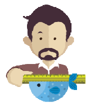
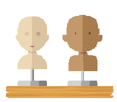
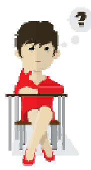
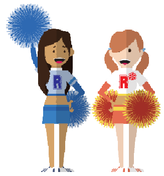
Contents



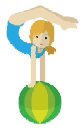


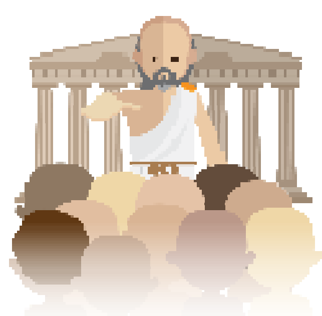




Chapter 11 - Gaining Credibility Through Root-Cause Analysis and Exception Handling 281

Chapter 12 - The Final Forecast Numbers: Reconciling Change & Chance *313*





2

Smarter Forecasting Is Mostly about Data

Improving Data Quality through Data Exploration and Visualization

Time series analysis consists of all the techniques that, when applied to time series data, yield, at least sometimes, either insight or knowledge, AND everything that helps us choose or understand these procedures.

JOHN W. TUKEY, pioneer data scientist (1980)

This chapter deals with the statistical basis for data analysis in demand forecasting. As in the modeling process presented in Chapter 1, you will find that

- data analysis is open-ended and iterative in nature
- the steps may not always be clearly defined
- the nature of the process depends on what information is revealed at various stages. At any given stage, various possibilities may arise, some of which will need to be explored separately
- As a statistical methodology, much data analysis in demand forecasting is informal and exploratory.

However, it is important to realize that

- an understanding of historical data in forecasting demand will be enhanced when we can identify key patterns in a time series
- analyzing data is part of the smarter forecasting process. For example, when data contain trends, contain seasonal patterns, or have outliers, the use of some commonly used projection techniques, such as moving averages is inappropriate
- data visualizations are beneficial in describing the shape or distribution of data patterns, model residuals and forecast errors
- assuming unrealistic (*non-robust*) distributions for forecast errors can lead to misleading results when assessing forecasting performance. Most testing procedures implicitly assume that the underlying data follow a ***normal***

***distribution**,* and this may not often be the case in demand forecasting applications

- most demand planners and forecasters use accuracy measurements to make oversimplified use of the arithmetic mean (as in the case of the MAPE, MAD, MASE) for (*non-outlier-resistant*) forecast performance measurement, not realizing that such data rarely follow conventional statistical assumptions

Smarter Forecasting Is Mostly about Data

When a multidisciplinary research study group at Princeton University undertook a study of the paired uses of electricity and gas in townhouses, it contacted the residents of Twin Rivers, a nearby planned community in New Jersey. Over a five-year study period, it learned how to eliminate three-quarters of the energy used by the furnace in quite ordinary, reasonably well-built townhouses, as chronicled in *Saving Energy in the Home: Princeton's Experiments at Twin Rivers,* edited by Robert H. Socolow (Cambridge, MA: Ballinger, 1977).

The purpose of the Princeton study, during a winter in the mid-1970s, was to examine differences in energy use and make comparisons with structural aspects of the 152 individual townhouses and the behavioral aspects of their inhabitants. As an applied statistician (a.k.a. data scientist), I took great delight in being a participant and was intrigued by later looking at the results and the data from the study. I was a resident at Twin Rivers at the time, not realizing that some new analysis techniques used on the data would eventually be published in 1977 in the ground-breaking book *Exploratory Data Analysis* by data science pioneer John W. Tukey (1915–2000).

The data were gathered automatically through a special device that was hooked up to the landline telephones and the energy sources in the home. There were questions to be answered periodically about our lifestyle, the details of which have long escaped my memory. Nevertheless, some novel uses of graphing techniques with schematic data plots (**data visualization**) can be found throughout this book. These techniques, new at the time, have now become a familiar part of many business statistics books.

Exploring Data Patterns

Studying the patterns in the data improves the forecaster's chances of successfully modeling data for forecasting applications. Through **exploratory data analysis** (EDA), a demand forecaster can start the important task of finding **factors** (drivers of demand) that are generally quantitative in nature.

Tukey likens EDA to detective work: "A detective investigating a crime needs both tools and understanding. If he/she has no fingerprint powder, the detective will fail to find fingerprints on most surfaces. If detectives do not understand where criminals are likely to have put their fingers, they will not look in the right places." A planned forecasting and modeling effort that does not include provisions for exploratory data analysis often misses the most interesting and important results; but it is only a first step, not the whole story.

Exploratory data analysis means looking at data, absorbing what the data are suggesting, and using various summaries and display methods to gain insight into the process generating the data.

Many business forecasting books describe a variety of classical ways to summarize data. For the practitioner, an entertaining yet informative cartoon guide covering these is Gonick and Smith's *A Cartoon Guide to Statistics,* published in 1993. For example, the familiar histogram is widely used in practice. In addition, there are a number of lesser-known techniques that are specifically useful for analyzing large quantities of data that have become accessible as a result of the increased flexibility in data management, computer processing, and predictive analytics. Because of their potential value to demand forecasting, we describe them in some detail.

Learning by Looking at Data Patterns

Because most forecasting techniques require data, a forecaster analyzes the availability of data from both external (outside the company) and internal (within the company or its industry) sources. For example, one potential source of internal data is a corporate data warehouse or **Enterprise Resource Planning** (ERP) system, which normally contains a rich history of product sales, shipments, prices, revenues, expenses, capital expenditures, and marketing programs.

The availability of **external data** is improving rapidly. Most of the required demographic factors (age, race, sex, households, and so forth), forecasts of **economic indicators**, and related variables can be readily obtained from computerized data sources and from industry and government publications on the Internet.

With the explosion of Internet websites, potential sources of valuable data are becoming limitless. With unstructured data, the need for data mining tools has become a necessity for exploring potential sources of data for consumer analyses and predictive modeling purposes.

Judging the Quality of Data

The analysis of data for forecasting purposes requires a careful consideration of the quality of data sources. **Data quality** is important for modeling, because a model based on historical data will be no better than the quality of its data source.

Definitions may vary because of changes in the structure of an organization, accounting procedures, or product and service definitions. As part of the forecasting process, throughout this book various kinds of demand data will be related to economic/demographic indicators, survey data, and other external data in a number of examples and spreadsheet exhibits.

There are several criteria that can be applied to data to determine their appropriateness for modeling:

- **Accuracy**. Proper care must be taken that the required data are collected from a reliable source with proper attention given to accuracy. **Survey data** exemplify the need to ensure accurate data: Survey data are collected by government agencies and research firms from questionnaires and interviews to determine future plans of consumers and businesses. The **consumer confidence index** in Figure 2.1 is the result of a survey made by the Conference Board (*https://www.conference-board.org/data/consumerconfidence.cfm*).These data have certain limitations because they reflect only the respondents' anticipation (what they expect others to do) or expectations (what they themselves plan to do), not firm commitments. Nevertheless, such information may be regarded as a valuable aid to forecasting demand directly or as an indication of the state of consumer confidence concerning the economic outlook.

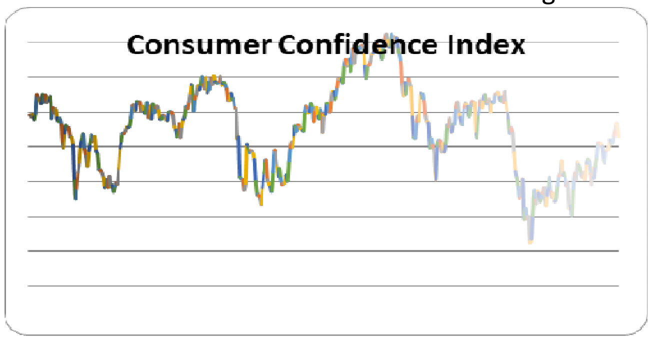


Figure 2.1 Time plot of the Conference Board's Consumer Confidence Index - US (1977 - 2014 (Units: 1985 = 100 - *Source*: https://www. Conference-board.org/); the variation in the Consumer Confidence Index shows the dominant 89% trend-cycle effect.

Based on an ANOVA decomposition (cf. MS Excel Data Analysis Add-in > ANOVA: Two-Factor without Replication), we determine that the total variability in the data is made up of 1% seasonality (Month effect), 89% trend-cycle (Year effect) and 10% unknown. When modeling we can ignore models with seasonal forecast profiles, like the Holt-Winters exponential smoothing family, but when using these data as an explanatory factor or driver of demand in causal models, we recognize its value in explaining primarily economic cycle and trending behavior.

- **Conformity**. The data must adequately represent the phenomenon for which it is being used. If the data purport to represent economic activity, the data should show upswings and downswings in accordance with past historical **business cycle** fluctuations. Data that are too smooth or too erratic may not adequately reflect the patterns desired for modeling. The Federal Reserve Board **Index of Industrial Production** (Figure 2.2) is an example of a cyclical indicator of the economy. It is evident that the data are consistent with economic *expansions* and *contractions* (to be discussed in Chapter 7). The index of industrial production measures changes in the physical volume of the output of manufacturers, mineral suppliers, and electric and gas utilities. The index does not cover production on farms, in the construction industry, in transportation, or in various trade and service sectors. Since the U.S. **Federal Reserve Board** (FRB) first introduced the index in 1920, it has been revised from time to time to take account of the growing complexity of the US economy, the availability of more data, improvements in statistical processing techniques and refinements in methods of analyses. Such indices are now widely available in many other national economies.

The ANOVA decomposition of the original (unadjusted) FRB Index of Industrial Production suggests that the total variability in the data is made up of 10% seasonality (Month effect), 89% trend-cycle (Year effect) and 1% irregular (Not identified). When using these data as an explanatory factor or driver of demand in causal models, we recognize its value in explaining primarily trend and economic cycles.

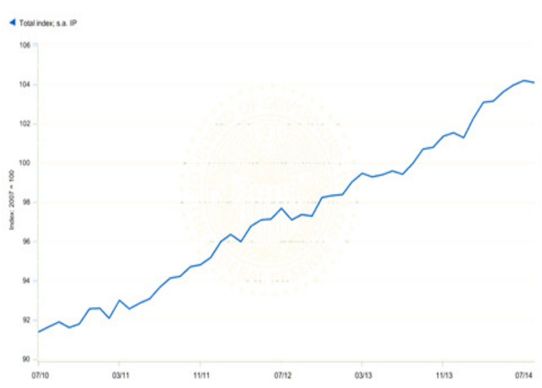


Figure 2.2 Time plot of the monthly seasonally adjusted index of industrial production from the Federal Reserve Board, July 2010–August 2014 (Units: 2007 = 100). (*Source*: Board of Governors of the Federal Reserve System - ***http://www.federalreserve.gov/***)

- **Timeliness**. It takes time to collect data. Data collected, summarized, and published on a timely basis are of greatest value to the forecaster. Often preliminary data are available first, so that the time delay before the data are declared official may become a significant factor. Demographic data may fall into this category for many users. The monthly housing starts data shown in Figure 2.3 are demographic data reported by contractors and builders for use by government and private industry. Such external data are of course, subject to adjustment because of data collection delays and reporting inaccuracies.

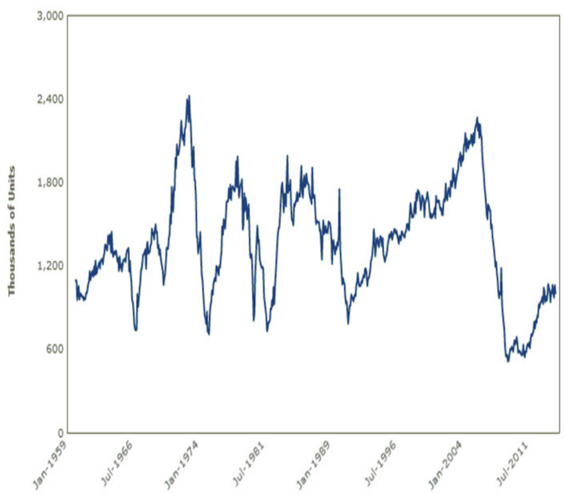


Figure 2.3 Time plot of monthly U.S. new housing starts (seasonally adjusted annual rates). (*Source*: *www.census.gov*)

Once a factor has been identified as a key driver of demand, we should not assume that its usefulness is good for all times. Data need to be analyzed on an ongoing basis for quality assurance. For example, a preliminary ANOVA decomposition, while useful to assess the strength of a seasonal or trend-cycle effect, can point to a change in the composition of components of variability. For the ten-year period 1960–1969 (Figure 2.4 *left frame*), the total variability in the data is made up of 0.1 % seasonality, 64.5% trend and 35.4% irregular (unknown). On the other hand, for the ten-year period 2004–2013 (Figure 2.4 *right frame*), the total variability in the data is made up of 0.3 % seasonality, 96.5% trend and 3.2% irregular (unknown). Because these are seasonally adjusted data, no significant seasonal effect can be present. The non-trending variability in the early period, masked by the deep dip, is suggested by the high irregular contribution. When using these data as an explanatory factor or driver of demand in causal models, we recognize its value in explaining primarily economic and demographic consumer behavior.

- **Consistency**. Data must be consistent throughout the period of their use. When definitions change, adjustments need to be made in order to retain logical consistency in historical patterns. The monthly demand for refrigerator sales, shown in Figure 2.5, is an example of internal data that would be made available to a demand forecaster of a consumer goods manufacturer. It shows a consistent (seasonal with business cycle?) pattern. If the data pattern shows abrupt level changes or unusual variation, the forecaster should check into how the data are defined.

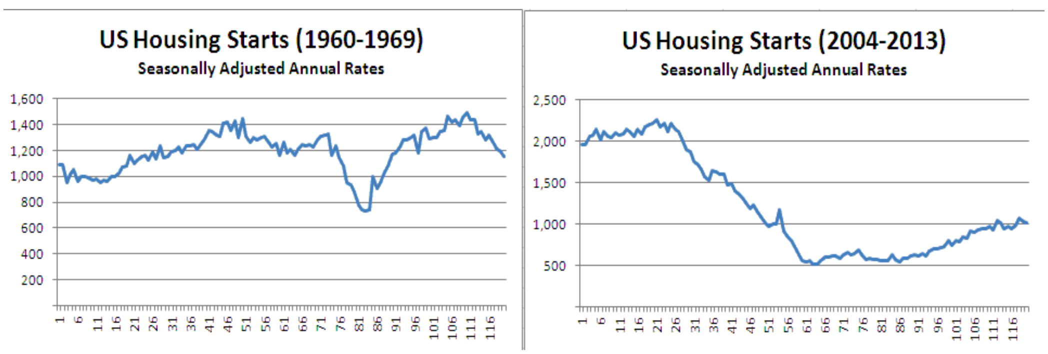


Figure 2.4 Time plots of monthly new housing units started in the United States (seasonally adjusted annual rates for periods 1960–1969 and 2004–2013). (*Source*: www.census.gov)

In Figure 2.5 (*left frame*), the total variability in the refrigerator data is made up of 81% seasonality, 7% trend and 12% Other (unknown). It can be seen that the second seasonal peak appears unusually low (domain expert's explanation: lack of inventory resulting in lost sales). A simple interpolation between the seasonal peaks makes the data more representative of the demand.

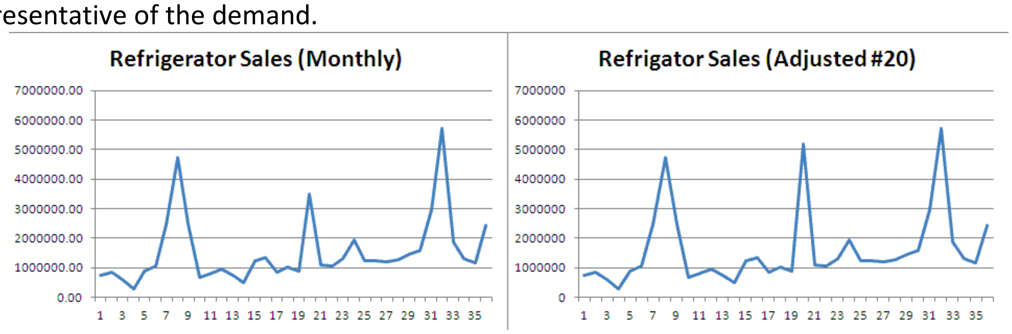


Figure 2.5 Time plots of monthly sales of home refrigerators.

It is important to recognize that modeling with unexamined data can distort the expected forecast profile (*change*) as well as the uncertainty (width of the prediction limits: *chance*). With the adjustment, the total variability in the data is now made up of 87%

seasonality, 4% trend and 9% Other (unknown). The seasonality is still dominant but the unknown component, comprised of the uncertain variation is significantly reduced—the impact of just a single data point!

When using these data as demand in causal models, we recognize the need for explanatory factors explaining primarily seasonality or habit in consumer behavior, along with a measure of the nature of the uncertainty (Other).

The consequences of jumping into the modeling phase before thoroughly investigating the quality of the data and checking for anomalies can be quite severe. For instance, if we blindly applied a credible trend-seasonal model (e.g. Holt-Winters exponential smoothing or an ARIMA $(011)(011)_{12}$ "airline" model) to the original data, we would see that the seasonal peak profile becomes level and the uncertainty range is wide; on the other hand with a single point adjustment based on informed judgment and the same trend-seasonal model, the results become much more credible and practical (Figure 2.6). Not only is the forecast profile representative of what may lie in the future, but the prediction limits have become much narrower, indicating a tamed (reduced) uncertainty. The variability in the Other category of the exploratory ANOVA decomposition is reduced and our forecasts become more precise.

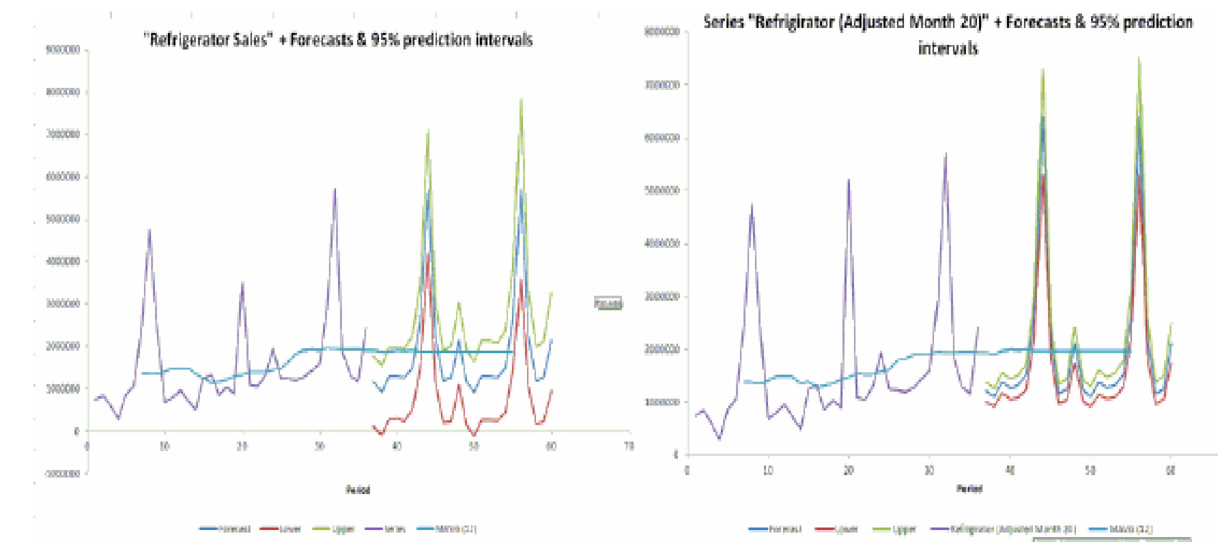


Figure 2.6 Improvement in the forecast profile for refrigerator sales for original (*left*) and outlier-adjusted (*right*) period #20.