

Chapter 2

Data Sets

Data comes in many forms. The current age of big data floods us with numbers accessible from the Web. We have trading data available in real time (which caused some problems with automatic trading algorithms, so some trading sites impose a delay of 20 min or so to make this data less real-time). Wal-Mart has real-time data from its many cash registers enabling it to automate intelligent decisions to manage its many inventories. Currently a wrist device called a Fitbit is very popular, enabling personal monitoring of individual health numbers, which have the ability to be shared in real-time with physicians or ostensibly EMT providers. The point is that there is an explosion of data in our world.

This data has a lot of different characteristics. Wal-Mart might have some data that is very steady, with little trend. Things like demand for milk, which might have a very predictable relationship to population. You might need to know a bit more about age group densities, but this kind of data might be big if it is obtained from grocer cash registers across a metropolitan region, but is likely to be fairly stable. Other kinds of data follow predictable theories. A field is a science when it is understood to the extent that it can be mathematically modelled. (That's why some deny that economics is a science—it can be mathematically modelled, but it is often not correct in its predictions.) Other types of data behave in patterns impossible to define over a long term.

Real data, especially economic data, is very difficult to predict [1]. Examples include the Internet bubble, the Asian financial crisis, global real estate crises, and many others just in the past 15 years. Nobody predicted failure of Long-Term Capital Management in the 1990s, nor Enron and WorldCom a few years later, nor Lehman Brothers, Northern Rock, and AIG around the global real estate crisis. Physics and engineering benefit from consistent behavior of variables, and can be

Electronic supplementary material The online version of this chapter (doi:[10.1007/978-981-10-2543-3_2](https://doi.org/10.1007/978-981-10-2543-3_2)) contains supplementary material, which is available to authorized users.

modeled to the extent that rockets can be sent to the moon and Mars and be expected to get there. Casinos have games that, given the assumption that they are fair, have precise probabilities. But when human choice gets involved, systems usually get too complex to accurately predict.

The field of complexity [2] was discussed by John Holland, and has been considered by many others. Holland viewed complicated mechanisms as those that could be designed, predicted, and controlled, such as automobile engines or pacemakers, or refineries. Complicated mechanisms can consist of a very large number of interacting parts, but they behave as their science has identified. Complex systems, on the other hand, cannot be completely defined. Things such as human immunology, global weather patterns, and networks such as Facebook and LinkedIn consist of uncountable components, with something like order emerging, but constantly interacting making precise prediction elusive. Ramo [3] described how this emerging world has grown in complexity to the point that a new paradigm shift has occurred that will change the world. People have Internet connections that can no longer be controlled.

Big data thus consists of an infinite stream of constantly generated data. Prices of stocks change as buyers and sellers interact. Thus trading data is instantaneous. Data mining tools described in this book can be applied to this real-time data, but in order to describe it, we need to look at it from a longer perspective. We will focus on data related to business, specifically Chinese trading data. While we will be looking at a small set of monthly data, that is to make it more understandable. We will also look at daily data, but the methods presented can be applied to data on any scale. Applying it to real-time data makes it big data.

We will consider three data series. The first is the price of gold, which represents a view of investor conservatism. The second is the price of Brent crude oil. We will use monthly data for the period 2001 through April 2016 for this overview. The third data set consists of four stock indices: the S&P 500 represents conservative US investor views; NYSE is the New York Stock Exchange; Eurostoxx is a European stock index; MXCN is the Morgan Stanley Capital Index for a composite of Chinese investment options will be presented in Chap. 4.

2.1 Gold

The price of gold in US dollars per troy ounce is shown in Fig. 2.1.

We view the price of gold as an indicator of investor conservatism—a high price of gold indicates concern over investment opportunities. The time series shows a very steady behavior from the beginning of 2001 until a bit of a jump in 2006, then on to a peak in early 2008, a slight drop with the realty bubble in 2008 followed by an unsteady rise to nearly \$1800 per ounce in 2011, followed by a drop from late 2012 to its hovering around \$1200 per ounce in early 2016. Thus the data from 2001 through 2005 demonstrates linearity, the kind of data that linear regression can forecast well. That is, until it starts to display nonlinear behavior, as it does

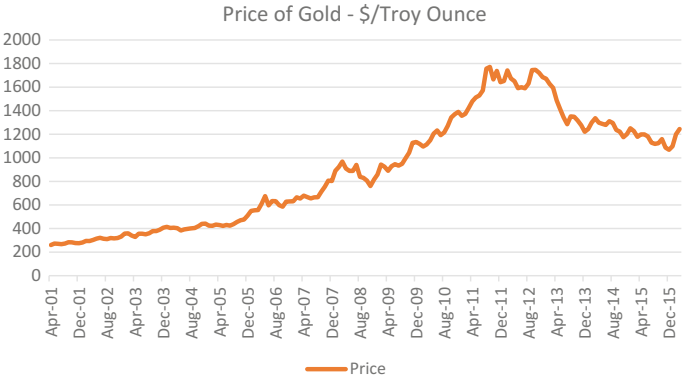


Fig. 2.1 Gold

starting in 2006. Figure 2.2 shows this linear trend, explaining 0.957 of the change in the price of gold by the formula $Trend = 254.9194 + 3.787971 \cdot Time$ (with $Time = 1$ in January 2001).

For the entire range 2001 through 2015 the linear regression for this data explains nearly 80 % of the change. This time series has interesting cyclical behavior that might lead to improvement over the linear model. It is viewed as an input to forecasting other variables, such as oil or stock indices.

In general, economic time series tend to display a great deal of nonlinearity, consisting of a series of switchbacks and discontinuities. The linearity displayed by the price of gold from 2001 through 2005 was thus atypical. The factor that led to nonlinearities in 2006 was probably the precursor to the high levels of risk perceived in the economy, especially in real estate, that led to near-economic collapse in 2008.

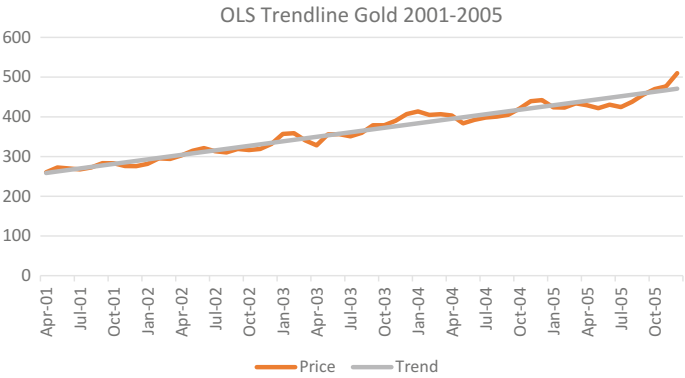


Fig. 2.2 OLS trendline gold 2001–2005

2.2 Brent Crude

The price of crude oil reflects reliance upon fossil fuel for transportation. There are many forces that work on this time series, including general economic activity, control by producers (OPEC), alternative sources (non-fossil fuels, new production from fracking), and the policies of many governments. This series was very linear through 1972, but has had very interesting behavior since 1973, when OPEC became active in controlling member production. There have been many governmental (and OPEC) policy actions (to include wars) that have seen high degrees of nonlinearity in the price of crude oil. Another major technological change was the emergence of fracking around 2006. Figure 2.3 displays the time series for Brent North Sea oil, one source of crude oil.

In 2001 the thinking was that peak oil had been reached, so that the price could be expected to rise as supply was exhausted. You can see a slight increase through the middle of 2006. This was followed by a noted drop in price, followed by a steep increase until the middle of 2008 as the global economy surged. The sudden drop in 2008 came from the bursting of the real estate bubble, seeing a dramatic decrease from a high of about \$135 per barrel in early 2008 to just over \$40 per barrel later in the year. The price came back, reaching over \$120 in 2011, and remaining in the 100–120 range until the middle of 2014, when the price sharply dropped to around \$50 per barrel, and after a short recovery dropped again into the \$30s.

As a time series, crude oil has very volatile behavior. Time itself thus does not explain why it changes. We will run a linear regression, which explains just over 40 percent of the change by r-squared measure. But clearly there are other factors explaining the price of oil, such as general economic activity as well as perceptions of risk. We will demonstrate the behavior of ARIMA and GARCH models that take advantage of cyclical time series content along with trend. We will demonstrate multiple regressions for this time series as well. Keep in mind that our purpose is to demonstrate forecasting methods, not to explain why this particular series behaves as it has.

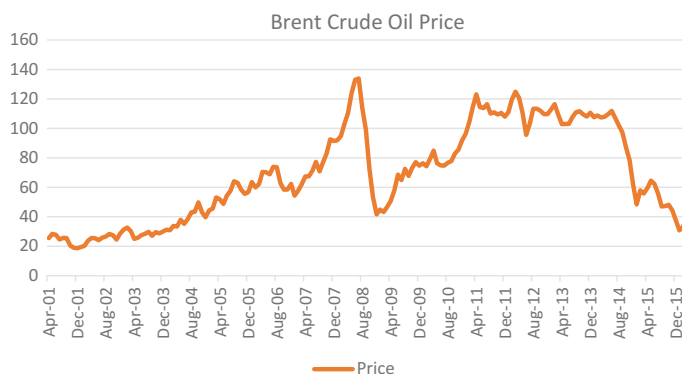


Fig. 2.3 Brent crude oil

2.3 Stock Indices

The S&P 500 is a relatively conservative stock index based upon a market basket of 500 stocks listed on the New York stock exchange. Its monthly behavior is displayed in Fig. 2.4.

This index displayed negative returns from the beginning of 2001 through 2003, recovered slightly through 2007, suffered a severe decline through the middle of 2009, after which it has slowly recovered to a new peak in early 2015 and has been fairly steady since. We view this series as a measure of relatively conservative investor confidence. The simple linear regression indicates that time by itself explains 52 % of its change, but there are obvious cyclical behaviors that might provide greater forecasting accuracy from ARIMA or GARCH models.

We have other indices as well, to include the New York Stock Exchange index, a broader measure of which the S&P is a component. Results are very similar to the S&P 500, as shown in Fig. 2.5.



Fig. 2.4 S&P

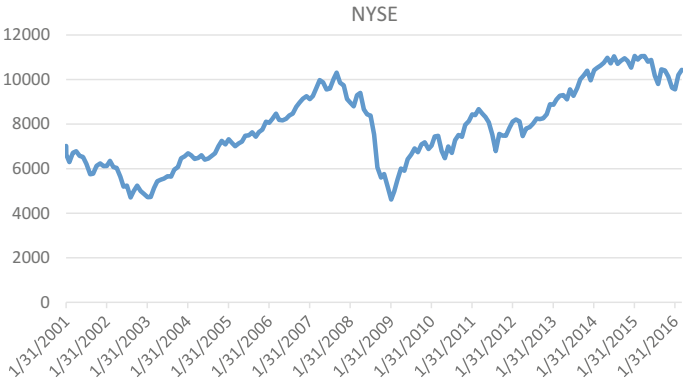


Fig. 2.5 NYSE

We include Eurostoxx (see Fig. 2.6) reflecting European stock investment and the Shenzhen index representing Chinese stock investment.

Eurostoxx can be seen to have not recovered from either their 2001 data set start, or the peak before the 2008 real estate bubble collapse. Yet it has seen something of an increase since the end of 2012. The linear regression model provides an r-square measure of only 0.045, or 4.5 %. Clearly other factors explain its performance beyond time.

We will model the monthly price of the MSCI (Morgan Stanley Capital International) China Index. A description of this time series is found at www.MSCI.com/China. This index was launched in October 1995, with back-tested data calculated. MSCI is intended to reflect Chinese mainland equity markets, to include trade on the Shanghai, Shenzhen, and Hong Kong exchanges, of both state-owned and non-state-owned shares. Monthly data for the period January 2001 through April 2016 yielded the time series shown in Fig. 2.7.

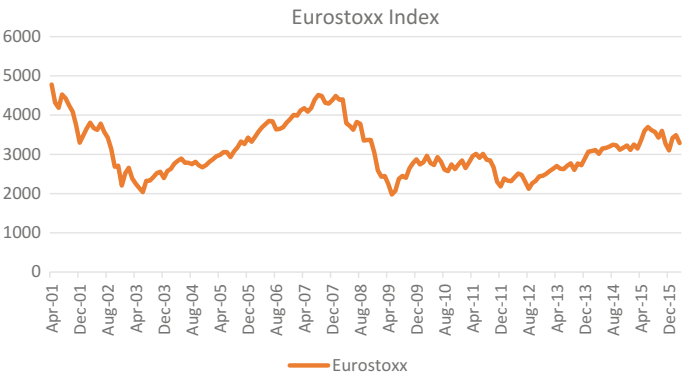


Fig. 2.6 Eurostoxx



Fig. 2.7 MSCI China

2.4 Summary

The data we have presented are economic time series, which usually involve high levels of nonlinearity over a long enough period of time. Sometimes these series display stability, such as the price of gold in the period 2000 through 2005. In that case, simple models, as presented in the next chapter, work quite well. These include moving average and ordinary least squares regression. But economic time series usually become more interesting, and thus harder to forecast. One approach is to build causal regression (covered in Chap. 4) or regression tree models (Chap. 5). Autoregressive moving average (ARIMA) and generalized autoregressive conditional heteroscedasticity (GARCH) models are more powerful tools to reflect cyclical behavior, and will be covered in subsequent chapters.

References

1. Makridakis S, Taleb N (2009) Decision making and planning under low levels of predictability. *Int J Forecast* 25:716–733
2. Holland JH (1995) *Hidden order: how adaptation builds complexity*. Perseus Books, Cambridge
3. Ramo JC (2016) *The seventh sense: power, fortune and survival in the age of networks*. Little, Brown and Company, NY

Predictive Data Mining Models

Olson, D.L.; Wu, D.D.

2017, XI, 102 p. 54 illus., 48 illus. in color., Hardcover

ISBN: 978-981-10-2542-6