Prognostication

Michael Sachs

Medieval alchemists are infamous for devising bizarre schemes for turning lead into gold. The lure of endless riches drew charlatans and swindlers, but also serious people who felt there was some hidden truth about nature to be found. Today, market forecasting has taken the place of transmutation. Now, as then, there are many forecasting techniques that are snake oil. However, the quality and quantity of data available do afford a unique opportunity. Markets are, in some sense, a distillation of human behavior. Unlike many other behavioral studies, they come with a detailed, easily accessible dataset. This elevates the pursuit of market forecasting from a get rich quick scheme to the study of human nature.

Of course there is nothing wrong with getting rich quick. As Oscar Wilde put it: "When I was young I thought that money was the most important thing in life; now that I am old I know that it is."

In what follows, I have applied some techniques from signal processing and information theory to market data. Although I have yet to devise any forecasting scheme, there are tantalizing hints in my analysis of some underlying structure.

The Data

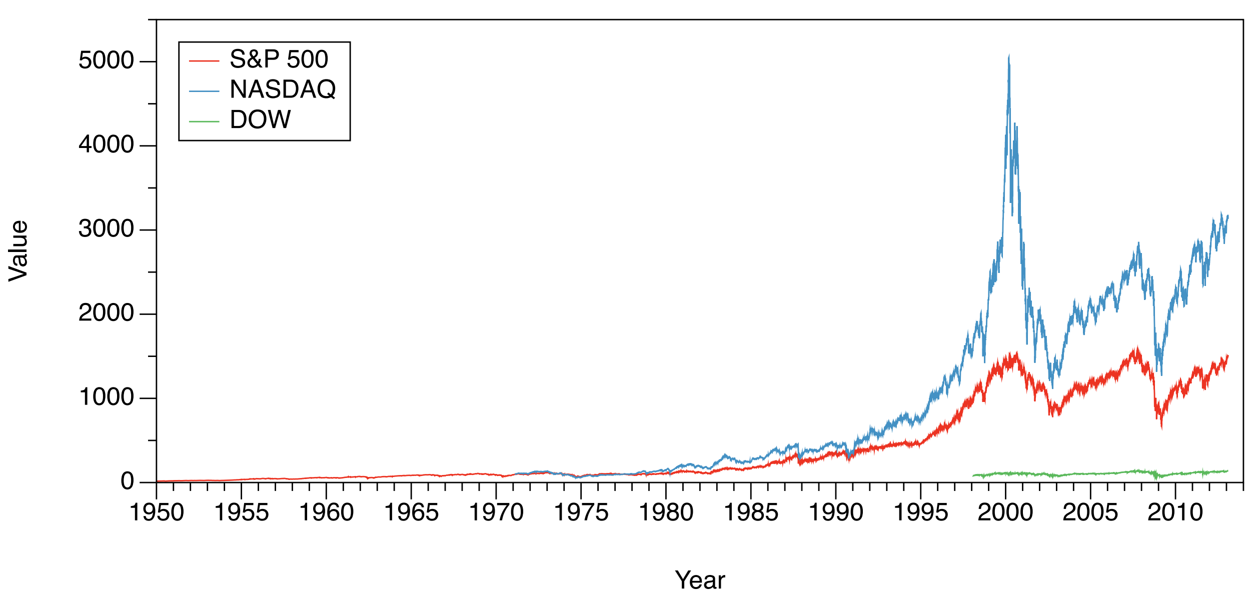

Figure 1: The values of the three major market indicies. **Blue** NASDAQ, **Red** S&P 500, and **Green** Dow ETF.

The market data that is most familiar is the price vs. time of a stock or market index (Figure 1). This data is difficult to analyze for two reasons. First, it is not stationary. Until the late 1980s the values of the S&P 500 and the NASDAQ grew at a slow pace. Then in the 1990s they began to increase exponentially. Then in 2001 they decrease and began a period of oscillation. Second, it is not clear how the values of each index compare with one another. Why is the value of the Dow so much less than the other indices when all three are supposedly generated by similar underlying dynamics?

What is needed is a stationary value that normalizes the difference in value between these quantities. The trailing returns, $R(t)$ (Equation 1), do exactly that. The results of applying Equation 1 to all three indices are shown in Figure 2.

$$R(t) = \frac{P(t) - P(t - \Delta t)}{P(t - \Delta t)}. \tag{1}$$

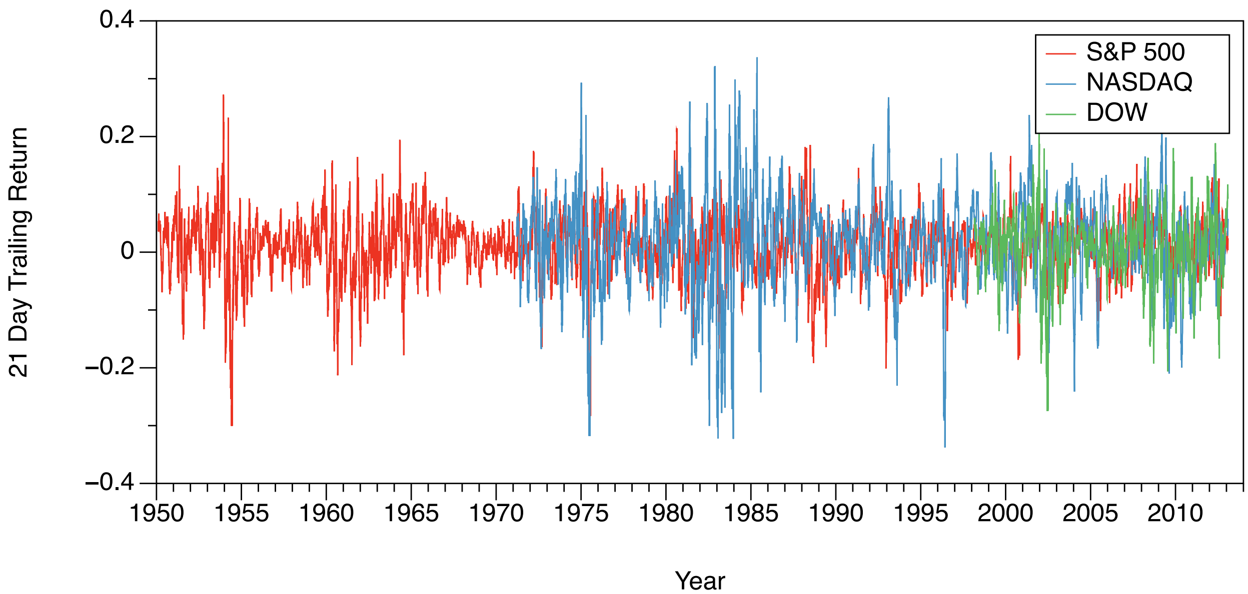

Figure 2: The 21 day trailing returns of the three major market indices. **Blue** NASDAQ, **Red** S&P 500, and **Green** Dow ETF.

Trailing returns are stationary around zero and all indices vary between -1 and 1. These properties make $R(t)$ much easier to analyze than the value of the asset.

Fourier Transforms

The first tool that is used in the analysis of any stationary time series is the Fourier transform. This technique is based on the fact that any discrete time series X can be rewritten as a sum of sines and cosines

(Equation 2).

$$X_n = \sum_{n=0}^{N-1} x_n \exp^{-i2\pi k \frac{n}{N}} . \qquad (2)$$

The x_n are a measure of the amount, or power, of each frequency present in the signal. Equation 2 can be inverted to extract the x_n in terms of the X_n; this is a Fourier transform. Figure 3 shows the results when this transformation is applied to the 21 day trailing returns of the S&P 500.

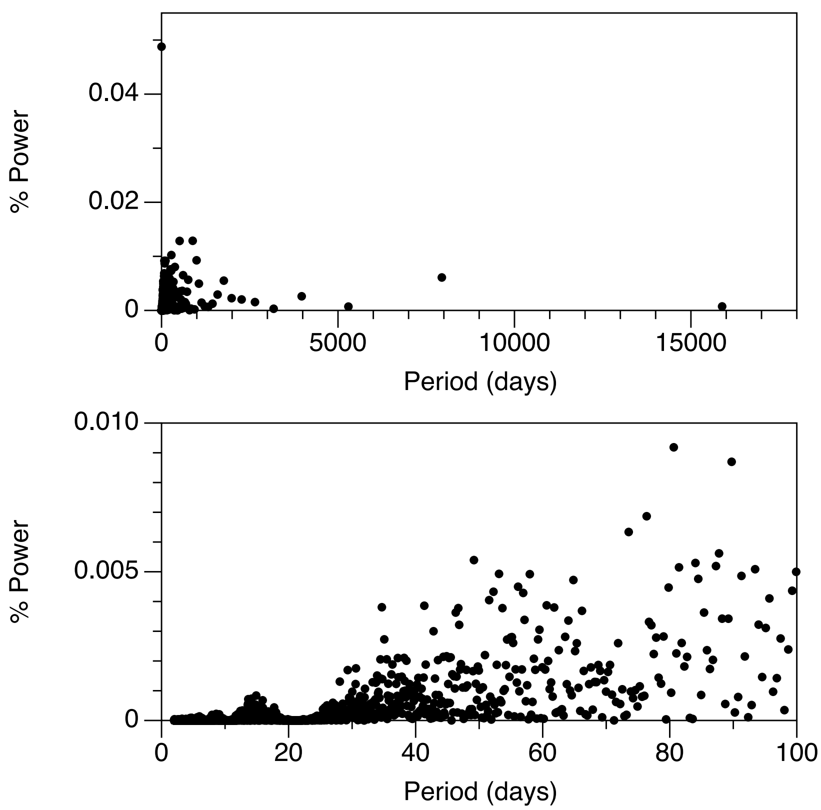

Figure 3: The normalized power spectrum of the 21 day trailing returns of the S&P 500. **Top** The entire spectrum. **Bottom** The spectrum for short periods. The power is expressed as a percentage of the total power in the spectrum.

The top of Figure 3 shows what one would expect: there is very little periodic structure in the S&P 500 at long times. This is pretty clear from Figure 1. However, at short time-frames there does seem to be some periodicity, as can be seen in the bottom of Figure 3. The first question that arises is if this structure is an artifact of the Fourier transform itself. To answer this, compare the results in the bottom of Figure 3 to Figure 4 which is the Fourier transform applied to a time-series of the same size as the one used in Figure 3 but drawn from a gaussian random variable with the same mean and standard deviation as the Figure 3 time-series. The gaussian time-series is truly random and devoid of all periodic structure.

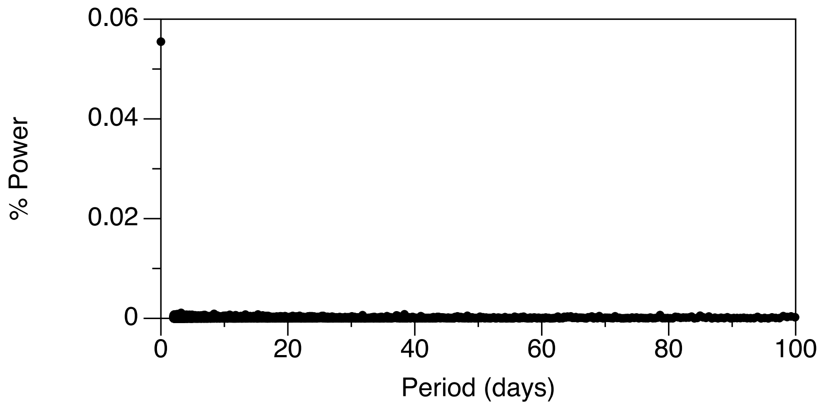

Figure 4: The short period power spectrum of a Gaussian random variable with the same length, mean, and standard deviation as the 21 trailing returns of the S&P 500.

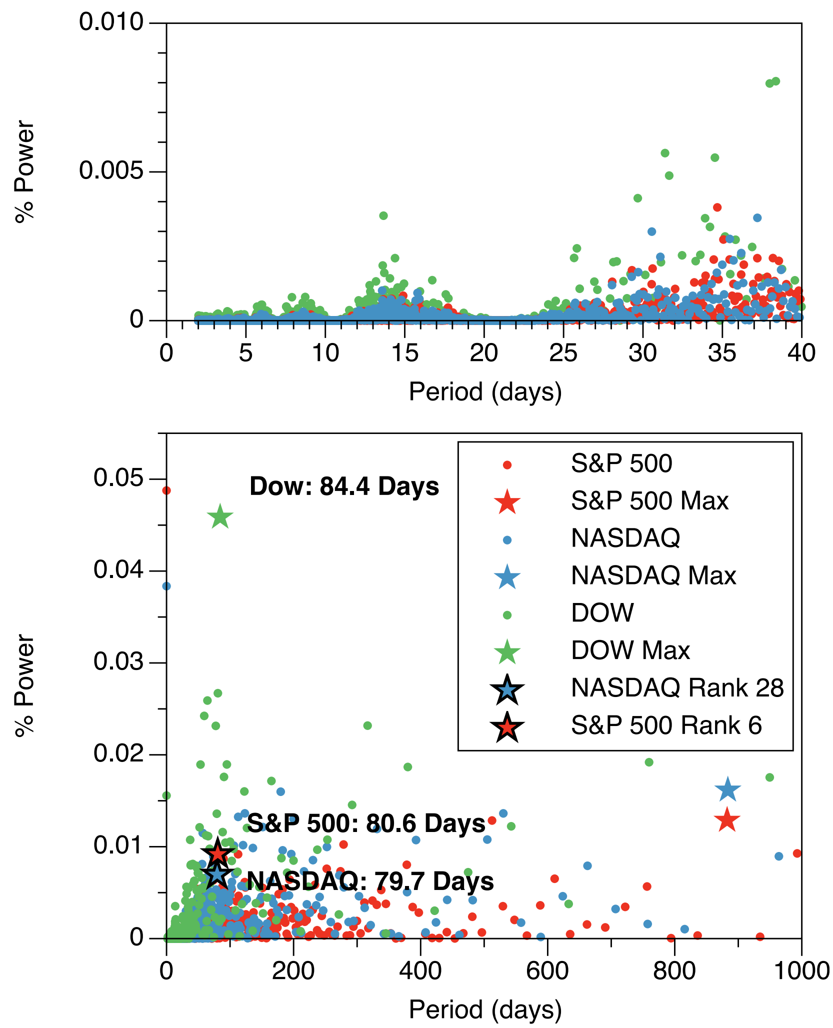

Figure 5: **Top** The short period power spectrums of the 21 day trailing returns of the S&P 500 (red), the NASDAQ (blue) and the Dow ETF (green). **Bottom** Maxima in the power spectrums.

How does the S&P 500 compare with the NASDAQ and Dow ETF? The Fourier transforms of the 21 day trailing returns for all three are shown in the top of Figure 5. It seems like all three indices share a similar structure, at least at short periods. None of the three show any periodicity at 21 days, the length of time for the trailing returns. This phenomenon is not unique to the 21 day returns. Figure 6 shows the period around the trailing return time for the Dow ETF at several trailing return times. In every case, there is no periodic structure at periods equal to the trailing return time.

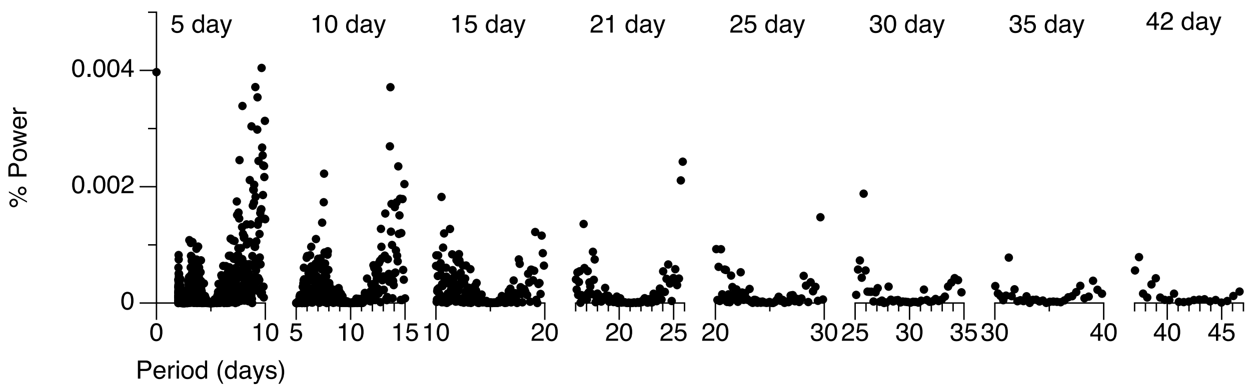

Figure 6: The period around the value of trailing days in the Dow ETF. Shown for several different numbers of trailing days. In each case there is no power for the period that corresponds to the trailing days of the returns.

Another interesting feature of the Fourier transforms is shown in the bottom of Figure 5. The largest power (excluding the zero-mode power) for the Dow ETF 21 day trailing returns is at $T \simeq 84.4$ days. The S&P 500 and NASDAQ also have fairly strong peaks (6th ranked and 28th ranked respectively) at around $T = 80$. It seems that there is a strong periodicity in the 21 day returns at a period of $T \simeq 80$ days, especially in the Dow ETF. Is this peak effected by changing the length of the trailing days measured like the minimum is in Figure 6? Figure 7 shows the range of periods around $T = 80$ for several different trailing times of the Dow ETF. The peak seems to be fixed at $T \simeq 84$ days no matter how many days of trailing time is used (the open stars in the first and last plot in Figure 7 are the second highest peak instead of the first). This seems to indicate that there is a strong periodicity in the Dow ETF trailing returns at $T \simeq 84$ days regardless of the number of trailing days used to calculate the returns. At this point I have no theories as to why this seems to be the case. It may be an effect of the Fourier transform. If it is an actual signal it would mean that if you wait 80 days or so your trailing returns will (with some strong probability) return to where they started.

Clearly the features of these Fourier transforms are interesting and bear more scrutiny. Even if these features can be verified it is unclear what role they could play in a forecasting algorithm.

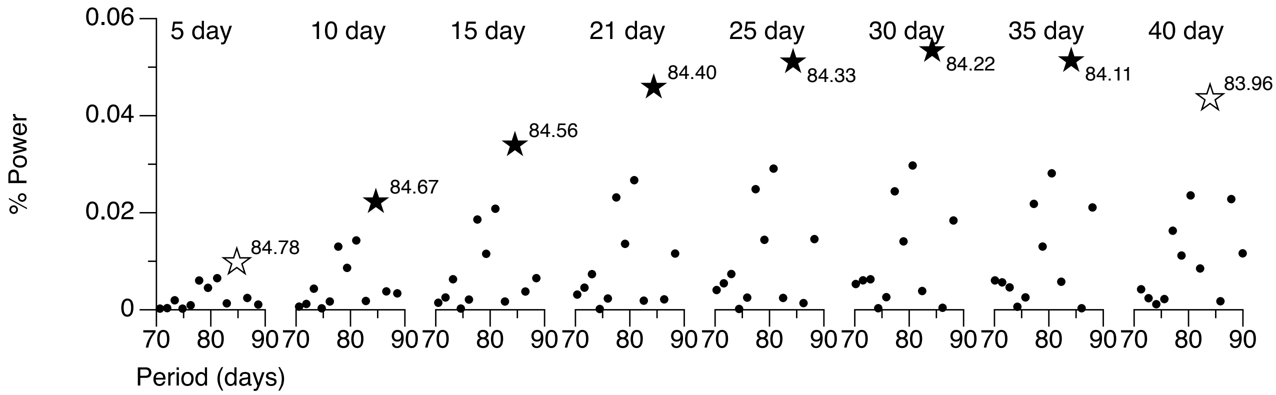

Figure 7: The period around 80 days for several different values of the number of trailing days of the returns of the Dow ETF. For 10 trailing days to 35 trailing days the peak of the power spectrum is around 84 days. For 5 trailing days and 42 trailing days the peak at 84 days is the second highest in the spectrum.

Shannon Entropy

The idea of entropy in information was first introduced by Shannon [1948] as a measure of the amount of information in a signal. It is usually defined as:

$$H(X) = -\sum_i P(X_i) \log P(X_i). \tag{3}$$

In Equation 3, X is a discrete random variable, and $P(X_i)$ is the probability of the variable being X_i. The logarithm is usually taken as base two, which results in H being in bits. Highly random signals will have a high entropy, while signals that are predictable will have a low entropy. The definition of entropy in Equation 3 can be difficult to apply to signals that are not binary. In order to avoid this difficulty a new definition called permutation entropy [Brandt and Pompe, 2002] can be employed. The permutation entropy is defined as:

$$H(n) = -\sum_i p(\pi) \log p(\pi). \tag{4}$$

$p(\pi)$ is the probability of finding permutations π of order n in the signal. The logarithm is base two. For example: assume that our signal is $x = (4, 7, 9, 10, 6, 11, 3)$. Looking at each pair of numbers – this is order $n = 2$ – there are four pairs where the first number is smaller than the second. There are two pairs where the first number is bigger than the second. These are the only possibilities at $n = 2$. There are six total pairs. The permutation entropy is:

$$H(2) = -\frac{4}{6} \log \frac{4}{6} - \frac{2}{6} \log \frac{2}{6} \simeq 0.918. \tag{5}$$

The permutation entropy in Equation 4 is bounded: $0 \leq H(n) \leq \log n!$. It is convenient to define a new value $h_n = H(n)/\log n!$ which

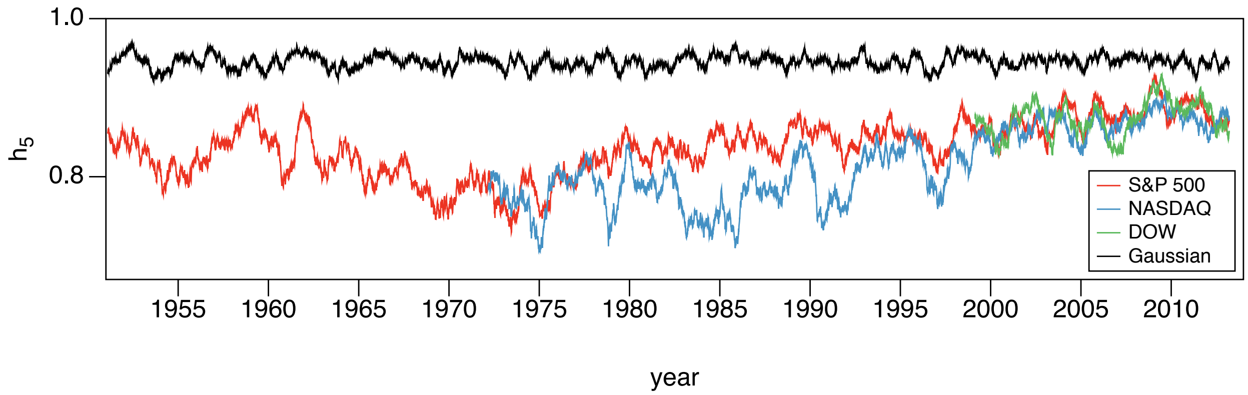

Figure 8: The permutation entropy $h_n = H(n)/\log n!$ for $n = 5$ over a trailing window of $T_{win} = 252$ days for all three market indices and a gaussian random variable with the same length, mean, and standard deviation as the S&P 500. **Blue** NASDAQ, **Red** S&P 500, **Green** Dow ETF, and **Black** Gaussian.

is zero for a completely predictable signal and one for a completely random signal.

The results of applying this to the 21 day returns of our three market indices are shown in Figure 8. Here h_n was calculated for $n = 5$ on a moving window of $T_{win} = 252$ days. The first interesting feature in Figure 8 is the fact that h_5 is large but still < 1. This implies that the indices examined here are not completely random; there is some underlying structure. For comparison, the results from a gaussian random variable with the same length, mean, and standard deviation as the S&P 500 is also plotted. The second interesting feature in Figure 8 is that, over long time periods, the amount of structure in the indices varies. It seems as if markets – at least to the degree they are reflected by the indices – have been getting more random. This could be because the number of trades executed has increased over time. More trades implies that whatever correlations exist in the data are erased faster. More research is needed to determine if this is indeed the case.

The value of h_n plotted in Figure 8 was calculated for a specific order n and window T_{win}. How does changing these values effect the results? In Figure 9 a 10 year section of h_n for various values of n and T_{win} is shown. For n chosen too low, the signal seems very random because not enough of the structure is sampled. For n chosen too high, there are not enough permutations of length n present in the window to completely sample them. This is why the window size strongly effects the $n = 7$ results. At $n = 5$ (bold in Figure 9) the behavior of h is similar to its neighbors at $n = 4$ and $n = 5$. Also at $n = 5$, doubling the window has a weaker effect than at higher values of n. Because of these factors I have chosen to use $n = 5$ and $T_{win} = 252$.

It does seem like h_n is telling us something about the structure in our indices. There are periods of relative randomness, followed by

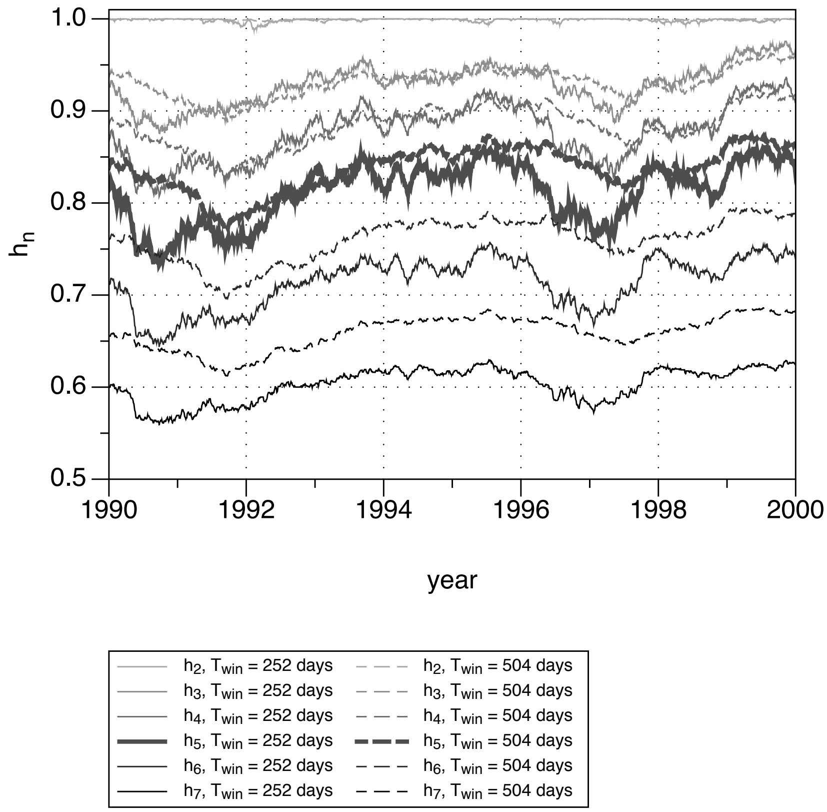

Figure 9: The effect on h_n of varying n and T_{win}.

periods of relative order. Is there a relationship between the returns and the the value of h_n? My first pass at answering this question involved calculating the correlations between the indices and h_n. The results are shown in Figure 10 which shows the returns (top), h_n (bottom), and the correlation of these two quantities over a trailing 21 day window (middle). The correlations in the middle plot of Figure 10 are always strong, but they vary between being correlated and anti correlated. I am still looking for a pattern here. There is some preliminary indication that strong correlation peaks or valleys ($> |0.5|$) are somehow related to peaks and valleys in the returns. However, whether there is a peak-peak/valley-valley or peak-valley/valley-peak correlation seems to shift. There is clearly more to be done here. Perhaps correlations between the correlation and the returns are the next thing to examine. Also, varying the window of the correlations – perhaps taking the window sizes from peaks in the Fourier data – could prove informative.

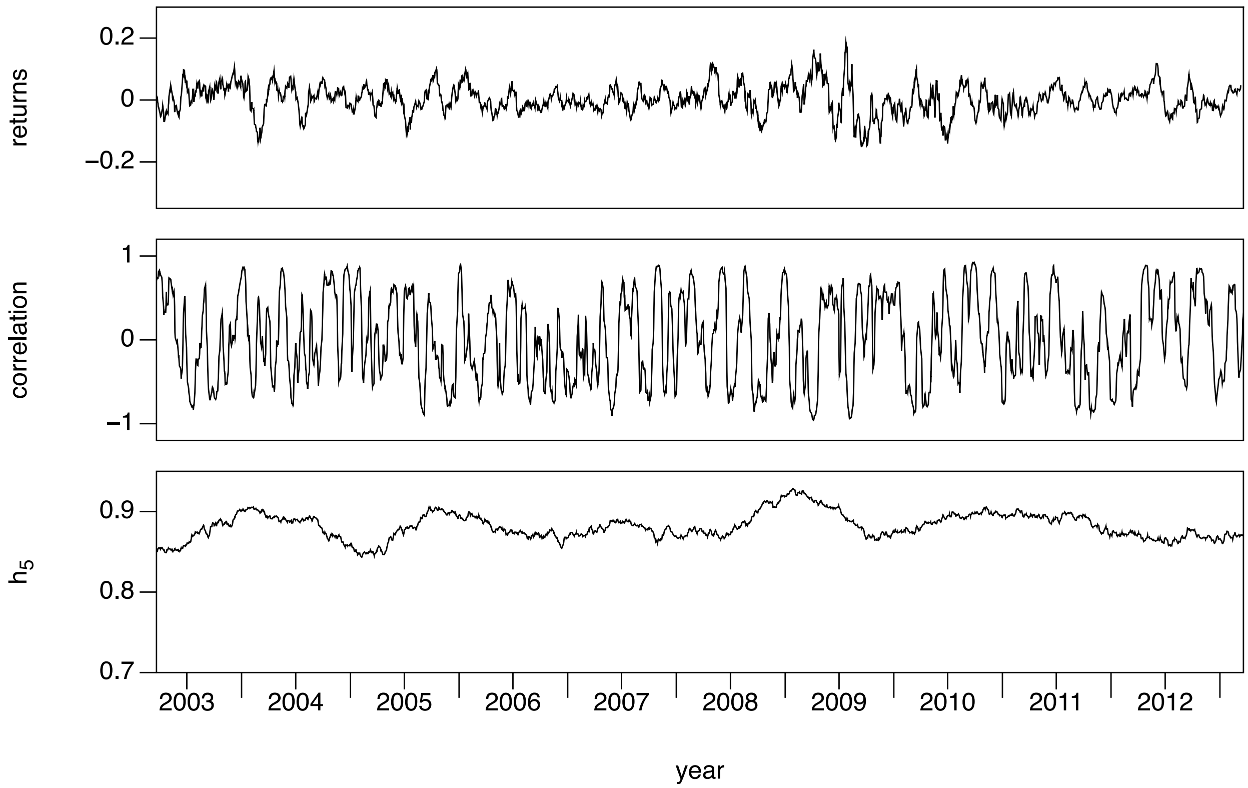

Figure 10: 21 day trailing returns of the S&P 500 compared to h_5 and the correlation between these two over a 21 day window. **Top** The 21 day trailing returns. **Middle** The correlation over a 21 day window. **Bottom** h_5 for the returns.

Epilog

As of this writing – perhaps unsurprisingly – a forecasting algorithm for the markets is still unrealized. There are tantalizing hints of structure in the Fourier transforms and the permutation entropy of market returns. Perhaps these hints can be exploited to find probabilities for use in the Kelly criterion. Another idea is to use the structure in the measures I have used here to constrain models. Perhaps then the models could be used to generate probabilities.

The process of examining these quantities has been very interesting and rewarding. The reward however, has so far been entirely intellectual. I hope that further investigations will prove financially rewarding as well!

References

C. Brandt and B. Pompe. Permutation entropy: A natural complexity measure for time series. *Phys. Rev. Let.*, 88(17):174102, 2002.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, July/October 1948.