

# A General Approach to Digital Analysis exemplified by Stock Market Indices

Peter N. Posch<sup>a,\*</sup>, Welf A. Kreiner<sup>b</sup>

First Version: June 2004. This Version: 31st May 2005- Preliminary

## Abstract

We propose a general methodology for using digit-distributions as an approach to examine arbitrary datasets. Using the Newcomb-Benford-Law as a starting point we develop a more general framework for digital analysis. We propose two measures based on this framework, namely the Digital-Fit-Factor (DFF) and the Mantissae-Distortion-Factor (MDF). Using these approaches we demonstrate the use for index comparison on the S&P 500, the Dow Jones Industrial Average and the Nikkei 225 index. To demonstrate the use of these measures we construct portfolios and measure the performance compared to the index itself. Our measures exceed the index by more than 10 percentage points per year. Furthermore these measures require only a very small proportion of the available information and are thus very efficient.

**Keywords:** Stock Market Index, Digital Analysis, Newcomb-Benford-Law

**JEL-Classification:** M49, H20, C49

<sup>a</sup> Department of Finance. University of Ulm, 89081 Ulm, Germany. Phone: +49-731-50-23596. E-Mail: peter.posch@uni-ulm.de

<sup>b</sup> Department of Chemistry. Phone: +49-731-50-22876. E-Mail: welf.kreiner@chemie.uni-ulm.de

\* Corresponding author.

# 1 Introduction

Stock Markets are considered to have a large, if not the largest, impact on the modern world economy. The information on how major economies change, parts of the national stocks are gathered in an index. Naturally, there are several methods of picking stocks for an index inclusion and the announcement of an index change often also changes the market prices of both the included and excluded stocks. The calculations of the indices themselves are reported daily in newspapers, hourly on radio and television and real-time in the internet and on news-channels. Anyone who ever listened to a conversation in a bar, in subways or gyms would not doubt that besides the weather forecast the changes in the national stock market index are worth being discussed extensively.

Although the index companies claim, that the composition of their product is based on quantitative variables such as free-float (i.e. the number of shares that are freely available to the investing public) or trading volume (i.e. number of shares traded each day), there are a lot of qualitative factors which have to be fulfilled in order to gain access to the index, e.g. quarterly balance reports or certain corporate governance rules. The index itself is seldomly calculated as the arithmetic mean of the market prices of the stocks included, but are more often subject to more complicated schemes.

In this paper, we demonstrate a general methodology for digital analysis of economic measures. The generality of this approach leads to two measures which can be applied to any set of economic variable in order to compare them within time or within economic-units such as countries, states etc.

Our measures are based on an approach known as 'Digital Analysis', which is already used to detect frauds in tax sheets and company reports. The most widely used distribution in this field is known as the 'First-Digit-Law' or 'Benford-Law'. This name ignores the original discovery of the law by Simon Newcomb, so we will refer to this phenomenon as Newcomb-Benford Law (NBL).

We will demonstrate the power of our approach by examining the Dow Jones Industrial Average Index and the Standard and Poor's Composite Index over time. The comparability on the economic-unit level is demonstrated by adding the Nikkei 225 Index.

The remainder of this paper is organized as follows. The ongoing section gives a short introduction to the Newcomb-Benford Law and then introduces the measures used. Section 3 applies these measures to stock market indices. The results given in this section are mainly descriptive and assumptions on further utility are left to the reader. Conclusions are drawn in Section 4.

## 2 The Newcomb-Benford Law

Newcomb [1881] noticed that the first pages of logarithmic tables were more worn than the later ones.<sup>1</sup> From a statistical view-point he concluded that 'the law of probability of the occurrence of numbers is such that all mantissae of their logarithms are equally probable.' While his work remained largely unnoticed, Benford [1938] discovered the same property fifty years later and studied a wide collection of data-sets. Since then, many mathematicians have tried to prove the so-called First-Digits- or Benford-Law. Some of these proofs are quite convincing, but they all lack a proper definition of the underlying probability space. This problem was solved by Hill [1996] with the definition of an appropriate  $\sigma$ -Algebra and the derivation of a probability measure. Furthermore, Hill [1996] gives a very plausible explanation with a statistical derivation of the First-Digit phenomenon.

### 2.1 Basics

Newcomb [1881] states that, for certain natural numbers, the mantissae of their logarithms are equally distributed.<sup>2</sup> From this assumption he derived his first digit law, which gives the probability for a first Digit  $d \in \{1, 2, \dots, b-1\}$ , where  $b \in \mathbb{N}, b \geq 2$  gives the number-base. Since most empirical data-sets are given in the decimal base ( $b = 10$ ), we will focus on that number-system in the remainder of this paper, but the results are valid for arbitrary bases other than 10 as well. The function  $D_k(x)$  denotes the  $k$ -th significant digit of  $x \in \mathbb{R}$ , e.g.  $D_1(3.1415) = D_1(31415) =$

---

<sup>1</sup>Before the invention of calculators these tables were used to multiply numbers by adding its logarithms.

<sup>2</sup>Note that Newcomb uses the term 'Mantissae' in the sense of logarithmic mantissae, i.e. the fractional part of the logarithm of a real number.

$D_1(0.0031415) = 3$  and  $D_4(3.1415) = D_4(31415) = D_4 = (0.0031415) = 1$ .

$$Prob(D_1(x) = d) = \log_b \left( \frac{d+1}{d} \right) \quad (1)$$

For positions exceeding the first digits this equation generalizes to

$$Prob(D_n^{(b)} = d_n) = \sum_{k=b^{n-2}}^{b^{(n-1)}-1} \log_b (1 + (k \cdot b + d_n)^{-1}) . \quad (2)$$

## 2.2 Counting

The frequency of the occurrence of first digits is obtained from a simple counting process,  $k$  digits. Using an automatic procedure for this counting process one has to account for the significance of the digits, e.g. i.e. 0.001 has a one as first significant digit. To gain uniqueness our routine uses mantissae of numbers instead, i.e. every number  $x \in \mathbb{R}$  is splitted into a part  $m \in [0, 10[$  and a part  $10^k$  for a given  $k \in \mathbb{N}$ . e.g. consider the number  $x = 100 \cdot \pi = 314.15\dots$ . We can write this number as  $\pi \cdot 10^2$  (note that  $\pi \in [1, 10[$ ). The benefit of this notation is that the decimal point is fixed, so a counting-routine based on  $m$  is unique and furthermore includes all the necessary information.

## 2.3 Digital Fit Factor (DFF)

Using non-linear methods we fit the density-function-vector of the Newcomb-Benford-Law whose entries are given by

$$Count_i \cdot N^{-1} = \frac{(Digit_i + 1)^{1-DFF_k} - Digit_i^{1-DFF_k}}{10^{1-DFF_k} - 1} \quad (3)$$

where  $N$  is the number of observations,  $Digit_i$  contains the possible digits and  $k$  is the depth of the digital analysis. For the analysis of the first significant digits this is  $Digit = 1, 2, \dots, 9$ , while an examination of on-going positions includes the zero as possible significant digit. For a detailed deviation of equation (3) see Kreiner [2003].

The density function is thus defined as  $C/N = D$  where  $C$  is the  $1 \times 10$  vector of counted values for each digit  $0, 1, 2, \dots, 9$ ,  $N$  is the constant number of observations and  $D$  is the  $1 \times 10$  vector of possible digits, i.e.  $D' = (0, 1, 2, \dots, 9)$ . Analysing the first significant digits, the first row of each vector is obsolete since 0 is not a possible first digit. In this case the vectors reduce to  $1 \times 9$  vectors. Note that both  $N$  and

*Digit* are given a priori, i.e. independently of the current counted digit frequency and that the left-hand side of equation (3) gives the percentage of first digits  $i$  in the dataset.

Using non-linear fit-procedures we obtain estimators for the Digital-Fit-Factor  $DFF_k$  of digital depth  $k$  and its variance  $VAR[DFF_k]$ . Depending on the depth of the analysis undertaken  $k$ , varies between 1, i.e. the first-significant digit, and the maximal possible length  $k_N$  given by the current data-set. In the remainder of this paper we will refer to  $DFF$  as Digital-Fit-Factor for 1st significant digits, but it is crucial to note, that the analysis described above can be performed for arbitrary digit positions  $k \geq 2$ . In this case, the  $DFF$ -Value will be denoted by  $DFF_k$ -Value. It is notable, that the information included rises with  $k$ , i.e. in the calculation of  $DFF_1$  only the first significant digit of each observation is included whereas for  $DFF_4$  the first four digits are used etc. For any  $k$  the  $DFF_k$  is greater than zero by definition.

Since the Newcomb-Benford-Law approaches the uniform-distribution with raising digit position  $k$ , e.g. for a first significant digit one the Newcomb-Benford-Probability (NBP) is about 30.1%, while the uniform distribution is at 1/9, but the NBP for a fourth significant digit one is  $\approx 10.014\%$  while the uniform probability is 1/10 (since zero is a possible fourth digit), less and less 'new' information is added in comparison to the uniform distribution. See table 1 for details. Following this observation and citing an unpublished result by K. Schuerger that almost every random variable with continuous density function will show such an asymptotic behaviour, we will not focus on  $DFF$ -Values for positions greater than one. Note that this restriction is without loss of generality, since the 'additional' information of the Newcomb-Benford-Law is greatest for  $DFF_1$  and declines afterwards.

| d | NBP for .. significant digit $d$ |              |              |              |
|---|----------------------------------|--------------|--------------|--------------|
|   | 1st                              | 2nd          | 3rd          | 4th          |
| 0 | -                                | .11967926840 | .10178436490 | .10017614540 |
| 1 | .30102999570                     | .11389010390 | .10137597840 | .10013689220 |
| 2 | .17609125910                     | .10882149850 | .10097219580 | .10009767100 |
| 3 | .12493873650                     | .10432956040 | .10057293260 | .10005849910 |
| 4 | .09691001300                     | .10030820300 | .10017808800 | .10001937270 |
| 5 | .07918124605                     | .09667723578 | .09978757705 | .09998028723 |
| 6 | .06694678975                     | .09337473577 | .09940130907 | .09994123464 |
| 7 | .05799194701                     | .09035198920 | .09901920603 | .09990223654 |
| 8 | .05115252247                     | .08757005334 | .09864118340 | .09986328509 |
| 9 | .04575749054                     | .08499735264 | .09826716225 | .09982437469 |

Table 1: Newcomb-Benford-Probability of Digit  $d$  at 1st to 4th position. Generated using  $P(D_n = d) = \sum_{k=10^{n-1}}^{10^n-1} \log(1 + \frac{1}{10^{k+d}})$  and rounded to ten significant digits.

The Digital-Fit-Factor gives an indication on the skewness of the digital distribution function. If the distribution is degenerated in the sense that it has only one mass point, e.g. all first significant digits are equal to 3, the distribution function is not well defined and the non-linear fit is not meaningful in a statistical sense. With at least two mass points present, e.g. the first significant digits where either 1 or 3 but never 2,4,5..., the  $DFF$  can be calculated in most cases with sufficient accuracy.

We used an  $F$ -test on the non-linear estimation of the DFF and report the  $p$ -values in the empirical analysis below.<sup>3</sup> We recommend using a confidence level based on the purpose of the digital analysis, e.g. for the size of lakes in the Chicago area a confidence level of 10% might be sufficient while in the use of fraud detection a 1% confidence level could be taken to avoid false positives.

The DFF is independent of the number of observations  $N$ . This is because the left-hand side of equation (3) sums to one, i.e.  $C/N \cdot \vec{1} = 1$  or  $C \cdot \vec{1} = N$ . E.g. if

---

<sup>3</sup>The  $p$ -value -as usual- refers to the confidence level of a  $F$  or  $t$ -Test, i.e. for a given test  $p$  gives the probability that the test-value occurred only randomly. We will refer to a 'high-significant' confidence level for  $p \leq 1\%$  and 'significant' for  $p \leq 5\%$ .

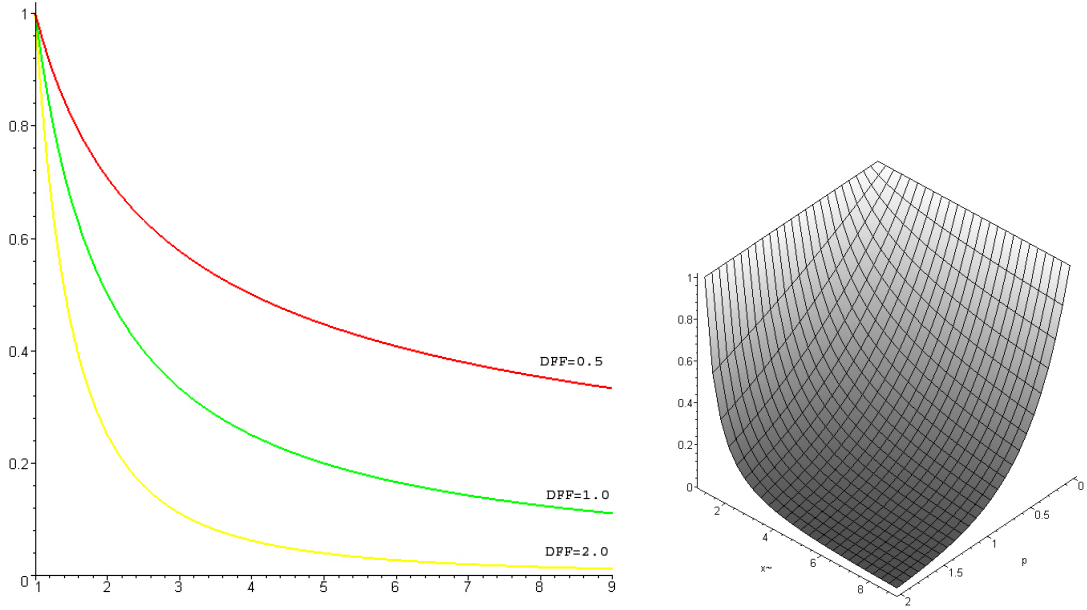


Figure 1: **Graphical representation of the Digital-Fit-Factor (DFF)** The left hand picture shows two-dimensional graphs for different DFF-values, while the right hand graph gives a three-dimensional representation.

one has a data-set with  $N = 500$  of which there are 250 first digits one and 250 first digits three ( $Count_1 = 250$  and  $Count_3 = 250$ ) this results in a DFF-value of around 1.68 (p-value: 0.008), while the same result holds for  $N = 100$  and  $Count_1 = 50$ ,  $Count_3 = 50$ . Thus the DFF is based on the *proportion* rather than the absolute counts. Figure 1 shows some distributions with their DFF-values.

A DFF-value of 1 results the digit-frequency of the Newcomb-Benford-Law, whereas values lower than one indicate a more flat and values greater than one a more step distribution. Uniformly distributed digits would result in a DFF-value of 0, but cannot be estimated with the procedure described above since the values of  $C$  in equation (3) are all the same.

## 2.4 Mantissae Distortion Factor (MDF)

While the DFF analysis specific digit positions in a dataset, an analysis of joint-probabilities would be another possibility, i.e. one could look at the first two digits probability or the first three etc. With the DFF we used a discrete measurement of digits since the deviation from an uniformly distributed digit frequency decreases within the position.

Using the joint-distribution we use the whole information available, i.e. the measure of the Mantissae-Distribution. Although the Mantissae-Distribution is the most general form of the Newcomb-Benford-Law, most of the current literature focuses on the discrete distribution of separated or pairly-joint digits. Following Posch [2004a] who generalizes an idea of Nigrini [1992], we will use the Mantissae-Distortion-Factor (MDF) Model for this part of the digital analysis. The MDF is mainly based on the comparison of the observed and the expected mean of mantissae. While the former is given by the current sample, the latter is calculated using a geometric sequence. Let  $x > 0$ . A geometric sequence  $(x^n)_{n \in \mathbb{N}}$  follows the Newcomb-Benford Law, if and only if  $\log(x)$  is irrational (for a proof see e.g. Posch [2004b]). Let  $M(x_i)$  denote the mantissae of  $x_i \in \mathbb{R}$ . Then the MDF measures the relative deviation of the acutal mean (AM) of mantissae to the expected mean (EM) as given by equation (4).

$$MDF = \sum_{i=1}^N M(x_i) \cdot \left( \frac{10^{\frac{N+1}{N}} - 1}{(10^{\frac{1}{N}} - 1) \cdot N} \right)^{-1} - 1 \quad (4)$$

A perfect fit to the NBL distribution results in a MDF of zero. The greater the deviation from the expected mean the greater the absolute value of the MDF. The sign of the MDF gives the direction of the deviation. If the MDF is positive we have higher values than the NBL suggests and the converse is true if the MDF is negative .E.g. the EM of a geometric sequence with  $N = 100$  is about 3.9638. The greatest observable mean of mantissae is 9.9999 since all mantissae are - per definitionem - smaller than 10. The resulting MDF is then  $1.5228$ . The smallest mean of mantissae is 1 resulting in a MDF of  $1/3.9638 - 1 = -.7477$ . Table 2 gives values of the MDF for different sizes of  $N$ .

| N      | EM                                    |
|--------|---------------------------------------|
| 10     | 4.4759                                |
| 50     | 4.0193                                |
| 100    | 3.9638                                |
| 500    | 3.9197                                |
| 1000   | 3.9142                                |
| 5000   | 3.9098                                |
| 100000 | 3.9087                                |
| Limit  | $9 \cdot \ln(10)^{-1} \approx 3.9087$ |

Table 2: **Values of the expected mantissae mean** This table shows values for the expected mantissae mean (EM) depending on the number of observations (N). The limit as  $N$  approaches infinity is given in the last row.

### 3 Stock Index Analysis

In this section we demonstrate how digital analysis in general and the two measures of DFF and MDF in particular can be used to examine the behaviour of stock market indices. Doubtlessly, these indices are a major measure of a countries performance and have great political impact. Usually there are two variables which are of public interest: The current level of the index and its change over a specific period of time. While the former is not informative per se, the latter is often used to describe and compare countries' economic wealth in terms of stock markets. Further more, investors use such information to gain access to investment projects. One problem which arises in using changes of the index level is the comparability over time and across economies. The usual way to incorporate information of the volatility of performance is the use of variances. This does, however, not solve the problem but rather makes it worse. Since the variance is not scale-independent it cannot be used for comparison.

Using digital analysis and particularly the measures of DFF and MDF leads to a more structural approach in comparing two countries and their stock index performance. We will demonstrate this general procedure with two commonly used stock indices, namely the Standard & Poor's Composite (S&P500) Index and the

Dow Jones Industrial Average (DJIA).

### 3.1 S&P 500

The S&P 500 is widely regarded as one of the best single gauges of the US equity market. It includes 500 companies of the US economy and focuses on the large-cap segments. Using the companies included, the index covers 80% of the US equities and thus is an ideal proxy for the total market.

We demonstrate the digital analysis using yearly data ranging from 2000 to 2003. The dataset is publically available on the Standard and Poor’s homepage. We used price data of the ultimo of each year (usually the 29th of December).

| Year | DFF      | SD[DFF]  | AM       | MDF      | MKTCAP     | PI       |
|------|----------|----------|----------|----------|------------|----------|
| 2000 | 0.518324 | 0.211013 | 4.203701 | 0.072466 | 11,735,261 | 1,320.28 |
| 2001 | 0.630884 | 0.22914  | 3.903374 | -0.00415 | 10,581,281 | 1,161.02 |
| 2002 | 0.842466 | 0.182088 | 3.558841 | -0.09205 | 8,066,633  | 875.395  |
| 2003 | 0.499894 | 0.242187 | 4.170775 | 0.064066 | 10,137,988 | 1,095.89 |

Table 3: **Digital and Market Measures for S&P500 Index (yearly basis)**

This table shows both Digital Measures, namely the Digital-Fit-Factor (DFF) and the Mantissae-Distortion-Factor (MDF) and its components the Expected Mantissae Mean (EM) and the Actual Mantissae Mean (AM) and market measures. The latter are the Market Capitalization (MKTCAP) and the Price Index (PI), where the latter is usually referred to as the index value.

Table 3 shows descriptive data of the yearly S&P 500 including the market capitalization (MKTCAP) and the price index (PI). Figure 2 gives a graphical representation of the yearly analysis while figure 5 uses daily data of the year 2003. Although it seems that the index-value and the DFF are positively correlated this correlation is not significant for the yearly data. Using the daily data of figure 5 the correlation of these measures is in fact negative at -0.9832 (p-value: 0) while the correlation of the MDF and the index is positive at 0.9848 (p-value: 0). Using the return of the S&P 500 this correlation is persistent, but not significant for the DFF.<sup>4</sup> This analysis shows the close connection of the index and the distribution of

<sup>4</sup>The correlation of DFF and the return of the index is at -0.0801 (p-value: 0.20) and for the MDF at 0.1107 (p-value: 0.07).

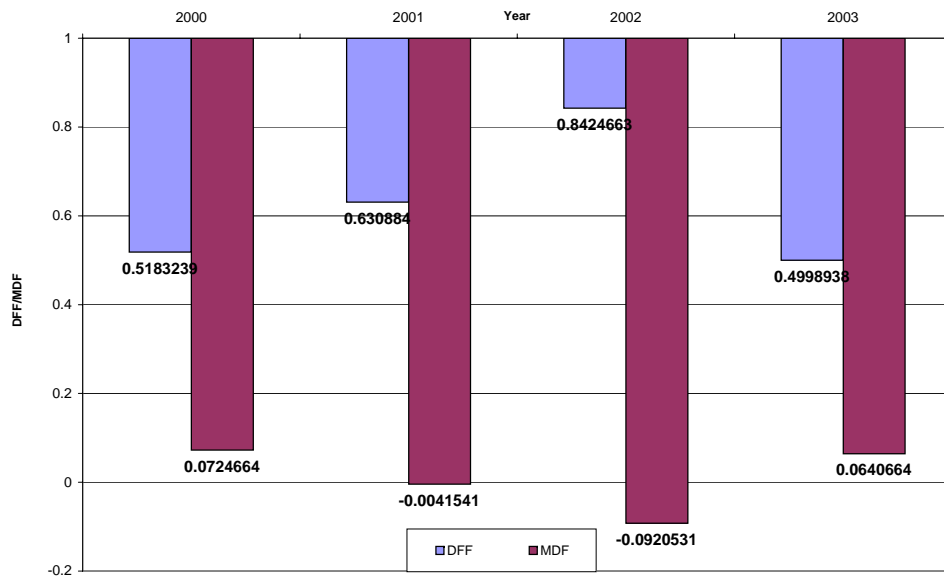
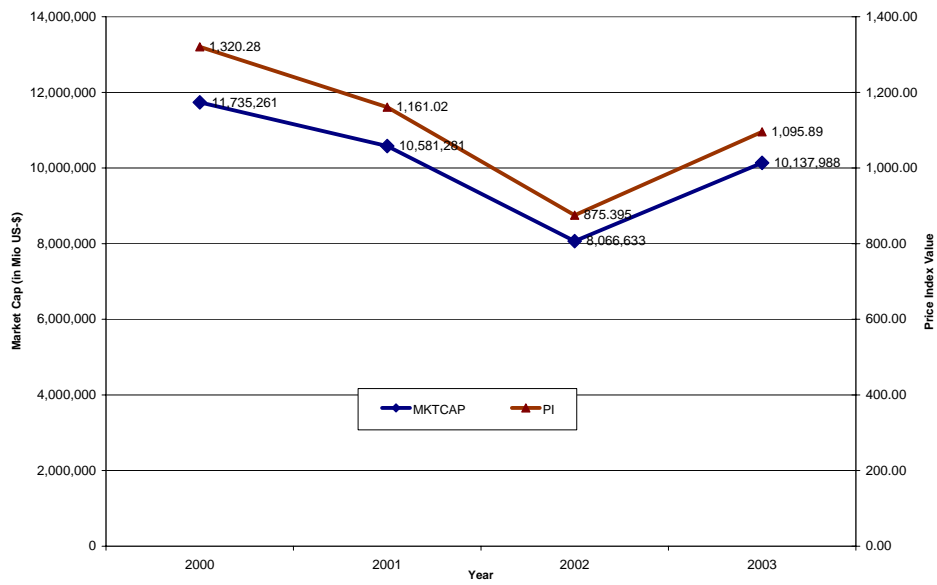


Figure 2: Graphical representation of digital analysis using the S&P500 (yearly data)

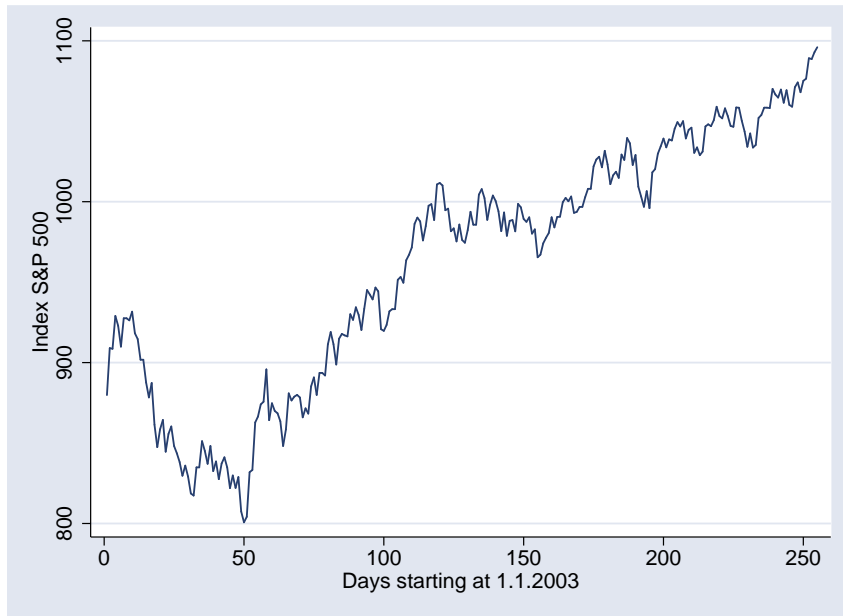


Figure 3: **S&P 500 Index**

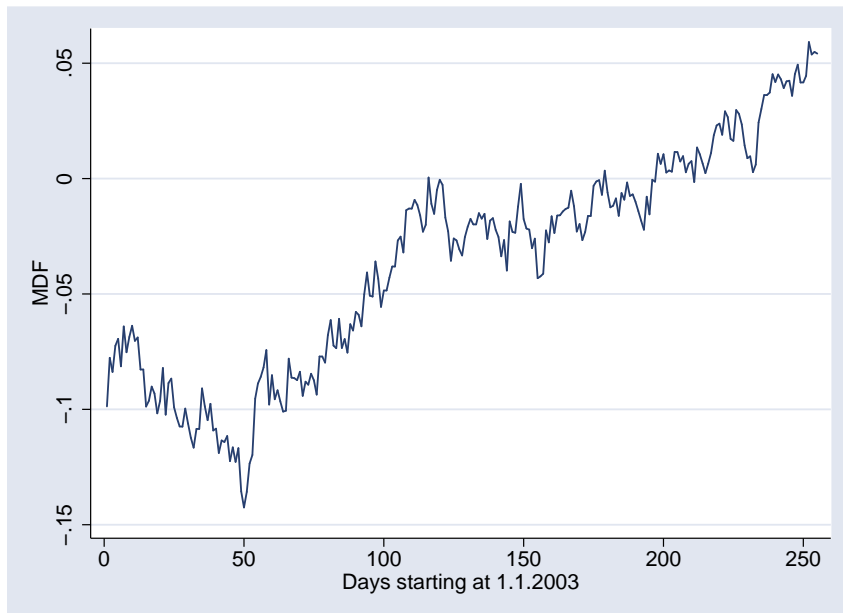


Figure 4: **MDF of S&P 500**

first digits of its stocks. Keep in mind that all the information used to compute the DFF are the first digits of the stock prices. We do not make assumptions on the index composition nor any other measure, just pure digits.

Now the questions arises: a) is this behaviour is stable and b) could it be used for investment strategies etc. The second part of this question is beyond the scope of this paper. To answer the first question, we now examine the Dow Jones Index to avoid 'cultural' differences which might occur when using a non-US-index.

## 3.2 Dow Jones

The Dow Jones index is one of the oldest stock market indices. As the homepage of Dow Jones indicates: 'When Charles H. Dow first unveiled his industrial stock average on May 26, 1896, the stock market was not highly regarded. Prudent investors bought bonds, which paid predictable amounts of interest and were backed by real machinery, factory buildings and other hard assets.'

Today the Dow Jones Industrial Average (DJIA) is maintained and reviewed by editors of The Wall Street Journal. To gain continuity composition changes are relatively seldom.

The only paper known to us which uses digital analysis of stock markets is Ley [1996] who discusses the digital distribution of daily returns of the Dow Jones Industrial (DJIA) Index. He uses the  $\chi^2$ -Test as goodness-of-fit measure. Although the  $\chi^2$ -tests are not significant he concludes a close relationship between the returns and the NBL-distribution. Considering the assumptions of the  $\chi^2$ -test that the observation must be independent, we doubt his conclusion that the rejection of the test is due to a 'weakness of the Newman-Pearson-Theory'. Furthermore Ley [1996] uses a direct comparison of the actual digits observed in the returns of the DJ and the expected ones due to the NBL. In our methodology we fit a whole digit-distribution and compare it to the NBL-distribution. Furthermore we do not use returns.

Table 4 shows descriptive data for the MDF of the index on the daily basis over the year 2003. Figure 6-8 show a comparison of the Index itself, its DFF and MDF-Values. Note that the DFF-Value contains almost the whole information of the index.

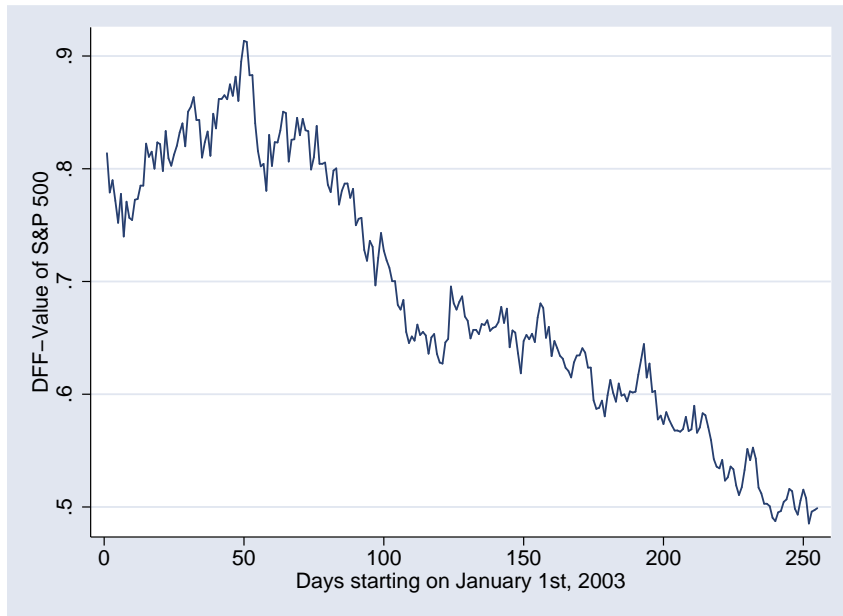


Figure 5: **DF**F of S&P 500

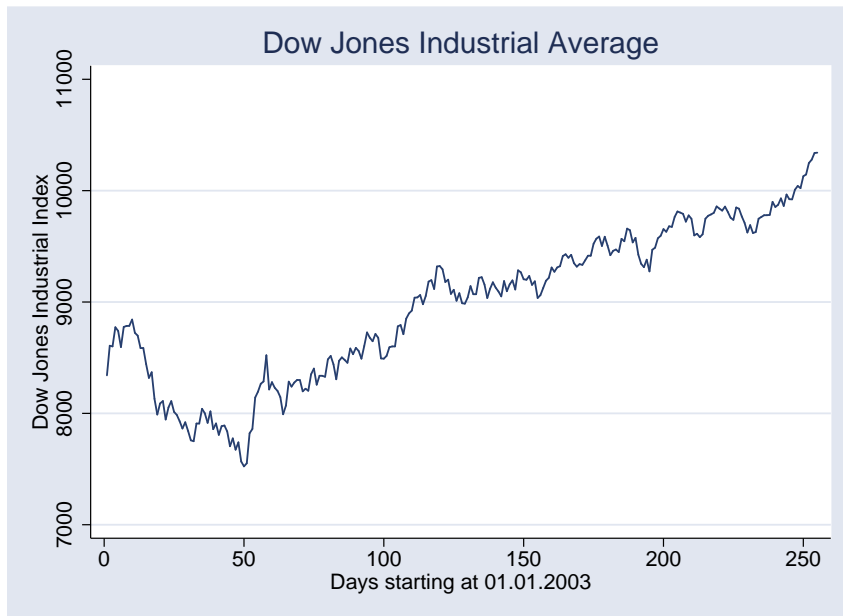


Figure 6: **Dow Jones Industrial Average Index**

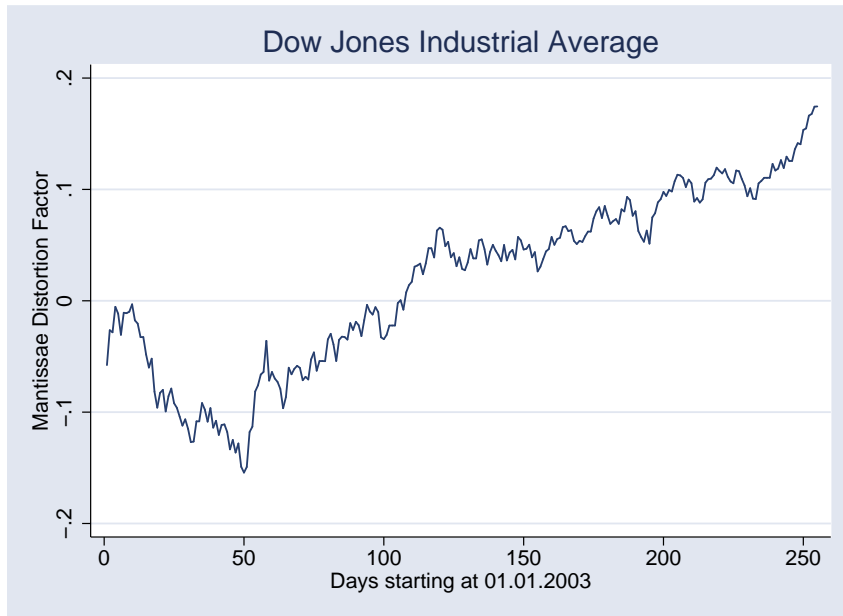


Figure 7: **MDF of Dow Jones Industrial Average**

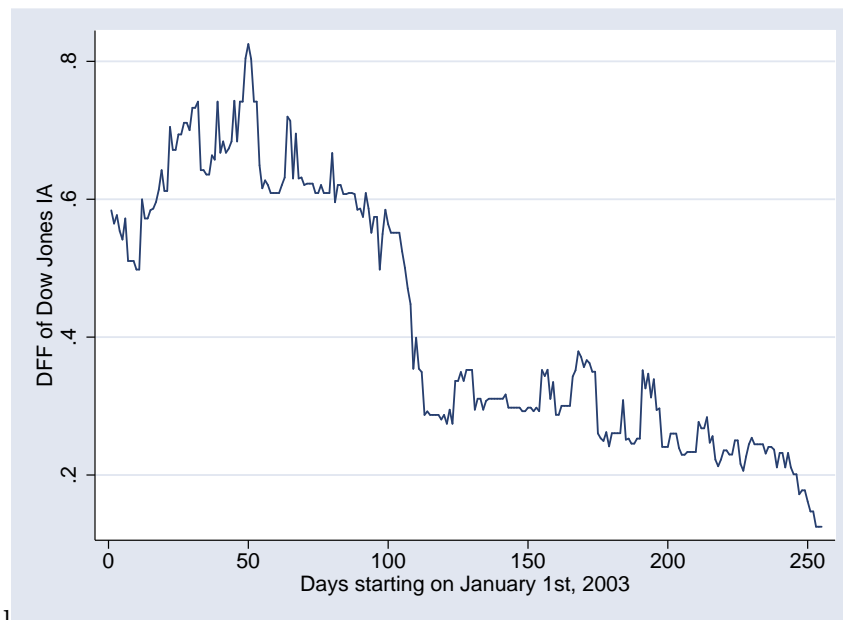


Figure 8: **DFF of Dow Jones Industrial Average**

|               | DJIA   | S&P500 | Nikkei |
|---------------|--------|--------|--------|
| mean          | 0.4225 | 0.8148 | 0.6841 |
| median        | 0.3470 | 0.8139 | 0.6588 |
| sd            | 0.1842 | 0.0379 | 0.1150 |
| 5% percentile | 0.2111 | 0.7490 | 0.5029 |
| 95%percentile | 0.7198 | 0.8836 | 0.8617 |

Table 4: **DDF - values of the Indices (Daily Basis)** The table shows the DDF-Values of several Indices on daily basis starting at the 1st of January 2003 ending at the 31st of December 2003. DJIA denotes the Dow Jones Industrial Average Index (USA), S&P500 the Standard and Poor's 500 Index (USA) and Nikkei the Nikkei 225 (Japan).

|               | DJIA     | S&P500   | Nikkei 225 |
|---------------|----------|----------|------------|
| mean          | 13417.06 | 961.6147 | 9879.089   |
| median        | 13612.55 | 984.03   | 10225.22   |
| sd            | 1107.275 | 77.05303 | 1260.215   |
| 5%percentile  | 11630.3  | 831.9    | 7907.19    |
| 95%percentile | 14991.65 | 1066.62  | 11800.4    |

Table 5: **Descriptive Statistics of Stock Market Indices.** The table shows the Index-Values of several indices on daily basis starting at the 1st of January 2003 ending at the 31st of December 2003. DJIA denotes the Dow Jones Industrial Average Index (USA), S&P500 the Standard and Poor's 500 Index (USA) and Nikkei the Nikkei 225 (Japan).

|               | DJIA    | S&P500  | Nikkei  |
|---------------|---------|---------|---------|
| mean          | 0.0180  | -0.0360 | 0.0008  |
| median        | 0.0381  | -0.0231 | -0.0001 |
| sd            | 0.0792  | 0.0480  | 0.0217  |
| 5% percentile | -0.1149 | -0.1135 | -0.0315 |
| 95%percentile | 0.1255  | 0.0419  | 0.0361  |

Table 6: **MDF-values of the Indices (Daily Basis).** The table shows the MDF-Values of several indices on daily basis starting at the 1st of January 2003 ending at the 31st of December 2003. DJIA denotes the Dow Jones Industrial Average Index (USA), S&P500 the Standard and Poor's 500 Index (USA) and Nikkei the Nikkei 225 (Japan).

|               | DJIA    | S&P500  | Nikkei  |
|---------------|---------|---------|---------|
| mean          | 0.090%  | 0.092%  | 0.082%  |
| median        | 0.097%  | 0.101%  | 0.037%  |
| sd            | 1.038%  | 1.068%  | 1.351%  |
| 5% percentile | -1.630% | -1.524% | -2.385% |
| 95%percentile | 1.888%  | 1.950%  | 2.112%  |

Table 7: **Returns of Indices (Daily Basis)**. The table shows the Daily Returns of several indices on daily basis starting at the 1st of January 2003 ending at the 31st of December 2003. DJIA denotes the Dow Jones Industrial Average Index (USA), S&P500 the Standard and Poor's 500 Index (USA) and Nikkei the Nikkei 225 (Japan).

Looking at the correlation of the Dow Jones Index gives the same picture as before. The DFF and the index-value are highly significant correlated at -0.9689 (p-value: 0), while the MDF and the index-value range at 0.9985 (p-value: 0). Again the (daily) returns of the index show the same correlation sign, but are less significant.<sup>5</sup> This shows that the direction of the correlation is stable across the two indices examined.

With this result it is possible to use standard econometric methods to predict stock index values using the digital measures of DFF and MDF. We undertook preliminary analysis which showed stable regression results using these measures. Since the purpose of this paper is to show a general methodology of digital anylsis these results are not reported.

Instead we show the possiblity of 'cross-cultural' comparision using the Nikkei 225 index.

### 3.3 Nikkei

The Nikkei Stock Average is the most widely watched index of the japanese stock market activity. It is calculated continuously since September 1950 and is based on the Dow Jones method.

Since the former stock indices, the S&P500 and the Dow Jones, are based on the US market, the examination of the Nikkei has several implications. Firstly, even

<sup>5</sup>The values for the DFF are -0.0856 (p-value: 0.1131) and for the MDF 0.1737 (p-value: 0.0721)

if the markets were correlated, this correlation is not perfect in the sense of exact co-movements. Secondly, presumably there are effects in each market which drive a country's performance.

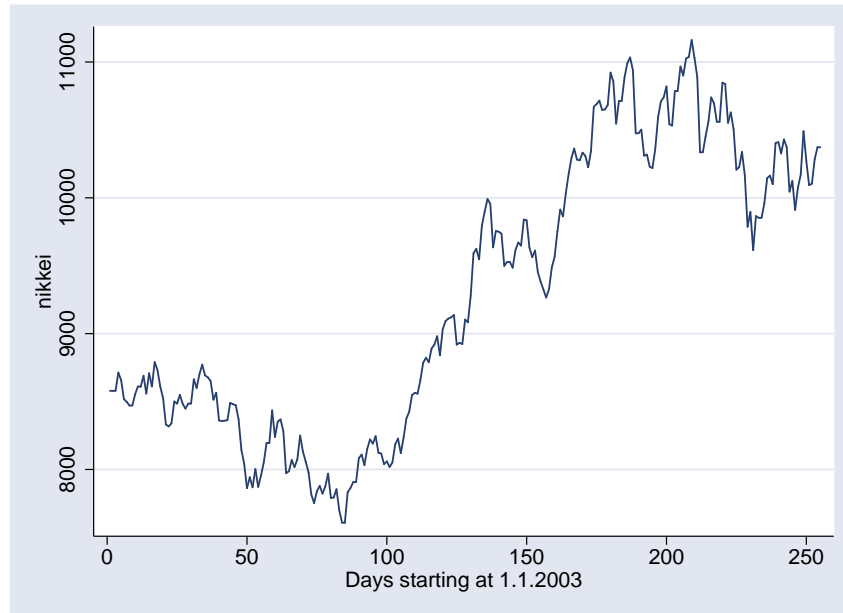


Figure 9: **Nikkei 225 Index**

Like the US-indices the Nikkei 225 and the measures of DFF and MDF are highly correlated.<sup>6</sup> Looking at the DFF figures of the three indices we see a decreasing DFF-value over time for the US-indices while the Nikkei-DFF is relatively stable. Analytically this is shown in table 4 in the percentiles. While the Nikkei difference of the 95% and the 5% percentile is very narrow (about three standard deviations) the US-indices' are much wider. Examining the level of the DFF it is crucial to note that the Nikkei-DFF has the highest value over time while the Dow Jones has the lowest. Although the DFF is independent of the number of stocks included in an index, one reason for this observation might be that it is more probable to observe same first digit behaviour in a narrowly defined market - such as the Dow Jones - than in a widely defined one. With this interpretation in mind the S&P 500 could serve as a benchmark. Then one could say that the Dow Jones does not represent the US-Market because of its significantly lower DFF, while the Nikkei shows the behaviour of a foreign market.

<sup>6</sup>The DFF-index-correlation is: -0.8508 (p-value: 0) and the MDF-index-correlation is : 0.8145 (p-value: 0)

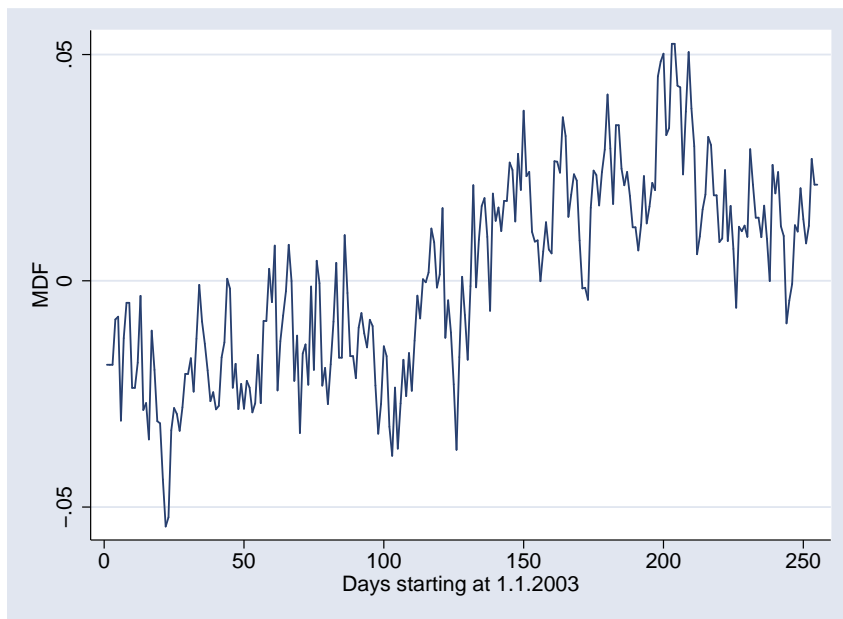


Figure 10: **MDF** of Nikkei 225

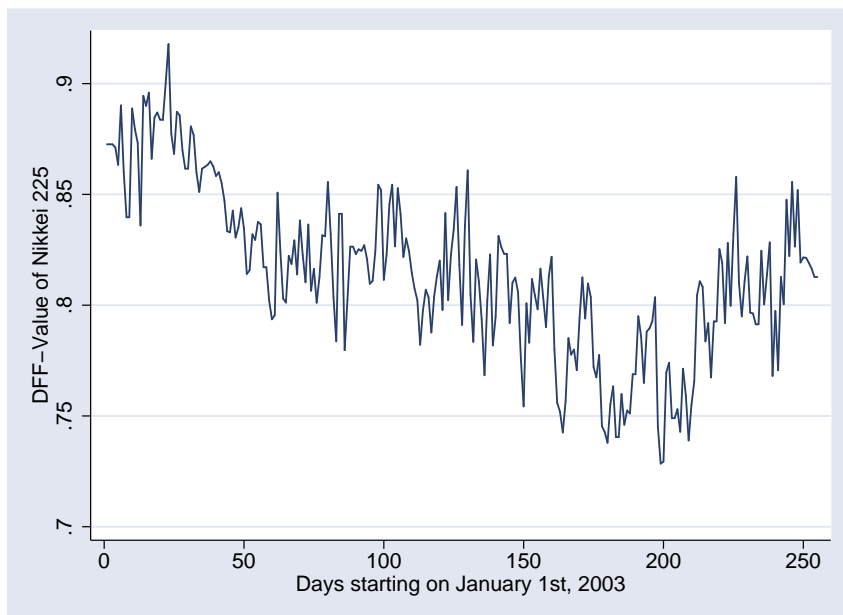


Figure 11: **DFF** of Nikkei 225

The daily returns of each index are almost the same as shown by table 7, but the Nikkei's are more volatil and slightly smaller than the US-indices' ones. This is the expected result when using the MDF for interpretation. The absolute values of the MDF are highest for the S&P 500 yielding the highest returns of the index. Almost equally followed by the Dow Jones but significantly lower than the MDF of the Nikkei (see tabel 6 for details).

## 4 Portfolio Composition

To demonstrate the use of the DFF index we simulated the following portfolios. Starting with an initial portfolio value of 100 we trade if and only if the index change is beyond a specific threshold. Consider the following example.

| Day | SP500  | DFF of SP500 | PF using Index | PF using DFF |
|-----|--------|--------------|----------------|--------------|
| 1   | 879.82 | 0.81         | 100            | 100          |
| 2   | 909.03 | 0.78         | 100            | 100          |
| 3   | 908.59 | 0.79         | 99.9           | 98.56        |
| 4   | 929.01 | 0.77         | 102.15         | 100.77       |

On Day 1 the Change of the SP500 cannot be calculated since values before that date are not included in this dataset. On day 2 the change is calculated as the value of day 2 divided by the value of day 1 minus 1:  $879/909-1=3.32\%$ . The value of the Portfolio is 100 times the acutal change that is 103.32. At the end of day 2 the portfolio composition is revised. If the observed index change is beyond the given threshold we adjust the portfolio by the proportion of the index change, e.g. take a threshold of 2.5%. Since the change observed on day 2 (3.32%) is beyond the threshold we adjust the portfolio and since the change is positive we buy 3.32% of the index to the portfolio. This strategie assumes that we expect the observed trend to be persistent, i.e. if the index raises we expect further increases if it falls we expect further decreases.

The level of the threshold takes account of the changes we are willing to accept without adjusting our portfolio, e.g. a threshold higher than the maximum observed change (e.g.  $\geq 10\%$  as a daily change of more than 10% is very unlikely) will result in not adjusting at all. In this case all simulated portfolios have the same value.

Whereas a very small threshold (e.g. 0.01%) result in a daily adjustment.

Since the choice of the threshold level influences the value of the portfolio over time its specification is crucial. Our purpose is to demonstrate the use of the DFF in portfolio composition, so we simulate over a wide range of possible thresholds and look at the mean performance over all simulation steps. The range of thresholds starts at

The following table shows descriptive data of the mean of all simulations.

|                | S&P 500 | DFF of S&P 500 | Nikkei | DFF of Nikkei | Dow Jones | DFF of Dow Jones |
|----------------|---------|----------------|--------|---------------|-----------|------------------|
| Mean           | 7.65%   | 22.03%         | 5.75%  | 5.09%         | 4.11%     | 10.75%           |
| Median         | 10.58%  | 25.02%         | 0.00%  | 0.00%         | 0.00%     | 1.09%            |
| SD             | 9.60%   | 1.95%          | 1.13%  | 1.20%         | 8.03%     | 1.55%            |
| 5% Percentile  | 9.10%   | 9.28%          | 9.10%  | 8.94%         | 9.14%     | 9.18%            |
| 95% Percentile | 20.65%  | 53.26%         | 27.39% | 27.80%        | 17.47%    | 35.11%           |

Table 8:

What is notable is that the US Indices' DFF perform on a very high level while the Nikkei's DFF results in less performance than the index investment. Furthermore the standard deviation (SD) as a measure for the risk is smaller for the US indices than for the Nikkei.

The following pictures give a graphical overview of the mean performance over time.

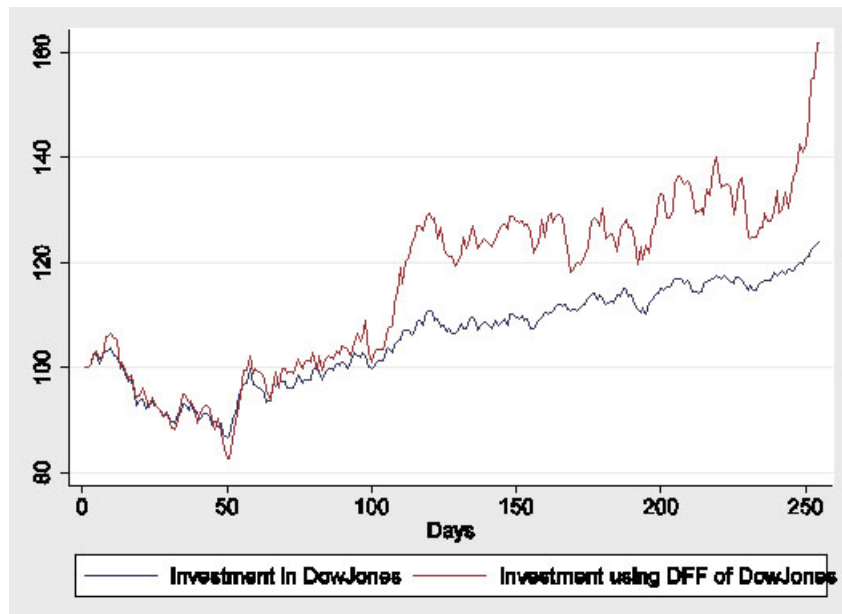


Figure 12: Dow Jones Index - Portfolio Comparison

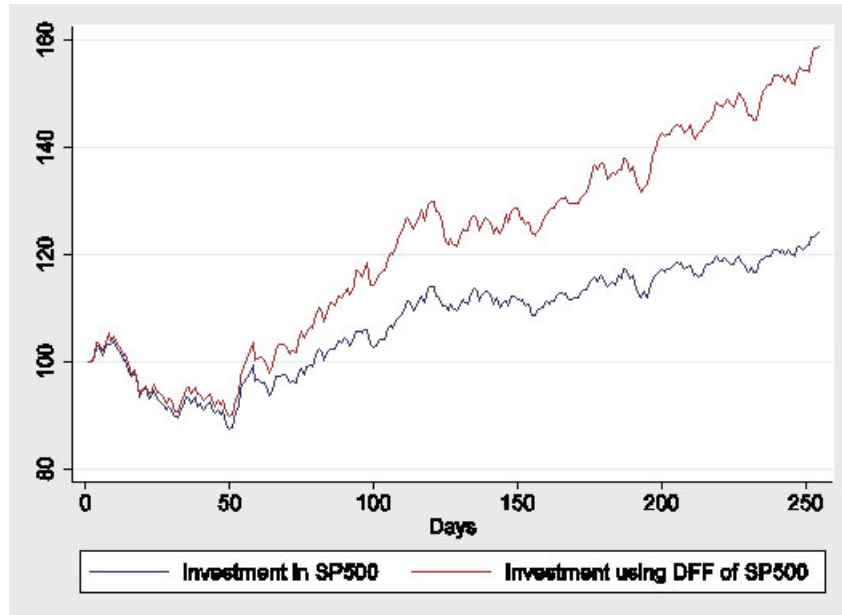


Figure 13: S&P 500 Index - Portfolio Comparison

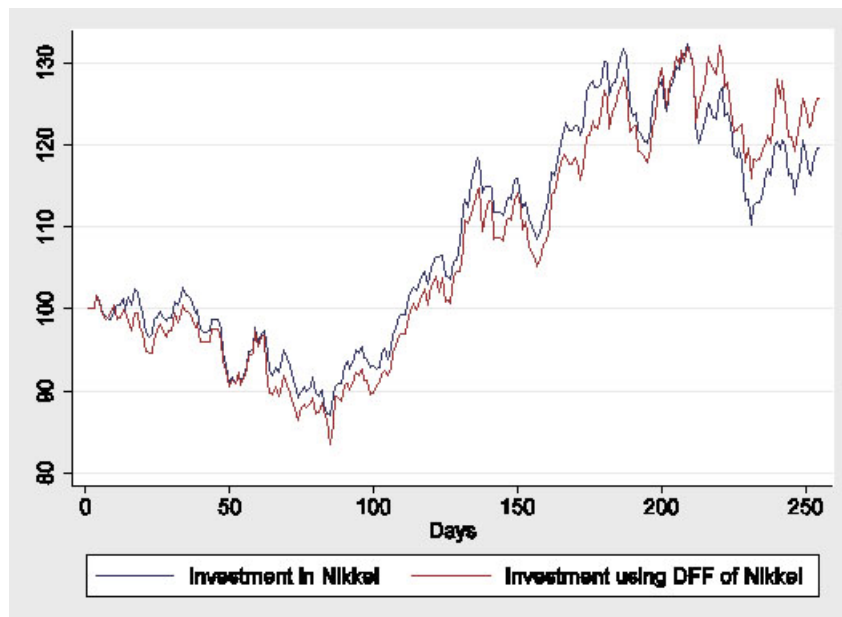


Figure 14: Nikkei 225 Index - Portfolio Comparison

## 5 Conclusion

In this paper we demonstrate a general methodology for digital analysis of economic measures. The generality of this approach leads to two measures which can be applied to any set of economic variable in order to compare them within time or within economic-units such as countries, states etc. Our measures are based on an

approach known as 'Digital Analysis', which is already used to detect frauds in tax sheets and company reports. The most widely used distribution in this field is known as the 'First-Digit-Law' or 'Benford-Law', where the latter expression suppresses the original discovery by Simon Newcomb, so we will refer to this phenomenon as Newcomb-Benford Law (NBL).

The Digital-Fit-Factor (DFF) is based on the probability of digit distribution and fitted using non-linear techniques while the Mantissae-Distortion-Factor (MDF) uses joint probabilities and its deviation from the Newcomb-Benford-Law.

We demonstrate the power of our measures examining the Dow Jones Industrial Average Index and the Standard and Poor's Composite Index over time. The comparability on the economic-unit level is demonstrated by adding the Nikkei 225 Index.

Further research should focus on the impact of both the DFF and the MDF on other economic variables, such as growth rates of unemployment, cash transfers etc. in order to explain differences between national and international markets. Furthermore the use of digital analysis in the field of index tracking and portfolio composition should be examined, because the close correlation of the measures of the MDF and DFF and the index-value suggests them to have a great impact on the returns.

## References

- Frank Benford. The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4):551–572, 1938.
- Theodore P. Hill. A statistical derivation of the significant-digit law. *Statistical Science*, 10(4):354–363, 1996.
- Welf A. Kreiner. On the newcomb-benford law. *Z. Naturforschung*, 58a:618–622, 2003.
- Eduardo Ley. On the peculiar distribution of the u.s. stock indices first digits. *The American Statistician*, 50(4):311–314, 1996.
- Simon Newcomb. Note on the frequency of use of the different digits in natural numbers. *Amer. J. Math.*, 4:39–40, 1881.
- Mark J. Nigrini. *The detection of income evasion through an analysis of digital distributions*. PhD thesis, Dept. of Accounting, University of Cincinnati, 1992.
- Peter N. Posch. Benford or not-benford? how to test for the first digit law. Working Paper, <http://www.mathematik.uni-ulm.de/dof/pnposch>, 2004a.
- Peter N. Posch. A survey on sequences and distribution functions satisfying the first-digit-law. Working Paper, <http://www.mathematik.uni-ulm.de/dof/pnposch>, 2004b.