DFA ANALYSIS OF FINANCIAL DATA

by

Bronson Argyle

Submitted to Brigham Young University in partial fulfillment
of graduation requirements for University Honors

Department of Physics and Astronomy
Brigham Young University
August 2008

Advisor: Dr. Gus Hart

Signature:_____

Honors Representative: Dr. Bruce Collings

Signature:_____

ABSTRACT

DFA ANALYSIS OF FINANCIAL DATA

Bronson Argyle
Physics Department
Bachelor of Science

Recent developments in the field of biophysics, both in findings and methods, have consequences that extend not only into physics in general, but may have application in a rigorous mathematical analysis of financial markets. Specifically, we apply the interpretative power of the Detrended Fluctuation Analysis to an Exchange Traded index Fund (ETF) mirroring the S&P 500. Not only do we verify the observation of positive long-range correlations, but we also characterize the effects of bin size on the DFA output. As a final application, we briefly examine the possibilities of using the results of a localized DFA to assess a measure of corporate health.

ACKNOWLEDGEMENTS

I would like to express my gratitude to both the university for the opportunity to perform this research as well as the many individuals who have provided support and technical assistance. Specifically, I would like to thank Dr. Gus Hart, who has profoundly affected my life in his capacity as teacher, mentor, and friend.

TABLE OF CONTENTS

LIST OF FIGURES

I Introduction

First developed by GoldBerger and others, Detrended Fluctuation Analysis serves to quantify the fractality of an underlying data signal [1]. Though the literature is replete with documentation [2] [3], it may be beneficial to offer a brief introduction to the theoretical and mathematical foundations of this approach. Given a raw data set $x(t)$, we first remove large trends by subtracting a running, integrated average M, thus

$$y(t) = \sum_{i=1}^{k} [x(t) - M]. \tag{1}$$

This integration effectively smooths the signal and is shown in the top plot of Figure 1. We then separate $y(t)$ into N/τ equal size nonoverlapping boxes, where N is the total number of data points and τ is the fluctuation function parameter.

In order to quantify the fluctuations present in the signal, a linear approximation is first constructed in each τ-size box using the Ordinary Least Squares method of estimation. Thus, we find the estimated model $\hat{y}(t) = \hat{\beta}_1 t + \hat{\beta}_0$ in each box. Though the regression may be much more rigorous, i.e. of higher order, a first order model is the convention. We will provide further reasoning for using the linear fit after the initial outline of the method.

Given the estimated regression $\hat{y}(t)$ corresponding to a given discrete box, we construct the detrended fluctuation function $F(\tau)$ as the root mean square deviation between $y(t)$ and $\hat{y}(t)$. Thus,

$$F^2(\tau) = \frac{1}{N} \sum_{t=1}^{N} |(y(t) - \hat{y}(t)|^2. \tag{2}$$

The summation is performed over all boxes to give an average fluctuation $\langle F^2(\tau) \rangle$ as a function of τ. This same analysis is then performed for all possible time scales (τ values). For example, the orange lines in Figure 1 correspond to an average fluctuation for boxes of a given size (τ_{orange}) and produce the orange dot in Figure 2; the blue lines correspond to an average fluctuation for boxes of a different given size ($\tau_{blue} = 2\tau_{orange}$) and produce the blue dot in Figure 2. We expect a power-law behavior given by,

$$\langle F^2(\tau) \rangle^{\frac{1}{2}} \sim \tau^{\alpha}. \tag{3}$$

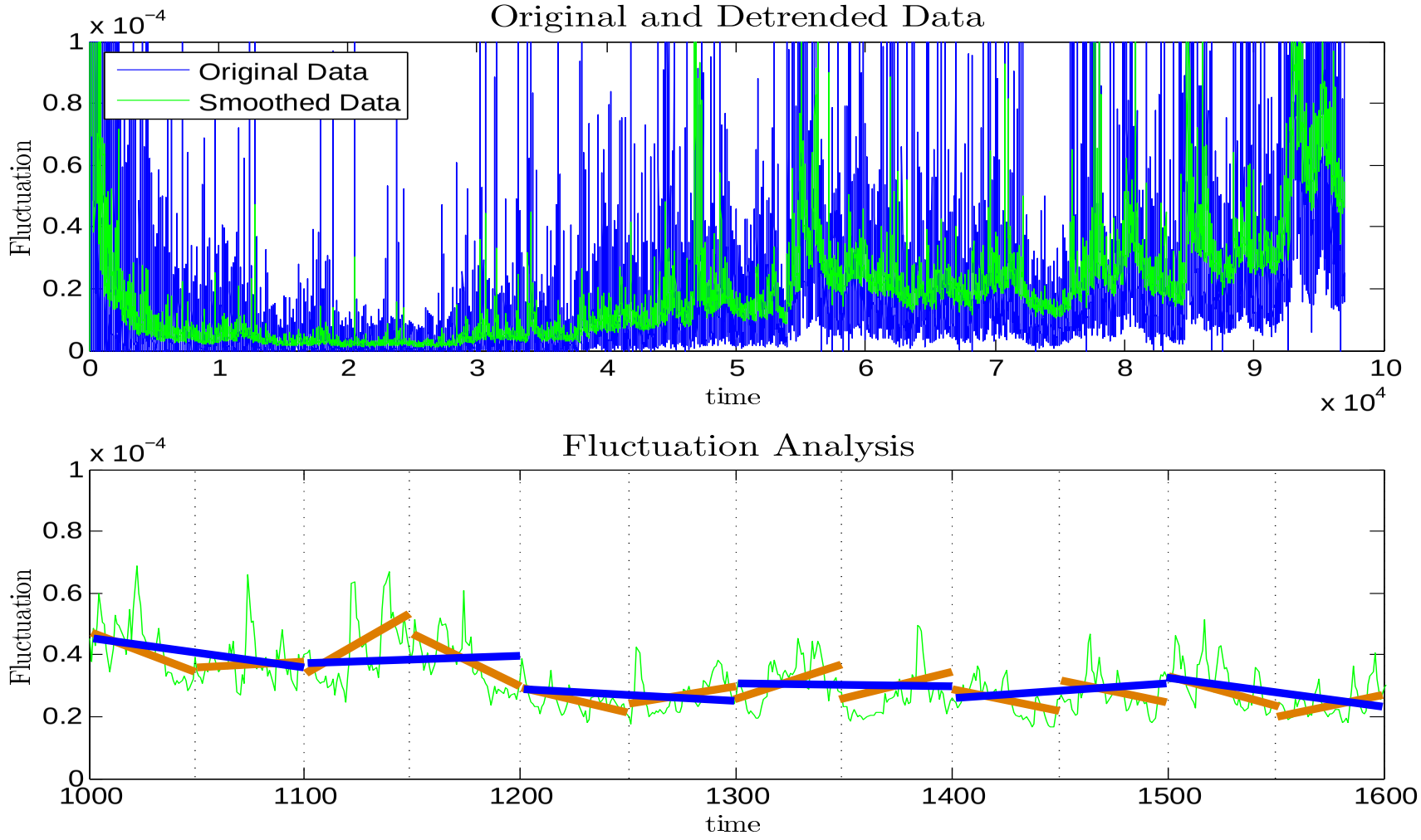

Figure 1: The top plot depicts both the raw data signal (blue) and the integrated signal (green) according to equation (1). The integrated data is then separated according to a given window size (the section from t=1000 to 1600 is depicted in the lower plot).

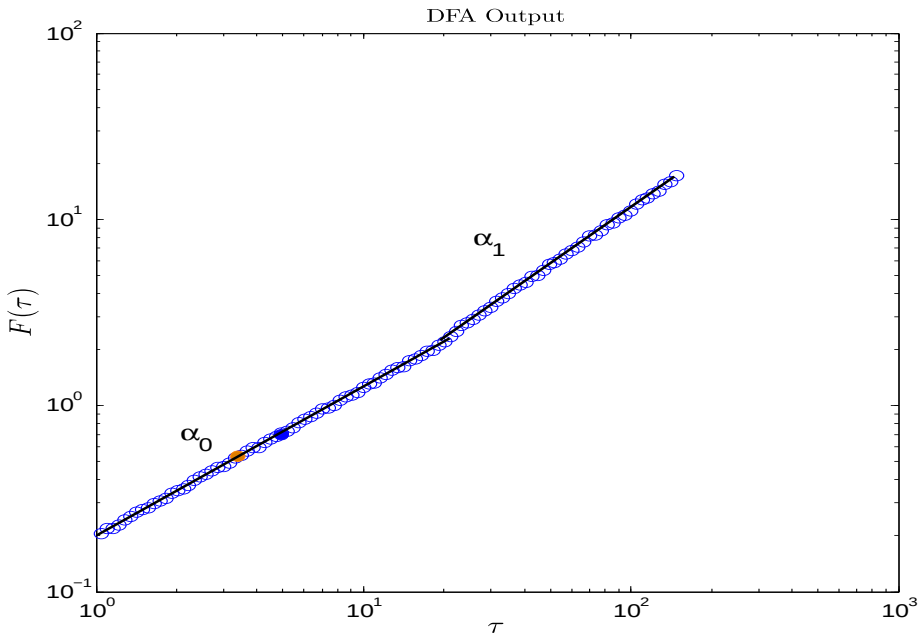

Figure 2: A root-mean-square fluctuation is calculated from a linear fit in every window in the bottom plot of Figure 1. We then plot the average fluctuation as a function of window size.

Examining the output of the DFA as a log-log relation, the α exponent emerges as a linear slope parameter. The interpretation of this slope parameter is straight forward:

$\alpha = 0.5$ corresponds to no long term temporal correlations and is indicative of random walk-based white noise in which each data point is completely uncorrelated with any previous data,

$0 < \alpha < 0.5$ is indicative of long-range power-law anti-correlations, sometimes referred to as "anti-persistent" correlations [4],

$0.5 < \alpha < 1$ is indicative of long-range power-law correlations, and

$\alpha > 1$ indicates the existence of correlations, though they no longer obey a power-law form. For example, $\alpha = 1.5$ corresponds to the integration of white noise commonly referred to as "brown noise."

It has been shown that α is directly related to the Hurst exponent [5] [6] and the signal fractal dimension [9]. Further, the proceeding interpretation is only valid within a finite range of the possible values of τ. Intuitively, as the time window shrinks, it becomes increasingly difficult to demonstrate fractality. Similarly, repeating fluctuations (reoccurring movements in the data which would constitute fractal-like relations) become more prevalent as the size of the time window increases, i.e. τ increases. Empirically, Ausloos defines this *scaling range* to be

roughly between $log(\tau) = 1$ and $log(\tau) = 2.6$ for most real or virtual foreign exchange currency (FEXC) rates [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]. We assume that such constraints on the values of τ are valid in the S&P 500 as well.

As a final note in the explanation of the DFA, it has been shown that the results of the analysis are surprisingly robust against more accurate box fits. For example, using a cubic model $\hat{y}(t) = \hat{\beta}_3 t^3 + \hat{\beta}_2 t^2 + \hat{\beta}_1 t + \hat{\beta}_0$, and proceeding in a similar fashion to estimate the β coefficients using a least squares algorithm, produces a DFA output only slightly different than the original linear estimation [22].

Though the usage of the DFA in biophysics requires a slightly different technique, application of the DFA to any data presupposes discrete time data, which advances at a constant rate, i.e. time steps are identical throughout the data. Further, it is assumed that there are no missing data points; as pertaining to financial markets, this means that the analysis ignores weekends, holidays, and market downtime. Because actual trades rarely occur according to a perfect time pattern—for example, trades only occurring at precisely every 5 sececonds—this assumption is almost always violated in the raw data. Thus, in order to obtain a constant sampling frequency and for the results of the DFA to be valid, data is usually discretely binned; the bin mean, rather than the data itself, is then examined.

Peculiarly, both the foreign exchange markets and interbeat heart intervals (commonly referred to as "RR intervals" and the initial application of the Detrended Fluctuation Analysis in biophysics) demonstrate a distinct pivot in the linear fitting of the DFA output [2] [7]. Two separate sets of points emerge with different α values, depending on the range of τ examined. Though the explanation may not shed light on the pivot in RR intervals, an interesting explanation has been given for the FEXC data. Ausloos recognizes that large τ values are usually fitted with α near to .5, indicating a random walk or no fluctuation correlations. As the time scale shrinks, however, persistent or anti-persistent relationship emerge. Ausloos proposes that persistent power-law relations ($.5 < \alpha < 1$) have corresponded with free-market (and "runaway") conditions, whereas anti-persistent behavior ($0 < \alpha < .5$) has accompanied strong political controls. Such a novel connection, that is, a relationship between the Hurst exponent and public policy, may prove helpful in examining the effects of various governmental controls.

As was previously mentioned, the Detrended Fluctuation Analysis has produced intriguing results in the field of biophysics. Particularly, it has been demonstrated that the relative health of an individual can be gauged (including the diagnosis of atrial fibrillations and congestive heart failure), simply based on the slope parameter of a DFA of the corresponding RR data [23]. Can this principle be

applied elsewhere? How robust is the analysis to parameter changes (i.e. binning size, window size, etc.)? We seek first to replicate previous work by others, identifying long range correlations in the fluctuations of security prices. Second, we examine the relative effects of bin size and data-filtering. Finally, we briefly explore the possibility of extracting meaningful information from the localized slope parameter, i.e. a possible metric of the corporate "health" of an underlying security.

II Preliminary Normalization

In order to answer these questions, we perform a DFA of historic trade data for an ETF following the S&P 500 (SPY). We chose to use all recorded during-hours trades occurring from January 1993 to December 2002. This data was obtained using the Wharton Research Data Services (WRDS) available via the University of Pennsylvania and constitutes tick-by-tick trade records, spanning the nine year range, for SPY. The high volume and resolution (recorded trades) make SPY an ideal candidate for DFA. See Figure 3. Further, we assume major trends, i.e.

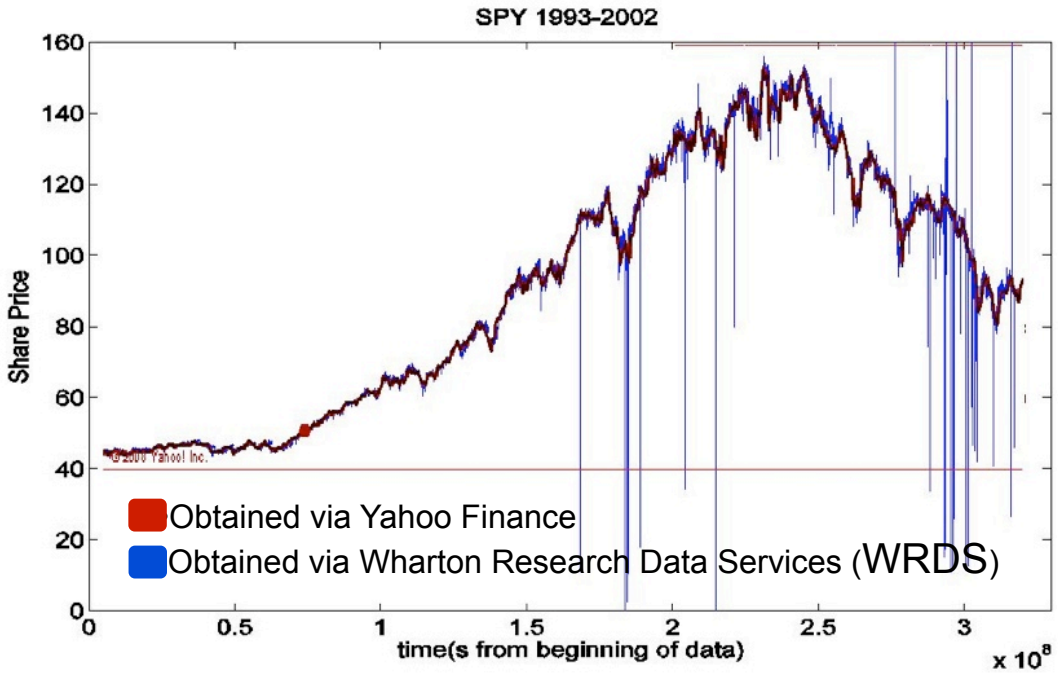

Figure 3: This is the raw dataset which was used. Notice the large number of outliers in the original data (blue).

inflation, are adequately addressed by the detrending process previously described. Notice the red line is the same data obtained via Yahoo Finance [8]. As we will demonstrate, the large number of outliers have a peculiar effect on the output of the

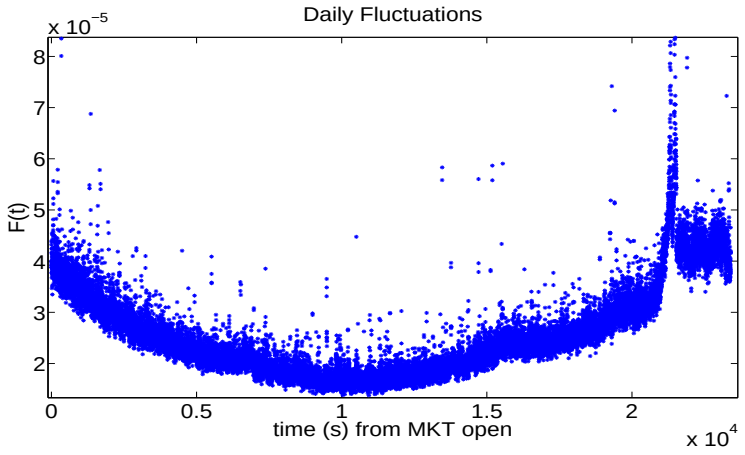

Figure 4: This is the average daily fluctuation. We note the higher uncertainty at the opening of the market, which decreases to reach a minimum around noon, and then increases towards the close of the market.

DFA. Though data is usually quite dense (especially beyond 2002, where there are approximately 15-30 data points per second), we extrapolate between sparse data points such that our shareprice matrix S has trade information for every second that the market is open. Fluctuations were then calculated similar to Liu et al.[21], except that S corresponds to the share price and does not explicitly reflect the number of outstanding shares. Thus, we define the fluctuation G at time t according to

$$G(t) = ln(S(t + \Delta t)) - ln(S(t)) \cong \frac{S(t + \Delta t) - S(t)}{S(t)}. \tag{4}$$

which represents a normalized fluctuation from one transaction price (at time t) to the next (at time $t + \Delta t$). We chose to not address *interday* fluctuations, though it may be fruitful to examine the results of expanding A (defined by equation (5)) to include *intraweekly* trends. As has been noted, the intraday fluctuations must be removed in order to avoid spurious, faux-correlations [21] (see Figure 4). We thus calculate the intraday pattern vector A in the following manner:

$$A(t_{\text{day}}) \equiv \frac{\sum_{j=1}^{N} |G(t_{\text{day}})|}{N}. \tag{5}$$

where t_{day} represents a specific time throughout the trading day, $G(t_{\text{day}})$ represents the fluctuation at time t_{day} given via equation (4), and the index j ranges over all N trading days in the dataset. In short, given a specific intraday time, the corresponding entry $A(t_{\text{day}})$ represents the average fluctuation at a given time of all trading days. Thus, $A(t_{\text{day}})$ is calculated for every second of the trading day, that is, $t_{\text{day}} \in [0, 23400]$. We then form the normalized fluctuation vector $g(t)$ by dividing every entry in $G(t)$ by the corresponding entry of $A(t_{\text{day}})$, that is

$$g(t) = \frac{G(t)}{A(t^*)} \quad \text{where} \quad t^* = t \mod 23400. \tag{6}$$

III Preliminary Analysis

We first perform the DFA on all nine years of data and obtain the output shown in Figure 5. We used 1 minute bins. Notice that a slope parameter of $\alpha_0 = .61$ corresponds to long-range power-law correlations; that is, large fluctuations are followed by large fluctuations, and small fluctuations are followed by small fluctuations. Further, note that the DFA seems to break apart for smaller window sizes.

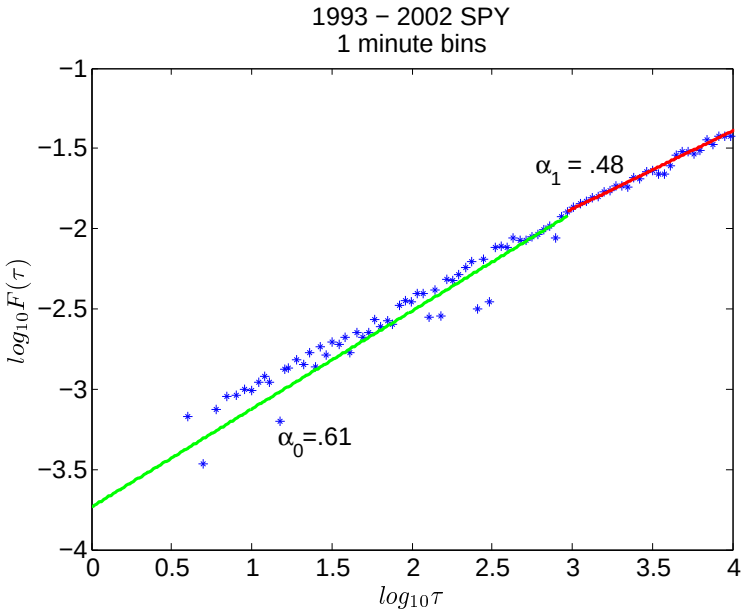

Figure 5: DFA performed on the entire dataset (1993-2002, inclusive). 1 min bins were used in order to achieve homogenous sample spacing. We note that $\alpha_0 = .61$ is indicative of long-range power-law correlations. Also, note the seeming disintegration of the analysis with decreasing window size.

IV Effects of Bin Size on DFA

In an effort to better understand the increase in noise for smaller windows, we perform the DFA on the entire dataset for various bin size. We consider three different bin sizes: 60 s, 300 s, 600 s for the data sample spanning 1993-2002; respective DFA results are shown in Figure 6. We note the increase in chatter noise with increasing bin size. Though we may preemptively suppose that this is demonstrative of a weakness of the analysis, it is in fact due to noise in the original dataset.

Figure 7 shows the same analysis performed on a filtered dataset. A simple mid-range filter is applied to the stock price set (we include all prices S such that $20 < S < 200$). Notice that the filtered DFA is smooth and we see a total disappearance of the short-range chatter. Further, the various DFAs (60 s, 300 s, and 600 s bins, respectively) are almost identical. A more rigorous analysis should be performed in order to more forcefully draw the conclusion of bin-size invariance

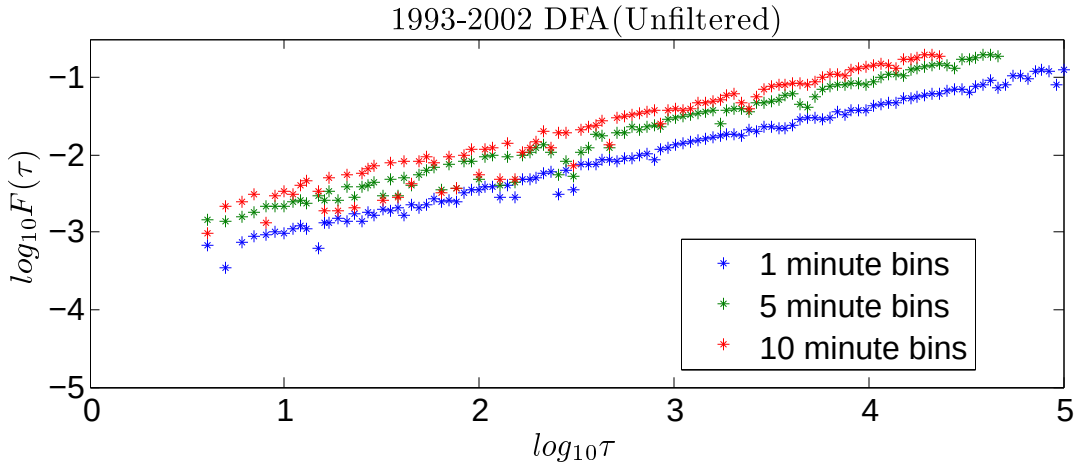

Figure 6: DFA performed on entire dataset using different bin sizes (60 s, 300 s, and 600 s). Notice the increasing noise at the lower end of the analysis output as we increase bin size. This would seem to suggest that the DFA becomes less and less robust against outliers with increasing bin size. In reality, this deterioration seems to be a result of noise in the original dataset.

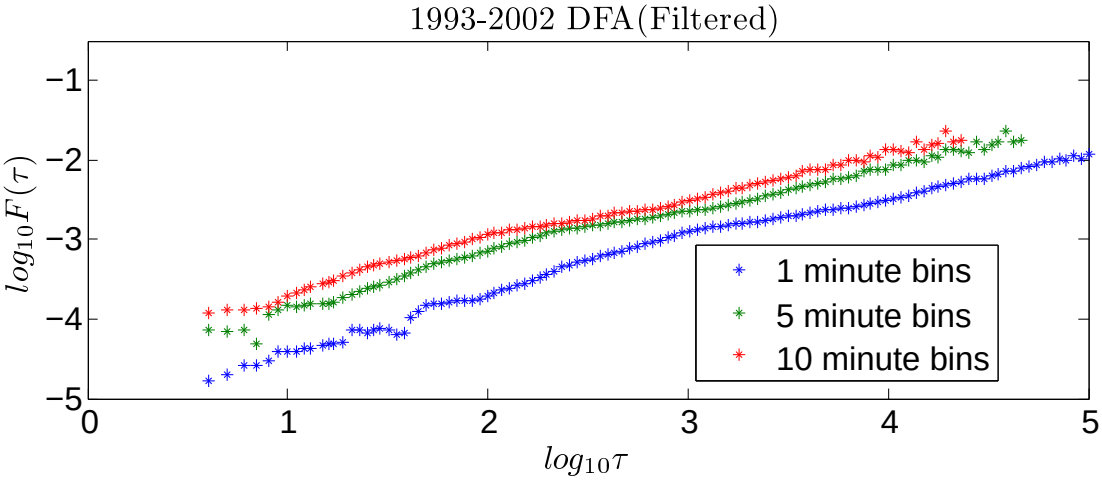

Figure 7: DFA performed on filtered dataset using different bin sizes (60 s, 300 s, and 600 s). Notice the relatively smooth DFA output, though the general shape and intercept is much different than the unfiltered data (compare with figure 6).

(an analysis including a wider range of bin sizes performed over a wider range of securities), but these preliminary results suggest that the DFA is robust against moderate variance in data bin sizes. Though this is noteworthy, it is also interesting to recognize the dramatic shift that occurs as a result of the filter (notice the change in intercept and overall shape).

V Localized DFA

The DFA examined in the previous section was performed on all nine years of SPY data. What are the results if we examine, instead, a localized DFA? For instance, what are the results of examining a year-by-year DFA spanning 1993-2002? We perform such an analysis for the raw (unfiltered) data as well as the filtered data (using the same mid-range filter as before). The results are shown in Figure 8 using 5 minute bins. Though a more powerful examination should be performed on these localized DFAs, there are two notable features.

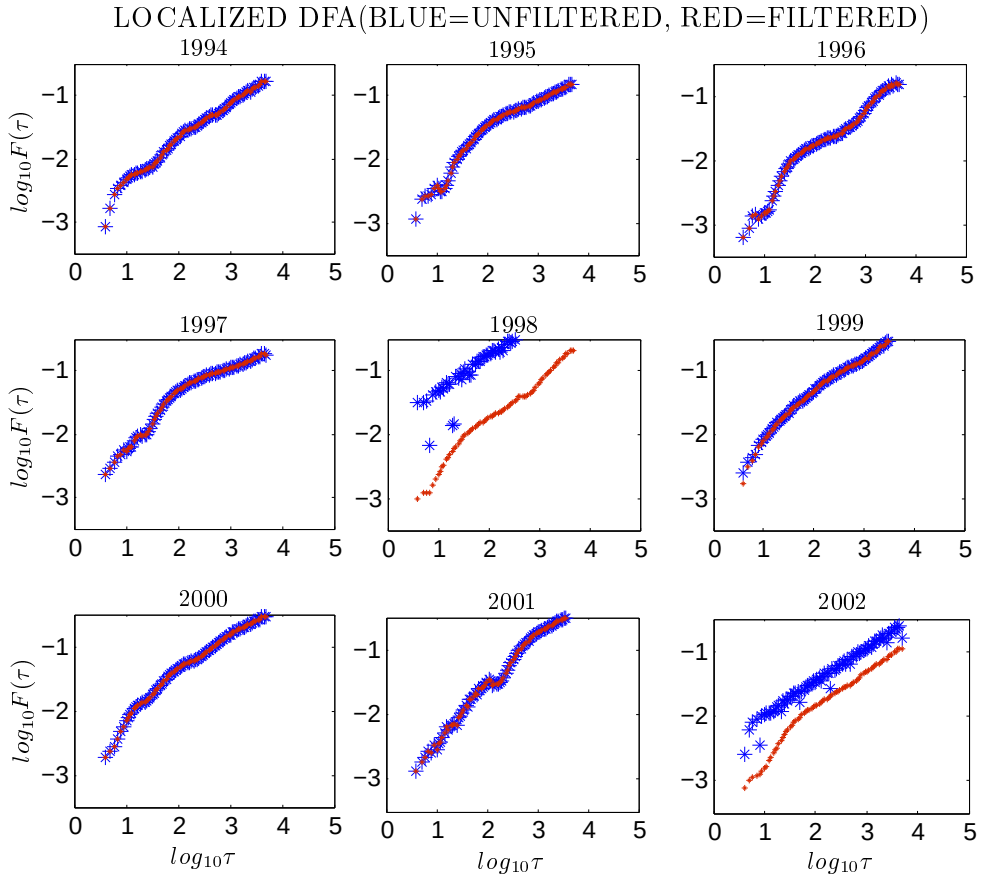

Figure 8: Localized DFA is reported for the years 1994 - 2002, inclusive. Blue is the original dataset without any filtering. Red is the same analysis but *after* a simple mid range filter, excluding every share price below 20 and above 200. Notice the strong overlap between the two analyses, except for the years 1998 and 2002, which contain most of the outliers.

We note strong agreement between the filtered and unfiltered data (note the almost perfect overlapping between the two sets, red (filtered) over blue (unfiltered)). This is surprising given the results of Figure 6 and 7. Namely, the

localized DFA is surprisingly robust against outliers; we propose that this is due, in part, to the fact that the localized DFA allows outliers to be safely "contained" within a single DFA and thus do not affect the other localized DFAs. We note that the two analyses that are different (occurring in 1998 and 2002) correspond to periods with a high number of data outliers. Filtering effects are demonstrated by the differences between filtered and unfiltered data points in the 1998 and 2002 DFAs. The strong agreement in all other years results from a low number of original outliers within those years. This seems to suggest that DFAs performed on datasets which are temporally shorter are more robust against outliers in the original dataset than DFAs on temporally longer data because we can effectively cull out the DFAs that have been affected by the outliers.

The second notable feature is the variance that we observe between the localized DFAs. Though the DFAs performed on 1994 and 1995 are relatively similar, compare either of these with 1998 or 2001. It would seem that there is something fundamentally different about the year 2001 than, say, 1996. What is the nature of this difference? First, the variance may be due to noise, not the noise caused by incorrect data points (which have been filtered out), but the tick-by-tick noise fluctuations inherent in the "real" data points. Simply put, the stochastic movements of the price may explain the differences in DFA output.

Though this explanation is entirely possible, it is also possible that there is information in a given localized period that is not in a subsequent period, and the presence and absence of this information is manifested in the heterogenous DFA outputs.

Further, it may be possible to connect this "localized information" with a metric of corporate health, i.e. P/E ratios, firm capital, investment outlays, etc. A more rigorous analysis should be pursued to examine these various possibilities.

VI Conclusion

The Detrended Fluctuation Analysis has been employed with relevant results to financial data. Using the DFA, we have recognized long-range correlations in SPY. Further characterization of the DFA method has demonstrated that the analysis is robust against data outliers (though this robustness seems to wane as the underlying data set grows temporally because of the inclusion of data point outliers). Finally, the diversity of DFA output when the analysis is performed on successive data sections may be demonstrative of localized information. To examine this possibility, more in depth analysis should be performed on a wider multiplicity of securities. Such analysis may ultimately demonstrate the DFA's ability in financial markets, as in biophysics, to distinguish between health and sickness.

References

[1] GoldBerger, A. L. et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101, e215e220 (2000). [Circulation Electronic Pages ¡http://circ.ahajournals.org/cgi/content/abstract/101/23/e215¿ (13 June 2000); see also ¡http://www.physionet.org¿.]

[2] GoldBerger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23):e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/cgi/content/full/101/23/e215]; 2000 (June 13).

[3] Shieh S.-J., International Journal of Theoretical and Applied Finance 9 787-799(2006).

[4] M. Ausloos, *Econophysics and Sociophysics: Trends and Perspectives* (Berlin, 2006). Chpt 9.

[5] Malamud, B. D., Turcotte, D. L., J. Stat. Plan. Infer. 80, 173-196 (1999).

[6] Carbone A., Castelli G., and Stanley H.E., Physica A 344, 267-271(2004).

[7] Qiu T., Zheng B., Ren F., and Trimper S., Physica A 378, 387-398(2007).

[8] http://finance.yahoo.com/q?s=spy last visited March 08.

[9] Mandelbrot, B. B., J. Business 36, 294-298 (1963).

[10] Vandewalle, N., Ausloos, M., Physica A 246, 454-459 (1997).

[11] Ivanova, K., Ausloos, M., Physica A 265, 279-286 (1999).

[12] Ausloos, M., Ivanova, K., Braz. J. Phys. 34,504-511 (2004).

[13] Ausloos, M. Physica A 285, 48-65 (2000).

[14] Vandewalle, N., Ausloos, M., Int. J. Phys. C 9, 711-720 (1998).

[15] Ausloos, M., Vandewalle, N., Boveroux, Ph., Minguet, A., Ivanova, K., Physica A 274, 229-240 (1999).

[16] Ausloos, M., Ivanova, K., Physica A 286, 353-366(2000).

[17] Ausloos, M., Ivanova, K., Int. J. Mod. Phys. C 12, 169-196 (2001).

[18] Ausloos, M., Ivanova, K., Eur. Phys. J. B 27, 239-247 (2002).

[19] Ivanova, K., Ausloos, M., Eur. Phys. J. B 20, 537-541 (2001).

[20] Ausloos, M., Ivanova, K., in New Directions in Statistical Physics - Econophysics, Bioinformatics, and Pattern Recognition, (Ed. L.T.Wille), Springer Verlag, Berlin, 2004) 93-114

[21] Liu Y., Gopikrishnan P., Cizeau P., Meyer M., Peng C.-K., and Stanley H.E., Phys. Rev. E 60, 1390 (1999)

[22] Vandewalle, N., Ausloos, M., Int. J. Comput. Anticipat. Syst., 1 (1998), pp.342-349

[23] Goldberger, A. L., Amaral, L. A., Hausdorff, J. M., Ivanov, P., Peng, C. K., and Stanley, H. E. (2002) Proc. Natl. Acad. Sci. USA 99, Suppl. 1, 24662472.

[24] Peng C.-K., Buldyrev S.V., Havlin S., Simons M., Stanley H.E., Goldberger AL, Phys Rev E 49, 1685-1689 (1994)