
University of Surrey

Department of Computing

Reuters Data Analysis

*Brand Protection: Comparing the Provenance
of two Time Series*

Khurshid Ahmad

Saif Ahmad

November 21, 2003
(Pre-final Draft)



OBJECTIVE:

To compare two time series in order to ascertain whether or not the two share a common source.

DATA

Two sets of time stamped data were analysed: one was generated/marketed by Reuters.com and the other by Bridge.com. Each data set comprises time variation of the *bid* and *ask* price for a financial instrument, Euro (€) and US Dollar (\$) exchange rate on a single day (July 29, 2003). There are 31873 data points from Reuters and 29652 data items from Bridge. The data is time stamped as and when the two vendors receive it. A sample set of 8 data arrivals (for the first minute) in Table 1 show that the two vendors receive data at different (*time*) points – data which is not in agreement.

Table 1 Sample of *Tick Data Arrival* for the two Vendors (00:00:00 – 00:00:57)

Time (hh:mm:ss)	Reuters		Bridge	
	Bid	Ask	Bid	Ask
00:00:00			1.1483	1.1488
00:00:03			1.1485	1.1490
00:00:06			1.1482	1.1487
00:00:24	1.1488	1.1496	1.1481	1.1486
00:00:27			1.1481	1.1486
00:00:45			1.1483	1.1488
00:00:48			1.1483	1.1488
00:00:57	1.1487	1.1497		

In order to ascertain the provenance of Bridge data (or Reuters data for that matter) we can carry out visual inspection. Most computer-based analysis systems use time-series analysis techniques to analyse a set of ordered data, which is available either continuously or at discrete time points. In order to use the analysis techniques, especially wavelet analysis and signal processing techniques, on the unordered data (e.g. as in Table 1) it is necessary to use data *compression* that aggregates the movement in the dataset over a certain period of time. The *compression* acts as a surrogate for the original. Table 2 shows a 1-minute *compression* (for the first 3 minutes) of both the Reuters data and Bridge data: here the maxima (*High*) and minima (*Low*) of the data in the minute and the value at the start (*Open*) and the end (*Close*) of the minute act as the surrogate for other data during the minute. There are special diagrammatic conventions to organise this display of the surrogates – called *Japanese Candlesticks*.

Table 2 One-minute Compressed Data for the two Series (for the first three minutes)

Time (hh:mm)	Reuters (<i>Bid</i>)				Bridge (<i>Bid</i>)			
	Open	High	Low	Close	Open	High	Low	Close
00:00	1.1488	1.1491	1.1487	1.1491	1.1483	1.1485	1.1481	1.1483
00:01	1.1491	1.1492	1.149	1.1492	1.1482	1.1486	1.1482	1.1484
00:02	1.1491	1.1493	1.149	1.1493	1.1485	1.1485	1.1483	1.1483
00:03	1.1492	1.1493	1.1491	1.1492	1.1485	1.1486	1.1483	1.1486

METHOD:

Our method relies on comparing the behaviour of two time series to establish whether one is a replica of the other. The hypothesis we have is this: one vendor, say V_A , may use the other vendor V_B 's data and perhaps randomly add/subtract from the data at different time points. There could be systematic additions to and subtractions from the data value as well. This means that V_A will preserve the overall patterns of movement in the data even after subtraction or addition to the original data V_B . These patterns include cyclical effects, correlation between current and immediate past values, *short-term* fluctuations, and *long-term* trends. If the patterns in V_A and V_B are the same, then V_A has the same provenance as V_B .

There are many different ways of finding the key elements in a time series – characteristic cyclical patterns, auto-correlational patterns, fluctuations and trends, and related power spectra. We will use the so-called wavelet analysis to extract these key elements – a kind of summary of the time series.

Our method comprises the following steps:

- (I) Compress the data on a minute-by-minute basis to get *Open (O)*, *High (H)*, *Low (L)* and *Close (C)* values for every minute.
- (II) Perform *Pearson Product Moment Correlation* on *OHLC* values for the two series.
- (III) Introduce lag/lead in the surrogates of one of the series and apply Method II.
- (IV) Process the results of Method II and III, in terms of its stationary and non-stationary components using wavelet transform (WT) and compare statistic related to the components.
- (V) Generate a textual summary for the two time series and compare the summaries visually.

PROCEDURE:

The original data comprises *bid* and *ask* prices (*tick* data) marketed by Reuters.com and Bridge.com for Euro (€) and US Dollar (\$) exchange rate on July 29, 2003. This is shown in Figure 1.

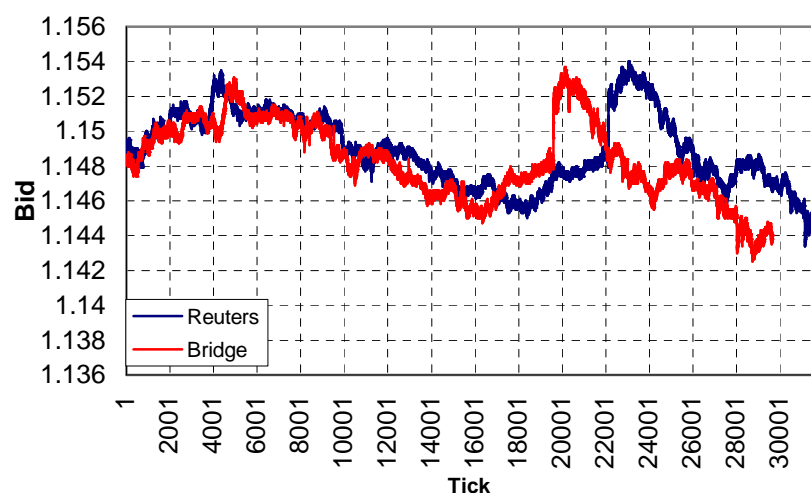


Figure 1 Raw *Tick* Data with around 30,000 Data Points each

Compute Surrogates – Apply Method I

Data *compression* of the two series essentially yields *four* new time series for each: Open, High, Low and Close data values. Figure 2a and Figure 2b show candlestick plots; comprising 4 time series each, of the Reuters and Bridge *bid* data for the first fifteen minutes. A *solid* candle body shows that the Close is *lower* than the Open price whereas a *hollow* candle body shows that the Close is *higher* than the Open price.

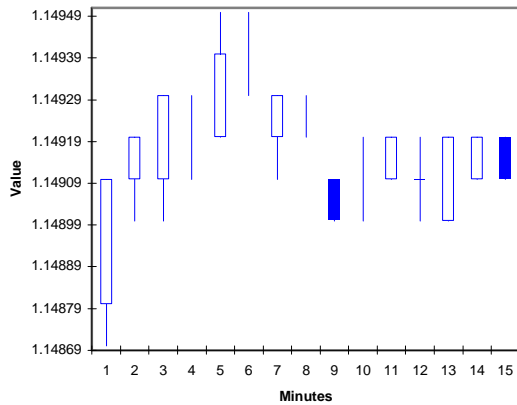


Figure 2a Candlestick Plot for Reuters (Blue series)

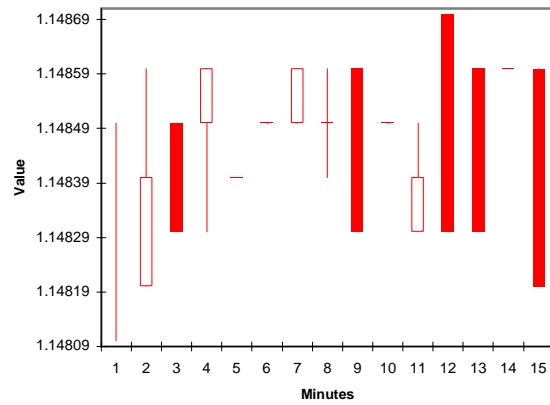


Figure 2b Candlestick Plot for Bridge (Red series)

Compute Correlation – Apply Method II

We perform a correlation analysis on all four *OHLC* arrays for the two datasets (Method II). This is a standard correlation method where the differences in each and same discrete time point in the two series are computed resulting in the so-called *Pearson Product Moment Correlation*. We observe a high correlation of approximately 80 % amongst each of the four surrogates for the two datasets. The correlation for *High* value is the highest (80.16 %); therefore we will concentrate on this series for the two datasets for further analysis.

Transformed Value Computations – Apply Method III

There is a possibility that one time series was shifted at the origin and reported thus. An inspection by eye shows that the peaks and troughs are separated in the *High* value of the surrogate: it appears that there is a 50 minute lag between Reuters data and that of Bridge; the *High* values in Reuters data lag behind Bridge by 50 minutes (Figure 3a).

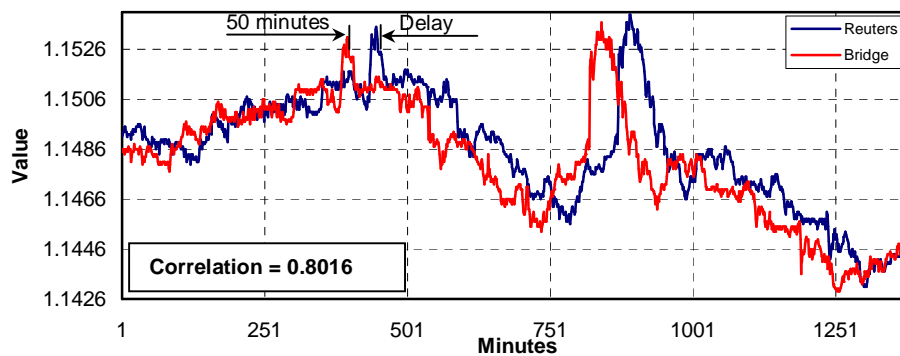


Figure 3a The Plot and Correlation of the ‘High’ Values of the Two Series in their Un-shifted Form

The correlation between the *High* time series was computed by removing the lag in the Reuters data: the results show that the new correlation is over 99% (see Figure 3b)!

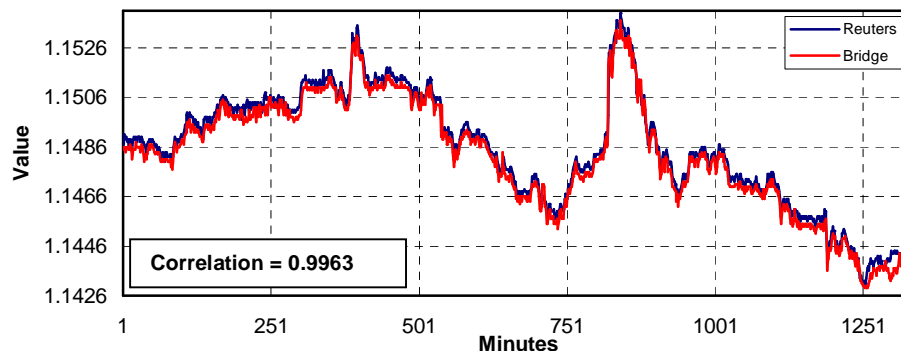


Figure 3b The Plot and Correlation of the ‘High’ Values of the Two Series in their Shifted Form (after shifting Bridge series fifty units forward)

Wavelet Analysis – Apply Method IV

Given the strong correlation in the original surrogate (*OHLC*) series of the two vendors (approx. 80%) and even higher correlation when an origin shift is applied to the Bridge data (approx. 99%), suggests that the provenance of the two datasets is similar. However, the numerical differences, and some are significant, suggests that it is not just simply a shift of origin – to put more strongly, and perhaps intuitively, vendor V_A supplies vendor V_B 's data by the simple device of copying the data from V_B and delaying its own data release by a fixed time interval. The data values are not the same and this could be accounted for by the addition of random noise to the plagiarised data. Random noise added (or subtracted) from a series is tantamount to introducing a *short-term* fluctuation on an otherwise stable system, which may have its own naturally occurring fluctuations.

The so-called Discrete Wavelet Transform (DWT) is used for analysing data with localised fluctuations. The DWT represents a signal as a sum of *approximations* (A_s) and *details* (D_s) that are localized in time and frequency. These individual *approximations* and *details* are obtained by recursively convolving the signal with a bank of *low-pass* and *high-pass* filters. These filters are unique to a single *prototype* or *mother* wavelet. Each successive recursion represents the highest to lowest frequency component of the original signal. Our system employs *Daubechies* six-coefficient filter banks to recursively convolve the original signal.

In this manner, we compute various (wavelet) *scales* associated with a time series. Higher *scales* or low frequency components provide a series-wide information about a time series and the lower *scales* or higher frequency components comprise *detailed* information which is otherwise emasculated in the original time series. Figure 4¹ shows the discrete wavelet transform *pyramidal* paradigm for a *level-N* decomposition. This iterative procedure (*level* of decomposition) can be performed as many times as desired. For a *level-N* decomposition, the fluctuations (high and low frequency phenomenon) are captured by decompositions D_1 to D_N , which are called the *details* of the signal. We can see from Figure 4 that the frequency of the wavelet components decreases after each filtering iteration. Therefore D_1 will represent very

¹ See Appendix B

short-term fluctuations (higher frequency phenomenon) whereas D_N will represent *long-term* fluctuations (lower frequency phenomenon). A_N is called the *highest-level approximation* and it captures the overall trend in the signal. Since D_1 to D_N and A_N are all decompositions (components) of the original signal S , adding them up gives us back the original signal. Eq (1) below shows this:

$$S = (D_1 + D_2 + \dots D_N) + A_N \quad (1)$$

Manipulating Eq (1) gives us,

$$A_N = S - (D_1 + D_2 + \dots D_N) \quad (2)$$

Eq (2) shows that once all the *short* and *long-term* fluctuations are removed from the original signal, we are left with the *long-term* trend A_N . Figure 5² shows the Reuters series (blue) in the centre while surrounding it are some of its various wavelet components for a *level-10* wavelet decomposition. Very *short-term* fluctuations are shown to the right of the original signal (D_1 , D_2 and D_3) while *long-term* fluctuations (D_5 , D_6 and D_7) are shown to the left of the signal S . A_{10} (shown in green) is the *long-term* trend in the original signal S .

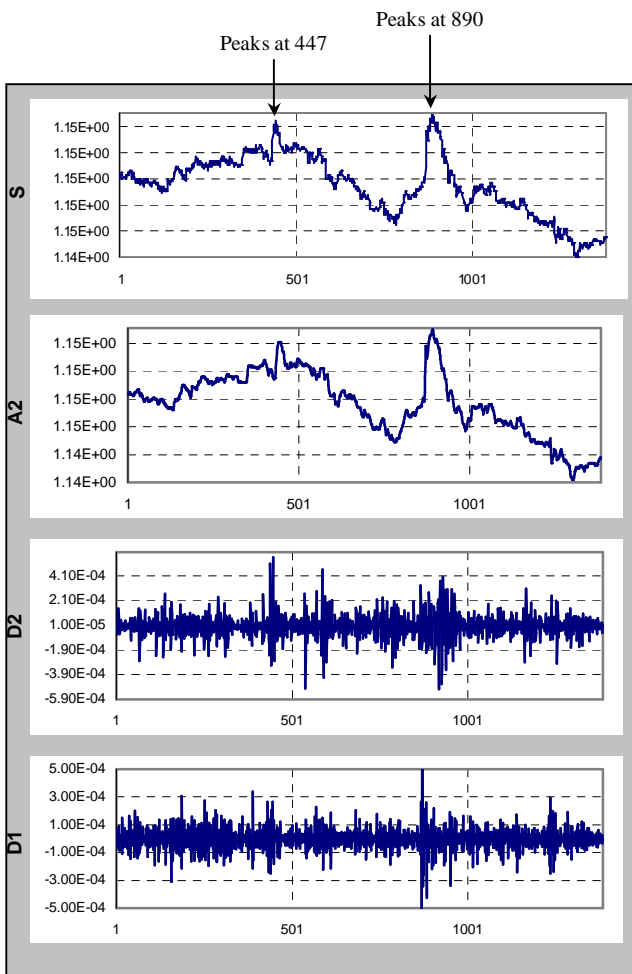


Figure 6a Short-term Fluctuations in the Reuters Data

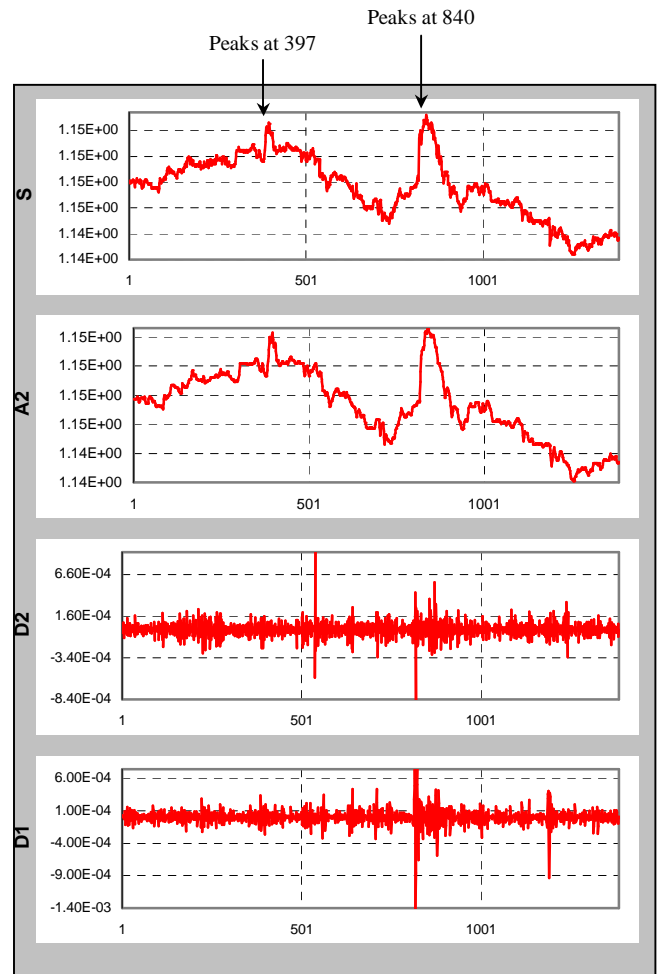


Figure 6b Short-term Fluctuations in the Bridge Data

² See Appendix C

Figure 6a and Figure 6b show a *level-2* wavelet decomposition for the Reuters and Bridge *High* data. Since we have kept the *level of decomposition* (or *level of wavelet filtering*) to only 2, we are able to separate out only the very *short-term* fluctuations from the two series. These are represented by D1 and D2 in Figure 6a and Figure 6b. Though the *highest-level approximation* A2 preserves the basic shape of the original series, we can still observe that it has smoothed out quite considerably as compared to S.

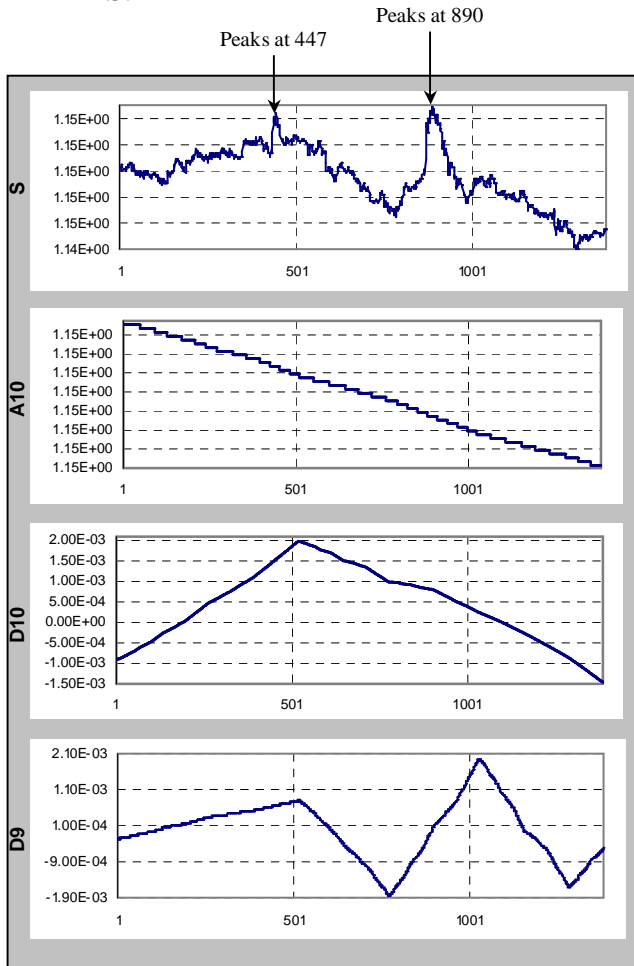


Figure 7a Long-term trend in the Reuters Data

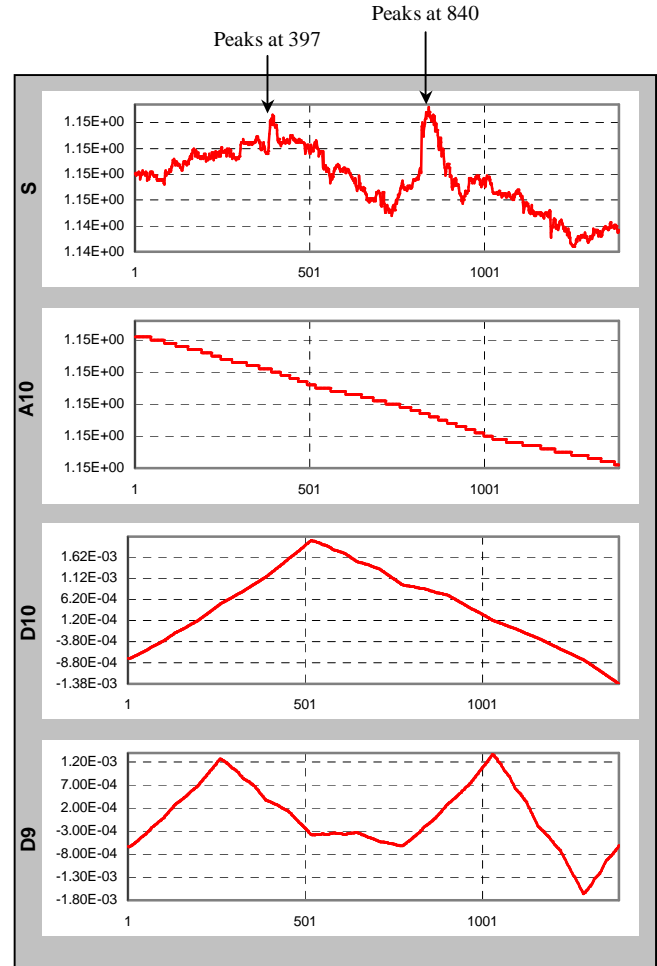


Figure 7b Long-term trend in the Bridge Data

As we keep on filtering the data, using the discrete wavelet transform (DWT), we keep removing the *short-term* and *long-term* fluctuations from the series and the *highest-level* approximation, A_N ($N = \text{level of wavelet decomposition}$) tends to be so smooth (*stable*) that it can be regarded as the *long-term* trend.

This is shown in Figure 7a and Figure 7b where we just show the *level 9 – level 10* results of a full *level-10* decomposition performed on the two series. It can be observed from Figure 7a and Figure 7b that the *long-term* trend, A10 is very similar for both the cases, thus giving further evidence that the provenance of the two datasets is the same.

We apply Method II to calculate the correlation between the two datasets after each level of wavelet filtering³. We observe (as expected) that the correlation between the

³ This analysis is performed on the *un-shifted* series

two datasets dramatically tends towards one (99.97 % to be precise) as we recursively filter them to remove the random noise and *short* and *long-term* fluctuations. Figure 8 shows the *rise* in correlation between the two datasets after each level of wavelet filtering.

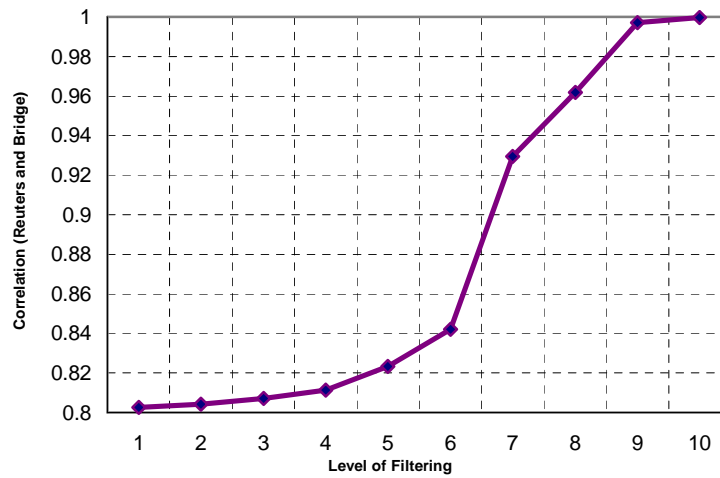


Figure 8 Correlation Versus Wavelet Filtering Level

Generate a Textual Summary of the two Time Series – Apply Method V

In the previous section, we performed wavelet analysis on the *un-shifted* series and showed how it can be used to determine the provenance of two time series by performing a correlation analysis between the two after each level of wavelet filtering (Figure 8). In this section, we run both the series through our *time series summarization module* to generate natural language summaries of the input data and compare these summaries by eye. Here we perform the analysis after shifting the Bridge series fifty minutes forward: we expect the summaries to make more sense after we introduce a shift of origin in the Bridge series.

Table 3 Output of the *Time Series Summarization Module* (after shifting the Bridge Series fifty minutes forward)

Feature	Reuters	Bridge
Trend	Downtrend $S_L'(t) = -(2.96 \times 10^{-6})t + 1.150$	Downtrend $S_L'(t) = -(3.09 \times 10^{-6})t + 1.149$
Dominant Cycle	205 minutes Starts in the 17 th minute	205 minutes Starts in the 17 th minute
Key Turning Points	517 th , 645 th , 836 th , 900 th , 1028 th , 1285 th and 1334 th minute	517 th , 645 th , 836 th , 900 th , 1028 th , 1285 th , and 1334 th minute
Variance Change⁴ Occurs in the	907 th minute	816 th minute

⁴ See Appendix A for Variance Change

Table 3 shows the program output for Reuters and Bridge data after *shifting* the Bridge series forward by fifty minutes. There are a lot of common features in the *summary* for the two series. The *trends* for the two series are very similar and have almost the same value for the *slope* and *intercept*: this is represented by the right-hand-side of $S_L'(t)$ in Table 3. Also, the dominant cycle of 205 minutes is detected in both the series and it even starts at the same point for the two. There is a 100 % agreement between chief turning points. The only parameter, which has not matched so well, is the Variance Change, which shows a value of 907 for Reuters and 816 for Bridge. This difference could be attributed to the introduction of a small random noise in one of the two series, which might have caused *short-term* fluctuations at arbitrary positions in the series.

CONCLUSION:

Our initial finding is that there are striking similarities between the two time series either by using methods that assume stationarity (Method *II* and *III*) or by using methods that assume that the two time series are non-stationary (Method *IV* and *V*). This confirms that the provenance of the data provided by the two vendors is similar. It appears that a random noise (of a small magnitude) was added to or subtracted from the original data. This synthetically fabricated data was then released with a certain time delay.

The purpose of using formal time series analysis methods on sequential data is to learn "something" about the nature of the system generating the data. Future work would therefore focus on the analysis of this random noise and an attempt to replicate it. This would not only improve our understanding of the underlying phenomenon that govern the behaviour of non-stationary time series but it may also help in copyright protection and data security.

Appendix A

A NOTE ON VARIANCE CHANGE:

In many fields, such as the physical sciences and economics the hypothesis that a process is composed of many features which occur at different *scales* is quite natural. Recently, attention has been given to identifying and modelling the so-called long memory processes. Stationarity is described as a quality of a process in which the statistical parameters (mean, variance and standard deviation) of the process do not change with time. In practise, one may question if the process are truly stationary, or composed of several stationary segments. The stationarity of almost any statistical characteristic fails when applied to financial data. However, where the non-stationarities occur, interestingness begins, and with tools like the wavelet transform, capable of taming the non-stationarities, interesting *local* patterns can be discovered in the data. Whitcher et al (1998) showed that the time where a non-stationarity occurs coincides with the time at which a sudden shift in the variance takes place.

We have developed an algorithm that uses the DWT (employing *Daubechies six-coefficient* filter banks) to locate a variance change in a time series. \tilde{P} is called the normalized cumulative sum of squares (NCSS) index and is defined as

$$\tilde{P}_k = \frac{\sum_{t=L_j-1}^k \tilde{w}_{j,t}^2}{\sum_{t=L_j-1}^{N-1} \tilde{w}_{j,t}^2}, \quad k = L_j - 1, \dots, N - 2 \quad (3)$$

where N is the total number of samples, L is the length of the filter used for the DWT analysis (6 in our case), and W_j is the *level-j* DWT of the *volatility* of the original signal. Test statistic $\tilde{D} = \max(\tilde{D}^+, \tilde{D}^-)$ where

$$\tilde{D}^+ = \max_k \left(\frac{k - L_j + 2}{N - L_j} - \tilde{P}_k \right) \quad (4)$$

and

$$\tilde{D}^- = \max_k \left(\tilde{P}_k - \frac{k - L_j + 1}{N - 1} \right) \quad (5)$$

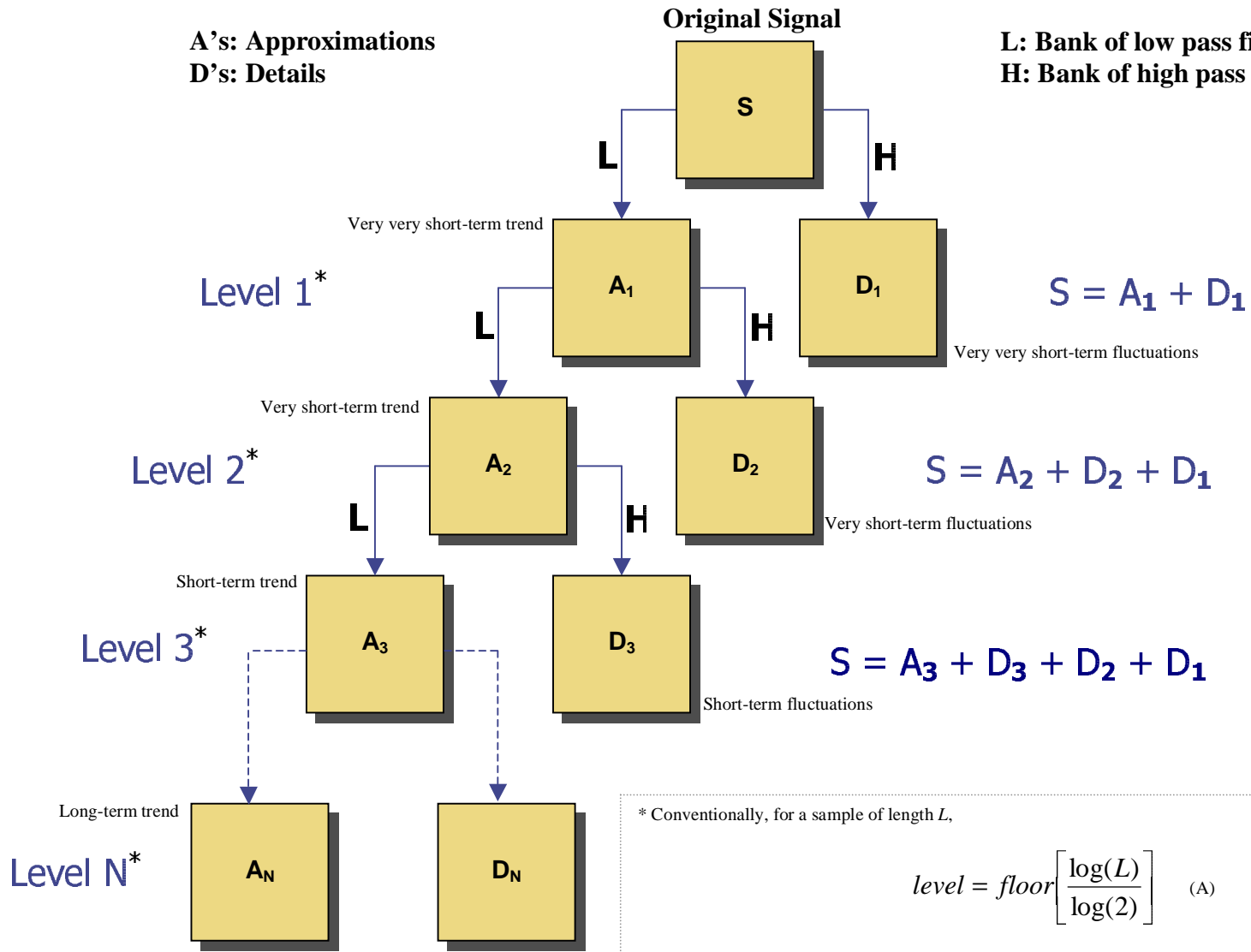
The location at which variance change occurs is \hat{k} which is the abscissa of the peak value of \tilde{D} . Also, \hat{k} is calculated on W_1 i.e., the *level-1* DWT of the *volatility* for a signal.

Discrete Wavelet Transform (DWT)

A's: Approximations
D's: Details

L: Bank of low pass filters
H: Bank of high pass filters

FREQUENCY



* Conventionally, for a sample of length L ,

$$level = \text{floor} \left[\frac{\log(L)}{\log(2)} \right] \quad (A)$$

The *floor* in Eq (A) ensures that the *level* remains an integer: if the result is in decimals, only the number to the left of the decimal point will be considered as the *level*.

Figure 4 Discrete Wavelet Transform showing three levels of Decompositions

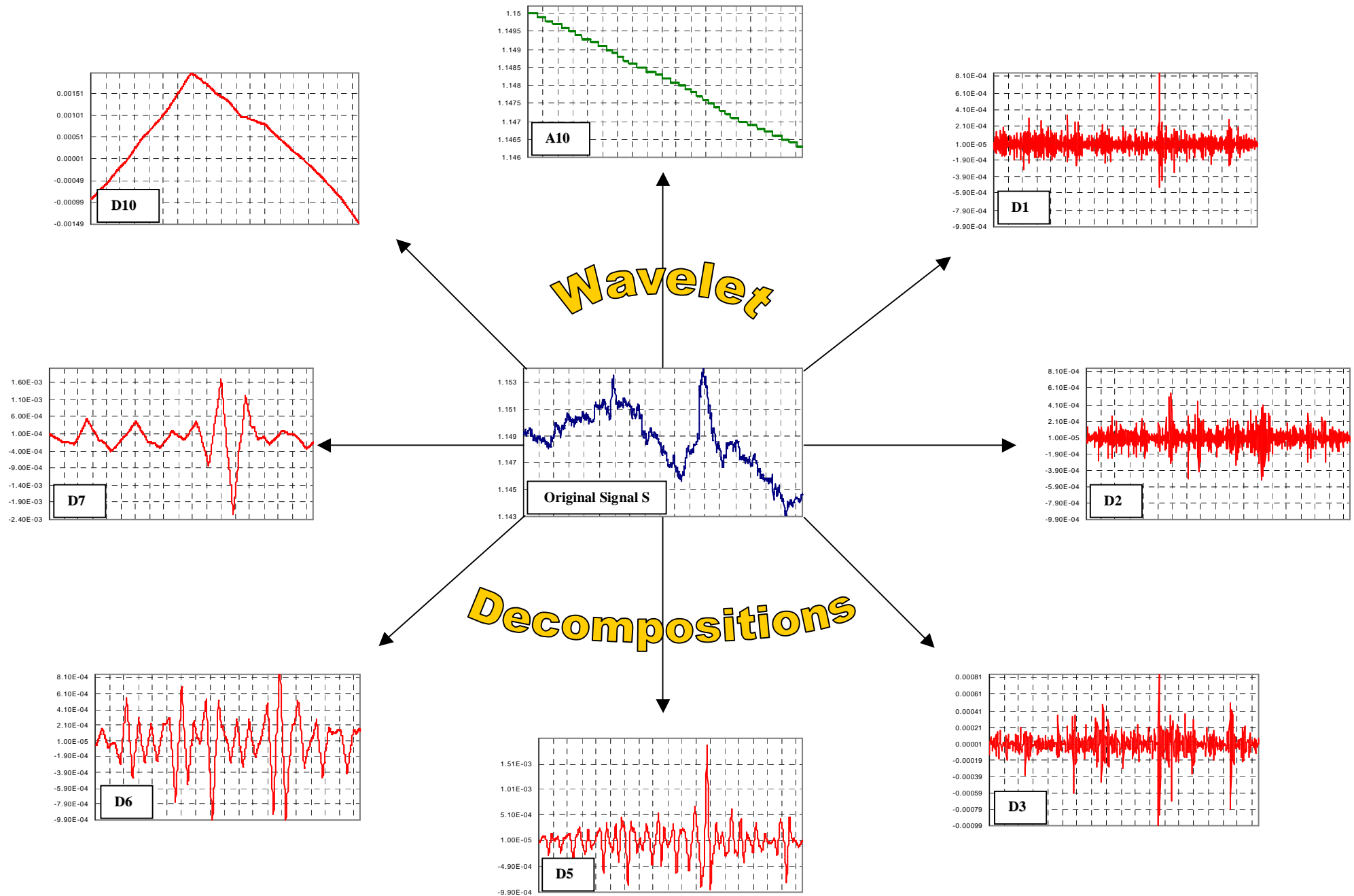


Figure 5 Reuters High Time Series (blue) and its Wavelet Details (red) and Approximation (green) for a level-10 Analysis