Financial Time Series: Market analysis techniques based on Matrix Profiles

Eoin Cartwright[1], Martin Crane[1], and Heather J. Ruskin[1]

Modelling & Scientific Computing group (ModSci)
School of Computing, Dublin City University
Dublin 9
Ireland
https://www.dcu.ie/computing

Abstract. The *Matrix Profile* (*MP*) algorithm has the potential to revolutionise many areas of data analysis. In this article several applications to financial time series are examined. Several approaches for the identification of similar behaviour patterns (or *motifs*) are proposed, illustrated and results discussed.
While the *MP* is primarily designed for single series analysis, it can also be applied to multi-variate financial series. It still permits initial identification of time periods with indicatively similar behaviour across individual market sectors and indexes, together with assessment of wider applications such as general market behaviour in times of financial crisis. In short, the *MP* algorithm offers considerable potential for detailed analysis, not only in terms of motif identification in financial time series, but also in terms of exploring the nature of underlying events.

Keywords: Financial Time Series · Matrix Profile · Time Series Motifs

1 Introduction

Time series *motifs* (repeated, matched or partially-matched sequences) occur both within and between individual time series[1]. Motif discovery is the task of extracting previously unknown recurrent patterns from such data-sets[2] with applications ranging from Music[3] to Seismology[4] and, of course to Finance, facilitating attempts to assess the importance of historical events and predict future trends.

In the financial domain a wide range of motif discovery approaches have been explored to date including that of Piecewise Aggregate Approximation (*PAA*)[5], used to investigate historical Standard and Poors *S&P500* index data. In addition, a *Motif Tracking Algorithm* was used to examine motifs in a West Texas intermediate (*WTI*) crude oil daily price time series (a popular indicator of oil prices in general)[6].

A Spatio-Temporal Pattern-Mining approach was also applied to the examination of company portfolios, where, for each company examined, this was taken to correspond to a moving trajectory over a two-dimensional financial grid (for discretized size and *price-to-book* ratio)[7]. A set of similar financial trajectories taken over the same time period was then considered to be a motif. For a more detailed review of currently available Motif Discovery and Evaluation techniques for Financial applications, see e.g. [8]

A new data construct based upon an efficient *Nearest Neighbours* discovery method and designated the *Matrix Profile* (*MP*)[10] has significant advantages over other time series data mining techniques, offering considerable flexibility in application. Here we investigate its potential to offer additional insight on financial series analysis over different timescales and scenarios.

To demonstrate relevance, *MP* plots are used to identify similar patterns (*motifs*) within a single series. The impact, on plot evolution of increasing *motif length*, is also examined, where this can indicate persistence of given behaviour over longer timescales. Additionally, histogram plots of *MP* data can illustrate whether the proportion of matches (*motifs*) or mismatches (*discords*) is greater for a given financial time series.

The examination of multi-dimensional *MP* plots for localised minima allows the combination of different measures for a single financial series to be explored. Additionally, periods of similar behaviour both within and across market sectors can be demonstrated in representative time series, while individual stocks contributing to a given index can also be investigated. *MP* use is illustrated for the financial crisis period, January 2007 to January 2009, and verified against the raw series.

2 The Matrix Profile: An Introduction

The *Matrix Profile* (*MP*) is a novel algorithm (proposed by the Keogh research group) that has proven useful for numerous data mining and time series analysis tasks[11]. As the *MP* is highly scalable for time series *sub-sequence all-pairs-similarity* search[10], it efficiently identifies time series motifs and discords (i.e. mis-matches). Thus the examination of *MP* plots can aid interpretation of distinctive or recurring patterns in financial time series.

The main advantages (amongst others) of the *MP* algorithm are that it

- Returns an exact solution for motif discovery
- Requires only one input parameter (sub-sequence length m)
  - For example, a similarity/distance threshold does not need to be specified (unlike for many other similar algorithms)
- Has a time complexity that is constant in sub-sequence length
  - Thus it can be constructed in a deterministic timeframe, an important consideration for time sensitive financial applications
- Incorporates flexibility
  - No assumptions are made about the underlying data
  - Is incrementally maintainable

For an input time series with a given sub-sequence length m the *MP* returns four results. These are:

Matrix Profile Index (MPI)

– For every index i or time point in the examined series the MPI contains a pointer to another index j (in the original series) indicating the start location of the nearest neighbour sub-sequence (or similar behaviour pattern)

Matrix Profile (MP)

– For every index i in the examined series, a record of the Z-normalized Euclidean distance to the nearest neighbour sequence (as indicated by the MPI) **Note:** Zero distance implies exact match

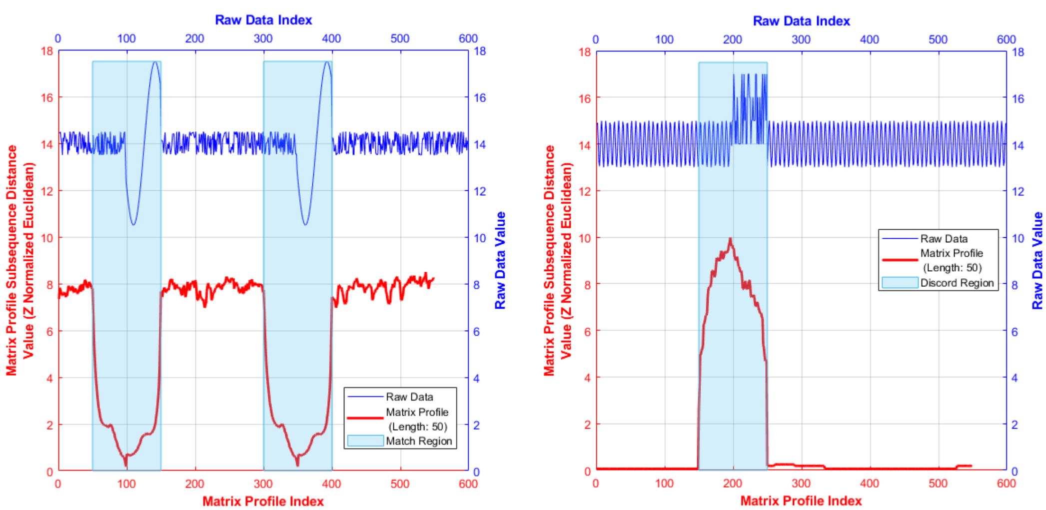
Motif Index (M_i)

– For the given series, M_i records the start location index of the sub-sequence that has the *lowest* sub-sequence distance value of MP i.e. closest match in terms of distance or 'classical' time series motif

Discord Index (D_i)

– D_i records the start location index of the sub-sequence that has the *highest* sub-sequence distance value of MP i.e. poorest match in terms of distance or 'classical' time series discord

A sample MP plot (*red* line) based on a synthetic input series (*blue* line) is shown in Figure 1. Illustrated is a MP with (a) a *matching region*, i.e. low MP distance values and (b) a *mis-match region* corresponding to high MP distance values. One important feature of the MP utilised in the following analysis is that *exact matches* of content are not necessary to obtain meaningful results, as a localised MP minimum value can be used to identify a close match even if the MP distance value considered is non zero.



(a) MP with Motif region (b) MP with Discord region

Fig. 1: Sample synthetic Matrix Profiles [11]

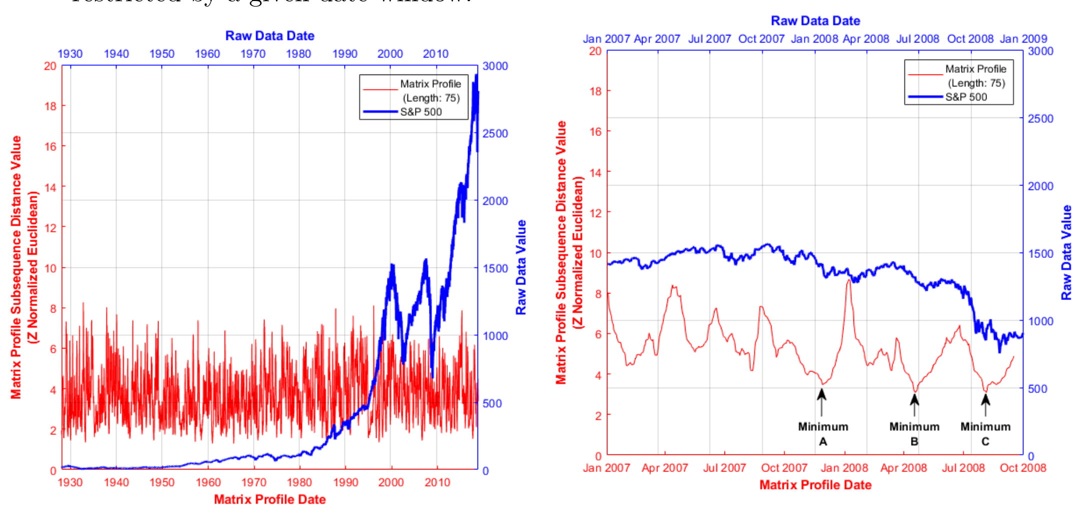
3 Matrix Profile Analysis of Financial Time Series

When investigating the Matrix Profile of a financial times series, a typical focus is on regions (as highlighted by lower *MP* distance values) indicating similar behaviour at some other point in the data series, as financial markets show evidence of auto-regression[13]. The nature of this behaviour can be characterised by shorter or longer sub-sequences or by common 'shapes', indicative of standard financial features of the original series. Examples include *Pennant*, consisting of significant rise or fall in the series followed by a period of consolidation and the *Triple Bottom* which occurs when the reduction in series values creates three distinct troughs, at around the same price level, before breaking out and then reversing the trend[14,15,16].

Constructing a sub-sequence of length m (to create the given *MP*) and starting at the index value indicated by the lowest *MP* distance value (i.e. the closest match), it is possible to explore whether similar regions occur at regular intervals or can be associated with external events such as, for example, a *FED* rate announcement.

3.1 Single Series motif identification

Financial data are inherently noisy however, so the *MP* interpretation is inevitably affected to some degree[12]. Figure 2a shows a *MP* for the full *S&P500* time series (available at time of writing [17]) labelled by both date and original series index, while Figure 2b shows a subset of the original *S&P500* series restricted by a given date window.



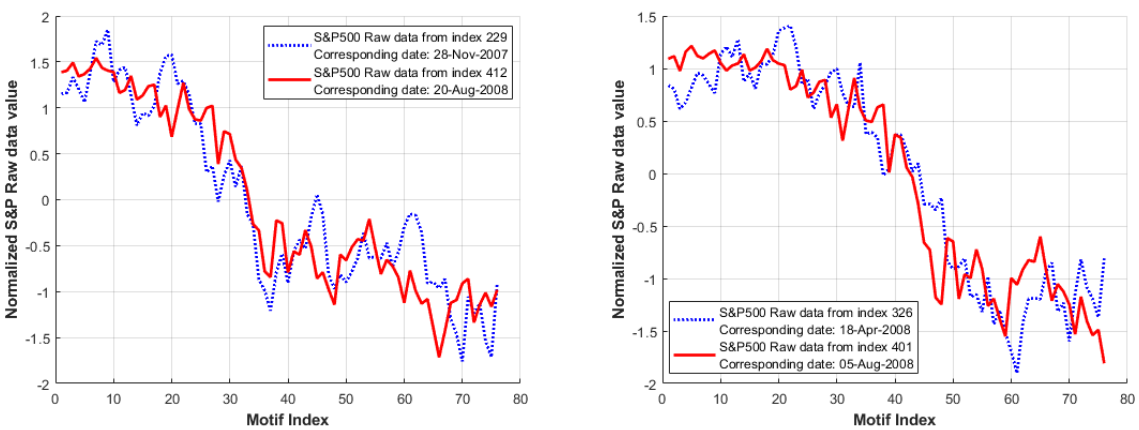
(a) January 1928 to March 2019

(b) January 2007 to January 2009. Further *MP* minima location detail is contained in Table 1

Fig. 2: *S&P500* series and associated *MP* distance values

Figure 2b thus shows *MP* patterns illustrated in greater detail, facilitating relation of these patterns to market conditions occurring within the given time-frame. The window chosen and used for further analysis reflects the considerable stress experienced in the global marketplace at this time[18] corresponding to initial confidence issues in the American sub-prime property market. This sparked a global liquidity crisis[19] that caused many financial institutions to collapse and triggered large systemic interventions in the form of bailouts from both government and global financial institutions, such as the *IMF* in order to re-establish system stability.

Figure 2b (*red* series) illustrates three points of interest highlighted as points ***A***, ***B*** and ***C*** (with further detail in Table 1). Low *MP* values indicate similar behaviour of the *S&P500* index (*blue* series) at some other point in the time window examined (obtained from the corresponding *MPI*). Thus, *MP* plots can highlight behavioural similarities which can be less obvious from the raw series data.



(a) Raw data of *S&P500* series identified by low *MP* value location ***A*** (Index 229) in Figure 2b

(b) Raw data of *S&P500* series identified by low *MP* value location ***B*** (Index 326) in Figure 2b

Fig. 3: Raw data of *S&P500* series indicated as motif locations by low *MP* values in Figure 2b & Table 1. Here the blue series indicates the sub-sequence identified visually from low *MP* values while the red sub-sequence represents the nearest 'match' as indicated by the corresponding *MPI* value

To demonstrate more fully, Figure 3 indicates typical motifs obtained from the raw *S&P500* series (as indicated by the *MP* and *MPI* values of Figure 2b & Table 1). These are constructed by the generation of a sub-sequence of *MP* length ($m = 75$) to facilitate display of longer term sub-sequences within the length bounds enforced by the *MP* algorithm (minimum and maximum constraints relative to the series length apply). An initial sub-sequence from the start index of the minimal *MP* distance value (visual inspection) is compared with a second sub-sequence, which starts at the nearest 'matching index' as indicated by the corresponding *MPI* value (Table 1).

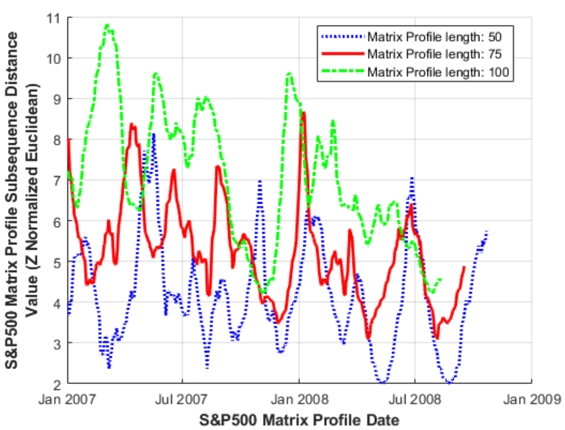
Note that although several local *MP* minima locations are identified in Figure 2b, only two raw data sequences are displayed in Figure 3 as, in this particular case, the remaining localised *MP* minima form a 'classic' motif (i.e. closest match in terms of distance). This can be seen in Table 1, where for minima locations ***B*** and ***C*** the *MPI* values are reversed i.e. marking the same sub-sequence.

Local *MP* Minima location	Identified *MP* Minima Index	Identified *MP* Minima Date	*MPI* Value of Identified Index	*MPI* Date of Identified Index
A	229	28 November 2007	412	20 August 2008
B	326	18 April 2008	401	05 August 2008
C	401	05 August 2008	326	18 April 2008

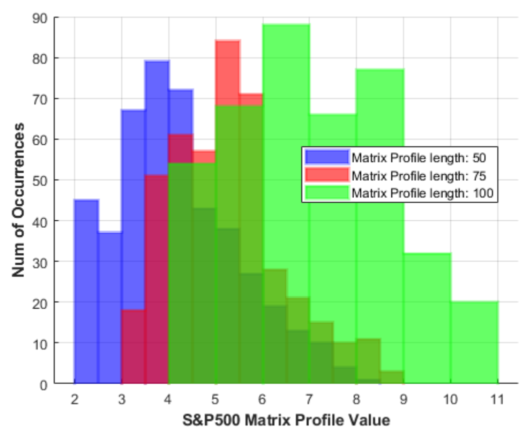
Table 1: *MP* minima details of reduced *S&P500* series as highlighted in Figure 2b. *Matrix Profile Index* (*MPI*) values i.e. location of matching index are also shown

3.2 Single series *MP* evolution over length

As the *MP* sub-sequence length increases, the average *MP* distance value for that sub-sequence length also appears to increase, indicating a less-exact match (in terms of average Z-normalized Euclidean distance) over the entire length of the *MP* (Figure 4). This result is intuitive, as the shorter the sub-sequence length the more readily it is matched[20].



(a) *MP* of *S&P500* series



(b) Histogram of *MP* data of *S&P500* series

Fig. 4: *MP* and Histogram of the *S&P500* series over increasing sub-sequence lengths. January 2007 to January 2009

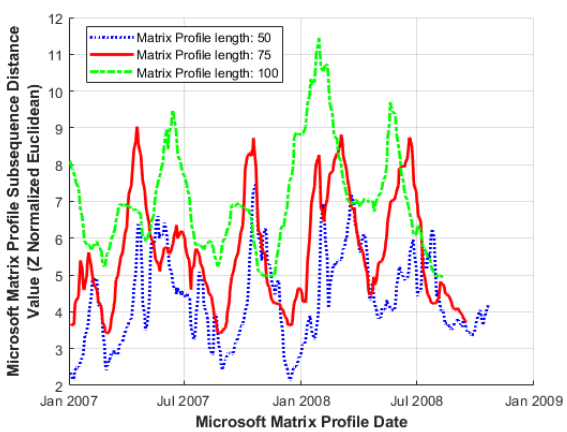
As *MP* sub-sequence length is increased, the frequencies of *MP* motif match and discord values correspondingly decrease, (Figure 4b). However, where found these large *MP* distance values (occurring at approximately the same index in *MP* plots of shorter and longer sub-sequence lengths) may indicate the existence of longer term trends in the data, despite more volatile behaviour observed at shorter *MP* sub-sequence lengths.

It should be noted that increase in *MP* sub-sequence length does not necessarily result in a clearer, 'less noisy' *MP* structure (particularly for the multivariate cases examined, Sections 3.3 and 3.4) in individual series. Hence both a range of sub-sequence lengths and *MP* distance minima are needed for balanced interpretation.

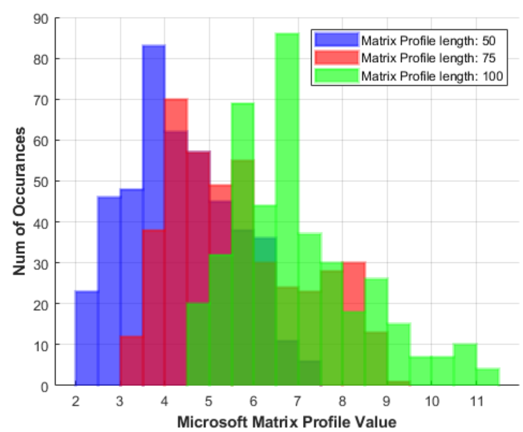
This is further illustrated by a histogram plot of the same *MP* data in Figure 4b. The entire histogram (of overall distance to repeats) is shifted to the right (for given sub-sequence length). The global behaviour of the *MP* can be linked to the distributional morphing. The shorter *MP* length (of 50) here with higher frequency of occurrence of matches/discords, is closer to the Normal (or Gaussian) form. For higher sub-sequence length (of 100 here) the distribution is flatter, indicating larger variation in motif and discord distance values. However, an examination of detailed motif shape in these longer sequences may prevent over-reliance on short term volatility, while capturing longer term patterns of growth or stability, with corresponding reduction in transaction costs.

The *MP* distance histogram also highlights the fat-tailed distribution of many financial market series data, (where a right-skew indicates a higher proportion of discords and a left-skew a higher proportion of motifs). Figure 4a shows *MP* line series plots of increasing sub-sequence length while part 4b illustrates their corresponding histogram values. These plots are based upon *S&P500* share value data, again for the time window of January 2007 to January 2009.

The same generalised behaviour as for Fig. 4 is observed in Figure 5, however in this case with a higher proportion of increased *MP* distance values (or discords) as indicated by a skewed distribution to the right. This occurs for all *MP* sub-sequence lengths examined so far and indicates that series behaviour is consistent over longer timescales.



(a) Microsoft *MP* series

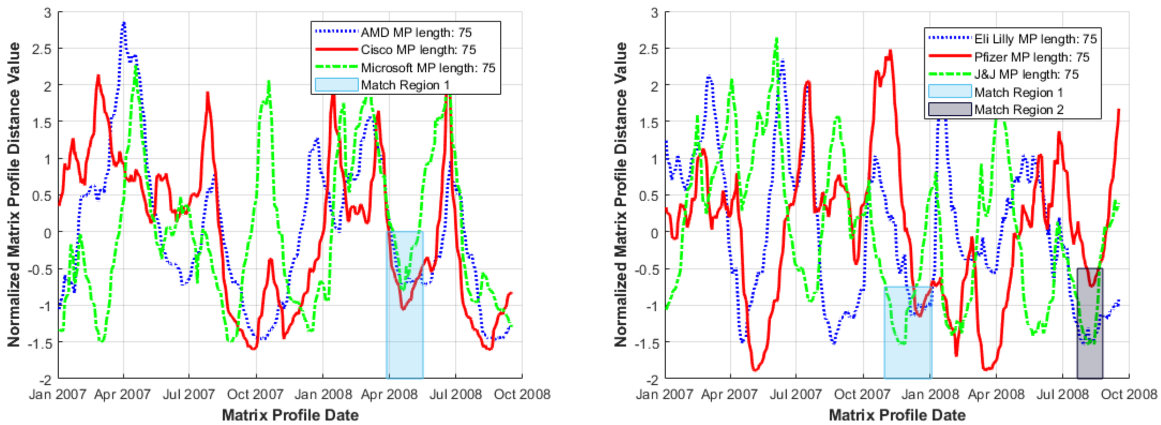


(b) Histogram of Microsoft *MP* series

Fig. 5: *MP* and Histogram of the Microsoft series over increasing sub-sequence lengths. January 2007 to January 2009

3.3 Multi-variate Series

In attempting to characterise wider market behaviour the *MP* single-series approach must be expanded to multi-variate series. Applications for finance include investigation of multiple companies within the same market sector, as opposed to an individual stock or index considered independently.



(a) Normalized *MP* series of sample Tech companies (b) Normalized *MP* series of Pharmaceutical companies

Fig. 6: Sample set of normalized Matrix Profiles across individual market sectors, local *MP* minima coherence indicated by coloured rectangles. 2007 to 2009

Single Sector Figure 6 illustrates the *MP* plot of stock series for influential companies within (a) Technology and (b) Pharmaceutical sectors chosen at random from several top 10 lists based on Market Cap, percentage annual return and market value[21,22]. Although fluctuations in amplitude are large, coherent movements at lower *MP* distances are observed over short time-frames, (i.e. local minima regions correspond across series).

Clearly, both the occurrence and values of these local *MP* minima over the shortest timeframe are of interest for motif identification and verification. The main considerations are i) the time duration to when similar behaviour is repeated (i.e. *when* a match occurs) and ii) distance range (indicating how close a match it is). Thus a visual choice of the point at which a generalised local minima region occurs in a multi-variate *MP* series plot is made based upon obtaining the best combination of local minima over the shortest timeframe and restricting *MP* minima spread to be as low as possible. We consider these to be match regions as highlighted by shaded areas in Figures 6a and 6b for example.

Occurrence of a motif within an identified match region may be shifted slightly from series to series, either with respect to starting index or by extension, date. In consequence, plots can be constructed to start at a specific index (where a given series feature may overlap slightly with a similar or matching feature in another series) or at a specific date, where shifts between series may be clarified.

It should be noted that, due to total *MP* series variance and the fact that areas of interest are small compared to the overall plot size involved, visual *MP* distance plot analysis is a limited technique. These plots become harder to interpret and sectors of interest more challenging to identify as multiple series are added, so that typically a small series set only is examined. However, consistent behaviours such as reduced volatility, less precise matching (increased *MP* distance) and better-defined *MP* structure are observed generally for long as well as shorter sub-sequence lengths.

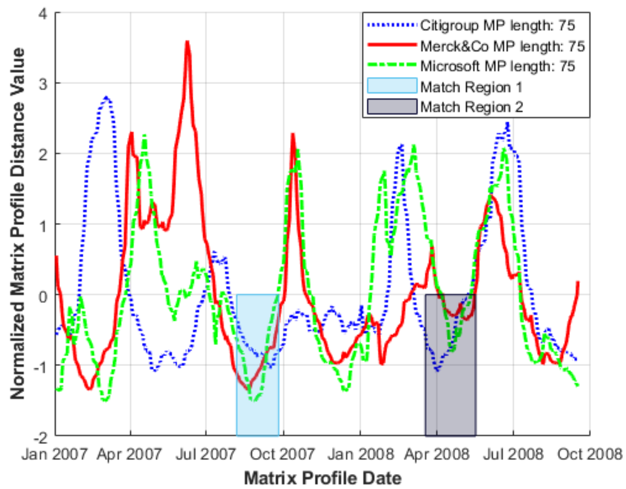


Fig. 7: Sample set of normalized Matrix Profiles across multiple market sectors, local *MP* minima coherence indicated by coloured rectangles. January 2007 to January 2009

Multi Sector Expanding the approach to multiple sectors (including indexes) can be useful in illustrating more generalised market behaviour where, for example, large events such as global shocks can generate coherence that is reflected in the behaviour of the corresponding *MPs*. To illustrate this, a range of leading sectoral companies were chosen, again from several top 10 lists based on Market Cap, percentage annual return and market value. These sectors span Information Technology (*Microsoft*), the Pharmaceutical industry (*Merck&Co*) and the Finance sector (*Citigroup*)[21,22,23]. *MP* line plots, constructed for the same time window of January 2007 to January 2009 are shown in Figure 7, together with coincident local *MP* minima that occur within narrower time intervals (shaded match regions).

3.4 Stocks within an Index

Matrix Profile plots are also useful in examining the influence of individual stock series on the index to which these contribute. Comparison of *MP* index series against several *MP* plots of individual companies (chosen to cover a wide range of sectors trading within that index) serves to characterise convergence of lower *MP* distance values (Figure 8).

Within the time window examined, short periods occur where localised *MP* minima coincide with those of the *S&P500* suggesting coherent behaviour; (for raw data analysis see section 3.5). Table 2, moreover, shows the shift in *location* (and by extension *timing*) of *MP* minima occurrence within these series.

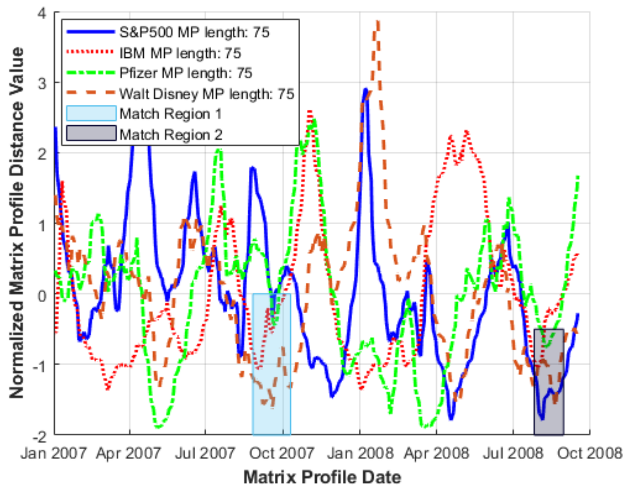


Fig. 8: Multi-Sector *MP* plots including S&P500 index, local *MP* minima coherence indicated by coloured rectangles. January 2007 to January 2009. Dates of occurrence of low minima regions in Fig. 8 are summarised in Table 2

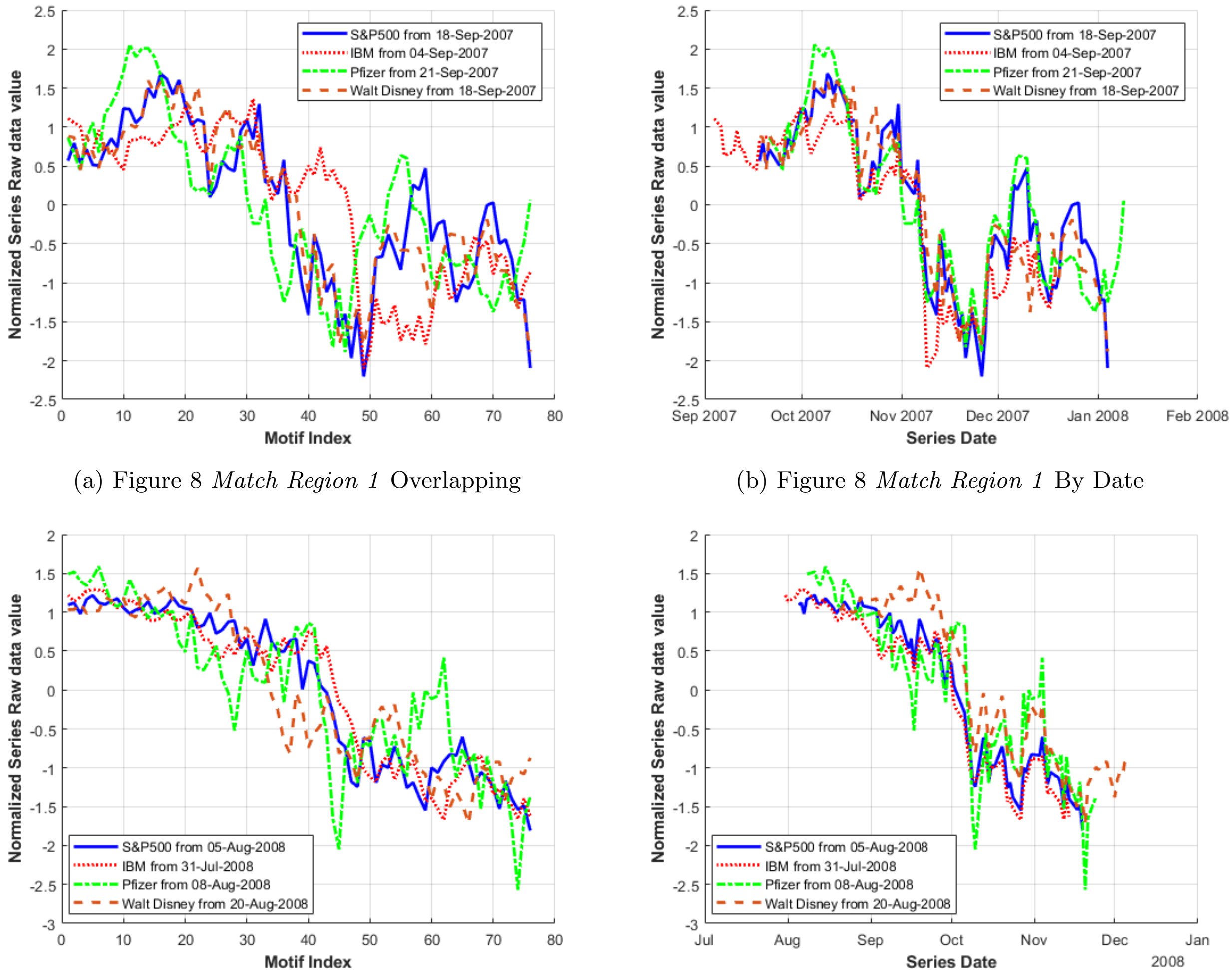
Series	**Sector**	**Match Region 1**		**Match Region 2**	
		Identified *MP* Minima Index	**Identified *MP* Minima Date**	**Identified *MP* Minima Index**	**Identified *MP* Minima Date**
S&P500	Various	179	18 September 2007	401	05 August 2008
IBM	Information Technology	169	04 September 2007	398	31 July 2008
Pfizer	Pharmaceutical	182	21 September 2007	404	08 August 2008
Walt Disney	Entertainment	179	18 September 2007	412	20 August 2008

Table 2: Identified *MP* minima dates and indexes of match regions 1 & 2 (i.e. localised *MP* minima coherence) as highlighted in Figure 8

For some series the *MP* minima occur before that of the *S&P500* minima indicating a leading influence upon the index while others are identified shortly afterwards indicating that underlying series subsequently reflect index movement. Only *three* sub-series are currently included of course so, given that other stock series may be influential, a comprehensive analysis would need to consider additional index components and combinations thereof.

3.5 Reviewing the raw data

In the multi-variate cases examined thus far, a low *MP* value at approximately the same index as for multiple series is taken to be a good indication of similar behaviour. Strictly, however, the *MP* algorithm in its current form examines each series independently so that an extreme *MP* value may indicate either a close match (*motif*) or mismatch (*discord*) within a *single* series. For example, series X and Y may both have a low *MP* value coinciding at index x indicating two matches (one within each series) but these are independent, so that event type *motif shapes* may differ. *MP* plots for several series indicate regions of *possible* consistency, so for real behaviour to be characterised event types in the raw series must be related to *MP* matches.



(a) Figure 8 *Match Region 1* Overlapping

(b) Figure 8 *Match Region 1* By Date

(c) Figure 8 *Match Region 2* Overlapping

(d) Figure 8 *Match Region 1* By Date

Fig. 9: Motif of Stocks within an Index. i.e. original data sub-sequences with starting indexes obtained from *MP* minima located in *Match Regions 1* & *2* in Figure 8

A motif as a repeated identifiable sub-sequence has a minimum of two parts, namely the initial sequence (as indicated by the index of the localised *MP* minimum) and the corresponding *matching* sequence obtained from the *MPI* (indicating the start point of the nearest neighbour sub-sequence). Figure 3b illustrates the two parts of a sample classic motif of the *S&P500* series found by locating low *MP* distance values in the time window of January 2007 and January 2009. In the multi-variate case considered here, one subsection (or motif part) per series only is shown for clarity.

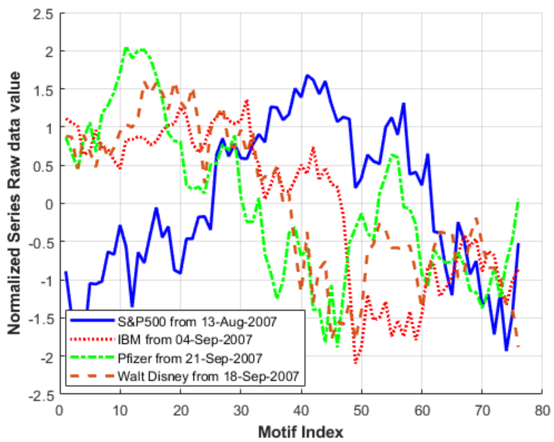
The two complementary approaches of the analyses consider 1) Nature of the behaviour of the sub-sequences (indicated by shape) i.e. *Event Type*, and 2) *Timing*. Of interest with respect to 1) for a set of sub-sequences considered in isolation is whether such events match in terms of length, magnitude and location. Alternatively sub-sequences may exhibit amplification or damping over an extended period. In terms of 2), interest centres on whether a motif sub-sequence leads, lags or coincides with other sub-sequences in terms of event *timing*.

Underlying motif sub-sequences in the original series of the *MP* plots (Figure 8) exhibit localised *MP* minima of index-contributing stocks across multiple market sectors. In Figures 9a and 9c the motif sequences for each series are plotted according to the *motif* sub-sequence index (i.e. overlapping). Again, illustrative of similar behaviour (in terms of shape), a large drop in value occurs approximately halfway through each of the motif sub-sequences. In Figure 9a it initially appears that both the *IBM* (red) and *Pfizer* (green) series are reacting at a later point in time to the *S&P500* (blue series). However, when plotted according to date (Figure 9b) it can be seen that the large drop in value actually occurs over the same time window of November 6$^{\text{th}}$ to 12$^{\text{th}}$ 2007 for all series.

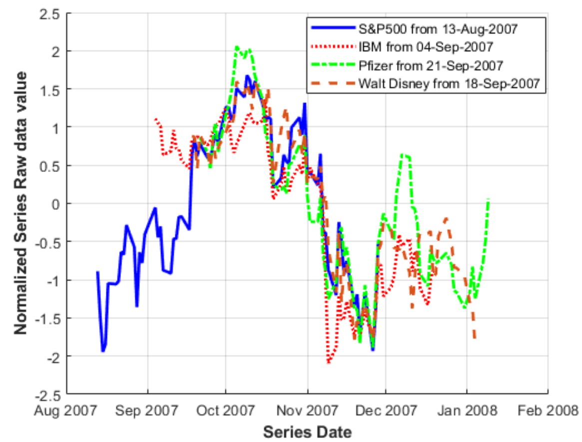
To place this in context, this corresponds to a period when a deepening liquidity crises sparked by issues in the American sub-prime property market[24] began to accelerate globally (as illustrated by the run on the Northern Rock bank in England in September 2007). Despite initial action by the *FED* over 2007 to increase liquidity in short-term money markets through larger open market operation interventions (as described [25]), the peak of market values was reached in October 2007. However fears of losses at *Citigroup* in combination with poor market sentiment prompted a more generalised sell-off (as reflected in Figures 9a and 9b).

Similar behaviour is observed in Figures 9c and 9d, in this case with the Pfizer and Walt Disney (brown) series reacting slightly after the *S&P500* (over the period of October 1$^{\text{st}}$ to 10$^{\text{th}}$ 2008). This corresponds to the US Congress opening its first hearing on the growing financial crisis when stocks then tumbled further (the *Dow Jones* index dropped below 10,000 for the first time in 4 years[26]) coinciding with the realisation by investors that the credit crisis was spreading around the globe and the recent (September 29$^{\text{th}}$) rejection by US Congress of a proposed $700bn bailout plan would not stabilise the situation. However, as the country's financial system continued to deteriorate, several representatives changed their minds and the legislation was signed off on October 3$^{\text{rd}}$ 2008[27].

Overall coherent behaviour is observed for the *S&P500* series and individual stocks, particularly when plotted by date (as initial lag between series is no longer evident).



(a) Overlapping



(b) By Date

Fig. 10: Motif of Stocks within an Index. i.e. original data sub-sequences with starting indexes obtained from *MP* minima in Figure 8 *Match Region 1* (using an alternative *S&P500* index)

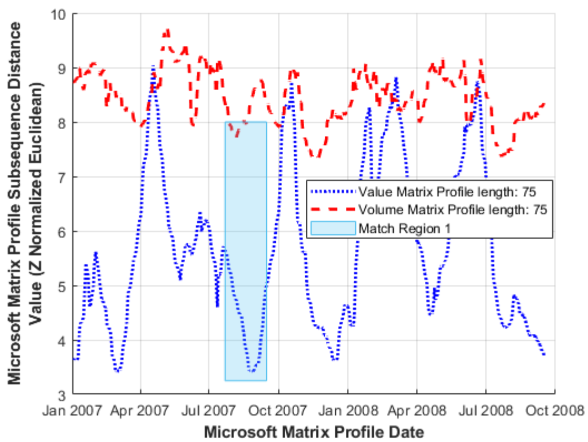
When examining Figure 8 to identify suitable lowest*MP* minima match regions an alternative lower index value of the *S&P500 MP* than that chosen for *Match Region 1* is also available. This gives a reduced *MP* value (i.e. a closer match in terms of *Euclidean* distance to some other point in the *S&P500* series). Incorporating this alternative *S&P500 MP* minima value (154) occurring on 13$^{\text{th}}$ August 2007 into Figures 9a and 9b gives the plots displayed in Figure 10. Plotting according to date (Figure 10b) the *S&P500* series corresponds quite well with the remaining series in the region where dates overlap. Figure 10a illustrates that *Event Type* (when considered as motif shape) does differ significantly between the series in question.

3.6 Multidimensional analysis of a single Stock

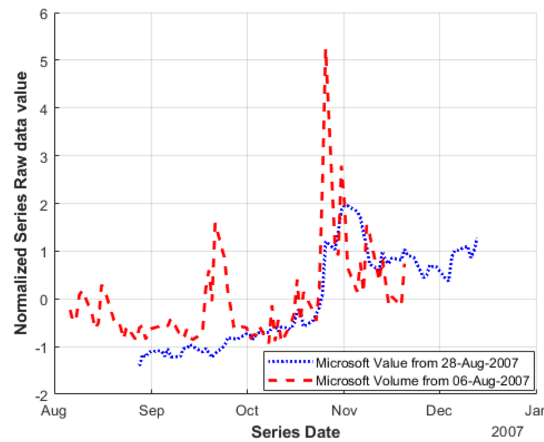
In addition to utilising the *MP* for multi-variate analysis of separate series spanning differing market sectors, the approach can also be applied to the combination of series based upon different measures of a single company or index.

In Figure 11a the *MP* in two measures of Microsoft stock (*value* & *volume*) are illustrated (again for the time window of January 2007 to January 2009)[28]. A match region (co-incidence of *MP* minima) is identified while raw data subsequence values shown in Figure 11b appear to indicate a large increase in both series occurring at approximate dates (October 26$^{\text{th}}$ 2007 for *volume* and 1$^{\text{st}}$ November for share *value*).

Although both series are based upon the same stock, the previous flexibility to display raw data sub-sequences by date of the identified *MP* minima still applies to features identified in both series (in this case applying to when these occur). Here, it illustrates timing of occurrence of features identified in one series relative to another. Figure 11b highlights reasonable alignment for increase in both share value and trading volume.



(a) *MP* Microsoft share volume and value



(b) Raw data Microsoft share volume and value based upon *MP* minima

Fig. 11: Motif of differing measures (Value and Volume) of Microsoft stocks. January 2007 to January 2009

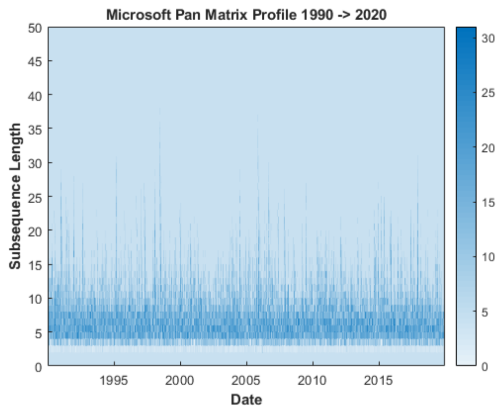
Examination of other combinations of commonly-used company measures such as *price-to-book* and *price-to-earnings* ratios is also possible.

3.7 Motif Length Selection Considerations & Long Vs Short Term behaviour

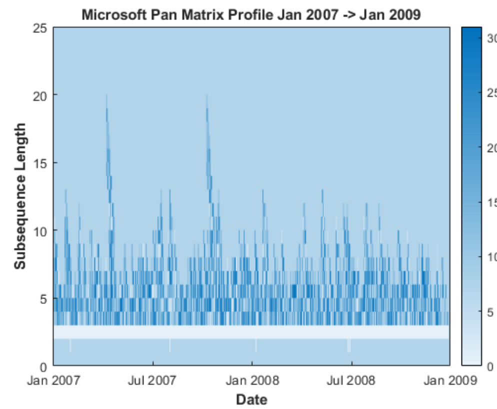
An important consideration for selection of the motif or sub-sequence length for analysis is whether interest is focused on short or long-term behaviour (shorter or longer motif lengths respectively). The large number of motif locations found for shorter *MP* lengths can obscure particular trends, while the reduced number of motifs returned for longer lengths can facilitate identification of extended match regions. Recent developments on the length selection process providing an illustration of the motif content (by *MP* length) include the *SKIMP*[29] algorithm.

SKIMP allows optimised generation of a set of *MPs* for a user-provided length range. The new structure, known as a *Pan Matrix Profile*(*PMP*), can be plotted as a heat-map indicating both the *location* and *length* of motifs in a data set, as illustrated in Figure 12a. Larger motif length locations are indicated by spikes while more frequent motif lengths correspond to areas of increased intensity. *PMP* plots can also provide an indication on common features of financial time-series i.e. may contain a large number of smaller length motifs even over a varying time window as shown in Figures 12a and 12b. This suggests a shorter *MP* length may be more applicable for financial series analysis.

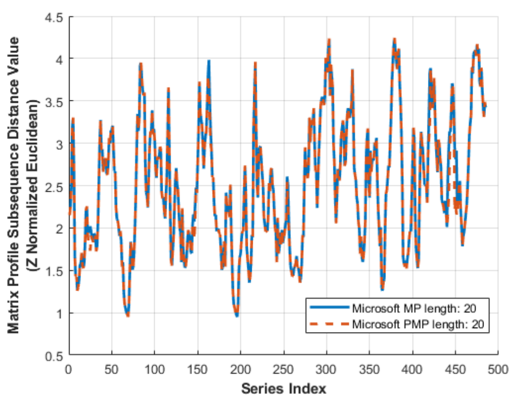
Thus a *PMP* can provide an alternative method when obtaining start locations for Motif behaviour investigations over reduced timescales, important as *MP* plots can become noisy at lower sub-sequence lengths (particularly in the multi-variate case). To illustrate this (within a single series initially), a motif length of 20 was chosen from Figure 12b as a suitable length for probing underlying raw series behaviour.



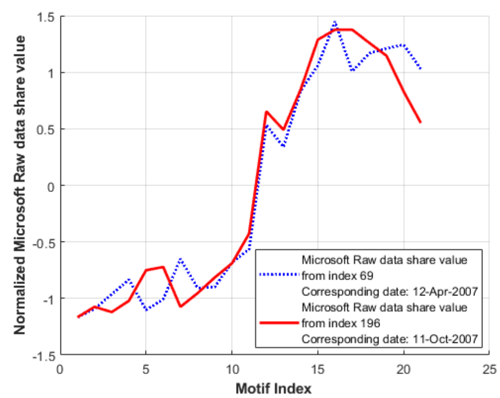
(a) Microsoft Pan Matrix Profile 1990 To 2020



(b) Microsoft Pan Matrix Profile 2007 To 2009



(c) Microsoft Pan Matrix Profile vs standard Matrix Profile



(d) Raw data of *Microsoft* series identified by index of peaks in the Microsoft Pan Matrix Profile (Figure 12b) and corresponding index of low *MP* value locations (Figure 12c)

Fig. 12: Microsoft Pan Matrix Profile (*PMP*) and underlying motif identification

A comparison between the standard *MP* and *PMP* is shown for the given sub-sequence length in Figure 12c illustrating close correlation (as anticipated). The location of peaks within the *PMP* plot (indicating motifs of greater sub-sequence length), are identified by the index which corresponds to localised MP minima values in Figure 12c. The underlying raw data sequences are isolated based upon these indexes and are displayed in Figure 12d. In this case the two locations correspond to the 'classical' motif as the *MPI* indexes refer to each other.

Expanding this approach for the multi-variate case, the same scenario (and individual series) of stock behaviour within an index was considered. Using an initial *S&P500 Pan Matrix Profile* plot (Figure 13a) a sub-sequence length of 13 was chosen for further analysis. For the *S&P500*, indices 134 and 246 exhibit peaks, corresponding to motifs of above average length. These are taken as approximate start locations for finding *MP* minima within the individual *matrix profiles* (Figure 13b). The alternative of only examining *matrix profiles* for low *MP* minima occurrence was not adopted as they become too noisy at this low sub-sequence resolution. Thus the indexes chosen from the *PMP* serve as regions previously considered as local match regions (Section 3.3) when examining the corresponding *MP* plots generated for this sub-sequence length (Figures 13c, 13d).

Figure 13b displays the full set of *MP* plots for these series (within the time window examined) with match regions centred on these indexes highlighted. Figure 13b also serves to illustrate further the noisy nature of financial *MP* plots at lower sub-sequence lengths, particularly for the multivariate case as here. For clarity, sub sections of Figure 13b illustrating the localised *MP* minima for match regions 1 and 2 are displayed in Figures 13c & 13d with the identified minima indexes and corresponding dates shown in Table 3.

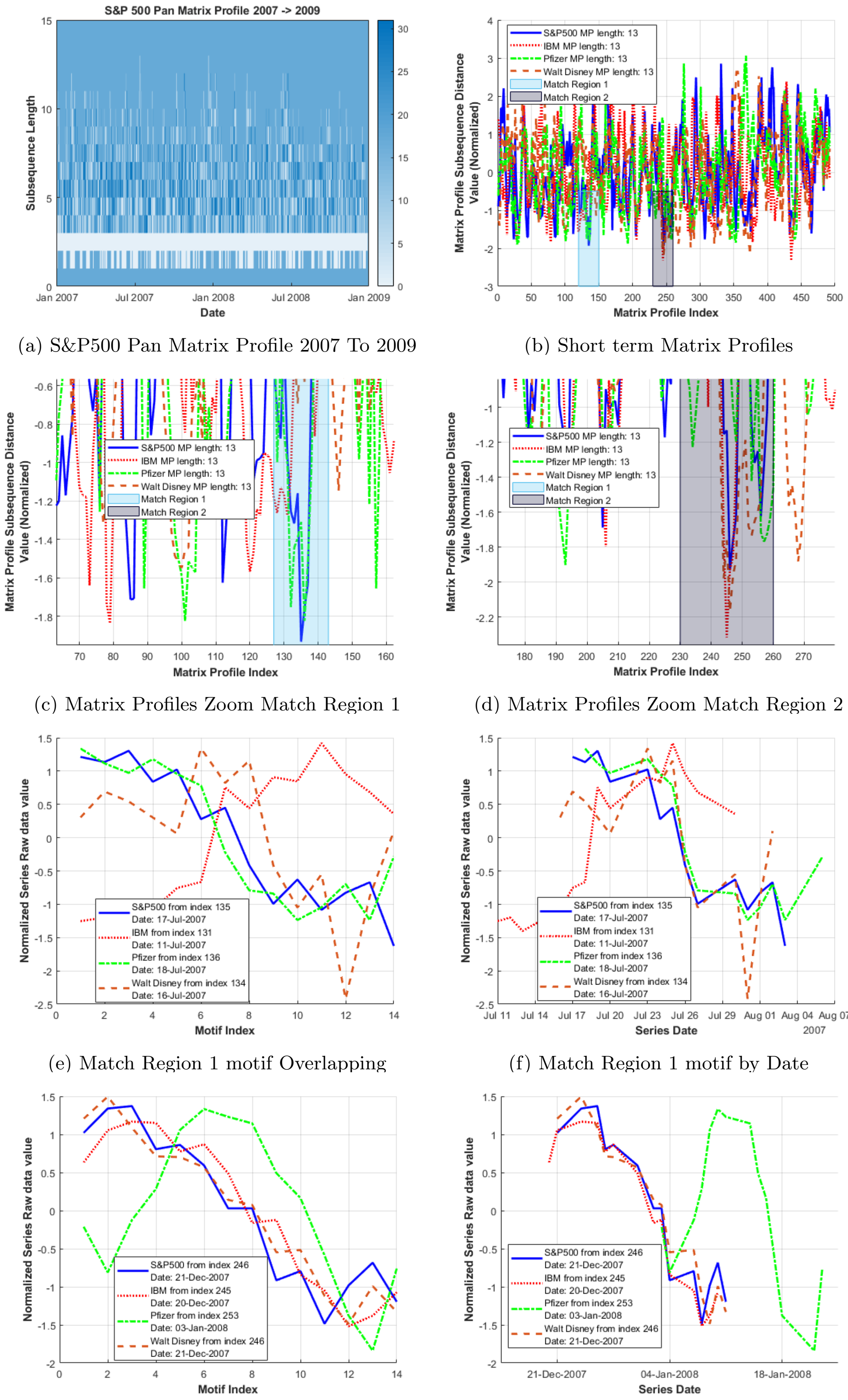
Series	**Sector**	**Match Region 1**		**Match Region 2**	
		Identified *MP* Minima Index	**Identified *MP* Minima Date**	**Identified *MP* Minima Index**	**Identified *MP* Minima Date**
S&P500	Various	135	17 July 2007	246	21 December 2007
IBM	Information Technology	131	11 July 2007	245	20 December 2007
Pfizer	Pharmaceutical	136	18 July 2007	253	03 January 2008
Walt Disney	Entertainment	134	16 July 2007	246	21 December 2007

Table 3: Identified *MP* minima dates and indexes of proposed match regions 1 & 2 as identified from *PMP* plot (Figure 13a) and highlighted in Figure 13b

For region 1, when plotted according to sub-sequence index (Figure 13e), independent raw data sub-sequences are not in particularly good agreement. However, when plotted according to date (Figure 13f) basic behaviour is similar for all series although the sharp reduction in value from 25$^{\text{th}}$ to 27$^{\text{th}}$ July 2007 is not as pronounced for *IBM*.

For region 2 raw data sub-sequence shapes appear to correspond quite well when plotted according to sub-sequence index (Figure 13g). However, when plotted by date in this case (Figure 13h) the *Pfizer* series briefly demonstrates coherent behaviour but, in general, lags relative to the other series.

Figure 13h also illustrates that the *MP* minima location has a disproportionately greater effect at these lower resolutions causing a larger shift (relative to motif length) as seen previously in section 3.5 for example. Further when plotting by date, there is less likelihood of an overlap region.



(a) S&P500 Pan Matrix Profile 2007 To 2009

(b) Short term Matrix Profiles

(c) Matrix Profiles Zoom Match Region 1

(d) Matrix Profiles Zoom Match Region 2

(e) Match Region 1 motif Overlapping

(f) Match Region 1 motif by Date

(g) Match Region 2 motif Overlapping

(h) Match Region 2 motif by Date

Fig. 13: Stock within an Index, short term Pan Matrix Profile (*PMP*) analysis

4 Conclusions

In this work we have explored the potential of the *Matrix Profile* (*MP*) algorithm, to offer additional insight on financial series analysis by practical demonstration of motif identification and behaviour characterisation. Construction of *MP* series plots *within* a single series can illustrate longer-term trends around a given date (identified from low *MP* values), while *MP* series distributions reflect the percentage of motif matches and discords in the underlying series.

In multiple series analyses, the coincidence of local *MP* minima values can illustrate similar behaviour (i.e. *motif shape*) across single market sectors, as well as more generalised market behaviour (based on a set of companies spanning multiple sectors). The relationship between index data and individual stock data can also be examined using the *MP*. Additionally the combination of series based upon different measures of a single company or index can be investigated using this approach, providing insight for example on whether a company is under or over valued. The relationship between local *MP* minima and the behaviour of the series they represent is also explored through examination of raw data sub-sequences (based on the identified *MP* minima location and known *MP* sub-sequence length). This is demonstrated for both the single and multi-variate case.

The choice of sub-sequence length for analysis is an important consideration. The *Pan Matrix Profile* (*PMP*) algorithm (an extension of the *Matrix Profile*), applied to financial series, demonstrates how this decision can be informed by motif location and length in a given data set. Additionally, it can be used to simplify interpretation of *MP* plots by using short sub-sequence range to probe regions of interest. Nevertheless, a more comprehensive automated method for determining localised *MP* minima is clearly desirable, while the robustness of the general methods should be tested on additional time series, such as market rate curves and commodities for example.

Moreover, while the work presented here has focused on interpretation of independent *MP* plots for the multivariate case, recent work on extending the *MP* algorithm, such as *mSTAMP*[30] and *Ostinato*[31], suggests that examining all underlying series simultaneously is within reach. This would facilitate automation of a process to illustrate occasions where series are conforming with market behaviour, additionally highlighting potential hedging opportunities through the identification of series (within the set examined) that do not exhibit this behaviour.

References

1. A. Mueen, E. Keogh, Q. Zhu, S. Cash & B. Westover (2009). Exact Discovery of Time Series Motifs. In: *SIAM International Conference on Data Mining*, pp.35-53, doi:*https://epubs.siam.org/doi/10.1137/1.9781611972795.41*, pp.473-484
2. N. Castro & P.Azevedo (2012) Significant motifs in time series. In: *Statistical Analysis and Data Mining*, pp.35-53, doi:*https://doi.org/10.1002/sam.11134*

3. D. F. Silva, C. M. Yeh, Y. Zhu, G. Batista & E. Keogh (2019) Fast Similarity Matrix Profile for Music Analysis and Exploration. In: *IEEE Transactions on Multimedia*, Volume 21, pp.29-38, doi:*10.1109/TMM.2018.2849563*
4. N. Shakibay Senobari, G. Funning, Z. Zimmerman ,Y. Zhu & E. Keogh (2018) Using the similarity Matrix Profile to investigate foreshock behavior of the 2004 Parkfield earthquake. In: *American Geophysical Union Fall Meeting*, pp.S51B-03 *https://ui.adsabs.harvard.edu/abs/2018AGUFM.S51B..03S/abstract*
5. P. Ferreira, P. Azevedo, G. Silva & M. Brito (2006). Mining Approximate Motifs in Time Series. In: *Discovery Science*, pp.89-101, isbn: *978-3-540-46493-8*
6. W. Wilson, P. Birkin & U. Aickelin (2008). The motif tracking algorithm. In: *International Journal of Automation and Computing*, pp.32-44, doi: *10.1007/s11633.008.0032.0*
7. D. Xiaoxi, J. Ruoming, D. Liang, V.E. Lee, & J.H. Thornton (2009). Migration Motif A Spatial Temporal Pattern Mining Approach for Financial Markets. In: *15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp.1135-1144, doi: *10.1145/1557019.1557142*
8. E. Cartwright, M. Crane & H. J. Ruskin (2018) Abstract: Motif Discovery & Evaluation Focus on Finance. In: *Econophysics Colloquium 2018* Palermo - Italy *https://sites.google.com/view/econophysics-colloquium-2018*.
9. E. Cartwright, M. Crane & H. J. Ruskin (2019) Financial Time Series: Motif Discovery and Analysis Using VALMOD. In: *International Conference on Computational Science* Faro, Algarve - Portugal, pp.771-778, doi: *10.1007/978-3-030-22750-0_77*
10. C. M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. Dau, D.F. Silva, A. Mueen & E. Keogh (2016). Matrix profile I: All pairs similarity joins for time series a unifying view that includes motifs discords and shapelets. In: *IEEE ICDM* 1, pp.1317-1322, doi: *10.1109/ICDM.2016.0179*
11. E. Keogh (2020). The UCR Matrix Profile Homepage: *https://www.cs.ucr.edu/ eamonn/MatrixProfile.html*.
12. J. Teall (2018) *Financial Trading and Investing*, Second Edition. Academic Press, pp.145-167 isbn: *9780128111161*
13. X. Gao, H. An, W. Fang, X. Huang, H. Li, W. Zhong (2014) Characteristics of the transmission of autoregressive sub-patterns in financial time series. In: *Scientific Reports*, 4, 6290, pp.2045-2322, doi: *https://doi.org/10.1038/srep06290*
14. Investopedia common chart pattern definitions: *https://www.investopedia.com/articles/technical/112601.asp*
15. Investopedia Pennant chart pattern definition: *https://www.investopedia.com/terms/p/pennant.asp*
16. Investopedia Triple Bottom chart pattern definition: *https://www.investopedia.com/terms/t/triplebottom.asp*
17. Yahoo Finance Historical S&P Index: *https://finance.yahoo.com/quote/%5EGSPC*
18. A. Meegan, S. Corbet, C. Larkin (2018) Financial market spillovers during the quantitative easing programmes of the global financial crisis (2007–2009) and the European debt crisis. In: *Journal of International Financial Markets, Institutions and Money*, 56, pp.128-148, doi: *https://doi.org/10.1016/j.intfin.2018.02.010*
19. T. Bracke, F. Michael (2012) The macro-financial factors behind the crisis: Global liquidity glut or global savings glut?. In: *The North American Journal of Economics and Finance*, 23, pp.185-202, doi: *https://doi.org/10.1016/j.najef.2012.01.001*
20. N.Castro, P. Azevedo (2012) Significant motifs in time series. In: *Statistical Analysis and Data Mining*, 5, pp.35-53, doi: *https://doi.org/10.1002/sam.11134*

21. Investopedia Website Technology Companies list: *https://www.investopedia.com/articles/markets/030816/worlds-top-10-technology-companies-aapl-googl.asp*
22. Investopedia Website Pharmaceutical Stocks list: *https://www.investopedia.com/investing/pharmaceutical-stocks/*
23. Investopedia Website Finance Stocks list: *https://www.investopedia.com/terms/f/financial_sector.asp*
24. OECD Financial Markets Highlights November (2007): *http://www.oecd.org/finance/financial-markets/39654572.pdf* In: *Financial Market Trends*, 93, Vol. 2007/2 pp.11-25, ISSN *0378-651X*
25. Bernanke B.S. (2017) The Recent Financial Turmoil and its Economic and Policy Consequences: *https://www.federalreserve.gov/newsevents/speech/bernanke20071015a.htm*
26. Yahoo Finance Historical Dow Jones Index October 26th 2004 to October 8th 2008: *https://finance.yahoo.com/quote/%5EDJI/history?period1=1098748800&period2=1223424000&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true*
27. Britannica Financial crisis of 2007–08 summary: *https://www.britannica.com/event/financial-crisis-of-2007-2008/Key-events-of-the-crisis*
28. Yahoo Finance Historical Microsoft Data: *https://finance.yahoo.com/quote/MSFT*
29. F. Madrid, S. Imani, R. Mercer, Z. Zimmerman, N. Shakibay, A. Mueen & E. Keogh (2019). Matrix Profile XX: Finding and Visualizing Time Series Motifs of All Lengths using the Matrix Profile. In ICBK, 1, pp.175-182, doi: *10.1109/ICBK.2019.00031*
30. C-C.M. Yeh, N. Kavantzas & E. Keogh (2017) Matrix Profile VI: Meaningful Multidimensional Motif Discovery. In: *IEEE ICDM*, 1, pp.565-574, doi: *10.1109/ICDM.2017.66*
31. K. Kamgar, S. Gharghabi & E. Keogh (2019) Matrix Profile XV: Exploiting Time Series Consensus Motifs to Find Structure in Time Series Sets In: *IEEE ICDM*, 1, pp.1156-1161, doi: *10.1109/ICDM.2019.00140*