FEATURES AFFECTING HURST EXPONENT ESTIMATION ON TIME SERIES

MWANGI JANE WAIRIMU

August 7, 2013

PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE DEGREE IN MASTER OF SCIENCE IN APPLIED STATISTICS OF JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY.

DECLARATION

This is to certify that this research project is my original work and has not been presented for any award in any other university or institution of higher learning. Information from all other sources has been duly acknowledged.

Students Name - Mwangi Jane Wairimu

Registration Number - Sc382-1220/2011

Signature. . . . . .

Date. . . . . .

Approval.

This project has been submitted for examination with our approval as the candidate's supervisors.

DR KIHORO . J . M.

Signature.....

Date.....

DR WAITITU . A . G

Signature.....

Date.....

Jomo Kenyatta University of Agriculture and Technology

Department of Statistics and Actuarial Sciences.

Acknowledgment.

I would like to give thanks to the Almighty God for giving me the grace to do this research project. I also wish to thank my supervisors Dr Kihoro J. M and Dr Waititu A. G. for their consistent support and encouragement in the past three years. Their initial ideas, insightful suggestions, and wise counsel have made the completion of this work possible. I have learned a lot from working with them. Their active attitude towards research, earnest, preciseness, and humor are rare assets to find. I am also grateful for the time they put into reading, amending, correcting errors, and putting forward constructive comments and suggestions in the writing of this project. I also acknowledge my husband Sam, and my classmate Ann, for their constant support, care, and love during my research. I am grateful to the Department of statistics and actuarial sciences for providing a great environment for studying and researching.

Nomenclature

ACF	AutoCorrelation Function
ACF	Autocorrelation Function
ADF	Augmented Dickey Fuller
AIC	Akaike Information Criterion
AR(p)	Autoregressive Process of order p
ARIMA	Autoregressive Integrated Moving Average
ARMA	Auto Regressive Moving Average
BIC	Bayesian Information Criterion
DFA	Detrended fluctuation analysis
DJIA	Dow Jones Industrial Average
DJIM	Dow Jones Islamic Market Index
GRETL	Gnu Regression, Econometrics and Time - series Library
H	Hurst Exponent
KTDA	Kenya Tea Development Authority

MA(q)	Moving Average Process of order q
PACF	Partial Autocorrelation Function
R/S	Rescaled Range Ratio
SBC	Schwarz Bayesian Criterion
SE	Standard Errors
SHI	Shanghai Index
SZBI	Shenzhen B-shares
SZI	Shenzhen A-shares
V/S	Variance Rescaled Statistic

Contents

Abstract

Forecasting is becoming increasingly relevant to producers and consumers in all markets today. Both for producers and consumers, forecasts are necessary to develop bidding strategies as well as negotiating skills at the market in order for both parties to maximize benefits. Due to fluctuations of the Kenya shilling strength, weather conditions, politics and many other variables in the markets, prices of goods are highly volatile. The volatility of prices following a pattern makes it a time series phenomenon. Everyone with a new prediction method wants to try it out on returns from a speculative asset, such as stock market prices. Papers continue to appear attempting to forecast stock returns usually with very little success. This project is aimed at estimating the amount of predictability of a time series data using the Hurst exponent index. The Hurst exponent (H) is a dimensionless estimator for the predictability of a time series. Initially defined by Harold Edwin Hurst to develop a law for regularities of the Nile water level, it now finds applications in financial data such as stock prices. The Hurst Exponent can be interpreted as a measure of the trendiness: To be more specific, different values of Hurst exponent imply fundamentally different price behaviors. It is a statistical measure used to classify time series into either a random series or a trend reinforcing series. The larger the index value is, the stronger the trend. In this study we have investigated the features of time series associated with different estimates of Hurst exponent. It is shown that series with large Hurst exponent can be predicted more accurately than those series with Hurst exponent value close to 0.50. The main focus of this study is therefore to determine: Why Hurst exponent index swings from persistence to anti-persistence. Estimating of the Hurst exponent for time series data plays a very important role in research of processes which show properties of auto-correlation. There

are many methods for estimating the Hurst exponent two of these methods, Higuchi method and Aggregated Variance method were used in this project. These methods are already researched. The lengths of the realizations are from 120 to 3000 values. The estimates of H are calculated for each time series using the methods mentioned above. Generated data was decomposed and Hurst exponent values estimated for each component. Samples of different sizes n from four different data sets are selected and the H estimates obtained for each sample. The minimum sample required for Hurst index estimation was obtained. Secondary data from GRETL software and real data of Tea prices for the past 10 years obtained from K.T..D.A, has been used to verify the statistical properties associated with Hurst Exponent.

Chapter 1

Introduction

1.1 Background of the study

Hurst was a hydrologist who worked on the Nile River Dam project for about 40 years during the early years of the last century, **?**. He tried to find out the ideal features for reservoir design. An ideal reservoir should discharge certain amount of water every year and should never overflow. However, the inflow of the reservoir varies due to changes in the climatic conditions. If the inflow of the reservoir is too low then releasing fixed amount of water would make it dry. Thus, he was confounded with the problem of fixing the water discharge policy, such that the reservoir will never be emptied nor will it overflow. In developing such a model, Hurst studied the inflow of water from rainfall. He measured how reservoir level rises and falls around its average and recorded range of the variations. If the series were random, the range would increase with the square root of time. To standardize the analysis, Hurst created a dimensionless ratio by dividing high-low Range of the reservoir by the Standard deviation of the time series (R/S ratio). Hurst watched that many natural phenomena like, rainfall, temperatures, river flow follow a biased random walk, which is a combination of random walk, trend and

noise. Time series predictability is a measure of how well future values of a time series can be predicted, where a time series is a sequence of observations X_t , $t = 1, 2, ...N$. Time series predictability indicates to what extent the past can be used to determine the future in a time series. A time series process with high predictability, is such that its future values can be predicted very well from the past values whereas a time series process with low predictability cannot be predicted very precisely, and its past values provide only a statistical characterization of the future values. In practice, a given time series is not simply deterministic or stochastic, but rather some combination of both. Predictability can be viewed as the signal strength of the deterministic component of the time series to the whole time series. In this study, the deterministic component is estimated by the estimates of the H index, whereas the stochastic component can be estimated by the residuals of the time series models. Thus, the time series predictability can be measured by the index H. Measuring the predictability of a time series is useful because it can make a prediction of the accuracy of the forecast if prediction is done. Therefore prediction of a time series with low predictability, such as a random walk time series, can be avoided. For a low predictability time series, past observations are of little use in predicting future values, and the future values are determined randomly or by unknown factors.

The Hurst exponent was initially used in hydrological studies but it has been applied to many research fields. The use of this exponent has also become popular in the financial studies largely due to work of Peters (1991, 1994) as quoted by, **?** . The Hurst exponent provides a measure for long-term memory and predictability of a time series, **?**. The objective of this paper will be to develop some insights on time series movements by comparing Hurst exponent (H) in different situations. The value of Hurst exponent will give some clue whether present value of the series depends on past values of the time series. According to the theory of Brownian motion, $H = 0.5$ implies an

independent innovation process i.e. the events are mutually independent throughout time, where past values do not influence the present values, **?**. A persistent market return series is characterized by presence of a long memory. In a persistent market, a positive (negative) change in a period is likely to be continued in the same direction in the following period. The strength of the trend-reinforcing behavior increases as the value of the Hurst exponent approaches unity. The closer the Hurst exponent is to 0.5, the more the price movement appears random and become more unpredictable. Anti-persistent processes are mean-reverting, which means that if the market has been in a particular direction in the previous period, it is more likely to move in the reverse direction in the following period. The strength of this mean reverting behavior depends on how close the Hurst exponent is to zero. Hurst observed that the H value directly depends on range to standard deviation ratio (R/S ratio), **?** . Thus R/S ratio can provide a method of classifying time series, which can be useful in identifying which markets have greater predictability. While forecasting a time series, we need to know whether the time series under study is predictable or not. If the time series is random, its H index will be below 0.5 and no forecasting method will be successful. A time series with a large R/S ratio has trending characteristics and such series is more predictable than a series with a low R/S ratio. There have been numerous publications in the area of nonlinear time series modeling and prediction over the last ten years, but few have studied the predictability of a time series.

The reason why Hurst Exponent is such a valuable asset in technical analysis is that it provides a means of classifying time series in terms of predictability.

- $H \leq 0.5$ This indicates a time series with anti-persistent behavior. This means an increase of a value is followed by a decrease and a decrease is followed by an increase. This means future values will have a tendency to return to a longer term mean value. The strength of this mean reversion increases as H approaches zero.

- $H = 0.5$ indicates a random (a Brownian time series). In a random walk, there is no correlation between any element and a future element and there is a 50% probability that future return values will go either up or down. The time series is an independent identically distributed stochastic process, following normal distribution.

- $H \geq 0.5$ indicates persistent behavior i.e time series is trending. An increase will tend to be followed by an increase and a decrease by a decrease. The larger the value of H the stronger the trend, **?**.

This analysis only means that series falling in the last category give better prediction results than series falling in the first two categories.

1.2 Problem statement

Taking data and directly modeling it to do forecasting can be very frustrating when one lands into data that cannot be predicted. Previously researchers have used residual analysis tests like the sample ACF , Portmanteau test, Turning point test, Difference-sign test and Normality checking in order to determine whether there exist dependence in observations so that they are able to predict future values. However residual analysis tests do not guarantee predictability of the data. In this study it is shown that there is need to subject data to predictability test before beginning to model and forecast. Hurst exponent is determined for four different sets of data in order to test for their predictability, then features as to why this index swings from persistence to anti persistence are investigated.

1.2.1 Problem justification

Hurst Exponent is a different measure from volatility. An index or fund may have a relatively low volatility but can still have a Hurst Exponent close to 0.5. Mature markets may have H values close to 0.5 than emerging markets indicating that they could be more efficient but less predictable. Hurst Exponent estimation will help us determine which assets to forecast and which ones to ignore. This is particularly useful in cases where time series models have higher predictability.

The contribution of this research is three fold.

- Analyze the behavior of Hurst exponent on simulated time series.
- Apply the concepts found on simulated data to secondary data.
- Compare these behaviors with real data.

1.3 Objectives

On the estimation of Hurst exponent on time series data;

1.3.1 Main objective

To investigate why H index swings from persistence to anti persistence.

1.3.2 Specific objectives

1. To determine the Hurst exponent of different sets of data in order to classify them in terms of predictability.
2. To investigate the impact of the size of the series on Hurst Exponent estimation and determine the minimum sample required for the estimation of Hurst exponent.

3. To investigate the effect of spikes on Hurst Exponent.

1.4 Research hypothesis

1. The size of data contributes to predictability of a time series.

2. Spikes increase predictability of data.

3. Any sample size is valid for Hurst exponent estimation.

Chapter 2

Literature review

2.1 Hurst exponent

Hurst exponent (normally denoted by H) is used in areas such as applied mathematics, fractals and chaos theory, long memory processes, and spectral analysis. It has different but related meanings in different contexts. Hurst exponent is a measure of whether the data is a pure random walk or has underlying trends and hence, it is considered as a measure of predictability of a series. Random Gaussian process with an underlying trend should have some degree of auto correlation, **?**. If this auto correlation has a very long (infinite) decay or long range correlations, it is referred to as a long memory process with a Hurst exponent value $0.5 < H < 1.0$. This long memory behavior could be due to a sudden impact that affects a process. In such process, although the impact is sudden, the underlying process takes some time to come back to its normal behavior. This is due to the memory which is carried through with the process itself. For example, although a large buy or sell order can cause a sudden change in stock price, stock price behavior takes some time to come back to its normal operation. Hurst estimate can be used as an indication of this type of behavior of processes. It can always be used

to compare behaviors of memory-less processes like random walks, **?**. It is a measure of persistence i.e the characteristic or tendency of underlying series to continue in its current direction. If the Hurst exponent value is between 0.5 and 1, the process can be considered as a persistence series (which has positive auto correlation) meaning that if the process has an increment between times $t-1$ and t, then there is a high possibility of having an increment between times t and $t+1$ as well. If H is between 0 and 0.5, it is an anti-persistence series (which has negative auto-correlation). In other words, if the process shows an increase between times $t-1$ and t, there is a high possibility of having a decrease in between t and $t+1$. If it is equal or closer to 0.5, this implies that it is a random and unpredictable series. This behavior is called "mean reversion" , **?**.

Hurst exponent is a critical characteristic quantity to understand the underlying price dynamics in a financial or commodity market. Thereby the correct or precise estimation of Hurst exponent is one of fundamentally important problems in computational financial economics literature, since an inaccurate estimate may mislead researchers to arrive at a dubious or even fallacious conclusion on whether a specific market is persistent with long memory, or efficient whose price behaviors follow random walks, or anti-persistent featured with a mean reverting process. In current literature, there are various tools proposed to estimate Hurst exponent but, **?** compared two methods R/S and V/S by means of Monte Carlo simulations, and disapproved some scholars (e.g. Andrew Lo) who underestimated the R/S method. According to, **?** both methods can overestimate the Exponents for some cases, but underestimate them for the others.

?, used five different estimation methods: Aggregated Variance Method, Higuchi Method, Peng Method, R/S Method and Periodogram Method on the data of absolute returns of the DJIA , S&P 500 and the Islamic index DJIM. He showed that there exists lower Hurst index when Periodogram method was used, on DJIA, S&P 500 and DJIM producing the value of Hurst index as 0.736, 0.6849 and 0.6189 respectively. Most of

the Hurst exponent index for the other methods, on the DJIA and S&P 500, took the values of between 0.85 to 0.97 in most cases.

H value of 0.50 signals presence of Brownian motion. When the value of H lies between $0 < H < 0.5$, it suggest trend reversing characteristics in the series. Conversely, value of H within the range of $0.5 < H < 1$ suggest presence of trend in the series. The power of the trend increases until value of H reaches its upper ceiling value of one, **?**.

Application of Hurst exponent in financial time series has gained momentum with the work of Peters as quoted by, **?** who estimated Hurst exponent for monthly returns on the S&P 500 from January 1950 to July 1988. From a sample of individual stocks, Peters noted that Hurst exponents varied from 0.54 for Consolidated Edison to 0.75 for Apple Computer. As the H values are greater than 0.5, he found persistence among stock returns than would be expected if stock prices followed a Brownian motion.

As for the values of Hurst exponent, there are several crucial implications. If H is equal to 0.5, the random walk is implied, **?**. Therefore, if one arrives at the value of H of 0.5, one has a weakly efficient market. The same author says that a persistent process is characterized by Hurst exponent significantly higher than 0.5 and implies rejection of independence which in turn rejects random walk and consequently implies an efficient market.

?, quotes Corazza and Malliaris who studied some foreign currency markets and found that Hurst exponent was statistically different from 0.5 in most of the samples. Besides, they also found that the Hurst exponent is not fixed but it changes dynamically overtime. They interpreted that foreign currency returns follow either a fractional Brownian motion or a Pareto-Levy stable distribution.

?, analyzed the Hurst exponent for all trading-day periods of the Dow-Jones index from January 1930 to May 2004. They found that the periods with large Hurst exponents could be predicted more accurately. This suggests that stock markets do not

show random walk in all periods. Some periods have strong trend structure and this structure can be learned to benefit forecasting. They classified financial data series of different periods and experimented with back-propagation neural networks to show that series with large H can be predicted more accurately. The authors inferred that the H provides a measure for predictability.

Extreme values or outliers are the values of a time series which are unusually different compared to other data. These values could distort the overall underlying movement of a time series by affecting the trend. It is necessary to detect and correct for outliers in order to improve modeling of the three time series components (trend, seasonal and irregular components), **?**.

According to **?**, the estimator can be underestimated or overestimated and special caution is suggested when a repeating pattern of significant jumps is present in time-dependent Hurst exponent estimates.

2.2 Decomposition

A time series is a collection of the values of some quantitative characteristic observed at regular intervals of time. These observations may be primary data or indices produced from them. A time series can be classified into two groups, the stock series which shows the output of some activity measured at different points of time, and the flow series reflecting the measured activity over a given period. Time Series can be decomposed into four main unobserved components, **?**.

Trend (T) component indicates the long term tendency, represents the structural variations of low frequency in a time series.

Seasonal component (S) is that part of variations in a time series which represent intra-year fluctuations more or less stable year after year with respect to timing,

direction and magnitude. It is also referred to as the seasonality of a time series. It reflects normal variations that recur every year to the same extent, e. g. weather fluctuations that are representative of the season, length of months, Christmas effect, etc. It may also include calender related systematic effects that are not regular in their annual timing and are caused by variations in the calender from year to year.

Cyclical component (C) indicates the medium term fluctuation. The cyclic component is worth examining only in case of very long time series. In accordance with the general practice, the trend component is assumed to include also the cyclic component. Sometimes the trend and cyclic components together are called trend-cycle.

Irregular component (I) includes unpredictable effects, which are considered as random variables; it is assumed that the expected value of these factors is zero (for additive model) or one (for multiplicative model). The irregular component of a time series is the residual time series (remaining component) after the trend, the seasonal and the cyclic components have been removed. Decomposition methods can be used to separate out these components of a time series. The function "decompose" in R is called to break data into various components. The additive model for a given time series X_1, X_2, ...X_n is the assumption that these data are realizations of random variables X_t that are themselves sums of the four components

$$X_t = T_t + S_t + C_t + R_t \; t = 1, 2, ...n$$

where T_t is the trend and a monotone function of t and C_t reflects some non random long term cyclic influence. Think of the famous business cycle usually consisting of recession, recovery, growth and decline. S_t describes some non random short term seasonal component, whereas R_t is a random variable grasping all the deviations from an ideal non-stochastic model suitable if the amplitudes of both the seasonal and irregular components do not vary as the level of the trend varies. A multiplicative model is

suitable if the amplitudes of both the seasonal and irregular variations increase as the level of the trend increases. A multiplicative model cannot be used when the original time series contains very small or zero values for each of its components. In this case, a pseudo additive model (a combination of additive and multiplicative models) is used. Pseudo additive model assumes that seasonal and irregular variations are both dependent on the trend but independent of each other. The pseudo-additive model continues the convention of the multiplicative model to have both the seasonal factor and the irregular factor centered around one, **?**.

2.3 Stationarity

A stationary process is one whose statistical properties do not change over time, **?**. More formally, a strictly stationary stochastic process is one where given t_1 . . . , t_l the joint statistical distribution of $X_{t1}, ..., X_{tl}$ is the same as the joint statistical distribution of $X_{t1+\tau}$, . . . $X_{tl+\tau}$ for all l and τ . This is an extremely strong definition: it means that all moments of all degrees (expectations, variances, third order and higher) of the process, anywhere are the same. It also means that the joint distribution of (X_t, X_s) is the same as (X_{t+r}, X_{s+r}) and hence cannot depend on s or t but only on the distance between s and t, i.e. $s—t$. Since the definition of strict stationarity is generally too strict for everyday life a weaker definition of second order or weak stationarity is usually used. Weak stationarity means that mean and the variance of a stochastic process do not depend on t (that is they are constant) and the auto covariance between X_t and $X_{t-\tau}$ can only depend on the lag τ (τ is an integer, the quantities also need to be finite). Hence for stationary processes, X_t, the definition of auto covariance is

$$\gamma(\tau) = cov(X_t, X_{t+\tau}),$$

for integers τ . It is vital to remember that, for the real world, the auto covariance of a stationary process is a model, albeit a useful one. Many actual processes are not stationary as we will see in section 3.2. Having said this much, fun can be had with stationary stochastic processes! One also routinely comes across the auto correlation of a process which is merely a normalized version of the auto covariance to values between -1 and 1 and commonly uses the Greek letter ρ as its notation: $\rho(\tau) = \gamma(\tau)/\gamma(0)$, for integers τ and where $\gamma(0) = cov(X_t, X_t) = Var(X_t)$.

?, identified lack of ergodicity, stationarity, and independence, and identified the degree of early persistence of the Chinese stock markets when they were more regulated. They used index series are from the Shanghai (SHI) stock market and Shenzhen A-shares (SZI) and B-shares (SZBI) , before and after the various de-regulations and re-regulations. By computing the Hurst exponents, they identified the market's later degrees of persistence. The empirical evidence revealed that SHI, SZI, and SZBI are moderately persistent with Hurst exponents slightly greater than 0.5. They showed that these stock markets were more persistent before the de-regulations, but that they now move like geometric Brownian motions, i.e. the markets have become more efficient in recent times.

According to **?**, R/S is a method constructed for stationary time series, but there are some methods like DFA that are immune to non-stationarities and thus can be used for both stationary and non-stationary data sets.

2.4 Forecasting

In the time domain applications, **?**, expressed X_t as a linear combination of previous values X_{t-1}, X_{t-2} .....X_{t-p} of the currently observed series. According to him, the outputs X_t may also depend on lagged values of another series, say Y_{t-1}, Y_{t-1},

.....Y_{t-q}, that have influence. It is easy to see that forecasting becomes an option when prediction models can be formulated in this form. Time series smoothing and filtering can be expressed in terms of local regression models. Extensions to filters of infinite extent can be handled using regression in the frequency domain. In particular, many regression problems in the frequency domain can be carried out as a function of the periodic components of the input and output series, providing useful scientific intuition into fields like acoustics, oceanographic, engineering, bio-medicine, and geophysics. The assumption of linearity, stationarity, and homogeneity of variances over time is critical in the regression. Time series forecasting forecasts future events based on known past data. One example is to predict the opening price of a stock based on its past performance. Some popular models are auto regressive moving average (ARMA) and auto regressive integrated moving average (ARIMA). Lento (2009) as quoted by **?**, tried to find synthesis between technical analysis and fractal geometry. He argued that Hurst exponent was developed from the field of fractal geometry and provides a statistical technique to identify the nature of any dependencies in a time series. Moreover, Technical analysis has developed various trading rules that are premised on the belief that past price data reveals patterns that can be used to predict future prices. Based on this logic, and further empirical analysis he found that time series with high H resulted in higher profits in case of trending trading rules and time series with low H resulted in higher profits from contrary trading rules.

?, found that the persistence of the United States' stock market is not constant. The series of Hurst exponents calculated using the most recent ninety days of returns changes dramatically with respect to time. Both of the data that he used i.e, the S&P 500 and the Russel 2000 experience periods of above normal persistence as well as go through periods of anti persistence. The market persistence for large capitalization stocks and small capitalization stocks behave in much the same manner over short periods of time.

Previous research show that these effects are small over short durations.

Chapter 3

Methodology

3.1 Introduction

It is reasonable to ask whether a given financial time series is predictable before modeling the data and trying to forecast its development. The Hurst Exponent gives us a clue as to whether the data at hand is predictable or not. This index is not so much calculated as estimated. Its accuracy and fidelity are also limited as there are various factors that influence this index. In order to compare Hurst index estimates, time series data (ARMA process) was simulated using R Statistical package. This data was then decomposed, and Hurst exponent estimated for each component. This was repeated on several data sets as we searched for the component that really contributes to the value of H index.

There are many methods of estimating Hurst exponent in the literature: Rescaled range analysis, Absolute values of the aggregated series (Ag-gabs), Aggregated variance (agg-var), Differenced variance method (diff-var), Higuchi method and others. In this study, Higuchi and the Aggregated Variance method were used because they are less complex to apply and the estimates can be computed easily using some software pack-

ages in R and GRETL. Secondary time series data obtained from GRETL software, i.e Bollerslev and Ghysels Exchange Rate Data, 1974 Daily Observations, on Nominal return on Mark/Pound exchange rate, and The Dow Jones Industrial Average (DJIA) also known as the Industrial Average index were also used. These indices are selected as they are some of the major indices in the world. DJIA index data (1980-1989) is the second oldest market index in the US after the Dow Jones Transportation Average, introduced by the same co-founder. Dow Jones Industrial Average is the most widely used indicator of the overall condition of the stock market, a price-weighted average of 30 actively traded blue chip stocks, primarily industrials. The 30 stocks are chosen by the editors of the Wall Street Journal (which is published by Dow Jones & Company), a practice that dates back to the beginning of the century, **?**. The Dow was officially started by Charles Dow in 1896, at which time it consisted of only 11 stocks. The Dow is computed using a price-weighted indexing system, rather than the more common market cap-weighted indexing system. Simply put, the editors add up the prices of all the stocks and then divide by the number of stocks in the index. (In actuality, the divisor is much higher today in order to account for stock splits that have occurred in the past). These sets of data were selected in such a way that the former has low H index while the latter has higher H index. Finally real data on Tea prices from KTDA was used to verify the application of Hurst Exponent to the real world market. The following steps will be followed for different data sets selected:

1. Stationarity tests.

2. H index Estimation for different samples and components of time series.

3. Induction of Spikes and estimation of H.

4. Model identification and parameter estimation .

5. Forecasting.

3.2 Stationarity

A stochastic process (a collection of random variables ordered in time t) is said to be (weakly) stationary if its mean and variance are constant over time, i.e. time invariant (along with its auto-covariance), **?**. Such a time series will tend to return to its mean (mean reversion) and fluctuations around this mean will have a broadly constant amplitude. Alternatively, a stationary process will not drift too far away from its mean value because of the finite variance. By contrast, a non-stationary time series will have a time-varying mean or a time-varying variance or both. e.g a random walk model is a non-stationary process. There are reasons why we need to test for stationarity before modeling some of which are:

1. The stationarity or otherwise of a series can strongly influence its behavior and properties - e.g. persistence of shocks will be infinite for non-stationary series. In this project, H index of non-stationary data and stationary data were compared, and conclusions made as to whether Hurst exponent is affected by the stationarity or non-stationarity of data.

2. Spurious regressions. If two variables are trending over time, a regression of one on the other could have a high R^2 even if the two are totally unrelated.

3. If the variables in the regression model are not stationary, then it can be proved that the standard assumptions for asymptotic analysis will not be valid. In other words, the usual "t-ratios" will not follow a t-distribution, so we cannot validly undertake hypothesis tests about the regression parameters.

Most time series display some kind of trending behavior. It is therefore necessary to find out whether the series are stationary or not. If the series is non-stationary we attempt to reduce it to an apparent stationary series. This can be done by: taking logs,

or ratios, or differencing the series X_t with respect to t. In our study, first differences were taken to achieve stationarity. According to some authors, evidence of long-term memory could be spuriously caused by non-stationarity in the time series itself, **?**.

There are several tests of stationarity, we focus on a test which has become popular over the past years: This is the unit root tests called Augmented Dickey-Fuller tests (ADF). This unit root test as implemented in GRETL, uses the t-statistic on φ as in the following regression,

$$\Delta y_t = \mu_t + \varphi y_{t-1} + \sum_{i=1}^{p} \gamma_i \Delta y_{t-i} + \varepsilon_t. \tag{3.1}$$

This test statistic is one of the best known and probably the most widely used unit root test. It is a one-sided test whose:

- Null hypothesis is $\varphi = 0$ (There is unit root and time series in non-stationary), versus the
- Alternative $\varphi < 0$ (Unit root does not exist), **?**. Large negative values of the test statistic lead to the rejection of the null. Under the null, y_t must be differenced at least once to achieve stationarity; under the alternative, y_t is already stationary and no differencing is required. One peculiar aspect of this test is that its limit distribution is non-standard under the null hypothesis: moreover, the shape of the distribution, and consequently the critical values for the test, depends on the form of the t term. Once stationarity of the series is achieved, H indices of the series are estimated.

3.3 Methods of H index Estimation.

Many methods of estimating the Hurst Exponent exist in literature but in our research, Aggregated variance and Higuchi's methods are used. These methods are less complex to apply and are also estimable by computer soft wares such as R software (fractals library package). The methods also work well in many empirical cases. As far as the estimation is concerned, there is yet a perfect method that is agreed by all researchers. Each method has its own drawbacks and cannot be used as a sole estimator in all cases. Comparison of the two methods is done to establish the method that gives better H estimates. In some cases estimates on H are outside the defined range $(0 < H < 1)$. This is probably due to the calculation method and the properties of the time series. We summarized some two heuristic estimators that were used throughout this study quoted by, **?** and shown below:

3.3.1 Aggregated variance

The entire series is partitioned into m contiguous groups. These are groups that share the boundaries. Within each group, the variance (relative to the mean of the entire series) is evaluated. A measure of the variability of this statistic between groups is calculated. The number of groups, m, is increased and the process is repeated. The observed variability changes with increasing m in a way related by theory to the Hurst parameter H of the input series. For the methods used here, a log-log plot of variability versus number of groups is, ideally, linear, with a slope related to H, so H can be determined by linear regression.

Divide the original time series into blocks of size m and average within each block, that is consider the aggregated series;

$$x^m(k) = \frac{1}{m}\sum x_i$$

for successive values of m. The index k, labels the block. Then take the sample variance of

$$x^m(k), k = 1, 2, ...$$

within each block. This sample variance is an estimator of $var(x^m)$. Since, for fractional Gaussian noise and fractional ARIMA, $var(x^m) \sim \sigma^2 m^\beta$as $m \to \infty$ where

$$\beta = 2H - 2 \leq 0$$

we can obtain an estimate for β, or H, by proceeding as follows.

For a given m, divide the data, $x_{1,}, x_2....x_N.$, into N/m blocks of size m, calculate

$$x^m(k), k = 1, 2, ...N/m.$$

and its sample variance

$$var(x^m) = \frac{1}{\frac{N}{m}}\sum_{k=1}^{N/m}(x^m(k))^2 - \left(\frac{1}{\frac{N}{m}}\sum_{k=1}^{N/m}(x^m(k))^2\right)^2$$

Repeat this procedure for different values of m and plot the logarithm of the sample variance versus $log(m)$. Since $\widehat{var(x^m)}$is an estimate of $var(x^m)$, the resulting points should form a straight line with slope.

$$\beta = 2H - 2, -1 \leq \beta \leq 0$$

In practice, the slope is estimated by fitting a line to the points obtained from the plot.

It is assumed here that both m and N are large, and that $m << N$, so that both the length of each block, and the number of blocks is large. If x has (short-range or) no dependence, the slope obtained should equal -1.

3.3.2 Higuchi's method

This method was suggested by Higuchi (1988). It involves calculating the length of a path and, in principle, finding its fractal dimension D. The method is in fact very similar to the method of absolute values of the aggregated series. It involves taking the partial sums,

$$Y(n) = \sum_{i=1}^{n} x_i$$

of the original time series X_i $i = 1, 2.....N$, (e.g., producing fractional Brownian motion from fractional Gaussian noise) and then finding the normalized length of the curve, given by:

$$L(m) = \frac{N-1}{m^3} \sum_{i=1}^{m} \left[\frac{N-i}{m} \right]^{-1} \sum_{k=1}^{\frac{N-i}{m}} Y\left(i+km\right) - Y\left(i+\left(k-1\right)m\right)$$

where N is the length of the time series, m is essentially a block size and [.] denotes the greatest integer function. Then,

$$EL\left(m\right) \sim C_H m^{-D}$$

where $D = 2–H$. Thus a log-log plot of $L\left(m\right)$versus m should produce a straight line with a slope of $D = 2–H$.

The two methods named above have been used with R software package.

3.4 Sample size on H index estimation.

In the literature, no work on the minimum sample size required for the estimation of H index was found. In this paper the results of H index for different sizes of data on different data sets are presented. The values of H for all the data sets were varied within the interval $0 < H < 1$. The length of sample sizes tried varied from 95 to about 2528. The sizes of data were obtained by taking the whole size n, half of the data $\frac{n}{2}$, a quarter of the data $\frac{n}{4}$, an eighth of the data $\frac{n}{8}$, a tenth of the data $\frac{n}{10}$, a sixteenth of the data $\frac{n}{16}$, $n = 100$ values, $n = 97$ values, $n = 96$ values and $n = 95$ values. Two things are illustrated:

1. Whether the Hurst exponent values measured from short series differs from the index for the whole set of data.

2. Whether any size of data may be used to estimate the H index.

3.5 Decomposition of time series and Estimation of H index.

There are four components (forces that determine the observed values) of a time series.

- **Secular Trend**

These are the long term movements which give the general way in which the data moves over a long period of time. A graph of observed values versus time may show small ups and downs, but the long term movement eliminates these minor variations and looks at the bigger picture. A graph of this trend is called a trend curve.

- **Cyclic Movements**

These are movements that happen in regular long-term cycles. In business, for example, cycles consist of alternating periods of recession and inflation, recovery and prosperous times. Cyclic movements are large scale and should be very clear in the data.

- **Seasonal Movement**

These are also movements that happen in regular cycles, but repeat yearly. The patterns are caused by either natural conditions such as weather fluctuations or man-made conditions such as business, administrative, political procedures, festive seasons etc. Both cyclic and seasonal variations are referred to as periodic variations. Although the periodicity is annual in business and economic theory, the idea of seasonality can be extended to include periodicity over any interval of time such as daily, hourly, weekly, etc depending on time.

- **Irregular Movements**

These are variations that are not accounted for by trend, cyclic, or seasonal movements. They are unpredictable and are as a result of chance events such as strikes, floods, earthquakes, etc.

A time series can be decomposed into its components: trend, cycles (including seasonal) and irregular components:

$$X_t = T_t + S_t + R_t \tag{3.2}$$

Where;

X_t =value of the series at time t.

T_t =trend component of the series.

S_t =cyclical or seasonal component with a period S.

R_t =random effect for which we have no explanation.

This idea is very old but it is still widely used. Put bluntly, non predictive information is useless to any research work, and it therefore makes sense to isolate the predictive information from the non-predictive one. It will turn out that most of the information we collect over a long period of time is non predictive, so that isolating the predictive information must go a long way towards separating out those features of the sensory world that are relevant for the time series behavior.

H index for all the components was calculated using the methods above and a conclusion made on the component(s) that really contribute to a high or a low index. Removing one or several components is also studied and observations made on any changes in the index.

3.6 Spikes on Hurst exponent

Spikes are abrupt and generally unanticipated extreme changes in the data values. Within a very short period of time, the system value can increase substantially and then drop back to the previous level. This is especially so for commodities that normally have extreme volatility. Price spikes are temporary features in a financial time series. The authors in literature find different reasons for price spikes: exogenous shocks ('supply shocks'), the structure of bidding and also long-term trends in market factors are found to cause high spot power prices, **?**. Many experimentally obtained series are substantially inhomogeneous. In order to illustrate this statement, a simulation is performed by adding several jumps of random magnitude and sign (up or down) at random locations in the simulated series and the Hurst exponent of the resulting series calculated and compared to the original H index. This was done to investigate whether

price spikes need a different kind of treatment than "normal" prices.

3.7 Forecasting

In time series forecasting, the first question we want to answer is whether the time series under study is predictable. If the time series is random, all methods are expected to fail, **?**. We want to identify and study those time series having at least some degree of predictability. From the literature, we know that a time series with a large Hurst exponent has strong trend, thus it's natural to believe that such time series are more predictable than those having a Hurst exponent close to 0.5. Prediction is one of the fundamental problems in research problems today. The best prediction model is the one that gives the forecasts with minimum errors. Researchers have come up with prediction methods that sometimes fail terribly. In this study, measuring the predictability of the data is insisted before testing any new model that comes up in the market today. A forecast is characterized by its origin and the lead time. The origin is the time from which the forecast is made (usually the last observation in a realization) and the lead time is the number of steps that the series is predicted denoted by h. The forecast of future observation X_{t+1} made from origin t going ahead h periods is denoted by $X_h(h)$. The aim is to show that the forecasts of series with H index less than 0.5 have higher Mean Squared Errors than for the series whose H index is more than 0.5. A lead time $h = 12$ was used. Before forecasting, models were fit on the data sets whose forecasts were required and parameters of estimation in the models identified. These models and parameter estimation were fit on the GRETL software.

3.7.1 Model identification and parameter estimation

The theory of stationary process to fit appropriate models in the time series was used. The first step involves model identification. For $AR(p)$, $MA(q)$, and ARIMA (p, d, q) models, there was need to know before hand their orders by identifying the values of p, q, and d. For seasonal ARIMA $(p, q, d)(P, D, Q)$, there was also need to know the values of P, D, and Q.

Model identification may be done using several methods but the two main methods are:

- Use of ACF and $PACF$

For stationary time series X_t the theoretical $(ACFs)$ and $(PACFs)$ are compared with those computed from the observed series. The idea is to compute the sample $ACFs$ and sample $PACFs$ and visually inspect their correlogram and compare it with that of the theoretical equivalents. By inspection of the auto correlation function of the suitably stationary series a tentative model can be selected. Non-stationarity is often indicated by an auto correlation plot with very slow decay i.e the tendency of the ACF not to die out quickly is an indication that a root close to unity exists. The degree of differencing d necessary to achieve stationarity is reached when the ACF dies out fairly quickly. Having tentatively decided what d should be, the general appearance of the ACF and $PACF$ of the differenced series is studied so as to provide clues about the choices of the orders of p and q for the AR and MA operators.

- Use of Information Criterions

Alternative approach to model selection is the use of information criterion such as Akaike Information Criterion (AIC) and/or Bayesian Information Criterion (BIC) as quoted by, **?**. The criterions use maximum likelihood method in their estimations. The models that yields minimum values for the criterions are preferred, and the AIC or BIC

values are compared among various models as the basis for selection of the model. The estimation of $ARIMA$ parameters in practice is not straightforward. However there exists computer algorithms for $ARIMA$ parameter estimation. Information Criterions were used to fit models on the data sets chosen for forecasting.

Chapter 4

Empirical Results

In order to compare Hurst indices estimates using the two methods proposed above, four sets of data were used.

1. Simulated data from $GRETL$ software denoted x
2. Two sample data sets from $GRETL$ denoted D and B
3. Real Data for Tea prices from 2002 to 2011 obtained from $KTDA$ denoted T

The plots of the four data sets are shown in fig 4.1 to fig 4.4

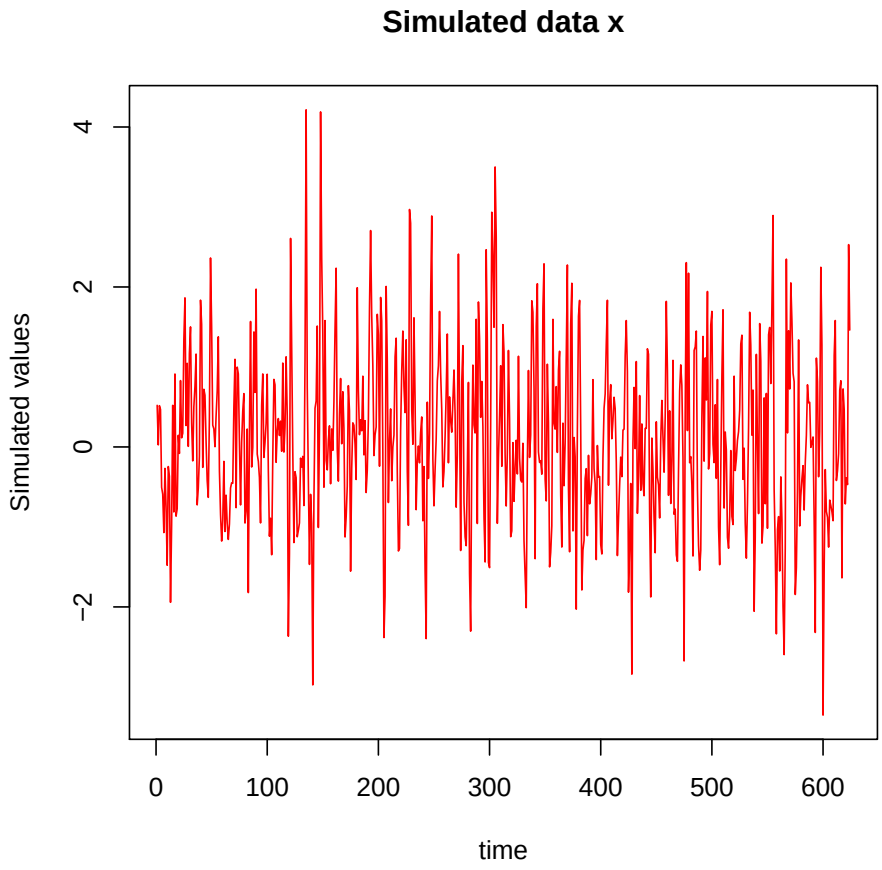

Figure 4.1: Graph of Simulated data

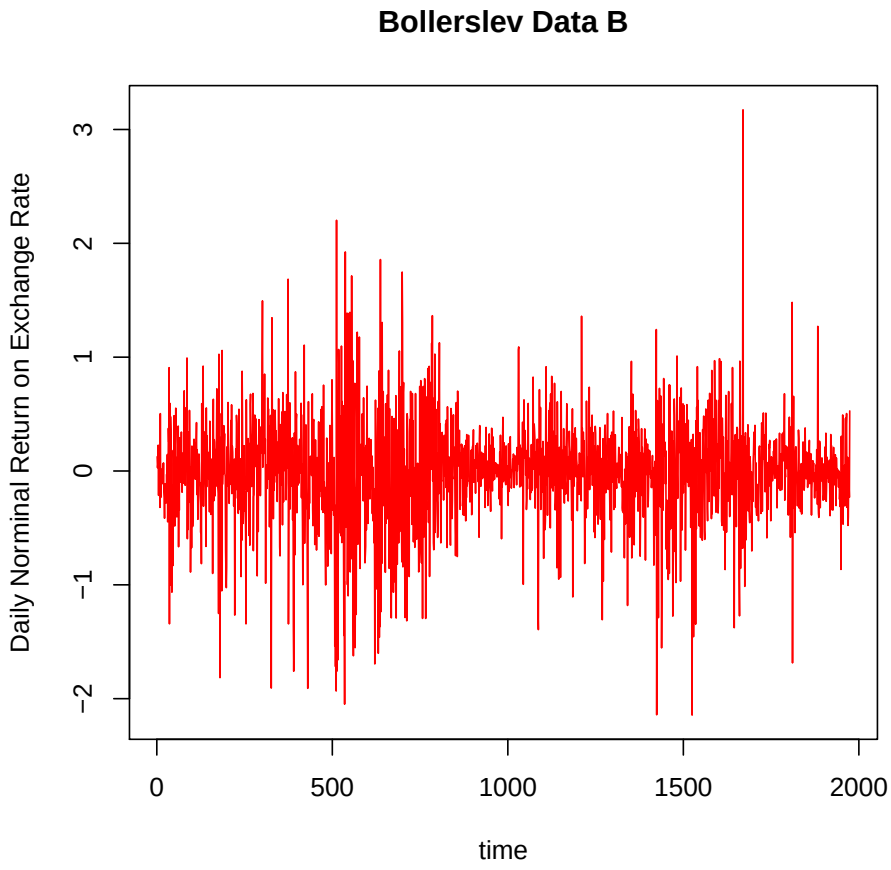

Figure 4.2: Graph of Bollerslev data

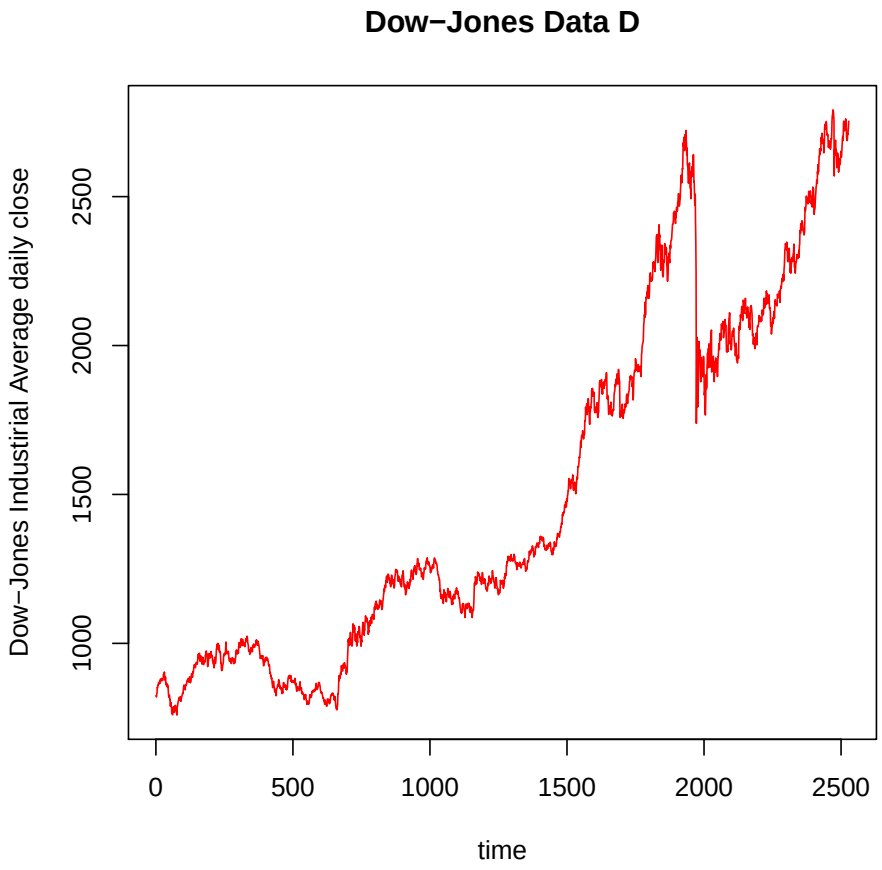

Figure 4.3: Graph of Dow-Jones data

Figure 4.4: Tea prices graph

4.1 Stationarity Tests

Before data was subjected to any analysis, it was first subjected to stationarity tests. Stationarity simply means that the fundamental form of the data generating process remains the same over time. Here, the weak form of stationarity is referred to. The stationarity or otherwise of a series may influence its behavior and properties e.g. persistence of shocks will be infinite for non-stationary series. There is need therefore to test whether the sets of data to be used are stationary or not. From the plots of the sets of data used in this project, we may deduce that some data sets like in figure 4.1 and

4.2 are stationary since their plots are jagged and they have a structure that does not wander too far from the mean of the data, meaning that things that happened recently are relatively more important than things that happened along time ago. There are several tests of stationarity, we focused on a test which has become popular over the past years: This is the unit root tests (Dickey-Fuller test) discussed in section 3.2. The hypothesis on the test statistic φ in the equation 3.1 are :

$H_0 : \varphi = 0$ i.e there is a unit root, and the series is non-stationary against

$H_A : \varphi < 0$ i.e. there is no unit root and the series is stationary.

The Null hypothesis is rejected if we get large negative values (with p values closer to zero.) of the test statistic φ and conclude that the series is stationary. The results of this test on the four sets of data are:

```
        Augmented Dickey-Fuller Test


data:  x
Dickey-Fuller = -8.6161, Lag order = 8, p-value = 0.01
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test


data:  B
Dickey-Fuller = -12.2923, Lag order = 12, p-value = 0.01
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test


data:  D
```

```
Dickey-Fuller = -2.2143, Lag order = 13, p-value = 0.4876
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test


data:  T
Dickey-Fuller = -3.2098, Lag order = 4, p-value = 0.0896
alternative hypothesis: stationary
```

Table 4.1 shows the summary results of ADF test of stationarity for the data sets $x, B, D.$ and T

Table 4.1: Augmented Dickey Fuller test Results

Data sets	**Dickey-Fuller**	**lag order**	**p-value**
Simulated data x	-8.613	8	0.01
Gretlbollerslev data B	-12.2785	12	0.01
Dow-Jones data D	-2.2143	13	0.4876
Tea prices data T	-3.2098	4	0.0896

Since the first two sets of data x and B have large negative values, and p values are closer to zero, we reject the null hypothesis of non-stationarity, and conclude that the series are stationary. However Dow-Jones data D and tea prices data T fail the ADF test since the test statistic φ has small negative values, and the p-values are far from zero. We fail to reject the null hypothesis and conclude that the series are non-stationary.

4.2 Minimum sample for H index Estimation

At best, H index provides a broad measure of whether a time series has a long memory character or not. H does not provide the local information needed for forecasting.

There is a difference between saying that there is predictability and the ability to predict. H shows that there is predictability in a time series. It is therefore never used for prediction in any direct way. It is only useful in analyzing the behavior of market models. No work have we read that shows how big a data set is required to estimate the H index. The table 4.2 shows Hurst values for different sizes in different data sets and the minimum sample required for H estimation.

Table 4.2: Hurst Exponent on different sample sizes

Sample size	**Simulated**	**Gretlbollerslev**	**Dow-Jones**	**Tea prices**
Full size(n)	0.57848	0.6000	1.024	1.102
n/2	0.56497	0.6209	1.021	-
n/4	0.60471	0.6332	1.021	-
n/8	-	0.4956	0.9865	-
n/10	-	0.5318	1.009	-
n/16	-	0.5222	1.019	-
n=100	0.558137	0.6233	1.069	0.9975
n=97	0.527711	0.6160	1.067	0.9722
n=96	0.520716	0.6358	1.069	0.9633
n=95	n size too small	n size too small	n too small	n too small

Hurst exponent was estimated for different time series from simulations and real data using values of different sizes and it was found that the H index values of the full series is not significantly different from the value of the index when the series is split into smaller series of a half, a quarter, an eighth, a tenth, or a sixteenth of the data. From the table 4.2 it should be noted that there are some instances where the H index goes outside the defined range ($0 \leq H \leq 1$). This is especially so to the sets of data that are not stationary, and we could say that non - stationary data does not give the best H estimates. The estimation method used is also another reason why the index goes beyond the stipulated range. Some of the estimation methods over estimate the exponent and others under estimate it as we shall see later. From the table 4.2 a sample size of 96 is arrived at as the minimum sample that is required to estimate H index,

series that have values below 96 will not give any estimate of H.

4.3 Decomposition of data and Estimation of H index.

Time Series Decomposition simply means to break down a time series into trend, seasonal, cyclical and irregular components as explained in section 3.5. The trend component stands for long term trend, the seasonal component is seasonal variation, the cyclical component is repeated but non-periodic fluctuations, and the residuals is the irregular component. The purpose of decomposition in this study was:

- To give a summary description of the prominent features of a time series. If the irregular and seasonal components are eliminated from the series, then we are left with an index which may give a clearer picture of the more important features of the time series.

- Estimate the H index for each component of the time series, and make deductions as to which component contributes significantly to the predictability of the entire series. The fig 4.5 to fig 4.8 show the decomposed series of data sets x, B, D, and T.

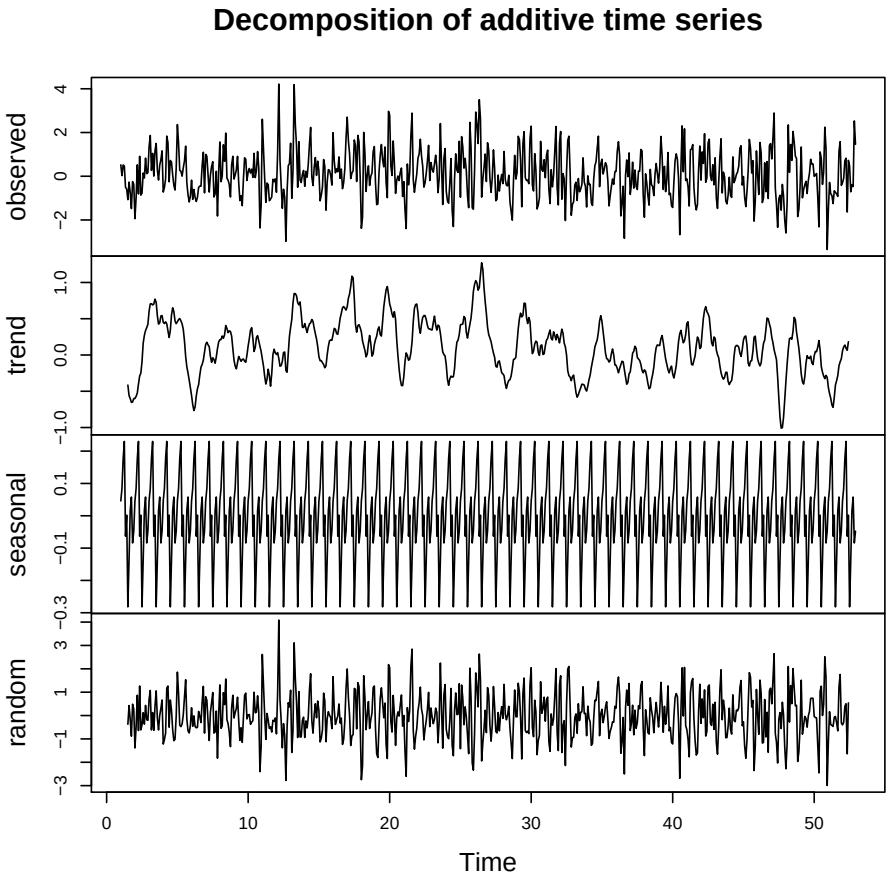

Figure 4.5: Decomposision of x

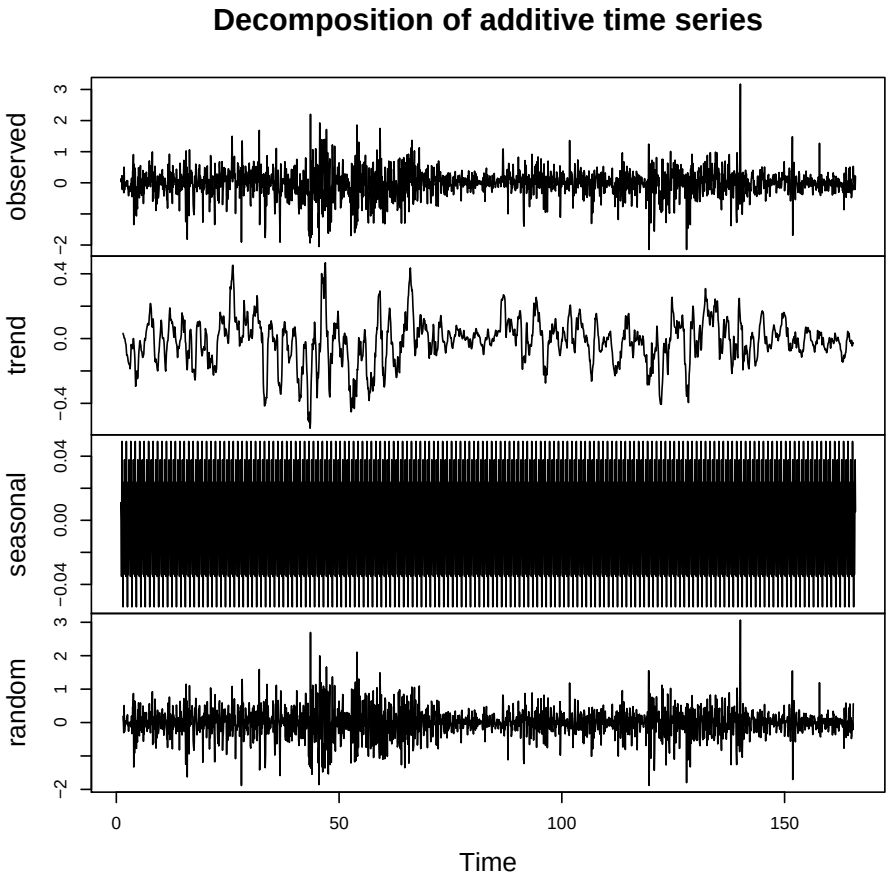

Figure 4.6: Decomposition of Bollerslev data

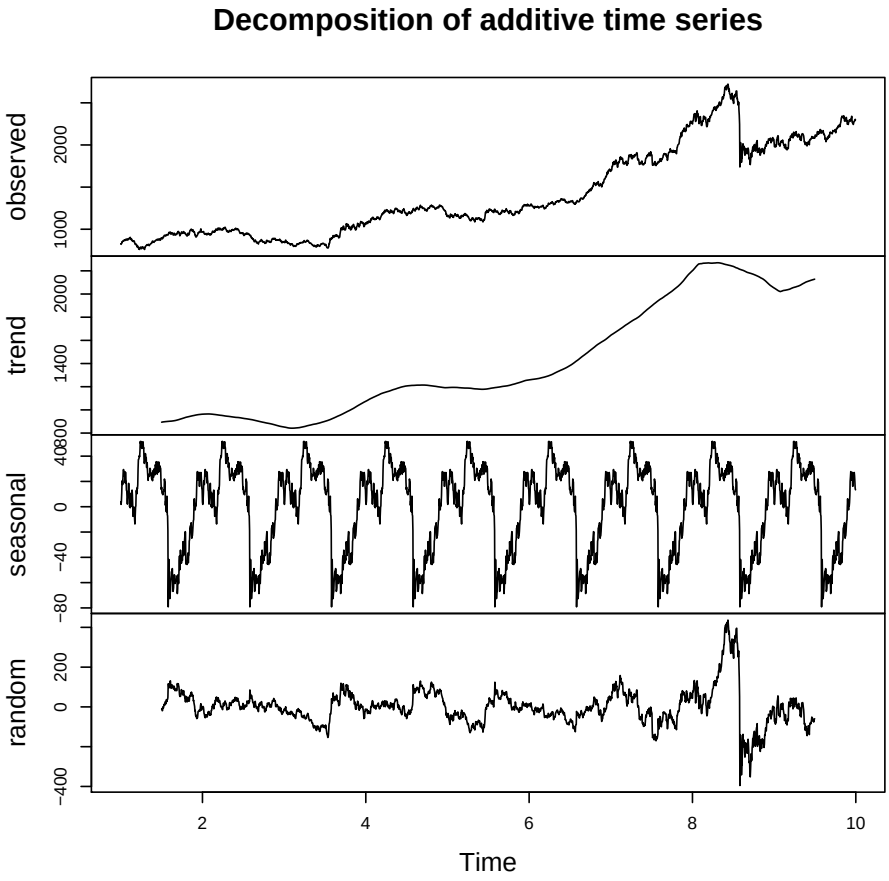

Figure 4.7: Decomposition of Dow-Jones data

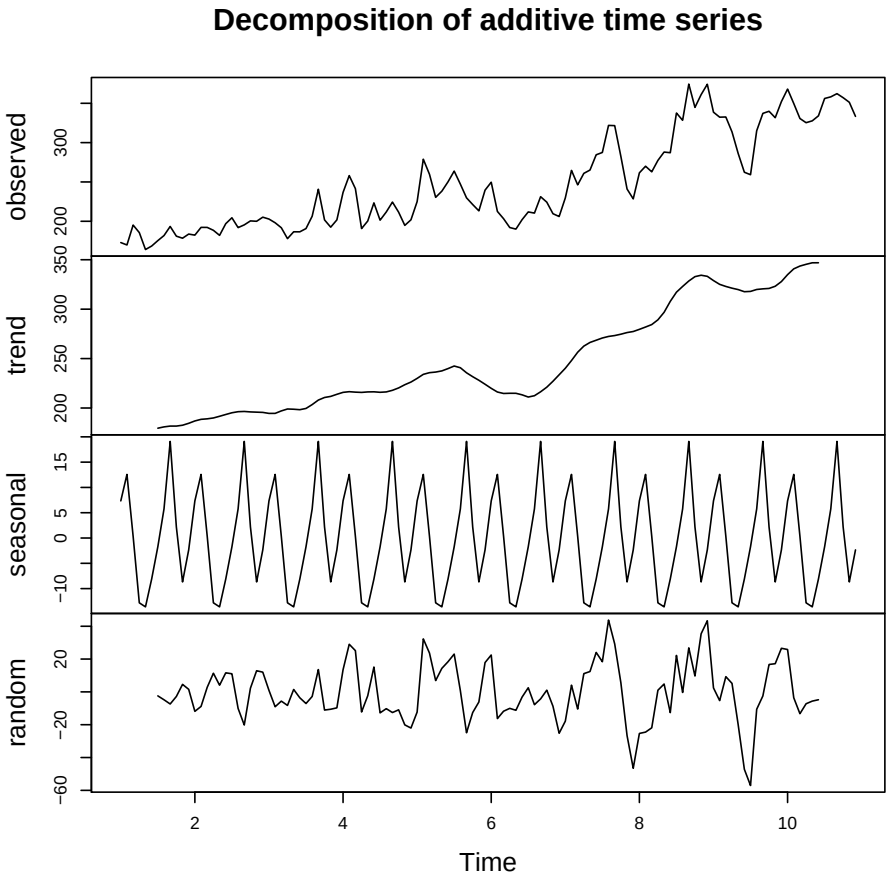

Figure 4.8: Decomposition of Tea prices

H index for each component of the decomposed series was estimated and compared to that of the entire series. Table 4.3 shows the different H indices for different components of the four data sets used in this project.

Table 4.3: Hurst Exponent on different Components.

Data/Method	**Whole data**	**Trend**	**Random**	**Seasonal**
x(Higuchi)	0.756	0.827	-0.002	-0.002
x(Agg-Var)	0.421	0.543	-0.324	0.024
B(Higuchi)	0.562	0.609	-0.007	0
B(Agg-Var)	0.336	0.429	-0.15	0.002
D(Higuchi)	0.993	0.995	0.286	-0.003
D(Agg-Var)	0.9	0.966	-0.419	-0.311
T(Higuchi)	0.987	0.988	0.055	0.013
T(Agg-Var)	0.621	0.859	-0.371	0

From the table a few things can clearly be deduced:

- The two methods of estimating the exponent used do not give the same values of the index. The Higuchi method over-estimates the index as compared with the Aggregated Variance. It is important to know the properties of the method one uses to estimate before making conclusions about the predictability of data.

- There are several instances when the H index goes below the stipulated range,

$$0 \leq H \leq 1$$

  especially when estimation involves the random and seasonal components. From the literature this is attributed to the estimation method used and also the fact that the series in question may be non stationary.

- Trend is the only component that has a high predictability since its H index is not significantly different from the H index of the entire series. Trend may therefore be said to solely contribute to the H index estimation and hence to the predictability of the entire time series.

- One method is not enough to make conclusions about the predictability of data. If two or more indices are above the 0.5 threshold, its enough to say that the

series is predictable. This was tested in section **??** where two time series were classified in terms of their predictability.

- What really makes the H exponent such a valuable asset in technical analysis is that it provides a means of classifying time series in terms of predictability. Out of the four sample sets used, only two are predictable i.e, Dow-Jones data D, and Tea prices data T since their indices for the two methods used are above 0.5. This indicates persistent behavior i.e the time series is trending. An increase will tend to be followed by an increase and a decrease by a decrease. Bollerslev data B from Gretel, and simulated data x both have low predictability since at least one of the H index is below 0.5. Thus on an overall basis, these series closely follows a random walk and therefore prediction of trend is not easy. The series have Anti-persistent behavior, this means an increase in a value will be followed by a decrease or a decrease followed by an increase. This behavior is sometimes called "Mean reversion" which means future values will have a tendency to return to a longer term mean value. The strength of this mean reversion increases as H tends to zero. Sub section 4.6.3 shows that only the data sets with a H index above 0.5 give better forecasts.

4.4 Stationarity on Hurst Exponent

Most methods of the Hurst exponent estimation are applied only to stationary time series, while a lot of natural, technical and information processes are non-stationary. The main type of non-stationarity, which occurs in practice, is the existence of trend and cyclical components. An attempt was made to carry out an analysis of the effect of stationarity on Dow-Jones data D and Tea prices data T. Both series are trending upwards, and non-stationarity forces us to remove the common trend. This was done

by first differencing. The first difference of x is the difference between x and its lag. This Differencing is done with R software package. The H index is then computed and compared with the original index.

Fig 4.9 and Fig 4.10 show the differenced series of the two data sets D and T .

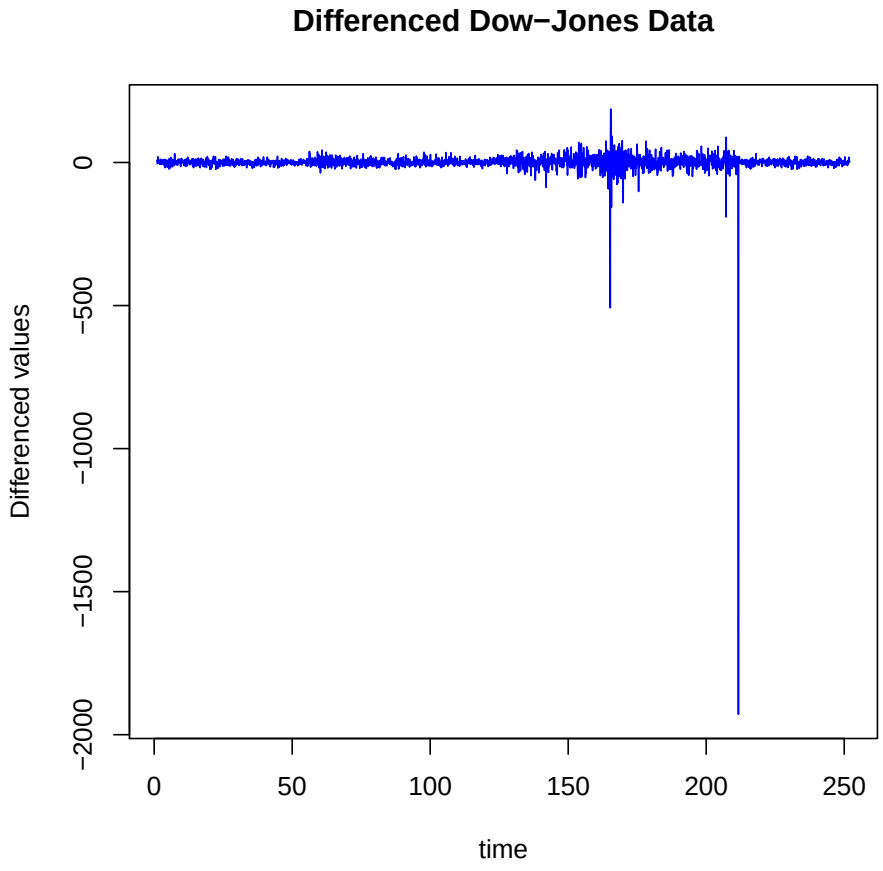

Figure 4.9: Graph of Differenced Dow-Jones data

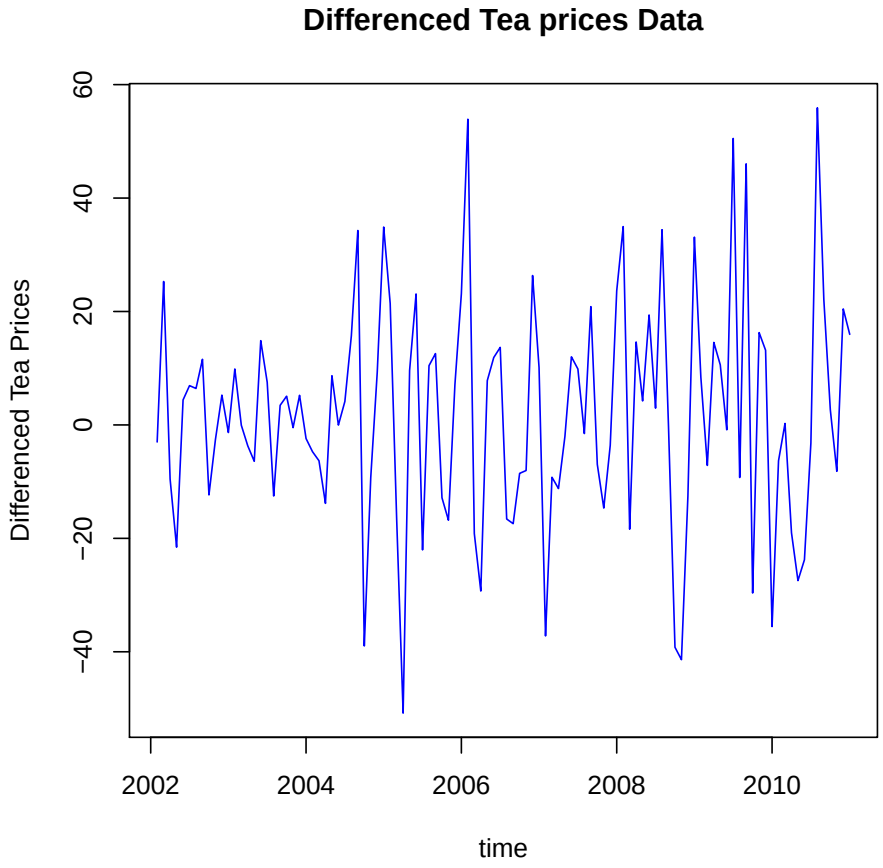

Figure 4.10: Graph of Differenced Tea prices

For the differenced series denoted DD and DT, the ADF test of stationarity was performed again to confirm that the data were now stationary.

```
        Augmented Dickey-Fuller Test


data:  DD
Dickey-Fuller = -14.3594, Lag order = 14, p-value = 0.01
alternative hypothesis: stationary

        Augmented Dickey-Fuller Test
```

```
data:  DT
Dickey-Fuller = -4.8602, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

The results of ADF test on the differenced series of Dow-Jones and Tea prices are shown above.

From the results, large values of the test statistic φ (-14.3594) , and (-4.8602) with p-values of 0.01 (closer to zero) leads us to a non rejection of the null hypothesis of non-stationarity and conclude that the series are stationary.

H indices of non-stationary series D and T were then compared to H indices of differenced series DD and DT.

Table 4.4 shows the H indices of the non-stationary series D and T and the stationary series DD and DT.

Table 4.4: H values for Stationary and Non-stationary series.

Data/Method	**Non stationary Data**	**Differenced Data**
D(Higuchi)	0.993	0.604
D(Agg-Var)	0.9	0.142
T(Higuchi)	0.987	0.488
T(Agg-Var)	0.621	0.571

It was noted that the original data sets have exaggerated H values, i.e values very close to unity denoting highly predictable series. Differenced series have a lower H index than the originally non-stationary series. Stationarity therefore does affect the H index estimates. Values of H closer to one for series that are non-stationary does not imply very high predictability. If the series were stationary and their H indices approached one, that is the only time that we may conclude that the entire data set is highly predictable.

There is also a big difference between the H estimates of Dow-Jones original series D and the differenced series DD. We see from the fig 4.9 that the data has some spikes

and this may have contributed to the big drop of the index as we shall see in the next section.

4.5 Spikes on Hurst exponent.

The presence of large inhomogeneities in the series, such as large-magnitude abrupt changes in the series variable (jumps or spikes), and their magnitude may lead to spurious results in detecting Hurst exponents. Sometimes extremely high prices price spikes occur in the time series. There is no clear (fixed or changing) threshold to differentiate between normal prices or price spikes in terms of their intensity. As illustrated in the methodology, a simulation was performed by adding jumps to an earlier simulated series. Several jumps of random magnitude and sign (up or down) were added at random locations in series and the Hurst exponent of the resulting series calculated and compared to the original H index. The new set of data with spikes is shown in fig 4.11.

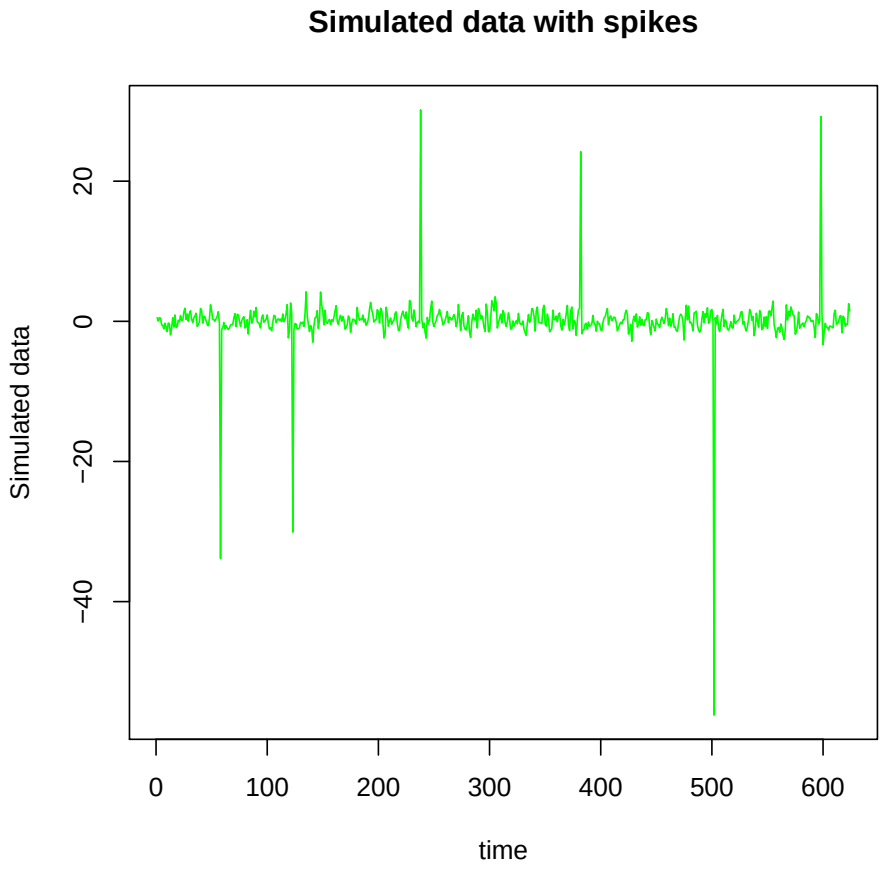

Figure 4.11: Graph of Simulatad data with spikes

Table 4.5 compares the H index for a series with no spikes with a series which has spikes.

Table 4.5: Hurst Exponent on Spikes.

x(Higuchi)	**x(Agg-Var)**	**x1(Higuchi)**	**x1(Agg-Var)**
0.756	0.421	0.479	0.226

The results of the H index estimation reveal that the data with spikes ($x1$) has lower H index compared to the original data x that had no spikes. On this perspective, these results disapprove a common saying that: Price spikes are inherent in the price process; they have the same behavior as average-level prices. Therefore "there are no spikes in

real sense", **?**. The Aggregated variance method of estimating H is greatly affected by spikes and the previous predictable data suddenly becomes unpredictable when the spikes are introduced. Our results suggest that many of the published results of such series with spikes, in order to be meaningful, need to be revised and substantiated by additional evidences of predictability before conclusions are drawn. Presence of spikes in the series decrease the estimated value of H and hence the predictability of data.

We also suggest that before testing for predictability, limit exceeding price spikes should be filtered by may be replacing them with an (average) value from the same series, or with a value from a special formula, but this is subject to further research.

4.6 Forecasting

Finally, we tried to check whether the accuracy of forecasts in a time series is dependent on our tool for measuring predictability (H index). This was done by selecting two sets of data, one with a H index greater than 0.5, meaning that it has higher predictability, and the other with an index less than 0.5, implying lower predictability. Two data sets x and T were used. The software R offers the function predict( ), which is a generic function for predictions from various models. In order to use predict( ), one has to save a "fit" of the model of the series. An appropriate ARIMA model was therefore first fitted on the data sets selected for forecasting. The simulated data x with a H index of $(0.756, 0.421)$ and the Tea prices data T with a H index of $(0.987, 0.621)$ were used. Earlier it was noted that Higuchi method over-estimates the H indices compared to the Aggregated Variance method. With that in mind, then the simulated data x has a lower predictability. For the Tea prices data T, since both estimates are above 0.5, the data is taken to have higher predictability.

4.6.1 Selection of ARIMA (p,d,q) Model.

A seasonal ARIMA model is classified as an ARIMA $(p, d, q) \times (P, D, Q)$ where P is the order of auto regressive terms, Q the order of the moving average part, and D is the number of seasonal differences. In identifying the model, the first step is to determine whether or not a seasonal difference is needed. If the seasonal pattern is both strong and stable over time, (e.g the tea prices are high during the first and the last quarter of the year and declines at the second and third quarter of the year), then a seasonal difference must be used so as to prevent the seasonality component from disappearing in future forecasts.

Using the iterative technique in the $GRETL$ software, several models on data x were tried in search for the model that gave the lowest values of AIC, SBC and the log likelihood. The model that best suited data x was ARIMA $(1, 0, 2)$. The table 4.6 below shows the models tested for the Simulated data.

Table 4.6: Best model for data x.

ARIMA(p, d, q)(P, D, Q)	**AIC**	**SBC**	**Log likelihood**
$(0, 0, 1)(0, 1, 1)$	1793.418	1811.079	-892.7092
$(1, 0, 0)(1, 1, 1)$	1828.292	1850.368	-909.1462
$(1, 0, 1)(1, 0, 1)$	1779.777	1806.384	-883.8886
$(1, 0, 2)(0, 0, 0)$	1777.685	1799.858	-883.8426

The same model fitting was done for the Tea prices data T, and the table 4.7 shows that the model that had the lowest values of the criterion was ARIMA$(0, 1, 0)(1, 1, 0)$.

Table 4.7: Best model for data T.

ARIMA(p, d, q)(P, D, Q)s	**AIC**	**SBC**	**Log likelihood**
$(1, 1, 1)(1, 1, 0)$	983.8109	997.1751	-486.9055
$(1, 1, 0)(0, 1, 0)$	1015.569	1023.587	-504.784
$(0, 1, 0)(1, 1, 0)$	980.638	988.6565	-487.319
$(0, 1, 1)(1, 1, 0)$	982.3743	993.0657	-487.187

4.6.2 Estimation of the parameters of the model.

Parameter estimation involves using very complicated computation algorithms to arrive at coefficients which best fits the selected $ARIMA$ model. Most common methods use maximum likelihood estimation or non-linear least square estimation. Due to its complication, a high quality software program $GRETL$ was used. Estimation of parameters and their SE for both models was done with GRETL software and the results are shown below :

Model 1: ARMA, using observations 1950:02–2001:12 ($T = 623$)

Dependent variable: u

Standard errors based on Hessian

	Coefficient	Std. Error	z	p-value
const	0.103785	0.0536027	1.9362	0.0528
ϕ_1	0.0995681	0.420807	0.2366	0.8130
θ_1	0.328501	0.418013	0.7859	0.4319
θ_2	−0.122737	0.191515	−0.6409	0.5216

Mean dependent var	0.104114	S.D. dependent var	1.091624
Mean of innovations	0.000057	S.D. of innovations	0.999546
Log-likelihood	−883.8426	Akaike criterion	1777.685
Schwarz criterion	1799.858	Hannan–Quinn	1786.302

			Real	Imaginary	Modulus	Frequency
AR						
	Root	1	10.0434	0.0000	10.0434	0.0000
MA						
	Root	1	−1.8143	0.0000	1.8143	0.5000
	Root	2	4.4908	0.0000	4.4908	0.0000

Model 1: ARIMA, using observations 2003:02–2011:12 ($T = 107$)

Dependent variable: $(1 - L)(1 - L^s)$TEAPRICES

Standard errors based on Hessian

	Coefficient	Std. Error	z	p-value
const	−0.337443	1.48117	−0.2278	0.8198
Φ_1	−0.541332	0.0790169	−6.8508	0.0000

Mean dependent var	−0.264019	S.D. dependent var	27.29737
Mean of innovations	0.075889	S.D. of innovations	22.55596
Log-likelihood	−487.3190	Akaike criterion	980.6380
Schwarz criterion	988.6565	Hannan–Quinn	983.8886

			Real	Imaginary	Modulus	Frequency
AR (seasonal)						
	Root	1	−1.8473	0.0000	1.8473	0.5000

4.6.3 Forecasting

Forecasting is a common goal in any analysis. The main goal in this study is to show that a time series with the H index closer to unity gives better forecasts than the series whose index is below 0.5. To verify the forecast accuracy, the percentage standard errors of the predicted values of two sets of data with different levels of predictability

data x and data T were used. In both data sets, h was set at $h = 12$ i.e 12 points ahead forecast. The last 12 values in each data set, and the "forecast " package in R software were used. The standard errors of the forecasts of the two series x and T were studied to check whether the data with a higher H index (data T) gives better forecasts than the data with lower H index (data x).

The actual values, predicted values, standard errors and percentage standard errors of data x are shown in Table 4.8 below.

	pred	se
1	-0.3152124	0.997276
2	0.1503517	1.087460
3	0.1076985	1.089951
4	0.1021897	1.089992
5	0.1014782	1.089993
6	0.1013863	1.089993
7	0.1013744	1.089993
8	0.1013729	1.089993
9	0.1013727	1.089993
10	0.1013727	1.089993
11	0.1013727	1.089993
12	0.1013727	1.089993

Table 4.8: Forecasts for Data x and their S.E.

Time	**Actual x values**	**Predicted x values**	**S.E**	**% S.E**
1	-0.899516	-0.3152124	0.997276	110.9
2	-0.86258	0.1503517	1.087460	126.1
3	0.708886	0.1076985	1.089951	153.7
4	0.827718	0.1021897	1.089992	131.6
5	-0.637081	0.1014782	1.089993	171.1
6	0.725484	0.1013863	1.089993	150.2
7	0.471817	0.1013744	1.089993	231.0
8	-0.712247	0.1013729	1.089993	153.0
9	-0.387347	0.1013727	1.089993	281.4
10	-0.470479	0.1013727	1.089993	231.7
11	0.529577	0.1013727	1.089993	205.8
12	0.461480	0.1013727	1.089993	236.2

The real data of Tea prices T was also predicted and the results of the actual, predicted, Standard Errors, and percentage Standard errors are shown in Table 4.9

```
       pred       se
1  348.1153 19.74900
2  340.0399 29.09391
3  332.5844 35.14983
4  325.7010 39.58706
5  319.3460 43.00937
6  313.4788 45.72472
7  308.0619 47.91792
8  303.0608 49.71103
9  298.4435 51.18989
10 294.1806 52.41751
11 290.2449 53.44166
12 286.6113 54.29938
```

Table 4.9: Forecasts for Tea Prices and their S.E.

Time	**Actual values**	**Predicted values**	**S.E**	**% S.E**
Jan 2011	368.10	348.1153	19.74900	5.365
Feb 2011	349.75	340.0399	29.09391	8.318
Mar 2011	330.65	332.5844	35.14983	10.63
Apr 2011	325.33	325.7010	39.58706	12.17
May 2011	327.68	319.3460	43.00937	13.13
Jun 2011	334.05	313.4788	45.72472	13.69
Jul 2011	356.14	308.0619	47.91792	13.45
Aug 2011	358.17	303.0608	49.71103	13.88
Sept 2011	362.27	298.4435	51.18989	14.13
Oct 2011	357.14	294.1806	52.41751	14.68
Nov 2011	351.27	290.2449	53.44166	15.21
Dec 2011	333.36	286.6113	54.29938	16.29

From the table 4.8, the predicted values on Data x deviate too much from the actual values. This deviation depends on the H index of the entire series which was below 0.5. If the H index was closer to 1.000, we would expect the deviation to be smaller. But if H lies between 0.5 and 0, the deviation explodes to very high values. One of the estimation methods (Aggregated Variance) gave a H index of 0.421, denoting low predictability. The $S.E$ values confirms that our time series is in deed not predictable.

For the series T whose H index was above 0.5 in both estimation methods, $(0.987, 0.621)$, it was found that the forecasts were far better compared to the forecasts of x. We may confidently say that series with Hurst exponents values close to one can be predicted more accurately than those with Hurst Exponents values close to 0.5. This implies that some series have stronger trend structure than others and this trend can be detected by estimating the Hurst Exponent. This suggests that there is need to subject data to predictability tests before we begin to forecast. Since the Hurst exponent provides a measure for predictability, we can use this value to give guidance on data selection before forecasting. Time series with large Hurst exponents can be identified before a model for prediction is build. This will save time and effort and lead to better forecasting.

Chapter 5

Conclusion and Recommendation

All the time series studied in this project shows that Hurst exponent provides a measure of predictability of time series, this index can therefore be used as a guide for data selection while analyzing trend and deciding a trade. Hurst exponent can also be interpreted as a measure of mood on the market. The lower the exponent the more nervous the investors are. At best H exponent tells us that there exists long memory in a process. H does not provide the local information needed for forecasting, nor can it provide much of a tool for estimating periods that are less random, it only classifies data according to the strength of predictability. It is every investor's wish to be able to know the outcome of any investment decision one is about to make. This is usually impossible as the outcome is bound to be affected by several aspects surrounding the market environment. In this study, this wish as far as forecasting of a series is concerned is made almost tenable. The Hurst exponent can be used to classify financial time series and guide the investor on series that may be predicted more accurately than others. The investor is also made aware of some aspects that might influence the results of the H index estimation e.g spikes, stationarity and the estimation method used. These are some of the features that make H index swing from persistence to anti-persistent

behavior. However, like any other method the predictability of a series H does not guarantee perfect forecasts, as much as we state that some data sets are predictable the forecasts still have some degree of errors. The logic is that if returns were perfectly predictable, many investors would use them to generate unlimited profits and avoid losses and there would exist a "money-machine" producing unlimited wealth which cannot occur in a stable economy. Every one with a new prediction method wants to try it out on returns from speculative assets such as stock market prices. Papers continue to appear attempting to forecast stock returns usually with very little success. By highlighting on the features that affect Hurst Exponent estimation, this project makes a contribution to this field of time series forecasting by warning the researchers who come up with new prediction models against using them before testing the time series for predictability. If this aspect is ignored prediction methods are bound to fail.

In this study we used the time series values themselves. In an applied sense we may recommend that we have some transform of the series and test whether this may increase or decrease the amount of predictive information. We could create a portfolio consisting of stocks with particular Hurst Exponent values and investigate their profit generating characteristics. If a particular stock has its Hurst Exponent drop below a threshold value, all investment positions in this asset could be closed.

Bibliography