Time-series-based Financial Analysis led us down a blind alley – could Big Data Analysis repeat the same mistake?

GDI Barr[1], TJ Stewart[2] & BS Kantor[3]

Abstract

This paper considers the new thrust of Statistical Analysis and Operations Research in the area of so-called Big Data. It considers the general underlying principles of good statistical modelling, particularly from the perspective of Pidd (2009), and how some initiatives in the Big Data area may not have applied these principles correctly. In particular, it notes that demonstrations of applicable techniques, which are purported to be appropriate for Big Data, frequently use data sets of stock market share prices and derivatives because of the huge quantum of high frequency share price data which is available. The paper goes on to critique the frequent use of such historical stock market price data to forecast stock market prices using time series analysis and details the limitations of such practice, suggesting that a large volume of work done towards this end by statisticians and financial analysts should be treated with circumspection. It is seen as unfortunate that a large contingent of extremely able students are directed into areas that encourage time series modelling of stock market data with the promise of forecasting what is essentially unforecastable. The paper also considers which approaches may be appropriately applied to model and understand the process of share price determination, and discusses the contributions of the Nobel-prize winning economists Fama and Shiller.
The paper then concludes by suggesting that the Big Data initiative should be treated with some caution and further echoes the sentiments of Pidd; namely that the focus in Operations Research and Statistics should remain firmly on creative modelling, rather than on the singular pursuit of large amounts of data.

1 Introduction

The statistical analysis of very large data sets, so-called Big Data (Pidd 2017) or Data Analytics has become a new thrust in Statistical Analysis and Operations Research. In some contexts this is easy to understand. If retailers, particularly online retailers, knew every purchase that each consumer had made over their lifetime, they could target those consumers with offers and advertising far more effectively than through broad based advertising. If banks knew every detail of their client's spending and income patterns over their lifetimes, they would be in a vastly better position to determine whether a client was credit worthy compared to the bank's current position, where they may well know only a small number of their client's transactions. As such, the proper analysis of such Big Data would be an invaluable aid to their decision making and management planning.

A number of Big Data analyses have, however, not necessarily focused on this kind of data set, which is so obviously valuable. A much more common approach used by statisticians and data analysts has been to apply time-series based statistical analysis, often under the descriptive envelope of "neural networks" and "data mining", to stock market share price and derivatives data sets, sometimes in association with past accounting figures such as earnings and cash flows. Such data is then used to

[1] *Professor at University of Cape Town (UCT) in the Departments of Economics and Statistical Sciences. Consultant over a period of over 30 years to the Financial Securities Industry, to, inter alia, Investec Securities, Old Mutual Asset Management, Simpson Frankel, and Mathison and Hollidge stockbrokers.*
[2] *Emeritus Professor in the Department of Statistical Sciences and Senior Research Scholar, UCT; Part-time Professor of Decision Science, Alliance Manchester Business School; Editor-in-Chief, Journal of Multi-Criteria Decision Analysis*
[3] *Emeritus Professor in the department of Economics, UCT and formerly Dean of Commerce at UCT. Currently Chief Investment strategist at Investec Securities.*

estimate models and capture patterns of past behaviour which can then be used to extrapolate the purported momentum of these magnitudes into the future. In the case of share price data, observations are often made at high frequency, say daily data, and thus even for middle-sized stock markets, the quantum of data available is absolutely huge. In the case of stock exchange data, it can also be argued, that such data has the additional advantage of being sampled without sampling error, as each data point constitutes an actual recorded price transaction, a rarity in the statistical data space. Following on from this, a number of recent expositions of how advances have been made in the statistical analysis of Big Data have relied on high frequency share price data to try and expound these advantages, resulting in highly complex models of questionable usefulness (see, for example, Sornette & von der Becke (2011), as well as Pidd (2009)).

The notion that share prices or even past company accounting data form some sort of repeatable pattern over time, and that complex time series or neural network techniques can be then be used to forecast such magnitudes into the future is, we will argue, based on a spurious argument that is hard to justify. An insidious feature of the analysis and modelling of time-series based stock market prices, in particular, is that the glamour of the financial sector has attracted large numbers of highly competent university students, keen to discover how sophisticated mathematical models, combined with high powered computers, can somehow unravel the true underlying time-series behaviour of share prices. We will argue that this has been an unfortunate waste of mis-directed top talent and that the typical Mathematics of Finance course often misleads and misconstrues the appropriate application of statistical techniques to such data.
Moreover, although the huge amount of work in the area has led to a large number of projects and theses, a large number of these students are ultimately employed in the financial industry and thus the work becomes proprietary and hence is rarely published, and thus is not open to objective scientific scrutiny.

2 The Modelling of real world phenomena using mathematical models

We need to first consider more generally what role the mathematical modelling of quantitative data has in the support of decision making and management planning. Such modelling necessarily needs to factor in the impact of human behaviour on final outcomes, as opposed to modelling in the pure sciences (such as cosmology) for the purposes of understanding the world, the universe or nature. This context is well developed and described by Pidd (2009). In particular, he summarizes six principles which are relevant to the theme of this paper.

1. Model simple, think complicated
2. Be parsimonious, start small and add
3. Divide and conquer, avoid mega models
4. Use metaphors, analogies and similarities
5. Do not fall in love with data
6. Model building may feel like muddling through.

In thinking about modelling in our context, we must recognize three key components to any model construction, namely (i) the incorporation of human cognitive understanding and experience of the underlying systems, (ii) the use of data to validate emerging models, and (iii) the role of mathematics to ensure internal coherence and logic.

Any model development which skimps on any one of these three components must at very least be viewed with scepticism as a tool for planning or decision making. It is perhaps broadly accepted that models lacking substantial data validation are inadequate for quantitative prediction and planning. Ironically, though, it is perhaps less widely recognized that "data" can be soft (e.g. community experiences) as well as hard (statistical). A focus on hard data only may be viewed as "scientific" in some quarters, but may seriously overlook important processes. One is reminded of the maxim of Forrester that omitting structures or variables known to be important because numerical data are unavailable is actually less scientific and less accurate than using your best judgment to estimate their values. ''To omit such variables is equivalent to saying they have zero effect - probably **the** only value that is known to be wrong!'' (Forrester 1961), elaborated by J D Sterman (2002).

In the modern world of "big data" and "machine learning" there is a growing tendency to base models purely on quantitative data, ignoring soft data, understanding and logical coherence. In presenting his principle 5, Pidd states that the "model should drive the data and not vice versa". In a similar vein, we caution against data driven models that have not been validated by careful reflection, consistency with soft data and logical mathematical coherency.

Simply because ubiquitous data are available does not make the data relevant to planning and decision problems at hand. We will elaborate below on how financial data in particular has become seductive to data analysts without proper regard to the role of human behaviour in the determination of the data Systems structures involving human behaviour are highly dynamic, so that it is difficult to confirm that patterns in historical data are relevant to future behaviour. The onus is on the modeller to provide evidence of relevance.

Similar caveats apply to an overemphasis on sophisticated mathematical structures, in violation of Pidd's principles 1 and 2. Adoption of overly complicated mathematical structures unmotivated by the specific needs of system understanding and of the hard and soft data, hinders validation of the model in terms of these other components. Simply because, for example, a stochastic calculus model of a diffusion process invokes deeper axioms and mathematical proofs than a discrete time-finite state model, does not make the former a better model of the underlying process which is inevitably more complex than any model! In a context where one is modelling the results of human behaviour and human interaction, models are necessarily an approximation. In such a context all models are wrong, but which are more useful?

3 The Stock market as a source of Big Data

One of the most ready and plentiful sources of time-series data is that obtained from a stock market, primarily high-frequency share or derivative price data, a fact that has made it particularly seductive for Statisticians and Data Analysts. Unfortunately this super-plentiful source of data has led to multiple analyses that have ignored the principles of good modelling outlined above. To properly contextualise and understand the nature of stock exchange data, we first discuss the process of share price (or derivative price) determination and precisely what this price measures.

A share's price is the last (transacted) price which a buyer and seller agree to and thus represents the perceived value of that share at the transaction time point. Those traders who sell generally think prices will fall (share price was over-valued in their opinion) and those who buy generally believe prices will rise (share price was undervalued in their opinion), so the fact that transactions occur reflect the fact that a divergence of trader opinions exists. It is thus important to realise that the price determination process is a process which involves perceptions of value by the two transacting parties, rather than any objectively determined and unverifiable, true value.. There is thus an element of game theory in the price determination process, since each participant is primarily focused on the perceptions of the other market participants, rather than the more obvious task of simply trying to estimate a true value. This makes price determination a complex game-theoretic task which involves assessing the perceptions of all other market participants.

3.1 Plentiful Finance data attracts

The falling cost of computer storage space and increasing computer power has led to large quantities of share price data being available at low cost. For liquid shares, there are often several transactions per minute. Thus the Johannesburg Stock Exchange (JSE), for example, with its several hundred shares can collect many millions of data points, recorded at high frequency, every day. Moreover, this data is not an estimate of the share price, but an actual recorded true value at the last recorded time point. This large quantity of available data, not subject to sampling error, has always been a seductive draw for statisticians and data scientists. It is conjectured, in fact, that the reason that such a large quantum of research effort has been directed into portfolio analysis using share price and derivative price data is simply because the data was available.

The advent of analyses that draw statistical advantage from the new Big Data initiative has become a key component of modern statistics. Big Data analyses, which are often described under the broad heading of data mining or neural networks; thus see plentiful stock market data as an obvious testing ground. The actual mode of analysis is not particularly new, however, and both data mining techniques and neural network techniques applied to time-series data draw heavily on Box-Jenkins type time-series modelling which was developed in the late nineteen sixties (Box & Jenkins (1970)). For the most part, these analyses simply attempt to perpetuate past data patterns and behaviour into the future; for the stock market price case see, for example, Hajizadeh *et al* (2010), Hazam *et al* (2010) and Leigh (2002).

So is there validity in a momentum approach to modelling stock exchange data? We consider this issue in some detail, first considering the behaviour of the Google share price over the last 10 or so years.

3.2 The example of Google – can we model anything about a share price?

Google listed in August 19, 2004, at $85 a share and the return, exactly 12 years later on August 19th, 2016, represented a gain of 1 780%. Many tech stocks have outperformed Google over the same time period; Amazon achieved a 1 827% return, Netflix a 4 181%. An analyst considering Google in 2004, might have concluded that such was the quality of management and the promise of the technology that it represented a fantastic investment. Others might have concluded it was not worth buying. Analysts on the right side of this “valuation equation” were extremely well rewarded, or if they were on the wrong side, lived the future with extreme regret. The process did not end there. If one considers the past record of Google’s price rise, then at almost all points in the price path, Google has performed better than market expectations. If one looks at its extraordinary historic price profile one might conclude that quite obviously it continues to be a sure-fire winner as its price simply goes higher and higher. This is clearly not the case. The current position is no different from the position faced by investors at any point in the history of the share. At each point in time there has been an interaction between buyers and sellers who have priced Google at the price which gave an expected return commensurate with its perceived risk. Those who did that calculation and thought it was worth more than the current price bought, and those who thought it was worth less, sold. The fact that it has consistently done better over the last 12 years than participants as a whole expected it to, is simply unusual. Will it continue? Who knows, but those who think it will are quite able to go out and buy Google shares from someone who thinks the opposite.

The fact remains that those analysts - were they clever, or perceptive, or lucky? - who bought Google in August 2004, because they believed the business would achieve great things, were proved astonishingly correct.

3.3 Fundamental Analysis versus Time-Series Analysis of stock market data

So-called fundamental analysts of listed companies spend their time trying to work out, (given any information they can glean about the companies’ management, employees, prospects etc.) the value of a company, and the associated flow of earnings or cash flows that a company generates. Such fundamental analysis done by these “stock market analysts”, who are generally assigned to research particular stock market sectors, constitutes the backbone of research done by stockbroking companies. The standard approach typically estimates a company’s future cash flows and computes the present value of the stream of these cash flows; the so-called discounted cash flow (DCF) approach. Such an approach attracts its own set of problems, not only the difficulty of estimating future cash flows but also the sensitivity of the present value to the discounting factor which measures the return required by the market. Moreover, such an approach necessarily has to make the non-intuitive assumption that the life of a company, and hence the set of assets that constitutes the company, is finite. Such fundamental analysis plays a big part in share price behaviour, with analysts issuing ‘Buy”, “Sell” or “Hold” recommendations to their clients, which quickly become part of the body of public information.

Fundamental analysis often extends to an historical (and hence time-series based) consideration of the price-earnings (PE) ratio to assess whether the current ratio is above or below historical norms. Such analysis has, itself, been the subject of debate, particular around the extent that earnings used in the computation of PE should be smoothed over time (see the work of Nobel prize winner Robert Shiller (Campbell & Shiller (1998))) This time series based approach, given prominence by Shiller, has, itself, been the subject of much criticism, notably from Jeremy Siegel (2011, 2012); also see the critique of Shiller’s work by Kantor and Holdsworth (2014). We will argue below that although earnings constitute an observable reality and may be used in a smoothed or unsmoothed form, any attempt to forecast prices based on only past information faces a severe logical challenge.
Perhaps the most insidious form of such time-based share price analysis has been the use of so-called technical analysis. This method, adopted by “chartists”, assumes that share prices repeat particular picture-like patterns over time. These patterns include and are termed, *inter alia*, “Head and Shoulders”, “Inverse Head and Shoulders”, “Cup and Handle”, “Double Tops and Bottoms” and “Triangles”. They represent a primitive but pictorial univariate time-series approach to share price forecasting, the argument being that such patterns in share price are repeated over time. Chartists might argue, for example, that a market mood switch from negative to positive, or vice versa, might consistently result in a repeatable price pattern. One issue is, of course, that if any market participant believed that chartists constitute a significant portion of the market, then such patterns become important as they could be perceived as self-fulfilling and thus represent useful forecasts of the future.

In modelling terms, a chartist approach constitutes a rather unsophisticated time-series approach and has little relation to the time-series modelling of share-price behaviour that Mathematical Statisticians and Data Analysts have generally been involved in. This modelling approach typically takes a price series, such as the price of Google shares over the last 12 years, and fits a complex stochastic model to the price behaviour. These models, as might be expected, often fit the past data extremely well but this is clearly not necessarily helpful to those wanting to know whether it is a good idea to buy or sell Google shares now. The models are typically used to predict the future Google price; this generally involves estimating a time-series-based model on the basis of past data and determining whether the current model prediction is either lower than the current price (indicating the current price is too high) or higher than the current price (indicating the current price is too low). In some cases such models are used to explicitly predict the share price at some arbitrary time in the future.

We will discuss below that there are strong reasons to believe that any models that use past time-series data of share prices, or accounting magnitudes, hold limited value for forecasting future values.

4 Time Series models applied to stock market data

Let us take a step back and understand what time-series models of a share price are doing. They are fitting a structured mathematical process to the share price generating mechanism; that is, they are determining some mathematical pattern to the price series over time. In the univariate case no other data is used in the process; it is the pattern of the price and the price only which is assumed to be perpetuated across time. If one considers again the 12 years of Google data and assuming this data was used in this analysis, this would imply that at the beginning of this period (August 2004) the future price behaviour could be closely determined by this model. Although one can, of course, easily fit a mathematical model to any given time series, the idea that some model which relies on the price and the price alone could be used to produce future prices seems strange and unrealistic. It can only hinge on the notion that, in some way, prices have "momentum", rather than being determined by continuously changing and unpredictable perceptions of value. This notion was comprehensively tested by Fama & French (1993) who concluded that a momentum factor should not be included in their three factor model as few portfolios had statistically significant loading on it. This conclusion of no significant momentum factor was reiterated in the Fama & French (2015) paper. Note that it is perceptions of value which are central to the process of price determination - if a group of market participants decide that the value of a share should be higher than the current price, they tend to drive the price up, independently of whether there is any veracity in this perception. No one knows the true value and the market is a cauldron of differing views trying to determine the value, under a myriad of continually changing circumstances. Hence perceptions are a key element, and what becomes important is not what is true but what one thinks other people think is true. Hence if unrealistic perceptions swing to the positive and drive the price up, the seemingly rational investor might recognise this and simply sell at the high unrealistic price. Market participants are thus continuously applying short-term game theoretic models to each other, often more interested in the short-term effects of sentiment, rather than any harder, more fundamental take on actual value. The famous economist John Maynard Keynes spoke of the effect of "animal spirits" (Keynes 1936, p.161) determining value on the stock market when describing the run-up in prices and then the subsequent stock market crash of 1929. Under such circumstances, the role of emotion, perceptions and what the expected future pattern of emotions becomes critical, especially in times of crisis. If we get the twin-towers-event of September 11th, 2001, issues of value are thrown out the window – when prices drop sharply under those circumstances of extreme uncertainty, a trader might see what appears to be real value, but is most concerned about making the mistake of buying in too early; alternatively a trader might short-sell if he believes prices may fall further. The key point is that fundamental value becomes irrelevant under emotionally charged circumstances as participants struggle to assess the degree of uncertainty and the effect of increased risk. We now consider which mathematical models are available for tracking the path of share prices over time and how they model this uncertainty.

4.1 The Random Walk model

The most widely accepted "model" (often referred to as "no-model") of share price behaviour is the Random Walk model, usually expressed in Log form to account for the proportionality effect

$$Log(P_t) = \theta + Log(P_{t-1}) + a_t ,$$

where θ > 0, is a constant and a_t is a white noise term.

Rewriting gives

$$\%_{cc} P_t = \theta + a_t ,$$

where $\%_{cc}$ denotes continuously compounded percentage changes.

This states that continuously compounded rates of change in share price, usually referred to as return comprise 2 components, θ, a constant, often referred to as drift term, representing the expected return from the share in question and a white noise term. The θ receives an interpretation from the so-called Capital Asset Pricing Model[4] which breaks this expected return into 2 further components,

[4] **CAPM** *was introduced by Jack Treynor (1961, 1962), William F. Sharpe (1964), John Lintner (1965) and Jan Mossin (1966) independently, building on the earlier work of Harry Markowitz on diversification and modern portfolio theory. Note that ever more complex versions of this model have been formulated which break the expected return into an increasing number of factors; see, for example, Fama, EF & French KR (2015).*

$\theta = r_f + \beta\, E(r_m - r_f)$.

That is, investors can expect a return from a share equal to the risk free rate (r_f) and a component to compensate them for investing in the market (r_m) as a whole, which necessarily has inherent risk. This compensation for risk component comprises a share specific risk factor (β) multiplied by a general market risk factor, calculated as the excess of the expected general market return to the risk free rate. If β is equal to one, the share mimics general market risk, if $\beta > 1$, that is the share is riskier than the market, and if $\beta < 1$ the share is less risky than the market.

These equations might appear technically complex, but they reflect a simple truth. The change in share price is some constant plus a random white noise term. That, is the percentage change in a share price from one time period to another, where this time period might be a day, a week, a month or even a minute is best described by a constant and a white noise term. That is, the return over any period is expected to be positive and is randomly distributed around this expected constant. The driver behind this change is clearly any packet of information that arrives in the intervening time period in question, which might have some bearing on the price. We do, of course, assume that participants in the market are rational and that the share price will incorporate new information into the share price, as it arrives, so-called market efficiency (see, for example, the work by Nobel Prize winner Eugene Fama (Fama 1998))
One could then write,

$$\%_{cc} P_t = \textit{information arriving between } t\text{ - }1 \textit{ and } t\,.$$

If the participants are rational and information is efficiently incorporated into the share price, then these packets of new information can be assumed to be unexpected and randomly distributed. If there was any pattern in this information arrival or if this new information could be predicted, it would, of course, have already been incorporated into the share price.

These assumptions are generally accepted by analysts and market participants. They infer that the best predictor of the next period's share price (apart from that constant expected amount, θ) is the current period's share price. This rather unexciting truth infers that any attempt at predicting share prices could be rather fruitless. It requires one qualification. As discussed above, prices of shares do not arrive from thin air, they represent a consensus of fundamental analysts' and other market participants best stab at what a company is worth, taking into account all the company specific factors and macro-economic factors which they assume to be relevant. Fundamental analysts, as mentioned above, may use their estimates of earnings to assess the worth of a share or compute deviations of current PE from long-term, supposedly equilibrium-based, PEs. If any participant decides a share is worth more than the current price he can buy shares, if less than the current price he can sell what he has. The interaction of different analyst's (and other market participants) views' gives rise to some equilibrium price. But then, rationally, there must be a reward attached to this activity of share value determination. Hence fundamental analysts who are good, will get it right more of the time than those who are bad, and make money out of it.

4.2 Forecasting share prices

Why does anybody think share prices can be forecasted? To some extent it may be that share prices have an anthropomorphic character over time - the path of some animate creature - and that this is what market participants and analysts find so seducing and compelling. If one is given a chart of a steadily rising price, most will forecast that it will continue to rise. But the arguments laid out above, under conditions of market efficiency where prices simply reflect the current information flow, and the set of expectations of market participants regarding all future information flows, necessarily preclude this. Only in the case of "insider information" where a participant had company information which others did not have and illegally acts on it, could a participant predict a price movement. One excludes the scenario where a powerful player could hypothetically "corner" the market and manipulate the price for their own gain. Such attempts have consistently ended in failure, the most notable example was the attempt by brothers Bunker and Herbert Hunt to corner the silver market in 1979/80; They purchased over 43 million ounces of silver contracts on U.S. exchanges, the result being that they drove the price of silver up from US$6 an ounce in early 1979 to just over US$50 an ounce – its highest price ever – in January 1980, But they failed to pay a huge margin call of US$135 million in March 26 1980 when the price started to drop, and were forced into bankruptcy, Excluding these cases, and as discussed above, today's price can be assumed to simply reflects all available information and future expected information – the Random Walk model discussed above, a simple formulation which implies that "The best predictor of tomorrow's price is today's price".

Can we then do anything useful with an historical time series of share prices, or company accounting data, or do they simply constitute a fairly interesting record of past behaviour?

4.3 Share price data - truths and myths

A proper contextualisation of these issues requires one to establish what financial time series data does have to offer, and what it does not offer. Share price data (usually expressed in return form) is clearly important for the computation of risk (2^{nd} moment) as well as for the computation of higher moments (at least 3^{rd} and 4^{th}) as these moments critically do have "momentum". That is, these moments are auto-correlated, have some stability over time and are thus forecastable. Hence at a time of crisis (say the 2008-2009 credit crisis), volatility (2^{nd} moment) is high and continues to be higher than normal over some extended period, reflecting the higher perceived risk over a lengthy period. The risk of specific sectors also tends to be stable and predictable-correlated over time. Thus, for example, the furniture business, which is an inherently risky part of the economy, because it is primarily credit funded, has associated shares which reflect this high risk over time. In contrast, the food sector and food shares consistently reflect lower risk over time. The inter-relationships between sectors (say the correlation between the JSE food and furniture sectors in return form) have also reflected some sort of stability across time. So in these areas the analysis and interrogation of financial data can be very useful for predicting the future. One can then structure portfolios of different shares across different sectors with desirable risk and expected return attributes, and one can expect that these characteristics will be relatively stable over the future.

Where share price time series data has a particularly poor record is in the actual forecasting of share prices themselves (again, usually expressed in return form). However, a large portion of financial research suggests that share price returns/prices ARE in some way forecastable. The Big Data initiative seems to imply that the only thing holding back the forecasting process is raw computer power and ever larger quantities of data.
This is not the case. There is no record that, in any consistent way, stock market prices are forecastable. Prices are, anyway, as we discussed at length above, only *perceptions* of value. The information set affecting perceptions of value will change continuously in a random way, and hence share prices will change continuously in a random way, reflecting the changing perceptions and expectations of all future scenarios for perpetuity.
Importantly, note that this does not preclude rewarding the skilled "fundamental" analyst. If analyst X analyses company Y and concludes that other analysts have misconstrued the value of the share and it should, for example, be more highly valued and have a higher price, then X can always buy the share and hope that at some point in the future the share price will rise, reflecting his valuation. It is, of course, teams of analysts who strongly influence share price and compete away any obvious deviations from true value. A caveat to the notion that price reflects only perceived value is that

"price" is always taken as the last *recorded* transaction, that is the last time that a buyer and seller concluded a contract. In the case of relatively "thinly traded" shares, it is clear that the "current" price may be a price recorded some time previously.

A further caveat is that forecasting models of share price data often appear to support the conclusion that they would produce good forecasts. The model is applied to past data and often produces a good model fit; it is then demonstrated that if the estimated model was applied to a past set of data it would have turned a profit. However, the only way to properly test the efficacy of any model is to use a portion of data which has not been used in the estimation of that model. In the case of time-series price data one should rigorously test the model "out-of-period"; that is, test the model over a period of data which was not used to produce the model in the first place. In the case of share price models, their good record for within period data, will almost always fall away when properly tested out-of-period.

5 Attracting the good students

In the case of Finance; any model that could consistently forecast share prices would, of course, translate into untold wealth, and if complex mathematical models (or any models at all) could provide such forecasts, one might expect a correlation between wealth and expert mathematical model builders – unfortunately (at least that's what most mathematicians tell me ☺), no such correlation appears to exist.

However, the promise of such wealth has, we believe, quite inappropriately attracted the student cream-of-the-crop into Mathematics-of-Finance courses. These students were often aspirant actuarial students, who comprise a large component of South Africa's top matriculants. The actual percentage of qualifying actuaries from this group is low and a much large proportion are directed into finance related work. This has led to an unfortunate allocation of top South African mathematical and quantitative talent pursuing the questionable task of estimating the inestimable!

5.1 Why do financial companies employ people with strong mathematical backgrounds to use complex models to model share price data if they don't work?

Managers of finance sections who make decisions around share purchases for the large portfolios they manage, often have only a rudimentary grasp of complex mathematical methods touted as being useful for modelling share price behaviour. They are thus unable to evaluate the efficacy of such models effectively and, in particular, will have little understanding of the mathematical nuances inherent in such models. They do, however, fear that such models *could* be of some use and if they do not engage at least one mathematical modelling expert to look into developing such models, their company might be left behind. If such a person was turned down for employment and went on to another firm to produce a share selection methodology which was hugely successful, it would clearly undermine the credibility of the former employer. So mathematical modelling retains some foothold on the basis of the untested promises made by its developers. The question should always be asked. If a mathematician really developed a model which consistently produced returns in excess of those on average, why would such a person not work for him/her self?

6 Concluding comments - what can the new BIG Data initiative provide for Financial Analysis?

As outlined above, models estimated from comprehensive time series data on share prices, or company accounting magnitudes, are often used as a testing ground for techniques developed for Big Data, but cannot provide useful forecasts. If banks or advertisers had detailed personal data on purchasers, income and debt repayment history, it would be extremely advantageous for targeted advertising and credit extension decisions. However, detailed data on expenditure behaviour would not, by itself, help forecast expenditure in the future in the same way that knowing past share prices does not help us forecast future share prices. The fact that high frequency and comprehensive time series data sets have become increasingly accessible and that the cost of storing this data in smaller and smaller compartments is dropping inexorably, has been the seductive driver behind the new Big Data initiative. The actual statistical analysis techniques used for such data haven't changed substantially over the years, but the power of computers and the ability to access and process huge amounts of data certainly has.

Hence the analysis of large data sets will continue to have increasing importance in the future, but only in certain specific contexts. For example, in the area of finance and economics, in which the role of human behaviour is central, time series models will be of limited help, certainly for forecasting. Time series analysis may be useful for identifying underlying models of equilibrium in population dynamics or in the physical sciences, but when the equilibrium is driven by rational humans, the usefulness is less clear. Huge, high frequency data sets may facilitate certain things. Data from the SKA may facilitate the identification of unknown planets, data from satellites may facilitate ever more accurate meteorological models, data on consumer purchases may help Amazon inc. bombard potential buyers with increasingly effective advertising. But, as has always been the case for much smaller data sets, Big Data will not help us <u>forecast</u> phenomena controlled by humans competing for limited resources. Maybe Pidd (2017) should have the final word on Big Data; he concludes that we should " … not get carried away by big data" and realise that, rather than Big Data adding value, *per se*, " … people add value by creating models that use it."

Bibliography

BOX GEP & JENKINS GM, 1970, *Time Series Analysis: Forecasting and Control*, Holden Day, San Francisco.

CAMPBELL J & SHILLER R, 1998, *Valuation Ratios and the Long-Run Stock Market Outlook*, The Journal of Portfolio Management, **24(2),** pp. 11-26. In 2001, the authors produced an updated version, CAMPBELL J & SHILLER R, 2001, *Valuation Ratios and the Long-run stock Market Outlook: An update*, No. 8221, National Bureau of Economic Research.

FAMA EF, 1998, *Market efficiency, long-term returns, and behavioral finance*, Journal of Financial Economics, **49(3)**, pp. 283–306. Original paper is Fama EF, 1965, *Random walks in stock market prices*, Financial Analysts Journal, 21(5), pp. 55-59.

FAMA EF & FRENCH KR, 2015, *A Five-Factor Asset Pricing Model*, Journal of Financial Economics, 116, pp. 1–22.

FAMA EF & FRENCH KR, 1993, *Common risk factors in the returns on stocks and bonds*, Journal of Financial Economics, **33(1)**, pp. 3-56.

FORRESTER JW, 1961, *Industrial Dynamics*, MIT Press, Cambridge (MA).

HAJIZADEH E, ARDAKANI H & SHAHRABI J, 2010, *Application of data mining techniques in stock markets: A survey*, Journal of Economics and International Finance **Vol. 2(7),** pp. 109-118.

HAZEM ME & AWAD WA, 2010, *Fast Forecasting of Stock Market Prices by using New High Speed Time Delay Neural Networks*, International Journal of Computer and Information Engineering, **4(2)**, pp 138-144.

KANTOR B. & HOLDSWORTH C, 2014, *Nobel Prize revisited: Do Shiller's Models really have predictive power*, Applied Corporate Finance, **26(2)**, pp. 101-112.

KEYNES JM, 1936, *The General Theory of Employment, Interest and Money*, Macmillan, London.

KORAJCZYK R, 1999, *Asset Pricing and Portfolio Performance: Models, Strategy and Performance Metrics*, Risk Books, London.

LEIGH W, PAZ M & PURVIS R, 2002, *An analysis of a hybrid neural network and pattern recognition technique for predicting short-term increases in the NYSE composite index*, Omega., **30(2)**, pp. 69-76.

LINTNER J, 1965, *The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets*, Review of Economics and Statistics, **47(1),** pp. 13–37.

MOSSIN J, 1966, *Equilibrium in a Capital Asset Market*, Econometrica, **34**, pp. 768–783

PIDD M, 2017, *Making an Impact: in praise of small data and big thinking*, Impact, pp. 11-12.

PIDD M, 2009, *Tools for Thinking: Modelling in Management Science*, Wiley, 3rd Ed.

Sharpe WF, 1964, *Capital Asset Prices – A Theory of Market Equilibrium Under Conditions of Risk*, Journal of Finance**, XIX (3**), pp. 425–42.

SIEGEL J, 2012, *Dueling Prisms for Valuing Stocks*, New York Times, October 14, 2012.

SIEGEL J, 2011, *Shiller vs. Siegel*, New York Times, April 27, 2011.

SORNETTE D & VON DER BECKE S, *2011, Complexity clouds finance-risk models*, Nature, **471**, pp. 166.

STERMAN JD, 2002, *All models are wrong: reflections on becoming a systems scientist*, System Dynamics Review, **18(4)**, pp. 101-114

TREYNOR JL, 1961, *Market Value, Time, and Risk*, Unpublished manuscript dated 8/8/61, No. 95-209.

TREYNOR JL, 1962, *Toward a Theory of Market Value of Risky Assets*, Unpublished manuscript. Subsequently published as Chapter 2 of Korajczyk (1999).