平成 24 年度

筑波大学情報学群情報科学類

卒業研究論文

# 題目
# 色による特徴表現を用いた<br>高次元データ可視化手法の開発

主専攻　　知能情報メディア主専攻

著者　　小林 弘明

指導教員　三末和男，志築文太郎，高橋伸，田中二郎

# 要　旨

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，フル HD ディスプレイ程度の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 はデータの特徴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概観を読み取ることが可能な表現手法である．また量的データに用いるカテゴリ分割手法を開発した．カテゴリ単位でデータを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．本表現手法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，ユースケースとして実際の高次元データにおける傾向の分析を行った．

# 目 次

# 図 目 次

# 表 目 次

# 第1章　序論

## 1.1　データにおける次元

世の中の様々な分野に現れているデータは，複数の属性を含んでいる場合が多い．データ中の 1 つの属性は 1 つの次元 (Dimension) として考える事ができ，多くの属性を持つデータは多次元データ (Multidimensional data) または多変量データ (Multivariate data) と呼ばれる．また一般に，次元数が 10 以上のデータは高次元データ (High-dimensional data) と呼ばれている．例えば，日々の気象情報を記録した気象データの場合，日毎の気温や湿度，降水量などが，それぞれ 1 つの次元として保存されている．さらに風速や風向，日照量などを細かく記録していくと，10 以上の次元を持つ高次元データになることは珍しくない．

## 1.2　次元を構成するデータの種類

一般に，次元を構成するデータの変数はその特徴によって，名義変数 (Nominal variables)，順序変数 (Ordinal variables)，量的変数 (Quantitative variables) の 3 つに分類することができる．名義変数とは名詞的な値をとり，値が同一か否かを評価することだけに意味を持つ変数である．順序変数とは順序を付けて比較できる変数である．量的変数とは値の大小を比較すること，また値の平均を算出することが可能な変数である．

カテゴリデータ (Categorical data) とは，質的データ (Qualitative data) とも呼ばれ，名義変数または順序変数によって構成されるデータである．一方で，量的変数によって構成されるデータは，カテゴリデータの対として量的データ (Quantitative data) や数値データ (Numerical data) と呼ばれる．本稿では以後，カテゴリデータによって構成されている次元をカテゴリ次元，量的データによって構成されている次元を量的次元と呼ぶ．

カテゴリデータ及び量的データと各変数の構成関係は以下の通りである．

- カテゴリデータ
  - 名義変数　(e.g. 地名，血液型)
  - 順序変数　(e.g. 鉱物の硬度，レースの着順)
- 量的データ
  - 量的変数　(e.g. 距離，気温)

## 1.3 多次元データ分析の有用性

多次元データは気象や放射能などを観測して記録したデータや，日常生活の行動を記録したライフログ，飲食店の購買履歴や顧客のデータ，またアンケートデータなど，実に様々な領域で保存されている．

これらの多次元データを分析して得られた知見は，様々な分野に活かすことが可能である．例えば気象データを分析した場合，各観測地の気温や降水量の推移を分析することで，地球温暖化を始めとした異常気象の傾向を知ることができる．また各地の放射能濃度を記録したデータと気象データを併用して分析することで，高濃度に放射能汚染されている“ホットスポット”をあぶり出すだけではなく，ホットスポットにおける共通した気象の特徴が発見できるかもしれない．他にも商品の購買履歴データや SNS ユーザの発言を記録したデータを分析すると，ターゲット層の嗜好を踏まえた効果的なマーケティングなども可能になる．

## 1.4 可視化によるデータ分析

データ分析の際には，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．

例えば最も基本的かつ効果的な可視化の 1 つとして，散布図 (Scatterplot) が挙げられる．一般的な Scatterplot では，直交座標系を構成する各座標軸にデータの任意の 2 次元を割り当てて，データの各要素を直交座標系にプロットする．プロットされた点の位置や密集度合いにより，データの分布を読み取ることができる．例えば図 1.1 のように，気象データにおける日毎の最低気温と最高気温という 2 次元を Scatterplot で可視化する．この Scatterplot の横軸は東京における最高気温，縦軸は東京における最低気温，また 1 つの点は 1 日分のデータに対応している．この図から，2 つの次元の関係性について分析することが可能である．

## 1.5 高次元データの可視化における問題点

高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．主な手法は以下の 3 つに分類される．

- **次元の一部を可視化する手法.**

  この手法では，分析においてユーザが注目する次元のみを可視化する．Sips らの手法 [1] 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している．しかし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．高次元データにおいて，分析に必要な部分をどのように判定するか，またどの程度まで表示するかが難しい問題になる．

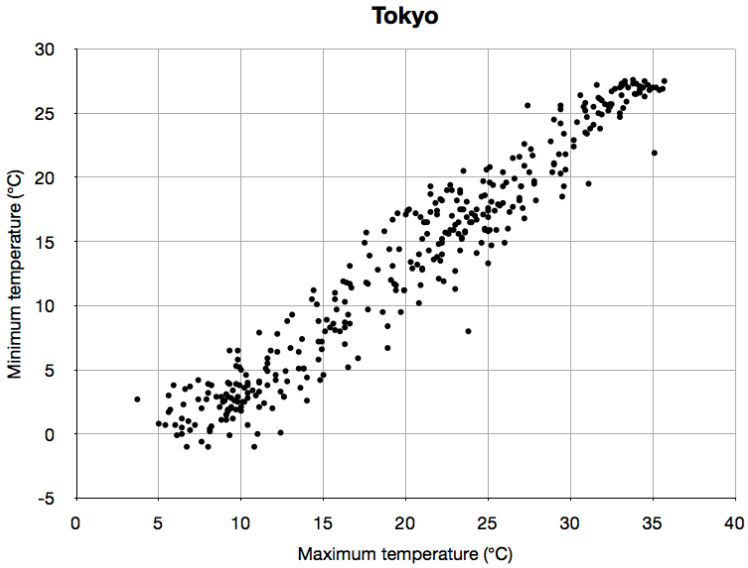



図 1.1: 散布図の一例

- **全次元を規則的に並べて可視化する手法.**

  データ全体を可視化して概観を得る手法である．例えば散布図行列 (Scatterplot Matrix)[2] 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 次元あたりの描画領域が限られてしまい，高次元データの全次元を可視化して概観を得るのは困難になる．

- **次元削減手法を適用し，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持するように低次元空間で可視化する手法.**

  次元削減手法の代表例としては，主成分分析による手法や多次元尺度構成法が一般に知られている．また RadViz[3, 4] は次元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．これらの手法の問題点としては，各次元の数値を直接読み取ることが難しくなる点が挙げられる．さらに次元を減らすことで情報量も減ってしまい，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，フル HD ディスプレイ (1920 × 1080) 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．

## 1.6 本研究の目的

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も閲覧可能にすることを目的とする．本研究では特に，30 次元以上の高次元データを対象とする．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．また本手法を適用することで，高次元データの概観を分析できるツールを開発する．

## 1.7 本研究の貢献

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なものである．
- 描画領域の大きさを変化させた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．複数種類の描画領域に対して，可視化結果からどの程度正確に情報を読み取れるかを，正答率を元に定量的に評価している．

# 第2章　関連研究

高次元データを可視化する手法としては，データの一部の次元または全ての次元を描画領域上に表示する手法が代表的である．これらは正確性の高い表現ができるという利点がある．一方で表示する次元数を増やす場合，1 次元に割り当てられる描画領域が狭くなるため，単位領域あたりが表現可能なデータ量を増やす，または操作インタフェースを工夫する必要がある．そしてこのアプローチは，データ量が非常に多い大規模データの可視化の際に生じる問題に対するアプローチと類似していると考えられる．

そこで本章では，大規模データを対象にした可視化手法の分類法 [5] を元に，本研究の関連研究を 3 つに分類する．本章で取り上げる主な可視化手法は，データの一部または全ての次元を一度に表示する手法である．主な対象データは多次元データ，多変量データ，高次元データ，大規模データである．加えて，高次元データに対して有効たり得る可視化手法についても言及する．

## 2.1　Atomic Visualizations

Atomic Visualizations は，1 つのマーカーが 1 つのレコードに対応した，最も基本的な表現手法の分類である．マーカーは手法における表現の最小単位のことを指している．またレコードは，一件ごとのまとまったデータの並びのことを指す．例えば日毎の気象データにおける「1 日」や，アンケートデータにおける「1 人」などが 1 つのレコードに対応する．

Atomic Visualizations の最も代表的な例としては，1.4 節で紹介した Scatterplot が挙げられる．Scatterplot においては 1 つの点が 1 つのマーカーであり，1 つのレコードに対応している．

### 2.1.1　Scatterplot Matrix

1.5 節でも挙げた Scatterplot Matrix[2] は，最も代表的な多次元データ可視化手法である．Scatterplot Matrix 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．例えば $N$ 次元データの場合，$N(N-1)$ 通りの Scatterplot を表示することで，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．通常は 2 次元の散布図を表示し，次元を切り替える際には，奥行きにあるもう 1 次元

との Scatterplot が 3 次元のアニメーションによって変遷する．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．

Scatterplot Matrix を用いた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てしまい，読解や分析が困難になる．

### 2.1.2 Parallel Coordinates Plot

Parallel Coordinates Plot(PCP)[7, 8, 9] も Scatterplot Matrix と同様，複数の次元に対して一度に概観を得ることが可能な可視化手法である．各次元に対して座標軸を用意し，座標軸を並列に並べる．そして，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．

FlowVizMenu[10] は複雑な多変量ネットワークデータを対象にした可視化手法である．FlowVizMenu では，Scatterplot Matrix と PCP を組み合わせてデータを表現する．独自のポップアップウィジェットである FlowVizMenu を用いて，インタラクティブな視覚的表現を実現している．

一般的な PCP においては，隣り合った次元同士でしか直接的に関係を見ることができない．一方で Heinrich らの手法 [11] では，PCP を縦に複数本並べることで，全次元ペアの隣り合わせを網羅している．これにより，次元軸を並び替えることなくデータの概観を得ることが可能である．しかし高次元データを扱おうとした場合，任意の次元対を探索するのが困難になる．

PCP は線の密集度合いや線の傾きにより，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という問題がある．

### 2.1.3 データ抽象化を行う可視化手法

データ抽象化 (Data Abstraction)[12, 13] とは，データ量を削減してデータの抽象度を高めながら，削減前と同等かそれに近い視覚的特徴を表現する手法である．扱うデータ量自体が減るため，高速なデータ処理や描画が可能である．データ抽象化の問題は，データの正確性が低下してしまう点である．

## 2.2 Aggregate Visualizations

Aggregate Visualizations は，1 つのマーカーが複数のレコードに対応する可視化手法の分類である．その性質上，Atomic Visualizations の手法と比べて空間効率が良いため，より高次元データ向きの手法であると考えられる．

本研究では，カテゴリデータを対象にした可視化手法は Aggregate Visualizations に分類する．カテゴリデータの変数に含まれる値は，その項目ごとにカテゴリとして区別され，1 つのカテ

ゴリには複数のレコードが含まれることが多い．カテゴリデータの可視化手法の多くは，1 つのマーカーが 1 つのカテゴリに対応しているため，これは本質的には Aggregate Visualizations であると言える．

### 2.2.1 Mosaic Plot を用いた多次元データ可視化手法

Mosaic Plot(Mosaic Display)[14, 15] は空間充填型のカテゴリデータ可視化手法である．描画領域を矩形に分割することにより，データ分布を各矩形の面積で表現する．分割する際の各矩形の幅 (または高さ) は，1 つの次元におけるカテゴリの比率によって決定する．この分割を縦横に再帰的に行うことにより，多次元カテゴリデータを表現することができる．例えば 2 次元データを可視化する場合，2 次元間におけるカテゴリの各組み合わせが Mosaic Plot 内の 1 つの矩形に対応付けされる．Mosaic Plot は，大きさの目立つ矩形などによりデータの大まかな特徴を把握することが可能であるが，3 つ以上の次元を表現すると，矩形の数が多くなってしまい，読解が難しくなる．

Mosaic Matrix[16, 17, 18] は，多次元データを表現するための Mosaic Plot の拡張である．これは Scatterplot Matrix と同様，全ての組み合わせ可能な次元対に対する Mosaic Plot を行列状に並べて表示する．Mosaic Matrix は Aggregate Visualizations であるため，データ量が増えても可読性を保つことができる．一方で並べる次元数が増えると，1 つの Mosaic Plot に割り当てられる描画領域が狭くなるため，Mosaic Plot 内の各矩形を区別するのが困難になる．

### 2.2.2 Parallel Coordinates Plot を拡張した手法

Parallel Sets[19] は，PCP と Mosaic Plot を組み合わせた多次元カテゴリデータの可視化手法である．PCP における座標軸上の各点を大きさを持った矩形として表現し，さらに矩形の幅を持った線として座標同士をつなぐ．このとき，1 個の矩形や 1 本の線は複数のレコードを表現している．またカテゴリに応じて矩形の色を設定することで，複数次元間の関係性を表現できる．

Angular Histogram[20] は，多次元データの各次元についてヒストグラムを作成し，PCP と同様に平行に並べて表示する手法である．さらに元となる PCP の線の傾きを元に，ヒストグラムの各棒の傾きを設定することで，隣り合う次元同士の関係性を表現する．

このように PCP を Aggregate Visualizations に拡張することで，大規模データを効果的に表現できる．一方で，高次元データを一度に可視化すると平行軸の幅が狭くなるため，可読性が低下してしまう．

VisBricks[21] は，等質的な次元を 1 つの次元群とし，また次元群内で等質的なレコードをクラスタとして 1 つのレンガで表現する手法である．アーチ状に次元群を並べた上で，中央にある次元群の任意のレンガをより詳細に可視化する．また，隣接した次元群間のレンガを Parallel Sets と同様に矩形でつなぐことで，次元群間の関係性を表現している．しかし詳細表示する次元やレコードはデータ全体の一部のみであるため，表示されていないレンガの概観を得ることができない．よってデータ全体の特徴を把握することは難しい．

### 2.2.3 テーブルベースの可視化手法

Table Lens[22] は，大規模データを表形式で可視化する手法である．大規模表データの一部をフォーカスとズームによって分析することができる．

ManyNets[23] は，大規模な多変量ネットワーク構造のデータを任意のサブネットワークに分割し，個々のネットワークを 1 行として行列状に表示する手法である．各列はユーザが定義した属性を表示し，列ごとにヒストグラムを表示することで，データ分布の表現やインタラクティブな検索機能を実現している．

これらの手法はテーブルベースという表現の性質上，同一画面上に表示できるネットワーク数は数十程度が限界であるという問題がある．

### 2.2.4 空間充填型の可視化手法

空間充填型の可視化手法は，任意のデータ量を一定の描画領域内でも表現できる手法である．2.2.1 節で紹介した Mosaic Plot は空間充填型の可視化手法である．またその他の代表的な例としては Treemap[24, 25] が挙げられる．Treemap は階層構造のデータを対象にした，空間充填型の可視化手法である．ノードを矩形として，親子関係は入れ子状に配置することで表現する．1 個のノードが複数のリーフを含んでいるという点から，Aggregate Visualizations の一種と言える．

Baudel らの手法 [26] も，四角形の描画領域内に全データを充填する可視化手法である．レコードの格納順序や充填基準について複数の手法を提案し，それらを組み合わせた空間充填アルゴリズムを開発している．

これらの手法は高次元データを対象にしたものではなく，そのままでは高次元データを可視化することはできない．一方で，一定の領域内で任意のデータ量を表現できるため，高次元データの可視化に対しても有用な表現である．

## 2.3 Density Plot Visualizations

Density Plot Visualizations は，色でレコードの密集度合いや分布を表現する可視化手法である．Fua らの手法 [27] 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 [28] 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値で分布を表すため，狭い描画領域で用いるには不向きな手法である．

# 第3章　高次元データ分析のための要求事項

本研究では，高次元データを分析するための要求事項を以下の 2 つに大別する．

1. 高次元データにおける概観の表現
2. 複数種類のデータ変数を持つデータへの対応

## 3.1　高次元データにおける概観の表現

Shneriderman によって提唱された Visual Information Seeking Mantra[29] によると，可視化における分析は，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に分析するものである，とされている．

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．データの次元数が増えるほど，分析に必要な次元の判断はより困難になる．もし概観から分析を行うことができれば，そこから得られた知見を元に，より詳細な分析を行っていくことが可能になる．

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としては，データが持つ全次元を一度に可視化する方法が挙げられる．例えば Scatterplot Matrix は，全ての組み合わせ可能な次元対の Scatterplot を一度に表示することにより，多次元データの概観を提示することが可能である．

しかし次元数の多い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という問題がある．そのため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なければならない．

## 3.2　複数種類のデータ変数を持つデータへの対応

高次元データを対象にする可視化手法は，任意の変数に対しても利用可能である必要がある．1.2 節で述べたとおり，データの各次元は，名義変数，順序変数，量的変数のいずれかの変数で構成されている．各次元は独立であるため，次元数が多い高次元データの場合，複数種類の変数で構成されていることが多い．よって，一部のデータ変数が扱えない手法では，多種多様な高次元データを分析するのは困難になる．

さらに，目的に応じた様々な分析を可能にするために，異なる種類の変数を比較できる表現を用いるべきである．高次元データを分析する際は，異種の変数同士を比べることもある．例えば気象データを分析する場合，名義変数の次元である観測地名と，量的変数の次元である気温の関係性を調べるためには，名義変数と量的変数の両方に対して利用可能な可視化手法が必要になる．

# 第4章　視覚的表現

## 4.1　表現の設計方針

本研究では，Aggregate Visualizations の手法と色によりデータの特徴を表現する手法に着目する．その上で，高次元データを概観から分析可能にするために，高い可読性を保ちながら高次元データの全次元を一度に可視化できる手法を設計する．

Aggregate Visualizations は狭い描画領域を有効活用したデータの可視化が可能である．またレコード数の多いデータに対しても高い可読性を保つことができる．Aggregate Visualizations の多くは，名義変数や順序変数から構成されるカテゴリデータを対象にした可視化手法である．また量的データはカテゴリデータに変換することが可能である．例えばアンケートデータであれば，人間の年齢という量的データを，10 代，20 代などの年代別というカテゴリデータに変換できる．このようにデータ処理を行うことで，Aggregate Visualizations を用いて全種類の変数のデータを可視化することができる．

色相の分布や割合でデータの特徴を表現する手法を用いることで，狭い描画領域でも高い可読性を保つことができる．データを色相の分布で表現する手法の例としては，Two-Tone Pseudo Coloring[30] が挙げられる．この手法は色相の分布のみで 1 次元のデータを表現しており，小さく細い矩形からでもデータの特徴を読み取り可能にしている．

## 4.2　色付き Mosaic Matrix

本研究では，高次元データの概観を得る手法として色付き Mosaic Matrix(CMM) を開発する．色付き Mosaic Matrix は Mosaic Matrix(2.2.1 節を参照) を拡張した手法である．色付き Mosaic Matrix では，独自の色付けによってデータの特徴を表現する．

色付き Mosaic Matrix に用いる色付き Mosaic Plot(CMP) には，任意の 2 次元を直交座標系の X 軸，Y 軸にそれぞれ割り当てる．以後は X 軸に割り当てられた次元を X 軸次元，Y 軸に割り当てられた次元を Y 軸次元と呼ぶ．また矩形の分割手順は図 4.1 のように，まず X 軸次元のカテゴリ比率で矩形を分割し，その後各矩形について，Y 軸次元のカテゴリ比率でさらに矩形に分割する．この分割手順は全ての色付き Mosaic Plot において共通とする．

色付き Mosaic Plot は Mosiac Plot と同様，各矩形の面積比によってデータの割合を表現している．色付き Mosaic Plot ではさらに，矩形毎に色を設定する．これにより色の割合や分布からデータの特徴を把握できる．

色付き Mosaic Matrix は，全ての組み合わせ可能な次元対に対する色付き Mosaic Plot を行

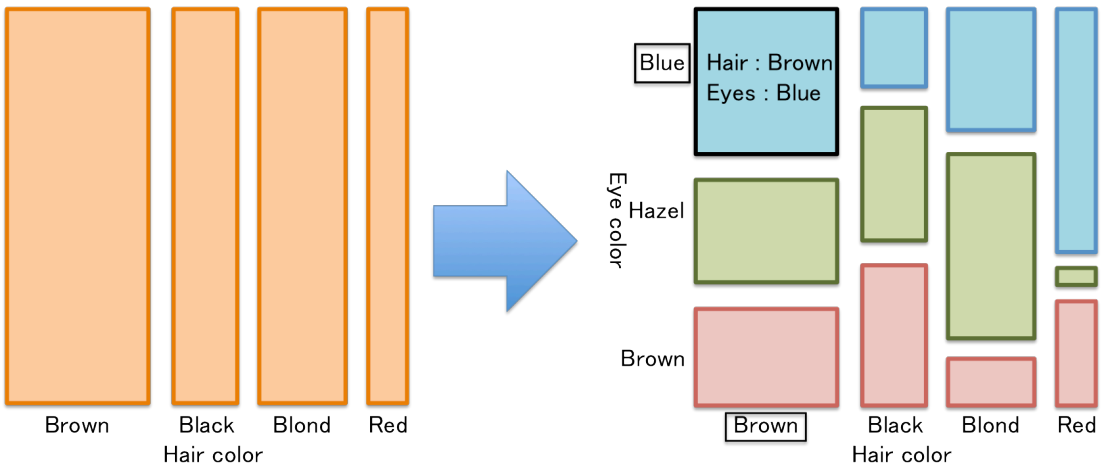



図 4.1: 色付き Mosaic Plot における矩形の分割順序

列状に配置した，空間充填型の可視化手法である．全次元対の色付き Mosaic Plot を俯瞰することにより，高次元データ全体としての特徴を把握する．

Mosaic Matrix 自体は多次元カテゴリデータの可視化手法であるため，そのままでは量的次元を扱うことが困難である．色付き Mosaic Plot は任意の高次元データを可視化するために，量的データをカテゴリデータへ変換する．さらに元々が量的次元同士の色付き Mosaic Plot については，相関係数を計算して表現に利用する．

## 4.3 色付けによる特徴の把握

高次元データを Mosaic Matrix を用いて可視化する場合，個々の Mosaic Plot の描画領域が非常に狭くなることが問題になる．そこで本手法では Mosaic Plot のカテゴリ毎に色付けの規則を設定し，それに従って矩形の色を決定することで，狭い描画領域における可読性を向上させる．

色付け規則の設定方法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．色の 3 要素である色相 (Hue)，彩度 (Saturation)，明度 (Brightness) をカテゴリ毎に変更することにより，色の規則性を実現する．

各矩形の色を計算するため，各次元内のカテゴリは昇順にして番号で区別する．各矩形のカテゴリについて，X 軸次元のカテゴリ総数を $n_X$，Y 軸次元のカテゴリ総数を $n_Y$ とおく．また，X 軸次元のカテゴリが $i$ 番目 $(1 \leq i \leq n_X)$，Y 軸次元のカテゴリが $j$ 番目 $(1 \leq j \leq n_Y)$ である矩形を $R(i, j)$ と表記する．

### 4.3.1　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．

X 軸次元のカテゴリに着目した色づけの場合，矩形 $R(i,j)$ について，

$$H = \frac{i}{n_X} \tag{4.1}$$

を計算し，$H$ に応じて色相を設定する．色相は緑から赤にかけてのグラデーションを採用し，例えば $H = 0$ であれば緑，$H = 0.5$ であれば黄，$H = 1$ であれば赤の色相を設定する．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 4.2 と図 4.3 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 4.2 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．一方で図 4.3 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．

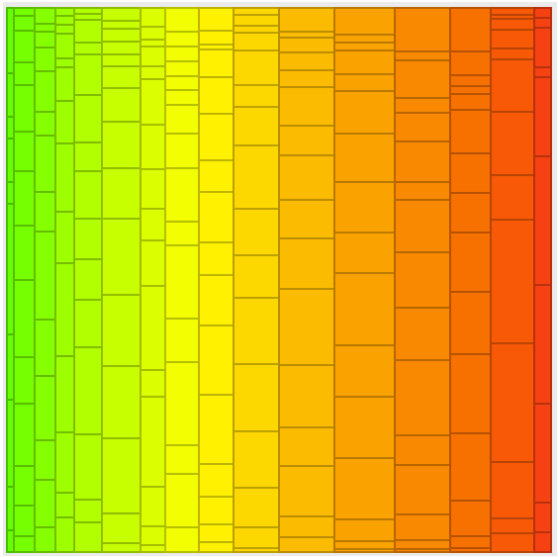


図 4.2: X 軸次元のカテゴリ分布に着目した色付け手法の一例

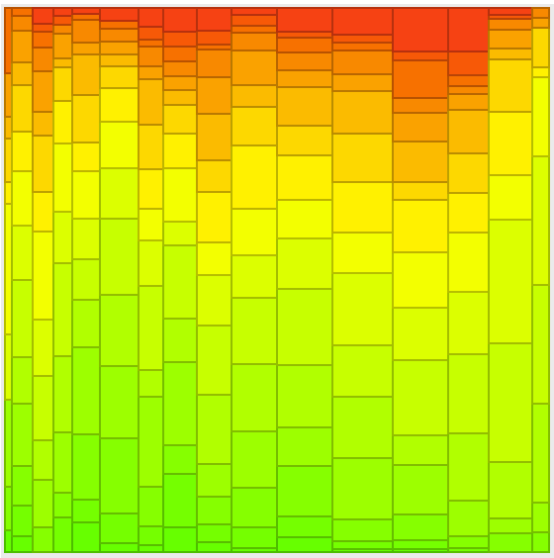


図 4.3: Y 軸次元のカテゴリ分布に着目した色付け手法の一例

### 4.3.2　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，元々が量的次元の対である Mosaic Plot 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，色相は Mosaic Plot 毎に共通のものとなる．

相関係数の計算は量的データをカテゴリデータへ変換する前に行う．高次元データのレコード数を $M$ とおく．対象となる任意の色付き Mosarc Plot について，$r$ 番目のレコードにおける X 軸次元の要素を $x_r$，Y 軸次元の要素を $y_r$ とおく．このとき $r$ は $1 \leq r \leq M$ を満たす自然数である．全レコードを元にデータ列 $\{(x_r, y_r)\}$ を生成し，このデータ列に対する相関係数 $C$ を以下の通りに求め，これを各 Mosaic Plot の相関係数とする．

$$C = \frac{\sum_{k=1}^{M}(x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{M}(x_k - \bar{x})^2}\sqrt{\sum_{k=1}^{M}(y_k - \bar{y})^2}} \tag{4.2}$$

なお，

$$\bar{x} = \frac{1}{M}\sum_{k=1}^{M} x_k$$
$$\bar{y} = \frac{1}{M}\sum_{k=1}^{M} y_k$$

である．

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．パターン 1 は図 4.4 及び図 4.5 に用いている色付け手法で，色相で相関の正負及び強弱を表現し，明度と彩度で Y 軸次元のカテゴリ区別を行っている．パターン 2 は図 4.6 及び図 4.7 に用いている色付け手法で，色相で相関の正負のみを表現し，明度と彩度で相関の強弱とカテゴリ区別を行っている．以下に，それぞれのパターンにおける色付け規則を述べる．

1. 色相は緑から赤のグラデーションを採用し，例えば．$C = 1$ であれば緑，$C = 0$ であれば黄，$C = -1$ であれば赤の色相を設定する．なお，相関係数が定義できない次元対の Mosaic Plot に関しては，青の色相を設定する．彩度と明度については，矩形 $R(i, j)$ について，

$$S = \min(\frac{2j}{n_X}, 1) \tag{4.3}$$
$$B = \max(\frac{2j}{n_X} - 1, 0) \tag{4.4}$$

   を計算し，$S$ に応じて彩度，$B$ に応じて明度を設定する．

   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. 色相は $C \geq 0$ であれば緑，$C < 0$ であれば赤の色相を設定する．彩度と明度については，矩形 $R(i, j)$ について，

$$S = \frac{|C|j}{n_X} \tag{4.5}$$
$$B = 1 - \frac{(1 - |C|)j}{n_X} \tag{4.6}$$

   を計算し，$S$ に応じて彩度，$B$ に応じて明度を設定する．なお相関係数を定義できない場合は，$C = 0$ として同様に計算し，色を設定する．

   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．一方でカテゴリの識別は困難な場合が多いが，これについては他の色付け手法を併用して補完することが可能である．

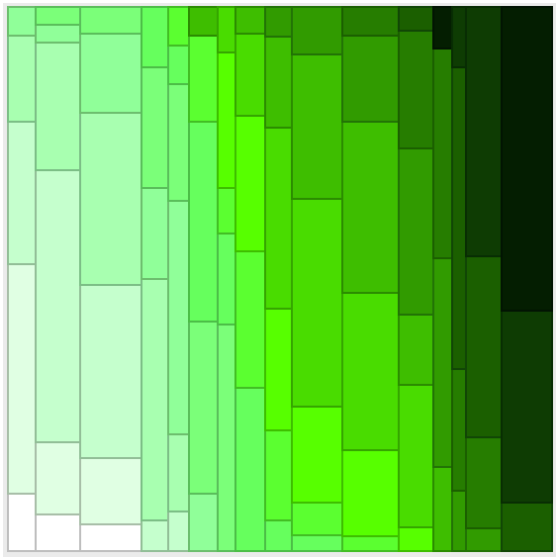


図 4.4: 色相で相関の正負及び強弱を表現する色付け手法の一例 (相関係数 $C = 0.969$)

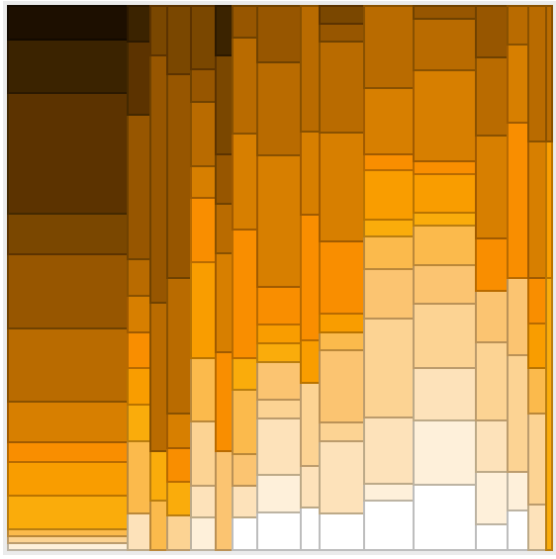


図 4.5: 色相で相関の正負及び強弱を表現する色付け手法の一例 (相関係数 $C = -0.514$)

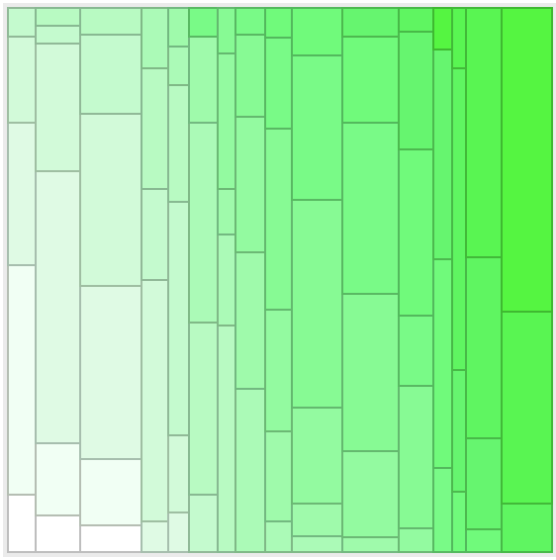


図 4.6: 色相で相関の正負のみを表現する色付け手法の一例 (相関係数 $C = 0.969$)

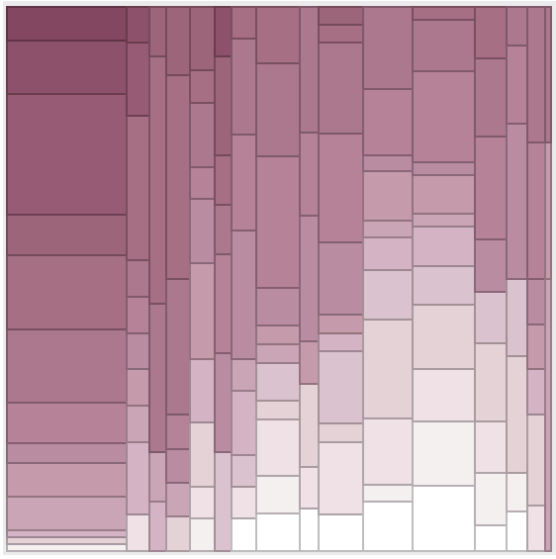


図 4.7: 色相で相関の正負のみを表現する色付け手法の一例 (相関係数 $C = -0.514$)

## 4.4 量的データに用いるカテゴリ分割手法

量的データをカテゴリデータへと変換するため，量的次元については値域全体を複数の値域に細分化することで，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．そこで本ツールでは，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．

以下，データの全レコードを $M$，着目する量的次元を $D$，$r$ 番目のレコードにおける次元 $D$ の値を $v_r$，分割数を $p(\geq 2)$，次元 $D$ における $i$ 番目のカテゴリを $c_i$ とおく．また，次元 $D$ における全レコード中の最大値を $v_{max}$，最小値を $v_{min}$，平均値を $\bar{v}$ とし，それぞれ以下の通りに定義する．

$$\begin{aligned} v_{max} &= \max_r v_r \\ v_{min} &= \min_r v_r \\ \bar{v} &= \frac{1}{M}\sum_{r=1}^{M} v_r \end{aligned}$$

### 4.4.1 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．次元 $D$ における $i$ 番目のカテゴリである $c_i$ の値域は，$w_p = \dfrac{v_{max} - v_{min}}{p}$ を用いて，

$$[v_{min} + (i-1)w_p, v_{min} + iw_p) \tag{4.7}$$

と表現できる．なお，$v_{max}$ は $c_p$ に含める．

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，他の値から大きく外れている外れ値 (Outlier) が含まれている場合，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という問題が生じる．

### 4.4.2 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．標準偏差は $\sigma = \sqrt{\dfrac{1}{M}\sum_{r=1}^{M}(v_r - \bar{v})^2}$ である．$c_i$ の値域は標準偏差を用

いて，

$$\begin{cases} \left(-\infty, \bar{v} + \alpha(2 - \frac{p}{2})\sigma\right) & (i = 1) \\ \left[\bar{v} + \alpha(i - \frac{p}{2})\sigma, \bar{v} + \alpha(i + 1 - \frac{p}{2})\sigma\right) & (2 \leq i \leq p - 1) \\ \left[\bar{v} + \frac{1}{2}\alpha p\sigma, \infty\right) & (i = p) \end{cases} \tag{4.8}$$

と表現できる．なお，$\alpha$ は定数であり，この値によって分割幅が変化する．

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．定数を適切に設定することで，データが集中している部分に焦点を当てた分割が可能である．一方で適切な定数が設定されていない場合，カテゴリの両端にデータが含まれない可能性が生じる．これは値が一部の値域に偏っており，そのため平均値と最大値 (または最小値) との間に差が殆どないようなデータで発生する問題である．

### 4.4.3 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．まず分割領域に該当するデータ量を $S$ とおく．本分割手法では以下の 2 通りのデータ量を採用する．

- $S = 0.95M$．データ量の 95%を分割領域とする．
- $S = \dfrac{p-2}{d}M$．分割数に応じて分割領域のデータ量を決定する．

求める分割領域を $[s_{min}, s_{max}]$ とおく．但し，$s_{max}$ の上限は $v_{max}$ とし，また $s_{min}$ の下限は $v_{min}$ とする．次に，$[\bar{v}, \bar{v}]$ を初期値とした値域を設定し，この値域内のデータ量を $S'$ とおく．$S \leq S'$ を満たすまで値域の両端または一端を広げていき，$S \leq S'$ を最初に満たした値域を分割領域とする．

分割領域の分割数を $p'$ とおき，もし $v_{min} = s_{min}$ または $s_{max} = v_{max}$ のいずれかを満たしている場合は $p' = p - 1$ とし，それ以外の場合は $p' = p - 2$ とする．分割領域を $p'$ 分割した場合，分割幅は $w_s = \dfrac{s_{max} - s_{min}}{p'}$ となる．一方で，最大値と最小値を基準にした分割手法 (4.4.1 節を参照) における分割幅は $w_p = \dfrac{v_{max} - v_{min}}{p}$ である．この 2 つの分割幅を計算した上で，最小の分割幅となる分割手法を採用する．

- $w_s \geq w_p \Rightarrow$ 最大値と最小値を基準にした分割手法を適用する．
- $w_s < w_p \Rightarrow i$ 番目のカテゴリ $c_i$ について，値域を以下の通りに設定する．
  - $v_{min} = s_{min}$ の場合．

$$\begin{cases} (s_{max}, v_{max}] & (i = p) \\ [s_{min} + (i-1)w_s, s_{min} + iw_s) & (otherwise) \end{cases} \tag{4.9}$$

– $s_{max} = v_{max}$ の場合.

$$\begin{cases} [v_{min}, s_{min}) & (i = 1) \\ [s_{min} + (i-1)w_s, s_{min} + iw_s) & (otherwise) \end{cases} \tag{4.10}$$

– その他.

$$\begin{cases} [v_{min}, s_{min}) & (i = 1) \\ (s_{max}, v_{max}] & (i = p) \\ [s_{min} + (i-2)w_s, s_{min} + (i-1)w_s) & (otherwise) \end{cases} \tag{4.11}$$

この分割手法では，データ分布に応じた必要最小限の分割幅を採用する．例えば図 4.8 のように，一様分布に近い次元に対しては最大値と最小値を用いてカテゴリを分割する．また図 4.9 のように，外れ値の影響などによりデータ分布が偏っている次元に対しては，データが集中している部分に焦点を当てたカテゴリ分割を行っている．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ることも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，各次元においてどの分割幅が採用されているか判断できないため，混乱および誤解を招く可能性がある．

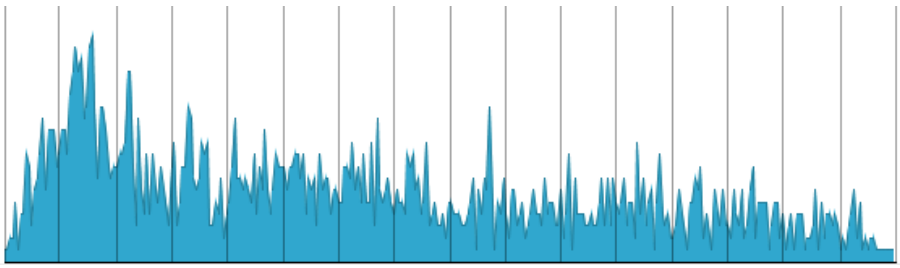


図 4.8: データが一様分布に近い次元に対するデータ量依存のカテゴリ分割例．横軸は値域，縦軸はデータ量.

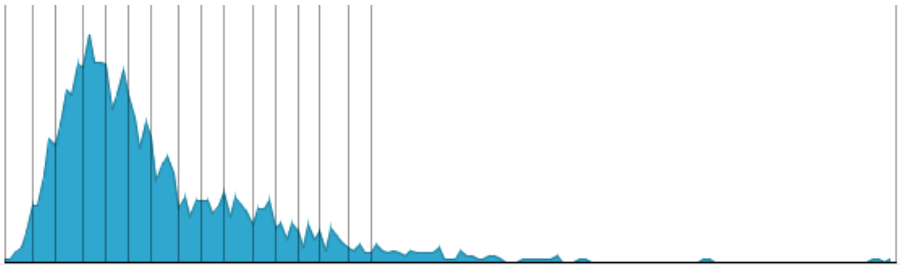


図 4.9: データ分布が偏っている次元に対するデータ量依存のカテゴリ分割例．横軸は値域，縦軸はデータ量.

# 第5章　ツールの開発

## 5.1　実装言語及びデータ形式

本ツールの実装言語には Java(Java™ Platform Standard Edition 6.0) 及び Processing[1]を使用する．processing.core.PApplet を継承することで Java に Processing を埋め込むことが可能であるため，本ツールの描画部分の実装に Processing を使用している．

データの読み込みは tsv 形式で行う．

## 5.2　ツールの設計

開発した表現手法を用いて，高次元データの全体を俯瞰し，そこで得られた知見を元に，更に詳細な分析を行うことが可能なツールを開発する．分析ツールの主な対象ユーザとしては，情報可視化またはデータ分析に関する専門的な知識を持つ人を想定する．

本ツールは，Visual Information Seeking Mantra[29] に沿った高次元データ分析を可能にするための表現及び機能を備える．概観から詳細へと掘り下げるドリルダウン式のデータ分析を可能にするため，本ツールでは高次元データの概観を得ることが可能な画面と，データの詳細な分析が可能な画面を設計する．またデータのフィルタリング機能を備えることで，より詳細なデータ分析を可能にする．

## 5.3　ツールに用いる表現手法

開発する分析ツールでは，データ全体を表示する Matrix View と，詳細を表示する Detail View を設ける．

### 5.3.1　Matrix View

Matrix View(図 5.1) 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 では描画領域を節約するため，また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．

[1] http://processing.org/

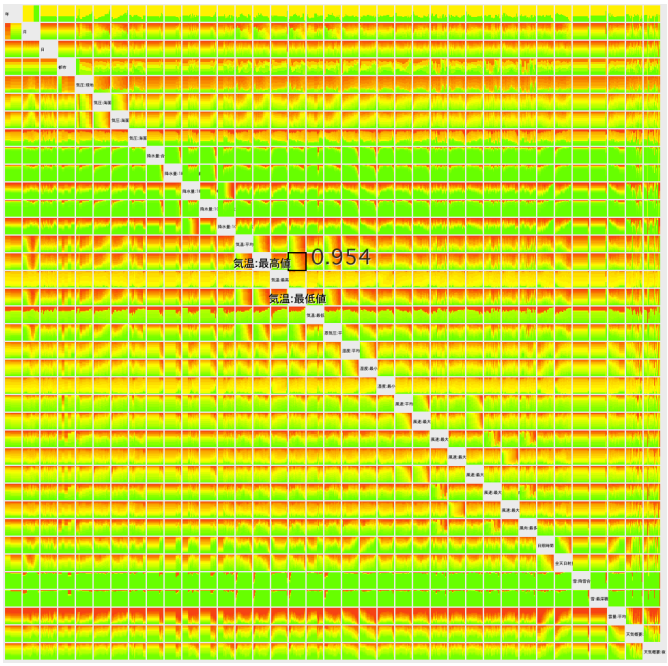



図 5.1: Matrix View の一例

### 5.3.2 Detail View

Detail View(図 5.2) 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．カテゴリ次元における各軸次元のカテゴリ名は，図 5.3 のように，X 軸次元は各カテゴリが対応する矩形の下に，Y 軸次元は等間隔に表示する．また量的次元においては，図 5.4 のように，カテゴリの切れ目となる値を表示する．

#### Area Graph

データ分布の読解を容易にするため，また各カテゴリの値域を視覚的に表現するために，Detail View では Area Graph(面グラフ) を用いて各軸の次元の分布を表示する．なお，Area Graph の表示は量的次元のみである．X 軸 Area Graph の場合，横軸が X 軸次元の値 (値軸) であり，高さが各値の頻度を表現している．Y 軸 Area Graph の場合，縦軸が Y 軸次元の値であり，横幅が各値の頻度を表現している．

Area Graph では図 5.2 のように，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．これにより，データ分布と各カテゴリの値域を関連付けて見ることができる．また Area Graph の各カテゴリ区間における色は．そのカテゴリが対応する最も面積が広い矩形

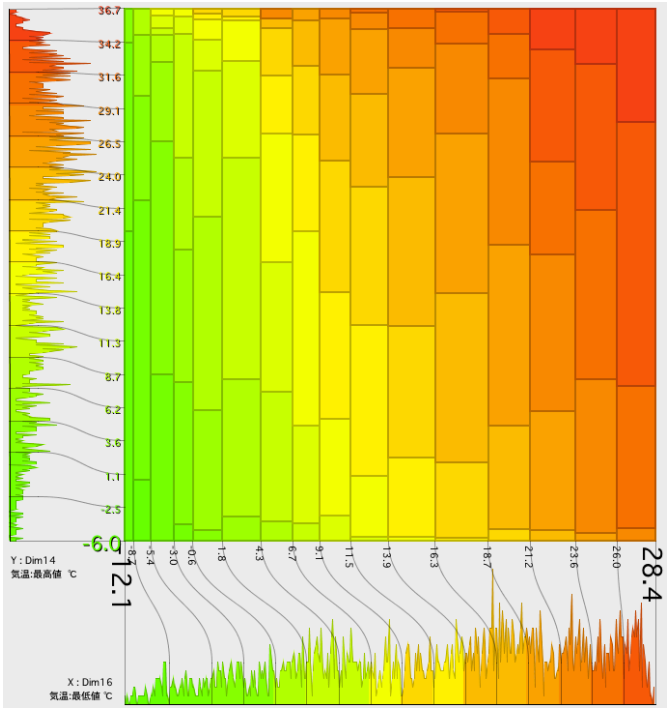



図 5.2: Detail View の一例


図 5.3: カテゴリ次元におけるカテゴリ名の表示

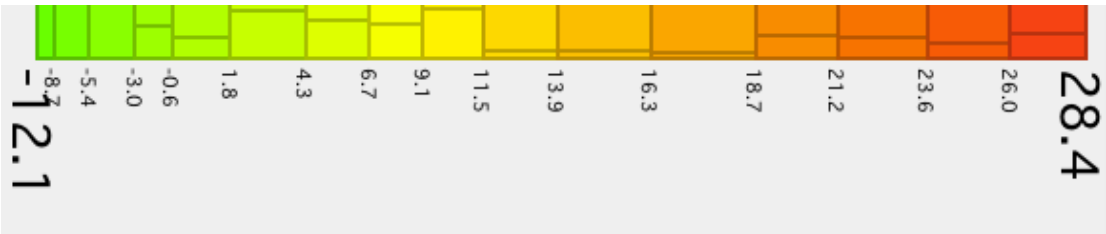



図 5.4: 量的次元におけるカテゴリ名の表示

と同じ色を採用している．

## 5.4　ツールのインタフェース

図 5.5 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．

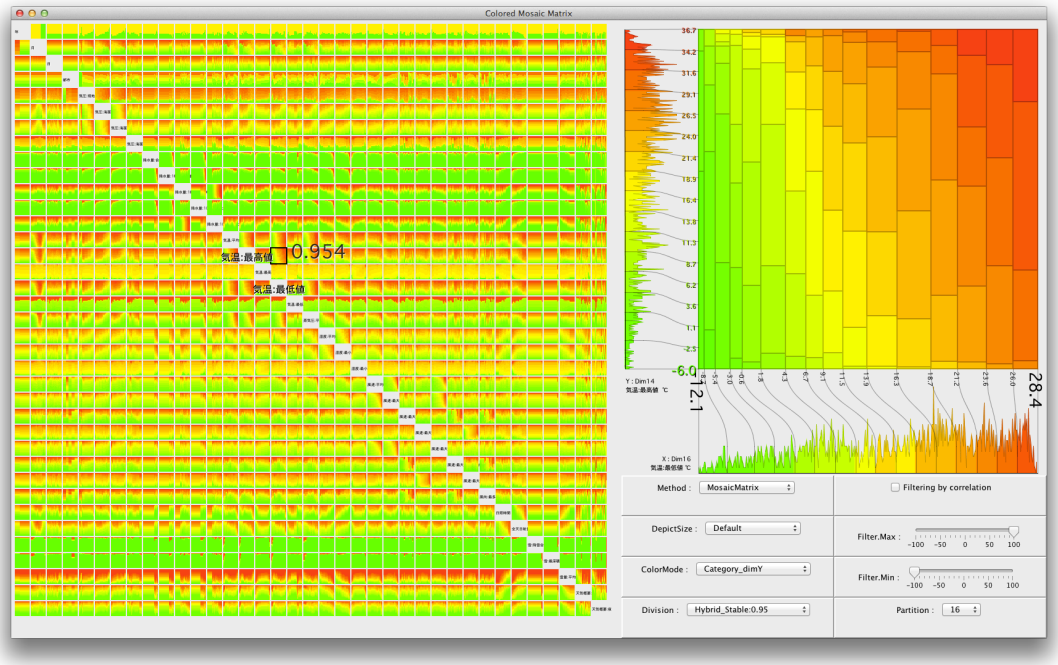



図 5.5: 開発した分析ツールのスクリーンショット

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．また対象とする色付き Mosaic Plot の両次元が量的データである場合，図 5.6 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図 5.7 のように表示する．またその矩形に対応する各軸のカテゴリについて，色付き Mosaic Plot の枠外に表示されているカテゴリ名を拡大する．このとき Area Graph についても，対応カテゴリの値域を彩度の変化によって強調する．

操作パネルでは，表現手法に用いる色付け手法及びカテゴリ分割手法を切り替えることが

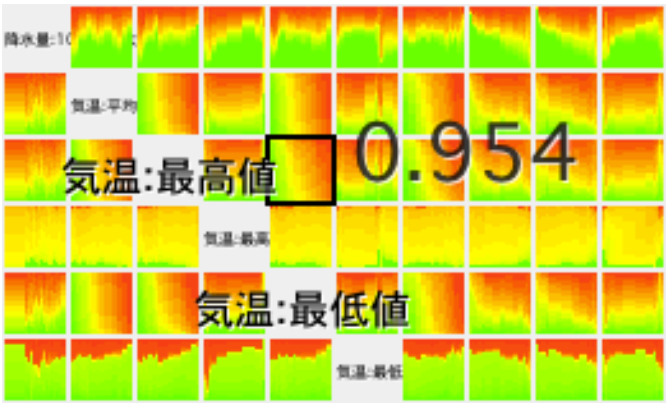



図 5.6: Matrix View におけるマウスオーバによる相関係数の表示例

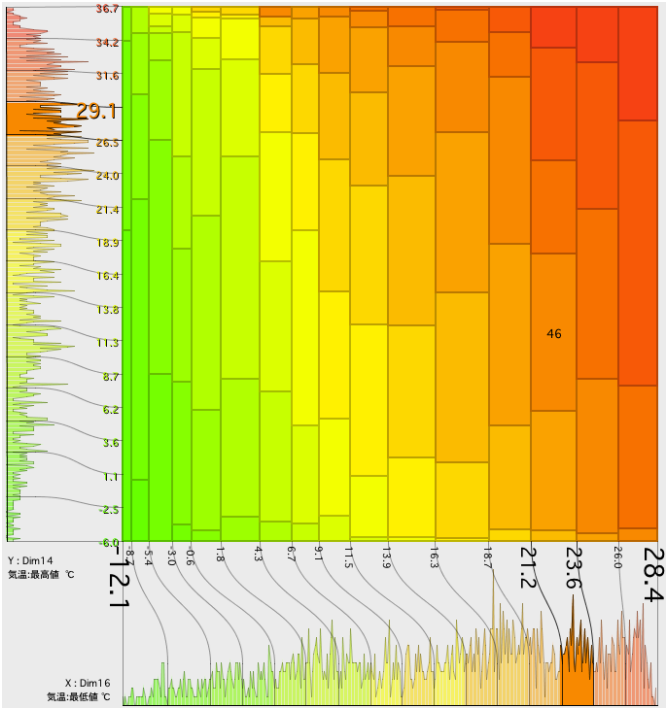



図 5.7: Detail View におけるマウスオーバによるレコード数の表示例

できる．また Matrix View に対して，相関係数によるフィルタリングを設定することができる．フィルタリングにおける相関係数の最大値と最小値は，スライダを用いて設定する．また可視化手法を切り替えることで，Scatterplot Matrix による可視化結果を表示することも可能である．

### 5.4.1 レコードの選択及びフィルタリング

Detail View では，矩形をマウスで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．この時，複数の矩形に対して AND 選択が可能である．

レコード群を選択している時は，図 5.8 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．このとき分割された 2 つの矩形については，色の彩度を変更することで区別を行なっている．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．

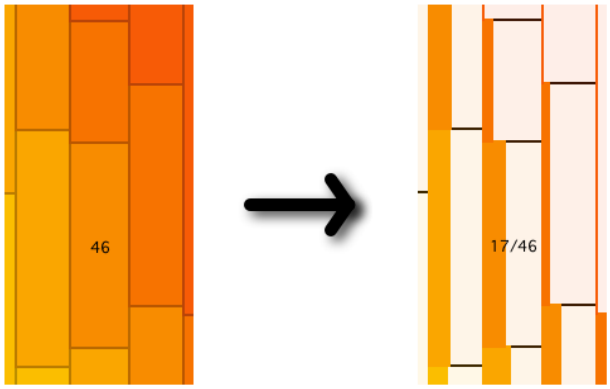



図 5.8: レコード選択前後における Detail View の変化 (左例が選択前，右例が選択後)

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．

# 第6章　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．

## 6.1　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．

### 6.1.1　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，コンピュータサイエンスを専攻する大学生，大学院生及び教員のうち，特に情報可視化を研究分野としている 7 名を選定した．7 名の被験者は全員日本人であり，うち男性が 4 名，女性が 3 名である．なお本章では被験者毎に ID を設定し，区別に用いることとする．

本実験は色を用いてデータの読解を行うため，被験者の色覚は正常である必要がある．そこで，実験を行う前に石原色覚検査表 [31] を用いて色盲検査を行い，全被験者の色覚が正常であることを確認した．

### 6.1.2　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である．

   - 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自由記述する．
   - その他，実験に対する意見を自由記述する．

### 6.1.3 実験に用いる計算機環境

全被験者に対して同一の計算機を用いて実験を行う．画面サイズは 15.4 インチ (対角) であり，ピクセル数は $1680 \times 1050$ である．

## 6.2 実験タスク

### 6.2.1 実験に用いるデータ

実験に用いるデータとして，16 次元のデータを作成した．図 6.1 は色付き Mosaic Matrix による実験用データの可視化結果であり，図 6.2 は Scatterplot Matrix による可視化結果である．なお，レコード数は $30,000$ レコードである．

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．

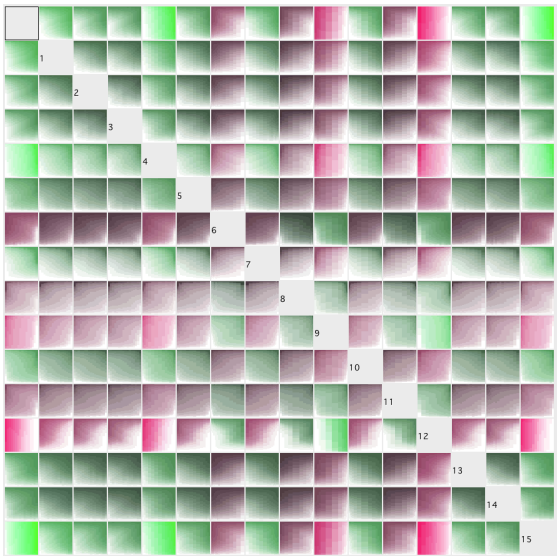



図 6.1: 色付き Mosaic Matrix による実験用データの可視化結果

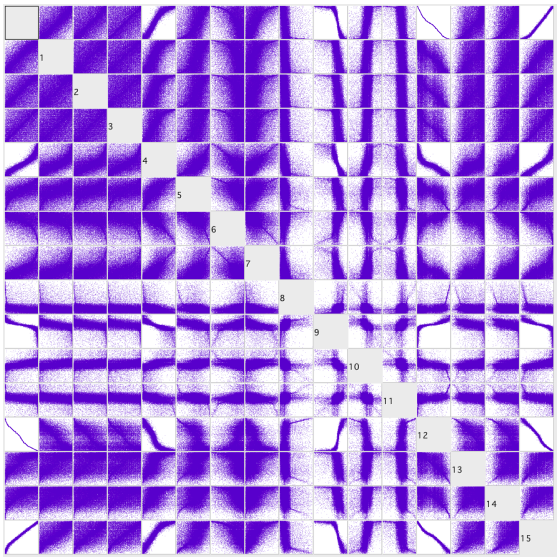


図 6.2: Scatterplot Matrix による実験用データの可視化結果

### 6.2.2 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 6.3 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．

各タスクの問および選択肢は，実験用データにおける 240 通りの次元対からランダムに選出する．ただし，問の次元対は選択肢にも含まれ，かつ 5 つの選択肢は全て異なる次元対となるように補正する．なお練習用タスクにおいては，問の表現手法及び選択肢が表現している次元対が表示される．これにより被験者は，正解がどの選択肢なのかを知ることができる．

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，カテゴリの分布に着目した色付け手法 (4.3.1 節を参照) と，次元間の相関に着目した色付け手法 (4.3.2 節を参照) を用いてタスクを行ってもらう．前者の色付け手法については，X 軸次元に着目した手法と Y 軸次元に着目した手法の 2 種類を採用する．後者の色付け手法については，色相で相関の正負を，明度と彩度で相関の強弱とカテゴリ区別を行う手法を採用する．

### 6.2.3 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 4 種類の条件パラメータを設定する．

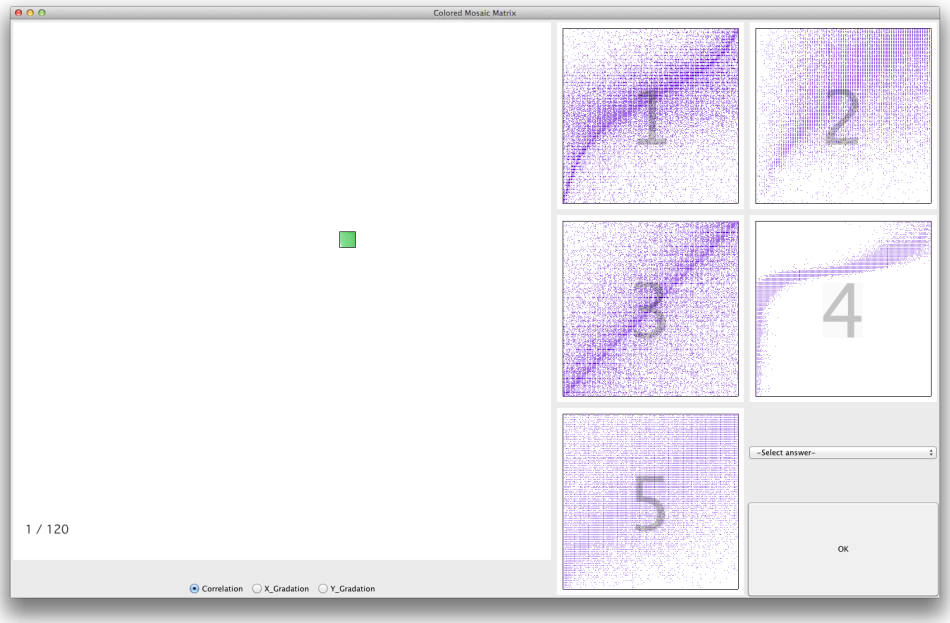



図 6.3: 本番用タスクにおける実験ツールのスクリーンショット

- 表現手法について．問の表現手法には，色付き Mosaic Plot または Scatterplot を用いる．
- 表現手法の描画領域の大きさについて．描画領域は正方形であり，辺のピクセル数は 700, 24，12，6 の 4 種類である．描画領域が 700 ピクセル四方のタスクは Detail View(5.3.2 節を参照) を，それ以外のタスクは Matrix View(5.3.1 節を参照) をそれぞれ想定したタスクとなる．そこで，700 ピクセル四方のタスクではカテゴリ名を表示する．
- Area Graph の有無について．描画領域が 700 ピクセル四方のタスクについては，Area Graph を表示する場合についてもタスクを行う．それ以外の描画領域については，Area Graph を表示しないタスクのみを行う．
- カテゴリ分割手法について．最大値と最小値を基準にした分割手法 (4.4.1 節を参照) は，最もデータ分布に関する誤解が少ない手法である．よって全タスクにおいてこの手法を用いる．
- カテゴリの分割数について．分割数は 16 分割，8 分割及び 4 分割の 3 種類を設定する．なおカテゴリ分割数は色付き Mosaic Plot のみのパラメータである．

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．

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

また，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．

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

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以下の 2 通りを採用する．

- Scatterplot で 1 セット → 色付き Mosaic Plot の 4 分割で 1 セット → 色付き Mosaic Plot の 8 分割で 1 セット → 色付き Mosaic Plot の 16 分割で 1 セット
- 色付き Mosaic Plot の 16 分割で 1 セット → 色付き Mosaic Plot の 8 分割で 1 セット → 色付き Mosaic Plot の 4 分割で 1 セット → Scatterplot で 1 セット

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．

## 6.3 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 6.4 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．

また図 6.5 の Scatterplot は，各被験者における全タスクの所要時間と正答率を図示したものである．図 6.5 より，所要時間と正答率の間には関係性がないことがわかった．

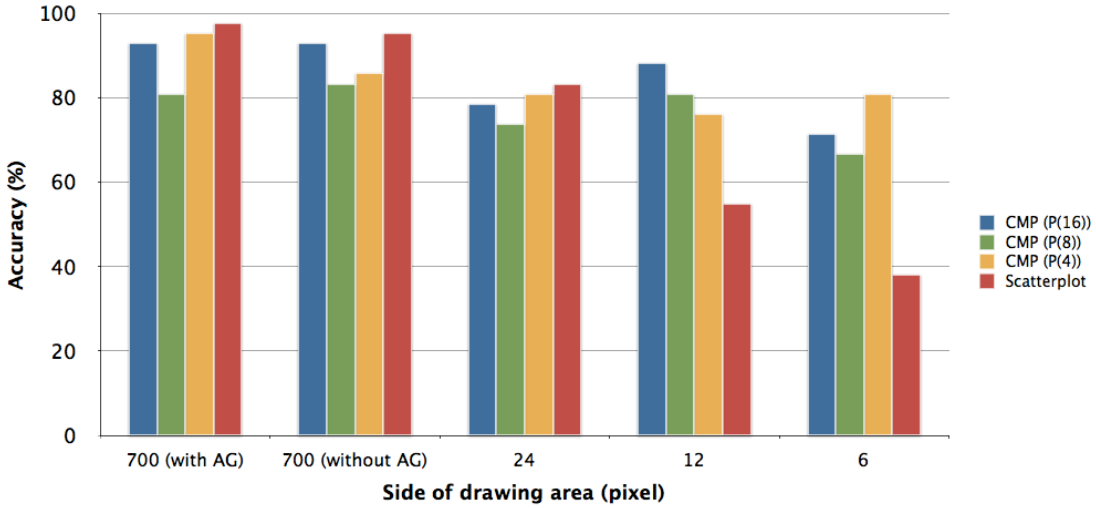



図 6.4: 各条件パラメータにおけるタスクの平均正答率

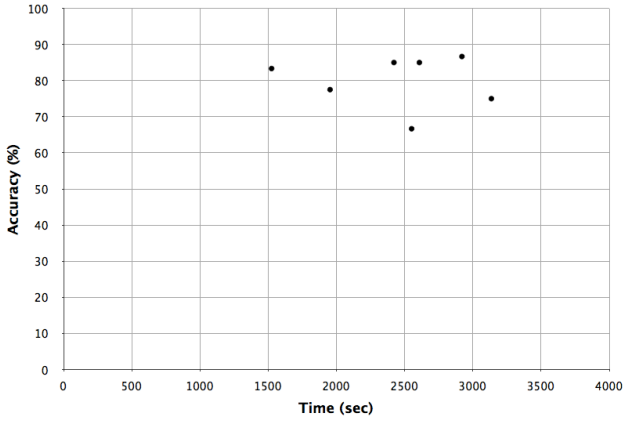



図 6.5: 各被験者における全タスクの所要時間と正答率

## 6.4 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 検定における計算には，Microsoft Excel 2008 for Mac に搭載されている TTEST 関数を用いる．なお，t 検定の自由度は $\nu = 6$ である．

### 6.4.1 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する．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．カテゴリ分割数の各組み合わせに対して，以下の仮説を立てて検定する．

$$\begin{cases} \text{帰無仮説 } H_0 : \mu(n_a) - \mu(n_b) = 0 \\ \text{対立仮説 } H_1 : \mu(n_a) - \mu(n_b) \neq 0 \end{cases} \tag{6.1}$$

有意水準 5%における t 検定の検定結果を表 6.1 に示す．表中の数値は両側 t 検定の p 値であり，$p < 0.05$ の場合は帰無仮説を棄却できる．

表 6.1: 各描画領域における CMP の平均正答率に関する t 検定の p 値

| | 描画領域の 1 辺 (pixel) 及び Area Graph の有無 | | | | |
|---|---|---|---|---|---|
| | 700 (with AG) | 700 (without AG) | 24 | 12 | 6 |
| $\mu(16)$ | 92.86% | 92.86% | 78.57% | 88.10% | 71.43% |
| $\mu(8)$ | 80.95% | 83.33% | 73.81% | 80.95% | 66.67% |
| $\mu(4)$ | 95.24% | 85.71% | 80.95% | 76.19% | 80.95% |
| $H_0 : \mu(16) - \mu(8) = 0$ | 0.1824 | 0.2308 | 0.4571 | 0.2894 | 0.7030 |
| $H_0 : \mu(16) - \mu(4) = 0$ | 0.6036 | 0.0781 | 0.7358 | 0.0465 | 0.4571 |
| $H_0 : \mu(8) - \mu(4) = 0$ | 0.1112 | 0.7358 | 0.1996 | 0.3559 | 0.0167 |

表 6.1 より，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 $\mu(16) \neq \mu(4)$，後者では $\mu(8) \neq \mu(4)$ となった．一方で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．

### 6.4.2 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．描画領域が 700 ピクセル四方の色付

き Mosaic Plot に関する，Area Graph の有無におけるカテゴリ分割数毎の平均正答率及び両側 t 検定の p 値を表 6.2 に示す．

表 6.2: Area Graph の有無におけるカテゴリ分割数毎の平均正答率及び t 検定の p 値

| | CMP (P(16)) | CMP (P(8)) | CMP (P(4)) |
|---|---|---|---|
| with AG | 92.86% | 80.95% | 95.24% |
| without AG | 92.86% | 83.33% | 85.71% |
| p 値 | 1.0000 | 0.8291 | 0.1030 |

表 6.2 より，カテゴリ分割数が 8 以上の色付き Mosaic Plot については，平均正答率に殆ど差がなかった．一方で分割数が 4 の場合については，Area Graph 有りのタスクに対する平均正答率が Area Graph 無しのタスクと比べて約 10 ポイント高かった．これは，4 分割のカテゴリでは細かな分布の把握が比較的困難であるため，Area Graph から得られる情報をより重要視した結果であると推測できる．

またカテゴリ分割数毎に，Area Graph の有無によって平均正答率に有意差が生じるか否かを検定した．その結果，全てのカテゴリ分割数において平均正答率の有意差は確認されなかった．

次に，アンケートから得られた意見の概要を以下に示す．

- Area Graph が表現として最も理解しやすく，学習コストも低かった．
- 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．

これらの Area Graph に関する意見から，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．

### 6.4.3 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．各被験者における色付けの利用率および全被験者における平均利用率を図 6.6 に示す．なお本節では，次元間の相関に着目した色付け手法を *Color1*，X 軸次元のカテゴリ分布に着目した色付け手法を *Color2*，Y 軸次元のカテゴリ分布に着目した色付け手法を *Color3* と略記する．図 6.6 についても同様の略記を用いる．

図 6.6 より，全ての被験者において色付けの利用率が *Color1*<*Color2*<*Color3* となった．また被験者 4 については，*Color1* は殆ど利用せず，*Color2* 及び *Color3* のみを用いてタスクに回答していた．

タスク終了後のアンケートにおいて，各色付け手法に対する利用頻度と貢献度を 5 段階のリッカート尺度を用いて回答してもらった．各色付け手法における利用頻度の集計結果を図 6.7，貢献度の集計結果を図 6.8 に示す．

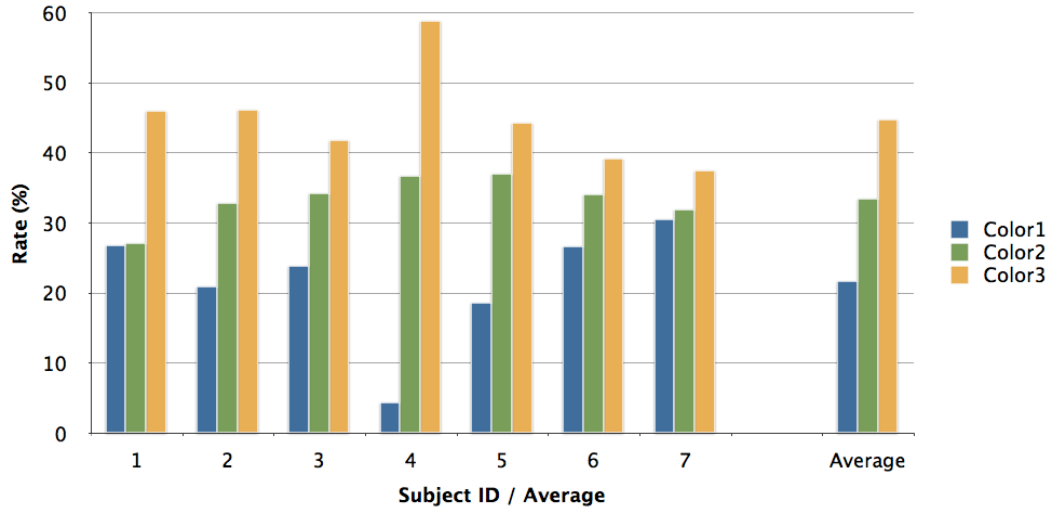



図 6.6: 各被験者における色付けの利用率および全被験者における平均利用率

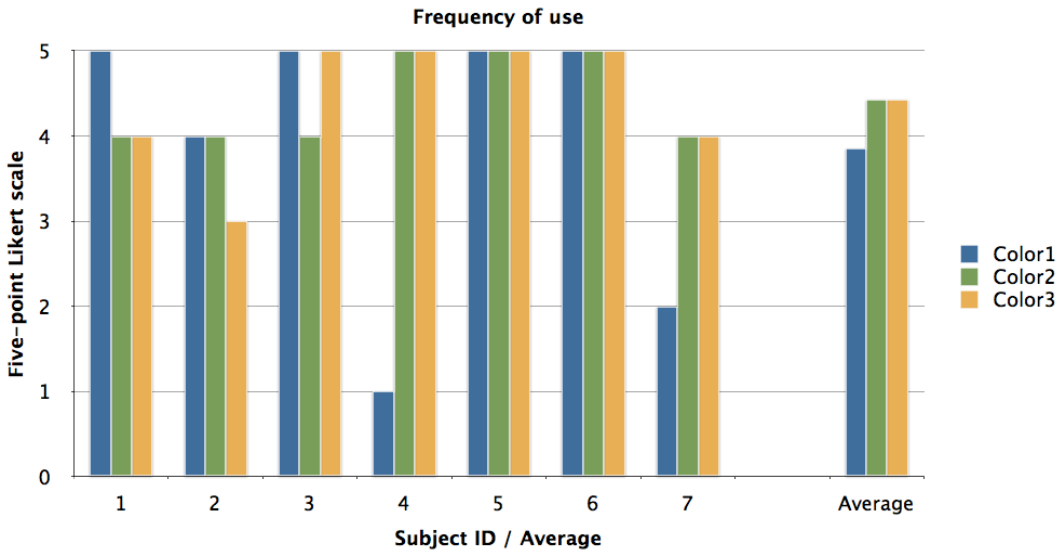



図 6.7: 5 段階リッカート尺度による各色付け手法の利用頻度

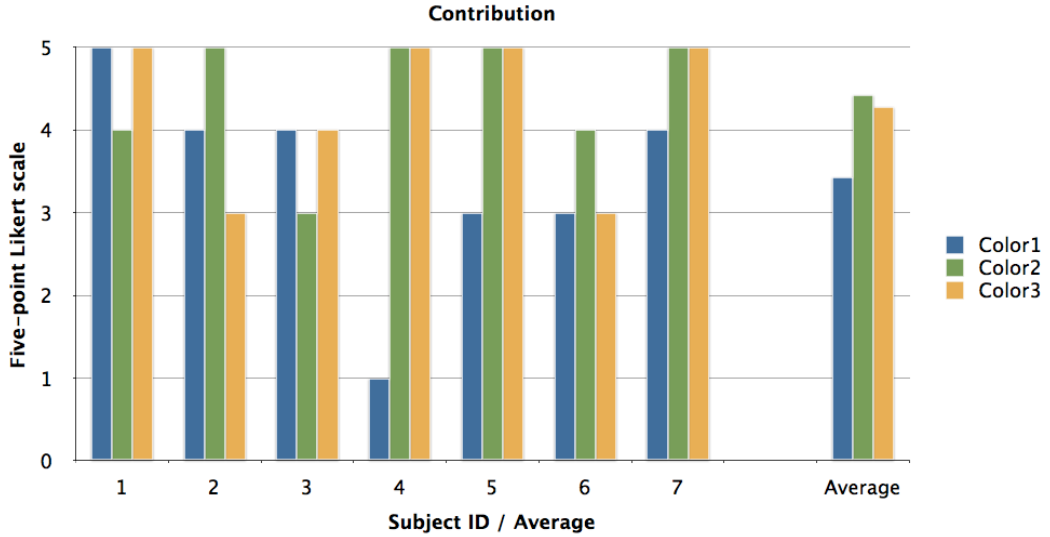



図 6.8: 5 段階リッカート尺度による各色付け手法の貢献度

図 6.7 の利用頻度について，平均値は *Color1* は他の 2 色より低い値となり，*Color2* 及び *Color3* については殆ど差が見られなかった．一方で被験者毎に比較すると，5 名の被験者が *Color1* に対して他色と同等かそれ以上の評価を付けており，これは実際に計測した利用率 (図 6.6) とは異なる結果となった．図 6.8 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．

実際に計測した利用率とアンケート結果が異なった原因としては，*Color1* に利用している色の種類が少なかったことが推測される．*Color2* 及び *Color3* では色相の分布を細かく見る必要があるため，色相が 2 種類のみの *Color1* と比べて読解に時間がかかる．よって実際に計測した利用率では *Color1* の割合が低くなったと考えられる．アンケートにおける *Color1* の利用頻度及び貢献度における評価の高さから，*Color1* 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．

また，各色付け手法に対するアンケートの意見の概要を以下に記す．

1. 次元間の相関に着目した色付け手法 (*Color1*) について．
   - 全体的な形状をイメージするのに役立った．
   - 色 (明度及び彩度) と傾き具合の関係性がわかり辛い．
2. X 軸次元のカテゴリ分布に着目した色付け手法 (*Color2*) について．
   - X 軸のデータ分布に偏りがある場合は色相で判別可能であり，見やすかった．
   - 均等に近い分布の場合は判断が難しい．
3. Y 軸次元のカテゴリ分布に着目した色付け手法 (*Color3*) について．

- 全体像を想像することができた．
- 矩形の分割順序がX軸カテゴリ →Y軸カテゴリであったため，*Color2* と比べると読解が難しかった．

*Color1* 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．*Color2* 及び *Color3* はデータを正確に読み取ることができ，また理解しやすいという意見を得られた．

### 6.4.4 Scatterplot との比較

色付き Mosaic Plot 及び Scatterplot について，4 種類の描画領域における平均正答率に対して検定を行う．検定に使用する色付き Mosaic Plot のカテゴリ分割数は各描画領域で平均正答率が最大となった分割数とし，いずれも Area Graph は非表示とする．

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．各描画領域に対して，以下の仮説を立てて検定する．

$$\begin{cases} \text{帰無仮説 } H_0 : \mu_C(n) - \mu_S(n) = 0 \\ \text{対立仮説 } H_1 : \mu_C(n) - \mu_S(n) \neq 0 \end{cases} \tag{6.2}$$

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 6.3 に示す．

表 6.3: 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 検定の結果

| 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 |

表 6.3 より，$n = 700$ 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．表 6.3 における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．

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．

# 第7章　ユースケース

本ツールを使用して実際のデータを分析するユースケースを示す．

## 7.1　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．

データの次元及び種類は表 7.1 に示す通りであり，その次元数は 37 次元である．本章では，日付や時刻についてもカテゴリデータとして扱う．

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．

図 7.1 に気象データの一部を示す．1 行が 1 レコード，1 列が 1 次元分を表現している．

| 年 | 月 | 日 | 都市 | 気圧 現地平均 | 気圧 海面平均 | 気圧 海面最低値 | 気圧 海面最低時分 | 降水量 合計 | 降水量 1時間最大値 | 降水量 1時間最大時分 | 降水量 10分間最大値 | 降水量 10分間最大時分 | 気温 平均 | 気温 最高値 | 気温 最高時分 | 気温 最低値 | 気温 最低時分 | 蒸気圧 平均 | 湿度 平均 | 湿度 最小値 | 湿度 最小時分 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2011 | 9 | 1 | 東京 | 1003.7 | 1007.8 | 1006.9 | 14:34 | 2.0 | 1.5 | 24:0 | 1.5 | 23:51 | 28.5 | 31.9 | 12:23 | 26.0 | 2:44 | 29.2 | 75 | 60 | 12:41 |
| 2011 | 9 | 1 | 大阪 | 993.4 | 1002.7 | 1000.9 | 23:27 | 5.0 | 2.5 | 17:19 | 2.0 | 16:29 | 27.2 | 29.9 | 11:20 | 25.1 | 2:23 | 29.7 | 83 | 70 | 14:59 |
| 2011 | 9 | 1 | 札幌 | 1004.8 | 1007.8 | 1006.4 | 11:38 | 0.0 | 0.0 | 22:26 | 0.0 | 21:36 | 26.0 | 32.0 | 13:12 | 22.4 | 4:22 | 24.2 | 73 | 42 | 13:44 |
| 2012 | 2 | 29 | 長野 | 972.6 | 1024.6 | 1021.2 | 23:45 | 4.0 | 1.0 | 10:21 | 0.5 | 10:22 | 0.6 | 4.6 | 21:20 | -1.6 | 6:03 | 5.4 | 85 | 48 | 0:09 |
| 2012 | 2 | 29 | 那覇 | 1010.0 | 1015.9 | 1013.4 | 0:27 | 0.0 | 0.0 | 15:40 | 0.0 | 14:50 | 19.1 | 21.5 | 11:39 | 17.6 | 23:17 | 15.2 | 69 | 57 | 8:52 |

図 7.1: 気象データの一例 (先頭から 22 次元分のみ)

## 7.2　ツールを用いたデータ分析

本ツールを用いて可視化した結果を図 7.2 に示す．図 7.2 は Y 軸次元のカテゴリ分布に着目した色付け手法を用いている．また量的次元に用いるカテゴリ分割手法は最大値と最小値を基準にした分割手法であり，カテゴリ分割数は 16 とした．なお画面左部の Matrix View における各色付き Mosaic Plot は，24 ピクセル四方の描画領域で表現している．

図 7.2 の Matrix View を見ると，色付き Mosaic Matrix において色相が横縞の模様になっていることが確認できた．これは色付けが Y 軸次元のカテゴリ分布に依存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

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

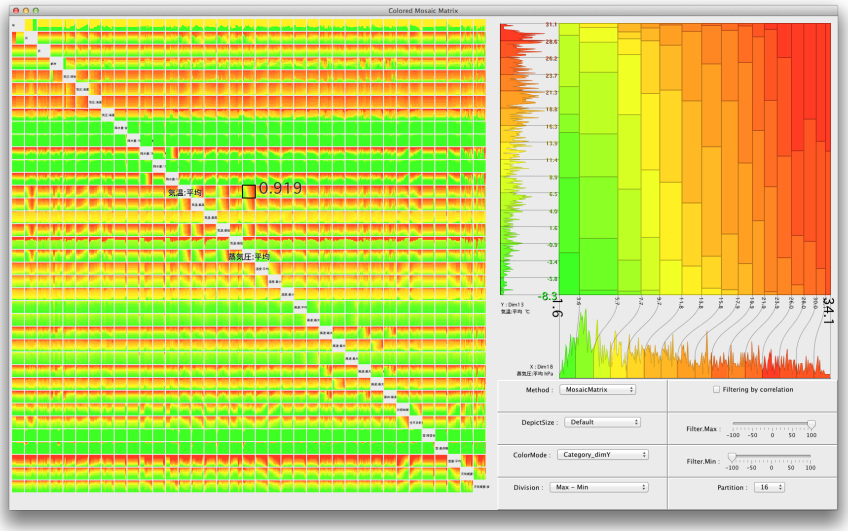



図 7.2: 本ツールによる気象データの可視化結果

次元においては，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．

次に次元間の相関で色相を決定する色付け手法に変更した図を図 7.3 に示す．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る．図 7.3 の Matrix View 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．図 7.3 の Detail View はこの次元対における色付き Mosaic Plot を表示しており，Area Graph を見ると両次元ともデータが偏っていることがわかる．また図 7.4 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．

緯度が最も高く北に位置している札幌おける気象データのみを Detail View から選択する．図 7.5 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 7.5 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．観測地が札幌であるレコードのみを対象として再描画した Matrix View が図 7.6 である．図 7.6 と図 7.3 を比べると，図 7.6 の方が最深積雪量と気温の間により強い負の相関関係があることが読み取れる．また図 7.7 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．

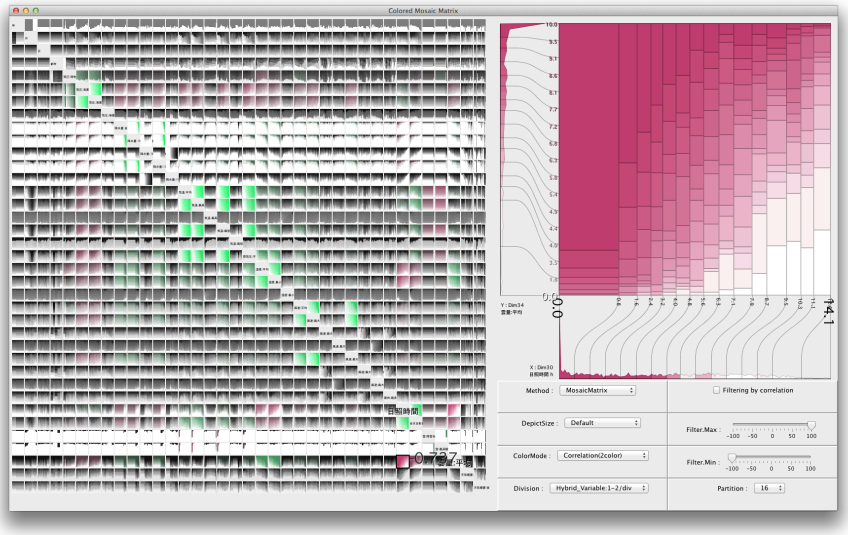



図 7.3: 次元間の相関に着目した色付け手法を用いた気象データの可視化結果

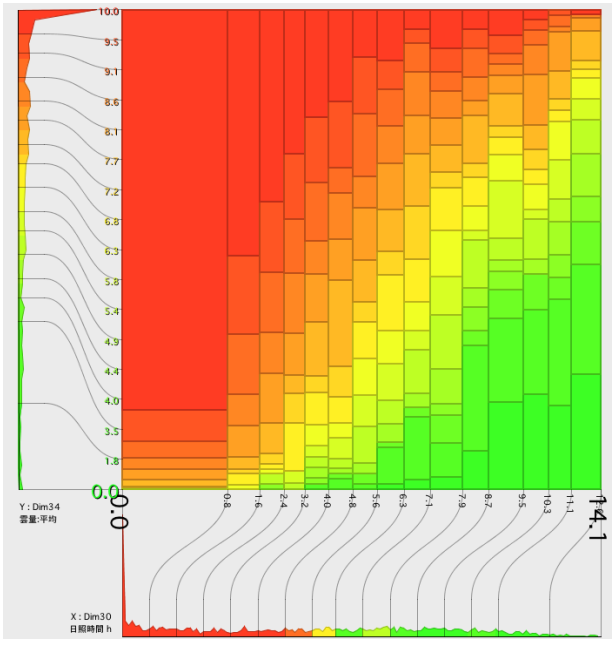



図 7.4: X 軸次元に日照時間，Y 軸次元に雲量平均を割り当てた Detail View

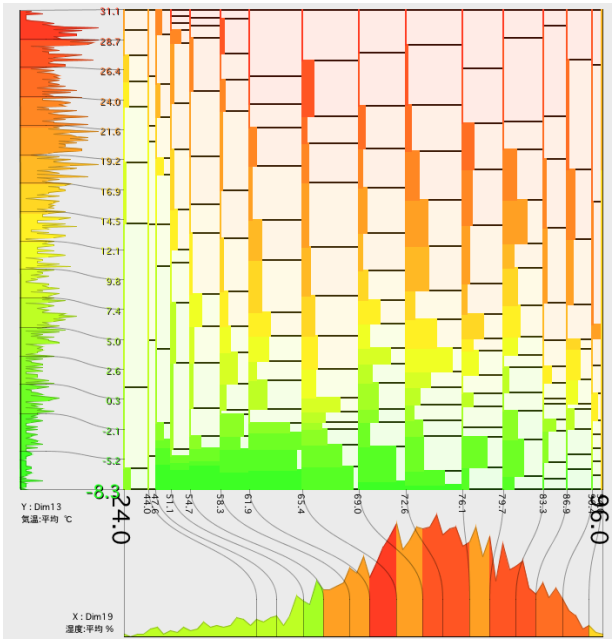



図 7.5: 観測地が札幌であるレコードのみを選択した状態の Detail View

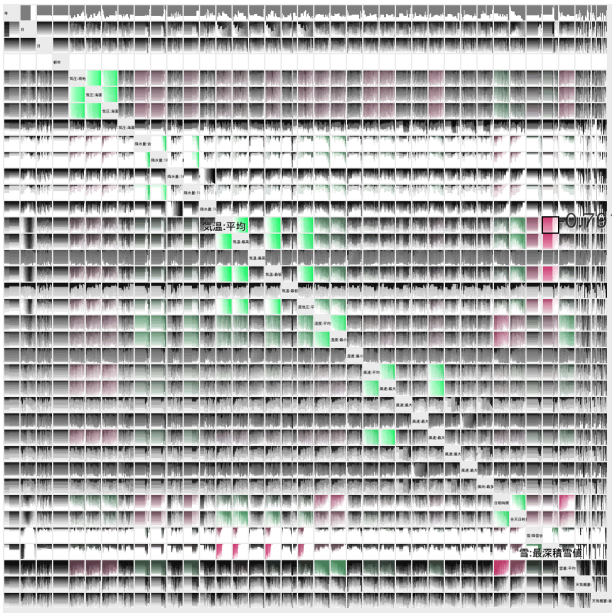



図 7.6: 観測地が札幌であるレコードのみで再描画した状態の Matrix View

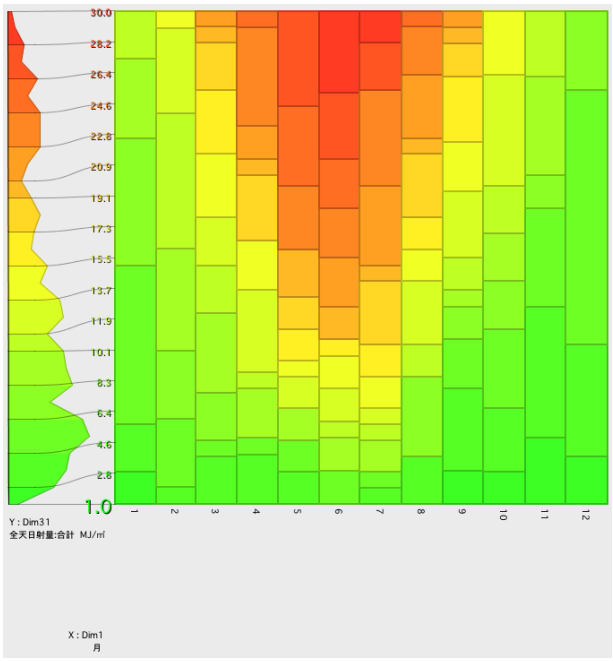



図 7.7: X 軸次元に観測月，Y 軸次元に全天日射量を割り当てた Detail View

表 7.1: 気象データの次元一覧

| 次元番号 | 次元名 | 単位 | データの種類 |
|---|---|---|---|
| 0 | 年 | | Categorical |
| 1 | 月 | | Categorical |
| 2 | 日 | | Categorical |
| 3 | 都市 | | Categorical |
| 4 | 気圧:現地平均 | $hPa$ | Quantitative |
| 5 | 気圧:海面平均 | $hPa$ | Quantitative |
| 6 | 気圧:海面最低値 | $hPa$ | Quantitative |
| 7 | 気圧:海面最低時分 | | Categorical |
| 8 | 降水量:合計 | $mm$ | Quantitative |
| 9 | 降水量:1 時間最大値 | $mm$ | Quantitative |
| 10 | 降水量:1 時間最大時分 | | Categorical |
| 11 | 降水量:10 分間最大値 | $mm$ | Quantitative |
| 12 | 降水量:10 分間最大時分 | | Categorical |
| 13 | 気温:平均 | $°C$ | Quantitative |
| 14 | 気温:最高値 | $°C$ | Quantitative |
| 15 | 気温:最高時分 | | Categorical |
| 16 | 気温:最低値 | $°C$ | Quantitative |
| 17 | 気温:最低時分 | | Categorical |
| 18 | 蒸気圧:平均 | $hPa$ | Quantitative |
| 19 | 湿度:平均 | $\%$ | Quantitative |
| 20 | 湿度:最小値 | $\%$ | Quantitative |
| 21 | 湿度:最小時分 | | Categorical |
| 22 | 風速:平均 | $m/s$ | Quantitative |
| 23 | 風速:最大 | $m/s$ | Quantitative |
| 24 | 風速:最大時風向 | | Categorical |
| 25 | 風速:最大時分 | | Categorical |
| 26 | 風速:最大瞬間 | $m/s$ | Quantitative |
| 27 | 風速:最大瞬間時風向 | | Categorical |
| 28 | 風速:最大瞬間時分 | | Categorical |
| 29 | 風向:最多 | | Categorical |
| 30 | 日照時間 | $h$ | Quantitative |
| 31 | 全天日射量:合計 | $MJ/m^2$ | Quantitative |
| 32 | 雪:降雪合計 | $cm$ | Quantitative |
| 33 | 雪:最深積雪値 | $cm$ | Quantitative |
| 34 | 雲量:平均 | | Quantitative |
| 35 | 天気概要:昼 | | Categorical |
| 36 | 天気概要:夜 | | Categorical |

# 第8章　結論

本研究では，フル HD ディスプレイ (1920 × 1080) 程度の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，Detail View を用いて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

本研究を行うにあたり，三末和男准教授には多大なご指導を頂きました．先生の丁寧なご指導のお陰で研究が順調に進み，無事に論文を執筆することができました．心から感謝しております．また志築文太郎准教授，高橋伸准教授，田中二郎教授には，研究室のゼミを通して様々な助言を頂きました．本当にありがとうございます．

インタラクティブプログラミング研究室の皆様には，公私共に大変お世話になりました．ゼミでの発表や日常生活の中で頂いた様々なご意見は，研究を進める上でも非常に参考になるものばかりでした．特に NAIS チームの皆様には，日々のゼミでご指摘を頂いたことはもちろん，普段の研究生活においても多大なご意見やご指摘を頂きました．深く感謝しております．

また，本手法を開発するきっかけを与えてくださった富士通研究所の皆様に感謝いたします．

そして，大学生活の中では沢山の人にお世話になりました．皆様のお陰で実りある学生生活を送ることができたことを感謝いたします．最後に，私が大学生活を送る上で，家族からは様々な面において援助をいただきました．心より感謝を申し上げます．

# 参考文献


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

[2] D. B. Carr, R. J. Littlefield, W. L. Nicholson and J. S. Littlefield. Scatterplot Matrix Techniques for Large N. In *Journal of the American Statistical Association*, Vol. 82, No. 398, pp. 424–436, 1987.

[3] L. Nováková and O. Štěpánková. Multidimensional clusters in RadViz. *SMO'06 Proceedings of the 6th WSEAS International Conference on Simulation, Modelling and Optimization*, pp. 470–475, 2006.

[4] J. Sharko, G. Grinstein and K. A. Marx. Vectorized Radviz and Its Application to Multiple Cluster Datasets. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 14, No. 6, pp. 1444–1451, 2008.

[5] B. Shneiderman. Extreme Visualization: Squeezing a Billion Records into a Million Pixels. *SIGMOD ' 08*, pp. 3–12, 2008.

[6] N. Elmqvist, P. Dragicevic and J.-D. Fekete. Rolling the Dice: Multidimensional Visual Exploration using Scatterplot Matrix Navigation. *IEEE Transactions on Visualization and Computer Graphics*, 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] A. Inselberg and B. Dimsdale. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *Proceedings of the 1st conference on Visualization (VIS '90)*, pp. 361–378, 1990.

[9] H. Siirtola. Combining parallel coordinates with the reorderable matrix. In *Proceedings of Coordinated and Multiple Views in Exploratory Visualization*, pp. 63–74, 2003.

[10] C. Viau, M. J. McGuffin, Y. Chiricota and I. Jurisica. The FlowVizMenu and Parallel Scatterplot Matrix: Hybrid Multidimensional Visualizations for Network Exploration. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 16, No. 6, pp. 1100–1108, 2010.

[11] J. Heinrich, J. Stasko and D. Weiskopf. The Parallel Coordinates Matrix. *Eurographics Conference on Visualization (EuroVis2012)*, pp. 37–41, 2012.

[12] Q. Cui, M. Ward, E. Rundensteiner and J. Yang. Measuring data abstraction quality in multiresolution visualizations. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 12, No. 5, pp. 709–716, 2006.

[13] J. Johansson and M. Cooper. A screen space quality method for data abstraction. *Computer Graphics Forum (EuroVis2012)*, Vol. 27, No. 3, pp. 1039–1046, 2008.

[14] M. Friendly. Mosaic Displays for Multi-Way Contingency Tables. In *Journal of the American Statistical Association*, Vol. 89, No. 425, pp. 190–200, 1994.

[15] H. Hofmann, A. P.J.M. Siebes and A. F.X. Wilhelm. Visualizing Association Rules with Interactive Mosaic Plots. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '00)*, pp. 227–235, 2000.

[16] M. Friendly. Extending Mosaic Displays: Marginal, Conditional, and Partial Views of Categorical Data. In *Journal of Computational and Graphical Statistics*, Vol. 8, No. 3, pp. 373–395, 1999.

[17] M. Friendly. *Visualizing Categorical Data*. SAS publishing, 2001.

[18] M. Friendly. A Brief History of the Mosaic Display. In *Journal of Computational and Graphical Statistics*, Vol. 11, No. 1, pp. 89–107, 2002.

[19] F. Bendix, R. Kosara and H. Hauser. Parallel Sets: Visual Analysis of Categorical Data. *IEEE Symposium on Information Visualization (InfoVis2005)*, pp. 133–140, 2005.

[20] Z. Geng, Z. Peng, R. S. Laramee, R. Walker, and J. C. Roberts. Angular Histograms: Frequency- Based Visualizations for Large, High Dimensional Data. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 17, No. 12, pp. 2572–2580, 2011.

[21] A. Lex, H.-J. Schulz, M. Streit, C. Partl and D. Schmalstieg. VisBricks: Multiform Visualization of Large, Inhomogeneous Data. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 17, No. 12, pp. 2291–2300, 2011.

[22] R. rao and S. K. Card. The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus+ Context Visualization for Tabular Information. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI '94)*, pp. 318–322, 1994.

[23] M. Freire, C. Plaisant, B. Shneiderman and J. Golbeck. ManyNets: an interface for multiple network analysis and visualization. In *Proceedings of the 28th international conference on Human factors in computing systems (CHI'10)*, pp. 213–222, 2010.

[24] B. Johnson and B. Shneiderman. Tree-maps: A Space-filling Approach to the Visualization of Hierarchical Information Structures. In *Proceedings of the IEEE Visualization*, pp. 284–291, 1991.

[25] A. Kobayashi, K. Misue and J. Tanaka. Edge Equalized Treemap. *16th International Conference Information Visualization (IV2012)*, pp. 7–12, 2012.

[26] T. Baudel and B. Broeksema. Capturing the Design Space of Sequential Space-Filling Layouts. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 18, No. 12, pp. 2593–2602, 2012.

[27] Y.-H. Fua, M. Ward and E. Rundensteiner. Hierarchical Parallel Coordinates for Exploration of Large Datasets. In *Proceedings of the conference on Visualization '99 (VIS '99)*, pp. 43–50, 1999.

[28] D. Feng, L.Kwock, Y. Lee and R. M. Taylor. Matching Visual Saliency to Confidence in Plots of Uncertain Data. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 16, No. 6, pp. 980–989, 2010.

[29] 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.

[30] T. Saito, H. N. Miyamura, M. Yamamoto, H. Saito, Y. Hoshiya and T. Kaseda. Two-tone pseudo coloring: compact visualization for one-dimensional data. *IEEE Symposium on Information Visualization (InfoVis2005)*, pp. 173–180, 2005.

[31] **石原色覚検査表 国際版** *38* **表**. (株) 半田屋商店. 2011.

# 付録

次ページ以降は，6 章の実験に利用した書類である．1 ページ目は実験の同意書である．2 ページ目から 4 ページ目までは実験に関する説明書である．5 ページ目以降はタスク終了後のアンケート用紙である．

# 高次元データ可視化手法に関する実験への協力のお願い

筑波大学　情報学群　情報科学類
小林 弘明

**研究の概要について**

本研究は，次元数の多い高次元データの傾向を，限られた画面領域内で把握するための可視化手法を開発するものです．高次元のデータは種類，量共に増え続けている一方で，データの可視化結果を表示する画面領域は限られています．本研究では，既存手法の色付けを工夫することで，次元数の多いデータを一度に俯瞰するための手法を開発します．

**被験者の必要性，方法，その成果について**

本研究では，開発手法の性質の調査や，既存手法との比較を行います．そのため，人を実験参加者とした実験が必要になります．本実験では，左側に表示される図を見て，それと同一のデータを現している散布図を，右側の 5 個の選択肢の中から選んでいただきます．

万が一，実験中に気分が悪くなったり，頭が痛くなったりした場合は，直ちに実験を止めて，実験実施者に声をかけてください．また，本実験は実験参加者の意思でいつでも中止することができます．本実験で得られた成果は，学術的利用目的のみに利用します．

**色覚検査の実施について**

本実験では色を用いた実験を行いますので，始めに石原色覚検査表を用いた色覚検査を受けていただきます．色覚検査は実験実施者が強要するものではなく，実験参加者はこれを拒否することができます．

**個人情報の保護について**

学会・論文などでデータを発表する際は，データおよびそれを統計的に処理したものだけを用います．実験参加者を表現するためには，記号・数字を用います．個人を特定できる情報は公表しません．ただし，実験参加者全体については，性別の実験参加者数，年齢の範囲，所属，国籍を公表することがあります．

---

同意書

私は，高次元データ可視化手法に関する研究について，研究の概要，被験者の必要性，方法，その成果，危険の回避，色覚検査の実施，個人情報の保護について十分な説明を受けました．

説明の際，本研究に協力することに同意しなくても何ら不利益を受けないこと，さらに，同意後も私自身の自由意思により不利益を受けず，いつでも撤回できることを聞きました．私は，このことを理解した上で被験者になることに同意します．

年　　月　　日

学籍番号 ＿＿＿＿＿＿＿＿＿＿＿＿

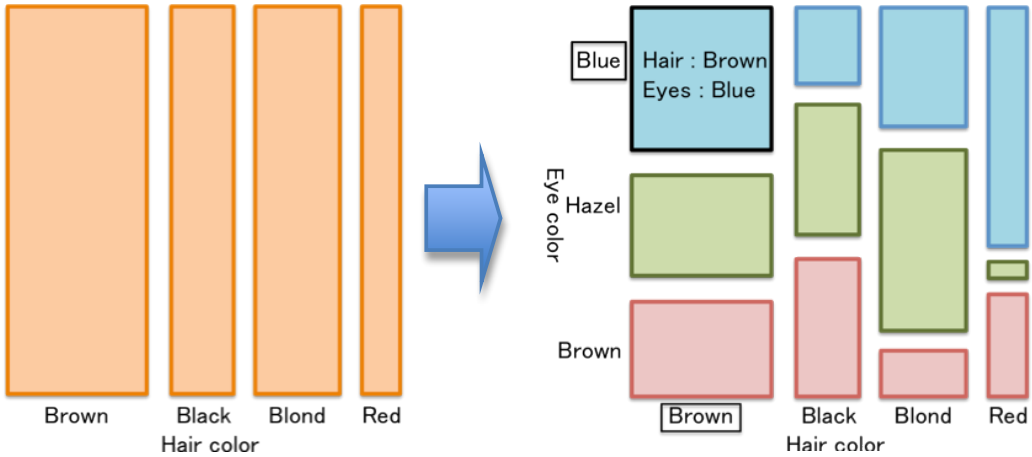
署名 ＿＿＿＿＿＿＿＿＿＿＿＿

■本研究が開発する可視化手法(以下，本手法)の概要

本手法は,Mosaic Plot を行列状に配置した Mosaic Matrix という手法を元にしています．Mosaic Matirx にデータの特徴が見える色付けを行うことで,個々が狭い描画領域であっても，高次元データの大局的傾向を読み取れるようになります．

■Mosaic Plot とは？

2 次元のカテゴリデータの可視化手法で，横幅が異なる 100%積み上げ棒グラフです．まず X 軸の各カテゴリの比率に応じて，矩形の横幅を決定して分割します(左下図)．続いて各矩形について,Y 軸の各カテゴリの比率で高さを決定して,さらに分割します(右下図)．このように分割することで，各矩形の面積がデータの比率を現します．右下図を例にすると，茶髪で目が青色の人は，金髪で目が茶色の人よりも多い，ということが読み取れます．



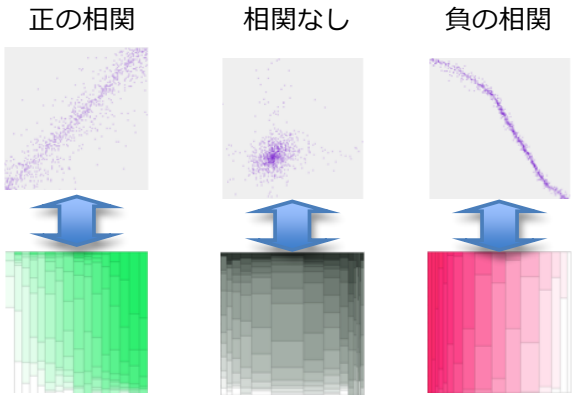

■色付けについて

本手法では，Mosaic Plot に独自の色付けを行います．本実験では 3 種類の色付けを切り替えて，データの特徴や分布を読み取っていただきます．

□手法 1 ：相関係数による色付け

次元同士の相関係数が正の場合は緑，負の場合は赤となるように色付けします．また Y 軸カテゴリの区別に明度と彩度を割り当てています．

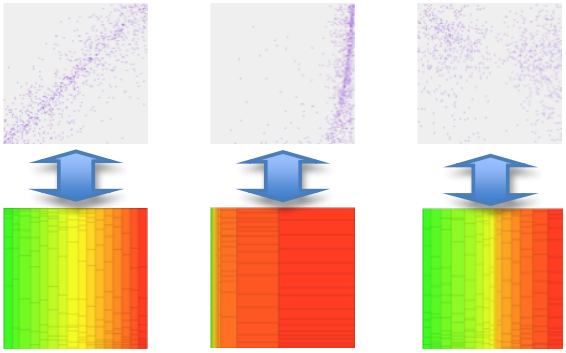
色が鮮やかであるほど強い相関を現し，グレーに近づくほど弱い相関であることを現しています．



□手法２：X 軸のカテゴリを区別する色付け

各矩形について，X 軸のカテゴリによって色相を決定します．Y 軸は関係ないため，縦縞の模様になります．

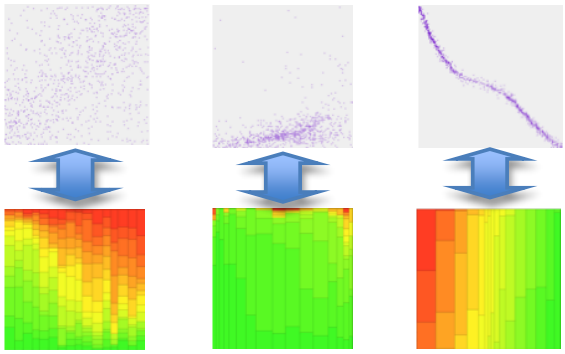
X 軸カテゴリのデータ量が均等であれば，緑から赤までの均等なグラデーションが見え，データが偏っている場合は，偏っている部分の色が多く見えます．



□手法３：Y 軸のカテゴリを区別する色付け

手法２の Y 軸版です．但し X 軸の時とは異なり，横縞の模様にはならないので，色の割合でデータの割合を読み取ります．

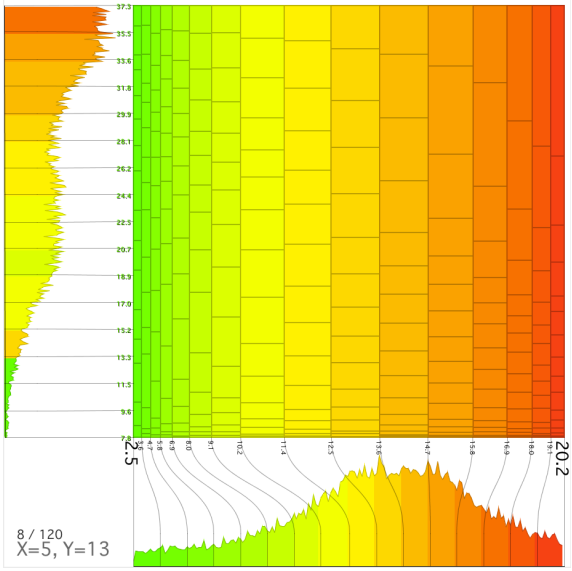
手法１でも，明度や彩度の違いによって,Y 軸の分布を読み取ることは可能です．手法３の場合は，色相で読み取ることが可能です．



■Area Graph(面グラフ)によるデータ分布の表示

本手法の Mosaic Plot では，データ分布をより理解しやすくするため，各次元の Area Graph を描画しています．

Area Graph はカテゴリ毎に線で区切られていて，それぞれ最も多い割合のカテゴリの色で塗られています．

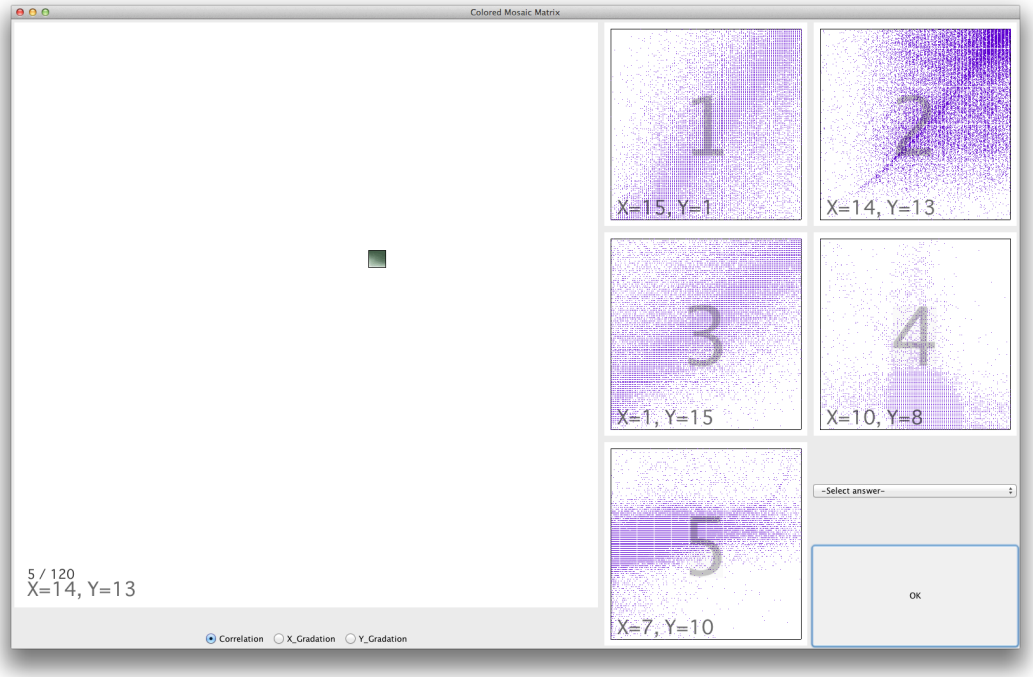


■本実験のタスク

本実験では，色付き Mosaic Matrix の可読性を調査するため，4 種類の大きさの Mosaic Plot に対して，その可読性を調査します．また既存手法である散布図に対しても，同様の実験を行います．

実験ツールの画面左には，Mosaic Plot または散布図が表示されます．画面右には，5 個の散布図が表示されます．左の図を見て，右図の中から同じデータだと思われる散布図を選んでいただきます．Mosaic Plot に関しては，3 つの色付け手法を切り替えることができます．色の割合や模様を参考に，正解だと思う散布図を選んでください．

本番の前に，練習モードで本手法と実験ツールに慣れていただきます．十分に慣れたといずれかが判断した時点で練習モードを終了し，本番モードでの実験を開始します．本番の全問題が終了すると，自動的にプログラムが終了します．その後で別紙のアンケートにお答えいただき，本実験は終了となります．




■注意事項

問題は全 120 問で，時間は無制限です．本実験は速度ではなく正確性を調査するものですので，正しい選択肢を選ぶように心がけてください．長時間の実験になることが予想されますので，途中で自由に休憩をお取りください．

また，ツールの仕様として，30 問目と 60 問目の終了時に 15 秒間ほど停止します．これはエラーではありませんので，暫くお待ちください．

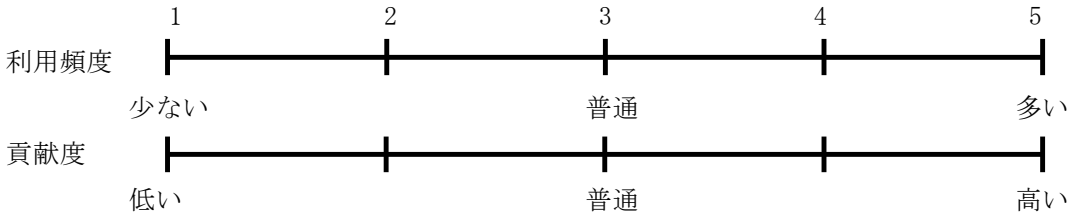
１．ご自身についてお答えください.

| | |
|---|---|
| 氏名 | |
| 性別　　□ 男　　□ 女 | 年齢　　歳 |
| 所属 | 国籍 |

２．本手法で用いた色付け手法についてお答えください.

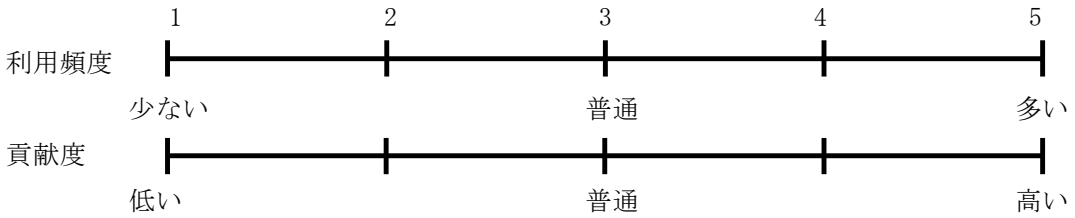
今回の実験では，3 種類の色付け手法を利用して，データの特徴や分布を読み取っていただきました．それぞれの色付け手法について，どの程度利用したか，回答を選ぶ際の判断材料としての貢献度について教えてください.

□手法１：相関係数による色付け




| |
|---|
| 理由： |

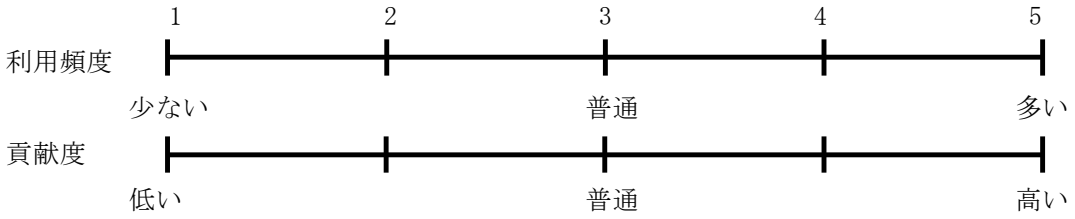
□手法２：X 軸のカテゴリを区別する色付け




裏面もあります

| |
|---|
| 理由： |

□手法３：Y 軸のカテゴリを区別する色付け




| |
|---|
| 理由： |

３．その他，本実験に関してご意見などありましたらご記入ください．

| |
|---|
| |

実験は以上です．

ご協力いただき，ありがとうございました．