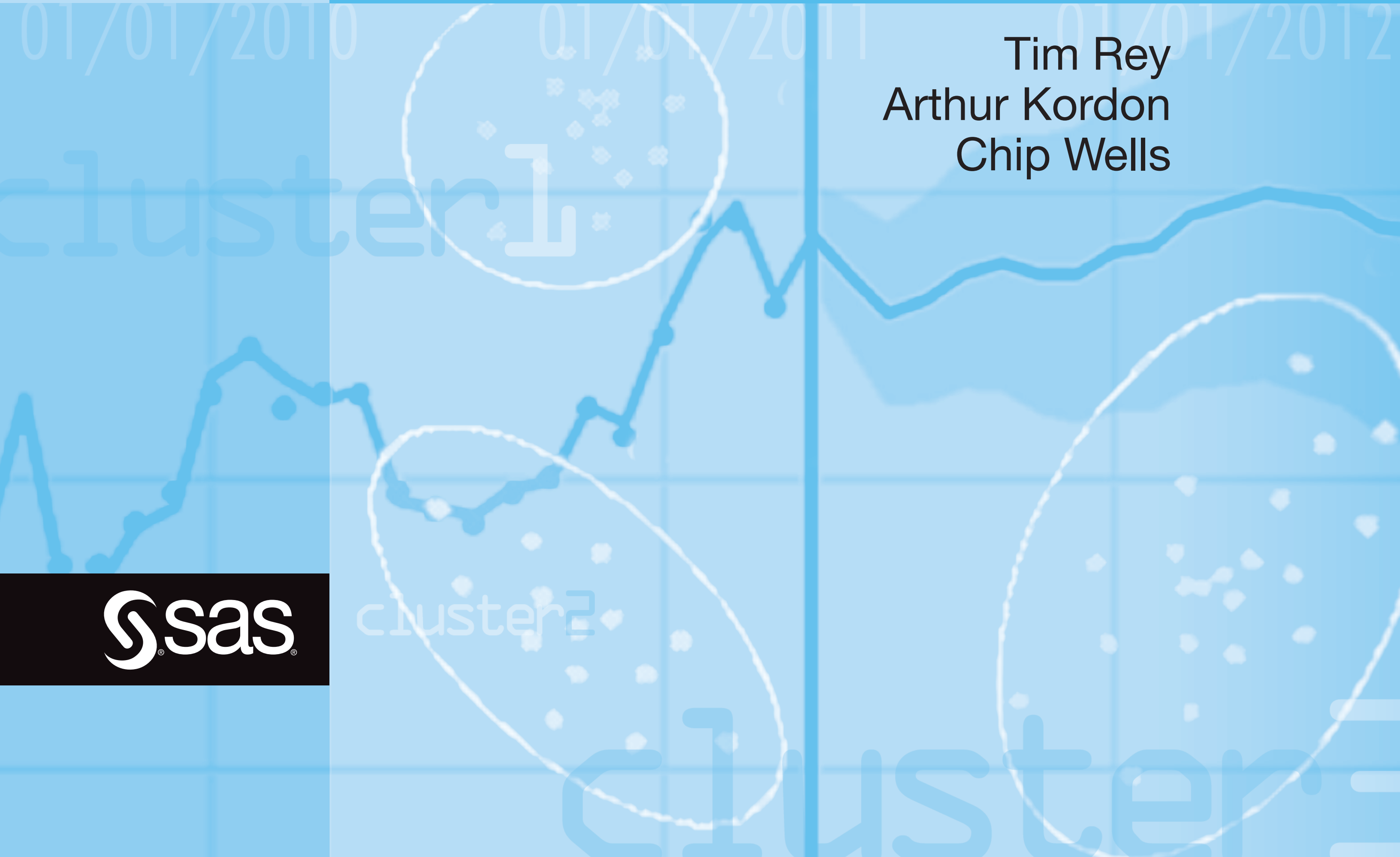
SAS® Press

Applied Data Mining for Forecasting Using SAS®

Tim Rey
Arthur Kordon
Chip Wells



From *Applied Data Mining for Forecasting Using SAS®*. Full book available for purchase here.

Contents

Preface .... **vii**

Chapter 1 Why Industry Needs Data Mining for Forecasting **1**

1.1 Overview 1

1.2 Forecasting Capabilities as a Competitive Advantage 2

1.3 The Explosion of Available Time Series Data 2

1.4 Some Background on Forecasting 3

1.5 The Limitations of Classical Univariate Forecasting 4

1.6 What is a Time Series Database? 4

1.7 What is Data Mining for Forecasting? 5

1.8 Advantages of Integrating Data Mining and Forecasting 6

1.9 Remaining Chapters 6

Chapter 2 Data Mining for Forecasting Work Process **7**

2.1 Introduction 7

2.2 Work Process Description 7
- 2.2.1 Generic Flowchart 8
- 2.2.2 Key Steps 9

2.3 Work Process with SAS Tools 16
- 2.3.1 Data Preparation Steps with SAS Tools 17
- 2.3.2 Variable Reduction and Selection Steps with SAS Tools 19
- 2.3.3 Forecasting Steps with SAS Tools 20
- 2.3.4 Model Deployment Steps with SAS Tools 21
- 2.3.5 Model Maintenance Steps with SAS Tools 22
- 2.3.6 Guidance for SAS Tool Selection Related to Data Mining in Forecasting 22

2.4 Work Process Integration in Six Sigma 23
- 2.4.1 Six Sigma in Industry 23
- 2.4.2 The DMAIC Process 24
- 2.4.3 Integration with the DMAIC Process 25

Appendix: Project Charter 26

Chapter 3 Data Mining for Forecasting Infrastructure **29**

3.1 Introduction 29

3.2 Hardware Infrastructure 29
- 3.2.1 Personal Computers Network Infrastructure 30
- 3.2.2 Client/Server Infrastructure 30
- 3.2.3 Cloud Computing Infrastructure 31

3.3 Software Infrastructure 32
- 3.3.1 Data Collection Software 32
- 3.3.2 Data Preparation Software 32
- 3.3.3 Data Mining Software 32
- 3.3.4 Forecasting Software 33

3.3.5 Software Selection Criteria 33
3.4 Data Infrastructure 34
3.4.1 Internal Data Infrastructure 34
3.4.2 External Data Infrastructure 34
3.5 Organizational Infrastructure 35
3.5.1 Developers Infrastructure 35
3.5.2 Users Infrastructure 36
3.5.3 Work Process Implementation 37
3.5.4 Integration with IT 37
Chapter 4 Issues with Data Mining for Forecasting Application 39
4.1 Introduction 39
4.2 Technical Issues 39
4.2.1 Data Quality Issues 39
4.2.2 Data Mining Methods Limitations 43
4.2.3 Forecasting Methods Limitations 45
4.3 Nontechnical Issues 46
4.3.1 Managing Forecasting Expectations 46
4.3.2 Handling Politics of Forecasting 47
4.3.3 Avoiding Bad Practices 49
4.3.4 Forecasting Aphorisms 49
4.4 Checklist "Are We Ready?" 50
Chapter 5 Data Collection 53
5.1 Introduction 53
5.2 System Structure and Data Identification 53
5.2.1 Mind-Mapping 54
5.2.2 System Structure Knowledge Acquisition 54
5.2.3 Data Structure Identification 55
5.3 Data Definition 57
5.3.1 Data Sources 57
5.3.2 Metadata 59
5.4 Data Extraction 61
5.4.1 Internal Data Extraction 61
5.4.2 External Data Extraction 61
5.5 Data Alignment 62
5.5.1 Data Alignment to a Business Structure 62
5.5.2 Data Alignment to Time 63
5.6 Data Collection Automation for Model Deployment 64
5.6.1 Differences between Data Collection for Model Development and Deployment 64
5.6.2 Data Collection Automation for Model Deployment 64
Chapter 6 Data Preparation 67
6.1 Overview 67
6.2 Transactional Data Versus Time Series Data 67
6.3 Matching Frequencies 75
6.3.1 Contracting 75
6.3.2 Expanding 78
6.4 Merging 83
6.5 Imputation 84

6.6 Outliers 87
6.7 Transformations 92
6.8 Summary 94

Chapter 7 A Practitioner's Guide of DMM Methods for Forecasting 95

7.1 Overview 95
7.2 Methods for Variable Reduction 96
Traditional Data Mining 96
Time Series Approach 97
7.3 Methods for Variable Selection 103
Traditional Data Mining 104
Example for Variable Selection 111
Variable Selection Based on Pearson Product-Moment Correlation Coefficient 112
Variable Selection Based on Stepwise Regression 114
Variable Selection Based on the SAS Enterprise Miner Variable Selection Node 116
Variable Selection Based on the SAS Enterprise Miner Partial Least Squares Node 118
Variable Selection Based on Decision Trees 120
Variable Selection Based on Genetic Programming 122
Comparison of Data Mining Variable Selection Results 124
7.4 Time Series Approach 125
7.5 Summary 134

Chapter 8 Model Building: ARMA Models 135

Introduction 135
8.1 ARMA Models 136
8.1.1 AR Models: Concepts and Application 136
8.1.2 Moving Average Models: Concepts and Application 149
8.1.3 Auto Regressive Moving Average (ARMA) Models 157
Appendix 1: Useful Technical Details 164
Appendix 2: The "I" in ARIMA 166

Chapter 9 Model Building: ARIMAX or Dynamic Regression Modes 179

Introduction 179
9.1 ARIMAX Concepts 180
9.2 ARIMAX Applications 185
Appendix: Prewhitening and Other Topics Associated with Interval-Valued Input Variables 195

Chapter 10 Model Building: Further Modeling Topics 205

Introduction 205
10.1 Creating Time Series Data and Data Hierarchies Using Accumulation and Aggregation Methods 206
Introduction 206
Creating Time Series Data Using Accumulation Methods 206
Creating Data Hierarchies Using Aggregation Methods 210
10.2 Statistical Forecast Reconciliation 216
10.3 Intermittent Demand 220
10.4 High-Frequency Data and Mixed-Frequency Forecasting 224
High-Frequency Data 224
Mixed-Interval Forecasting 232

10.5 Holdout Samples and Forecast Model Selection in Time Series ... 237
Introduction ... 237
10.6 Planning Versus Forecasting and Manual Overrides ... 240
10.7 Scenario-Based Forecasting ... 241
10.8 New Product Forecasting ... 244

Chapter 11 Model Building: Alternative Modeling Approaches ... 251

11.1 Nonlinear Forecasting Models ... 251
11.1.1 Nonlinear Modeling Features ... 251
11.1.2 Forecasting Models Based on Neural Networks ... 252
11.1.3 Forecasting Models Based on Support Vector Machines ... 254
11.1.4 Forecasting Models Based on Evolutionary Computation ... 256
11.2 More Modeling Alternatives ... 258
11.2.1 Multivariate Models ... 258
11.2.2 Unobserved Component Models (UCM) ... 263

Chapter 12 An Example of Data Mining for Forecasting ... 273

12.1 The Business Problem ... 273
12.2 The Charter ... 273
12.3 The Mind Map ... 274
12.4 Data Sources ... 274
12.5 Data Prep ... 275
12.6 Exploratory Analysis and Data Preprocessing ... 279
12.7 X Variable Imputation ... 283
12.8 Variable Reduction and Selection ... 287
12.9 Modeling ... 291
12.10 Summary ... 307

Appendix A ... 309

Appendix B ... 311

References ... 313

Index ... 317

From Applied Data Mining for Forecasting Using SAS® by Tim Rey, Arthur Kordon, and Chip Wells. Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

From *Applied Data Mining for Forecasting Using SAS®*. Full book available for purchase here.

Chapter 1: Why Industry Needs Data Mining For Forecasting

1.1 Overview 1
1.2 Forecasting Capabilities as a Competitive Advantage 2
1.3 The Explosion of Available Time Series Data 2
1.4 Some Background on Forecasting 3
1.5 The Limitations of Classical Univariate Forecasting 4
1.6 What is a Time Series Database? 4
1.7 What is Data Mining for Forecasting? 5
1.8 Advantages of Integrating Data Mining and Forecasting 6
1.9 Remaining Chapters 6

1.1 Overview

In today's economic environment there is ample opportunity to leverage the numerous sources of time series data that are readily available to the savvy decision maker. This time series data can be used for business gain if the data is converted first to information and then to knowledge—knowing what to make when for whom, knowing when resource costs (raw material, logistics, labor, and so on) are changing or what the drivers of demand are and when they will be changing. All this knowledge leads to advantages to the bottom line for the decision maker when times series trends are captured in an appropriate mathematical form. The question becomes how and when to do so. Data mining processes, methods and technology oriented to transactional type data (data that does not have a time series framework) have grown immensely in the last quarter century. Many of the references listed in the bibliography (Fayyad et al. 1996, Cabena et al. 1998, Berry 2000, Pyle 2003, Duling and Thompson 2005, Rey and Kalos 2005, Kurgan and Musilek 2006, Han et al. 2012) speak to the many methods and processes aimed at building prediction models on data that does not have a time series framework. There is significant value in the interdisciplinary notion of data mining for forecasting when used to solve time series problems. The intention of this book is to describe how to get the most value out of the host of available time series data by using data mining techniques specifically oriented to data collected over time. Previous authors have written about various aspects of data mining for time series, but not in a holistic framework: Antunes, Oliveira (2006), Laxman, Sastry (2006), Mitsa (2010), Duling, Lee (2008), and Lee, Schubert (2011).

In this introductory chapter, we help build the case for using data mining for forecasting and using forecasting as a competitive advantage. We cover the explosion of available economic time series data, the basic background on forecasting, and the limitations of classical univariate forecasting (from a business perspective). We also define what a time series database is and what data mining for forecasting is all about, and lastly describe what the advantages of integrating data mining and forecasting actually are.

1.2 Forecasting Capabilities as a Competitive Advantage

Information Technology (IT) Systems for collecting and managing transactional data, such as SAP and others, have opened the door for businesses to understand their detailed historical transaction data for revenue, volume, price, costs and often times even the whole product income statement. 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 decline of the cost of storage in recent years, architectural designs have in fact included saving various prorated levels of detail over time so that companies can fully take advantage of this wealth of information. IT infrastructures were initially put in place simply to manage the transactions. Today, these architectures should also accommodate leveraging this history for business gain by looking at it from an advanced analytics view point. Various authors have discussed this framework in detail (Chattratichat et al. 1999, Mundy et al. 2008, Pletcher et al. 2005, Duling et al. 2008).

Large corporations generally have many internal processes and functions that support businesses—all of which can leverage quality forecasts for business gain. This is beyond the typical supply chain need for having the right product at the right time for the right customer in the right amount. Some companies have moved to a lean pull replenishment framework in their supply chains. This lean approach does not preclude the use of high-quality forecasting processes, methods, and technology.

In addition to those who analyze the supply chain, many other organizations in a corporation can use high-quality forecasts. Finance groups generally control the planning process for corporations and deliver the numbers 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) 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 (NBD) incorporates medium- to long-range forecasts in the NPV (net present value) 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 about the market. Since every penny saved goes straight to a company's bottom line, it behooves a company's purchasing organization to develop and support high-quality forecasts for raw material, logistics, materials and supplies, and service costs.

Differentiating a planning process from a forecasting process is important. Companies do in fact need to have a plan to follow. Business leaders do in fact have to be responsible for the plan. But claiming 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 maintained over a long period of time. When reported to Wall Street, accuracy in the actual forecast 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, why not move towards a "one number" framework for the whole company? If finance, strategy, marketing and 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 or mismanagement if an organization is not totally aligned with the same numbers. Such cross-organizational alignment requires a more centralized approach that can deliver forecasts that are balanced with input from the business and financial planning parts of the corporation. Chase (2009) presents this corporate framework for centralized forecasting in his book called *Demand Driven Forecasting*.

1.3 The Explosion of Available Time Series Data

Over the last 15 years, there has been an explosion in the amount of time series-based data available to businesses. To name a few, Global Insights, Euromonitor, CMAI, Bloomberg, Nielsen, Moody's Economy.com, Economagic—not to mention government sources such as www.census.gov, www.statistics.gov.uk/statbase, www.statistics.gov.uk/hub/regional-statistics, IQSS database, research.stlouisfed.org, imf.org, stat.wto.org, www2.lib.udel.edu, and sunsite.berkeley.edu. All provide some sort of time series data—that is, data collected over time inclusive of a time stamp. Many of these services are available for a fee, but some are free. Global Insights (www.ihs.com) contains over 30,000,000 time series. It has been the authors' collective experience that this richness of available time series data is not the same worldwide.

This wealth of additional time series information actually changes how a company should approach the time series forecasting problem in that new processes, methods, and technology are necessary to determine which of the potentially thousands of useful time series variables should be considered in the exogenous or multivariate in an X forecasting problem (Rey 2009). Business managers do not have the time to scan and plot all of these series for use in decision making. Statistical inference is a reduction process and data mining techniques used for forecasting can aid in the reduction process.

In order to provide some structure to data concerning various product lines consumed in an economy, there has long been a code structure used to represent an economies market. Various government and private sources provide this data in a time series format. This code structure is called NAICS (*North American Industry Classification System*) in North America (www.census.gov/naics)**.** Various sources provide historical data in this classification system, but some also produce forecasts (Global Insights). For global product histories, an international system was recently deployed (ICIS—International Code Industry System). This system is at a higher level than the NAICS codes. For reference, there are cross-walk tables between the two (www.naics.com/). Both of these systems, among others, provide potential Y variables for a corporation's market forecasting endeavors. In some cases, depending on the level of detail being considered, these same sources may even be considered Xs.

Many of these sources offer databases for historical time series data but do not offer forecasts themselves. Other services, such as Global Insights and CMAI, do in fact offer forecasts. In both of these cases though, the forecasts are developed based on an econometric engine versus simply supplying individual forecasts. There are many advantages to having these forecasts and leveraging them for business gain. How to do so by leveraging both data mining and forecasting techniques will be discussed in the remainder of this book.

1.4 Some Background on Forecasting

A couple of important distinctions about time series models are important at this point. First, the one thing that differentiates time series data from transaction data is that the time series data contains a time stamp (day, month, year.) Second, time series data is actually 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, without regard to serial correlation, the business person can be misled. Therefore, rigorous statistical handling of this serial correlation is important. Third, there are two main classes of statistical forecasting approaches detailed in this book. 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 the seasonality of the Y itself are the only structures considered when building the univariate forecast model. In the second approach, where a multitude of time series data sources as well as the use of data mining techniques come in, various Xs or independent (exogenous) variables are used to help forecast the Y or dependent variable of interest. This approach is considered multivariate in the X or exogenous variable forecast model building. Building models for forecasting is all about finding mathematical relationships between Ys and Xs. Data mining techniques for forecasting become all but mandatory when 100s or even 1000s of Xs are considered in a particular forecasting problem.

For reference purposes, short-range forecasts are defined as one to three years, medium-range forecasts are defined as three to five years, and long-term forecasts are defined as greater than five years. Generally, the authors agree that anything greater than 10 years should be considered a scenario rather than a forecast. More often than not, in business modeling, quarterly forecasts are being developed. Quarterly data is the frequency that the historical data is stored and forecast by the vast majority of external data service providers. High-frequency forecasting might also be of interest even in finance where data can be collected by the hour or minute.

1.5 The Limitations of Classical Univariate Forecasting

Thanks to new transaction system software, businesses are experiencing a new richness of internal data, but, as detailed above, they can also purchase services to gain access to other databases that reside outside the company. As mentioned earlier, when building forecasts using internal transaction Y data only, the forecasting problem is generally called a univariate forecasting model. Essentially, the transaction data history is used to define what was experienced in the past in the form of trends, cycles, and seasonality to then forecast the future. Though these forecasts are often very useful and can be quite accurate in the short run, there are two things that they cannot do as well as the multivariate in X forecasts: They cannot provide any information about the "drivers" of the forecasts. Business managers always want to know what variables drive the series they are trying to forecast. Univariate forecasts do not even consider these drivers. Secondly, when using these drivers, the multivariate in X or exogenous models can often forecast further in time, with accuracy, then the univariate forecasting models.

The 2008–09 economic recession was evidence of a situation where the use of proper Xs in a multivariate in X "leading indicator" framework would have given some companies more warning of the dilemma ahead. Services like ECRI (Economic Cycle Research Institute) provided reasonable warning of the downturn some three to nine months ahead of time. Univariate forecasts were not able to capture these phenomena as well as multivariate in X forecasts.

The external databases introduced above not only offer the Ys that businesses are trying to model (like that in NAICS or ISIC databases), but also provide potential Xs (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 X variables to consider in a multivariate in X forecasting problem. Ellis provides a thought process that, when complemented with the data mining for forecasting process proposed herein, will help the business forecaster do a better job of both identifying key drivers and building useful forecasting models.

Forecasting is needed not only to predict accurate values for price, demand, costs, and so on, but it is also needed to predict when changes in economic activity will occur. Achuthan and Banerji—in their *Beating the Business Cycle* (2004) and Banerji in his complementary paper in 1999—present a compelling approach for determining which potential Xs 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 for specific industries. In doing so, they have identified key drivers as well. In the end, much of this work shows that, if we study them over a long enough time frame, we can see that many of the structural relations between Ys and Xs do not actually change. This fact 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.

1.6 What is a Time Series Database?

Many large companies have decided to include external data, such as that found in Global Insights, as part of their overall data architecture. Small internal computer systems are built to automatically move data from the external source to an internal database. This practice, accompanied with tools like the SAS® Data Surveyor for SAP (which is used to extract internal transaction data from SAP), enables both the external Y and X data to be brought alongside the internal Y and X data. Often the internal Y data is still in transactional form that, once properly processed, can be converted to time series type data. With the proper time stamps in the data sets, technology such as Oracle, Sequel, Microsoft Access or SAS itself can be used to build a time series database from this internal transactional data and the external time series data. This database would now have the proper time stamp and Y and X data all in one place. This time series database is now the starting point for the data mining for forecasting multivariate in X modeling process.

1.7 What is Data Mining for Forecasting?

Various authors have defined the difference between “data mining” and classical statistical inference (Hand 1998, Glymour et al. 1997, and Kantardzic 2011, among others). 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 Xs are orthogonal to one another, which leads to a perfect separation of the effects therein. So the data is then collected, the model is fit and all previously specified hypotheses are tested using specific statistical approaches. In this way, very clean and specific cause-and-effect models can be built.

In contrast, in many business settings a set of “data” often contains many Ys and Xs, but there was no particular modeling objective or hypothesis in mind when the data was being collected in the first place. This lack of an original objective often leads to the data having multi-collinearity—that is, the Xs are actually related to one another. This makes building cause-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 Xs that might predict the Y with a certain level of accuracy. Data mining on transactional data is then the process of determining what set of Xs best predicts the Ys. This is quite different than classical statistical inference using the scientific method. Building adequate prediction models does not necessarily mean that an adequate cause-and-effect model was built, again, due to the multi-collinearity problem.

When considering time series data, a similar framework can be understood. The scientific method in time series problems is driven by the economics or physics of the problem. Various structural forms can be hypothesized. Often there is a small and limited set of Xs that are then used to build multivariate in X 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 transaction data mining process. That is, given a set of Ys and Xs in a time series database, the goal is to find out what Xs do the best job of forecasting the Ys. In an industrial setting, unlike traditional data mining, a data set is not normally 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 here will be that time series data cannot be laid out in a “designed experiment” fashion. This book goes into much detail about the process, methods, and technology for building these multivariate in X time series models while taking care to find the drivers of the problem at hand.

With regard to process (previously discussed), various authors have reported on the process for data mining transactional data. A paper by Azevedo and Santos (2008) compared the KDD process, SAS Institute’s SEMMA (Sample, Explore, Modify, Model, Assess) process 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 Xs, and therefore some methodology is necessary to reduce the number of Xs 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 transaction data (Koller 1996, Guyon 2003). One of the main concentrations of this book will be on an evolving area of research in variable selection for time series type data.

At a high level, the data mining process for forecasting starts with understanding the strategic objectives of the business leadership sponsoring the project. This is often secured via a written charter that documents 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 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. Then there is often a need to treat missing data properly. This may be in the form of forecasting forward, backcasting or simply filling in missing data points with various algorithms. Often the time series database has hundreds if not thousands of hypothesized Xs in it. So, just as in data mining for transactional data, a specific feature or variable selection step is needed. This book will cover the traditional transactional feature selection approaches, adapted to time series data, as well as introduce various new time series specific variable reduction and variable selection approaches. Next, various forms of time series models are developed; but, just as in the data mining case for transaction data, there are some specific methods used to guard against overfitting, which helps provide a robust final model. One such method is 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 transaction 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.

This discussion shows how and why it is important that the subject matter experts' knowledge of a company's market dynamics is captured in a form that institutionalizes this knowledge. This institutionalization actually surfaces through the use of mathematics, specifically statistics, machine learning and econometrics. When done, the ensuing equations become intellectual property (IP) that can be leveraged across the company. This is true even if the data sources are in fact public, since how the data is used to capture the IP in the form of mathematical models is in fact proprietary.

The core content of the book is designed to help the reader understand in detail the process described in the previous paragraphs. This will be done in the context of various SAS technologies, including SAS® Enterprise Guide®, SAS Forecast Server and various SAS/ETS® time series procedures like PROC EXPAND, PROC TIMESERIES, PROC ARIMA, PROC SIMILARITY, PROC Xll/12, as well as the SAS® Enterprise Miner™ time series data mining nodes, and others.

1.8 Advantages of Integrating Data Mining and Forecasting

The reason for integrating data mining and forecasting is simply to provide the highest-quality forecasts possible. Business leaders now have a unique advantage in that they have easy access to thousands of Xs, and the knowledge about a process and technology that enables data mining on time series data. With the tools now available through various SAS technologies, the business leader can create the best explanatory (cause and effect) forecasting model possible, and this can be accomplished in an expedient and cost efficient manner.

Now that models of this type are easier to build, they then can be used in other applications, including scenario analysis, optimization problems, and simulation problems (linear systems of equations as well as non-linear system dynamics). All in all, the business decision maker is now prepared to make better decisions with these advanced analytics forecasting processes, methods and technologies.

1.9 Remaining Chapters

The next chapter defines and discusses in detail the process of data mining for forecasting. In Chapter 3, details are given about how to set up an infrastructure for data mining for forecasting. Chapter 4 covers issues with data dining for forecasting applications. This then leads to data collection in Chapter 5 and data preparation in Chapter 6, which has an entire chapter dedicated to the topic since 60–80% of the work lies in this step. Chapter 7 discusses the foundation for the actually doing data mining by providing a practitioner's guide to data mining methods for forecasting. Chapters 8 through 11 present a practitioner's guide to time series forecasting methods. Chapter 12 finishes the book by walking through an example of data mining for forecasting from start to finish.

From *Applied Data Mining for Forecasting Using SAS®* by Tim Rey, Arthur Kordon, and Chip Wells. Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

SAS Press
Applied Data Mining for Forecasting Using SAS®
Tim Rey
Arthur Kordon
Chip Wells
sas

From *Applied Data Mining for Forecasting Using SAS®*. Full book available for purchase here.

Index

A

Access (SEMMA work process) 19
accumulation methods
  creating data hierarchies using 210–211
  creating time series data using 206–210
ACF (autocorrelation function) plot
  AR models and 140–144
  ARIMAX models and 189
  ARMA models and 159–161
  MA models and 152–153
ADF (Augmented Dickey, Fuller) test 167–176
ADJMEAN option, SPECTRA procedure 110
AGGDATA option, HPFRECONCILE procedure 218
aggregation methods 210–216
AIC (Akaike's information criterion)
  about 20
  AR models and 145
  ARMA models and 162
  MA models and 155–156
  overfitting and 237, 239
Analyze phase (DMAIC) 24, 26
aphorisms, forecasting 49–50
application issues (data mining for forecasting)
  about 39
  avoiding bad practices 49
  data mining methods limitations 43–45
  data quality issues 39–43
  forecasting aphorisms 49–50
  forecasting methods limitations 45–46
  handling politics of forecasting 47–48
  managing forecasting expectations 46–47
  readiness checklist 51–52
AR (autoregressive) models
  about 135–136, 149
  AR 2 specification 144–146
  AR 3 specification 146–147
  AR 4 specification 147–149
  concepts and application 136–138
  model estimation, selection, and forecasting 144
  model identification 138–144
ARCH models 19
ARIMA Modeling and Forecasting task 21
ARIMA models
  See ARMA (autoregressive moving average) models
ARIMA procedure
  about 20, 22
  code example 141
  ESTIMATE statement 142, 184–185
  FORECAST statement 148, 194
  holdout samples and 238
  IDENTIFY statement 138, 140, 173, 182, 187
  MA models and 151
  mixed-interval forecasting and 235–236
  multivariate model example 259
  NOEST option 241
ARIMAX models
  See also ARMA (autoregressive moving average) models
  about 179–180
  applications for 185–195
  basic concepts 180–185
  interval input variables and 195–203
  UCM and 263
ARMA (autoregressive moving average) models
  about 14, 157–158
  AR models and 136–149
  high-frequency data and 225
  irregular variation 185–186, 265
  limitations of 46
  model estimation, selection, and forecasting 161–164
  model identification 159–161
  moving average models and 149–157
  SAS Enterprise Guide and 18
  SAS Forecast Studio and 21
  seasonal adjustment of time series data 108–109
  technical details about 164–166
  transfer function modeling and 166–176
augmenting lags 171
%AUTOARIMA macro 161
autocorrelation function plot
  See ACF (autocorrelation function) plot
automating data collection for model deployment 64–65
AUTOREG procedure 19, 22
autoregressive models
  See AR (autoregressive) models
autoregressive moving average models
  See ARMA (autoregressive moving average) models

B

back-propagation algorithm 253
backcasting process 85–86, 285–286
backshift operator 158, 164–165
Bartlett's Kolmogorov-Smirnov statistic 110
Base SAS 16, 22
Basic Forecasting task 21
Beating the Business Cycle (Achuthan and Banerji) 4
biases in forecasting 48
BIC
  See SBC (Schwartz's Bayesian criterion)
BLOCKSEASON statement, UCM procedure 232
Bloomberg service 58
Bottom Up reconciliation 216, 218–219
Box-Cox transformation method 20, 92
Breiman, Leo 43
building models
  See model building
business decisions, forecasting models and 15
The Business Forecasting Deal (Gilliland) 36, 49–50
business structures, aligning data to 62–63
Buzan, Tony 54
BY statement
  HPFRECONCILE procedure 219
  TIMESERIES procedure 212

C

CCF (cross-correlation function) plot
  about 179
  ARIMAX models and 181–185, 187–188
  data mining for forecasting example 291
  multivariate models and 258–261
  new product forecasting example 246–247
  variable selection and time series data 126–131
CFNAI (Chicago Fed National Activity Index) 12–13, 58–59

client/server infrastructure 30–31
cloud computing infrastructure 31
clustering algorithms limitations 44
CMAI (Chemical Market Associates, Inc.) 58
co-integration test 126, 130, 132, 291
coefficient of variation (CV) 40
competitive advantage 47
COMPRESS= option, TARGET statement (SIMILARITY) 245
concept mapping
    See mind-mapping
consensus planning 14
consultant services 37–38
contracting time interval frequencies 75–78
Control phase (DMAIC) 24, 26
Create Time Series Data task 18, 69–71
CRISP data mining process 5
cross-correlation function plot
    See CCF (cross-correlation function) plot
CROSSCORR option, IDENTIFY statement (ARIMA) 182, 187
Croston's method 220–222
curve fitting technique 105
CV (coefficient of variation) 40

D

data collection
    about 11–12, 15
    automation for model deployment 64–65
    data alignment in 62–63
    data extraction in 61–62, 64
    data sources for 57–59
    importance of 53
    metadata and 59–61
    software for 32
    system structure and data identification in 53–56
data hierarchies
    creating 210–216
    statistical forecast reconciliation and 216–217
data identification
    See system structure and data identification
data infrastructure 34–35
Data Mining and Modeling process (DOW Chemical Company) 5
data mining for forecasting
    about 1–6
    application issues 39–51
    infrastructure considerations 29–38
    method limitations 43–46
    work process considerations 7–28
data mining for forecasting example
    business problem 273
    data preparation for 275–279
    data preprocessing 279–282
    data sources 274–275
    exploratory analysis 279–282
    imputing variables 283–287
    mind map 274
    modeling 291–307
    project charter 273–274
    variable reduction and selection 287–291
DATA= option, TIMESERIES procedure 208
data preparation
    about 67
    data collection in 53–63
    for data mining for forecasting example 275–279
    for variable selection 104–111
    imputing missing data 84–87
    in DMAIC process 25
    in model development 8, 11–12
    matching frequencies 75–83
    merging data 83
    outliers 87–92
    software infrastructure for 32
    transactional data versus time series data 67–75
    transformations and 92–94
    using SAS Enterprise Guide 18
    using SAS Enterprise Miner 18–19
    using SAS/ETS 17–18
data quality
    about 39
    data consolidation and 40
    data glitches 41
    missing data 41
    noisy data 40–41
    time series data with short history 41–42
    time series outliers 42–43
data sources
    for data collection 57–59
    for data mining for forecasting example 274–275
Data Surveyor for SAP 4
data transformations
    data preparation and 92–94
    ESM procedure and 20
    selecting 81–83
DATASOURCE procedure 17
de-seasonalization of time series data 106–108
de-trending time series data 105–106
decision trees
    limitations of 44
    variable selection based on 120–122
DECOMP statement, TIMESERIES procedure 106
Define phase (DMAIC) 24, 26
deliverables
    in forecasting model deployment 15
    in forecasting model development 14–15
    in forecasting model maintenance 16
    in project definition 11
    in variable reduction and selection 13
demand-driven forecasting 15, 96
demand volatility 40
denominator order polynomials 158, 164–165
deployment
    See model deployment cycle
developer infrastructure 37–38
development
    See model development cycle
DFT (discrete Fourier transformation) 20
Dickey-Fuller test 126, 131
DIF= option, TIMESERIES procedure 104
differencing time series data 104–105, 167–168
digital filtering technique 105
DISAGGDATA option, HPFRECONCILE procedure 218
discrete Fourier transformation (DFT) 20
discrete wavelet transformation (DWT) 20

DMAIC process 24–26
DMM methods for forecasting
  about 95–96, 103
  time series approach 97–103, 125–134
  traditional data mining 96–97, 104–111
Durbin-Watson test 126
DWT (discrete wavelet transformation) 20
DyFor GP (Dynamic Forecasting Genetic Programming) 258

E

ECRI (Economic Cycle Research Institute) 4, 59
Eigenvector-Eigenvalue problem 96
END= option, ID statement (SIMILARITY) 245
Enterprise Guide
  See SAS Enterprise Guide
Enterprise Miner
  See SAS Enterprise Miner
ESM (Exponential Smoothing Models) 135
ESM procedure 20
ESTIMATE statement, ARIMA procedure 142, 184–185
estimation in model development 14
evangelical forecasting 48
event selection in model development 13
evolutionary computation 44, 256–258
Excel spreadsheets 61
EXPAND= option, TARGET statement (SIMILARITY) 245
EXPAND procedure 17–18, 232
expanding time interval frequencies 78–83
Explore (SEMMA work process) 18
exponential smoothing limitations 45
Exponential Smoothing Models (ESM) 135
external consultant services 37–38
external data
  data sources for 57
  extracting 61–62
  infrastructure for 34–35
extracting data 61–62, 64

F

Fisher's Kappa test 110
flowchart of work processes 8–9
FORECAST procedure 20, 22
Forecast Server/Studio
  See SAS Forecast Server/Studio
FORECAST statement, ARIMA procedure 148, 194
forecast time horizon 46, 225
forecast value analysis (FVA) 15
forecasting
  See data mining for forecasting
forecasting model deployment cycle
  See model deployment cycle
forecasting model development cycle
  See model development cycle
Friedman, Jerome 20
FVA (forecast value analysis) 15

G

garbage in, gold out trap 47
GARCH models 19
genetic programming (GP)
  forecasting models based on 256–258
  limitations of 44–45
  variable selection based on 122–124
Genetic Programming (Koza) 122
Gilliland, Michael 36, 49–50
Global Insight 2–3, 57–60
GP (genetic programming)
  forecasting models based on 256–258
  limitations of 44–45
  variable selection based on 122–124
GPLOT procedure 87
GPTools toolbox 122
gradient boosting machine 20
Guide Line Plot wizard 87

H

hardware infrastructure
  about 29
  client/server component 30–31
  cloud computing component 31
  PC network component 30
Haykin, Simon 252
high-frequency data 224–232
historical data, expectations based on 49
holdout samples 237–240
HPFDIAGNOSE procedure 222
HPFENGINE procedure 222
HPFRECONCILE procedure
  AGGDATA option 218
  BY statement 219
  data hierarchies and 232
  DISAGGDATA option 218
  RECDIFF option 218

I

ICIS (International Code Industry System) 3
ID statement
  SIMILARITY procedure 245
  TIMESERIES procedure 208, 212
identification, data
  See system structure and data identification
IDENTIFY statement, ARIMA procedure
  about 138, 173
  CROSSCORR option 182, 187
  NLAGS option 140
IMF (International Monetary Fund) 58
Import Data task 61
Improve phase (DMAIC) 24, 26
imputing missing data 84–87, 283–287
indexes, economic 58–59
infrastructure (data mining for forecasting)
  about 29
  data component 34–35
  hardware component 29–31
  organizational component 35–37
  software component 32–33
INPUT function 277
INPUT statement, SIMILARITY procedure 245
intermittent demand 220–223
internal data
  data infrastructure for 34–35
  data sources for 57
  extracting 61, 64
International Code Industry System (ICIS) 3

International Monetary Fund (IMF) 58
interval frequencies for time series data
- contracting 75–78
- expanding 78–83
- mixed-frequency forecasting 232–237

IRREGULAR statement, UCM procedure 268–269
irregular variation (ARMA models) 185–186, 265
IT services, integrating infrastructure with 37

J

JMP 17

K

KDD process 5
Koza, John 122, 256
KPIs (Key Performance Indicators) 37

L

lagged variables 111, 126, 171
LEVEL statement, UCM procedure 265–270
line segment approximations 20
log transformation method 20, 92
logistic transformation method 20, 92

M

MA (moving average) models
- about 135–136, 149–151
- limitations of 45
- MA 1 model 155–157
- model estimation, selection, and forecasting 154–155
- model identification 151–154

MABSDEV metric 98, 245
maintenance (model deployment cycle)
- avoiding bad practices in 49
- in model deployment 8, 16, 22

Makridakis, Spyros 49
manual overrides 240–241
MAPE metric 47, 285–286
MAS (Markov switching autoregressive) model 252
MDY function 277
MEANS procedure 71–72
Measure phase (DMAIC) 24, 26
merging data 83, 279–281
metadata 59–61
Middle Out reconciliation 216, 218–219
mind-mapping
- about 10–11
- data collection example 54–56
- data mining for forecasting example 274

missing data
- application issues with 41
- data mining for forecasting example 279–287
- imputing 84–87, 283–287

mixed-frequency forecasting 232–237
MMDDYY10. format 67
Model (SEMMA work process) 19
model building
- creating time series data 206–216
- data mining for forecasting example 291–307
- forecast model selection 237–240
- high-frequency data 224–232
- holdout samples 237–240
- intermittent demand 220–223
- mixed-frequency forecasting 232–237
- multivariate models 258–263
- new product forecasting 244–248
- nonlinear forecasting models 251–258
- planning versus forecasting 2, 240–241
- scenario-based forecasting 241–244
- statistical forecast reconciliation 216–219
- UCM models 263–272

model deployment cycle
- about 8
- data collection automation for 64–65
- deployment steps in 8, 15
- DMAIC process and 25
- maintenance in 8, 16
- performance tracking in 8, 16
- SAS tools for 21–22

model development cycle
- *See also* data preparation
- data collection in 64
- development steps in 8, 13–15
- DMAIC process and 25
- project definition in 8–11
- SAS tools for 17–21
- variable reduction and selection in 8, 12–13

MODEL procedure 19
Modify (SEMMA work process) 18
moving average models
- *See* MA (moving average) models

multilayer perceptron 253
multivariate models
- about 14, 135
- building 258–263
- limitations of 46

N

NAICS (North American Industry Classification System) 3, 62–63
network infrastructure 30
neural network models 135, 252–254
Neural Networks (Haykin) 252
new product forecasting 244–248
New Project wizard 84
NLAGS option, IDENTIFY statement (ARIMA) 140
NOEST option, ARIMA procedure 241
noisy data 40–41
nonlinear forecasting models
- based on evolutionary computation 44, 256–258
- based on neural networks 135, 252–254
- based on support vector machines 254–256
- features of 251–252

nonstationary series data
- about 166–167
- ADF test for 172–176
- diagnosing 168–176
- differencing 167–168

nonzero Lagrange multipliers 255
NOTSORTED option, ID statement (TIMESERIES) 212

O

ODS GRAPHICS statement 138
organizational infrastructure
  about 35
  developers component 35–36
  handling application issues 39–51
  integration with IT 37
  users component 36–37
  work process implementation 37
OUT= option, TIMESERIES procedure 208
OUTLIER statement, UCM procedure 88
outliers in time series data
  application issues with 42–43
  handling during data preparation 87–92
OUTPUT statement, VARMAX procedure 262
OUTSUM option, SIMILARITY procedure 245
overfitting, AIC and 237, 239

P

P option, SPECTRA procedure 110
PACF (partial autocorrelation) plot
  AR models and 140–144
  ARIMAX models and 189
  ARMA models and 159–161
  MA models and 152–153
Pareto-front GP 257–258
partial least squares (PLS) 118–120
PC network infrastructure 30
PCA (principal component analysis)
  limitations of 43–44
  PRINCOMP procedure and 19
  variable reduction and 12, 96
Pearson, Ronald 40
Pearson product-moment correlation 112–114, 127
performance tracking
  avoiding bad practices in 49
  in model deployment 16
  KPIs for 37
  managing expectations 46–47
Permanent Change model 189–190, 192–193
piecewise polynomials 105
planning versus forecasting 2, 240–241
PLOTS= option, TIMESERIES procedure 208
PLS (partial least squares) 118–120
PLS procedure 19–20, 22
politics of forecasting 47–48
PPI (Producer Price Index) 59–60
Prepare Time Series Data task 18, 79–83
preparing data
  See data preparation
prewhitening technique 196–203
principal component analysis
  See PCA (principal component analysis)
PRINCOMP procedure 19, 22
Producer Price Index (PPI) 59–60
product forecasting 244–248
production mode in model deployment 15
project charters
  samples of 26–27, 273–274
  writing 48
project definition in model development 9–11
project objectives 9
project roles 10
project scope 10
proximity effect 153

Q

quality, data
  See data quality
Query Builder task 83
QUERY wizard 277

R

radial basis function (RBF) 253
random walk 167, 188
rational polynomial transfer function
  about 179–180
  advantages of 135
  ARIMAX applications 185–195
  ARIMAX concepts 180–185
  ARMA example 158
  interval input variables and 195–203
RBF (radial basis function) 253
readiness checklist 50–51
RECDIFF option, HPFRECONCILE procedure 218
reconciliation
  data hierarchies and 211
  statistical forecast 216–219, 232
recurrent neural networks 254
REG procedure 19, 22, 235
Regression Analysis with Autoregressive Errors task 21

S

S option, SPECTRA procedure 110
Sample (SEMMA work process) 18
SAS Data Surveyor for SAP 4
SAS Enterprise Guide
  about 17
  ARIMA Modeling and Forecasting task 21
  Basic Forecasting task 21
  Create Time Series Data task 18, 69–71
  data mining for forecasting example 287
  data preparation using 18
  forecasting using 21–22
  Guide Line Plot wizard 87
  Import Data task 61
  Prepare Time Series Data task 18, 79–83
  Query Builder task 83
  Regression Analysis with Autoregressive Errors task 21
  Summary Statistics task 75–78
  TRANSPOSE wizard 276
  variable selection methods 129–131
SAS Enterprise Miner
  about 17
  data preparation using 18–19
  Decision Tree node 20, 120–122
  Explore tab 20
  forecasting using 22
  Gradient Boosting node 20
  Input Data Source node 60
  metadata options 60–61
  Model tab 20
  Modify tab 20
  Partial Least Squares node 118–120

SAS Enterprise Miner (*continued*)
- Principal Components node 20
- Regression node 20, 114–115
- StatExplore node 112–113
- Time Series Data Mining tab 20
- TS Dimension Reduction node 20
- TS Similarity node 20
- Variable Clustering node 20
- variable reduction and selection using 20, 111–112
- Variable Selection node 20, 116–118

SAS/ETS
- about 16
- data preparation using 17–18
- forecasting using 20–22
- variable reduction and selection using 19

SAS Forecast Server/Studio
- about 17, 21–22
- data mining for forecasting example 287, 291–298
- handling transactional data 69, 72–74
- holdout samples and 238
- imputing missing data 84–87
- intermittent demand examples 222
- outlier example 88–92
- scenario-based analysis 241–244
- Series View analysis platform 92–94, 242
- Stored Process Reports 132
- variable selection and 124–125

SAS/GRAPH 16
SAS High-Performance Forecasting 16
SAS/IML 16
SAS Microsoft Add-In 15, 17, 22
SAS/STAT
- about 16
- forecasting using 22
- variable reduction and selection using 19

SAS tools
- about 16–17
- for model deployment cycle 21–22
- for model development cycle 17–21
- selecting for data mining for forecasting 22

SAS Web Report Studio 22
SBC (Schwartz's Bayesian criterion)
- about 20
- AR models and 145
- ARMA models and 162
- data mining for forecasting example 303
- MA models and 155–156

scenario-based forecasting 15, 241–244
SDIF= option, TIMESERIES procedure 104
SEASON statement, UCM procedure 232, 267, 270
self-exciting threshold autoregressive (SETAR) model 252
Selukar, Rajesh 263
SEMMA work process 5, 18–19
Series View analysis platform 92–94
SETAR (self-exciting threshold autoregressive) model 252
SETMISSING= option, ID statement (TIMESERIES) 208
similarity metric
- data mining for forecasting example 291
- variable reduction methods and 97–98
- variable selection methods and 13, 126, 130

SIMILARITY procedure
- about 19–20, 244–245
- ID statement 245
- INPUT statement 245
- OUTSUM option 245
- TARGET statement 245
- VAR statement 98–99

singular value decomposition (SVD) 20
Six Sigma work process 23–28
SLOPE statement, UCM procedure 266, 270
SMAPE model 292, 297, 302
smooth transition regression (STR) model 252
software infrastructure
- about 32
- for data collection 32
- for data mining 32
- for data preparation 32
- for forecasting 33
- selection criteria 33

SPECTRA procedure
- about 22
- ADJMEAN option 110
- P option 110
- S option 110
- seasonality detection by 109–111
- VAR statement 110
- WEIGHTS statement 110
- WHITETEST option 110

square root transformation method 20, 92
stakeholder interests 48
START= option, ID statement (SIMILARITY) 245
STATESPACE procedure 21
stationary time series 104
statistical baseline
- for performance assessment 16
- tracking manual overrides 240–241

statistical forecasts
- manual overrides for 240–241
- reconciliation for 216–219, 232

statistical inference, data mining and 5
stepwise regression method
- limitations of 43
- variable selection based on 114–116, 132–134

stored processes, model deployment via 22
STR (smooth transition regression) model 252
structural equations 136
SUBSTR function 277
SUM option, MEANS procedure 71–72
Summary Statistics task 75–78
support vector machines (SVM) 254–256
SVD (singular value decomposition) 20
SVM (support vector machines) 254–256
symbolic regression 44
system structure and data identification
- data collection and 53–55
- in DMAIC process 25
- in model development 10–11, 14
- knowledge acquisition and 54–55

T

TARGET command 98
TARGET statement, SIMILARITY procedure 245
Temporary Change model 189–194
time series data
- *See also* ARMA (autoregressive moving average) models

about 4, 206
contracting intervals in 75–78
creating with accumulation methods 206–210
creating with aggregation methods 210–216
de-seasonalization of 106–108
de-trending 105–106
differencing 104–105, 167–168
ESM procedure and 20
explosion of availability 2–3
forecast model selection 237–240
global systemic variation in 196
holdout samples and 239
interval frequencies in 75–83, 232–237
limitations of forecasting 45
metadata considerations 59
outliers in 42–43, 87–92
preparing with SAS Enterprise Guide 18–19
seasonal adjustment by 18, 108–109
spectral analysis of 109–111
stationary 104
transactional data versus 3, 5, 67–75
transfer function models 126–129
trend analysis of 105
variable reduction in 12–13, 97–103
variable selection in 13, 125–134
with short history 41–42
Time Series Data Mining (TSDM) 19, 22
TIMESERIES procedure
about 207–208, 246
accumulation methods for 209–210
BY statement 212
Create Time Series Data task and 18, 69, 71
DATA= option 208
DECOMP statement 106
DIF= option 104
high-frequency data example 225–226
ID statement 208, 212
multivariate model example 259
OUT= option 208
PLOTS= option 208
SDIF= option 104
timestamps
aligning data to 63
transactional data and 67
tools
See SAS tools
Top Down reconciliation 216, 218, 232
training users 15
transactional data
accumulating for data hierarchies 211
cluster analysis of 96
time series data versus 3, 5, 67–75
timestamps in 67
variable selection methods 13
transfer function models
about 126–128
ARIMA models and 166–176
components of 128
order rules for 128–129
rational polynomial 135, 158, 179–203
transformation methods
See data transformations
TRANSPOSE wizard 276

trend analysis of time series data 105
TSDM (Time Series Data Mining) 19, 22

U

UCM (unobserved component models)
about 14, 135, 231
building 263–272
UCM procedure
about 21, 263, 271–272
BLOCKSEASONAL statement 232
IRREGULAR statement 268–269
LEVEL statement 265–270
OUTLIER statement 88
SEASON statement 232, 267, 270
SLOPE statement 266, 270
univariate forecasting 4, 14
unobserved component models
See UCM (unobserved component models)
user infrastructure 36–37

V

VAR statement
SIMILARITY procedure 98–99
SPECTRA procedure 110
VARCLUS procedure 19–20, 22, 96–103
variable cluster analysis 12, 96–103, 288–289
variable reduction and selection
about 12–13, 103–111
based on decision trees 120–122
based on genetic programming 122–124
based on Pearson product-moment correlation 112–114
based on SAS Enterprise Miner Partial Least Squares node 118–120
based on SAS Enterprise Miner Variable Selection node 116–118
based on stepwise regression 114–116, 132–134
comparison of methods for 124–125
data mining for forecasting example 287–291
time series approach 12–13, 97–103, 125–134
traditional data mining and 12–13, 96–97, 104–111
using SAS Enterprise Miner 20
using SAS/ETS 19
using SAS/STAT 19
VARMAX procedure
about 21–22
multivariate models and 261
OUTPUT statement 262
variable reduction and 96
VIP metric 118–120

W

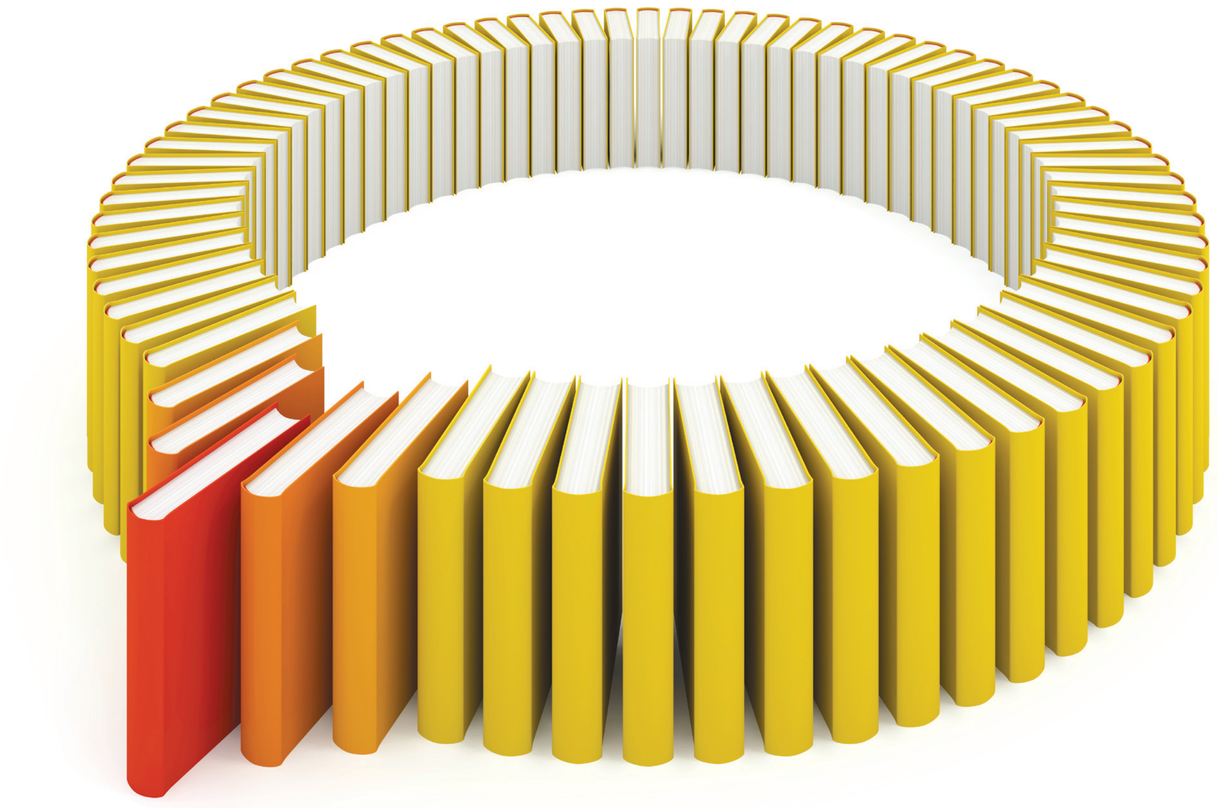
WEIGHTS statement, SPECTRA procedure 110
What-If scenarios 15, 241–244
white noise process
AR models and 136–137, 142, 145, 147
ARMA models and 159, 162
MA models and 152
WHITETEST option, SPECTRA procedure 110

work processes (data mining for forecasting)
implementing infrastructure for 37
integrating in Six Sigma 23–28
key steps in 9–16
process overview 7–9
SEMMA 5, 18–19
with SAS tools 16–22

X

X11 procedure
about 22
seasonal adjustment by 18, 108–109
X12 procedure
about 22
seasonal adjustment by 18, 108–109

From *Applied Data Mining for Forecasting Using SAS®* by Tim Rey, Arthur Kordon, and Chip Wells. Copyright © 2012, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.



Gain Greater Insight into Your SAS® Software with SAS Books.

Discover all that you need on your journey to knowledge and empowerment.

support.sas.com/bookstore
for additional books and resources.

sas®
THE POWER TO KNOW®

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. © 2013 SAS Institute Inc. All rights reserved. S107969US.0613