Vol.2013-HCI-152 No.23
2013/3/14

色による特徴表現を用いた高次元データの可視化

小林 弘明[1,a)]　三末 和男[2]　田中 二郎[2]

概要：可視化結果を表示するディスプレイの画面領域には限りがあるため，高次元データの概観を得ることは難しい．そこで本研究では，限られた画面領域において高次元データの概観を得ることを目的とし，そのための表現手法として色付き Mosiac Matrix を開発した．色付き Mosaic Matrix は高次元カテゴリデータの可視化手法であり，データの特徴を色を用いて表現する．量的データをカテゴリ単位で表現することにより，レコード数の多い高次元データの可視化を可能にした．評価実験によって可読性を調査した結果，本表現手法が高次元データの概観を得る手法として有用であることがわかった．

High-Dimensional Data Visualization Using a Color Representation of Features

Hiroaki Kobayashi[1,a)]　Kazuo Misue[2]　Jiro Tanaka[2]

Abstract: Due to the displays' limitation in size, it is difficult to obtain an overview of high-dimensional data in the area of the display that displays the results of visualization. In this paper, we aimed to obtain an overview of the high-dimensional data in a limited area of the screen. We developed Colored Mosaic Matrix as a method to obtain an overview of high-dimensional data. Colored Mosaic Matrix is a visualization method for high-dimensional categorical data, using a color representation of the features. By representing the quantitative data in units of categories, it enables the visualization of high-dimensional data with a large number of records. As a result of investigating the readability by experiments, we have found our method to be useful in obtaining an overview of high-dimensional data.

1. はじめに

データを分析する際には，人間の直感的理解を支援するために，データを可視化して視覚的に表現することが有効である．可視化はデータから知識を抽出するための非常に有効な手段である．目的やデータに応じた可視化手法を用いることにより，より複雑な分析を行うことが可能になる．

世の中の様々な分野に現れるデータの多くは，複数の次元を持つ多次元データである．さらに次元数が 10 以上の多次元データは，一般に高次元データと呼ばれている．これらの多次元データを可視化して分析するため，旧来様々な可視化手法が研究されてきた．

例えば Scatterplot Matrix[1] は，全ての組み合わせ可能な次元対の Scatterplot を行列状に並べて表示する手法である．この手法は次元が規則的に並んでいるため，分析に必要な次元の探索が容易である．しかし可視化結果を表示する画面領域には物理的な限界がある．よって高次元データを対象にした場合，1 組あたりの描画領域が限られてしまい，全次元を可視化して概観を得るのは困難になる．

このような問題に対するひとつの対策は，データの一部の次元のみを表示することである．Sips らの手法 [2] では，距離とエントロピーによる次元選別を用いることで，分析に有用な部分のみを表示する．また VisBricks[3] では，表示したい次元やレコードをユーザが選択することにより，複雑なデータを表現可能にしている．しかしこれらの手法で高次元データを扱う場合，可視化する部分の選択や判定が難しい場合も多い．分析に必要な部分をどのように判定

[1] 筑波大学情報学群情報科学類
College of Information Science, School of Informatics, University of Tsukuba
[2] 筑波大学システム情報系
Faculty of Engineering, Information and Systems, University of Tsukuba
[a)] hiroaki@iplab.cs.tsukuba.ac.jp

するか，またどの程度まで表示するかが難しい問題になる．

高次元空間全体にわたるプロット間の距離関係や密度分布を保ちつつ，低次元空間で高次元データを可視化する手法も存在する．例えば RadViz[4], [5] は次元の軸を円周上に配置することにより，高次元データを 2 次元の描画領域上で表現する．しかしこれらの手法では，各次元の数値を直接読み取ることが難しくなる上に，データに対する誤解を招く可能性もある．

大画面ディスプレイを使う事で描画領域そのものを増やせば，表示する次元数を増やしたり，また全次元を一度に可視化することも可能になる．しかし単純に画面を大きくするだけでは，視認しなければならない面積が増加し，また視線の移動距離も長くなるため，分析が困難になってしまう．分析をスムーズに行うためにも，デスクトップで容易に見渡せる程度の画面領域内で高次元データを可視化できることが望ましい．

本研究では特に，扱いの難しい 30 次元以上の高次元データを対象とし，高次元データの概観をフル HD ディスプレイ（1920×1080）程度の画面領域内で閲覧可能にすることを目標とする．限られた描画領域において表示できる次元数を増やし，かつ高い可読性を保つことにより，正確な高次元データ分析を可能にする．

本研究における貢献は以下の 4 点である．

- 高次元データの概観を一度に閲覧可能な手法を開発したこと．本表現手法により，今までは難しかった高次元データの概観を表現する事が可能になる．
- 高次元データの概観を色で読み取る手法を開発したこと．色を有効に使うことで，非常に小さな描画領域におけるデータの可読性を向上させる．
- 量的次元に用いるカテゴリ分割手法を複数提唱し，それらの視覚的表現への影響について考察したこと．このカテゴリ分割手法については，その他のカテゴリデータ可視化手法に対しても利用可能である．
- 評価実験を行うことで，具体的な描画領域と可読性の関係性について考察を行ったこと．これにより，複数種類の描画領域における可読性を定量的に評価した．

2. 関連研究

2.1 Scatterplot Matrix

Scatterplot Matrix[1] は，全ての組み合わせ可能な次元対の Scatterplot を行列状に並べて表示することで，データ全体を俯瞰する．SCATTERDICE[6] は Scatterplot Matrix を拡張した可視化手法で，多次元データにおける各次元を切り替える作業を，サイコロを転がすようにインタラクティブに操作することで実現する．これにより，一般的な 2 次元の Scatterplot が持つシンプルさを活かしつつ，次元の切り替えを直感的に行うことができる．これらの手法に共通する問題として，高次元データの全次元を同時に表現しようとした場合，各 Scatterplot の描画領域が狭くなってしまうことが挙げられる．データの次元数が増えると，単位面積あたりの点の数が増えるためオーバープロットが発生し，読解や分析が困難になる．

2.2 Parallel Coordinates Plot

Parallel Coordinates Plot（PCP）[7], [8], [9] も Scatterplot Matrix と同様，複数の次元を一度に俯瞰できる可視化手法である．各次元に対して座標軸を用意して並列に並べ，隣り合った座標軸における点を全て結んでいくことにより，1 レコードを 1 本の線で表現する．これらの PCP を用いた手法は，線の密集度合いや線の傾きによってデータ分布や隣接次元間の関係性を表現する．しかし隣り合う次元同士でしか直接的な比較が行えないため，次元数の多い高次元データでは比較がより困難になる．また，高次元データを PCP で可視化すると次元軸間の幅が狭くなり，可読性が低下してしまう．

2.3 色でレコードの密集度合いや分布を表現する手法

Fua らの手法 [10] では，PCP におけるデータの密度に応じて明暗をつけることでデータ分布を表現している．また Feng らの手法 [11] は不確実データを対象にしており，PCP や Scatterplot を拡張して色でデータの密度を表現している．しかしこれらの手法は色の位置で分布を表すため，狭い描画領域で用いるには不向きな手法である．

Two-Tone Pseudo Coloring[12] は色相の分布のみで 1 次元のデータを表現する手法である．これは多次元データの可視化手法ではないが，小さく細い矩形からデータの特徴を読み取ることができる．

3. 視覚的表現

3.1 表現の設計方針

可視化を用いた視覚的操作は，まず全体を俯瞰し，ズーミングやフィルタリングを行い，さらに必要に応じて詳細を見る，という流れが理想的である [13]．高次元データを分析する場合においても，まずは可視化によって高次元データの概観を得られることが望ましい．そこで本研究では，データが持つ全次元を一度に可視化する手法を設計する．これはデータの概観を得るための最も分析が容易かつ直感的な方法である．しかし高次元データの全次元を同一画面上に表示すると，1 次元あたりの描画領域が狭くなり，結果として可読性が低下してしまう．高次元データを可視化して概観を得るためには，データの描画における空間効率が良く，かつ可読性の高い手法を用いる必要がある．

本研究では，空間効率の良い空間充填型の可視化手法に着目し，高次元データを俯瞰できる手法を開発する．また量的データをカテゴリデータとして扱うことで，オーバープロットを軽減する．さらに狭い描画領域において高い可

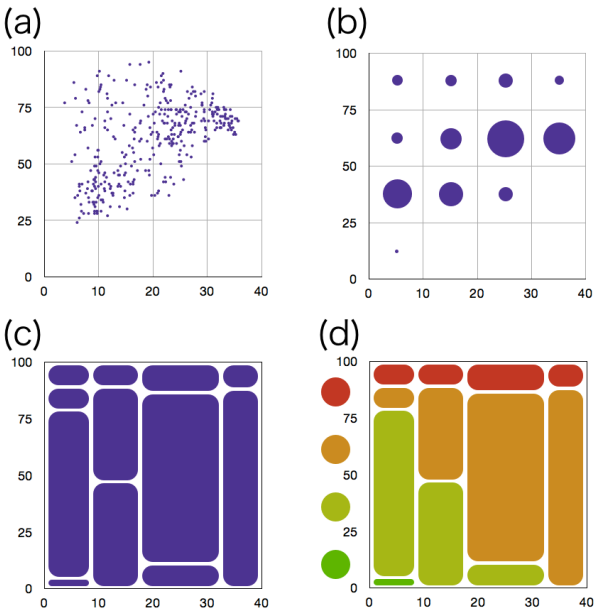

図 1 色付き Mosaic Plot の生成過程
Fig. 1 Generation process of Colored Mosaic Plot

読性を保つため，色相の分布や割合によってデータの特徴を表現する．

3.2 色付き Mosaic Matrix

まずは量的データをカテゴリデータへ変換する．図 1 の (a) と (b) は，変換前後のデータを模式的に表現したものである．量的データの各次元について，値域を細分化した上で各値域をカテゴリとして扱う．これにより少量の点でデータを表現できる．

次に，多次元カテゴリデータを対象にした空間充填型の可視化手法である Mosaic Plot[14], [15] を用いて，データの 2 次元分を可視化する．これは図 1(c) のように，各カテゴリの比率に応じて描画領域を矩形に分割する手法である．Mosaic Plot では，カテゴリの分布を各矩形の面積で表現する．しかし狭い領域では，各矩形の区別が困難になる．

我々は Mosaic Plot における各矩形の識別を可能にするため，図 1(d) のように各矩形に色付けを行う色付き Mosaic Plot を開発する．色付き Mosaic Plot は，色の割合や模様などの概観からデータの分布や特徴を読み取り可能にする．

さらに高次元データを可視化するために，全次元対の色付き Mosaic Plot を行列状に配置する．全次元対の色付き Mosaic Plot，すなわち色付き Mosaic Matrix を俯瞰することにより，高次元データ全体の特徴を把握できる．

3.3 色付けによる特徴の把握

色付け規則によってデータの見え方は変化し，それに応じて読み取れるデータの特徴も変化する．本研究では，データの特徴を俯瞰するための判断材料としてカテゴリの

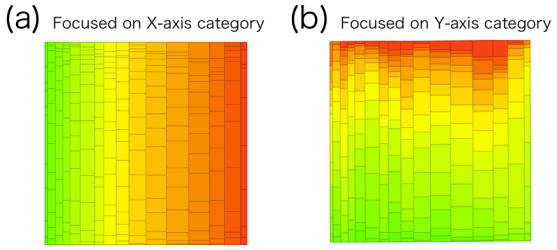

図 2 カテゴリに着目した色付け手法の例
Fig. 2 Examples of the coloring technique focused on category

分布と次元間の相関に着目し，それぞれに応じた複数の色付け手法を開発する．

3.3.1 カテゴリの分布に着目した色付け手法

この色付け手法の目的は，各 Mosaic Plot におけるカテゴリ分布を把握可能にすることである．しかし色のみで 2 次元分のカテゴリを同時に区別をしようとした場合，色の種類が増えることで色の区別が困難になる．そこで本色付け手法では，Mosaic Plot のいずれか 1 次元分のカテゴリに着目した色付けを行うことにより，可読性を向上させる．

色相はカテゴリの昇順に，緑から赤にかけてのグラデーションを採用する．Y 軸次元のカテゴリに着目した場合も，同様にして色相を求めて設定する．いずれの場合も，彩度と明度は全矩形において共通の値を設定する．

図 2 は，同じ Mosaic Plot に対して異なる色付けを行ったものである．図 2(a) のように X 軸次元で色付けを行った場合，X 軸次元の各カテゴリにおける幅は統一されているため，縦縞の模様に見える．縞模様の幅を見ることで，X 軸次元のカテゴリについて分布を読み取ることが可能である．例えば図 2(a) は黄色から赤にかけての色相が多く見られることから，高い値が多く存在していることがわかる．一方で図 2(b) のように Y 軸次元で色付けを行った場合，色相の分布から Y 軸次元のカテゴリについて分布を読み取ることが可能である．また各矩形の高さは X 軸，Y 軸の両次元のカテゴリに依存するため，色相の模様によって次元間の関係性を読み取ることが可能になる．

3.3.2 次元間の相関に着目した色付け手法

この色付け手法はデータの概観を推測可能にすることを目的とし，それぞれの Mosaic Plot に割り当てた 2 次元における相関係数を元に色相を決定する．この色付けは相関係数を計算して利用するため，元々が量的データの次元対に特化した色付けとなる．なお，相関係数は次元対に対して計算されるため，各矩形の色相は Mosaic Plot 内で同一のものとなる．この色付け手法では，各 Mosaic Plot における相関係数を用いて 2 通りの色付けを開発する．

(1) 色相を相関係数と色相環を対応付けることにより決定する．図 3 のように，相関係数が 1 の色相を緑，−1 の場合を赤とする．これにより，色相で相関係数の正負及び強弱を表現する．また明度と彩度を変更すること

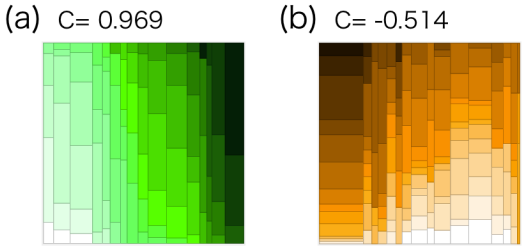

図 3 相関に着目した色付け手法の例 (パターン 1)
Fig. 3 Examples of the coloring technique focused on the correlation (pattern 1)

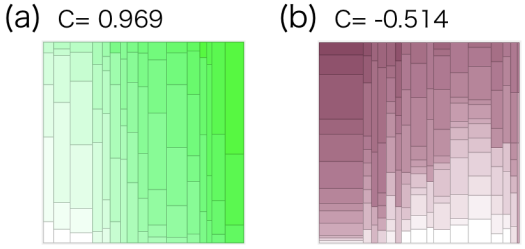

図 4 相関に着目した色付け手法の例 (パターン 2)
Fig. 4 Examples of the coloring technique focused on the correlation (pattern 2)

により，Y 軸次元のカテゴリ区別を行っている．この色付け手法の利点は，相関関係をある程度把握しながら，1 次元分のカテゴリを区別できる点である．しかし弱い相関同士を比較する場合，色付き Mosaic Plot の色相に差がほとんど現れないため判別が困難である．

(2) 図 4 のように，相関係数が 1 の色相を緑，−1 の場合を赤とする．また明度と彩度により，相関の強弱とカテゴリ区別を行っている．この色付け手法の利点は，相関の正負と強弱の判別がより容易である点である．

相関に着目した色付け手法は，カテゴリの識別が比較的困難となる．しかしこれは，他の色付け手法を併用することで補完できる．

3.4 量的データに用いるカテゴリ分割手法

量的データの次元については，値域全体を複数の値域に細分化し，各値域をそれぞれカテゴリとして扱う．カテゴリの値域はカテゴリ分割手法によって決定する．よって，カテゴリ分割手法は色付き Mosaic Plot における各矩形の形や大きさ，色付けなどに影響を及ぼす．そこで本表現手法では，データ分布や分析目的に応じて複数のカテゴリ分割手法を切り替え可能にする．

3.4.1 最大値と最小値を基準にした分割手法

このカテゴリ分割手法では，量的次元における最大値と最小値を元に値域を等間隔に分割する．そして分割された値域をそれぞれカテゴリとする．

この分割手法は最も直感的でわかりやすく，データ分布に関する誤解が少ないという利点がある．一方で外れ値が含まれる場合，各カテゴリの値域が広がってしまい，結果として殆どの値が一部のカテゴリに集中してしまう．カテゴリが一部に集中すると，詳細な分析が困難になる．

3.4.2 平均値と標準偏差を用いた分割手法

このカテゴリ分割手法は，外れ値の影響を減らすために，平均値と標準偏差を用いてカテゴリの値域を計算する．平均値をカテゴリの中心とし，標準偏差を元に両端以外のカテゴリの幅を決定する．なお両端のカテゴリについては，値域の下限または上限を設けない．

この分割手法では最大値及び最小値を値域の計算に利用しないため，直接的に外れ値の影響を受けないという利点がある．一方でデータが極端に偏っていると，カテゴリの両端にデータが含まれない場合がある．この場合，色付けによる誤解が生じる可能性がある．

3.4.3 データ量依存の分割手法

このカテゴリ分割手法は，データ量に応じて分割する領域と分割幅を決定する手法である．まず平均値を中心として一定のデータ量が含まれる値域を求め，これを分割値域とする．分割数を p とすると，分割値域内を等幅に $p-2$ 分割してカテゴリとする．また，分割値域外の値域をそれぞれカテゴリとする．

このカテゴリ分割手法は外れ値の影響が少なく，また一部のカテゴリにデータが集中しすぎないという利点がある．一方で色付き Mosaic Matrix を用いて高次元データ全体を俯瞰する場合，データがどの程度偏っているかを判断できないため，混乱および誤解を招く可能性がある．

4. ツールの開発

4.1 ツールの設計

開発した表現手法を用いて，概観から詳細へと掘り下げるドリルダウン式の高次元データ分析が可能なツールを開発する．本ツールでは概観を得ることが可能な Matrix View と，データの詳細な分析が可能な Detail View を設計する．

4.2 ツールに用いる表現手法

4.2.1 Matrix View

Matrix View は色付き Mosaic Matrix を用いることで，高次元データにおける全ての組み合わせ可能な次元対を行列状に表示する．色付き Mosaic Matrix の色を見ることで，高次元データの各次元対の特徴を把握する．Matrix View では一覧性を高めるために，それぞれの色付き Mosaic Plot における各軸のカテゴリ名は表示しない．

4.2.2 Detail View

Detail View では，任意の 1 つの次元対に対する色付き Mosaic Plot を詳細に表示する．対象の色付き Mosaic Plot を画面中央に大きく表示し，色付き Mosaic Plot の下部に X 軸次元，左部に Y 軸次元のカテゴリ名を表示している．

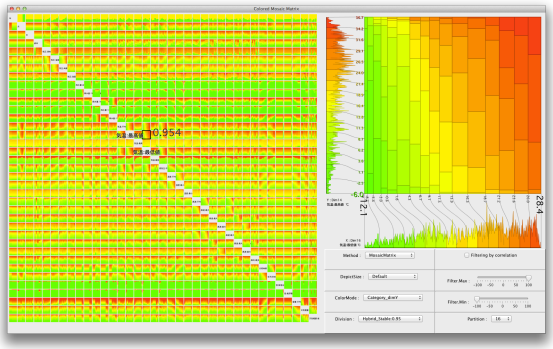

図 5 分析ツールのスクリーンショット
Fig. 5 Screenshot of the analytical tool

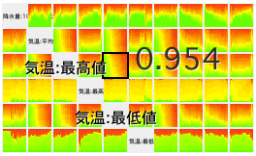

図 6 相関係数の表示例
Fig. 6 Example of displaying the correlation coefficient

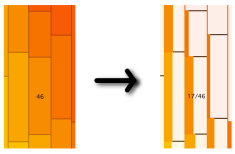

図 7 レコード選択による変化
Fig. 7 Changes due to the record selection

4.2.3 Area Graph

データ分布の読解を容易にするため，Detail View では Area Graph を用いて各軸の次元の分布を表示する．Area Graph は量的データにおける各値の頻度を表現する．また各カテゴリの値域を視覚的に表現するために，カテゴリの切れ目となる値とそのカテゴリ名を曲線で結んでいる．

4.3 ツールのインタフェース

図 5 は開発した分析ツールのスクリーンショットである．ツール画面の左部には Matrix View を，右上部には Detail View をそれぞれ表示している．またツール画面の右下部には，可視化手法の切り替えやフィルタリングを行うための操作パネルが配置されている．

Matrix View において，マウスポインタ直下にある色付き Mosaic Plot の各軸次元名を拡大表示する．また対象となる色付き Mosaic Plot の両次元が量的データである場合，図 6 のように，対象の次元対における相関係数をマウスポインタ付近に表示する．Matrix View と Detail View は関連付けされており，任意の色付き Mosaic Plot をクリックすることで，その色付き Mosaic Plot を Detail View で詳細表示する．

Detail View において，マウスポインタ直下にある矩形に対応するレコード数を表示する．さらに矩形をクリックすることで，その矩形に該当するレコード群を選択することが可能である．レコード群を選択している時は，図 7 のように，被選択レコードの量に応じて各矩形を 2 つに分割する．これにより，矩形内における被選択レコードの割合及び分布を色によって読み取ることができる．

また被選択レコードのみを対象として，ツールの画面を再描画することができる．特定の条件を満たすデータのみをフィルタリングして表示することで，データのより詳細な分析を行うことができる．

5. 評価実験

本実験は，色付き Mosaic Plot の有用性に関する評価を行うことを目的とする．有用性については，読解の正確性，読解の所要時間，各表現の理解しやすさの観点から評価する．さらに色付き Mosaic Plot の評価を元に，色付き Mosaic Matrix について考察する．

5.1 概要

高次元データを分析する上で必要となる，量的データの分布読み取りに関するタスクを設定する．色付き Mosaic plot と Scatterplot を組み合わせた実験ツールを用いて，被験者にタスクを行ってもらう．タスク終了後，色付け手法に関するアンケートに回答してもらう．タスクの正答率を元に，各条件下における本手法の可読性を評価する．

5.1.1 被験者

本ツールの対象ユーザとしては，可視化によるデータ分析を行う専門家を想定している．そこで本実験では被験者として，情報可視化を研究分野としている 7 名を選定した．なお実験を行う前に色盲検査を行い，全被験者の色覚が正常であることを確認した．

5.1.2 実験の手順

被験者は以下の手順の通りに実験を行う．

(1) 実験ツールに用いる表現手法に関する説明を受ける．
(2) 実験ツール及びタスクに慣れるまで，練習用のタスクを解く．このとき，必要に応じて追加説明を受ける．
(3) 本番用のタスクを全て解く．被験者は自由に休憩を取ることができ，また表現手法の説明書をいつでも見ることができる．
(4) 全てのタスクが終了した後，表現手法に関するアンケートに回答する．アンケートの内容は以下の通り．
- 各色付けの利用頻度及び貢献度を 5 段階のリッカート尺度を用いて評価する，
- 各色付けに対する評価の理由及び意見を記述する．
- その他，実験に対する意見を自由記述する．

5.2 実験タスク

5.2.1 実験に用いるデータ

実験に用いるデータとして，30,000 レコードの 16 次元データを作成した．作成したデータは量的次元のみで構成されており，一様分布または偏りのあるデータ分布とした．またレコードの構成については，相関係数の値の分布が一様分布となるように構成した．

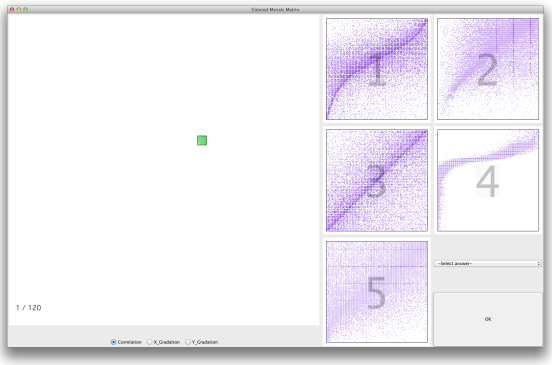

図 8 実験ツールのスクリーンショット
Fig. 8 Screenshot of the experiment tool

5.2.2 実験ツール

被験者には実験ツールを用いてタスクを実行してもらう．図 8 は，本番用タスクにおける実験ツールのスクリーンショットである．実験ツールの画面左部には色付き Mosaic Plot または Scatterplot が問として表示され，画面右部には 5 つの Scatterplot が選択肢として表示される．問の表現手法が示しているデータ分布について，同一のデータ分布を表現している選択肢を推測して選んでもらう．

色付き Mosaic Plot が用いる色付け手法については，被験者が任意に切り替え可能とする．これは実際の分析操作においても，色付け手法を切り替えることで分析を進めるためである．本実験では，次元間の相関に着目した色付け手法（*ColorC*），X 軸次元のカテゴリ分布に着目した手法（*ColorX*），Y 軸次元のカテゴリ分布に着目した手法（*ColorY*）を用いてタスクを行ってもらう．

5.2.3 タスクにおける条件パラメータ

問の表現手法について，タスク毎に以下の条件パラメータを設定する．

- 描画領域及び Area Graph の有無に対する条件の組み合わせは以下の 5 通りであり，これを 1 セットとおく．

$$\begin{cases} 700\text{ pixel} & \begin{cases} \text{with Area Graph} \\ \text{without Area Graph} \end{cases} \\ 24\text{ pixel} & \text{without Area Graph} \\ 12\text{ pixel} & \text{without Area Graph} \\ 6\text{ pixel} & \text{without Area Graph} \end{cases} \tag{1}$$

- 表現手法及びカテゴリ分割数 $P(n)$ に対する条件の組み合わせは以下の 4 通りである．

$$\begin{cases} \text{色付き Mosaic Plot} & \begin{cases} P(16) \\ P(8) \\ P(4) \end{cases} \\ \text{Scatterplot} \end{cases} \tag{2}$$

なお，カテゴリ分割手法については，全タスクにおいて最大値と最小値を基準にした分割手法を用いる．

以上を組み合わせることで，条件パラメータの組み合わせは全 20 通りとなる．本実験では各被験者に，全ての条件パラメータの組み合わせに対して 6 回ずつ，計 120 回のタスクを行ってもらう．タスクの順序による影響を減らすため，条件パラメータの出現順序は (2) 式を降順に行うパターンと昇順に行うパターンの 2 通りを採用する．いずれの場合も，各セット内での条件パラメータの順序はランダムとする．またそれぞれのタスクにおいて，問と選択肢に用いられるデータの次元対はランダムに選ばれる．

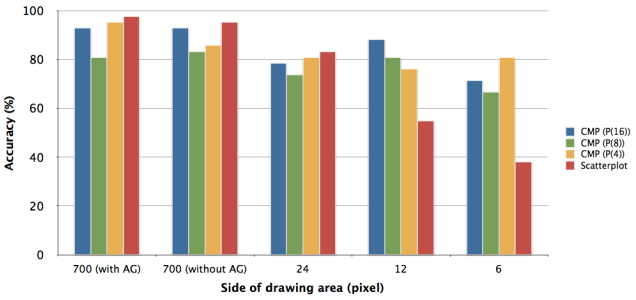

図 9 各条件パラメータにおけるタスクの平均正答率
Fig. 9 The average percentage of correct answers in task parameters for each condition

5.3 結果

各条件パラメータのタスクにおける平均正答率を集計した結果を図 9 に示す．横軸は描画領域及び Area Graph（AG）の有無，縦軸は正答率である．棒グラフの色は，表現手法及びカテゴリ分割数 $P(n)$ を表現している．

5.4 考察

読解の正確性，読解の所要時間，各表現の理解しやすさの観点から評価及び考察を行う．各タスクにおける平均正答率の比較には，1 対の標本に対する有意水準 5%の両側 t 検定を用いる．t 検定の自由度は $\nu = 6$ である．

5.4.1 描画領域の大きさとカテゴリ分割数の関係性

色付き Mosaic Plot の各描画領域におけるタスクの正答率について，分割数の違いによる有意差の有無を検定した．カテゴリ分割数が $P(n)$ のタスクにおける平均正答率を $\mu(n)$ とおく．

検定の結果，24 ピクセル四方以上の描画領域において，カテゴリの分割数による有意差は確認されなかった．12 ピクセル四方および 6 ピクセル四方の描画領域では有意差が確認され，前者では $\mu(16) \neq \mu(4)(p = 0.2894)$，後者では $\mu(8) \neq \mu(4)(p = 0.0167)$ となった．いずれの描画領域においても，他の組み合わせにおける有意差は確認されなかった．

5.4.2 Area Graph の影響

Area Graph の有無による正答率とアンケートで得られた意見を元に，Area Graph がデータ分布の読み取りにどの程度効果的であるかを考察する．まず t 検定を行ったところ，全ての描画領域において Area Graph の有無による

表 1　各表現手法の平均正答率及び t 検定の結果
Table 1　The average percentage of correct answers in each visualization method and the results of t-test

n	700	24	12	6
$\mu_C(n)$	92.86%	80.95%	88.10%	80.95%
$\mu_S(n)$	95.24%	83.33%	54.76%	38.10%
p 値	0.6036	0.8588	0.0177	0.0057

平均正答率の有意差は確認されなかった．一方で，アンケートからは『Area Graph が表現として最も理解しやすい』という意見や，『Area Graph 有りの場合，回答に対する確信をより高く持つことができた』という意見を得られた．これらの関する意見から，Area Graph は理解しやすい表現手法であり，またデータの読解を容易にする効果があると推測できる．

5.4.3　色付け手法の比較

本実験では，被験者毎かつタスク毎に各色付け手法を利用していた時間を計測して，色付け手法の利用率を計算した．計測結果より，全ての被験者において色付けの利用率が *ColorC*<*ColorX*<*ColorY* であったことがわかった．これは *ColorC* は他の色付け手法と比べて色の種類が少なく，読み取りに時間がかからなかったためと推測できる．

次にアンケートにおける色付けの利用頻度を集計したところ，平均値は *ColorC* は他の色付け手法より低い値となった．また *ColorX* 及び *ColorY* の平均値については殆ど差がなかった．一方で被験者毎に比較すると，5 名の被験者が *ColorC* に対して他の色付け手法と同等以上の評価を付けており，実際に計測した利用率とは異なる結果となった．また色付けの貢献度についても，利用頻度とおおよそ同様の結果となった．アンケートにおける自由記述からは，*ColorC* は全体像の把握に関して有用であるという意見が得られたが，一方で読み取りの難しさが欠点としてあげられた．また *ColorX* 及び *ColorY* はデータを正確に読み取ることができ，理解しやすいという意見を得られた．これらのアンケートにおける *ColorC* の利用頻度及び貢献度における評価の高さから，*ColorC* は他の色付け手法と同様に，判断材料として有用な手法であると判断できる．

5.4.4　Scatterplot との比較

Area Graph 非表示の色付き Mosaic Plot 及び Scatterplot について，各描画領域における平均正答率に対して検定を行った．カテゴリ分割数は平均正答率が最大のものを採用した．描画領域の 1 辺のピクセル数が n のタスクにおける平均正答率について，色付き Mosaic Plot の平均正答率を $\mu_C(n)$，色付き Mosaic Plot の平均正答率を $\mu_S(n)$ とおく．各描画領域における表現手法毎の平均正答率及び，両側 t 検定の p 値を表 1 に示す．検定の結果，$n = 700$ 及び $n = 24$ においては有意水準 5%で有意差が確認できなかった．一方で，$n = 12$ 及び $n = 6$ においては有意差を確認することができ，$\mu_C(n) \neq \mu_S(n)$ となった．各表現手法の平均正答率から，$\mu_C(n) > \mu_S(n)$ であると判断できる．

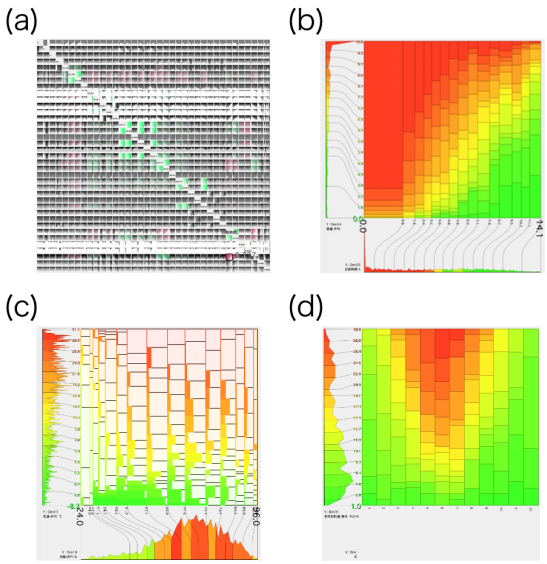

図 10　本ツールによる気象データの可視化例
Fig. 10　Cases visualization of the weather data with this tool

色付き Mosaic Plot 及び Scatterplot を用いたデータ分布の読解に関して，ある程度大きな描画領域においては有意差はなかった．一方で，12 ピクセル四方以下の非常に狭い描画領域においては，色付き Mosaic Plot の方が高い正答率となり，読解における正確性の高さが確認できた．また色付き Mosaic Plot は，全ての描画領域において高い可読性を維持することが可能であった．これは，色付き Mosaic Matrix が高次元データの概観を得るための手法として有用であることを示している．

6.　ユースケース

6.1　対象データ

気象庁ホームページ内[*1]から，札幌，東京，長野，大阪，那覇の観測地における気象データを取得し，本ツールを用いて可視化した．データの次元数は 37 次元である．データの取得期間は [2011/9/1-2012/8/31] の 366 日間である．1 つのレコードが 1 つの都市における 1 日の気象に対応しているため，データのレコード数は 1830 レコードである．

6.2　ツールを用いたデータ分析

図 5 は気象データを用いた可視化結果である．図 5 の Matrix View を見ると，色相が横縞の模様になっていることが確認できた．これは Y 軸次元のカテゴリ分布に着目した色付けを使用しているためである．例えば殆ど緑で表示されている行の次元は，日毎の降水量の次元や積雪量の次元であった．これは年間を通して雨または雪が降る日は少ないという結果を反映している．また雲量平均の次元にお

[*1] http://www.data.jma.go.jp/obd/stats/etrn/index.php

いては，赤から黄にかけての色相が多く見られた．これより雲量が多い日の割合が高いことがわかる．

図 10(a) は次元間の相関で色相を決定する色付け手法に変更した可視化結果である．ここでは各次元における外れ値の影響を減らすため，カテゴリ分割手法としてデータ量依存の分割手法を用いている．図 10(a) に着目すると，右下部に鮮やかな赤の色相で表現されている色付き Mosaic Plot があった．これは強い負の相関関係であることを示しており，各軸の次元名を確認したところ，X 軸次元は日照時間，Y 軸次元は雲量平均であった．また図 10(b) は，Y 軸次元のカテゴリに着目した色付け手法を用いた Detail View であり，X 軸次元には日照時間，Y 軸次元には雲量平均が割り当てられている．色付き Mosaic Plot の色相分布より，雲量平均が大きな値の日ほど日照時間が少なくなっていることがわかる．

最も北に位置する札幌における気象データのみを Detail View から選択する．図 10(c) はデータ選択状態における Detail View の一例であり，X 軸次元は平均湿度，Y 軸次元は平均気温である．図 10(c) より，札幌は他の観測地におけるデータと比べて，気温及び湿度が比較的低いことがわかる．ここで，観測地が札幌であるレコードのみを対象として再描画する．図 10(d) は札幌における観測月と日射量の関係を表した Detail View である．これより，11 月から 2 月にかけての冬季期間では，日射量も減少していることがわかる．

7. まとめ

本研究では，限られた画面領域にて高次元データの概観を得ることを目的とした表現手法である色付き Mosaic Matrix を開発した．色付き Mosaic Matrix はデータの分布を色を用いて表現することにより，限られた描画領域内でもデータの特徴を把握できる表現手法である．また量的データをカテゴリデータとして扱うために，複数のカテゴリ分割手法を開発した．本表現手法はカテゴリ単位でデータを表現するため，レコード数の多い高次元データでも可視化できる．

色付き Mosaic Matrix を用いて高次元データの分析を行うためのツールを開発した．色付き Mosaic Matrix により高次元データの概観を取得し，そこから得られた知見を元に，詳細な分析を行なっていくことが可能である．

評価実験では色付き Mosaic Plot を用いてデータ分布の読み取りタスクを設定することで，色付き Mosaic Plot の可読性を調査した．タスクの正答率より，色付き Mosaic Plot は描画領域の大きさに依らず高い可読性を維持できることを確認した．これにより，高次元データ分析における色付き Mosaic Matrix の有用性を示した．

本研究により，高次元データを一度に俯瞰し，そこから得た知見を元により詳細な分析を行うことが可能になる．これは今後の高次元データ分析及びその分析手法の発展に対する手助けとなる．

謝辞

本成果の一部は，株式会社富士通研究所からの受託研究によるものである．

参考文献

[1] D. B. Carr, R. J. Littlefield, W. L. Nicholson and J. S. Littlefield. Scatterplot Matrix Techniques for Large N. In *JASA'87*, Vol. 82, No. 398, pp. 424–436, 1987.

[2] M. Sips, B. Neubert, J. P. Lewis and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. In *IEEE-VGTC Symposium on Visualization*, Vol. 28, No. 3, pp. 831–838, 2009.

[3] A. Lex, H.-J. Schulz, M. Streit, C. Partl and D. Schmalstieg. VisBricks: Multiform Visualization of Large, Inhomogeneous Data. In *TVCG'11*, Vol. 17, No. 12, pp. 2291–2300, 2011.

[4] L. Nováková and O. Štěpánková. Multidimensional clusters in RadViz. In *SMO'06*, pp. 470–475, 2006.

[5] J. Sharko, G. Grinstein and K. A. Marx. Vectorized Radviz and Its Application to Multiple Cluster Datasets. In *TVCG'08*, Vol. 14, No. 6, pp. 1444–1451, 2008.

[6] N. Elmqvist, P. Dragicevic and J.-D. Fekete. Rolling the Dice: Multidimensional Visual Exploration using Scatterplot Matrix Navigation. In *TVCG'08*, Vol. 14, No. 6, pp. 1141–1148, 2008.

[7] A. Inselberg and B. Dimsdale. The plane with parallel coordinates. *The Visual Computer*, Vol. 1, No. 4, pp. 69–91, 1985.

[8] F. Bendix, R. Kosara and H. Hauser. Parallel Sets: Visual Analysis of Categorical Data. In *InfoVis'05*, pp. 133–140, 2005.

[9] Z. Geng, Z. Peng, R. S. Laramee, R. Walker, and J. C. Roberts. Angular Histograms: Frequency- Based Visualizations for Large, High Dimensional Data. In *TVCG'11*, Vol. 17, No. 12, pp. 2572–2580, 2011.

[10] Y.-H. Fua, M. Ward and E. Rundensteiner. Hierarchical Parallel Coordinates for Exploration of Large Datasets. In *VIS'99*, pp. 43–50, 1999.

[11] D. Feng, L.Kwock, Y. Lee and R. M. Taylor. Matching Visual Saliency to Confidence in Plots of Uncertain Data. In *TVCG'10*, Vol. 16, No. 6, pp. 980–989, 2010.

[12] T. Saito, H. N. Miyamura, M. Yamamoto, H. Saito, Y. Hoshiya and T. Kaseda. Two-tone pseudo coloring: compact visualization for one-dimensional data. In *InfoVis'05*, pp. 173–180, 2005.

[13] B. Shneiderman. The eyes have it: A task by data-type taxonomy for information visualizations. In *Proceedings of the Symposium on Visual Languages*, pp. 336–343, 1996.

[14] M. Friendly. Mosaic Displays for Multi-Way Contingency Tables. In *JASA'94*, Vol. 89, No. 425, pp. 190–200, 1994.

[15] M. Friendly. Extending Mosaic Displays: Marginal, Conditional, and Partial Views of Categorical Data. In *JCGS'99*, Vol. 8, No. 3, pp. 373–395, 1999.