The Simple Moving Average

© Malcolm Moore
14-Dec-2003

Introduction

Simple High School maths provides one of the most powerful tools to minimise trading noise and see a more realistic picture of what is going on.

To calculate an average price we would take a set number days and using these Close prices, calculate the Average price based on the number of days by summing up all the Close prices in the number set and then divide by the count of numbers. This gives us the average price for that set of Close prices!

By sequentially moving on one day at a time and repeating the process we would have what is called a Simple Moving Average (SMA). These value points can be plotted on the same graph as the End Of Day (EOD) Close prices.

The standard thinking is to plot these SMA points on the last period (day) that the moving average for that set of periods was calculated, as this will be the first time that the moving average values will be "useful". This is the (flawed) convention…

When these calculated points are joined sequentially, they form a stepped curve that follows the close prices from which it was originally derived. Unfortunately the (flawed) convention here is to simply join the dots and make a smooth curve.

In engineering terms the Simple Moving Average (SMA) over say 4 periods is:

SMA(4) = (Close(t-0) + Close(t-1) + Close(t-2) + Close(t-3))/4

In this case the function SMA is based on 4 periods, and as a function each period is shown as a time period one day earlier.

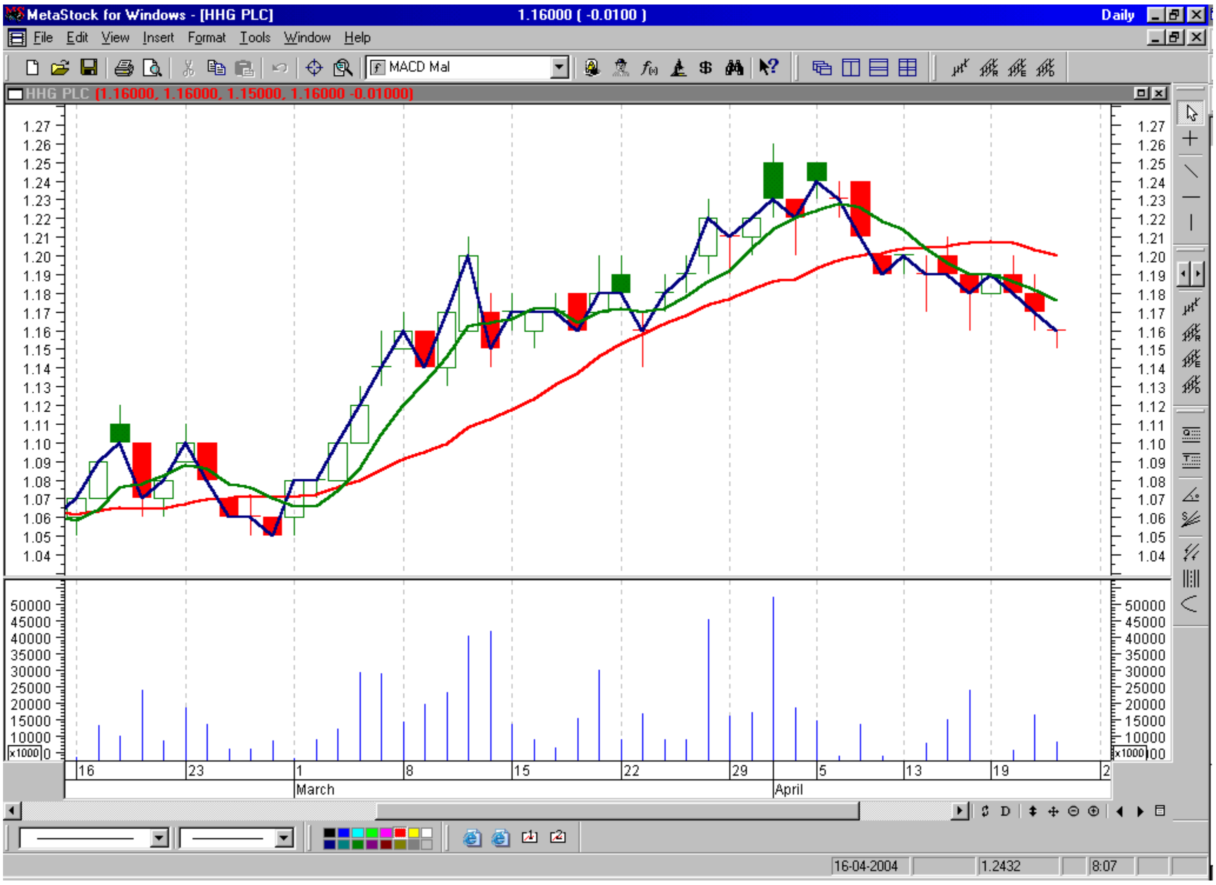
For those slightly more mathematically minded, the same equation can be written in a more compact and general form as follows (it means the same).

$$SMA(Periods) = \frac{\left(\sum_{0}^{Periods} Close(Period)\right)}{Periods}$$

This simply means that the Close prices for each Period is summed (by the capital Epselon sign), over a specified number of Periods, and divided by the total Periods equals the Simple moving Average over that number of Periods! Easy!

As an example, for a 12 day Simple Moving Average; the Close price of each End Of Day (EOD) Period would be added up (summed) with all the other Close prices over the Periods of interest (in this case 12), and then averaged to form a Simple Moving Average over 12 days to form the most recent SMA12 value.

The picture below gives a visual description of a share price in candlestick form and three Simple Moving Averages (SMA) based on 1 (blue), 5 (green) and 15 (red) days.



The dark blue jagged line is a one-day Simple Moving Average (SMA1) based on just the Close price of that day, and already it looks less noisy than the candlesticks in the background. The smoother (middle) green line is an SMA5 (based on the most recent five days), and the even smoother red line is an SMA15 (based on 15 most recent days).

Already the picture should be coming clearer! The SMA tool here smoothes out the jerky nature of the EOD results and provides us with curves that are much smoother.

The SMA is really a very primitive form of a Low (frequency) Pass Filter. Most texts on Indicator analysis stop here and move onto something else, without really understanding what they have unearthed in electronics engineering land!

The similarities in these two lateral applications are astounding, and the understandings of each can be applied liberally from one to another and with a little more lateral thinking, these can be applied in several other areas of life and research! Nerdy stuff eh?

Low (Frequency) Pass Filters

Filters are a fundamental part of mechanics and electronics and are used extensively in our everyday life, and examples include the padded soles on our shoes, tyres on a car, padding on a seat, a mattress. These are all mechanical low pass filters. In technical trading of securities the prices continually flutter and a moving average 'smooths' these out – removing the price fluctuations and giving the price a 'soft' visual feeling.

A Low Pass Filter passes zero frequency (a constant), and higher frequencies in the "pass band" till the frequency exceeds the "cut-off" frequency point. Above that frequency the amplitude of these higher frequencies is decreased (or attenuated) in a “transition band” until the frequency reaches the “stop band” area where the amplitude of these frequencies are sufficiently attenuated to have a negligible effect.

Electronic and mechanical filters can be engineered to have very steep “skirts” (no real explanation needed here) so that the cut-off frequency can be quite close to the stop band edge where here and above, the attenuation is high in the stop band area.

The “frequency response shape” and “transient response” of a filter can be engineered by changing the components in the filter, and this is a mature branch of electrical and mechanical engineering.

In engineering terms they are specified in terms like a defined maximum pass band ripple, and minimum stop band attenuation, and the ratio from pass band to stop band, and a time responsive shape, plus many other engineering based qualities.

The problem is to understand what is required and engineer the design of a filter to behave in a manner that suits the requirements! This is one of the areas where a Professional Electronics Engineer comes in to use as they usually have the expertise to understand and convert these requirements into a specification and then into a working design. These filters have a direct application in the technical analysis of stock data!

Digital Filters

In “technical trading” land, the application of a simple moving average is a direct application of a very primitive (simple) digital filter, so at this point, a little background knowledge with digital filters would never go astray!

A Digital Filter operates by taking sequential samples, passing these sampled values through a “shift register” (which moves the values on one stage at a time with the synchronising clock), then combines the result(s) from the various shift register outputs (often called “taps”) to provide a sequential summed sample output, which is usually then transformed back to an analogue (real life) output.

The advantages of a digital filter are that it can be programmed into a computer, or microprocessor, or digital signal processor (DSP)), the programming can usually be very easily changed and with all digital processes using memory. Digital filters are far more versatile than analogue filters.

By virtue of EOD trading figures, we already have sampled inputs from all traded stocks from that day, and in most cases every trading day for several years – so we will never be short of EOD trading data. This is an ideal starting point for using digital filters!

Think about it a little laterally, the EOD data on the earlier trading days is the data in the input “shift register”, and that is why learning a little about digital filters may prove to be very helpful! These EOD prices when added up and averaged form a Simple Moving Average (SMA) that is in reality a Non-Recursive algorithm of a “First Order” low (frequency) pass digital filter!

Non-Recursive Filters

These are the most primitive type of filter and they take a sample of each of the most recent (say) 21 days Close prices, then ‘weight’ these values with individual constants, then average these values and that is the new output for the most recent day.

It is called non-recursive because the ***output is <u>not</u> fed back*** into the input. It also has another name “Finite Impulse Response” (FIR) Filter, because the response does not have an infinitely long exponentially decaying tail as there is no feedback to cause this tail!

The equation for a non-recursive filter with seven stages is:

$$y(t) = (k_0x(t_0) + k_{t-1}x(t_{t-1}) + k_{t-2}x(t_{t-2}) + k_{t-3}x(t_{t-3}) + k_{t-4}x(t_{t-4}) + k_{t-5}x(t_{t-5}) + k_{t-6}x(t_{t-6})/7$$

The subscripts relate to the age of the sample, so you can see it stepping through the shift register, as the negative subscript number gets larger.

If this were EOD data then the time clock would be nominally 24 hours (a day), so each input would be one trading day earlier (older) than the most recent with the time subscript *(t_0).*

In the more general case, the equation can be written as below:

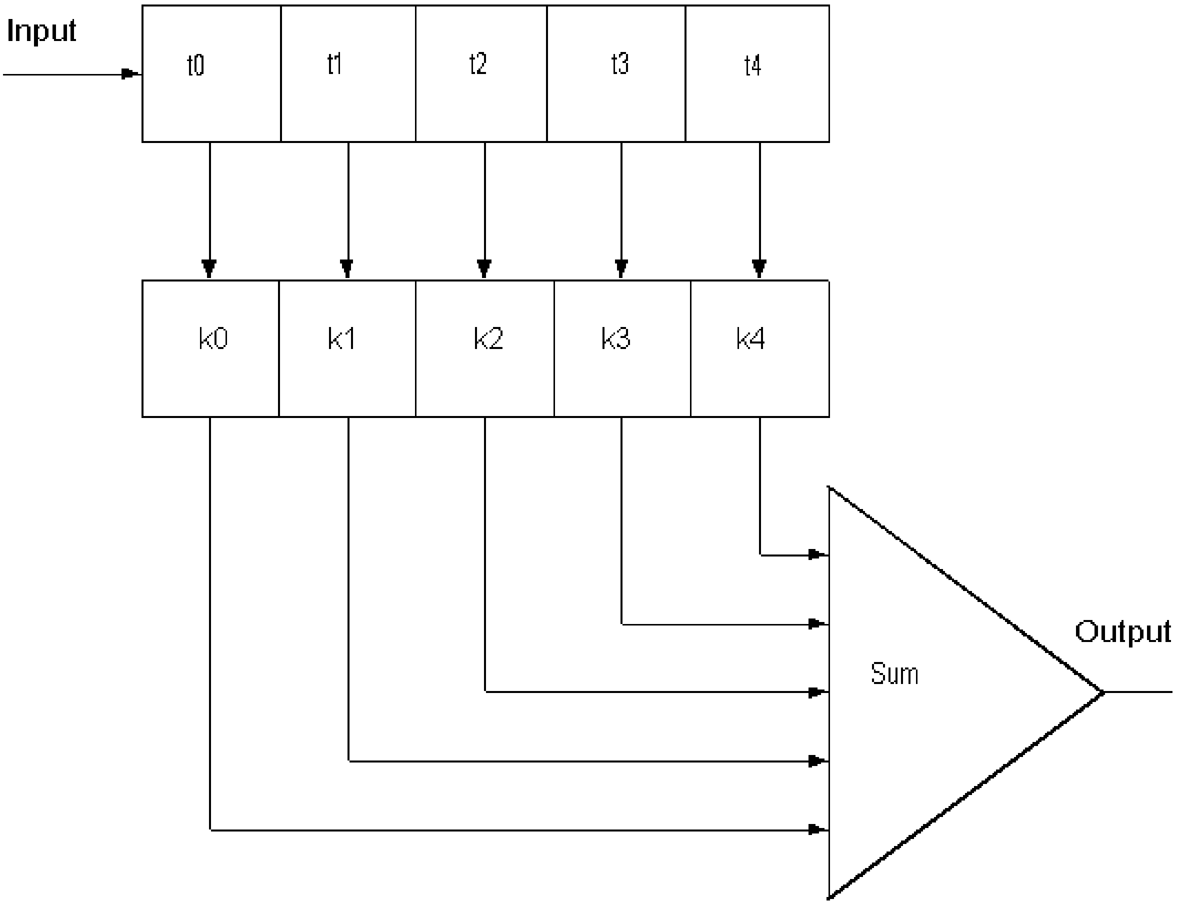
$$y(t) = \frac{\left(\sum_{0}^{n} k(t)x(t)\right)}{n}$$

In this equation the output (y) as a function of time (t) is the total sum of the product of the weighting constant (k) as a function of time (t) and the input (x) as a function of time (t), all divided by the total number of inputs counted.

Compare this maths to the maths used for a Simple Moving Average above and it is virtually identical. If the constants *k(t)* are set to unity, then it is identical!

The block schematic below visually shows the signal coming in at a fixed discrete level, and then stepping through a shift register from left to right (t0, t1, t2, ...).

At each time step, the taps off the shift register feed the delayed values through a multiplying constant array (k0, k1, k2, ... etc), and then the outputs of those are all summed and to provide the output. In this case this is a 5 stage FIR, and it does not take too much imagination to add on stages after t4, k4 and make it a 25 stage, 200 stage or whatever!



The SMA is a classic case of this non-recursive filter (FIR Filter) where the weighting *k(t)* is 1.0 for all values.

In this case, as all the constants are 1.0 there is no correction factor required. This means that if the input price was say $1.00 constantly, then after five periods, the output would be $1.00 also.

In cases where the weighting constants for all the values is not unity, and the average of these weighting values does not come out at unity, then a correction factor is required to align the output of the filter with the input.

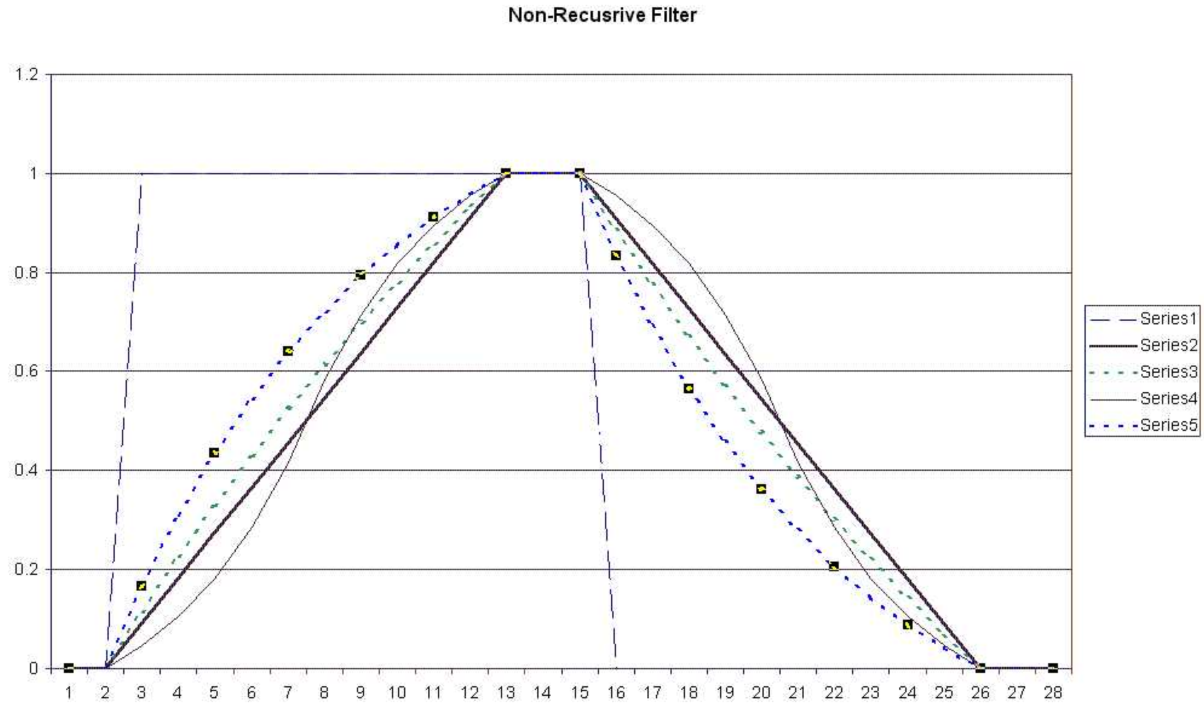
The “Weighted filter” is almost the same as the SMA but the weighting *k(t)* is usually greater for the more recent values and lesser for the older values. An example based on 11 stages could be constants of 1.300, 1.235, 1.173, 1.115, 1.059, 1.006, 0.9556, 0.9078, 0.8624, 0.8193, 0.7783 and in this case a correction factor of 1.0192 is required to approximate a unity gain filter.

If this was carried a bit further then another “Weighted filter” example based on 11 stages could be constants of 2.000, 1.7411, 1.5157, 1.3195, 1.1487, 1.000, 0.8705, 0.7578, 0.6597, 0.5743, 0.5000 and in this case a correction factor of 1.0988 is required to approximate a unity gain filter.

The “T” Weighted Filter is almost the same as the SMA but the weighting *k(t)* is greater for the more middle register constants.

For example a “T” Weighted filter based on say 11 stages could have weighting something like 0.600, 0.780, 1.014, 1.382, 1.714, 2.228, 1.714, 1.382, 1.014, 0.78, 0.600 and you would see that days 5, 6 and 7 have the heaviest weighting! In this case a correction factor of 1.2007 would create approximate a unity gain/loss filter.

To get a theoretical look at what is going on it is not hard to set these up in Excel and graph the results, and they come out as follows:



Series 1 (the dashed line) is to all intents and purposes Heavisides' Unit Step function, where the value jumps from 0 to 1 and stays there and in this instance is stepped back to 0 after everything has stabilised.

Series 2 (the thick line) is the SMA output – note that it is a ramp – so this kills the argument that in using an SMA the old values have a high impact on the final output – that folly simply is not true!

The older SMA values do have an impact but the effect of the impact is divided by the number of samples. This effective impact can be reduced by weighting the samples so that the older samples have a lesser effect (or weight), and this is shown in the next series.

Series 3 (the short dotted line) is the lesser of the Weighted outputs. It is the dotted line near the SMA output, and the weighting is somewhat exponential – hence the slight exponential shaped rise and fall shapes with time. Note that it like all the outputs in this case stops at the 11th step.

Series 4 (the longer dotted line) is the greater of the Weighted outputs It is the more heavily dotted line and its shape starts to approximate that of a 1st order exponential rise and decay. Note that after the eleventh step there is no continuing effect – so it is a truncated approximation.

Series 5 (the thin line) is a "T" Weighted response and in this case it has an "S" shape in response to a Unit Step input.

In this case the output kicks off more slowly, has a fast centre rise and comes into land at the new level quite nicely.

This shape is more akin to an overdamped step response seen from a 2nd order filter (but it is not 2nd order, it is a 1st order filter)!

What is noticed is that as the number of days (samples) is increased to make the moving average, the smoother SMA lines are but they drift to the right, lagging the change in prices over more days. This is upsetting for some as it means that the well-smoothed SMA of a changing price is in reality – highly inaccurate.

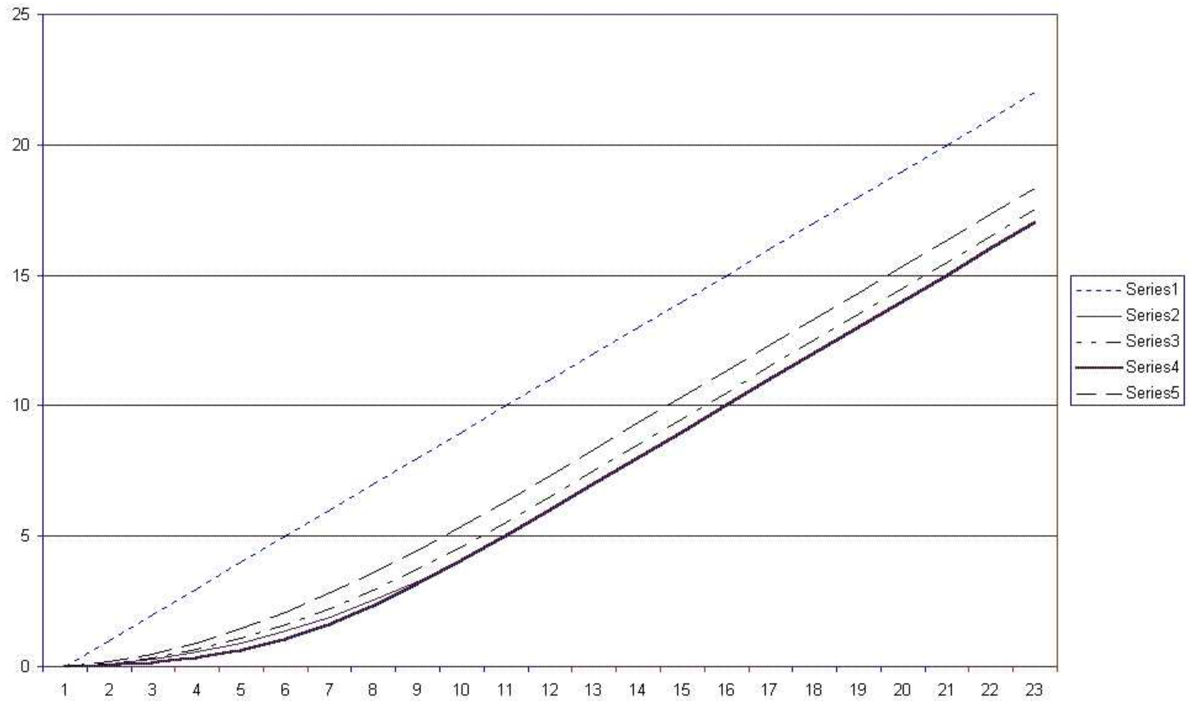
Slew Rate is another electronic engineering term relating to the speed at which a voltage (or current) can change value with time.

Slew Rate effectively limits the upper frequency response combined with available voltage swing of amplifiers, the limit clock speed of computers and the speed at which memory devices can be clocked.

In relation to technical trading on the stock market, the Simple Moving Average (SMA) in its most simple form where all the coefficients are ‘1’ this FIR low pass filter has a ramp output from a step input that emulates a slew rate limited operational amplifier as shown in the thick line above as ‘Series 2’.

So now we have some definitive answers that we did not have before. The FIR filter has a zero output error compared to an infinitely long step input, but it has a (time or ‘x’ related) calculable error in getting to a new stable level and the output response can be shaped by changing the coefficients (the weighting).

Now let’s bring this to a new level. It is common to have a ramp in share prices, so by putting in the clean Heaviside Unit Ramp as an input to this first order FIR filter we can gauge the response:



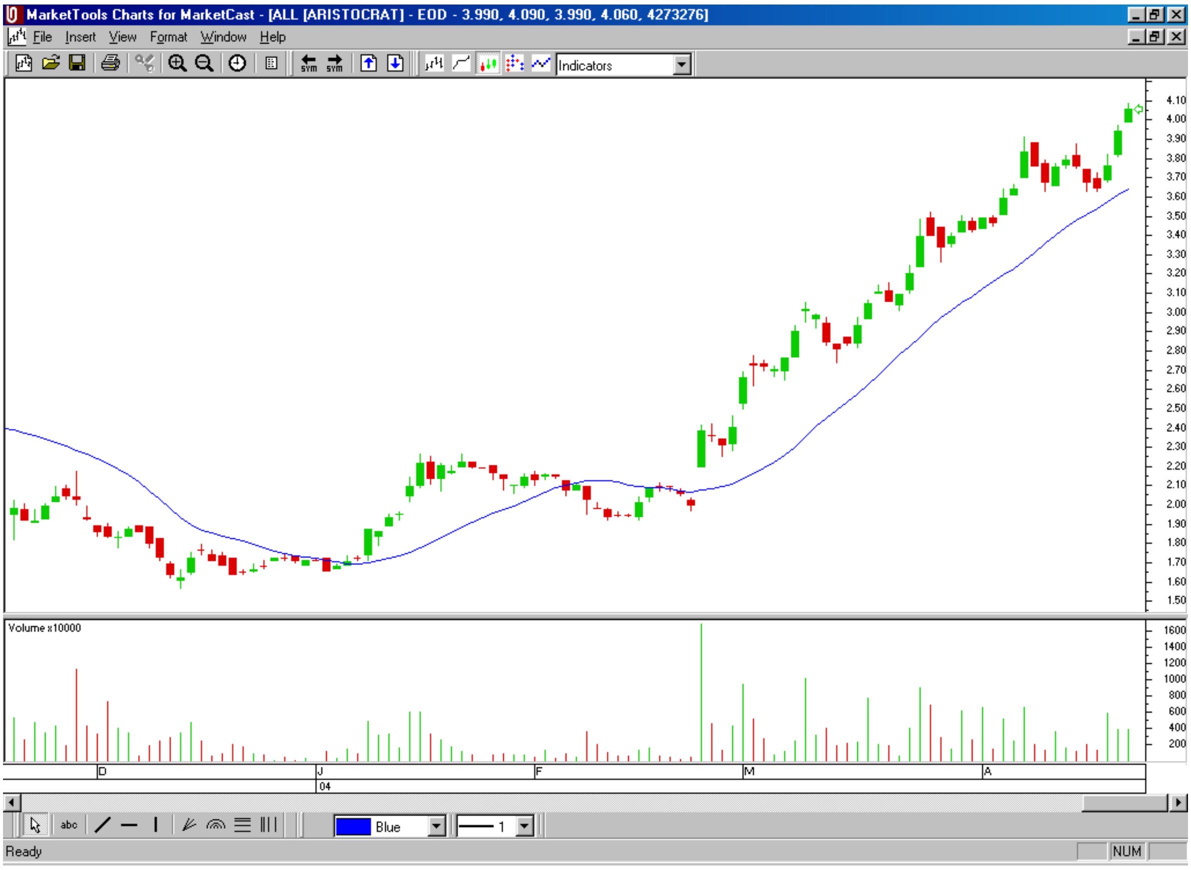
In this case the ramp is the small dotted line (Series 1) and the others are all outputs as per the data above. Series 2 is the SMA with even weighting, Series 3 is with a slight exponential weighting, Series 4 is with more exponential weighting, and Series 5 is with 'T' weighting. What this really shows is that the weighting affects two things: the time response and the frequency response – even though the number of stages is constant.

Compared to Series 1 (The Unit Ramp Input), in this case Series 2 (Simple Moving Average) and Series 4 (greater Weighted Moving Average) have virtually the same ramp error constant, but the Weighted Moving Average moves quicker towards a constant error (being slightly less reactive in the time domain in this case).

Series 3 (lesser Weighted Moving Average) has a slightly smaller error constant, making it more reactive to changes in input and Series 5 (T Weighting) has the smallest ramp error, making it most reactive to changes in the ramp input.

So now we know that the size of the error is dependent on the total number, values and ranges of the coefficients of the timed samples. We are not going any deeper here as we have the necessary basics to understand how a very simple FIR filter operates and what we can expect from this knowing the input that is driving it!

We can use this characteristic of first order filters to our advantage. In a graph of share prices, with all the trading dates shown, if the trading is hovering about a certain price, then the SMA will cut through the middle (average) of those prices. As the price ramps up or down at a constant rate per day, the SMA will follow the price but with a certain constant error. The picture below gives a good example of this:



MarketTools Charts for MarketCast - [ALL [ARISTOCRAT] - EOD - 3.990, 4.090, 3.990, 4.060, 4273276]

We now know that merely by looking at the Candlestick EOD above graph using the MarketTools Chart graphing package, of *Aristocrat Leisure* and a 20-day Simple Moving Average (SMA20) indicator that this Simple Moving Average indicator is a “First Order System”.

This knowledge is extremely important, as we know that a first order system has zero error compared to a constant, like mid January through to late February and it has a constant error compared to a ramp, like late February onwards on this chart.

The size of the error is related to the steepness of the ramp and the time-constant of the Moving Average. The steeper the ramp, the more the error. The longer the time constant, the larger the error (between the actual price value and the value of the moving average of that day).

It just so happens that the stock market is an object that is based in the time domain, and that digital filters are also clocked in the time domain, and the time response from filtering is of paramount importance for performing meaningful analysis. The ability to link between the time and frequency domains through using mathematical transforms opens up a huge variety of options that would otherwise not be possible.

Non-Recursive digital filters utilise a defined number of stages and because of that the response to a known excitation is finite.

Mathematically the Unit Impulse is the common foundation for times responses, hence these filters are called Finite Impulse Response (FIR) filters.

We have a dilemma here in that we wish to use a filter that has a minimum of delay (stages) and to respond (in many cases) to a huge number of delayed inputs. There are other digital filter structures that we can use that can take a lot less stages to effect our desired transient response. In our case we want the effect in a minimum of stages, because that implies a minimum of delay – but not necessarily so!

Non-Recursive (FIR) digital filters have a real use in recovering a digital stream and cleaning it up (equalising it) before it goes into further real digital processing. One classic example is when using optical fibre, in some cases the Group Delay (relative received frequency band time differential) is such that the transitions are lost.

The FIR filter can be set up as an ‘all pass’ filter, in that it passes the received signal and the tap weightings sum to zero. (It is a ‘zero forcing’ output.) When a transition comes in, it passed through the filter to the point that the taps re-constitute the signal as a sharp (but overall slightly delayed) transition.

The ‘kick-in’ to buy shares often follows a very similar pattern where for the first dew days there is a false start, where the share price may rise say 5% then a retracement, followed by a very nice rise well exceeding 5% and often going on to 100%. A little bit of lateral engineering could provide a good yardstick, using a relatively simple FIR filter. The one problem is that the EOD figures are spaced out too far to be much use in this case, and this type of filter could come in with a positive result about a week too late!

Now that the appetite is whetted, and we have started to see some of the limitations of FIR filters, it is time that we spread our wings. Digital filters are everywhere in several forms.

In this case we have discovered that a Simple Moving Average (SMA) is a direct implementation of a FIR digital filter, with a weighting of one (1) for all taps.

We have played a little with some tap settings and found that the time response can be significantly changed, and that these filters are effectively 1st order filters, (having a constant error to a ramp input, and a zero error to a constant level input).

So this then begs the question, “What is an Exponential Moving Average (EMA) equivalent to in terms of what sort of a digital filter?”

Comments and Corrections are welcome