芸術科学会論文誌 Vol. 10 No. 3 pp. 179 – 191

SAX 法による局所パターン抽出を導入した 時系列データの三次元可視化の一手法

井元麻衣子 [1]　伊藤貴之

お茶の水女子大学大学院

A 3D Visualization Technique for Time-Varying Data Applying Local Pattern Extraction By SAX method

Maiko Imoto　Takayuki Itoh

Graduate School of Humanities and Sciences, Ochanomizu University

{maiko_i, itot} @ itolab.is.ocha.ac.jp

概要

情報可視化の諸技術の中でも，時系列データの可視化は依然として活発な研究が続いている．時系列データ可視化手法の多くは，時系列データを折れ線で表現している．大規模な時系列データを二次元平面上に可視化すると，線分同士の交差が多くなるため，その観察が困難になる場合が多い．本論文では，大規模な時系列データを大局的にも局所的にも可視化するための，三次元可視化手法を提案する．本手法では x 軸に時間，y 軸に数値を割り当て，算出した類似度に従って大量の折れ線群を z 軸方向に等間隔に並べて配置する．本手法では，大局的な類似度に従って折れ線群を順列化すると同時に，SAX 法 (Symbolic Aggregate approXimation) を適用して局所パターンを抽出する．続いて本手法では，x 軸に時間，y 軸に数値を割り当てた三次元直交座標系を想定し，折れ線群を z 軸方向に等間隔に並べて配置する．本手法では，この折れ線グラフを 2 つの視点から表示する．一つ目の視点は，折れ線グラフ群全体を大局的に観察するための，xz 平面に垂直な視点である．ここでは折れ線グラフが表す各値を色で表現し，さらに SAX 法が抽出した局所パターンも重ねて表示する．この視点からの表示において，詳しく観察したい有限本数の折れ線群をユーザが対話的に選択することができる．二つ目の視点は，ユーザが選択した折れ線群を局所的に観察するための，xy 平面に垂直な視点である．本手法で二つの視点を組み合わせながら時系列データを可視化することにより，大規模な時系列データの全体像を眺めながら，興味のある少数の数値群を選択的に注視し，折れ線グラフ同士の相関性を発見することが容易になると考えられる．

Abstract

Time-varying data visualization is an especially active research topic in the field of information visualization. We represent time-varying data as polyline charts very often. At the same time, we often need to observe hundreds or even thousands of time-varying values in one chart. However, it is often difficult to understand such large-scale time-varying data if all the values are drawn in a single polyline chart. This paper presents a polyline-based 3D time-varying data visualization technique. This technique orders a set of polylines based on their global similarities, and extracts local patterns by applying SAX (Symbolic Aggregate approXimation) method. It then places the polylines along the Z-axis in the order, in a 3D orthogonal coordinate system where time is assigned to the X-axis, and values are assigned to the Y-axis. This technique provides two views to visualize the set of polylines. The first view is orthogonal to the XZ-plane, for global observation of the whole polylines. Here, it represents the values by colors, and overlays the local patterns extracted by SAX method. The technique provides a user interface to interactively select a set of polylines which he/she would like to observe in detail. The second view is orthogonal to the XY-plane, for local observation of the selected polylines. The technique makes easier to overview the large time-varying data, selectively observe interested polylines in detail, and discover relevancy among the interested polylines.

[1]現在，日本電信電話株式会社 サイバーソリューション研究所に勤務

1 はじめに

情報可視化とは，空間や形状をもたない情報を含めて，身の回りの一般的な情報を可視化する，きわめて汎用的な可視化技術の総称である．情報可視化の権威的存在である Shneiderman[1] は，情報可視化が対象とする手法なデータ構造を，「一次元・二次元・三次元・n 次元 ($n > 3$)・時系列・木構造・グラフ」の 7 種類であると提唱している．その中でも我々は，時系列データの可視化に着目している．時系列データを分析し，その要素同士の関連性や重要性を視覚的に捉えやすく可視化することで，既知の過去の事象に基づいた将来のモデルの構築が可能となり，その結果，未知の事象の予測が期待できると考えられる．

本論文において我々は，時系列データの中でも，以下のように記述可能な時系列データに着目する．

- 時系列データ $A = \{A_1, A_2, ...A_m\}$, ただし A_j は j 番目の事象に関する各時刻の実数値の集合であり，m は事象の総数である．
- $A_j = \{a_{1j}, ..., a_{nj}\}$, ただし a_{ij} は j 番目の事象に関する i 番目の時刻における実数値であり，n は実数値を観測した時刻の総数である．

我々の身の回りにおいて，上述のような時系列データの多くは折れ線グラフとして表現されている．近年発表されている情報可視化手法にも，折れ線グラフとして時系列データを可視化するものが多い．しかし，大規模な時系列データを二次元平面上に描画すると，線分同士の交差が多くなり観察が困難になる場合が多く，煩雑な可視化結果になってしまうことが多い．この問題を解決するために，一画面上に描画する折れ線の本数を減らして可視化する，似た形状の折れ線グラフ群をまとめて可視化する，など数多くの手法が提案されている．これらの手法では，可視化結果の煩雑さは軽減されるものの，逆に重要な情報を失う可能性がある．

画面上での折れ線どうしの交差をなくし，また，データの情報量を減らすことなく時系列データを観察するための一手段として，三次元空間中に折れ線を描画することが考えられる．しかし，三次元空間内にランダムな順番で折れ線を配置したのでは，ユーザは折れ線同士にどのような性質の相関性があるのかを視認しにくくなり，データ中のどこに注視すればいいのかを判断するのに時間がかかる．結果として，時系列データの観察に非常に多くの時間を要したり，時には折れ線同士の相関性を理解しきれないこともあり得る．また，三次元空間内に折れ線を配置する場合，視点をどこに置くかによってユーザに与える印象は大きく変わってくる．よって，時系列データを三次元空間内に配置して可視化するためには，様々な工夫が必要であり，研究の余地は十分にあると思われる．

そこで本論文では，折れ線同士の相関性を考慮した，大規模時系列データの三次元可視化の一手法を提案する．図 1 に，本手法による可視化のイメージを示す．本手法は，x 軸に時刻，y 軸に数値を割り当てられた三次元空間中において，z 軸正方向に折れ線を等間隔に配置することで，時系列データを可視化する．また本手法では，前処理として折れ線同士の相関性を見つけ出し，それにしたがった順番で折れ線を配置する．これによってユーザは，時系列データ中にどのような性質の相関性が存在し，そのうちのどの相関性が注視するに値するかを視認しやすくなる．結果として本手法は，時系列データ中の興味深い局所部分をユーザに抽出させやすくする．

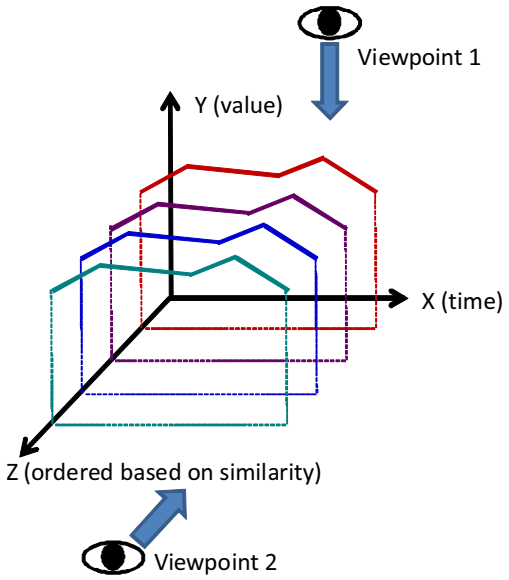

図 1: 本手法による可視化のイメージ．

図 2 に，本手法による可視化の実例を示す．我々の実装では，画面を左右に二分し，各々の領域に以下の描画結果を示す．

- 画面左半分 (図 2(左) 参照)：図 1 における xz 平面に垂直な視線 Viewpoint1 からの可視化結果を表示する．本論文では「真上視点部分」と称する．
- 画面右半分 (図 2(右) 参照)：図 1 における xy 平面に垂直な視線 Viewpoint2 からの可視化結果を表示する．本論文では「正面視点部分」と称する．

これらを併用することで本手法は，時系列データの「大局的な可視化」と「局所的な可視化」の相互利用を容易にする．我々が想定する本手法の典型的な操作手順は，以下の通りである．まず，ユーザは真上視点部分で折れ線グラフ群全体を大局的に観察する．続いて，詳しく観察したい局所的な少数の折れ線グラフ群を，クリック操作により選択する．すると本手法は正面視点部分に，選択した折れ線グラフ群を表示する．本手法では大局的に類似する折れ線にできるだけ近い z 座標値を与えるため，このクリック操作によって大局的に類似する折れ線群が選択されて，正面視点部分に表示される．ここで，真上視点部分では数値を色相で観察し，正面視点部分では数値を折れ線の高さで観察することになる．真上視点部分では折れ線どうしが画面上で重ならないので，多数の折れ線が互いに絡むことなく可視化されるのに対して，人間の視覚は色相の変化にあまり敏感ではないので，微細な数値結果を

視認することが難しい．それに対して，正面視点部分を用いて折れ線で数値を観察することで，色相で数値が表現された真上視点部分よりも詳細に数値変化を観察することができる．以上の仕組みによってユーザは，大規模時系列データの全体像を眺めながら，興味ある数値群を選択的に注視でき，折れ線同士の類似性（あるいは相違性）を注意深く観察できるようになる．

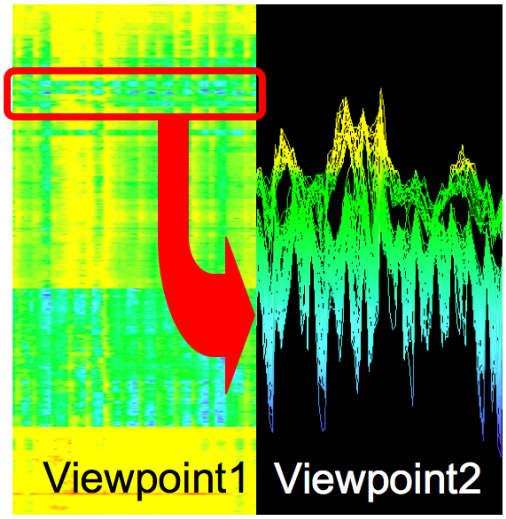

図 2: 本手法による可視化の実例.

また本手法では，SAX 法 [2]（Symbolic Aggregate approXimation）によって局所パターンを抽出し，それらをユーザの操作に合わせて真上視点部分に表示する．さらに本手法では，抽出した局所パターンに対して，数値変動や折れ線の形状に基づいたクラスタリングを適用し，いくつかのパターンに分類する．そして，出現回数が少ない例外パターン，または出現回数が多い頻出パターンを積極的にユーザに提示することで，ユーザにクリック操作を促す．この手法によって，ユーザは折れ線の大局的な類似性だけでなく，局所的なパターンによっても注視したい局所を見つけやすくなる．

以上をまとめると，本手法は以下の特徴において新規性を有する可視化手法であるといえる．

- 画面左側（真上視点部分）で時系列データ全体を俯瞰し，その中からユーザが興味を持った折れ線群を対話的に選択して，画面右側（正面視点部分）で注視する，という操作手順により時系列データを可視化する手法である．
- 大局的な類似度による並べ替えと，SAX 法による局所パターン抽出により，ユーザは真上視点部分から，大局的な特徴と局所的な特徴の両方に着目しながら，折れ線群を選択することができる．

我々は適用事例として，AMeDAS が観測した気温データの時系列データの可視化を試みた．この適用事例では，大規模な時系列データからユーザが目視だけでは気づきにくいと考えられる数値特徴を抽出することができた．

なお本論文の内容の一部は，既に著者自身による国際会議講演でも提案されている [3]．本論文はそれに対して，提案手法の技術的内容をより詳細化するとともに，適用事例における結果や考察を大幅に追加したものである．

2 関連研究

2.1 折れ線グラフによる時系列データの可視化

時系列データの可視化は活発に研究されており，その多くが二次元平面上に折れ線を描画する手法を提案している．大量の折れ線グラフ群を一画面に表示する際に，線分同士の交差が多くなり，表示結果が煩雑になるという問題点がある．それによって，ユーザは折れ線 1 本の数値変動を読み取りにくくなり，データの特徴を把握しにくくなってしまうと考えられる．

この問題点を解決するために，近年になって多くの手法が提案されている．その多くは，

1. 形状的に共通性のある折れ線だけを絞り込むようにして表示する．
2. 類似する折れ線のうち代表的な 1 本だけを表示し，形状の異なる折れ線どうしを表示する．

という 2 種類の考え方に基づいている．

前者の代表的なものとして Wattenberg らは，形状に基づくグラフ検索手法 [4] を提案している．ここでは，ユーザが求めるグラフ全体の形状を画面上で描画し，システムはこれに近い形状を持つ折れ線を複数提示している．また Hoccheiser[5] らは，複数の長方形を折れ線グラフ群に配置させるユーザインタフェースにより，類似する折れ線グラフを選択させる手法を提案している．この可視化表現では，縦横の長さを自由に設定することができる四角形を複数個用い，すべての四角形の内部を通るグラフのみを画面表示する．これにより，大規模なグラフ群の中から全体的な動向が類似しているグラフを抽出する．

後者の代表的なものとして Uchida ら [6] は，クラスタリングを利用した折れ線グラフの詳細度制御により，大局的に類似する折れ線のうち代表的な 1 本ずつだけを選び，異なる動きをしている折れ線どうしを画面表示する手法を提案している．本論文の本手法は，Uchida らの手法とは対照的に，大局的に類似する折れ線グラフ群を注視することを目的としており，それを選びやすくするために三次元可視化を導入した手法と位置づけられる．

また，大規模な時系列データの可視化においては，局所的な観察も重要である．Kincaid[7] は，長大な折れ線グラフの一部を滑らかに拡大表示する手法を提案している．この手法では，大局的な観察において，特異な動きをしていると考えられる局所的な区間を見つけ，その部分のみを拡大表示して観察することにより，大規模な時系列データの中から重要な箇所を見つけ出す．

2.2 折れ線グラフ以外による時系列データの可視化

あくまでも時系列データの可視化には，折れ線グラフが最もよく用いられており，多くのユーザの目に馴染んでいる．我々もそのような観点から，折れ線グラフをベースにした時系列データ可視化手法を提案している．

一方で折れ線グラフ形式以外にも，時系列データの可視化の研究は数多く発表されている．代表例として，時間軸を加えた三次元でのヒストグラム表示手法 [8]，時系列データの周期性を強調するために螺旋上に時間軸をとった表示手法 [9]，ThemeRiver[10] に代表される複数の折れ線グラフ要素の積み重ね表示手法，Two-Tone Pseudo coloring[11] に代表される二色塗り分け手法，などがあげられる．折れ線グラフをベースにした汎用的な時系列データ可視化手法と比べて，これらの手法は，特定の性質を有する時系列データの可視化，あるいは特定の実用目的に限定した時系列データの可視化，などに対して有効であると考えられる．

2.3 時系列データのクラスタリング

大規模な時系列データの中から数値特徴を抽出するためには，時系列データに特化したクラスタリング手法も有用である．本論文の本手法のうち 3.2 節および 3.3 節で提案する内容は，時系列データの順列化やクラスタリングに関するものであり，将来的には以下のクラスタリング手法との融合も考えられる．

Hartmut ら [12] は，金融市場の時系列データを題材として，分野が異なる 2 つの分野の時系列データに対してクラスタリングを行う手法を提案している．この可視化表現では，クラスタの個数が最適になるように設定した後，クラスタリングを実装している．このとき，クラスタの代表となる折れ線の形状を表示するのみでなく，クラスタに含まれる折れ線の形状をユーザに提示する．このとき，折れ線の形状がどちらの分野の時系列データであるかを色分けして表示している．これにより，ユーザは特異的な折れ線の形状を容易に見つけることができる．また，中村ら [13] は，時系列データをベクトル列として扱い，ベクトルの角度を比較して類似度を測定する類似度測定手法 AMSS(Angular Metrics for Shape Similarity) を提案し，瀧ら [14] は，AMSS を用いた検索のフィルタリングとして上界関数を導入した手法を提案している．その他にも，クモの生態系からヒントを得たアルゴリズムを用いたクラスタリング手法 [15]，ダイナミックタイムワピング (DWT:Dynamic Time Warping) 距離を用いて類似度を算出する手法 [16][17] なども提案されている．

2.4 SAX 法

本論文の本手法が採用する SAX 法（Symbolic Aggregation approXimation）[2] は，時系列データのパターン抽出および検索のために，時系列データを文字列によって表現する手法である．一般的な SAX 法の実装では，長さ n の時系列を任意長 w の文字列に変換する．このとき，変換後に計算した 2 データ間の距離が元データの距離の下限となるように距離尺度を定義している．これにより，変換後の距離の相対的大小が変換前と同じになる．また，次元圧縮を行うことで，データ処理に必要な計算コストが削減される．さらに，時系列データを 1 つの文字列で表現することにより，自然言語処理分野における文字列処理，言語解析などのアルゴリズムを適用することができる．

我々による SAX 法の実装では，図 3 に示すように，時間軸を等間隔に分割し，分割した時刻における折れ線の数値をアルファベットに変換に，折れ線を 1 つの文字列として表現する．このとき，アルファベットの出現回数が同程度になるように，y 座標値の境界値を設定する．例えば，図 3 では，折れ線のある局所的な部分は”CABCDE”という文字列で表現される．

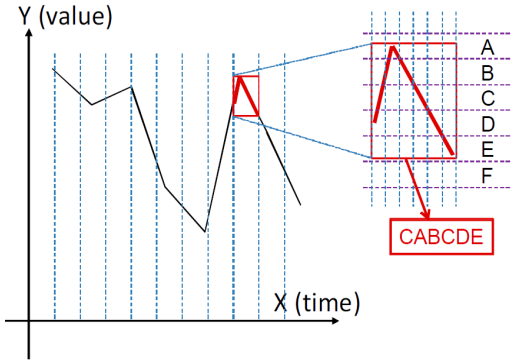

図 3: SAX 法を適用して数値群をアルファベット列で表現．

3 時系列データの三次元可視化手法の提案

1 章でも論じたとおり本章では，時系列データを構成する折れ線群を三次元空間に配置し可視化する一手法を提案する．本手法では，大局的な類似性が高いと思われる折れ線が近くに表示されるように，折れ線群を順列化し，その順列にしたがって各々の折れ線に z 座標値を与える．そして図 1 に示すように，x 軸を時間軸，y 軸を数値として，折れ線を z 軸上に等間隔に並べて可視化する．このとき，「xz 平面に垂直な視線方向を有する視点」「xy 平面に垂直な視線方向を有する視点」の二視点を併用することで，大局的な可視化と局所的な可視化の相互利用を容易にする．また，SAX 法によって局所パターンを抽出し，重ねて表示することにより，折れ線間の局所的な相関性を見つけやすくする．

3.1 概要

本手法は，以下の 4 つの処理手順で構成される．

1. デンドログラムを用いた折れ線の順列化

2. SAX 法を用いた局所パターンの抽出
3. 真上視点部分による折れ線グラフ全体の表示
4. 詳しく観察する特定の折れ線群の抽出，および正面視点部分による観察

以下の節では，これら 4 つの処理手順について説明する．

3.2 折れ線の順列化

時刻 i における j 番目の数値を a_{ij} とし，j 番目の数値群によって構成される 1 本の折れ線を $A_j = a_{1j}, \cdots, a_{nj}$ とする．このとき本手法は，以下のいずれかの判断基準により任意の 2 本の折れ線間の距離を算出し，デンドログラムを生成する．そして，生成されたデンドログラムによって折れ線群を順列化する．ここで距離の算出には，

- 同一時刻に近い値を有する傾向が大局的に見られる折れ線が近くに配置されるように，距離を算出する
- 同一時刻に近い値を有する傾向が局所的に見られる折れ線が近くに配置されるように，距離を算出する
- 同一時刻でなくてもいいから，また値が大きく異なってもいいから，大局的または局所的に類似形状を有する折れ線が近くに配置されるように，距離を算出する

など，いくつかのパターンを予め用意しておく．ユーザが必要に応じて表示するパターンを選択することで，目的にあった観点において類似する折れ線を比較観察できる．ここでは，同一時刻に近い値を有する傾向が大局的に見られる折れ線が近くに配置されるように折れ線を並べる場合について説明する．折れ線が N 本あるとする．このとき，任意の 2 本の折れ線 A_j，A_k 間の距離の算出には最短距離法を用いる．まず，時刻 i における 2 本の折れ線のユークリッド平方距離

$$\sqrt{(a_{ij} - a_{ik})^2} \tag{1}$$

を時刻 t_1 から時刻 t_n まで求め (図 4)，以下の式

$$S_n(j,k) = \sum_{k=1}^{n} \sqrt{(a_{ij} - a_{ik})^2} \tag{2}$$

で和 $S_n(j,k)$ を算出する．

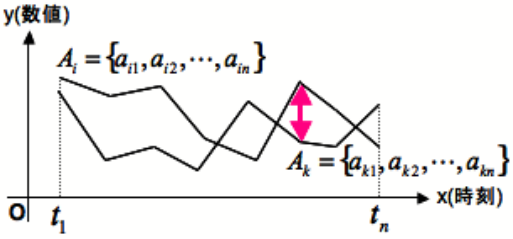

図 4: 2 本の折れ線のユークリッド平方距離．

次に，$S_n(j,k)$ の値が最も小さい 2 本の折れ線を併合し，1 つのクラスタ K_1 を生成する．ここで K_1 は，含まれている 2 本の折れ線の各時刻における数値の重心をその時刻での数値とし，それらにより生成される 1 本の折れ線とみなす (図 5)．

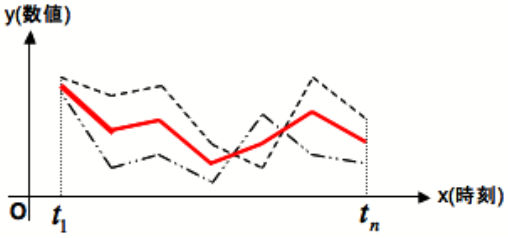

図 5: 2 本の折れ線の併合．

新しく折れ線を生成したことにより，折れ線の本数は N-1 本となる．この操作を繰り返し，折れ線の本数を減らしながらクラスタリングを行うことにより，デンドログラムを生成し (図 6)，数列群を順列化する．このとき，デンドログラムは折れ線を，$S_n(j,k) \leq S_n(l,m) \leq \ldots$ となるように，昇順に並べる．

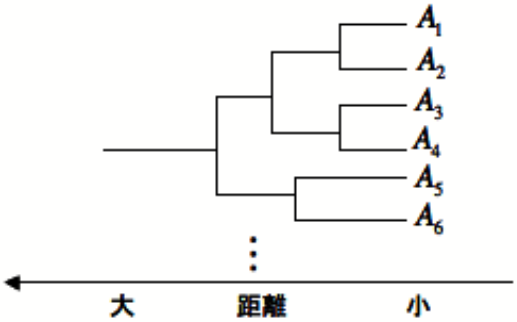

図 6: デンドログラムの生成．

3.3 SAX 法による局所パターンの抽出とクラスタリング

前節で説明した順列化によって本手法は，大局的に類似する数値群が隣接するように，時系列データを表現する折れ線群を並べ替える．それに対して時系列データの分析では，局所的な類似性，あるいは短時間の頻出パターン（または異端パターン）に注目することも重要であると考えられる．しかしながら，大規模な時系列データにおいて類似度を算出する場合，全ての局所間の組み合わせに対して類似度を計算していたのでは，膨大な計算コストがかかってしまう．そこで本手法では，時系列データの局所パターン抽出のために，2.4 節で紹介した SAX 法（Symbolic Aggregation approXimation）[2] を導入する．

本処理ではまず SAX 法を用いて，全ての時系列数値群をアルファベット列に変換する．続いてアルファベット列を等間隔（等文字数）に分割し，短時間のアルファベット列（局所パターン）を生成する．分割方

法については，周期性をもつ時系列データに対しては周期ごとで分割する，あるいはユーザが分割数を指定する，などが考えられる．現時点での我々の実装では，ユーザに分割数を指定させるようになっている．

続いて，以上の処理によって生成される局所パターンが何種類あるか数え上げる．多くの場合において，この局所パターンは数百種類，数千種類といった多種多様なものが生成されるため，この中から特定の局所パターンを対話的に指定するのは必ずしも容易ではない．そこで本手法では対話操作性の向上を目的として，上述の処理によって生成された局所パターンに対して，クラスタリングを適用する．現時点での我々の実装では，数値変動に着目したクラスタリング，折れ線形状に着目したクラスタリング，の 2 種類のいずれかを適用できるようになっており，ユーザにいずれかを選択させるようになっている．また，現時点の我々の実装では，非階層型クラスタリングの一手法である K-means 法を採用しているが，階層型クラスタリング手法を採用することも可能である．

3.4 折れ線グラフの全体表示

本手法では，xz 平面に垂直な視線方向を有する視点 Viewpoint1 と，xy 平面に垂直な視線方向を有する Viewpoint2 を併用することで，大局的な可視化と局所的な可視化の相互利用を容易にする．本手法ではまず，視点 Viewpoint1 により，折れ線グラフ群の全体を上から俯瞰し，画面左側に相当する真上視点部分 (図 2(左) 参照) に表示する．このとき，折れ線をリボン上の面で表現し，シェーディングを施すことで，立体感を表現しやすくする．

また本手法では，各時刻における色相を変化させることで，折れ線の各時刻の数値や変化量を表現する．具体的には，各時刻の数値を色で表現する場合には y 座標値が大きければ暖色，小さければ寒色を与え，各数値の変化 (微分係数) を色で表現する場合には傾きが正であれば暖色，負であれば寒色を与える．これにより，具体的な数値として観測することのできない y 座標値や微分係数を把握する．折れ線の各時刻の数値と各数値の変化のどちらに色を割り当てるかは，ユーザが時系列データの分析の目的に応じて選択・切替できるようになっている．例えば，図 7 は同じ時系列データを真上視点から俯瞰した図であり，図 7(左) は各時刻の数値に色を割り当てた場合，図 7(中央) は各時刻の微分係数値に色を割り当てた場合，図 7(右) は各時刻の微分係数値が正ならば赤，0 ならば緑，負ならば青を割り当てた場合である．図 7(中央) においては，似た色の変化をしている箇所が多くあることから，時系列データに周期性があるのではないかということが読み取れる．図 7(右) においても，同様に周期性を見つけることができる．さらに，時系列データの性質上微分係数値が 0 になることが非常に稀である場合，データの欠損や計測時における異常を知り得る場合もある．

真上視点部分では初期状態として，折れ線グラフ全体が一画面上に表示される位置に視点を配置するが，ユーザは視点を自由に操作できる．y 軸方向の平行移動で一画面上に表示する折れ線群の本数を調整し，x 軸・z 軸方向の平行移動で時刻や表示する折れ線群を変える．ユーザは y 軸方向の平行移動の操作はマウス右ボタンのドラッグ操作，x 軸・z 軸方向の平行移動の操作はマウス左ボタンのドラッグ操作によって行い，折れ線グラフ全体を観察する．

さらに我々の実装では，図 8(右) に示すように，SAX 法によって抽出された局所パターンのクラスタリング結果から，各クラスタを構成する代表的な局所パターンを表示するウィンドウを生成する．このウィンドウでは，画面左側に数値変動に着目したクラスタリング結果を，画面右側に折れ線形状に着目したクラスタリング結果を表示する．このとき画面左側・右側ともに，クラスタリング結果のうち，短時間の頻出パターン（または異端パターン）上位数パターンを表示する．図 8(右) の例では，画面左側・右側ともに，異端パターン上位 5 パターンを表示している．ユーザはダイアログ部品を載せたパネル（図 8(左) を参照）の中のラジオボタンを選択することにより，これらの局所パターンがどの折れ線の，どの時間帯で現れているのかを xz 平面に垂直な視線 Viewpoint1 からの可視化結果で観察する．

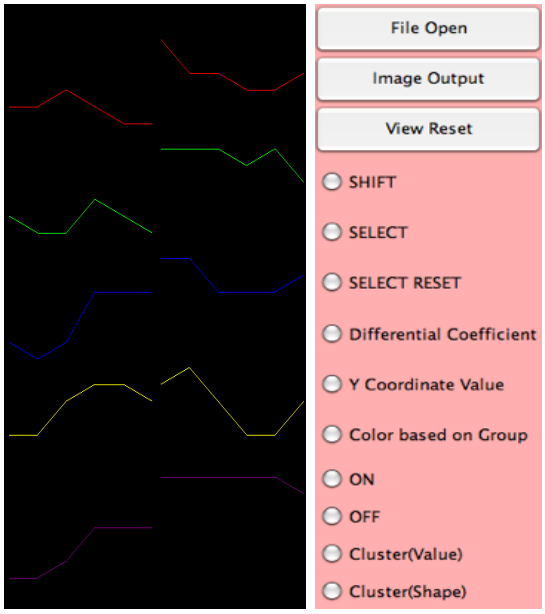

図 8: (左) 局所パターンのクラスタリング結果表示を選択するパネル．(右) 局所パターンのクラスタリング結果をユーザに提示するウィンドウ．この画面の左半分は，ラジオボタン Cluster(Value) を押すことで，数値変動に着目したクラスタリング結果を表示している．この画面の右半分は，ラジオボタン Cluster(Shape) を押すことで，折れ線形状に着目したクラスタリング結果を表示．

上記のダイアログウィンドウでユーザが特定のパターンを選択すると，真上視点部分では，これらのパターンが現れている箇所にユーザが興味を引くように，以下の 2 種類の表現

- パターンが現れている箇所を四角枠で囲む
- パターンが現れている箇所は明るく，現れてい

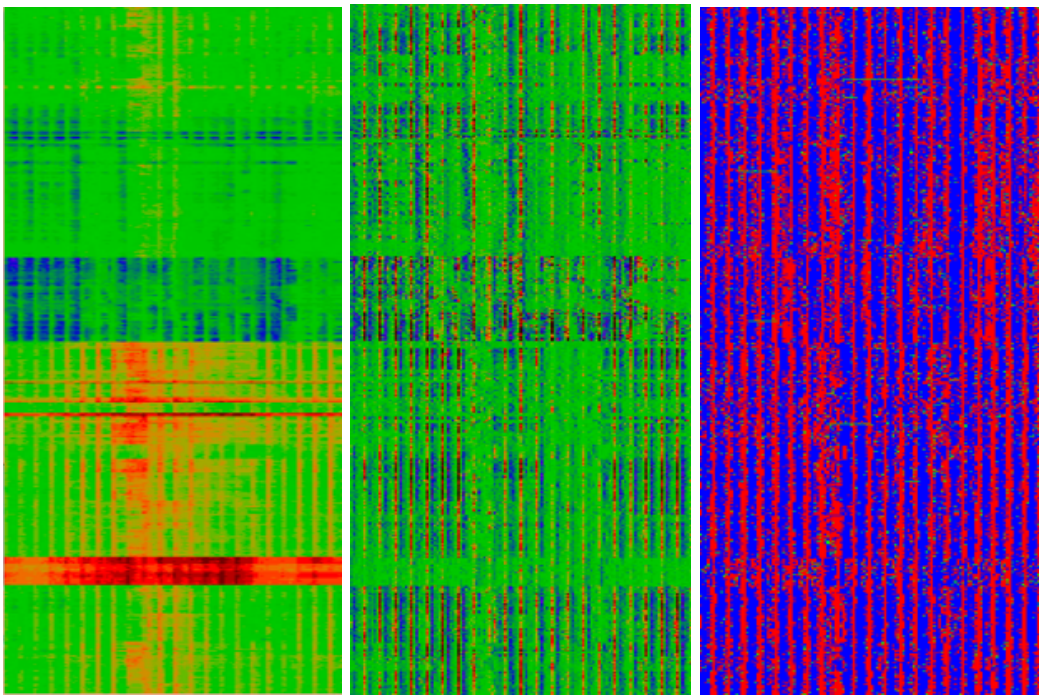

図 7: (左)y 座標値に対して色を割り当て．(中) 微分係数値に対して色を割り当て．(右) 緑色部分はデータの欠損である可能性が高い．

ない箇所は暗くする．

のいずれかを用いることで，パターンが現れている箇所を明示する．また，ユーザへの提示に関しては，これらの画面左側，右側の折れ線の色と四角枠の色を同じにすることで，ユーザはどの折れ線のどの時間帯にどのような数値変動，または形状をしていたのかを，真上視点部分を観察することで，短時間で把握することができる．

このように，局所的な類似性があると考えられる特定のパターンの箇所をユーザに提示することで，ユーザは興味深く観察すべき箇所を短時間に見つけ出すことができ，作業の効率化を図ることができる．また，大規模な時系列データから局所的な類似性を見落とすことなく観察することができ，重要な相関性を見つけ出せる可能性が広がると考えられる．

3.5 少数の折れ線群抽出

本手法では図 9 に示すように，特に着目したい折れ線群を発見したら，その近くに視点を移動させてズームアップし，2 回のクリック操作により少数の折れ線群の範囲を指定する．このとき，ユーザは複数の範囲を同時に選択することが可能である．また我々の実装では，一つの範囲に含まれる折れ線の本数を制限していない．

この操作に伴って，画面右上では視点 Viewpoint2 により，クリックされた範囲に含まれる折れ線群を可視化する．この場合も，視点は指定した折れ線群が全て一画面上に表示されるように配置しているが，拡大縮小・平行移動などの視点操作をユーザが自由にできるようになっている．

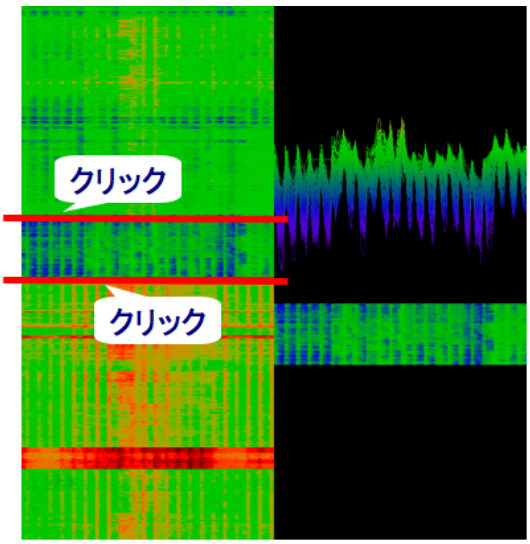

図 9: 少数の折れ線群の選択と表示．

また画面右下では，視点 Viewpoint1 により，クリックされた範囲に含まれる折れ線群を可視化する，という機能を必要に応じて選択的に起動できる．このとき x 軸・y 軸方向の平行移動で，時刻や表示する折れ線群を制御できる．また，このとき，真上視点同様に折れ線の各時刻の数値や各数値の変化を色で表現することができる．正面視点では，クリックした範囲 (グループ) が複数あった場合に，グループに色を割り当てることもでき，どのグループに含まれている折

れ線なのかが分かるようになっている．また，クラスタリングの結果をより分かりやすくユーザに提示するために，色を用いてクラスタを区別することもできる(詳細は後述)．また，このとき，画面右上と画面右下が連動して平行移動することで，2 つの視線から同一の折れ線群を観察することができる．

以上によって，大量の折れ線の中から注目する類似折れ線群を抽出し，その関連性や差異を観察できる．折れ線グラフ全体を上から俯瞰することで，重要な意味をもつと考えられる折れ線を見落とすという問題点は解決されると考えられる．

4 適用事例

我々は，日本の気象観測システム AMeDAS（Automated Meteorological Data Acquisition System）が 2006 年 1 月に観測した全国 916 地点の気温を集めたデータを本手法に適用した．本データでは 4 時間おきに観測した 31 日間の気温を時系列データとしている．結果として，

クリック操作による折れ線群の抽出

図 10 は，真上視点部分からのクリック操作によって折れ線群を抽出した例である．この折れ線群は，時間変動が非常に類似しており，総じて朝と夜の気温差が小さい．抽出された折れ線は主に秋田県・福島県・栃木県北部の気温であった．よって，内陸・日本海側で同じような気温変化をしていることが分かった．本手法を用いることで，ユーザはクリックして折れ線群を抽出するという比較的簡単な操作によって，大規模な時系列データの中から非常に類似した折れ線を容易に抽出することができた．

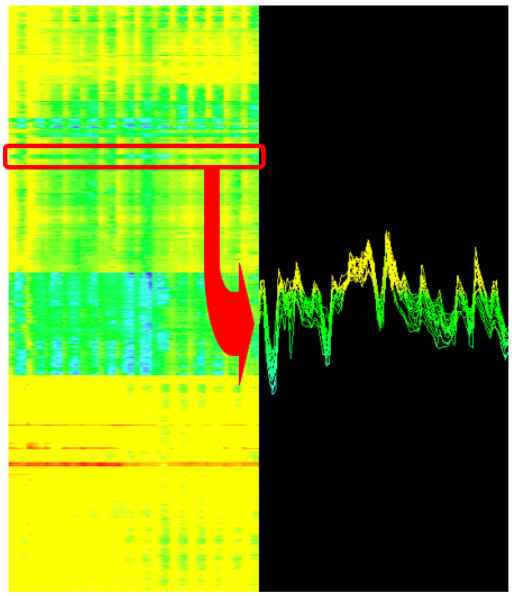

図 10: 非常に類似した数値変動の折れ線を抽出.

また，図 11 においても，図 10 と同様に真上視点からクリックして抽出した折れ線群は，時間変動が非常に類似している．抽出された折れ線は主に茨城県・石川県・奈良県の気温であった．よって，これらの地域は，この時期に気温差の小さい天候が続いていたことがわかる．また，地理的に離れている地域でも，同じような気温変化をしている時間帯があったことがわかる．ここで興味深い点として，図 11 においては，円で囲まれた時間帯においては，気温が高く上がった地域と低い気温が続いていた地域に分かれている．つまり，この期間においては，非常に類似する気温変化が見られた 3 つの地域について，ある特定の日だけでは全く異なる気温変化をしていた，ということがわかる．これは，真上視点からでの認識は難しいが，正面視点からでは明確に認識することができる．

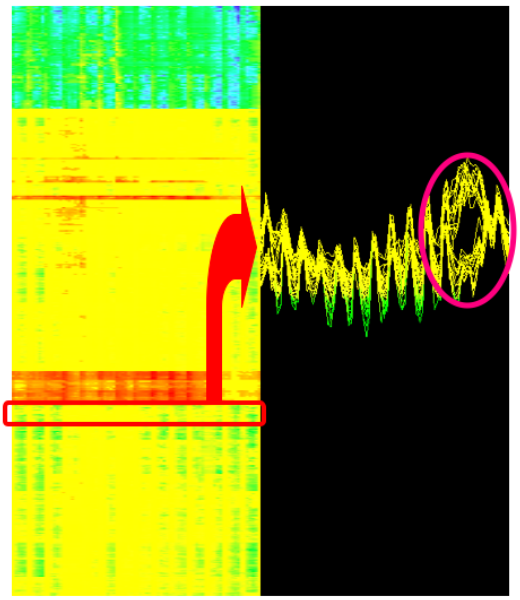

図 11: 大局的に類似するが局所的に差異のある折れ線群の発見.

図 12 は，図 10,11 にも示してある真上視点部分から，さらに別の折れ線群を選択した結果である．ここで図 12(上) の丸で囲んだ 4 箇所にわたって，青い折れ線が見られる．しかし 1 章でも論じたとおり，人間の視覚は色相の変化に敏感ではないので，この 4 箇所の数値の大小関係を色相のみから判別するのは難しい．しかし図 12(下) の折れ線グラフを観察することで，この 4 箇所の数値の大小関係を容易に視認できることがわかる．

以上の結果から，本手法の「真上視点からの大局的な可視化」と「正面視点からの局所的な可視化」を組み合わせることの有用性が示されている．

SAX 法による局所パターン抽出

図 13(左) は SAX 法を適用した可視化例である．我々は時系列データを 4 時間ごとに区切って，各々の区間に 7 種類の文字を割り当てた．そして，1 日の気温変動に相当する 6 文字を 1 つの文字列とし，出現回数が 300 以下であった文字列を異端パターンとし

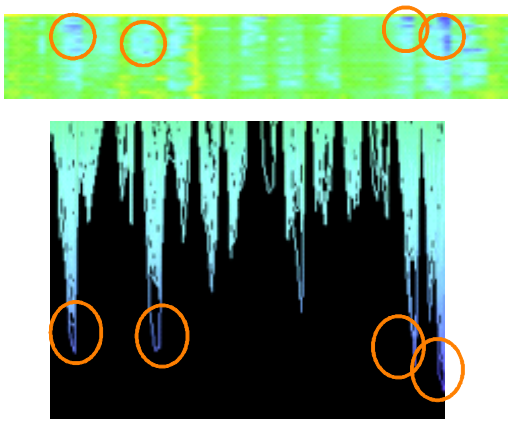

図 12: 正面視点部分の有効性を示す折れ線群の例.

て抽出した．さらに，抽出した異端パターンの中から 1 パターンを紫の枠で表示した．

ここで興味深い点として，真上視点からクリックした 2 つの範囲（以下,「グループ赤」「グループ青」と呼ぶ）は，全体的な気温の時間変動が非常に類似しており，また，桃色の四角形内部には，抽出されたパターンが多く存在する．一方で，橙色と黄色の四角形内部には，グループ青には抽出したパターンがないが，グループ赤には抽出したパターンが多く存在する．正面視点で描画してみると，グループ赤は 1 日の気温変動が他とは異なっており，日中気温が上がらなかったことがわかる．

なお，図 13(右) は真上視点部分の可視化結果において，紫の枠を表示していない場合を示したものである．このように，局所パターンを明示しない真上視点部分においても，ユーザが色の微細な変化を読み取ることは不可能ではない．しかし，大規模な時系列データの可視化において，このような局所パターンを色だけから発見するのは難しい場合が多いと考えられる．以上のことから，局所パターンを真上視点部分で明示する意義は大きいと考えられる．

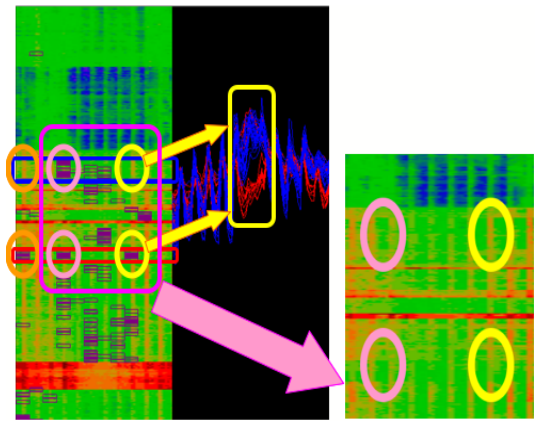

図 13: SAX 法を適用し，特定の局所パターンをユーザに提示した可視化例.

局所パターンのクラスタリング結果の提示

図 14，図 15，図 16 は SAX 法を適用した後に，文字列に対して K-means 法によるクラスタリングを適用して 25 種類の局所パターンに分類し，ユーザに提示した例である．

図 14 は，真上視点において出現回数が多い 5 つの局所パターン (頻出パターン) を色のついた枠を用いて表した可視化例である．枠の色は出現回数が多い順に，赤，緑，青，黄，紫となっている．頻出パターンを枠を用いて表すことで，1 月に頻出する気温変動をしていた日や観測所をユーザはすぐに見つけることができる．図 14(左) において，赤色の四角形で囲まれた箇所は，5 日間にわたって頻出パターンが続出したことを示している．しかし，この赤色の四角形に該当する地点では，その前後の日に頻出パターンがほとんど見られなかったこともわかる．このことから，赤色の枠が示す 5 日間では，非常に広い地域にわたって似たような気温変動が示されていたことがわかる．また，枠で囲まれなかった箇所は出現回数が少ない局所パターン (異端パターン) の気温変動をしている可能性があり，これらの箇所に着目して観察することが新たな発見につながり得る．例えば，紺色の四角形で囲まれた観測所は 1ヶ月のほとんどが緑色の枠で囲まれているのに対して，橙色，黄色，桃色の四角形で囲まれた箇所は緑色の枠で囲まれておらず，異端パターンであるとユーザは容易に想像しうる．図 14(右) は図 14(左) において橙色，桃色の四角形で囲まれた箇所を正面視点から表示したものである．これらは，最低気温は他の日と大きく変わらないものの，最高気温が他の日よりも高く，他の日と比べて暖かい日であったことが分かった．

図 15(上)，図 16(左) は，真上視点において出現回数が少ない 5 つのパターン (異端パターン) を色のついた枠で表示した可視化例である．枠の色は出現回数が少ない順に，赤，緑，青，黄，紫となっている．紺色の四角形で囲まれた観測所は，赤色の四角形で囲まれた日に異端パターンの気温変動をしていることがわかった．このことから，この日だけ他の日と気温変動が異なっている可能性が示唆される．これらの観測所の折れ線を正面視点 (図 15 右下) で観察してみると，他の日よりも気温が下がらなかった日であったことがわかった．このような異端パターンを，色のついた枠でユーザに提示しない場合 (図 15(左下) 参照)，色の変化が微小であるため，ユーザが発見しえない可能性は十分にある．以上のことから，異端パターンを真上視点部分で明示する意義は大きいと考えられる．

また，図 16(左) において，紺色の四角形で囲まれた 2 日は，どちらの日にも同じ色のついた枠が表示されている観測所が多くあったのではないかと推測できる．よって，似たような気温変動をしていた日であった可能性が高く，ユーザが興味を持つのではないかと考えられる．また，図 16(右) は，出現回数が少ない 5 つのパターンの代表的な形状を表示したものであり，折れ線の色と枠の色は対応している．それぞれのパターンの形状をユーザに提示することにより，異端パターンの中でも特にユーザが興味を持っているパターンがどこに現れているかを把握しやすくなり，より効率よく観察することができる．

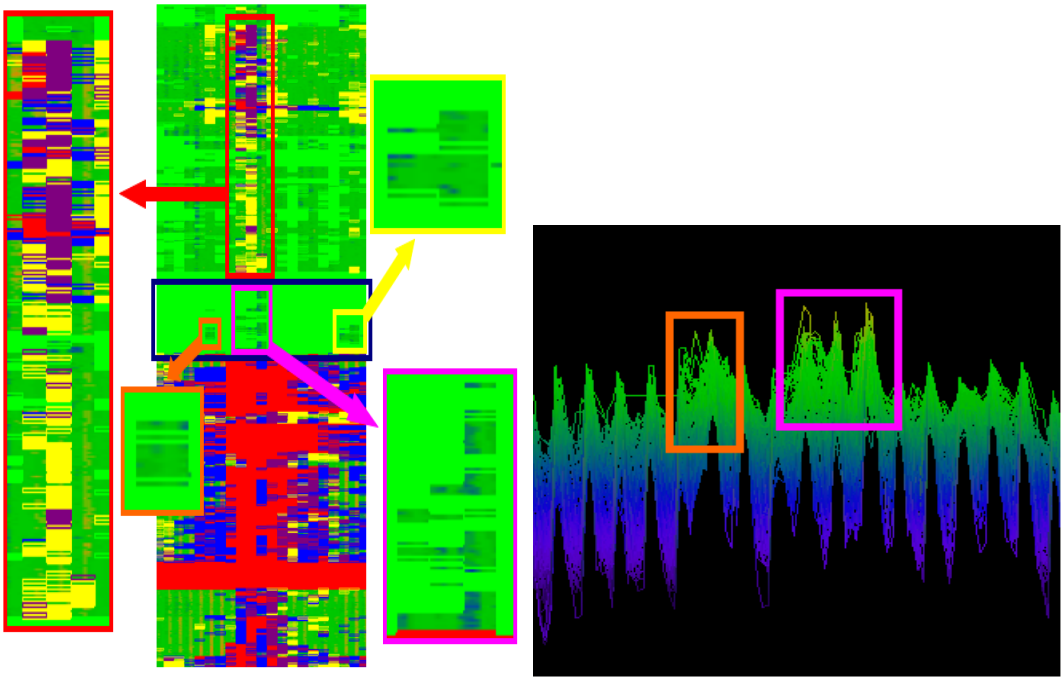

図 14: (左) 出現回数が多い 5 パターンをユーザに提示．(右) 出現回数が多いパターンに含まれていなかった箇所を正面視点で表示．

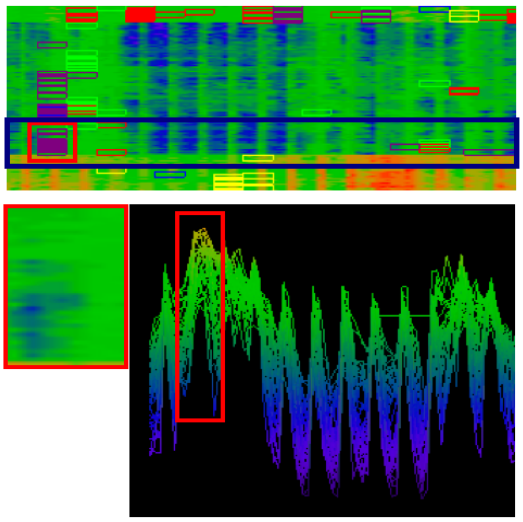

図 15: 出現回数が少ない 5 パターンをユーザに提示．

このように本手法を用いることで，まず大局的に時系列データを眺めて興味深い類似データ群を抽出し，続いてそれらを局所的に比較することが容易になる．

処理時間

上述のデータを用いた際の現時点での我々の実装による処理時間は，以下のとおりであった．

現在の実装では，3.2 節で論じた折れ線の順列化に，非常に大きな計算時間を要している．そこで今後の課題として，デンドログラム形成部分の実装の高速化が必要である．また，同じデータを何度も可視化するような用途に備えて，デンドログラム構築結果，および局所パターンの抽出結果をファイル出力する機能の開発が必要であると考える．一方で，局所パターンのクラスタリングの処理時間は相対的にみて非常に小さい．このことから，時系列データをいったん文字列化して抽出した局所パターンをクラスタリングする，という本手法の方針が処理時間の観点で妥当であることがわかる．

表 1: 処理時間の測定結果（単位：ミリ秒）

ファイル読み込み・折れ線の順列化	28234
折れ線の文字列化・局所パターンの抽出	8811
局所パターンのクラスタリング	268
折れ線等の描画	741

また現在の実装では，折れ線等の描画に 1 秒弱の処理時間を要しており，クリック操作による折れ線群選択や，表示すべき局所パターンの選択において，操作にストレスを感じることがある．この問題に対しては，描画部分の実装に GPU プログラミングを導入するなどの解決策が考えられる．

5 今後の展望

今後の展望として，以下の点が挙げられる．

折れ線順列化アルゴリズムの検討． 現時点での我々の折れ線順列化の実装では，ユークリッド空間で折れ線間の距離を算出しているが，今後は他の空間での距離算出も検討する予定である．また一般的に，デンドロ

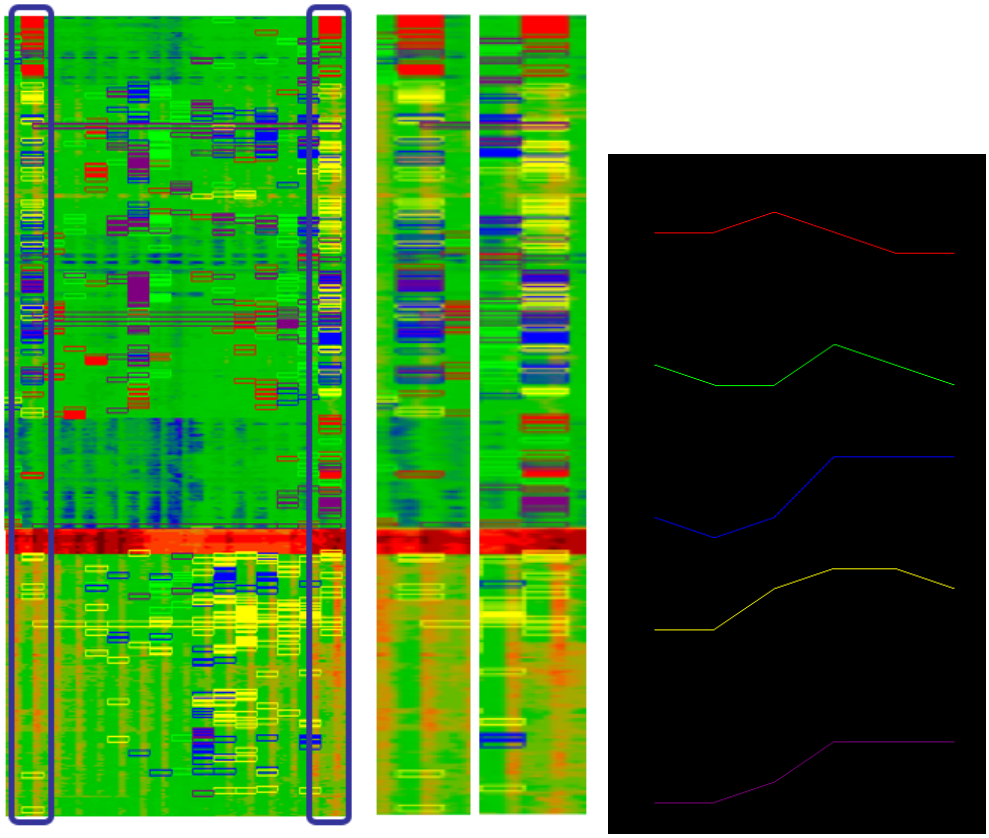

図 16: (左) 似たような気温変動をしていた日を推測．(右) 異端パターンの形状をユーザに提示．

グラムの生成結果は折れ線間の距離だけからは一意にならないため，折れ線の並べ順も一意には決定しないという問題点がある．それによって，全体の類似度が高い折れ線に近い z 座標値が与えられていない可能性があり，ひいては折れ線同士の相関性にユーザが気づけない可能性が懸念される．この問題については，順列化の過程において一意性の高いルールを加えることにより解決したい．

SAX 法の実装に関する検討． SAX 法は時系列を文字列で表すため，もとの数値の厳密性が失われてしまい，微細かつ特異な特徴が失われてしまう可能性もある．ここで，もとの数値の精度は，数値に対して割り当てるアルファベットの個数に依存する．我々は現在，アルファベットの個数を手動で決定しているが，それに対してアルファベットの最適な個数を自動設定する手法の確立を重要な課題の 1 つとしている．題材にする時系列データに応じて，比較実験などで経験的に結果を蓄えながら，最適な個数を自動設定できるようにしたい．また，Kmeans 法のクラスタ数についても，現在は手動で設定している．これについても，計算量と操作性の兼ね合いを考慮しながら，クラスタ数の自動決定手法の確立についても検討していきたい．我々は本手法について，ユーザとの対話的操作を重要視していることから，計算コストの低減も同時に実現する方向で検討すべきであると考えている．

三次元空間の表示に関する検討． 本手法は時系列データを表現する折れ線群を三次元空間に配置しているが，現時点での表示方法は xy 平面と yz 平面に写像した 2 種類の二次元可視化手法の組み合わせとなっている．これは製図において三面図だけを用いて三次元空間を表示していることに相当しており，必ずしも折れ線群を三次元空間に配置した意義を十分に発揮したとは言い切れない．例えばこの三次元空間を斜めから三次元らしく表示することで，時系列データ可視化における別の視認性を生むことはできないか，といった点について今後検討したい．

他の時系列データへの適用と新たなパターンの抽出． 現在我々が題材としている時系列データは AMeDAS の気象データのみである．我々は，様々な時系列データの可視化に適した手法の提案を目指しており，AMeDAS の気象データ以外の時系列データに対しても本手法を適用し，本手法の有用性を示したいと考えている．具体的には，以下のような時系列データへの適用を視野にいれている．

- 工業製品や基盤システムの計測情報．例として自動車の車載センサの時系列データや，原子力プラントの計測情報の時系列データなど．
- 生物・化学・生命情報に関する実験値．例として遺伝子マイクロアレイ実験から得られる発現量の時系列データなど．

- 連続的に測定した医療情報．例として血圧や脈拍数などの健康診断値を長時間にわたって計測して得られた時系列データなど．
- 経済・金融・流通などの社会科学的な時系列情報．例として株価や商品物流量の変動に関する時系列データなど．

また，現在題材としている AMeDAS の気象データにおいて，観測地ごとの気温変動パターン，または同日における気温変動パターンの相関性について検討したいと考えている．

ユーザテストによる定量的評価．本論文では我々自身による可視化結果の解説と評価のみを掲載しているが，一方で第三者によるユーザテストを評価に用いることも重要と考えている．一例として，本手法を用いてユーザに時系列データを提示し，その中に潜む現象に関する設問を与え，その正解率や回答所要時間などを計測することで，定量的に本手法の有効性を検証したい．また,「真上視点部分」または「正面視点部分」の片方だけを提示した場合と，両方を提示してユーザに対話操作させた場合の正解率や回答所要時間の比較，あるいは著者自身による先行研究 [6] との正解率や回答所要時間の比較，なども重要であると考える．

6 まとめ

本論文では，時系列データを表現する折れ線群を三次元空間 (xyz 空間) 内に配置し可視化する一手法を提案した．本手法では，大局的な類似性に基づいた並び順で折れ線群を表示することで，大局的に類似する時系列数値の抽出を容易にする．また，SAX 法によって抽出された頻出パターンや異端パターンに着目することで，局所的な特徴を有する時系列数値の抽出も容易にする．

参考文献

[1] B. Shneiderman, *The Eyes Have It: A Task by Data Type Taxonomy for Information Visualization*, IEEE Symposium on Visual Languages, 336-343, 1996.

[2] J. Lin, E. Keogh, S. Lonardi, B. Chiu, *A Symbolic Representation of Time Series, with Implications for Streaming Algorithms*, 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2003.

[3] M. Imoto, T. Itoh, *A 3D Visualization Technique for Large Scale Time-Varying Data*, 14th International Conference on Information Visualisation (IV10), 17-22, 2010.

[4] M. Wattenberg, D. Jones, *Sketching a Graph to Query a Time-Series Database*, SIGCHI Conference on Human Factors in Computing Systems Extended Abstract, 381-382, 2001.

[5] H. Hochheiser, B. Shneiderman, *Dynamic Query Tools for Time Series Data Sets: Timebox Widgets for Interactive Exploration*, Information Visualization, 3(1), 1-18, 2004.

[6] Y. Uchida, T. Itoh, *A Visualization and Level-of-Detail Control Technique for Large Scale Time Series Data*, 13th International Conference on Information Visualisation (IV09), 80-85, 2009.

[7] R. Kincaid, *SignalLens: Focus+Context Applied to Electronic Time Series*, IEEE Transactions on Visualization and Computer Graphics, 16(6), 900-907, 2010.

[8] R. Kosara, F. Bendix, H. Hauser, *Timehistograms for Large, Time-Dependent Data*, Eurographics/IEEE TVCG Symposium on Visualization, 45-54, 2004.

[9] M. Webera, M. Alexa, W. Muller, *Visualizing Time-Series on Spirals*, IEEE Symposium on Information Visualization 2001, 7-14, 2001.

[10] S. Havre, B. Hetzler, L. Nowell, *ThemeRiver: visualizing theme changes over time*, IEEE Symposium on Information Visualization, 115-123, 2000.

[11] T. Saito, H. Miyamura, M. Yamamoto, H. Saito, Y. Hoshiya, T. Kaseda, *Two-Tone Pseudo Coloring: Compact Visualization for One-Dimensional Data*, IEEE Symposium on Information Visualization, 173-180, 2005.

[12] H. Ziegler, M. Jenny, T. Gruse, D. A. Keim, *Visual Market Sector Analysis for Financial Time Series Data*, IEEE Symposium on Visual Analytics Scienceand Technology, 83-90, 2010

[13] 中村哲也, 瀧敬士, 野宮浩揮, 上原邦昭, *AMSS*:時系列データの効率的な類似度測定手法, 電子情報通信学会論文誌, J91-D(11), 2579-2588, 2008.

[14] 瀧敬士, 中村哲也, 上原邦昭, 時系列の類似性検索における上界関数による効率化, 電子情報通信学会論文誌, J91-D(12), 2926-2938, 2008.

[15] 山村雅幸, 亀田祥平, 時系列クラスタリングのためのスパイダーアルゴリズム, 情報処理学会バイオ情報学研究報告, 64, 65-68, 2006.

[16] 高橋範亘, 櫻井保志, 義久智樹, 金澤正憲, ダイナミックタイムワーピング距離を用いたセンサストリーム処理システムの設計と実装, 電子通信学会第 19 回データ工学ワークショップ (DEWS2008), 2008.

[17] T. Oates, L. Firoiu, P. Cohen, *Clustering Time Series with Hidden Markov Models and Dynamic Time Warping*, Proceedings of the IJCAI-99 Workshop on Neural, Symbolic and Reinforcement Learning Methods for Sequence Learning, 17-21, 1999.

井元 麻衣子
2009 年お茶の水女子大学理学部情報科学科卒業．2011 年お茶の水女子大学大学院人間文化研究科理学専攻情報科学コース博士前期課程修了．2011 年より日本電信電話株式会社サイバーソリューション研究所勤務．

伊藤 貴之
1990 年早稲田大学理工学部電子通信学科卒業．1992 年早稲田大学大学院理工学研究科電気工学専攻修士課程修了．同年日本アイ・ビー・エム (株) 入社．1997 年博士 (工学)．2000 年米国カーネギーメロン大学客員研究員．2003 年から 2005 年まで京都大学大学院情報学研究科 COE 研究員 (客員助教授相当)．2005 年日本アイ・ビー・エム (株) 退職，2005 年お茶の水女子大学理学部情報科学科助教授 (准教授)．2011 年同大学教授，同大学シミュレーション科学教育研究センター長兼任．ACM, IEEE Computer Society, 情報処理学会，芸術科学会，画像電子学会，可視化情報学会，他会員．