Side-Length-Independent Motif (SLIM): Motif Discovery and Volatility Analysis in Time Series—SAX, MDL and the Matrix Profile †

Eoin Cartwright 1,*‡‡, Martin Crane 2,‡‡ and Heather J. Ruskin 1,†‡

1 Modelling & Scientific Computing Group (ModSci), School of Computing, Dublin City University, D09Y074 Dublin, Ireland; heather.ruskin@dcu.ie
2 ADAPT Centre, School of Computing, Dublin City University, D09Y074 Dublin, Ireland; martin.crane@dcu.ie
* Correspondence: eoin.cartwright3@mail.dcu.ie
† Special Issue “New Challenges in Energy and Finance Forecasting in the Era of Big Data”.
‡ These authors contributed equally to this work.

Abstract: As the availability of big data-sets becomes more widespread so the importance of motif (or repeated pattern) identification and analysis increases. To date, the majority of motif identification algorithms that permit flexibility of sub-sequence length do so over a given range, with the restriction that both sides of an identified sub-sequence pair are of equal length. In this article, motivated by a better localised representation of variations in time series, a novel approach to the identification of motifs is discussed, which allows for some flexibility in side-length. The advantages of this flexibility include improved recognition of localised similar behaviour (manifested as motif shape) over varying timescales. As well as facilitating improved interpretation of localised volatility patterns and a visual comparison of relative volatility levels of series at a globalised level. The process described extends and modifies established techniques, namely SAX, MDL and the Matrix Profile, allowing advantageous properties of leading algorithms for data analysis and dimensionality reduction to be incorporated and future-proofed. Although this technique is potentially applicable to any time series analysis, the focus here is financial and energy sector applications where real-world examples examining S&P500 and Open Power System Data are also provided for illustration.

Keywords: financial time series; matrix profile; symbolic aggregate approximation (SAX); minimum description length (MDL); time series motifs

1. Introduction

A motif [1,2] is a repeated matched (or partially matched) sub-sequence taken from a larger parent time series (or set of time series). Given the ever-increasing prevalence of large, ‘Big Data’ sets, commonly seen now in the Energy and Financial sectors, for example, the importance of motif analysis to facilitate interpretation of underlying series behaviour and prediction of future trends is increasing [3]. Here we apply a combination of existing algorithms and principles, namely SAX, MDL and the Matrix Profile, in order to improve the identification of repeated behaviour occurring within a series while allowing for flexibility of sub-sequence length.

The approach (designated Side-Length-Independent Motif or SLIM) is distinct from that of other motif search algorithms, which permit a user-defined length range with motif side-length of A and B to be equal, whereas our method permits motif sides to be of different lengths, extending pattern recognition potential for the series. Additionally, the details recorded during this process (described in Section 2.2) provide insight into series volatility at a local level but also facilitate the comparison of overall volatility between series.

This technique combination (Algorithm in Section 2.2.4), while developed for financial series initially, yields tangible results in more application areas than existing methods. For
example, within the energy sector, the identification of motifs in power consumption data can represent patterns in user behaviour, improving forecasting of energy demand. In finance, these patterns or motifs represent repeated behaviour of a given series, such as a sharp rise in share or market-rate value with a slow decline or a gradual rise over a longer time period. Often these can take the form of commonly recognised ‘shapes’ (or behaviours) in financial series, such as Head and Shoulders, for example [4].

1.1. Literature Review

Numerous approaches to time series analysis and forecasting appear in the literature [5]. For example [6], where several forecasting models, such as Simple Exponential Smoothing (SES) and Autoregressive Integrated Moving Average (ARIMA), are compared against a machine learning Support Vector Regression (SVR) model using weekly crude oil price data from 2009–2017. Additionally, in [7], local Hurst Exponent signals were used to investigate an anti-correlation signature in the share price evolution of the Warsaw stock exchange (WIG20) index, which occurs around the maximum share value.

ARIMA and SES use all data, whereas Hurst (and motifs) are more local in their focus. Here we are concerned with the use of motifs in time series analysis, which facilitate the examination of underlying trends and processes [3].

Focus on motif discovery in time series has intensified since the early 2000s, leading to significant algorithmic improvements in terms of speed and efficiency. Applications include finance [8], health [9] and music [10], amongst others. Additionally, motif discovery algorithms are used as subroutines in many time-series data mining tasks, such as rule mining, clustering and classification [11–13].

In general, motif discovery algorithms can be divided into two groups, categorised by fixed and variable lengths with a further distinction made between the use of approximate and exact approaches (Figure 1).

Figure 1. Motif discovery algorithm highlights timeline. Green indicates fixed-length algorithms while blue represents variable length.

Approximate fixed-length motif discovery is largely based upon random projection (CK Algorithm [14]) and Symbolic Aggregate Approximation or SAX [2,15] techniques (discussed further in Section 2.1.1). Of note is the use of iSAX in the MrMotif [16,17] algorithm that derives a set of top-K motifs for a fixed length through increasing SAX resolutions.

Exact approaches initially concentrated on the use of early abandonment or Smart Brute Force (SBF) in the MK [1] algorithm, with further speed efficiencies gained by Quick Motif [18] for example. More recently, the use of the Matrix Profile (MP) [19,20], a highly efficient Euclidean Distance similarity search algorithm, has predominated as state-of-the-art for fixed-length motif discovery.

Variable-length approximate algorithms include grammar induction-based approaches, such as Sequitur [21] and DP-Sequitur [11], along with VMLD [22] and kBMD [23], which eliminate the requirement for a predefined sliding window length parameter. Exact algorithms that allow for variable-length include K-Motif [24] and VALMOD [25], along with SKIMP [26]. SKIMP is the first practical technique to find motifs and discords for all lengths through the creation of a Pan Matrix Profile (PMP), which can also be easily visualised as
a heatmap (see Figure 10d). Both VALMOD and SKIMP are part of a body of recent MP papers [20], providing an important resource for time series analysis.

1.2. Contribution

Here we propose a novel combination of several methods in order to investigate high-frequency changes (i.e., the volatility in financial and energy markets). Similarly, our approach offers increased flexibility in the identification of motif characteristics, such as side length and shape.

The main contribution of SLIM is the introduction of flexibility of motif side-length within an individual motif sub-sequence pair, allowing similar behaviour occurring over differing lengths to be identified and directly compared. This is demonstrated in Figure 2 with two identified motif pairs planted in a synthetic sample series, one of equal motif side-length (Figure 2b) and the other unequal (Figure 2c).

Existing variable-length motif discovery algorithms produce results over a range with motif pair sides equal. Thus, extra processing is required in order to compare behaviour over two different length values directly, whereas SLIM can produce individual motifs with variable side lengths, obtained through the temporary compression of similar sub-sequence values before close matches are identified.

Additional benefits of SLIM include allowing a visual comparison of relative volatility levels within and between series (an important consideration, especially in finance as volatility furnishes key aspects, such as return on investments and effective hedging [27]). The identification of localised sub-sequences distinguishing volatility related to sudden large events from a consistent increase, for example, is also facilitated by SLIM.

The basis for the SLIM approach utilises established techniques, such as Symbolic Aggregate Approximation or SAX [2,15,28], the Matrix Profile (MP) algorithm [19,20] and the principle of Minimum Description Length (MDL). A brief outline of the algorithms employed is provided in Section 2, with a detailed explanation of the new technique of applying MDL to SAX strings in Section 2.2. Finally, real-world examples are used to illustrate
technique application in Section 3. Future scope for improvements, such as more rigorous dimensionality reduction (offered by SAX for examination of the ‘Big Data’ sets that are becoming more prevalent), are also discussed in Section 4.

2. Methodology
2.1. Underlying Algorithms

The three main algorithms of interest are:

2.1.1. Symbolic Aggregate Approximation (SAX)

Symbolic Aggregate Approximation or SAX [2,15,28] is used to discretise time-series data into a symbolic string that reduces dimensionality while indexed by a lower-bounding distance measure. It has proven to be a particularly effective tool for motif discovery, underpinning many string analysis techniques for motif detection borrowed from the study of DNA sequences. SAX relies upon the following definitions:

Definition 1. A time series $T$, is a sequence of real-valued numbers $T = t_1,t_2,...,t_n$ where $n$ is the length of $T$ [2].

Definition 2. A SAX string $\hat{C}$, is a symbolic representation of a time series $\hat{C} = \hat{C}_1, \hat{C}_2, ..., \hat{C}_w$ assigned to a Piecewise Aggregate Approximation reduction of $T$, from $n$ to $w$ dimensions $C = C_1, C_2, ..., C_w$ (adapted from [2]). $w$ is the length of symbolic representation (i.e., no. of piecewise segments) where $w << n$.

Definition 3. Breakpoints are a sorted list of numbers $B = \beta_1, \ldots, \beta_a - 1$ such that the area under a $N(0,1)$ Gaussian curve from $\beta_i$ to $\beta_{i+1} = \frac{1}{a}$ ($\beta_0$ and $\beta_a$ are defined as $-\infty$ and $\infty$, respectively), while $a$ is the alphabet size [2].

In summary, an input series is first normalised and broken into a user-provided number of horizontal segments. An average of the series values within each segment is taken, and a symbol is then assigned, depending upon the value range that contains the average. The symbol value intervals (i.e., vertical segment size or breakpoints) are calculated according to the equal assignment of area under a Gaussian curve, which, in turn, is dependent upon the alphabet size provided. Figure 3 illustrates this transformation process. Note, symbols can be alphanumerics, producing SAX strings, such as bacbc, for example, or 21323, as shown here.

![Figure 3](image-url)
Numerous versions of SAX have evolved that improve and tailor the technique to particular applications. Examples include iSAX, which applies an indexing technique that is fast and scalable [30] and Symbolic Fourier Approximation (SFA) that introduces a symbolic representation based on the frequency domain, allowing for indexing of high-dimensional data-sets [31]. Here we are concerned with obtaining SAX strings from sample time series to serve as the basis for further pattern analysis.

2.1.2. Minimum Description Length (MDL)

The basic principle of Minimum Description Length (MDL) is that, given a limited set of observed data, the best description is one that permits the greatest compression. MDL is used in a wide range of disciplines, such as machine learning, data mining, biology and econometrics [32,33].

The MDL principle is applied here to SAX strings obtained from a sample series (see Figure 8, Section 3.1.1 for an illustration). Using MDL, a SAX series representation can be refined to highlight regions of stability/volatility, as well as to allow a length flexibility when identifying pattern repeats (motifs).

2.1.3. Matrix Profile (MP)

The Matrix Profile (MP), a novel algorithm due to [20], has already demonstrated considerable potential for numerous data mining and time series analysis tasks. It has been found to be highly scalable for time-series sub-sequence all-pairs-similarity search [19] and also efficiently identifies time series motifs and discords (i.e., mismatches). For a more comprehensive summary and further information on the MP (including extensions) see [20].

The MP can be represented as a pseudo-series where low MP values are indicative of close matches (in terms of Euclidean distance) to some other point within the examined series. The start location of the identified matching sub-sequence can be obtained from an associated Matrix Profile Index (MPI) value. When examining MP plots, low-value regions indicate matches or motifs, while high values illustrate mismatch regions, or discords (Figure 4).

Figure 4. Sample matrix profile representation (red series) of the S&P500 (blue) series between January 2008 and January 2010. Here low MP values indicate a close match with another sub-sequence (of MP length) at some other point within the S&P500. This point is given by the associated Matrix Profile Index (MPI) value.
Here the focus is on the interpretation of MP plots in financial and energy sector time series, where low MP plot values (based upon SAX series representations) highlight match regions or motifs.

2.2. Combined Methodology

There are two parts to our approach: Firstly, the MDL principle is applied to SAX strings, followed by the application of the Matrix Profile to an MDL-SAX string in order to identify match regions (or motifs).

As a motif is a pair of similar sub-sequences (or segments) of a larger time series, there must be a minimum of two sides or parts to the match. For clarity, the sides of a motif pair are designated as Sides A and B, where Side A is considered as the initial candidate segment, and Side B is the segment identified as a match of Side A.

2.2.1. Application of MDL to SAX strings

Series values are parsed into SAX strings, which are then compressed, where adjacent repeat elements are collapsed into one element, with the superscript value indicating the number of repeated elements, Figure 5.

**Definition 4.** An MDL-SAX string $\hat{M}$, is a compressed representation of a SAX string $\hat{M} = \hat{M}_1, \hat{M}_2, ..., \hat{M}_n$ obtained by the application of the Minimum Description Length principle, such that adjacent equal values $\hat{M}_i, ..., \hat{M}_n$ are represented by $\hat{M}_x^z$ where $x = |\hat{M}_1, ..., \hat{M}_n|$ (i.e., the number of compressed elements).

```
SAX  5 5 5 4 4 4 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 4 3 3
MDL-SAX  5^3 4 5 4^3 3^3 2 3 3 3 3 3 4 3 3
```

Figure 5. Illustration of MDL-SAX string created from a SAX time series representation. Consecutive equal SAX elements are collapsed into one element, with the superscript value indicating the number of repeats.

Thus, a sample SAX string of 555454433233333433 becomes $5^345^33^323^543^2$. The process is demonstrated for a real series (S&P500) in Figure 8, Section 3.1.1, for an initial time window of January 2008 to January 2010, with the resulting MDL-SAX value string displayed in Figure 8d.

Additional detail retained for further analysis includes the number of consecutive SAX string elements combined, the value difference between adjacent MDL-SAX string elements, along with SAX/MDL-SAX string and raw series indexes (Table 1). Note the length reduction between the SAX(c) and MDL-SAX(d) representations in Figure 8.

Thus, an MDL-SAX string is a compression of the original SAX string representation of the raw time-series data. For time series, MDL-SAX compression of the original SAX string has the net effect of ‘removing’ periods of stability while retaining the volatility profile (Figure 6a).
**Table 1.** Sample additional detail recorded during the construction of an MDL-SAX string. **SAXVal** is the value of the SAX series at a given index, **SAXValDiff** is the difference between the current and previously recorded SAX value, while **SymJoinNum** is the number of consecutive SAX series values combined through the application of MDL.

<table>
<thead>
<tr>
<th>Raw Series Date</th>
<th>SAX Series Index</th>
<th>SAXVal</th>
<th>SAXValDiff</th>
<th>SymJoinNum</th>
<th>RawSeries Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 January 2009</td>
<td>97</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>254</td>
</tr>
<tr>
<td>7 January 2009</td>
<td>98</td>
<td>4</td>
<td>−1</td>
<td>1</td>
<td>257</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>28 January 2009</td>
<td>104</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>271</td>
</tr>
<tr>
<td>29 January 2009</td>
<td>105</td>
<td>3</td>
<td>−1</td>
<td>2</td>
<td>272</td>
</tr>
</tbody>
</table>

Figure 6 illustrates a comparison between SAX and MDL-SAX representations of the S&P500 from January 2008 to January 2010, a window chosen for the volatility that reflects the considerable stress experienced in the global marketplace at this time [34,35] and extending previous work [36,37].

In the illustrated example, the MDL-SAX and SAX strings are plotted using both non-adjusted (Figure 6a) and adjusted (Figure 6b) scales to highlight the features of the SAX string captured by MDL-SAX. The overall shape and dynamics are well preserved, as the series examined has relatively few periods of stability. However, if an alternative series with low volatility is chosen, then a higher rate of compression would be observed, affecting the profile of the MDL-SAX string relative to the SAX string values.

2.2.2. Hyperparameter Selection: Influence of Alphabet Size Choice upon MDL Compression Rate

The choice of alphabet size used when creating the initial SAX string from the raw series will influence the compression rate when MDL is applied. Thus, while MDL compression is dependent upon the volatility level of the series in question, an increase in alphabet size chosen will reduce the overall compression rate.
This result is intuitive, of course, as a larger alphabet size requires a corresponding increase in SAX breakpoints, which in turn leads to an increased resolution of SAX values. As a given series segment is then represented by an increased range of SAX values, overall MDL compression is reduced. Figure 7 illustrates this for the S&P500 from January 2008 to January 2010, where the compression rate CR% is given as:

\[ CR\% = \left( \frac{L_{SAX} - L_{MDL-SAX}}{L_{SAX}} \right) \times 100 \]  

where \( L_{SAX} \) is the length of the SAX series and \( L_{MDL-SAX} \) is the length of the SAX series after MDL has been applied.

An increased alphabet may be required for a stable series in order to obtain a similar compression rate to that of a more volatile series, i.e., in order to compensate for slow increases or decreases represented by the same SAX value.

Thus, Compression Rate % vs. SAX Alphabet Size plots, as shown in Figure 7, allow a choice of a suitable alphabet size value to be made in order to achieve a desired compression rate when obtaining an initial MDL-SAX string from the raw series. Additionally, when used in conjunction with an examination of raw series motifs (as discussed in Section 2.2.4), these plots facilitate a choice based on the amount of compression required.

![S&P500 Jan-2008 to Jan-2010 MDL SAX Series Length % Vs Alphabet Size](image1)

![S&P500 Jan-2008 to Jan-2010 Compression Rate % Vs Alphabet Size](image2)

(a) MDL-SAX series length for increasing alphabet size. (b) Compression Rate % for increasing alphabet size.

Figure 7. Effect of MDL application to a SAX S&P500 series from January 2008 to January 2010. Accordingly, as the alphabet size used to create the SAX representation of the S&P500 series increases, the length of the resulting MDL-SAX series also increases (a) while the compression rate is reduced (b). A large alphabet size range is utilised here to illustrate the stability of the compression rate at higher alphabet values.

Of further note is the choice of the alphabet and segment size (i.e., the length of raw series that each SAX symbol represents) to provide a suitable compression rate (as here) when MDL is applied, as opposed to preservation of raw series features. The original SAX technique objective is dimensionality reduction [15], where the choice of alphabet size and segment number determines the maximum reduction in data while preserving features of an input series with a lower bound.

However, since the primary concern here is a desired compression rate upon the application of MDL, SAX series breakpoints may be, in consequence, more frequent in relation to the original raw series than the volatility demands. Of course, extremely large series (such as those for high-frequency financial trading) may require initial data reduction from the SAX representation, achieved by a suitable choice of segment size.
2.2.3. Motif Discovery

It should be noted that SAX or MDL-SAX representations of an original series can also be used as input to the MP algorithm allowing match regions (or motifs) to be identified (within the SAX or MDL-SAX strings), as indicated by low MP values.

This raises the question of why we choose to use the MP of a SAX string for motif detection at all, as opposed to available alternatives, such as comparing SAX string values directly, using, for example, some form of a sliding window. While this was investigated and is relatively simple, the number of trivial matches obtained is very large. Further, using a 1:1 ratio between SAX values and raw data means that searching is limited to finding an exact match (or ideal motif), which is a rare event [38].

The MP algorithm is more efficient as a motif candidate is obtained even for non-exact matches (in terms of Euclidean distance). It also incorporates an in-built exclusion zone principle that rejects trivial (or same position) matches. Finally, the efficiency of the MP is $O(n\log n)$ (where $n$ is the length of the input time series) [19], which is an improvement on a brute-force comparative approach based upon SAX strings.

2.2.4. Independent Side-Length Motif Discovery Process

Clearly, an MP can be obtained directly from raw data input with less effort than by first creating a SAX representation of the series. However, the use of SAX presents further in-depth opportunities for analysis through the application of MDL to the SAX string before the MP plot is generated. This allows motifs to be identified after the MDL compression has occurred, capturing higher-order match regions of similar behaviour.

Additionally, when returning to either the original SAX string or raw data without MDL applied (using values stored in Table 1, for example), variable-length introduction is possible without reference to the side-length of the motif pair as it relies solely on the compression that occurs within the individual segment of the MDL-SAX string that is identified as a close match. Thus, side-length-independent motifs can be obtained from the raw data where the length of Side A does not necessarily match that of Side B.

The process by which this is achieved is as follows Algorithm 1:

Algorithm 1: Side-Length-Independent Motif (SLIM) pseudo-code.

Data: Input raw time series
Result: Candidate motif locations with variable side-length

Step A: Transform raw input series into a suitable SAX representation
Step B: Compress the SAX series using MDL to create an MDL-SAX series
Step C: MDL-SAX series serves as input to the MP algorithm creating an MDL-SAX-MP series

while examining MDL-SAX-MP series do
  Identify low MDL-SAX-MP series values indicating close match, i.e., motifs
  • Sub-sequence length = length parameter input to the MP algorithm
  • Location of matching side of the motif pair is obtained from MPI value (i.e., motif sides A and B as described earlier)
  Obtain end point of MDL-SAX motif segment for each side of the motif pair
  • Allowing MDL compression to be removed when plotting corresponding SAX and raw series segments
end

2.2.5. Advantages

Our approach, designated the Side-Length-Independent Motif (SLIM), offers several advantages over the reported methods to date:

• Permits identification of motif pairs in which the length of each side is independent.
• Properties of the underlying algorithms are inherited.
  – Dimensionality reduction of SAX (if required).
  – Efficiency and scalability of the MP.
• Is independent of SAX and MP versions used and so can take advantage of further improvements to these algorithms.

3. Results and Discussions

In the following section we highlight potential applications and advantages of SLIM (Side-Length-Independent Motif identification) in the financial and energy sectors, two areas generating increasingly large data-sets.

In the energy sector, a set of hourly Open Power System Data (OPSD) relevant for power system modelling within the EU and neighbouring countries was considered [39]. For the financial illustration, we build upon previous work [36,37] and continue with S&P500 data, as it is widely regarded as the best single gauge of U.S. equities and serves as the foundation for a wide range of investment products [40].

For the S&P500, localised aspects of volatility are highlighted, while compression rate % vs. alphabet size plots are also used to compare volatility levels for power usage across several European countries.

3.1. Finance

3.1.1. Side-Length-Independent Motif Discovery

To demonstrate the steps outlined in Section 2.2.4, S&P500 index data between January 2008 and 2010 (Figure 8a) is converted into a SAX representation in Figure 8b. Conversion of these SAX values to an MDL-SAX series is shown in Figure 8c,d (using a single month for clarity).

An MDL-SAX-MP plot for the full time window was created (Figure 9a). Close matches, as indicated by low MP values within the MDL-SAX-MP series, are highlighted as points X, Y and Z, while corresponding motif pairs within the MDL-SAX series are shown in the top plots of Figure 9b–d.

The MDL criterion is then removed by returning to the start and end points of the SAX series that the MDL-SAX motif represents. The length of each side of the motif pair will vary independently as a result of the level of compression that occurred in the particular section of the MDL-SAX series, as shown in Figure 8c,d.
The resulting change in motif side-length of the SAX series is shown in the middle plots of Figure 9b–d, while the equivalent raw series segments are also included in the bottom plots for comparison.

Figure 9b shows the most compression and gives a greater length differential between the sides of the motif pair when MDL is removed. Since MDL compression is obtained from the combination of successive equal SAX values, its removal leads to flatter plots, as indicated by the variance in SAX value range between Sides A and B in the middle plot of Figure 9b.

Overall the behaviour of the SAX and raw data series correspond quite well, as shown in the middle and bottom plots of Figure 9b–d. Additionally, even though no normalisation has been applied to the raw series data, a close match between each side of the motif pair is still observed, particularly in Figure 9c,d.

This identification of similar behaviour (characterised by motif shape) within financial time series may be used to identify potential investment opportunities through identification of pattern repeats now flexibly interpreted with respect to match length.

Figure 9. Matrix profile of the S&P500 from January 2008 to January 2010, after application of MDL.
Motif pairs identified by MP minima indexes highlighted at points X, Y and Z in (a) are illustrated in (b–d) with variable lengths per motif side illustrated in both the SAX (middle) and raw data (bottom) sub-sequences.
3.1.2. Alternative Motif Identification Algorithms Comparison

To aid comparison to other motif identification methods, Figure 10 was constructed with SLIM contrasted to the MP, MrMotif and SKIMP. The choice of MP for the initial analysis (Figure 10a,b) was based on its current perception in the literature as state-of-the-art. The equivalent index from previously identified low MP values of the MDL-SAX-MP (i.e., start indexes of Side A in Figure 9c,d) was translated to an MP based upon S&P500 series data and matching raw series sub-sequences identified.

Overall, the behaviour (i.e., motif shape) is in good agreement between the original fixed-length MP (top) and variable-length SLIM motif plots (bottom) in Figure 10a,b. Along with an increase in sub-sequence length, even for the shorter side-length SLIM motif. We note that the index of the raw series MP plot used may not necessarily be a low MP value (in the context of the overall raw series MP). As in this case, we use a fixed index as a starting point and corresponding MPI to illustrate a match, rather than consulting the entire raw series MP for a global minimum (indicating the point of best match obtained).

**Figure 10.** Comparison of motifs obtained from MP and SLIM. (a) SLIM vs. MP from Figure 9c. (b) SLIM vs. MP from Figure 9d. (c) MrMotif [16,17] front end. (d) Pan Matrix Profile (PMP) heatmap created by the SKIMP [26] algorithm.

Figure 10c contains the results for the MrMotif [16,17] algorithm (based upon author-provided sample data), illustrating a set of fixed-length motifs obtained at increasing SAX resolutions. While in Figure 10d, a heatmap of a Pan Matrix Profile (PMP) created by the SKIMP [26] algorithm is shown. This indicates the location and lengths identified within...
the same S&P500 data-set used previously from January 2008 to January 2010. Of the algorithms examined, only SLIM returns a direct comparison of sub-sequences of differing lengths, without the need for an extra processing step.

3.1.3. Localised Volatility Analysis

The additional details recorded when MDL is applied to a SAX string (Table 1) also permit volatility analysis at the local level, with segments (or sub-sequences) of an overall series identifiable in terms of volatility match, for example (i.e., max, or min, as shown here). In finance, volatility is an important measure that represents the dispersion of returns for a given security or index and essentially measures risk [41].

Key segments are identified by the creation of a sliding window (of user-specified segment size) parsing through the previously created MDL-SAX combination table (Table 1). Values such as the sum of the absolute values of SAX value differences column (SAXValDiff), symbol join number (SymJoinNum) column and amplitude (calculated as the difference between the minimum and maximum SAX values within the sliding window) are recorded in a volatility summary table along with the MDL-SAX series index of the current location of the sliding window (see Table 2).

Table 2. Sample extract of volatility table details from MDL-SAX table.

<table>
<thead>
<tr>
<th>MDL SX SeriesIdx</th>
<th>SAXValDiffTotal</th>
<th>SymJoinNumTotal</th>
<th>SAXValAmplitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>26</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>26</td>
<td>10</td>
<td>17</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Table 2 now contains summary information on the original MDL-SAX table that can be used to identify volatility areas of interest in the original series. A large difference between SAX (or MDL-SAX) values (i.e., SAXValDiff) corresponds to a large shift in raw series value. Similarly, a high value of SymJoinNum (i.e., consecutive, unchanged SAX values) indicate series stability.

Overall levels of volatility in a segment can be ordered by SAXValAmplitude (maximum corresponding to highest volatility and vice versa), SymJoinNum (reflecting stability within the sliding window) or by SAXValDiff (indicating the number of changes within the sliding window).

Thus, the identification of segments is based upon a combination of values rather than a single standard deviation value, which is commonly used [41]. Additionally, the particular focus can be emphasised by the choice of primary column ordering. For example, prioritising SAXValAmplitude over SAXValDiff emphasises the significance of a single large event within the sliding window, as opposed to smaller but more numerous changes captured by SAXValDiff.

Figure 11 shows the MDL-SAX representation of the S&P500 between January 2008 and January 2010 with identified high (red) and low (green) volatility segments. Additionally, identified individual raw series segments and locations within the original series are provided for clarity. The standard deviation was also examined, returning values of 13.02 and 93.97 for the isolated low and high segments, respectively (confirming very different volatility levels).
Figure 11. S&P500 MDL-SAX representation (a) with identified high/low volatility segments. For clarity, the original S&P500 series from January 2008 to January 2010 is shown in (b), with isolated low (c) and high (d) volatility segments. SAX Alphabet size = 200, Num of SAX segments = series length, MDL-SAX segment length = 15.

Limitations to our SLIM approach mainly centre on the normalisation step applied during the application of SAX to the raw input series. In the initial step within the SAX algorithm, a normalisation (of Gaussian form) is applied to the input series, such that the range of original series values represented by each SAX symbol is larger for extreme raw series values than mid-range. This translates to increased sensitivity of SAX values to the raw series in the mid-range of the data (and corresponding higher SAXValDiff sum when applying a sliding window). This may result in the identification of more segments at the outer edges of the MDL-SAX series, particularly when looking for low volatility.

This effect can be mitigated somewhat through the choice of a large alphabet size, resulting in a reduced interval range when assigning SAX symbol values. However, even with small alphabets, the SLIM technique provides a good starting point for further analysis.

3.2. Energy Sector

3.2.1. Side-Length-Independent Motif Discovery

Application of SLIM to an energy sector example is illustrated for a set of 2020 Open Power System, hourly power consumption series for Germany and the UK [39], Figure 12. The same MDL-SAX process as outlined in Section 2.2.4 was followed. The sub-sequence length chosen for the MP algorithm was 48 (equivalent hours), permitting a search for similarities of 2 days in length within the compressed MDL-SAX string. An alphabet size of 20 was chosen in order to obtain a desired level of compression in the more volatile hourly data (see Figure 13 for relevant global compression rate % vs. SAX alphabet size plots).

Figure 12a shows motif segments identified from a sample low MP distance value occurring on 22 January 2020 (index 334 of the MDL-SAX-MP string) along with the matching segment identified (from the MPI), giving an index of 99. Overall a close correlation in behaviour between both motif sides is observed. When returning to non-compressed SAX and raw series from MDL-SAX, we observe that side length differs by 10 h, indicating similar power consumption behaviour occurring over a shorter timeframe.
This is a lower level of compression than previously observed and is reflective of the challenges encountered when dealing with increased volatility in the hourly energy data when compared to the daily financial data previously examined. Although the SAX alphabet size can be reduced to counter increased volatility in the raw data, there are limits to the efficacy of this tuning (in terms of distinguishing differences between motif Side A and B lengths).

In Figure 12b, the results of a different approach are illustrated. Rather than using a low MP distance value obtained from an MP plot of an MDL-SAX string as a starting point, the table obtained during the creation of the MDL-SAX string was consulted (equivalent to Table 1). A high SymJoinNum value was chosen as a starting point, providing an initial index of 621 in the MDL-SAX string (corresponding to 13 February 2020 in the raw series), with the alternative side of the match obtained from the MPI value at this point. A less accurate match may be obtained (an intuitive result as the MP distance value was greater in this case than that used in Figure 12a), but a length difference between motif sides is more likely with compression being removed. The approach is particularly useful where this is of importance, relative to the match accuracy from MDL-SAX-MP.

For highly volatile data, as here, less compression occurs in the creation of the MDL-SAX string, even for a small alphabet size in the initial SAX representation. Identifying different motif side-lengths is consequently more difficult when applying MDL. Figure 12 illustrates motifs of hourly power consumption with differing side lengths, representing potential patterns in user behaviour while allowing for flexibility in the match. Figure 12a shows overall matching behaviour while Figure 12b indicates a prolonged higher consumption level initially for Side A of the match, causing the remainder of the identified segment to be out of phase with that of Side B (a result of the initial choice of start location of Side A with high SymJoinNum value).

Figure 12. Sample motifs identified by SLIM in 2020 UK and Germany hourly power consumption data.

3.2.2. Globalised Volatility Analysis

Compression rate % vs. alphabet size plots also permit a visual analysis and comparison of relative volatility levels of series at a global level, an important feature as the availability of large data-sets increases. To demonstrate, Figure 13 shows relative volatility levels of German and UK hourly power consumption over a time span of 5 years. Here a lower overall compression rate is observed, indicating a higher level of volatility than previously observed for daily S&P500 data (Figure 7b), where maximum compression is approx 90%, as opposed to 75% here.
Furthermore, also of note in Figure 13 is the consistency of volatility levels observed, even for 2020, where differences compared to previous years might be expected due to the Covid-19 pandemic [42] triggering lock-downs in many countries and, by extension, alternative consumption patterns. Although a larger spread occurs for the UK in Figure 13b, the 2020 values still fall within the typical series distribution profile. In summary, despite different pandemic policies, neither Germany nor UK power consumption volatility appears to show a marked change from previous years.

4. Conclusions

In this work we have explored the novel use of a combination of several established data mining techniques for motif detection in time series. Specifically, these included Symbolic Aggregate Approximation (SAX), Minimum Description Length (MDL) and Matrix Profile (MP). Applications for finance and energy series are discussed.

The compression resulting from an application of MDL to SAX string representations of time series effectively removes periods of stability while retaining volatility. The compression rate achieved is a combination of the alphabet and segment size chosen during the creation of the initial SAX string and the volatility level of the series in question.

Construction of MP plots based on MDL-SAX representations permits the identification of motif pairs with an independent length per side. This is a highly useful feature for financial, as well as other series analysis, allowing similar behaviour (represented as motif shape) occurring over differing timescales to be identified. Example applications in the energy and financial domains are used in illustration with features, such as input tuning and motif side-length discussed in more detail.

Compression rate % vs. alphabet size plots provide a picture of the amount of compression obtained and act as an indicator of the overall volatility level within a given series or set of series. This technique can also be used for the identification and isolation of localised periods of high volatility or stability through the examination of additional detail on MDL-SAX representation.

Although SAX normalisation leads to some bias in terms of over-identification of significant matches at extremes of the data range, inputs may be tuned to optimise individual data-set analysis. Overall, the inherent algorithm properties are both flexible and highly scalable, with MDL-SAX independent of SAX and MP type, so that further potential for series analysis is considerable.

Future improvements include automation of low MP value selection and corresponding motif display. Additionally, given the dependence on the amount of compression within an individual segment of the SAX series representation, exact motif side-length
values can not currently be determined in advance, and this might usefully be a target for more detailed quantification.

Further work is also needed to assess the impact of an initial data reduction when converting the raw series to a SAX string in order to facilitate analysis of the ever-increasing volume of data generated for finance and other applications.

**Author Contributions:** Conceptualisation, methodology, software, formal analysis, investigation, visualisation and writing—original draft, E.C.; writing—review and editing and validation, M.C. and H.J.R. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was in part supported by the Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at DCU. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by the Science Foundation Ireland through the SFI Research Centres Programme.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.


**Conflicts of Interest:** The authors declare no conflict of interest.

**Abbreviations**

The following abbreviations are used in this manuscript:

- **SLIM** Side-Length-Independent Motif
- **SES** Simple Exponential Smoothing
- **ARIMA** Autoregressive Integrated Moving Average
- **SVR** Support Vector Regression
- **WIG20** Warsaw stock exchange index
- **SAX** Symbolic Aggregate Approximation
- **MDL** Minimum Description Length
- **MP** Matrix Profile
- **MPI** Matrix Profile Index
- **SFA** Symbolic Fourier Approximation
- **S&P500** Standard and Poor’s 500

**References**


