Using Moving Averages to Smooth Time Series Data

By [Jim Frost](https://statisticsbyjim.com/author/statis11_wp/) [10 Comments](https://statisticsbyjim.com/time-series/moving-averages-smoothing/#comments)

Moving averages can smooth time series data, reveal underlying trends, and identify components for use in statistical modeling. Smoothing is the process of removing random variations that appear as coarseness in a plot of raw time series data. It reduces the noise to emphasize the signal that can contain trends and cycles. Analysts also refer to the smoothing process as filtering the data.

Developed in the 1920s, the moving average is the oldest process for smoothing data and continues to be a useful tool today. This method relies on the notion that observations close in time are likely to have similar values. Consequently, the averaging removes random variation, or noise, from the data.

In this post, I look at using moving averages to smooth time series data. This method is the simplest form of smoothing. In future posts, I’ll explore more complex ways of smoothing.

What are Moving Averages?

Moving averages are a series of averages calculated using sequential segments of data points over a series of values. They have a length, which defines the number of data points to include in each average.

One-sided moving averages

One-sided moving averages include the current and previous observations for each average. For example, the formula for a moving average (MA) of X at time *t* with a length of 7 is the following:

MA_{7} = {\displaystyle \frac {X_{t-6}+X_{t-5}+X_{t-4}+X_{t-3}+X_{t-2}+X_{t-1}+X_{t}}{7}}

**[What is time series forecasting in data science?](https://www.quora.com/What-is-time-series-forecasting-in-data-science" \t "_blank)**

Time series is **a sequence of data points recorded in time order**, often taken at successive equally spaced points in time. Time series data can be taken yearly, monthly, weekly, hourly, or even by the minute.

Each successive value depends on the previous values. For example, weather temperature increases almost day by day or remains constant in the month of June. It is a kind of trend.

Some definitions associated with Time Series analysis are:

1. **Level**. The baseline value for the series if it were a straight line
2. **Trend**. The optional and often linear increasing or decreasing behavior of the series over time.
3. **Seasonality**. The repeating patterns or cycles of behavior over time.
4. **Noise**. The values which are not dependent on time or is very different from the rest of the values

**[What is arima in time series forecast?](https://www.quora.com/What-is-arima-in-time-series-forecast" \t "_blank)**

In order to understand what ARIMA is in time series forecasting, it is important to discern ARMA from ARIMA models and comprehend fundamental terminology (in bold) for forecasting.

Note that ARMA stands for “Autoregressive Moving Average” and ARIMA stands for “Autoregressive Integrated Moving Average”. The main difference in ARIMA models is that ARIMA models involve **differencing** the original time series data used for forecasting. The notation for ARMA is ARMA(p,q), whereas the notation for ARIMA is ARIMA(p,d,q). The “p” stands for the number of autoregressive terms, “d” is the number of differencing needed, and “q” is the number of lagged forecast errors in the prediction equation (of best fit). The only case in which an ARIMA model can be expressed as an ARMA model is when there is no differencing needed to make the time series **stationary**(basically, ARMA(p,q)=ARIMA(p,0,q)).

**Differencing** time series means forming new time series by subtracting observation 1 from time 2, observation 2 from observation 3, and so on. The point of this is to remove certain trends, such as seasonality, downward/upward trends (correlation), or inconsistent variance in time series data. You can observe such trends just by observing the data. For example, you can detect seasonality by observing if there are periodic patterns. You can also detect inconsistent variance by observing if the “spikes” in the data are of varying heights. Also observe if there are tiny or large spikes and jumps in the data, which implies autocorrelation. Ideally, narrow spikes with little to no flat areas and random jumps implies low autocorrelation. Contrary, time series data with wider spikes, flat areas, and non-random jumps imply high autocorrelation. Lastly, you can detect correlation by observing upward or downward trends. For example, if we tried to analyze infant mortality rates in the United states, the variance would be inconsistent because of changing factors throughout time. For example, medical technology, nutrition, and access to healthcare are catalysts to the recent plateauing infant mortality rates. The more volatile mortality rates during the beginning of the 20th century could likely be due to the lack of medicinal knowledge, no access to healthcare, epidemics, and poorer environmental regulation, implying high variance. The data would likely demonstrate higher spikes during the beginning of the 20th century, which would eventually flatten and create a more linear appearance by the beginning of the 21st century. The downward trend could imply that there is high autocorrelation in the data.

A**stationary** time series retains certain statistical properties such as consistent mean, variance, autocorrelation. Of course, given many of the time series trends we have nowadays, it’s much harder to find data that is stationary. Thus, we use time series analysis, such as differencing, boxcox, or log transformations to make the data stationary.

Last but not least, we must understand the p,d, and q terms of an ARIMA(p,d,q) model. A simple way to understand the “p” term is to estimate the number of lags. You can determine the number of lags by reading the PACF and ACF graphs for the time series. To determine the p value, or AR(p) component, observe where there is a gradual decay in the ACF and an abrupt drop in the PACF graph. The “d” term is simple the amount of times you need to difference an original time series data in order to mitigate any downward or upward trends in the data. To determine the q value, or MA(q) component, observe where there is an abrupt drop in the ACF graph and a gradual decay towards zero for the PACF graph.

I will eventually update my answer with diagrams (if that is preferable). If you have questions concerning which R code to use (I only have experience using R for time series analysis), I am more than happy to answer them separately through question requests. There are codes that can efficiently perform boxcox and log transformations, differencing, acf, pacf, and automatic model prediction with forecasting. I can also share other interesting tests to verify consistent variance, no serial correlation (autocorrelation) within the residuals, normality, and stationarity conditions.