筑波大学大学院博士課程

システム情報工学研究科修士論文

多変量データセットの特徴探索及び比較分析を支援する可視化手法の開発

小林 弘明

修士（工学）

(コンピュータサイエンス専攻)

指導教員　三末 和男

2015 年 3 月

概要

多変量データセットの分析における、特徴的な部分集合の探索および比較の支援を目的とした研究を行った。本論文の目的を達成するため、多変量データセットに含まれる特徴的な部分集合をインタラクティブに探索でき、かつデータ分布を視覚的に比較できる表現や機能を備える視覚的分析ツールを開発した。開発した分析ツールは、まず特徴的部分集合を探索する起点や補助となる情報として、多変量データセットが持つ変量間の関係性及びレコード間の関係性を可視化して提示する。またデータ間の差異の探索を支援するため、データ分布の差異が目立つような色付けを施した視覚的表現を提供する。さらに本分析ツールには、部分集合間の比較を支援するためのインタラクティブな機能を備えている。本論文では、開発した視覚的分析ツールの有用性を示すためのケーススタディを実施した。本分析ツールのケーススタディでは、ソーシャルネットワークサービスの一つであるブログの解析データの分析を行い、特定のユーザ集団に関する分析結果をまとめると共に本分析ツールの有用性を示した。

目 次

図 目 次

表 目 次

第1章　序論

1.1　多変量データセット

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。データ中の一つの属性は一つの変量（Variate）または次元（Dimension）として考える事ができ、多くの属性を持つデータは多変量データ（Multivariate data）または多次元データ（Multidimensional data）と呼ばれる。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、風向きなどが、それぞれ一つの変量として保存されている。

本論文では、意味的な観点から複数の部分集合に分割可能なデータの集合を特にデータセットと呼ぶこととする。また、多変量から成るデータセットのことを多変量データセットと呼ぶこととする。多変量データセットの一例として、関東地方の一都六県において観測された気象情報が含まれている気象データが挙げられる。関東地方の気象データセットの場合、データセット内には内陸山間部の気候特有の傾向と海沿い平野部の気候特有の傾向が両方含まれることになる。このように、多変量データセットはそれぞれの部分集合が持っている特徴を兼ね備えているため、全体の傾向や詳細な特徴の把握を慎重に行う必要がある。

多変量データセットを分析して得られる知見は、様々な分野に活かすことが可能である。例えば気象データセットを分析した場合、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セットと気象データセットを併用して分析することにより、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 SNS ユーザの発言を記録したデータセットを分析すると、特定の商品に関心を示している集団の傾向把握や、その集団の嗜好を踏まえた効果的なマーケティングも可能になる。

ところで、データを構成する変量には、その性質に応じた 4 種類の尺度水準が存在する [1]。名義尺度（Nominal scale）の変量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量である。順序尺度（Ordinal scale）の変量は、順序を付けて比較できる変量である。間隔尺度（Interval scale）の変量は順序尺度の要件を満たし、かつ加減演算の結果に意味がある変量である。比例尺度（Ratio scale）の変量は間隔尺度の要件を満たし、かつ変量間の比率や乗除演算にも意味がある変量である。本論文ではこれらの尺度水準の変量のうち、名義尺度および順序尺度の変量を質的変量として、また間隔尺度および比例尺度の変量を量的変量として扱う。多変量データセットの多くは、質的変量と量的変量の両方が含まれている。先の気象データの例であれば、気温や湿度は量的変量、風向きは質的変量である。

1.2　データの視覚的分析

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。可視化によってデータセットを目で見てわかりやすい形に表現することにより、データセットに含まれる有用な情報を効率的に探索して分析することができる。さらに可視化にインタラクティブな視覚的インタフェースを追加することにより、分析的推論を促進することができる [2]。

最も基本的な可視化手法の一つとして、散布図（Scatterplot）が挙げられる。一般的な散布図では、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 変量を散布図で可視化する。この散布図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、また一つの点は 1 日分のデータに対応している。この図から、二変量の関係性について分析することが可能である。なお、気象データにおける “1 日分の気象情報” のように、多変量データにおける一件ごとのまとまったデータの並びのことをレコードと呼ぶ。

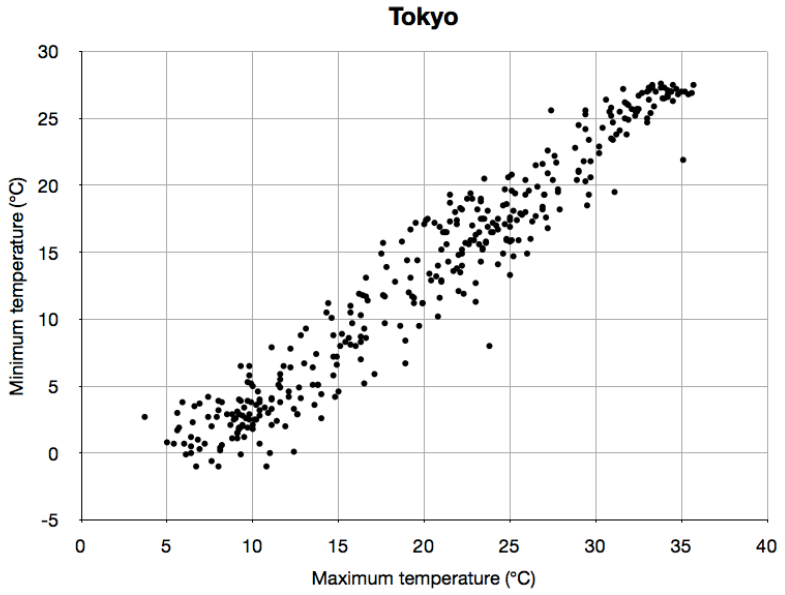

図 1.1: 散布図の一例。横軸は東京における最高気温、縦軸は東京における最低気温、また一つの点は 1 日分のデータに対応している。

1.3 多変量データセットの視覚的分析における課題

多変量データセットの分析需要から、これまで多変量データセットの可視化手法が数多く提案されてきた。例えば VisBricks [3] は、比較的多数の変量やレコードを含む多変量データセットを可視化・分析できるツールである。VisBricks では表示したい変量やレコードを一つのまとまり（ブロック）として、ブロック内やブロック間のデータの関係性を可視化できる。一方で、データセットをブロックとして細かく分割して表現しているため、複数のブロックが含まれるような特徴的な部分集合の探索は困難である。また変量の軸がブロックごとに独立しているため、ブロック間でのデータセットの比較が困難である。

現状の多変量データセットの視覚的分析における課題について整理する。まず一つ目の課題として、多変量データセットに含まれている特徴的な部分集合の探索が困難なことが挙げられる。多変量データセットは複数の部分集合が持っている特徴を兼ね備えているため、データセット全体を概観しただけでは特徴的な部分集合の発見が難しい。多変量データセットから特徴的部分集合を探索するためには、そのようなタスクを意図して設計された視覚的表現およびデータ操作のインタラクションが必要不可欠である。

二つ目の課題は、多変量データセットの多角的な比較が困難なことである。データセット全体から部分集合の特徴を発見するためには、着目している部分集合とデータセット全体（または他の部分集合）を比較してデータの差異を確認することが有効である。多変量データセットを比較するためには、“レコード的な観点からの比較” と “変量的な観点からの比較” の両方を連携して行う必要がある。また複数の部分集合を比較するためには、比較タスクを行いやすくするための視覚的補助が必要になる。しかし既存手法では、これらの比較のための支援が行われてこなかった。

1.4 本論文の目的とアプローチ

本論文では、多変量データセットの分析における特徴的な部分集合の探索および比較の支援を目的とする。目的を達成するためのアプローチとして、多変量データセットに含まれる特徴的な部分集合をインタラクティブに探索でき、かつデータ分布を視覚的に比較できる表現や機能を備える視覚的分析ツールを開発する。その分析ツールは、特徴的部分集合を探索する起点や補助となる情報として、多変量データセットが持つ変量間の関係性およびレコード間の関係性を可視化して提示する。さらにデータ間の差異の探索を支援するため、データ分布の差異が目立つような色付けを施した視覚的表現を提供する。本論文では、これらの表現と比較分析のためのインタラクティブな機能を備えた視覚的分析ツールを開発し、ソーシャルネットワークサービスの解析データを用いたケーススタディによりその有用性を示す。

1.5 貢献

本論文の貢献は以下の通りである。

1. データ分布の比較タスクを支援する表現として、比較対象間のデータ分布の差異を色から発見できるような視覚的表現を開発した点。開発した視覚的表現は、$L^*a^*b^*$ 色空間を利用した色付けを行うことにより、色相の違いによる色差の影響を考慮している。

2. 多変量データセットにおける特徴的部分集合の発見を支援する多変量データセットの視覚的分析ツールを開発した。開発した分析ツールは、多変量データセットが持つ変量間の関係性およびレコード間の関係性を可視化して提示する。さらに着目した部分集合同士を比較できるような視覚的表現およびインタラクションを提供する。これらの表現および機能により、特徴的部分集合の探索を可能にした。

3. 開発した視覚的分析ツールを用いてソーシャルネットワークサービスの解析データを分析し、特定のユーザ集団に関する集団的傾向をまとめた。

第2章　関連研究

本論文の関連研究として、多変量データまたは多変量データセットを対象とした可視化手法および分析手法について述べる。また多次元データを対象とした表現手法の多くは多変量データセットにも適用可能であるため、それらの既存手法についても言及する。

また本章では、分析における比較タスクを支援している表現手法についても紹介する。ここでは多変量データセット以外を対象とした研究についても言及する。

2.1　表形式の表現手法

多変量データの最も基本的な表現方法は、表形式での視覚的表現である。例えば Microsoft Excel [1]では、表の列に変量、行にレコードを対応付けることによって多変量データを表現できる。Excel は基本的かつ一般的に広く用いられているツールであるが、複雑なデータ操作を伴う分析は困難である。

Table Lens [4] は、大規模なデータを表形式にて可視化する手法である。Table Lens は複数のレコードを 1 行で表現可能であり、大規模な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 [5] は、大規模な多変量ネットワーク構造のデータセット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。

Suematsu らは、時系列付き多変量データを対象にしたヒートマップベースの可視化手法を提案した [6]。量的変量と質的変量を持つ多変量データに対して、一つの変量に 1 行を割り当てて類似度順に並べて可視化する。時刻を列に割り当てた上で、色によって各時刻における各変量の値を表現する。Suematsu らの手法は各変量の時間変化を確認する際に有効な表現手法である。

表形式による多変量データの視覚的表現は、比較的馴染み深く理解しやすいという利点がある。一方で、データセットの分割やデータセット間の比較などの複雑な分析操作は困難である。また視覚的表現自体の空間効率が悪いため、一度に表示できる変量数が少ないという問題がある。

[1]Microsoft, http://office.microsoft.com/excel/, 閲覧:2015.01.04

2.2 直交座標系を組み合わせた表現手法

直交座標を用いた可視化手法は、多変量データの一つの変量に対して一つの座標軸を割り当てた上で、座標軸を直交するように配置することによりデータを表現する。2 次元平面上の直交座標を用いれば 2 変量、3 次元空間上の直交座標を用いれば 3 変量を同時に表現できる。例えば 1.2 節にて紹介した散布図（図 1.1 を参照）は、2 次元平面上の直交座標系を用いた基本的な可視化手法の一つである。単一の直交座標系表現に対応付けられる変量の数には限りがあるが、複数の直交座標系表現を同時に用いることにより、多変量データ全体を可視化できる。

散布図行列（Scatterplot Matrix） [7] は、古くから用いられてきた代表的な多変量データの可視化手法である。散布図行列は、全ての組み合わせ可能な変量対の散布図を行列状に並べて表示する。例えば N 変量データの場合、$N(N-1)$ 通りの散布図を表示することで、データ全体を俯瞰する。

SCATTERDICE [8] は散布図行列を拡張した可視化手法である。SCATTERDICE では、多変量データにおける各変量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により実現する。通常は 2 次元の散布図を表示し、表示する変量を切り替える際には、奥行きにあるもう 1 変量との散布図が 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 次元の散布図が持つシンプルさを活かしつつ、変量の切り替えを直感的に行うことができる。

Colored Mosaic Matrix [9] は、Mosaic Plot [10, 11] を拡張した高次元データの可視化手法である。Mosaic Plot にデータ分布や特徴を読み取れるような色付けを行った上で、それらを散布図行列と同様に行列上に配置する。これにより、次元数が非常に多い高次元データのおおまかな分布や傾向を色のパターンから読み取ることができる。

直交座標系の可視化手法に共通する問題として、表現空間の次元数以上の変量間の比較が困難である点が挙げられる。例えば 2 次元平面上の直交座標系では 2 変量間の関係性しか表現することができない。そのため、多変量データセットの概観を直感的に得ることができない。また 3 変量以上の間の関係性を分析するためには、複数の図を同時に閲覧しなければならない。また直交座標系は、データセットから複数の部分集合を取り出して比較するには不向きな座標系である。複数の部分集合を同一の図上に可視化した場合、読解や分析が困難になるためである。

2.3 並行座標系を用いた表現手法

Parallel Coordinate Plot（PCP） [12, 13] も散布図行列と同様、複数の変量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。各変量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、一つのレコードを 1 本の線で表現する。

一般的な PCP においては、隣り合った変量同士でしか直接的に関係を見ることができない。一方で PCP を拡張した Heinrich らの手法 [14] では、軸の並び順を変えた PCP を縦に複数本

並べることにより、全変量ペアの隣り合わせを網羅している。これにより、座標軸を並び替えることなく全ての変量ペア間の関係性を直接閲覧できる。

PCP は線の密集度合いや線の傾きにより、データ分布や隣接変量間の関係性を表現する。しかし PCP では、座標軸間の線の分布から関係性を読み取る必要があるため、隣り合っていない変量間の関係性を把握することが難しい。また、PCP は一つの線が一つのレコードに対応した表現であるため、レコード数が多い多変量データセットの比較は視覚的に困難である。

Parallel Sets [15] は、PCP と Mosaic Plot を組み合わせた多変量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、レコードが持っているカテゴリ間のつながり具合に応じて、座標軸間の矩形を幅のある帯によって繋ぐ。このとき、1 個の矩形や 1 本の帯は複数のレコードに対応している。またカテゴリに応じて矩形の色を設定することにより、複数変量間の関係性を表現できる。Parallel Sets は質的な変量から成る多変量データの表現手法として優れており、その表現の性質上、視覚的混雑度を抑えながらレコード数の多い多変量データを集約的に表現できる。一方で、量的な変量の表現を行うには不向きな手法である。また複数のデータセットを同時に座標軸上に表示した場合、データセットの区別ができなくなってしまう。そのため、本論文が目的としているようなデータの比較を行うことができない。

Angular Histograms [16] は、多変量データの各変量についてヒストグラムを作成し、PCP と同様に並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、ヒストグラムの各ビンの傾きを設定することにより、隣り合う変量同士の関係性を表現する。Angular Histograms は複数のレコードを一つのヒストグラムのビンによって集約的に表現できる。一方で、Angular Histograms は Parallel Sets と同様、複数のデータセットを比較できるような表現手法ではないため、本論文が目的とするようなデータの比較には不向きである。

Andrienko らは、データの部分集合が持つ特徴を探索するための表現手法を提案した [17]。Andrienko らの手法では、各変量の値をクラスタリングして円によって表現する。さらに複数の部分集合の分布を色を変えて重畳表示することにより、各変量のおおまかな分布を部分集合間で比較できる。Andrienko らの手法は量的な変量の比較は可能だが、質的な変量の表現ができない。また円による分布の表現は直感的ではなく読み取りが難しいため、注意深く探索しなければ各変量において差のある部分を発見できない。

Lex らは、量的な変量から成る多変量データセットを可視化して比較できる分析ツールを開発した [18]。分析ツールに用いている表現は Parallel Sets をベースとしており、変量をクラスタリングした上で変量クラスタ間を帯で繋ぐことによって多変量データセットを可視化している。また帯の色を変化させることにより、クラスタごとの帯の差異を比較できる。Lex らのツールは多変量データセットにおける変量クラスタ間の関係性を比較できる。一方で、レコード的な観点から多変量データセットの部分集合を分割して、それらを視覚的に比較することができない。

2.4 次元削減を用いた表現手法

多変量・多次元のデータを表現する手法として、次元削減手法を適用し、多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にて可視化する手法が開発されてきた。次元削減手法の代表例としては、主成分分析（Principal Component Analysis）による手法や多次元尺度構成法（Multi Dimensional Scaling）が一般に知られている。またRadViz [19] や Vectorized RadViz [20] は、変量を円周上に配置することにより、多変量データを 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、各変量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。また、質的な変量では距離関係を定義できないため、質的な変量を可視化することができない。

変量数・次元数の多いデータを表現するアプローチの一つとして、分析においてユーザが注目する次元のみを可視化する方法が挙げられる。Sips らの手法 [21] 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。これにより、データセットにおける特徴的な部分に焦点を絞った分析が可能である。しかし Sips らの手法では散布図や散布図行列を基本としたデータ表現を採用しており、またデータの比較操作に関する支援は行われていない。そのため、多変量データセットにおける複数の部分集合を比較するには不十分である。

2.5 複数の座標系を組み合わせた表現手法

Siirtola は、PCP と並び替え可能な行列表現を組み合わせた表現手法の有効性を実証実験により示した [22]。実験に用いた表現手法には、行列表現側からデータ選択やハイライト、ソートなどのデータ操作を行うと、PCP 側の表現にもその操作が反映されるようなインタラクションを備えている。Siirtola は実験の結果から、複数のビューを組み合わせることにより分析タスクが効率化できることを実証した。

1.2 節にて例示した VisBricks [3] は、比較的多数の変量やレコードを含む多変量データセットを可視化・分析できるツールである。VisBricks では表示したい変量やレコードを一つのまとまり（ブロック）として、ブロック内やブロック間のデータの関係性を可視化できる。一方で、変量の軸がブロックごとに独立しているため、ブロック間でのデータの比較が困難である。またブロック内での量的変量の表現には PCP が用いられているが、2.3 節で述べた通り、PCP はデータの比較には不向きな手法である。

Elias ら [23] は、分析の専門家の意見に基づいたダッシュボード形式の視覚的分析ツールを開発した。Elias らのツールは分析環境における可視化のスナップショットとテキスト記述からレポートを作成できる機能を備えている。しかし Elias らのツールは、本論文が対象としている多変量データセットの分析に焦点を当てた手法ではない。そのため、Elias らのツールを用いて変量や部分集合を比較分析することは困難である。

Viau らは、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を開発した [24]。Viau らが開発した手法では、散布図行列と 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。複数のビューから対象データを様々な角度から閲覧することができるが、集約的な表現を行っていないため、複数のデータセットを表示して比較するには不向きな手法である。

GPLOM [25] は散布図行列と同様、行列を用いて多変量データを可視化する手法である。GPLOM の最大の特徴は、変量の尺度水準に合わせて最適なチャートを自動選択する点である。量的変量のペアの表現には散布図を、質的変量のペアの表現にはヒートマップを、量的変量と質的変量のペアの表現には棒グラフを利用する。このように利用するチャートを使い分けることによって、質的変量と量的変量の両方を併せ持つ多変量データにも対応でき、また専門知識を持たなくても理解しやすい表現ができる。GPLOM ではデータ選択時のハイライト表現により、多変量データの一部を他の部分と比較することができる。一方で、GPLOM は多変量データセットを対象とした分析ツールではない。そのため、本論文が目的としているような任意の部分集合同士の比較を支援していない。

Domino [26] はデータセットから意味を持った部分集合を抽出したり組み合わせることが可能な分析ツールである。分割ブロック、数値ブロックおよび行列ブロックを基本的な視覚的ユニットとしており、ユーザがブロックを組み合わせることによってデータセットを可視化して分析できる。Domino はデータセットから部分集合を探索するツールとして有能であるが、部分集合間の比較という観点では支援を行なっていない。

2.6 分析における比較タスクを支援している表現手法

Malik らは、時空間情報付きの多変量データセットの相互関係を分析するツールを開発した [27]。Malik らが開発した分析ツールでは、棒グラフにおいて値が大きな部分や差のある部分のみに色を付けることにより、データが持つ特徴の発見を視覚的に支援している。一方で、データセットから特徴的な部分集合を探索する際の視覚的支援は行なっていない。

Kehrer らは、量的な変量と質的な変量の両方が含まれるような多変量データを対象とした、small-multiples による表現を用いた比較分析ツールを開発した [28]。前後関係を持つような意味的階層にデータを分離した上で、絶対参照および相対参照を用いて前後の差異を比較できるような可視化を行う。Kehrer らの手法では質的な情報から部分集合を作成できるが、量的な観点から部分集合を作成することができない。

Diversity Map [29] は、多数の多変量データセット間の差異を可視化する手法である。PCP のように一つの座標軸に一つの変量を割り当てた上で、各変量に含まれる値を矩形によって表現する。差異の大きな矩形ほど濃く目立つ色が割り当てることにより、データセットにおける差異の発見を支援している。Diversity Map は差異の発見は可能であるが、その詳細を表現から読み取ることができない。またデータセットの分割という観点において支援を行なっていない。

多変量データを対象とした表現手法ではないが、表現の比較を支援している研究について紹介する。VAICo [30] は画像を比較分析するための表現手法である。VAICo は画像データのコンテキスト情報を提供し、大規模な画像セットにおける差異・類似点を可視化する。画像の差分領域位を検出した上で、画像セット間の差分をハイライト表示する。これにより、画

像セット間で差分がある位置とその詳細な情報をひと目で把握できる。

第3章　多変量データセット分析の要求事項

本論文では、多変量データセットを分析する際の要求事項を以下の三点に大別する。

1. 多変量データセットを視覚的に分析するための表現への要求事項
2. 特徴的部分集合の探索のための要求事項
3. 多変量データセットの比較のための要求事項

3.1　多変量データセットを視覚的に分析するための表現への要求事項

1.1 節にて述べたとおり、データを構成する変量には複数の尺度水準が存在しており、その変量の尺度水準に応じて適した表現手法は異なっている [1]。例えば GPLOM [25] は、変量の性質に応じた表現手法を組み合わせて行列上に配置することにより、量的変量と質的変量を併せ持つ多変量データの分析を支援している。GPLOM のように、複数の尺度水準の変量が含まれるような多変量データセットを比較するためには、変量の尺度水準に適した表現手法を組み合わせる必要がある。

Shneiderman は大規模データを対象にした可視化手法を 3 種類に分類した [31]。以下にそれぞれの特徴を示す。

- Atomic Visualizations は、一つのマーカが一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。なお、マーカは手法における表現の最小単位のことを指している。代表的な手法では、散布図行列 [7] や PCP [12] が Atomic Visualizations に分類される。Atomic Visualizations の手法は個々のレコードの詳細を確認できる利点がある。一方で、変量数やレコード数の多い多変量データセットを可視化すると視覚的混雑度が高くなり、データの読み取りが困難になってしまう。

- Density Plot Visualizations は、色などによりレコードの密集度合いや分布を表現する可視化手法の分類である。Fua らの手法 [32] は Density Plot Visualizations に分類される手法の一例であり、データの密度に応じた色の明暗によってデータ分布を表現している。また Feng らの手法 [33] は不確実データを対象にしており、色のぼかしを利用してデータの密度を表現している。これらの Density Plot Visualizations の手法は色の位値やぼかしなどによって分布を表すため、読み取り時に曖昧性が発生してしまう。

- Aggregate Visualizations は、一つのマーカが複数のレコードに対応する集約的な可視化手法の分類である。Aggregate Visualizations に分類される手法の例として、Parallel Sets [15] や Angular Histograms [16] は、レコードを集約的に可視化することによって分析を支援している。集約的な可視化手法は視覚的混雑度を抑えながら多数のレコードを表現できるため、多変量データセットを概観する手法として適している。

以上より、多変量データセットを概観を表現する際は、集約的な表現手法を利用できることが望ましい。またデータセットの詳細を確認するために、Atomic Visualizations のような個々のレコードを確認できる表現も利用できることが望ましい。

3.2 特徴的部分集合の探索のための要求事項

視覚的な分析のプロセスにおいて Keim らは、まず可視化結果から分析でき、そこからさらに分析操作を行えることが重要である、としている [34]。このプロセスに沿った分析を行うためには、多変量データセット内において他とは異なる傾向を持つ変量や、類似の変量同士などを可視化結果から探索できる必要がある。さらにその上で、次の分析の起点となるような特徴の発見を支援できる必要がある。

複数の変量を扱う際の主な課題の一つは、どこから探索を開始するかを決めることである [35]。場合によっては、一度にデータ全てを閲覧し、次に興味深い方向を指し示すことができる興味深い点を探すのが良い。また、他のデータから孤立しているデータ点、いわゆる外れ値に対して興味を持つこともある。それはデータセットの中で最も面白い部分である場合もあれば、ただのタイプミスであることもある。分析の際は、外れ値のような他とは異なるデータの部分集合を確認できる必要がある。

多変量データセットから特徴を発見するためには、変量的な観点からの分析だけではなく、レコード的な観点からの分析も必要になる。これらの分析タスクを行うためには、両観点から多変量データセットを閲覧するための視覚的表現が必要である。また同時に、特徴的な部分集合を発見した後にさらに部分集合を絞り込んでいけるような、ドリルダウン式の分析操作が可能でなければならない。

3.3 多変量データセットの比較のための要求事項

多変量データの分析において、複数のデータ集団を用いた比較はよく行われるタスクである [18]。このような比較タスクを実現するためには、複数のデータセットが持つ変量同士を同時に比較し、特徴的な部分を持つデータ集合を探索できる必要がある。

多変量データセットを表現する際には、散布図行列 [7] や VisBricks [3] など、複数のチャートを並べることによって複数のデータセットを表現する手法が広く用いられてきた。これらの手法は同時に複数のデータセットを確認できる利点がある。一方で、異なる座標軸上のデータセット同士を目で見て比較することは困難である。変量内の比較を視覚的に支援するため

には、同一座標軸内に複数のデータセットを表現することが望ましい。また特に量的な値を比較する際は、表現の位置や長さによる比較が有効である [36]。変量間の比較を視覚的に支援するためには、位置や長さによってデータを比較できるような座標系および表現を用いることが望ましい。

多変量データセットを分析する際、3.2 節で述べた Keim らの視覚的分析プロセスに則った比較を行うためには、差異のある部分などの特徴的な部分をひと目で発見できる必要がある。例えば VAICo [30] では、データセット間の差分を詳細とともにハイライト表示している。また Diversity Map [29] では差異を矩形の色から確認できる。これらの手法のように、比較分析を円滑に進めるためには、データセット間の差異をひと目で確認できる表現を用いることが望ましい。

第4章　Blade Graph: 量的変量の比較のための視覚的表現

本章では、多変量データセットにおける量的変量の比較に用いる視覚的表現について述べる。

4.1　視覚的表現の要件

3.1 節にて述べた多変量データセットを視覚的に分析するための表現への要求事項、および 3.3 節にて述べた多変量データセットの比較のための要求事項に基づき、比較のための視覚的表現における要件を整理する。表現の要件は以下の通りである。

- **集約的な表現であること。**PCP のように視覚的混雑度が高くて視認しづらい表現を用いると、表現間の対応付けを把握することが難しくなり、結果として表現間の比較が困難になる。そのため、データを比較する際に用いる表現は、視覚的混雑度が低くて視認しやすい表現が必要である。集約的な表現手法は、視覚的混雑度を抑えながら多数のレコードを表現できるという利点がある。以上より、多変量データセット内の一つの変量における分布などの情報を一つの集約的な表現によって可視化できることが望ましい。

- **視覚的な比較が容易であること。**VisBricks [3] を始めとした複数のチャートを並べる表現を用いて比較を行う場合、チャート間の距離が離れているため視覚的な比較は難しい。量的な値を比較する際は、表現の位置や長さによる比較が有効である [36]。変量間の比較を視覚的に支援するためには、位置や長さを用いて変量間を視覚的に比較できるような座標系、およびその座標系において利用可能な表現を採用する必要がある。

- **複数の部分集合の情報を表現でき、かつそれぞれの情報を読み取れること。**部分集合の情報を独立して読み取ることができなければ、データ全体としての傾向を把握できない。その結果、3.2 節にて述べた理想的なプロセス [34] に沿った視覚的分析ができなくなる。複数の部分集合を比較するためにも、まずは比較対象となる部分集合の情報をそれぞれ読み取れる必要がある。

- **部分集合間の差異を読み取れること。**複数の部分集合を別々の座標軸によって可視化した場合、比較対象の距離が離れてしまうため、目で見て比較することが困難になる。VAICo [30] のように、一つの表現上に差異をハイライト表示することにより、分析時に直感的かつ迅速に差異を確認できる。部分集合間の比較を行うためには、部分集合間の差異を顕在化できる表現手法を用いることが有効である。

本論文では、上記の要件をすべて満たすような表現手法を開発する。

4.2 視覚的表現の開発

本論文では、一つの変量の分布を表現する手法である Braided Graph [37] を拡張した、データ分布の比較に特化した表現手法である Blade Graph を開発する。Braided Graph は、複数の部分集合における 1 変量内の値の分布をそれぞれ一つの面グラフによって表現した上で、各面グラフを同一座標軸上に重畳表現する手法である。Braided Graph は複数の部分集合の分布を同一座標軸上で可視化することにより、各データセットの分布の差異を確認できるという利点がある。一方で、Braided Graph の色付けは誤解を与えやすく、また複雑な模様から情報を読み取る必要があるため、各部分集合の分布の把握や分布間の差異の発見が難しい。そこで本論文では、上記の Braided Graph の欠点を解消するような、量的な 1 変量の分布を表現する手法である Blade Graph を開発する。最終的には Blade Graph を並行座標軸上に配置することにより、多変量データセットの量的変量を表現する。表現を並行に並べることにより、高さや長さを用いて変量間を視覚的に比較できる。

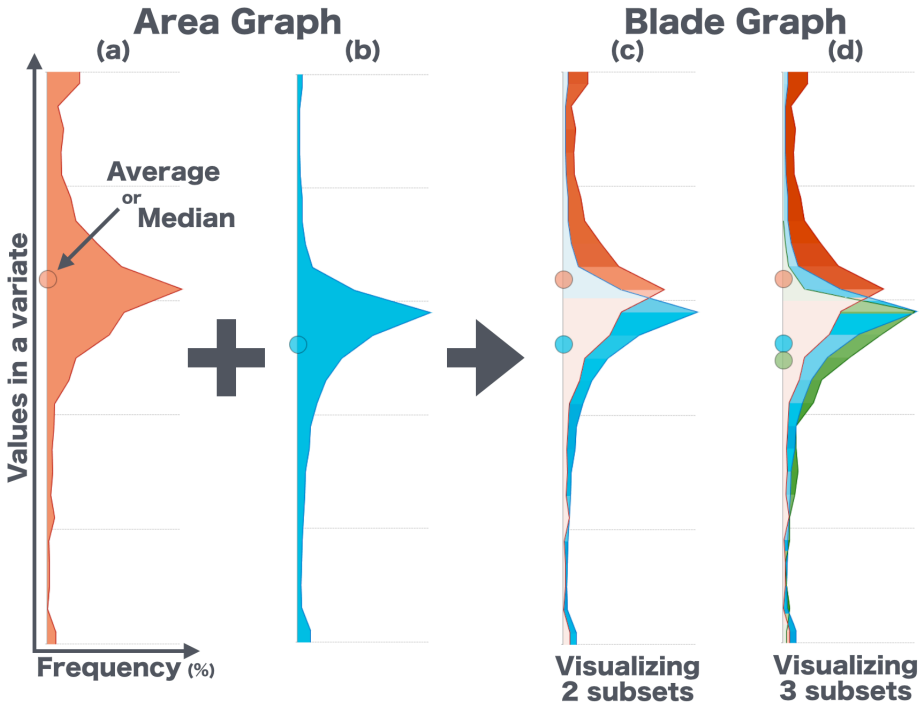

図 4.1: Blade Graph の生成例。

4.2.1 Blade Graph の特徴

本論文にて開発する Blade Graph では差の大きい部分が目立つような色付けを行う。これにより、データ分布の比較および特徴的部位の発見を支援する。Blade Graph は以下の点において Braided Graph を拡張している。

1. 分布を読取る際の参考値として、平均値または中央値を重畳表示する。
2. 異なるレコード数の部分集合間の分布を比較できるよう、割合を用いて高さを正規化する。
3. $L^*a^*b^*$ 色空間を用いることにより、知覚的均等性を考慮した色付けを行う。
4. 部分集合間での差異が大きい部分ほど強調されるような色付けを行う。

図 4.1 に Blade Graph の生成例を示す。図 4.1 中の (a) や (b) のように、まず一つの面グラフにより一つの部分集合における量的な 1 変量のデータ分布を表現する。(c) は Blade Graph の一例であり、(a) と (b) を重畳表示したものである。また (d) は三つの部分集合を同時に表現している Blade Graph の一例である。Blade Graph では、部分集合ごとに異なる色相を割り当てることにより、それぞれの部分集合の分布を区別可能にしている。さらに分布を読取る際の参考値として、平均値または中央値を表示している。なお、Blade Graph は順序尺度の変量についても表現可能な手法である。ただし順序尺度の変量を Blade Graph により表現する場合、参考値は中央値のみを採用した上で、縦軸の意味については読み取り時に十分考慮する必要がある。

以降の 4.2.2 節では、拡張事項の 2 を踏まえた Blade Graph を構成する面グラフの生成アルゴリズムについて述べる。4.2.3 節では、拡張事項 3 の $L^*a^*b^*$ 色空間について述べる。また 4.2.4 節では、拡張事項の 4 を踏まえた Blade Graph の着色アルゴリズムについて述べる。

4.2.2 Blade Graph を構成する面グラフの生成アルゴリズム

Blade Graph を構成する面グラフは、図 4.1 中の (a) や (b) のように、分布を表現するヒストグラムの各ビンの頂点を結んだ折れ線グラフにより生成する。n 変量 m レコードから成る多変量データセットの全体集合を U、第 p レコードにおける第 q 変量の値を v_{pq}、第 p レコードを $r_p = (v_{p1}, v_{p2}, ..., v_{pn}) \in U$ とする。このとき、l 個の部分集合 $G_1, G_2, ..., G_l \subseteq U$ における量的な第 d 変量を Blade Graph によって可視化することを考える。第 x 部分集合 G_x（$1 \leq x \leq l$）における第 d 変量内を k （$k \geq 1$）個のビンに分割する場合、i（$1 \leq i \leq k$）番目のビンの高

さ $m_{G_x,d}(i)$ は（4.1）式から（4.3）式により与えられる。

$$h = \frac{max\{v_{pd}|1 \leq p \leq m\} - min\{v_{pd}|1 \leq p \leq m\}}{k} \tag{4.1}$$

$$\delta_{pi} = \begin{cases} 1 & (r_p \in G_x \wedge i = 1 \wedge 0 \leq v_{pd} \leq h) \\ 1 & (r_p \in G_x \wedge i \geq 2 \wedge h(i-1) < v_{pd} \leq hi) \\ 0 & (\text{otherwise}) \end{cases} \tag{4.2}$$

$$m_{G_x,d}(i) = \frac{\sum_{p=1}^{m} \delta_{pi}}{|G_x|} \tag{4.3}$$

なお、上記式中の $|G_x|$ は部分集合 G_x の要素数とする。$m_{G_x,d}(i)$ はそのビンに対応した領域に含まれる値の相対頻度であり、$\sum_{i=1}^{k} m_{G_x,d}(i) = 1$ が成立する。

次に、各部分集合を表現する面グラフを同一座標軸内に重畳するため、ビンの高さを正規化する。ビンの高さを正規化することにより、データサイズの異なる部分集合間のデータ分布を比較できるようになる。分析の目的や状況に応じて、以下のいずれかの正規化処理を行う。以下、数式中に登場する記号は前段落と同様の意味とし、正規化後の $m_{G_x,d}(i)$ を $\acute{m}_{G_x,d}(i)$ と表記する。

1. **各変量内での正規化：** $\acute{m}_{G_x,d}(i) = \dfrac{m_{G_x,d}(i)}{\max\limits_{y,i} m_{G_y,d}(i)}$

   それぞれの変量におけるビンの高さの最大値を用いてビンの高さを正規化する。変量に着目したデータ分布の比較を行う際に有効な正規化処理である。

2. **全変量間での正規化：** $\acute{m}_{G_x,d}(i) = \dfrac{m_{G_x,d}(i)}{\max\limits_{y,q,i} m_{G_y,q}(i)}$

   全ての変量におけるビンの高さの最大値を用いてビンの高さを正規化する。複数の変量間でのデータ分布を比較する際に有効な正規化処理である。

3. **部分集合のサイズを考慮した全変量間での正規化：** $\acute{m}_{G_x,d}(i) = \dfrac{|G_x| m_{G_x,d}(i)}{\max\limits_{y,q,i}\{|G_y| m_{G_y,q}(i)\}}$

   全変量間での正規化の際に、各部分集合のデータサイズに応じた重み付けを行った上でビンの高さを正規化する。値の頻度の割合ではなく、値の頻度そのものを用いた正規化を行っているため、部分集合のデータサイズを考慮した比較を行う際に有効な正規化処理である。

4.2.3 $L^*a^*b^*$ 色空間の利用

Blade Graph の色付けには、知覚的均等性を重視した $L^*a^*b^*$ 色空間 [1]を利用する。Blade Graph では差異が大きな部分を強調する都合上、色相の違いによる見え方への影響を抑える必要がある。例えば HSB 色空間を用いた場合、同じ明度・彩度であっても、黄色は紫色よりも明るく見えてしまう。そこで $L^*a^*b^*$ 色空間を用いることにより、知覚的均等性を考慮した色付けを行う。

$L^*a^*b^*$ 色空間では、空間内のユークリッド距離が色差の知覚量を表している。L^* は色の明度に対応する座標である。L^* 座標の範囲は $[0, 100]$ であり、$L^* = 0$ は黒、$L^* = 100$ は白となる。a^* および b^* は色味に対応する座標であり、範囲は L^* の値に応じて決定する。a^* は正が赤色、負が緑色の色味に対応している。b^* は正が黄色、負が青色の色味に対応している。また、$\theta_{a^*,b^*} = \arctan(b^*/a^*)$ を色相角と呼ぶ。

$L^*a^*b^*$ 色空間を実際に描画する際は、まず $L^*a^*b^*$ 色空間を XYZ 色空間に変換した後、さらにそれを sRGB 色空間などに変換する。$L^*a^*b^*$ 色空間から XYZ 色空間への変換は、以下の（4.4）式から（4.7）式により行われる [38]。

$$X = X_n f^{-1}(\frac{L^* + 16}{116} + \frac{a^*}{500}) \tag{4.4}$$

$$Y = Y_n f^{-1}(\frac{L^* + 16}{116}) \tag{4.5}$$

$$Z = Z_n f^{-1}(\frac{L^* + 16}{116} - \frac{b^*}{200}) \tag{4.6}$$

$$f^{-1}(t) = \begin{cases} t^3 & (t > \frac{6}{29}) \\ 3(\frac{6}{29})^2(t - \frac{4}{29}) & (\text{otherwise}) \end{cases} \tag{4.7}$$

なお、上記式中の X_n、Y_n および Z_n は基準となる白色点（ホワイトポイント）の XYZ 色空間での値である。例えば D65 光源 [2]であれば、$\{X_n, Y_n, Z_n\} = \{95.047, 100.0, 108.883\}$ が採用される [39]。

次に、XYZ 色空間から sRGB 色空間への変換を行う。変換は以下の（4.8）式から（4.13）式により行われる [40]。

[1]JIS Z 8781-4:2013 測色－第 4 部：CIE 1976 L ＊ a ＊ b ＊色空間, http://kikakurui.com/z8/Z8781-4-2013-01.html, 閲覧:2015.01.04

[2]D65 光源: 国際照明委員会が規定している標準光源の代用となる光源の規格の一つ。

$$M^{-1} = \frac{1}{100}\begin{pmatrix} 3.2406 & -1.5372 & -0.4986 \\ -0.9689 & 1.8758 & 0.0415 \\ 0.0557 & -0.2040 & 1.0570 \end{pmatrix} \tag{4.8}$$

$$\begin{pmatrix} f_R \\ f_G \\ f_B \end{pmatrix} = M^{-1}\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \tag{4.9}$$

$$\grave{\delta} = 0.0031308 \tag{4.10}$$

$$R = \begin{cases} 255 \times (1.055 f_R^{1/2.4} - 0.055) & (f_R > \grave{\delta}) \\ 255 \times 12.92 f_R & (\text{otherwise}) \end{cases} \tag{4.11}$$

$$G = \begin{cases} 255 \times (1.055 f_G^{1/2.4} - 0.055) & (f_G > \grave{\delta}) \\ 255 \times 12.92 f_G & (\text{otherwise}) \end{cases} \tag{4.12}$$

$$B = \begin{cases} 255 \times (1.055 f_B^{1/2.4} - 0.055) & (f_B > \grave{\delta}) \\ 255 \times 12.92 f_B & (\text{otherwise}) \end{cases} \tag{4.13}$$

4.2.4 色付けによる差異の顕在化

Blade Graph では、複数の面グラフに異なる色相を割り当てた上で、面グラフに囲まれた領域の色を塗り分ける。Blade Graph の描画の際は、Braided Graph と同様、ビンごとに部分集合の描画順序を決める。ビンの高さが高い順に部分集合のビンの描画を行うことにより、ビンの高さに依らず全ての部分集合のデータ分布が閲覧できるようになる。その際、部分集合間での差異が大きい部分ほど強調されるような色付けを行うことにより、データ分布の比較を支援する。

Blade Graph では、$L^*a^*b^*$ 色空間における L^* を用いて差異の強調を行う。また部分集合ごとに θ_{a^*,b^*} を割り当てることにより、部分集合の区別を行う。部分集合 G_x における変量 d 内のビン i の色について、L^* は以下の（4.14）式により与えられる。なお特筆がない場合、本節の数式中に登場する記号は 4.2.2 節にて定義した記号と同様のものとする。

$$L^* = 95 - 45 \times \frac{\acute{m}_{G_x,d}(i) - \min_y \acute{m}_{G_y,d}(i)}{\max_y \acute{m}_{G_y,d}(i)} \tag{4.14}$$

求めた L^* に対して a^* 座標および b^* 座標の範囲を算出し、範囲内における中心からの最大半径 $\lfloor r \rfloor$ を決定する。$\lfloor r \rfloor$ は L^* と比例関係にあり、例えば $L^* = 50$ では $\lfloor r \rfloor = 79$、$L^* = 95$ で

は $\lfloor r \rfloor = 154$ の値を取る。a^* および b^* は以下の式により与えられる。

$$a^* = \lfloor r \rfloor \cos \theta_{a^*,b^*} \tag{4.15}$$

$$b^* = \lfloor r \rfloor \sin \theta_{a^*,b^*} \tag{4.16}$$

色の明度を決定付ける L^* の値は、そのビンにおける高さの最大値と最小値に対する $\acute{m}_{G_x,d}(i)$ の大きさに依存する。ビン内での最小値に近い部分集合ほど L^* が大きくなり、結果として高明度・低彩度の目立ちにくい色で該当領域が塗りつぶされる。逆に、最小値との差が大きい部分集合の領域は L^* が小さくなり、結果として低明度・高彩度の目立つ色で領域が塗りつぶされる。

第5章　分析ツールの開発

開発した表現手法を用いて、多変量データセットに含まれる特徴的部分集合の探索および比較が可能な分析ツールを開発する。分析ツールの主な利用者としては、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。

5.1　実装言語

本ツールの実装言語には Java（Java™ Platform Standard Edition 7.0）及び Processing[1]を使用した。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。本ツールが読み込む多変量データセットは tsv 形式のものを対象としている。なお、ツールの実装および実行環境として利用した計算機は MacBook Pro（Retina, 15-inch, Late 2013）および MacBook Air（11-inch, Mid 2012）であり、いずれも Mac OS X Yosemite（version 10.10.1）を搭載している。

5.2　分析ツールの開発

開発した分析ツールのスクリーンショットを図 5.1 に示す。本節では、分析ツールの設計、ツールを構成するパネル、および分析のためのインタラクションについて述べる。

[1]Processing, http://processing.org/, 閲覧:2015.01.04

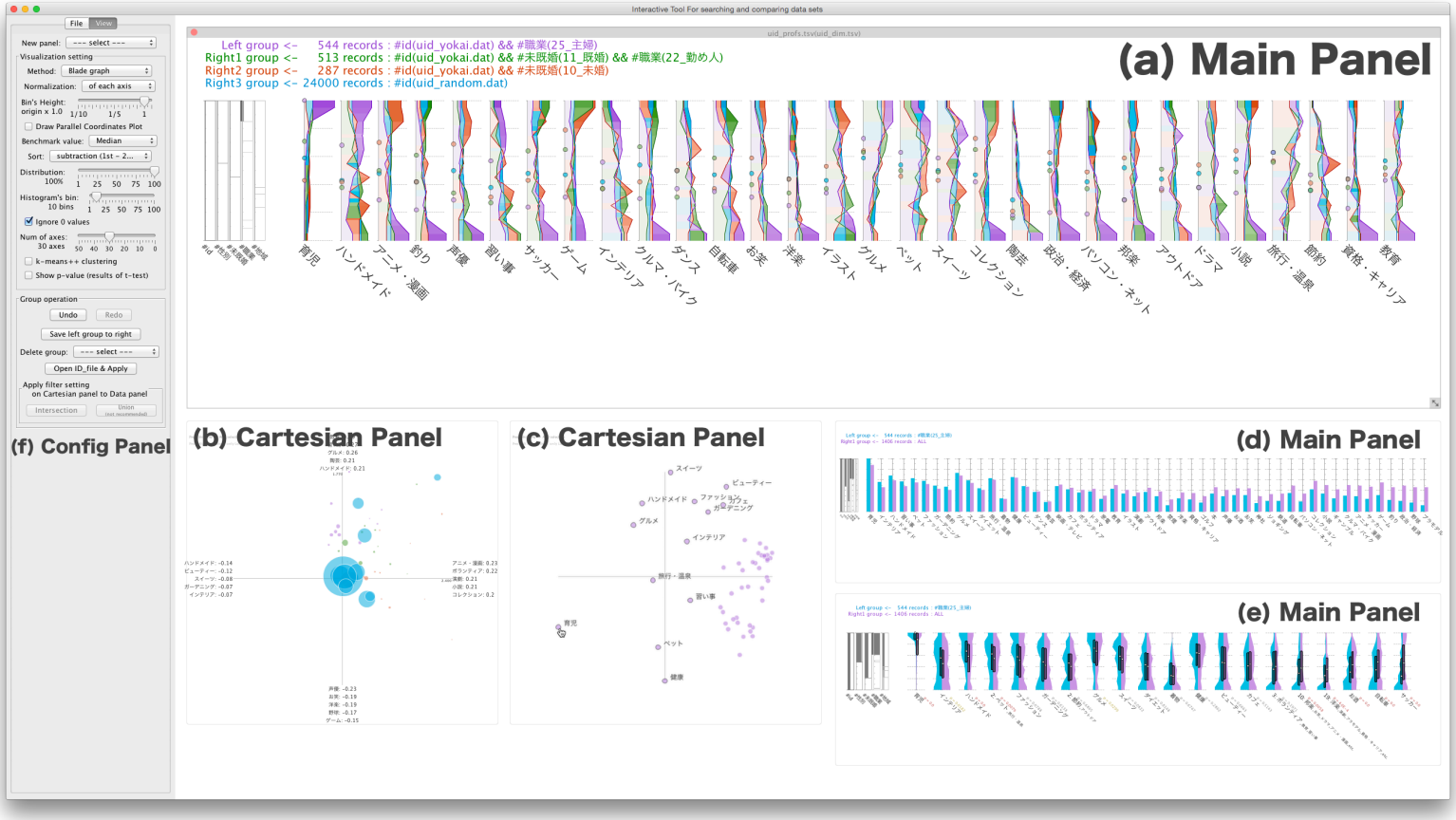

図 5.1: 開発した分析ツールのスクリーンショット。

5.2.1 分析ツールの設計

本論文において開発する分析ツールでは、質的な変量と量的な変量のそれぞれに適した表現手法によって各変量を表現する。また、PCP [12] のような並行座標軸を用いた表現手法は、座標軸が横並びに揃っているため変量間の比較が視覚的に容易である。さらに、多変量データセットにおける複数の部分集合を同一座標軸上に表示することによって、複数のデータセットを同時に探索でき [41]、かつそれらの比較が容易になる。そこで本分析ツールには、並行座標軸上に複数の部分集合を表現するデザインを採用した。

また本分析ツールでは、並行座標軸をベースとした表現に加え、直交座標軸を用いた表現を補助的に利用する。Siirtola は、複数のビューをインタラクティブに結びつけた可視化手法の有効性を実証した [22]。そこで本分析ツールにおいても、複数の表現からのインタラクティブなデータ操作を可能にする。

本分析ツールは、2 種類のパネルを組み合わせることによって、多変量データセットの比較分析を支援する。まず Main Panel は、多変量データセットの概観や各変量における分布を並行座標軸を用いて可視化するためのパネルである。Cartesian Panel は、Main Panel では表現しきれない情報を表現するために設計されたパネルであり、直交座標軸上にレコード間や量的変量間の関係性を可視化する。ツールの利用者は簡単にパネルの追加または削除ができ、かつパネルの配置や大きさも自由に変更できるようにする。これにより、分析目的に応じた柔軟なデータ探索および比較が可能である。さらに本分析ツールには、比較分析に必要なインタラクションを備える。

以降の 5.2.2 節では Main Panel について、5.2.3 節では Cartesian Panel について、5.2.4 節ではツールのインタラクションについて、それぞれ詳細に説明する。

5.2.2 Main Panel

Main Panel では、並行座標軸を用いて多変量データセットの概観や各変量におけるデータ分布を可視化する。一つの座標軸に一つの変量が割り当てられており、座標軸の下に対応する変量のラベルが記されている。図 5.1 中の (a)(d)(e) は、分析ツール上に配置された Main Panel の一例である。パネル左側の積み上げ棒グラフを用いて質的変量を可視化し、右側の並行座標軸に量的変量を可視化する。量的変量を表現する手法は、分析目的や用途に応じて切り替え可能である。またレコードとしての繋がりを表現するため、PCP を重畳表示できる。

量的変量の表現手法には、主に 4 章にて述べた Blade Graph を用いる。また Blade Graph に加えて、用途に応じて、棒グラフ、Box Plot [42] および Violin Plot [43] への切り替えが可能である。図 5.1 の (a) は Blade Graph、(d) は棒グラフ、(e) は Violin Plot を用いた Main Panel の一例である。Blade Graph は部分集合間の比較に特化した表現手法であり、色や形から部分集合のデータ分布を確認しつつ比較することができる。棒グラフは一つの値を表現する最も基本的な表現手法の一つであり、描画領域が狭い場合においても比較的読み取りやすいという利点がある。Box Plot は矩形および上下に伸びる直線を用いて 5 種類の要約統計量を表現する手法であり、視覚的混雑度を抑えながらデータ分布の読み取りに必要な情報を提供できる。

Violin Plot は Box Plot を拡張した手法であり、Box Plot と同様に 5 種類の要約統計量を表現しながら、より細かなデータ分布を表現できる利点がある。

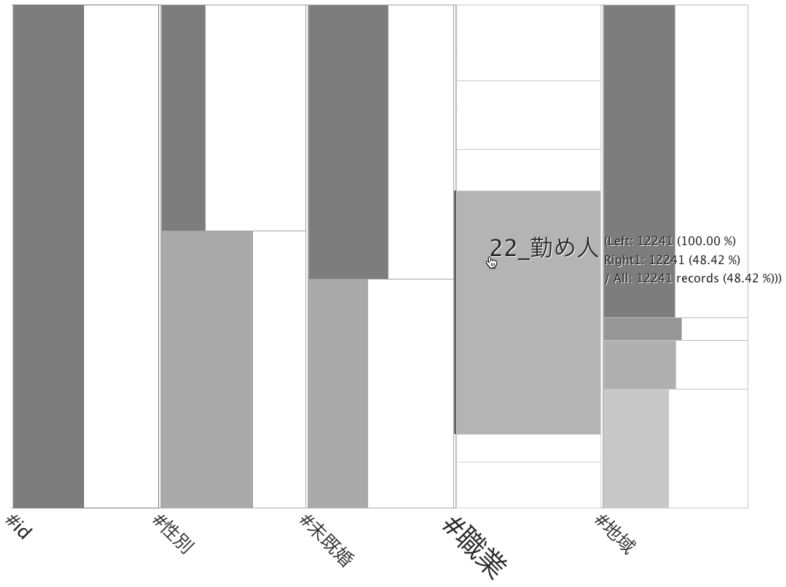

図 5.2: 横幅が変化する積み上げ棒グラフによる質的変量の表現。

Main Panel では、図 5.2 のように、一つの質的な変量をそれぞれ一つの積み上げ棒グラフにより表現する。例えば図 5.2 中では、“#id”、“#性別”、“#未既婚”、“#職業” および “#地域” の 5 つの質的変量を表現している。積み上げ棒グラフでは、変量中の一つのカテゴリが一つの矩形に割り当てられている。矩形の高さはカテゴリの比率に応じて決定しており、高さの比から変量内のカテゴリの構成比を読み取ることができる。矩形の横幅は、データセット全体に含まれるカテゴリ量に対する部分集合内のカテゴリ量に応じて決定する。矩形の横幅を比較することによって、各カテゴリの変化量を比較できる。積み上げ棒グラフでは無彩色のみの配色を採用している。これは Blade Graph において色相を部分集合の区別に用いる都合上、誤解を与えないようにするための配慮である。

本分析ツールでは、Main Panel とデータセットが紐付けられている。そのため、選択中の部分集合のみを母集団として新たな Main Panel を生成したり、全く別のデータセットを読み込んだ Main Panel を生成することができる。これらの機能により、画面上に複数のデータセットを表示して分析できる。また、データセットを絞り込んでいくようなドリルダウン式の分析が可能である。

5.2.3 Cartesian Panel

Cartesian Panel では、直交座標軸上にレコード間や量的変量間の関係性を可視化する。多変量の情報を 2 次元の座標軸上にて表現するための手法として、本ツールでは多次元尺度構成

法または主成分分析を用いる。なお Cartesian Panel は Main Panel と紐付けられており、紐付けられた Main Panel が可視化している多変量データセットを可視化の対象とする。例えば図 5.1 中の (b) および (c) は、分析ツール上に配置された Cartesian Panel の例であり、いずれも図 5.1(a) の Main Panel と紐付けられている。

多次元尺度構成法は、分類対象物間の関係性を低次元空間における点の位置関係により表現する手法である。多次元尺度構成法では、似たものは近くに、異なるものは遠くに配置されるように座標変換が行われる。多次元尺度構成法は量的変量間の類似性を把握する際の判断材料として有効である。一方で、座標軸に意味を持たないため、多次元尺度構成法における点の分布から意味を持った情報を読み取ることができない。

図 5.3 は多次元尺度構成法を用いて量的変量の類似性を可視化した一例である。点が密集している部分から離れた点は、他とは異なる傾向を示している可能性がある。またそれぞれ近くに配置されている点同士は、似た傾向を示す可能性がある。Cartesian Panel の多次元尺度構成法を用いることにより、これらの “傾向の可能性” の発見が期待できる。

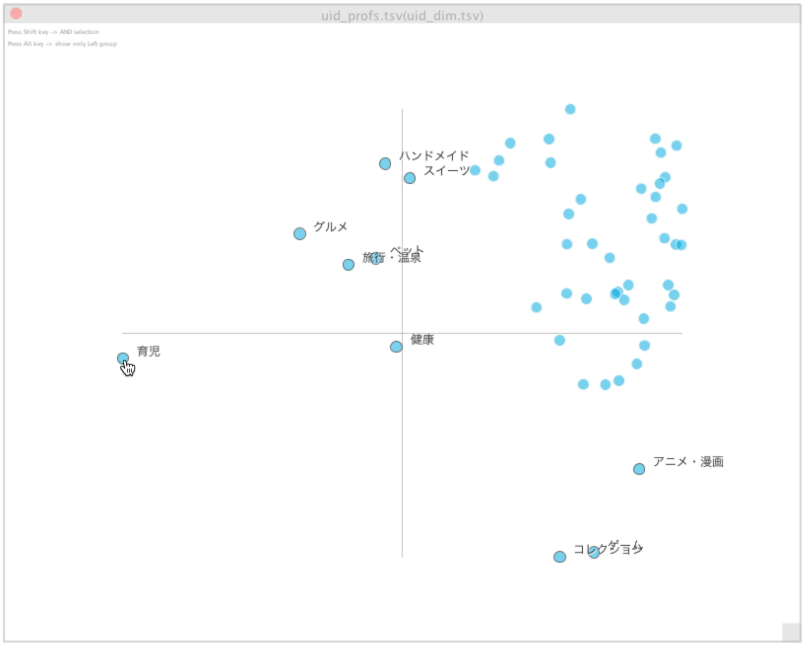

図 5.3: 多次元尺度構成法を用いて量的変量の類似性を可視化している Cartesian Panel の一例。

Cartesian Panel では、主成分分析の結果を表示することも可能である。主成分分析は多次元尺度構成法と同様、分類対象物間の関係性を低次元空間における点の位置により表現する手法である。主成分分析では、情報量（＝データセットの分散）が多くなるように合成変数（主成分）を生成した後、最も分散が大きくなる方向に軸を取って可視化する。主成分分析は多次元尺度構成法とは異なり、軸に意味を持たせた手法である。そのため、各点の位置からおおまかな意味を読み取ることが可能である。

レコードを主成分分析により可視化する際は、視覚的混雑を避けるため、事前に k-means++

法 [44] によるクラスタリングを行えるようにする。事前にレコードをクラスタリングすることにより、点の分布の把握や点同士の比較が容易になるという利点がある。k-means++法は k-means 法における初期クラスタ中心を決定するアルゴリズムに改良を加えた非階層クラスタリング手法である。重み付き確率分布を用いた初期クラスタ中心の選出アルゴリズムにより、初期クラスタ中心がなるべく離れるよう、かつ外れ値に影響されすぎないように設計されている。クラスタリングを行う際は、まず一つのレコードが持つ量的変量の値を一つの次元としたレコードベクトルを生成する。このレコードベクトルを k-means++法によってクラスタリングすることにより、類似した複数のレコードを 1 個のクラスタとして集約できる。

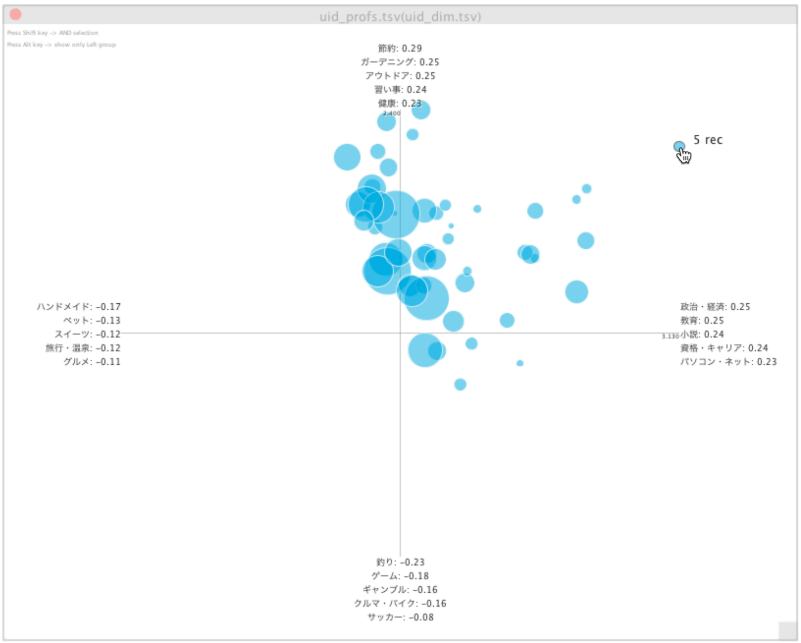

図 5.4: 主成分分析を用いてレコードと量的変量の関係性を可視化している Cartesian Panel の一例。一つの部分集合に絞って主成分分析を計算している。

図 5.4 は主成分分析を用いて部分集合に含まれるレコードを可視化した一例である。図 5.4 では分類対象となるレコードの数が多いため、事前にクラスタリングを行っている。一つの円（点）が一つのクラスタに対応しており、各円の面積はクラスタの大きさに比例している。また各軸には、軸の主成分に対応した個体値と、その主成分における主成分係数が正に大きい変量と負に大きい変量がそれぞれ表示されている。

図 5.5 は、二つの部分集合に含まれるレコードを同時に Cartesian Panel 上に可視化した一例である。円の色相は Main Panel において部分集合に割り当てられた色相と対応付けられている。各色の円の分布を見比べることにより、それぞれの部分集合における傾向を読み取ることができる。円の数の割合は各部分集合におけるレコード数に応じて決定できるようにする。例えば部分集合 G_a のレコード数を $|G_a|$、部分集合 G_b のレコード数を $|G_b|$ とした場合、

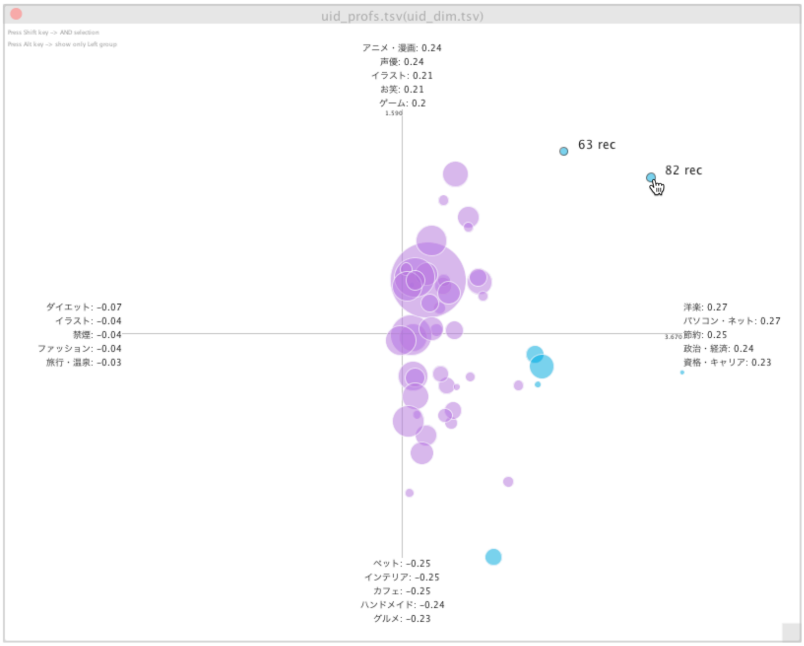

図 5.5: 主成分分析を用いてレコードと変量の関係性を可視化している Cartesian Panel の一例。複数の部分集合に対して主成分分析を計算している。

それぞれの部分集合に割り当てるクラスタ数 k_a および k_b は以下の式により与えられる。

$$\theta = 1 - \frac{2|G_a|}{|G_a| + |G_b|} \tag{5.1}$$

$$\acute{k_a} = \lfloor 0.5 + K\frac{\arccos\theta}{\pi} \rfloor \tag{5.2}$$

$$k_a = \begin{cases} 1 & (\acute{k_a} \leq 1) \\ K-1 & (\acute{k_a} \geq K) \\ \acute{k_a} & (\text{otherwise}) \end{cases} \tag{5.3}$$

$$k_b = K - k_a \tag{5.4}$$

なお K はクラスタの総数、すなわち Cartesian Panel に表示する円の数である。レコード数との単純な正比例ではなく上記のような補正をかける式を採用することにより、同時に表示したい部分集合のレコード数の間に大きな差があった場合でも、各部分集合に割り当てる円の数をある程度確保できるようになる。

なお、Cartesian Panel では視覚的混雑を避けるため、通常状態において各円のラベルは表示しない。マウス操作により円を選択した場合に限り、その円のラベルを表示する。例えば、図 5.3、図 5.4、および図 5.5 では、縁が黒い円が選択された状態になっている。視覚的混雑を

避けることにより、円の数が多い場合においてもそれぞれの円を視認できる。

5.2.4 分析のためのインタラクション

本分析ツールは、マウス操作により多変量データセットの一部を抽出する機能を備える。表示されている積み上げ棒グラフや Blade Graph、Cartesian Panel 上の円などをクリックまたはドラッグすることによって、該当部分の値を持つデータセットを抽出できる。また抽出したい ID の一覧が記載された外部ファイルを読み込むことにより、該当 ID を持つレコードのみで構成される部分集合を抽出できる。抽出したデータの部分集合は各表現に反映され、またその部分集合を保存することも可能である。これらの操作により、任意のデータセットを比較することができる。

量的な変量の座標軸におけるソート機能とフィルタリング機能について述べる。特定の条件に基づいて座標軸をソートすることにより、特徴的な変量の軸が特定の位置に集中するため特徴を発見しやすくなる。座標軸のソート条件としては、各量的変量における平均値の降順、部分集合間の差の降順、t 検定の p 値の昇順などを備えており、分析者が目的に応じて切り替え可能にする。また、データセットの変量数が非常に多くなると、画面領域上に全ての情報を可視化できない場合がある。そこで本ツールでは、変量数が非常に多いデータセットを効率的に分析するため、また可視化結果を効果的に提示するために、画面上に表示する変量数を変更可能にするフィルタリング機能を備える。ソート機能とフィルタリング機能を併用することにより、データセットの特徴的な部分を効率的に閲覧することができる。

分析的推論を支援するツール機能として、k-means++法による量的変量のクラスタリングを備える。まず各量的変量の最大値と最小値に基づいて値を正規化した上で、それぞれの量的変量を 1 本の変量ベクトルとみなす。例えば M 列の変量、N 行のレコードから成る多変量データセットの場合、N 次元の変量ベクトルが M 本生成される。この変量ベクトルを k-means++ 法によってクラスタリングすることにより、類似した複数の変量を 1 個の変量クラスタとして集約できる。本機能は複数の類似した変量を 1 本の座標軸に集約して表現するため、情報量を保ちながら特徴的な情報を表現できるとともに、類似した変量の探索が容易になるという利点がある。

また分析的推論を支援するための別機能として、t 検定によってデータセット間の有意差の有無を計算し、その結果を可視化結果に反映させる機能を備えている。まず変量ごとの平均値が最大の部分集合に対して、他の部分集合とそれぞれ t 検定を行う。その結果に応じて、各変量のラベル上部に t 検定の p 値を表示する。このとき、有意水準 1%において全て有意差があれば赤色、有意水準 5%において全て有意差があれば黄色、それ以外は灰色によって p 値が表示される。これにより、データセット間の違いを視覚的表現から定量的に判断することができる。

図 5.1(f) の Config Panel から実行可能な機能は主に以下の通りである。

- 新規データファイルの読み込み。
- 新規パネルの生成。

- 選択中の Main Panel に関する設定。量的変量に用いる可視化手法の変更、Blade Graph に用いる正規化処理の変更、ビンの高さの調節スライダ、PCP の有無、表現やソートに利用する基準値の選択（平均値または中央値）、軸のソート手法の選択および実行、分布の表示領域に関するフィルタリング設定スライダおよびチェックボックス、表示する変量数の調節スライダ、クラスタリングの有無、および t 検定の結果表示の有無。
- フィルタリングに関する Undo および Redo 操作。
- 選択中の部分集合の保存。
- 保存中の部分集合の削除。
- 外部 ID ファイルの読み込み。
- Cartesian Panel でのレコード選択状態の Main Panel への反映。

第6章　ケーススタディ

開発した分析ツールを使用して実際のデータを分析するケーススタディとして、ソーシャルメディアの一つであるブログの解析データから特定のユーザ集団に関する分析を行う。

ソーシャルメディア上では多種多様な属性のユーザが記事を書いており、その記事には各自の嗜好に関する内容が含まれている。また、ユーザは他のユーザが書いた記事を読むことができ、必要に応じてコメント等のフィードバックを与えることができる。このようなコミュニティが積み重なることによってソーシャルメディアは成り立っている。

ソーシャルメディアを解析して得たデータからは、“各属性のユーザが何に対して高い関心を持っているのか” という集団的傾向を知ることができる。ここから得られる知見は、マーケティング分野への幅広い応用が期待できる。例えば、ある商品を売りたいとき、その商品がどのような客層に好まれているかを知ることにより、顧客層を絞った効果的な広告を打つことができる。また逆に、客層毎の嗜好を把握することにより、顧客層が好みそうな商品やサービスを開発・提供することができる。このように、ソーシャルメディアから得られる集団的傾向は、利益を増やす上で重要な要素となり得るだろう。

6.1　対象データセット

本ケーススタディにて扱う多変量データセットは、ブログユーザに関する情報を記したものである[1]。ここで言うユーザの情報とは、ユーザのデモグラフィックな属性（ユーザ属性：表 6.1 を参照）、および 51 種類の各趣味に対するユーザのスコア（趣味スコア）である。趣味スコアは、“それぞれの趣味に関連する記事をどの程度書いているか” をユーザごと・趣味ごとに数値化したものである。各趣味について、関連する記事を多く書いているほど、その趣味のスコアが高くなるように集計されている。趣味スコアの値域は [0,1] の実数値であり、その趣味に対して全く記事を書いていない場合はスコアが 0 になっている。

データは tsv 形式で保存されており、各行（レコード）が各ブログユーザ、各列が各ユーザ属性または趣味スコアに対応している。今回は 25,280 レコード、つまり 25,280 人分のブログユーザのデータセットを扱うことを考える。このデータは 5 種類のユーザ属性と 51 種類の趣味から構成される、56 変量の多変量データセットである。

また上記の多変量データセットに加え、分析対象とする集団の ID 一覧が記載された外部ファイルを扱う。今回外部ファイルとして、以下の外部 ID ファイルを用いる。

[1] ブログユーザに関する情報を記した多変量データセットは、株式会社富士通研究所から提供を受けたものである。

表 6.1: 対象データセットにおけるユーザ属性のカテゴリ一覧

属性名	カテゴリ名	ユニーク数
ユーザ ID	数値	25,280
性別	男性、女性	2
結婚歴	未婚、既婚	2
職業	中高生、大学生、勤め人、シニア、その他、主婦	6
居住地	東京、大阪、愛知、不明	4

- “uid_random.dat”：ブログからランダムサンプリングしたユーザ 24,000 人の ID リスト。以降、このリストに含まれる ID の集団を一般集団と表記する。
- “uid_yokai.dat”：ブログにおいて “妖怪ウォッチ” に言及しているユーザ 1,406 人の ID リスト。以降、このリストに含まれる ID の集団を妖怪集団と表記する。

本ケーススタディの目的は、この妖怪集団に関する集団的傾向の知見を得ることである。

6.2 観察

まず、変量的な観点から妖怪集団および一般集団の概観を得るため、Blade Graph を用いた Main Panel を確認した（図 6.1 を参照）。図 6.1 ではスコアが 0 のユーザの分布を不可視状態にした上で、妖怪集団（橙色）と一般集団（水色）の中央値の差分が大きい上位 15 趣味を表示している。図 6.1 において特に左側に配置されている育児やハンドメイド、プラモデル、コレクションなどは、一般集団と比べて高スコアのユーザが多い趣味である。各趣味の Blade Graph を確認すると、育児、ハンドメイド、コレクション、ゲームにおいて、橙色の妖怪集団が高スコア部分において突出していた。一方でアニメ漫画は、両集団の中央値は殆ど変わらないものの、低スコア部分に妖怪集団が集中していた。

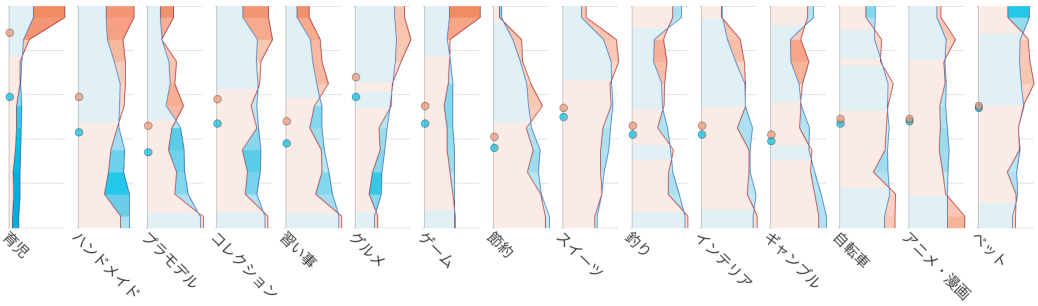

図 6.1: 妖怪集団（橙色）および一般集団（水色）を Blade Graph により表現した Main Panel。

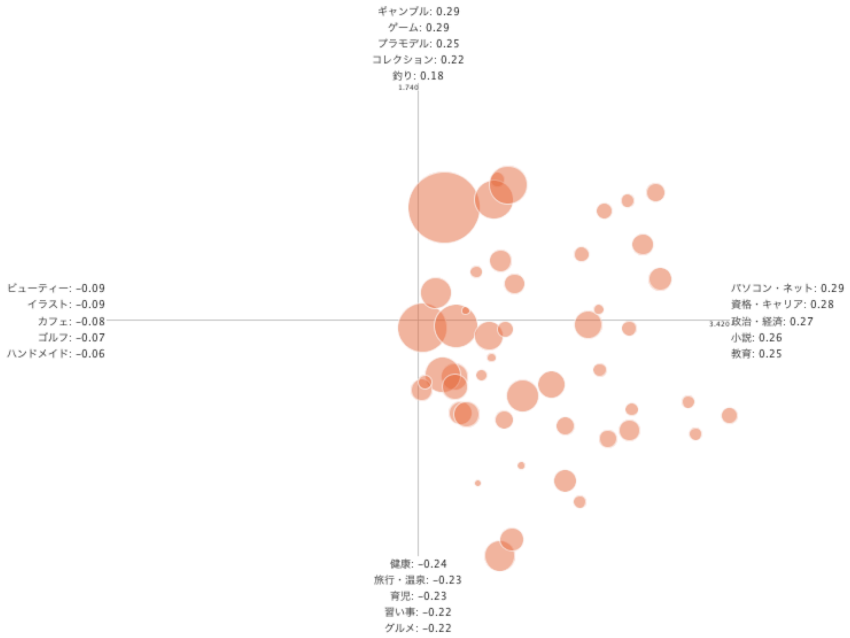

図 6.2: 妖怪集団のレコードを主成分分析を用いて表現した Cartesian Panel。

次に、レコード的な観点から妖怪集団の傾向を調査するため、妖怪集団のレコードに対して主成分分析を行った。図 6.2 は、妖怪集団のレコードを主成分分析を用いて表現した Cartesian Panel である。妖怪集団のクラスタは主に第一・第四象限に分布していた。横軸に対応する第一主成分に対して正方向に影響が大きな趣味は、パソコンネット、資格キャリア、政治経済、小説、教育である。縦軸に対応する第二主成分に対して正方向に影響が大きな趣味は、ギャンブル、ゲーム、プラモデル、コレクション、釣りである。第二主成分に対して負の方向に影響が大きな趣味は、健康、旅行温泉、育児、習い事、グルメ である。なお下線付きの趣味は、先ほどの Main Panel（図 6.1 を参照）において妖怪集団が一般集団よりも高い中央値を示していた趣味である。

さらに、妖怪集団に含まれるユーザの属性分布を Main Panel 中の積み上げ棒グラフより確認した（図 6.3 を参照）。妖怪集団と一般集団の分布を比較したところ、性別についてはほぼ差が見られなかった。結婚歴は未婚（20.41%、287/1,406）に対して既婚（79.59%、1,119/1,406）の方が包含率が高かった。職業は勤め人（45.87%、645/1,406）が最も多く、次点で主婦（38.69%、544/1,406）であった。逆に、妖怪集団に中高生（1.71%、24/1,406）や大学生（0.92%、13/1,406）は殆ど含まれていなかった。勤め人内での妖怪集団の含有率は、既婚が 79.53%（513/645）、未婚が 20.47%（132/645）であり、勤め人の約 8 割が既婚者であった。また、妖怪集団における地域差は殆ど見受けられなかった。以上より、妖怪集団の中でも、「主婦」「既婚勤め人」「未婚」の 3 集団が多数を占めていることがわかった。そこでこの 3 集団に着目して、集団のより詳細な傾向の分析を行った。以降、“妖怪ウォッチ” に言及している主婦の集団を妖怪主婦集団、“妖怪ウォッチ” に言及している既婚かつ勤め人の集団を妖怪既婚勤め人集団、“妖怪

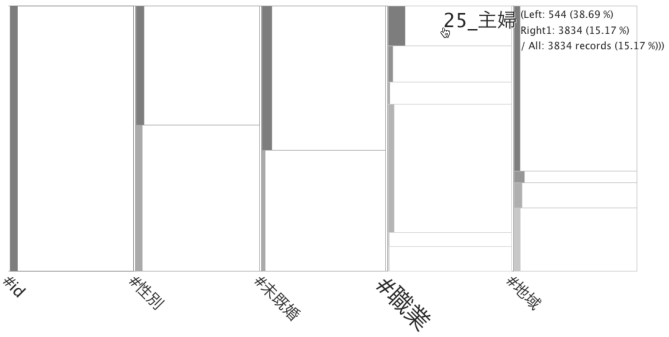

図 6.3: 妖怪集団を選択した状態の Main Panel 中の積み上げ棒グラフ。

ウォッチ” に言及している未婚の集団を妖怪未婚集団とそれぞれ略記する。

6.2.1 妖怪主婦集団、妖怪既婚勤め人集団および妖怪未婚集団の比較

まずは妖怪主婦集団（544 レコード）、妖怪既婚勤め人集団（513 レコード）、妖怪未婚集団（287 レコード）および一般集団（24,000 レコード）の間において比較を行った。図 6.4 は、妖怪主婦集団（紫色）、妖怪既婚勤め人集団（緑色）、妖怪未婚集団（橙色）および一般集団（水色）を Blade Graph により表現した Main Panel である。図 6.4 中の Blade Graph では、スコアが 0 のユーザの分布を不可視状態にしている。量的変量における座標軸は、表示中の部分集合において最大の中央値と 2 番目に大きな中央値の差の大きい順とした。この並び順では、突出した中央値を持つ部分集合があるような変量ほど左側に配置される。このとき、それぞれの趣味の Blade Graph における色の濃い部分を確認することにより、他の集団とは異なる特徴を示している部分を発見できた。例えば、育児やハンドメイドは紫色の妖怪主婦集団が最も高い中央値を示しており、また他の集団とも異なる分布を示していることが育児の

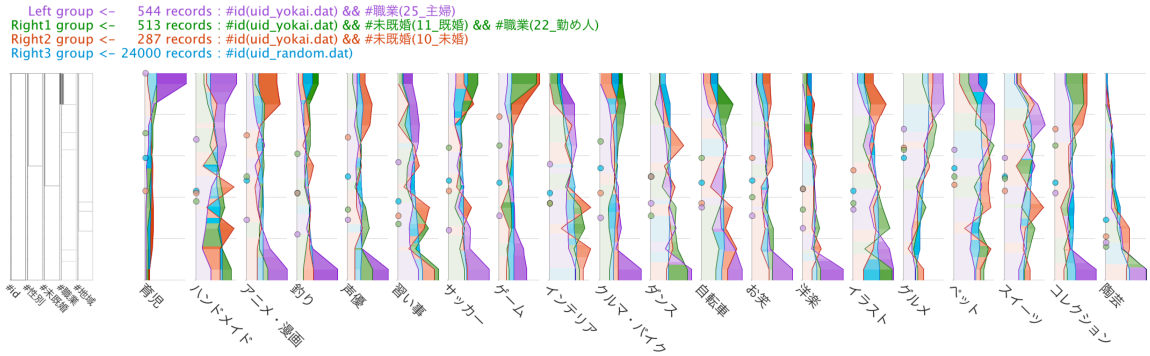

図 6.4: 妖怪主婦集団（紫色）、妖怪既婚勤め人集団（緑色）、妖怪未婚集団（橙色）および一般集団（水色）を Blade Graph により表現した Main Panel。

Blade Graph から読み取れた。一方で、多くの趣味において紫色は下の方に突出して分布していたことから、妖怪主婦集団は他の集団とは全体的に異なる嗜好を持っていると推測できる。緑色の妖怪既婚勤め人集団は、ゲームやコレクション、自転車に対して比較的高いスコアを持つユーザが多く見られた。橙色の妖怪未婚集団は、アニメ漫画、イラストに対して高い関心を示すユーザを多数確認できた。また緑色の妖怪既婚勤め人集団と同様、ゲームやコレクションについても高いスコア分布を示していることが読み取れた。

次に、これらの 4 集団のレコードをクラスタリングした上で主成分分析の結果を Cartesian Panel 上に可視化した（図 6.5 を参照）。第一主成分に対して正方向に影響が大きな変量は、アニメ漫画、ボランティア、演劇、小説、コレクションである。第二主成分に対して正方向に影響が大きな変量は、教育、習い事、グルメ、陶芸、ハンドメイドである。青色の一般集団は、中央付近に寄っているクラスタと第一象限に配置されているクラスタが存在していた。紫色の妖怪主婦集団は第二主成分である縦軸の正方向に多く配置されていることが確認できた。緑色の妖怪既婚勤め人集団および橙色の妖怪未婚集団は中央付近の第一象限に多く分布していることが確認できた。

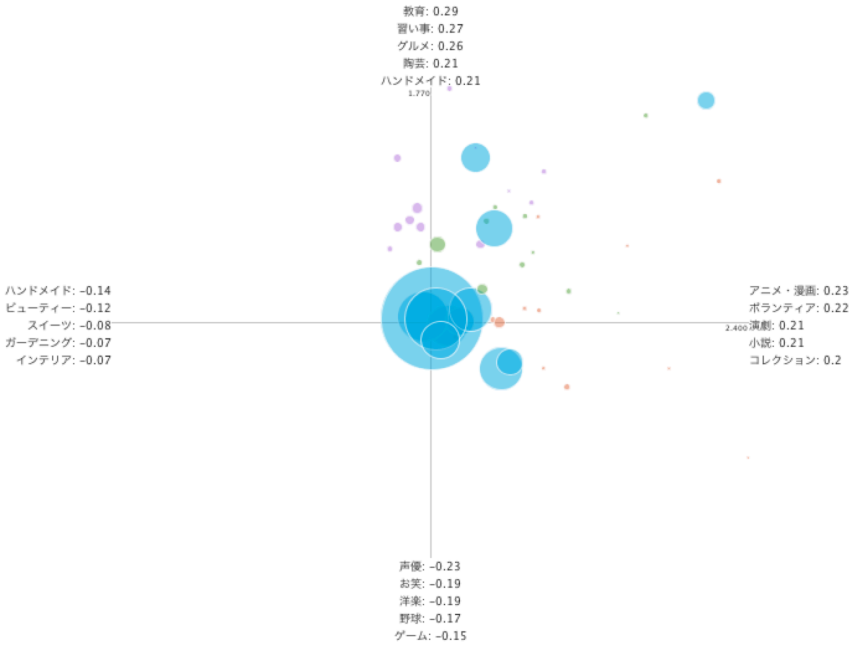

図 6.5: 妖怪主婦集団（紫色）、妖怪既婚勤め人集団（緑色）、妖怪未婚集団（橙色）および一般集団（水色）のレコードを主成分分析を用いて表現した Cartesian Panel。

次に、一般集団以外の 3 集団のみでの比較を行った。図 6.6 は、妖怪主婦集団（紫色）、妖怪既婚勤め人集団（緑色）および妖怪未婚集団（橙色）を Blade Graph により表現した Main Panel である。ソートなどの条件は図 6.4 と同様である。紫色の妖怪主婦集団は、育児、ハンドメイド、習い事、インテリアにおいて、他の 2 集団と比べて高スコアが多い分布となっていた。緑色の妖怪既婚勤め人集団は、クルマバイク、自転車、サッカー、釣り、ゲームなどにお

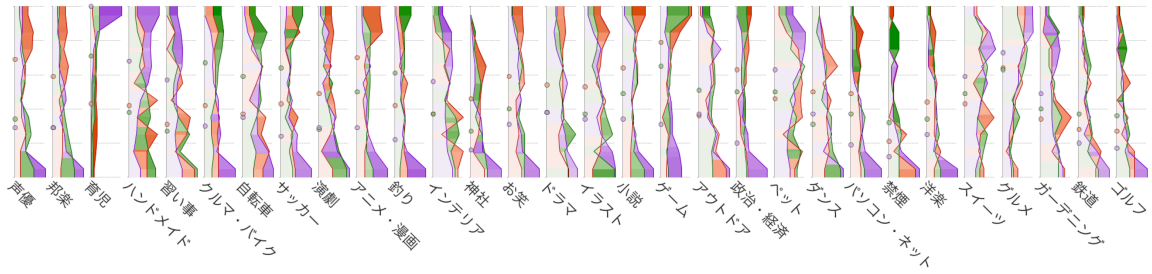

図 6.6: 妖怪主婦集団（紫色）、妖怪既婚勤め人集団（緑色）および妖怪未婚集団（橙色）を Blade Graph により表現した Main Panel。

いて高スコアが突出した分布を示していた。橙色の妖怪未婚集団は、声優、邦楽、演劇、アニメ漫画、ゲームなどにおいて高スコアが多かった。

次に、一般集団以外の 3 集団のレコードに対する主成分分析を行い、その結果を Cartesian Panel 上に可視化した（図 6.7 を参照）。第一主成分における正に大きな係数を持つ趣味は、習い事、育児、スイーツ、グルメ、ペットであり、負に大きな係数を持つ趣味は、アニメ漫画、映画テレビ、ゲーム、ギャンブル、邦楽である。第二主成分における正に大きな係数を持つ趣味は、アウトドア、節約、クルマバイク、ボランティア、パソコンネットである。各色の円を確認したところ、紫色の妖怪主婦集団は第一象限にすべてのクラスタが配置されていた。緑色の妖怪既婚勤め人集団と橙色の妖怪未婚集団は似た分布を示しており、第二象限を中心に分布していた。

さらに、妖怪既婚勤め人集団と妖怪未婚集団の差異を調査するため、この 2 集団のレコードに対する主成分分析を行って Cartesian Panel 上に可視化した（図 6.8 を参照）。図 6.8 から、橙色の妖怪未婚集団は主に第一象限、緑色の妖怪既婚勤め人集団は中央付近から第一主成分の正方向に対してそれぞれ分布していることが読み取れた。第一主成分は、ボランティア、アウトドア、政治経済、コレクション、小説が正に大きな係数の趣味である。第二主成分における正に大きな係数を持つ趣味は、お笑、お酒、声優、ドラマ、映画テレビであり、負に大きな係数を持つ趣味は、資格キャリア、教育、陶芸、着物、政治経済である。

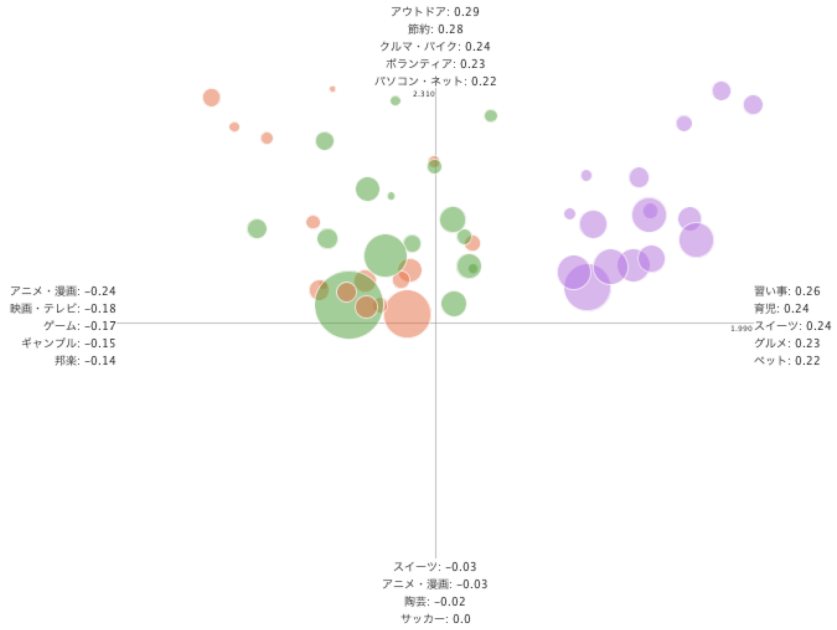

図 6.7: 妖怪主婦集団（紫色）、妖怪既婚勤め人集団（緑色）および妖怪未婚集団（橙色）のレコードを主成分分析を用いて表現した Cartesian Panel。

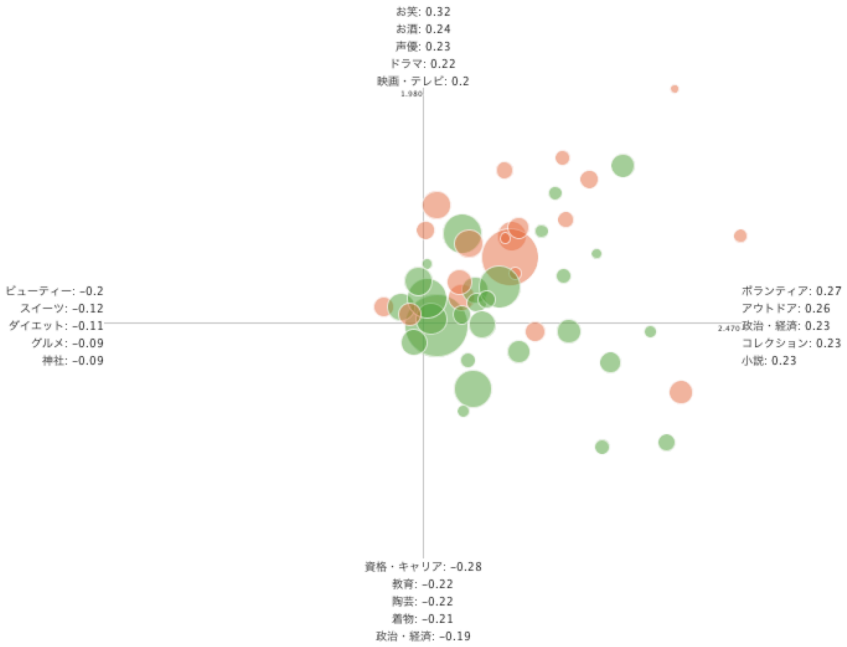

図 6.8: 妖怪既婚勤め人集団（緑色）および妖怪未婚集団（橙色）のレコードを主成分分析を用いて表現した Cartesian Panel。

6.2.2　妖怪主婦集団に関する比較観察

まずは妖怪主婦集団の概観を得るため、妖怪主婦集団（橙色）と一般集団（水色）の中央値を棒グラフにより表現した（図 6.9 を参照）。図 6.9 における趣味の並び順は、妖怪主婦集団と一般集団の差が大きい順であり、左ほど妖怪主婦集団の方が高い中央値である趣味が配置されている。図 6.9 より、育児やハンドメイド、習い事、グルメ、インテリアなどの趣味は一般集団と比べて高スコアを示していることがわかった。

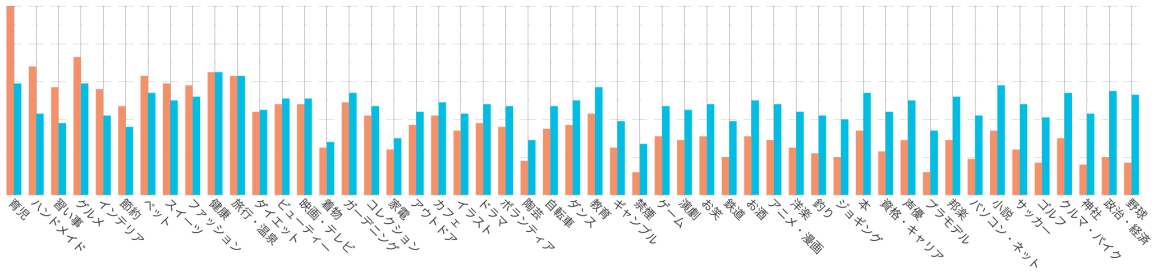

図 6.9: 妖怪主婦集団（橙色）と一般集団（水色）の中央値を棒グラフにより表現した Main Panel。

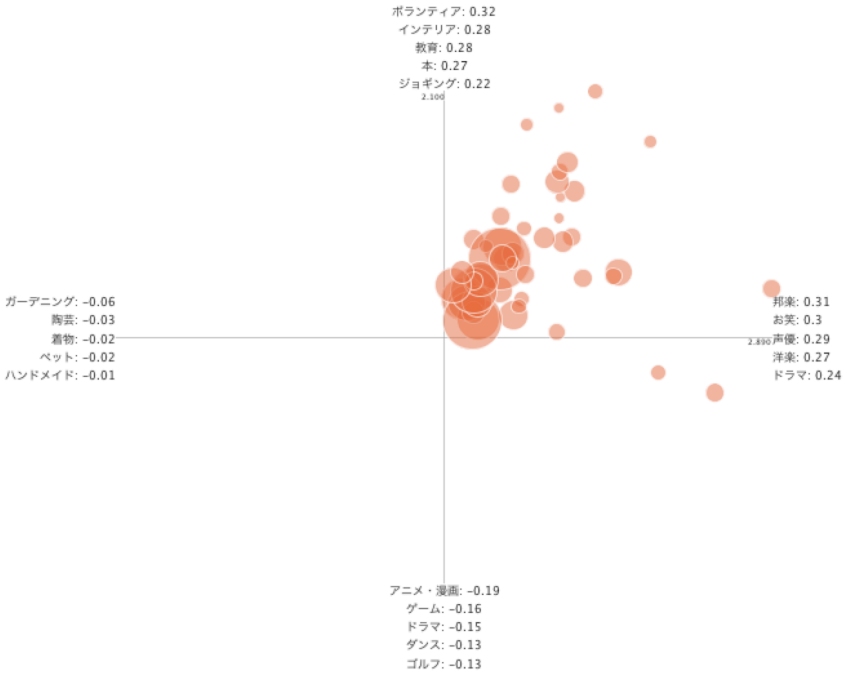

図 6.10: 妖怪主婦集団（橙色）のレコードを主成分分析を用いて表現した Cartesian Panel。

次に、妖怪主婦集団に関する主成分分析の結果（図 6.10 を参照）を調査した。図 6.10 では、妖怪主婦集団の第一主成分が横軸、第二主成分が縦軸に対応付けられている。図 6.10 より、

妖怪主婦集団は第一象限に多く分布していることが判明した。第一主成分における正に大きな係数を持つ趣味は、邦楽、お笑、声優、洋楽、ドラマである。第二主成分における正に大きな係数を持つ趣味は、ボランティア、インテリア、教育、本、ジョギングである。

図 6.11 は、妖怪主婦集団における趣味を $k = 10$ にてクラスタリングした結果である。ペットと温泉旅行のクラスタや、インテリア・ファッション・ハンドメイドのクラスタを確認できた。また図 6.12 は妖怪主婦集団に関する趣味の多次元尺度構成法の結果である。図 6.12 を確認すると、育児が他の趣味と異なる傾向を示していた。

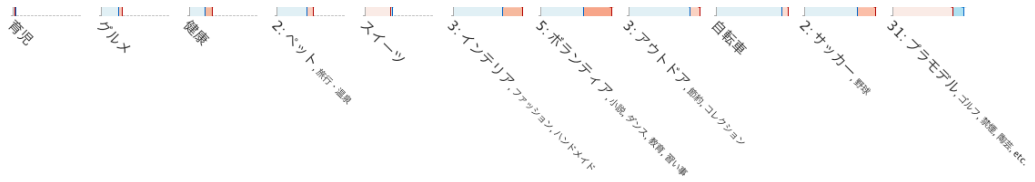

図 6.11: 妖怪主婦集団における趣味を $k = 10$ にてクラスタリングした結果。

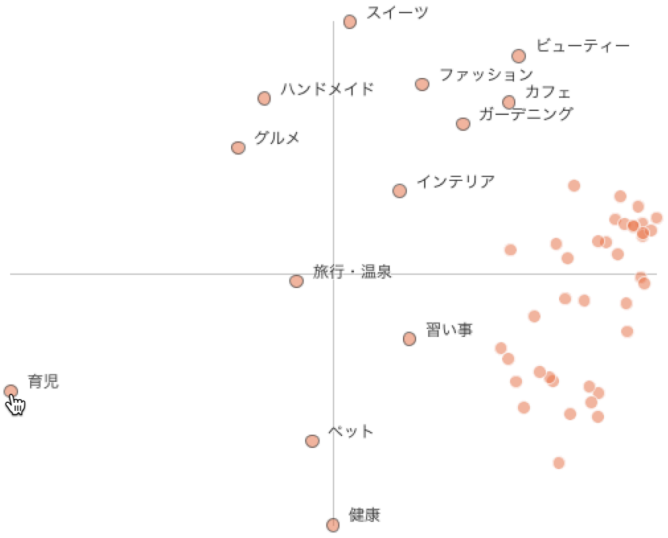

図 6.12: 妖怪主婦集団（橙色）における量的変量（趣味）を多次元尺度構成法により二元空間上にて表現した Cartesian Panel。

ここでさらに、妖怪主婦集団の中でも育児に対して特に高いスコアであった集団を観察するため、育児の Blade Graph からレコードのフィルタリングを行った。妖怪主婦集団に属していて育児のスコアが 0.87 以上であった 408 レコードから成る集団に紫色、妖怪主婦集団の中で育児のスコアが 0.87 未満であった 136 レコードから成る集団に緑色を割り当てた上で、両集団の分布を可視化した Blade Graph を Main Panel 上に表示した（図 6.13 を参照）。紫色の集団における中央値と緑色の集団における中央値の差分を計算して、差が大きい順に趣味の軸を並べ替えた。図 6.13 の (a) は、育児のスコアが高い紫色の集団の方が高い中央値であった上

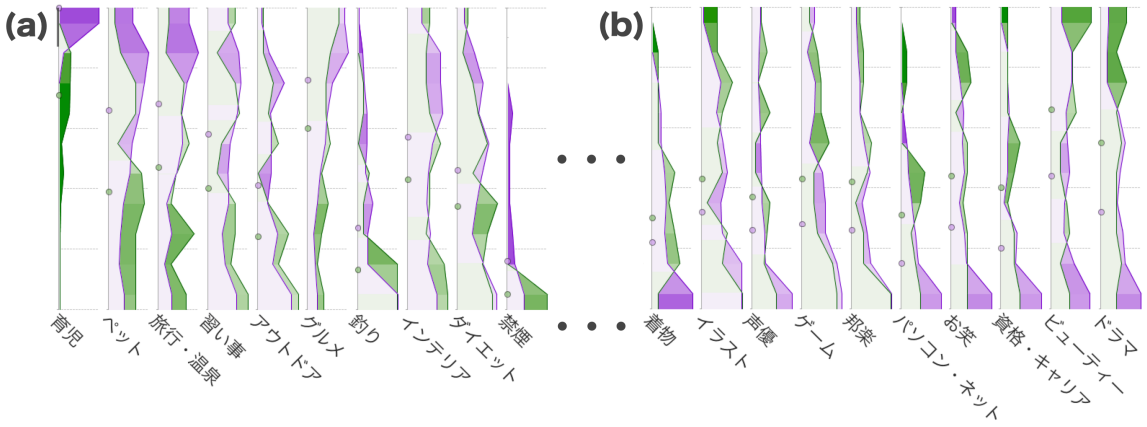

図 6.13: 妖怪主婦集団において、育児のスコアが高い集団（紫色）とそれ以外の集団（緑色）の分布を Blade Graph により表現した Main Panel の一部。

位 10 趣味を表示している。図 6.13(a) より、育児のスコアが高い妖怪主婦集団は、育児以外にペットや旅行温泉、習い事、アウトドアにおいても比較的高いスコアを示していることがわかった。また図 6.13 の (b) は、育児のスコアが低い緑色の集団の方が高い中央値であった上位 10 趣味を表示している。図 6.13(b) より、育児のスコアが低い妖怪主婦集団は、ドラマやビューティー、資格キャリア、お笑において比較的高スコアであることがわかった。また、ハンドメイド、ファッション、スイーツに関しては、両集団とも高い中央値を示していた。

図 6.14 は、育児のスコアが高い妖怪主婦集団（紫色）と育児のスコアが低い妖怪主婦集団（緑色）のレコードを主成分分析を用いて表現した Cartesian Panel である。図 6.14 を確認したところ、紫色の円は第二象限に多く配置されており、緑色の円は第一象限に多く配置されていた。第一主成分において正に大きな係数の趣味は、アニメ漫画、お笑、声優、ドラマ、映画テレビである。第一主成分において負に大きな係数の趣味は、自転車、育児、スイーツ、ペット、ガーデニングである。第二主成分において正に大きな係数の趣味は、節約、資格キャリア、家電、アウトドア、グルメである。

次に、妖怪主婦集団を対象とした比較分析を行った。比較対象には、一般主婦集団と妖怪集団を用いた。それぞれの集団要素の違いにより、どのような趣味の傾向の差異が発生しているのかを調査した。以降では、それぞれの比較において観察された事項について述べる。

図 6.15 は、妖怪主婦集団（橙色）、一般主婦集団（水色）および妖怪集団（紫色）の分布を Blade Graph により表現した Main Panel の一部である。妖怪主婦集団の中央値が大きい順に変量を並び替えた上で、上位 15 趣味の Blade Graph を表示している。図 6.15 より、育児、グルメ、ハンドメイド、ファッション、習い事、インテリアは他の 2 集団よりも高い中央値を示していた。これらの Blade Graph では、高スコアの部分に橙色が突出している様子が確認できた。一方で、健康、ペット、ガーデニングにおける高スコア部分は、水色の一般主婦集団が最も突出していた。また趣味間で分布の形を比較すると、3 集団とも育児は他の趣味とは特に異なる形を示していた。

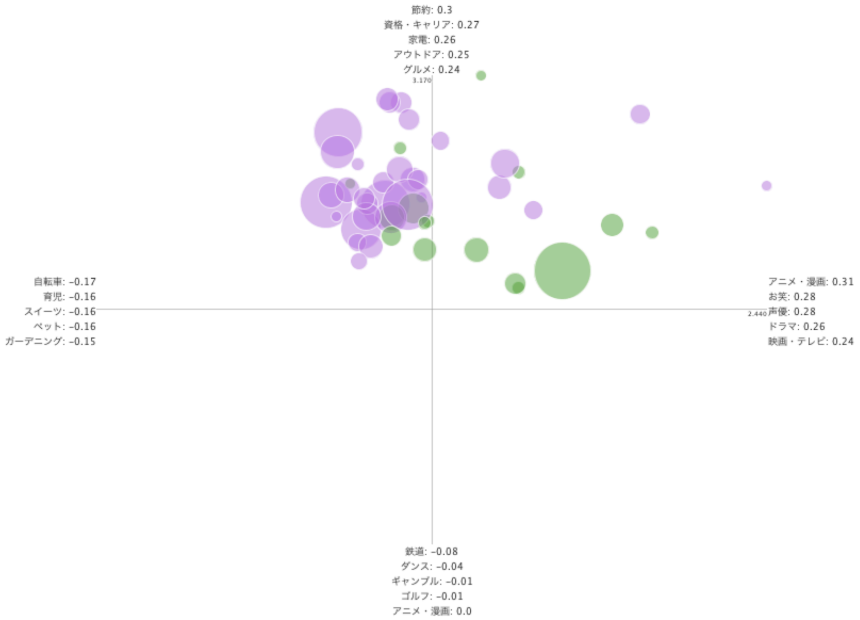

図 6.14: 妖怪主婦集団において、育児のスコアが高い集団（紫色）とそれ以外の集団（緑色）のレコードを主成分分析を用いて表現した Cartesian Panel。

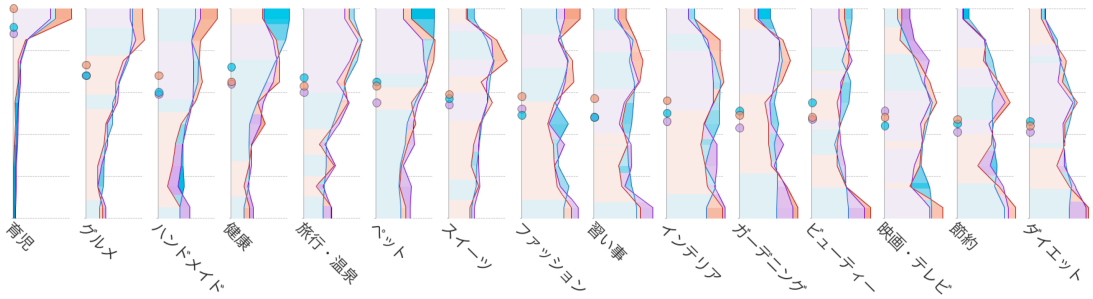

図 6.15: 妖怪主婦集団（橙色）、一般主婦集団（水色）および妖怪集団（紫色）の分布を Blade Graph により表現した Main Panel の一部。

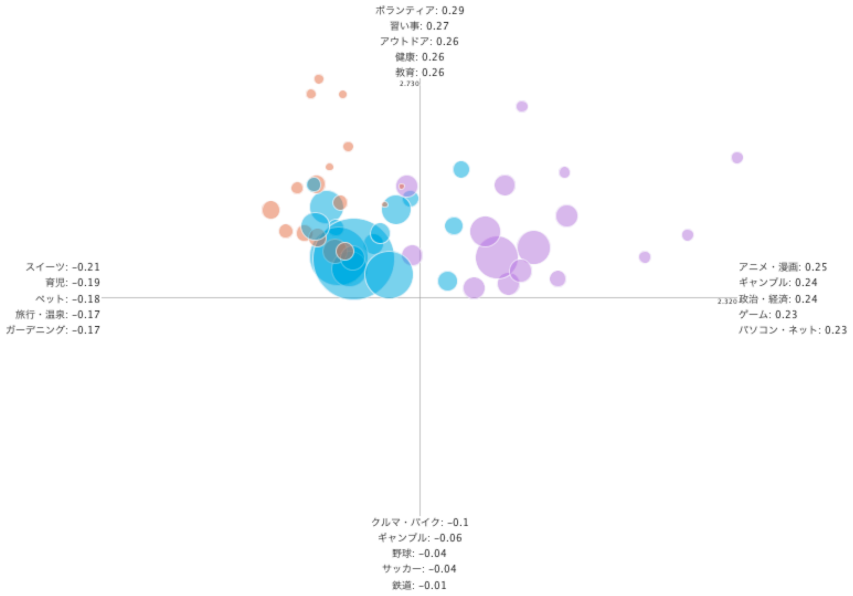

図 6.16: 妖怪主婦集団（橙色）、一般主婦集団（水色）および妖怪集団（紫色）のレコードを主成分分析を用いて表現した Cartesian Panel。

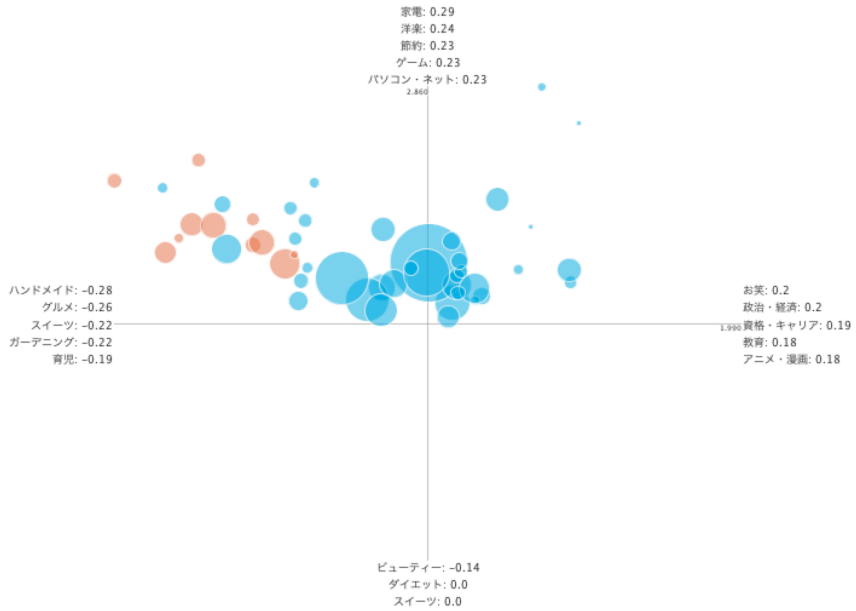

図 6.17: 妖怪主婦集団（橙色）と一般主婦集団（水色）のレコードを主成分分析を用いて表現した Cartesian Panel。

図 6.16 は、妖怪主婦集団（橙色）、一般主婦集団（水色）および妖怪集団（紫色）のレコードを主成分分析を用いて表現した Cartesian Panel である。まず妖怪主婦集団（橙色）は第二象限に分布しており、第二主成分の正方向に大きな値を持つクラスタが確認された。一般主婦集団（水色）は中央寄りの第二象限に分布していた。妖怪集団（紫色）は第一象限に多くのクラスタが集中しており、第一・第二主成分共に正方向に大きな値を持つクラスタが確認された。第一主成分において正に大きな係数の趣味は、アニメ漫画、ギャンブル、政治経済、ゲーム、パソコンネットである。第一主成分において負に大きな係数の趣味は、スイーツ、育児、ペット、旅行温泉、ガーデニングである。第二主成分において正に大きな係数の趣味は、ボランティア、習い事、アウトドア、健康、教育である。

妖怪主婦集団と一般主婦集団の分布が近似していたため、この 2 集団に着目してさらに調査を行った。図 6.17 は、妖怪主婦集団（橙色）と一般主婦集団（水色）のレコードを主成分分析を用いて表現した Cartesian Panel である。図 6.17 より、一般主婦集団は第一・第二象限にクラスタの分布が広がっていることがわかった。一方、妖怪主婦集団は第二象限のみに分布が広がっていた。第一主成分において正に大きな係数の趣味は、お笑、政治経済、資格キャリア、教育、アニメ漫画である。第一主成分において負に大きな係数の趣味は、ハンドメイド、グルメ、スイーツ、ガーデニング、育児である。第二主成分において正に大きな係数の趣味は、家電、洋楽、節約、ゲーム、パソコンネットである。両集団において、第二主成分を起因とする分布の差異はあまり見られなかった。

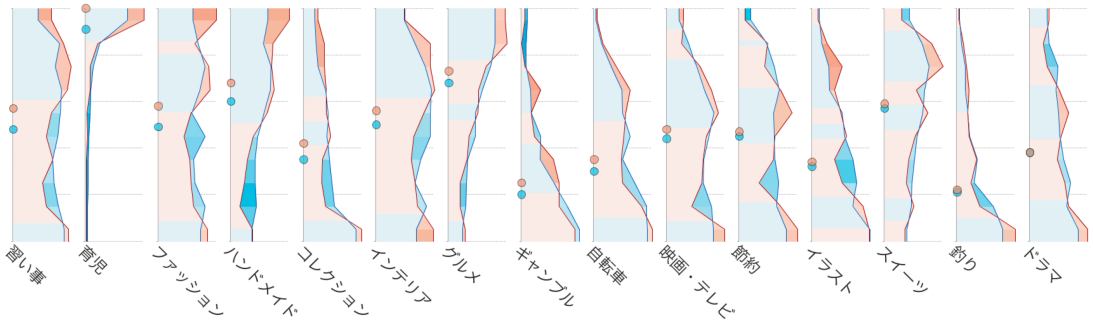

図 6.18: 妖怪主婦集団（橙色）と一般主婦集団（水色）の分布を Blade Graph により表現した Main Panel の一部。

図 6.18 は妖怪主婦集団（橙色）と一般主婦集団（水色）の分布を Blade Graph により表現した Main Panel の一部である。各趣味における評価値を、評価値 = (妖怪主婦集団における中央値) – (一般主婦集団における中央値) として、評価値が大きい順に趣味を並べている。図 6.18 の各 Blade Graph より、習い事、育児、ファッション、ハンドメイドなどは橙色の妖怪主婦集団の方が高スコアの割合が高いことがわかった。集団間での分布の形を比較したところ、上記の 4 趣味などはスコア上部が突出していたが、集団間に大きな差異は見られなかった。

6.2.3　妖怪既婚勤め人集団に関する比較観察

まずは妖怪既婚勤め人集団の概観を得るため、妖怪既婚勤め人集団（橙色）と一般集団（水色）の中央値を棒グラフにより表現した（図 6.19 を参照）。図 6.19 における趣味の並び順は、妖怪既婚勤め人集団と一般集団の差が大きい順であり、左ほど妖怪既婚勤め人集団の方が高い中央値である趣味が配置されている。図 6.19 より、釣りやゲーム、コレクション、サッカー、クルマバイクなどの趣味は一般集団と比べて高スコアを示していることがわかった。

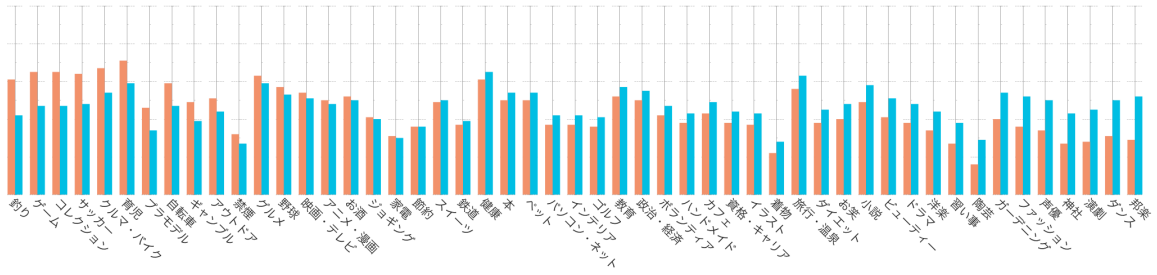

図 6.19: 妖怪既婚勤め人集団（橙色）と一般集団（水色）の中央値を棒グラフにより表現した Main Panel。

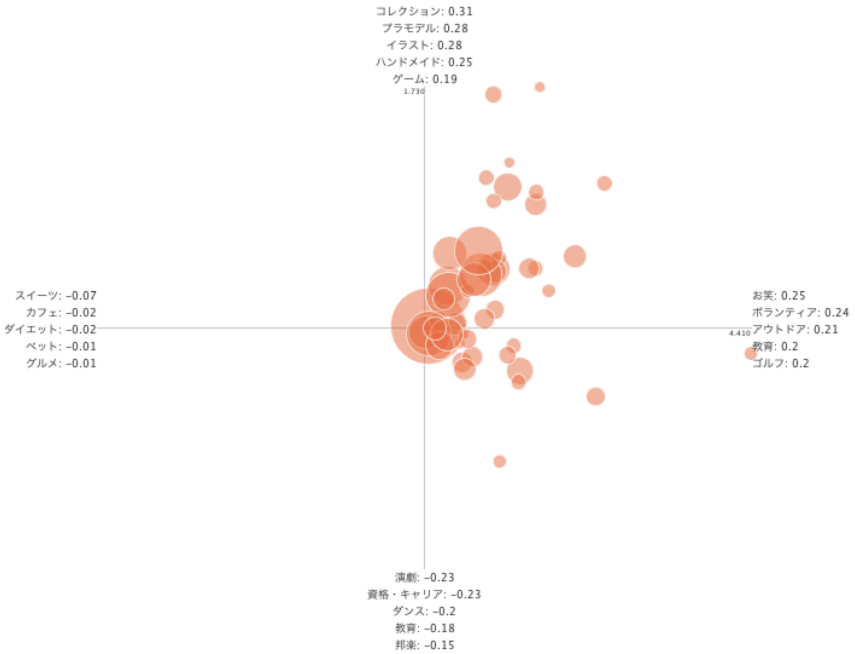

図 6.20: 妖怪既婚勤め人集団（橙色）のレコードを主成分分析を用いて表現した Cartesian Panel。

次に、妖怪既婚勤め人集団に関する主成分分析の結果（図 6.20 を参照）を調査した。図 6.20

を確認すると、妖怪既婚勤め人集団は第一象限に多く分布していた。第一主成分における正に大きな係数を持つ趣味は、お笑、ボランティア、アウトドア、教育、ゴルフである。第二主成分における正に大きな係数を持つ趣味は、コレクション、プラモデル、イラスト、ハンドメイド、ゲームである。

図 6.21 は、妖怪既婚勤め人集団における趣味を $k = 10$ にてクラスタリングした結果である。コレクションとゲームのクラスタや、アウトドア・自転車・クルマバイク・釣りのクラスタ、資格キャリア・政治経済・教育のクラスタなどを確認できた。また図 6.22 は妖怪既婚勤め人集団に関する趣味の多次元尺度構成法の結果である。図 6.22 から、育児やゲーム、コレクションなどが他の趣味と異なる傾向を示していることが確認できた。

図 6.21: 妖怪既婚勤め人集団における趣味を $k = 10$ にてクラスタリングした結果。

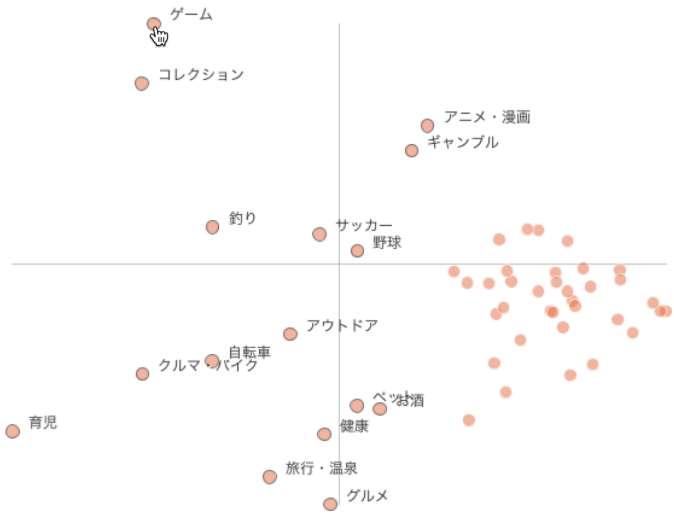

図 6.22: 妖怪既婚勤め人集団（橙色）における量的変量（趣味）を多次元尺度構成法により二元空間上にて表現した Cartesian Panel。

ここでさらに、妖怪既婚勤め人集団の中でも育児に対して特に高いスコアであった集団を観察するため、育児の Blade Graph からレコードのフィルタリングを行った。妖怪既婚勤め人集団に属していて育児のスコアが 0.81 以上であった 130 レコード（妖怪既婚勤め人集団において育児スコアが上位 25%に含まれるユーザ集団）から成る集団に紫色を割り当てた。また同様に、妖怪既婚勤め人集団の中で育児のスコアが 0.81 未満であった 383 レコードから成る

集団に緑色を割り当てた。その上で、両集団の分布を可視化した Blade Graph を Main Panel 上に表示した（図 6.23 を参照）。紫色の集団における中央値と緑色の集団における中央値の差分を計算して、差が大きい順に趣味の軸を並べ替えた。図 6.23 の (a) は、育児のスコアが高い紫色の集団の方が高い中央値であった上位 10 趣味を表示している。図 6.23(a) より、育児のスコアが高い妖怪既婚勤め人集団は、育児以外に資格キャリアやアウトドア、自転車、小説などにおいても比較的高いスコアを示していることがわかった。また図 6.23 の (b) は、育児のスコアが低い緑色の集団の方が高い中央値であった上位 10 趣味を表示している。図 6.23(b) より、育児のスコアが低い妖怪既婚勤め人集団は、ビューティーや禁煙、プラモデル、ドラマにおいて比較的高スコアであることがわかった。また、健康、サッカー、ゲーム、グルメに関しては、両集団とも高い中央値を示していた。

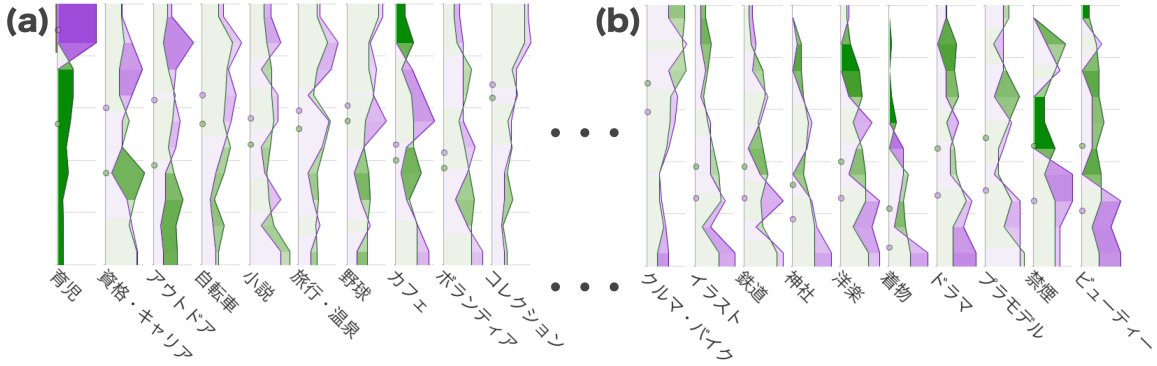

図 6.23: 妖怪既婚勤め人集団において、育児のスコアが高い集団（紫色）とそれ以外の集団（緑色）の分布を Blade Graph により表現した Main Panel の一部。

図 6.24 は、育児のスコアが高い妖怪既婚勤め人集団（紫色）と育児のスコアが低い妖怪既婚勤め人集団（緑色）のレコードを主成分分析を用いて表現した Cartesian Panel である。図 6.24 を確認したところ、緑色の円が第一主成分の正方向に比較的広く分散していたものの、両集団において大きな分布の差は見られなかった。第一主成分において正に大きな係数の趣味は、資格キャリア、お笑、演劇、政治経済、パソコンネットである。第二主成分において正に大きな係数の趣味は、プラモデル、ゴルフ、自転車、ギャンブル、釣りである。第二主成分において負に大きな係数の趣味は、邦楽、声優、映画テレビ、本、ドラマである。

次に、妖怪既婚勤め人集団（橙色）を対象とした比較分析を行った。比較対象には、妖怪集団（水色）と一般既婚勤め人集団（紫色）を用いた。それぞれの集団要素の違いにより、どのような趣味の傾向の差異が発生しているのかを調査した。以降では、それぞれの比較において観察された事項について述べる。

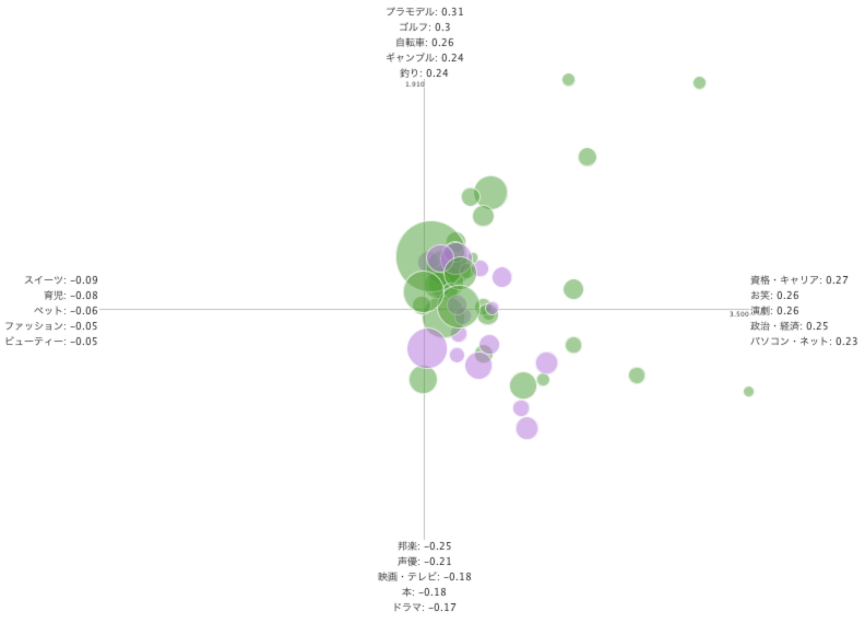

図 6.24: 妖怪既婚勤め人集団において、育児のスコアが高い集団（紫色）とそれ以外の集団（緑色）のレコードを主成分分析を用いて表現した Cartesian Panel。

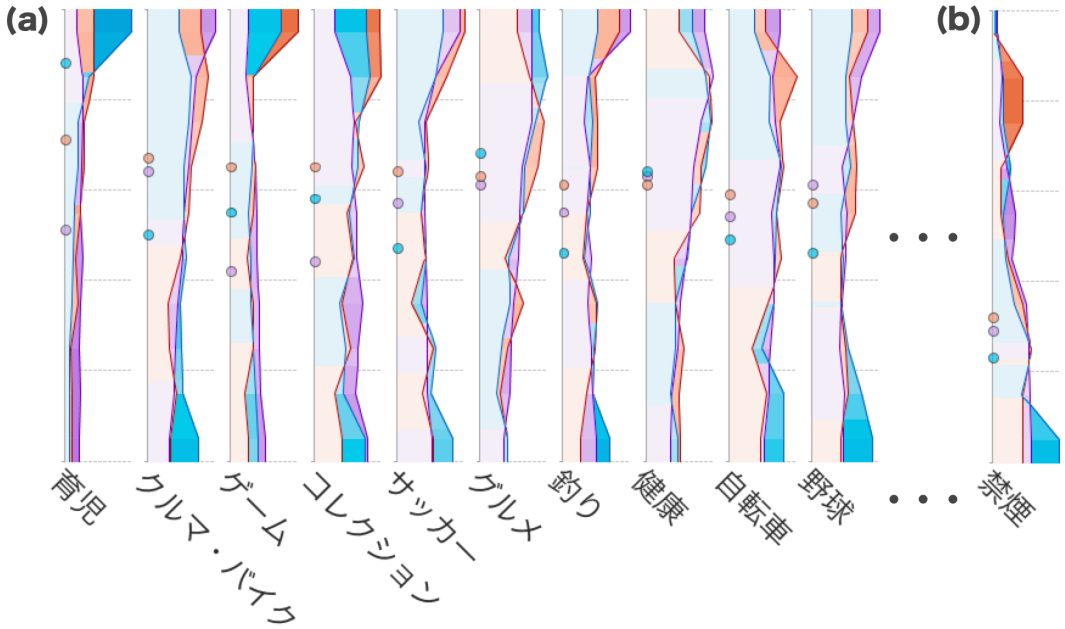

図 6.25: 妖怪既婚勤め人集団（橙色）、妖怪集団（水色）および一般既婚勤め人集団（紫色）の分布を Blade Graph により表現した Main Panel の一部。

図 6.25 は、妖怪既婚勤め人集団（橙色）、妖怪集団（水色）および一般既婚勤め人集団（紫色）の分布を Blade Graph により表現した Main Panel の一部である。妖怪既婚勤め人集団の中央値が大きい順に変量を並び替えた上で、図 6.25(a) にて上位 10 趣味の Blade Graph を表示している。ゲームおよびコレクションは、橙色の妖怪既婚勤め人集団と水色の妖怪集団が共に高スコアにおいて突出した分布となっていた。また、クルマバイクや釣りは、橙色の妖怪既婚勤め人集団と紫色の一般既婚勤め人集団が共に高スコアとなっていた。図 6.25(b) は、上位以外の趣味で特徴的な分布を示していた禁煙の Blade Graph である。ここから、禁煙の高スコア部分は橙色の妖怪既婚勤め人集団が比較的多く分布していることが読み取れた。

図 6.26 は、妖怪既婚勤め人集団（橙色）、妖怪集団（水色）および一般既婚勤め人集団（紫色）のレコードを主成分分析を用いて表現した Cartesian Panel である。橙色の妖怪既婚勤め人集団と水色の妖怪集団は、第一象限に多くのクラスタが分布していた。一方で紫色の一般既婚勤め人集団は、主に中央付近と第四象限に分布していた。第一主成分において正に大きな係数の趣味は、教育、小説、政治経済、ボランティア、資格キャリアである。第二主成分において正に大きな係数の趣味は、アニメ漫画、映画テレビ、ギャンブル、ゲーム、ドラマである。第二主成分において負に大きな係数の趣味は、ハンドメイド、カフェ、着物、ボランティア、陶芸である。

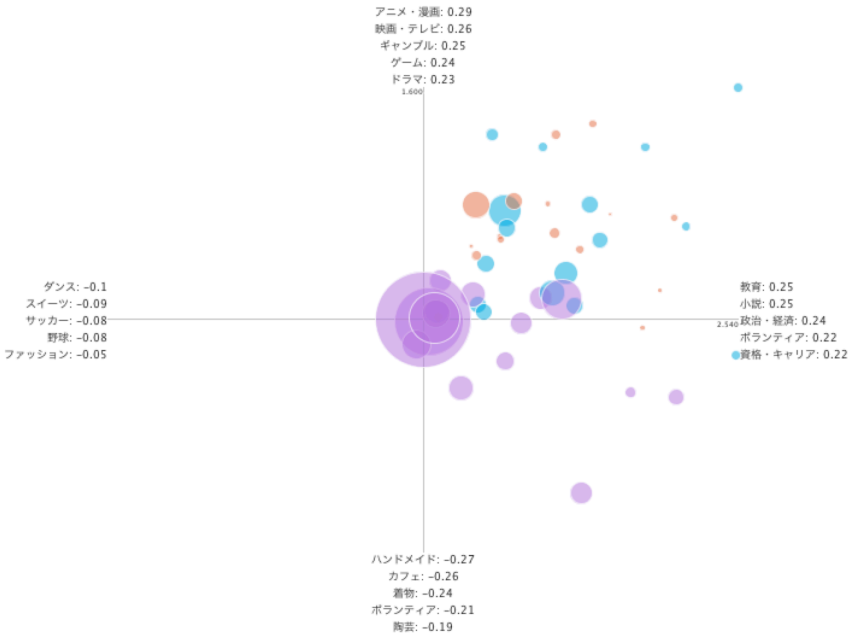

図 6.26: 妖怪既婚勤め人集団（橙色）、妖怪集団（水色）および一般既婚勤め人集団（紫色）のレコードを主成分分析を用いて表現した Cartesian Panel。

図 6.26 において妖怪既婚勤め人集団と妖怪集団の分布が近似していたため、この 2 集団に着目してさらに調査を行った。図 6.27 は、妖怪既婚勤め人集団（橙色）と妖怪集団（水色）のレコードを主成分分析を用いて表現した Cartesian Panel である。橙色の妖怪既婚勤め人集

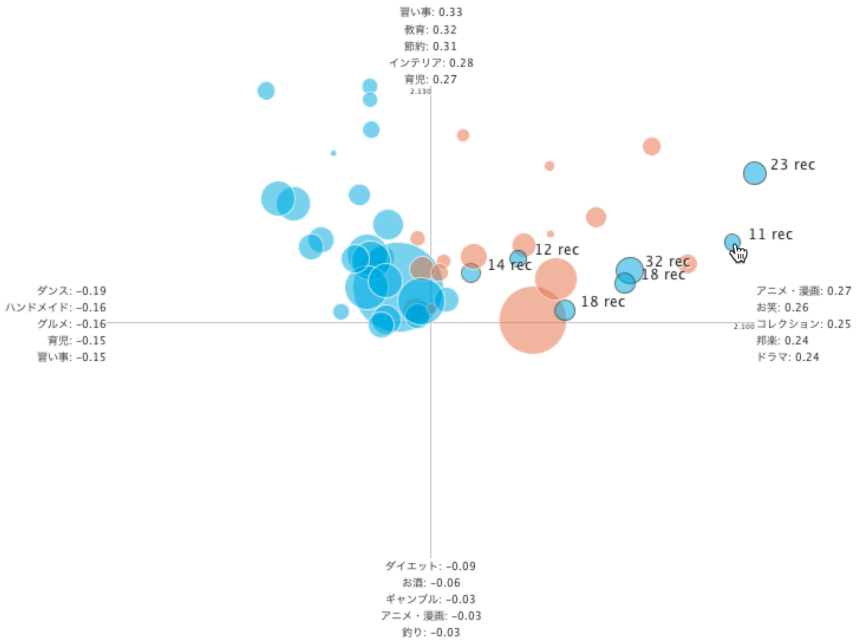

図 6.27: 妖怪既婚勤め人集団（橙色）と妖怪集団（水色）のレコードを主成分分析を用いて表現した Cartesian Panel。

団は、第一象限の広い範囲に殆どのクラスタが分布していた。一方で水色の妖怪集団は、妖怪既婚勤め人集団と似た傾向を示すクラスタと、第二象限に配置されているクラスタが確認できた。ここで、前者のクラスタ（図 6.27 中で縁が黒くなっている円のクラスタ）を選択した上で、この部分集合の職業を確認した。その結果、これらのクラスタに含まれるレコードの職業はすべて勤め人であることが判明した。なお図 6.27 において、第一主成分において正に大きな係数の趣味は、アニメ漫画、お笑、コレクション、邦楽、ドラマである。第一主成分において負に大きな係数の趣味は、ダンス、ハンドメイド、グルメ、育児、習い事である。第二主成分において正に大きな係数の趣味は、習い事、教育、節約、インテリア、育児である。

6.2.4　妖怪未婚集団に関する比較観察

まずは妖怪未婚集団の概観を得るため、妖怪未婚集団（橙色）と一般集団（水色）の中央値を棒グラフにより表現した（図 6.28 を参照）。図 6.28 における趣味の並び順は、妖怪未婚集団と一般集団の差が大きい順であり、左ほど妖怪未婚集団の方が高い中央値である趣味が配置されている。図 6.28 より、ゲーム、コレクション、アニメ漫画、声優、プラモデルなどの趣味は一般集団と比べて高スコアを示していることがわかった。

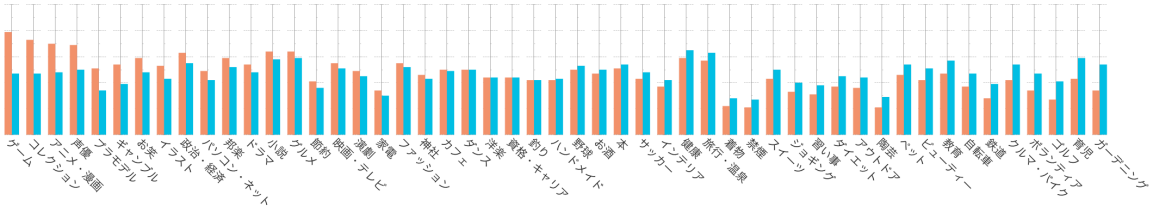

図 6.28: 妖怪未婚集団（橙色）と一般集団（水色）の中央値を棒グラフにより表現した Main Panel。

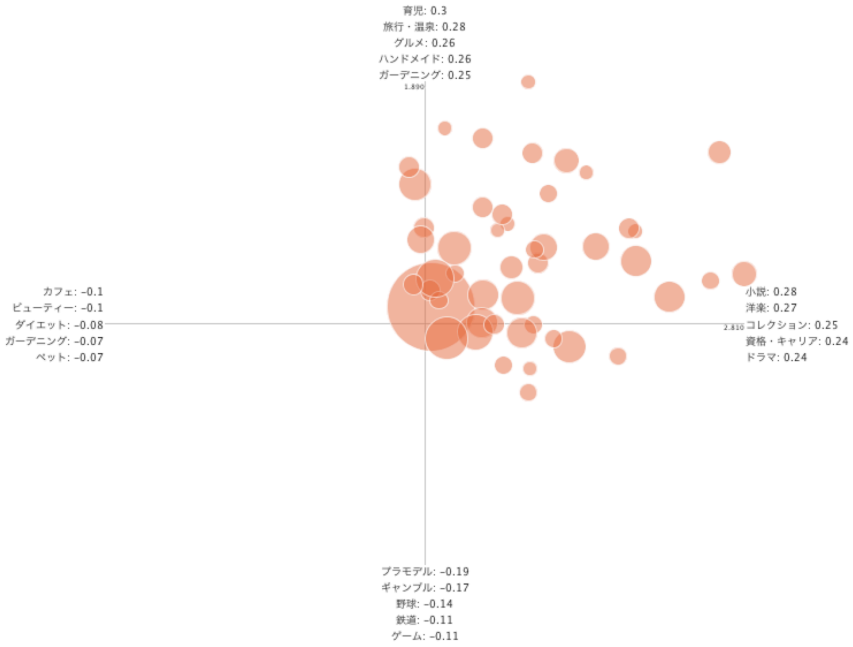

図 6.29: 妖怪未婚集団（橙色）のレコードを主成分分析を用いて表現した Cartesian Panel。

図 6.30: 妖怪未婚集団における趣味を $k = 10$ にてクラスタリングした結果。

次に、妖怪未婚集団に関する主成分分析の結果（図 6.29 を参照）を調査した。図 6.29 を確認すると、妖怪未婚集団は第一象限に多く分布していた。第一主成分における正に大きな係数を持つ趣味は、小説、洋楽、コレクション、資格キャリア、ドラマである。第二主成分における正に大きな係数を持つ趣味は、育児、旅行温泉、グルメ、ハンドメイド、ガーデニングである。

図 6.30 は、妖怪未婚集団における趣味を $k = 10$ にてクラスタリングした結果である。声優・アニメ漫画・コレクション・ゲームのクラスタ、アウトドア・ボランティア・教育・演劇のクラスタなどを確認できた。また図 6.31 は妖怪未婚集団に関する趣味の多次元尺度構成法の結果である。図 6.31 から、声優、アニメ漫画、コレクション、ゲーム、グルメなどが他の趣味と異なる傾向を示していることが確認できた。

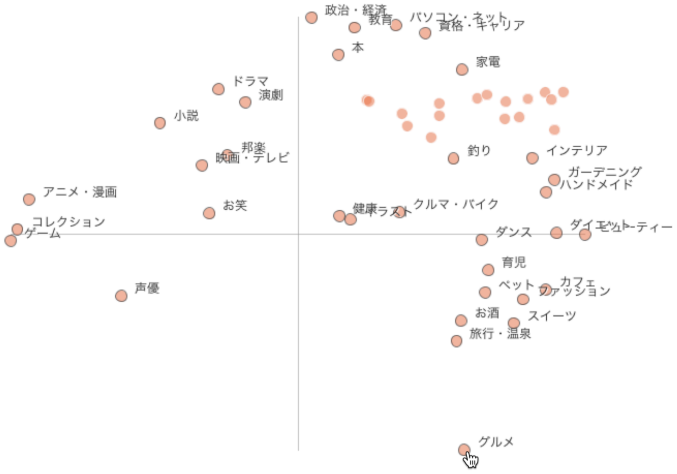

図 6.31: 妖怪未婚集団（橙色）における量的変量（趣味）を多次元尺度構成法により二元空間上にて表現した Cartesian Panel。

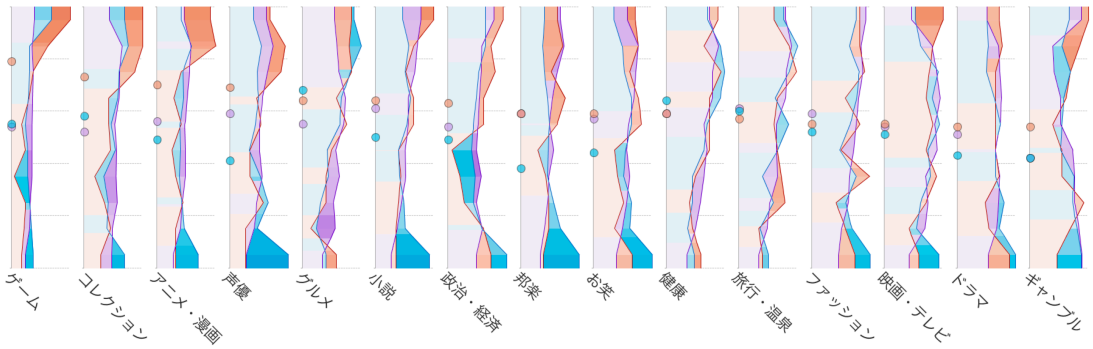

図 6.32: 妖怪未婚集団（橙色）、妖怪集団（水色）および一般未婚集団（紫色）の分布を Blade Graph により表現した Main Panel の一部。

次に、妖怪未婚集団（橙色）を対象とした比較分析を行った。比較対象には、妖怪集団（水色）と一般未婚集団（紫色）を用いた。それぞれの集団要素の違いにより、どのような趣味の傾向の差異が発生しているのかを調査した。以降では、それぞれの比較において観察された事項について述べる。

図 6.32 は、妖怪未婚集団（橙色）、妖怪集団（水色）および一般未婚集団（紫色）の分布を Blade Graph により表現した Main Panel の一部である。図 6.32 では、妖怪未婚集団の中央値が大きい順に変量を並び替えた上で、上位 15 趣味の Blade Graph を表示している。ゲームとコレクションは水色の妖怪集団も比較的高スコアの分布だが、妖怪未婚集団はそれ以上に高スコアが突出した分布であることがわかった。他にもアニメ漫画や小説、映画テレビにおいて妖怪未婚集団は高スコアの分布であった。

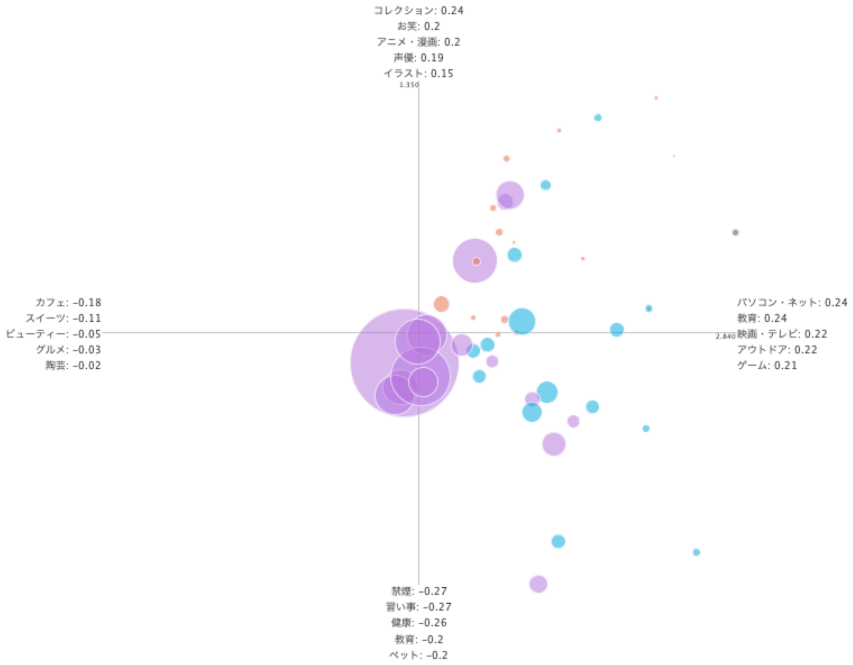

図 6.33: 妖怪未婚集団（橙色）、妖怪集団（水色）および一般未婚集団（紫色）のレコードを主成分分析を用いて表現した Cartesian Panel。

図 6.33 は、妖怪未婚集団（橙色）、妖怪集団（水色）および一般未婚集団（紫色）のレコードを主成分分析を用いて表現した Cartesian Panel である。橙色の妖怪未婚集団は、主に第一象限にクラスタが分布していた。水色の妖怪集団は、第一・第四象限に多くのクラスタが分布していた。紫色の一般未婚集団は、主に中央付近と第一・第四象限に分布していた。第一主成分において正に大きな係数の趣味は、パソコンネット、教育、映画テレビ、アウトドア、ゲームである。第二主成分において正に大きな係数の趣味は、コレクション、お笑い、アニメ漫画、声優、イラストである。第二主成分において負に大きな係数の趣味は、禁煙、習い事、健康、教育、ペットである。

6.3 考察

まず妖怪集団は、育児、ハンドメイド、プラモデル、コレクション、ゲームに対して、一般集団と比べて高い関心を示している（図 6.1 を参照）。また妖怪集団に対する主成分分析（図 6.2 を参照）では、第二主成分に対して大きな分散が見られた。ここから、ブログ上で “妖怪ウォッチ” に言及している集団は、以下の 2 種類のいずれかまたは両方に該当する集団であると考えられる。

- ゲームやアニメに関連して “妖怪ウォッチ” に関心を示している集団。
- 育児に関連して “妖怪ウォッチ” に言及している集団。

また妖怪集団において、既婚の主婦ユーザが約 39%（544/1,406）、既婚の勤め人ユーザが約 36%（513/1,406）、未婚のユーザが約 20%（287/1,406）をそれぞれ占めている。先に述べた妖怪集団の特徴は、それぞれの集団における特徴が混在した結果であると推測できる。そこで本節では、それぞれの集団において観察された傾向や特徴について述べる。

まず 6.2.1 節では、妖怪主婦集団、妖怪既婚勤め人集団および妖怪未婚集団の間での比較を行った（図 6.6 を参照）。妖怪主婦集団は他の集団と比べて、育児、ハンドメイド、習い事に対して高い関心を示している。また他集団と比べて低スコアが突出している趣味も多く存在していることから、妖怪主婦集団は関心が偏っていることが推測できる。図 6.8 の主成分分析からも、第一主成分の正方向である習い事や育児などに関心を示している集団であることがわかる。主成分分析の結果から見ると、妖怪既婚勤め人集団と妖怪未婚集団は比較的似た傾向であることが伺える。図 6.6 の各趣味における Blade Graph からも、両集団が似た分布を示していることが読み取れる。特にゲームは両集団共に強い関心を示している。両集団で異なる部分として、声優、邦楽、アニメ漫画などは妖怪未婚集団の方が顕著に高スコアが多い分布となっている。一方、妖怪既婚勤め人集団の方が関心の高い趣味としては、クルマバイク、自転車、サッカー、釣りなどが挙げられる。また妖怪主婦集団ほどではないものの、妖怪既婚勤め人集団は育児に対しても比較的高い関心を示している。

図 6.15 より、妖怪主婦集団は一般の主婦集団と比べても育児に対する関心が高いことが伺える。育児に加えて、グルメやハンドメイド、ファッションに対しても、妖怪主婦集団は一般主婦よりも高い関心を示している。一方で健康やペットに対しては、妖怪主婦集団は一般主婦集団ほどの関心を示していない。また妖怪主婦集団と一般主婦集団に対する主成分分析の結果（図 6.17 を参照）から、妖怪主婦集団は一般主婦集団と比べ、第二主成分の負方向に偏っていた。つまり妖怪主婦集団は、ハンドメイド、グルメ、スイーツ、ガーデニング、育児などには比較的興味がある一方で、お笑、政治経済、資格キャリア、教育、アニメ漫画にはあまり興味を示さない傾向にあることがわかった。以上より妖怪主婦集団は、ゲームやアニメ漫画などの娯楽に対する関心が低い一方で、特に育児やハンドメイドなどの家庭的な趣味に対する関心が特に高いユーザの多い集団であることが推測できる。

図 6.25 より、妖怪既婚勤め人集団は一般の既婚勤め人集団と比べ、特にゲームやコレクションに対して高い関心を示していることがわかる。また育児についても、妖怪既婚勤め人集団の方が一般の既婚勤め人集団よりも高い関心を示している。他に既婚勤め人集団が持つ傾向と

して、禁煙に対する関心の高さが挙げられる。これは他の集団では確認されなかった傾向である。図 6.26 にて示した主成分分析の結果から、妖怪既婚勤め人集団は一般の勤め人集団と比べ、特に第一主成分の正方向に対して高い値を持っていることが読み取れた。“妖怪ウォッチ”に言及している既婚勤め人の集団は、アニメ漫画やゲームだけではなく、映画テレビやギャンブル、ドラマに対しても、一般の既婚勤め人集団より高い関心を持っている可能性があることを意味している。以上より妖怪既婚勤め人集団は、育児への関心とアニメ漫画やゲーム的な関心の両方を持っている集団であることが推測できる。

図 6.32 より、妖怪未婚集団は一般の未婚集団と比べ、ゲームやコレクション、アニメ漫画、声優、グルメに対して強い関心を示していることが判明した。また図 6.33 にて示した主成分分析の結果から、妖怪未婚集団は一般の未婚集団と比べ、禁煙や習い事、健康などに対する関心が弱いことが推測できる。図 6.6 より、妖怪未婚集団は他の集団と比べて育児への関心が低いことがわかる。以上より妖怪未婚集団は、アニメ漫画やゲーム、声優などの娯楽的な関心が高い集団であることが推測できる。

また 6.2.2 節の妖怪主婦集団に関する比較分析では、育児に高い関心を持っている妖怪主婦集団と育児に強い関心を持っていない妖怪主婦集団を比較した。図 6.13 および図 6.14 より、育児に関心の高い妖怪主婦集団は、旅行温泉やアウトドアを始めとした、複数人を対象とするような趣味や家庭的な趣味に対して比較的興味を持っている。一方の育児に対する関心の低い妖怪主婦集団は、美容や資格キャリアなど、個人を対象とした趣味に対して比較的高い関心を示す傾向にある。以上より妖怪主婦集団において、育児に対する関心の高さはそのユーザの関心が誰に向いているのかを推測する材料になると考えられる。

同様に 6.2.3 節では、妖怪主婦集団ほどではないものの育児への関心が比較的高かった妖怪既婚勤め人集団においても、育児に関心が高い集団と関心が高くない集団の比較を行った。図 6.23 および図 6.24 より、妖怪既婚勤め人集団においては妖怪主婦集団において見られたような特徴は現れず、各集団において見られる傾向には大きな差がないことがわかる。両集団において差異が見られた趣味は以下の通りである。育児に関心の高い妖怪既婚勤め人は、資格キャリア、アウトドア、自転車などに対しても比較的高い関心を示す傾向にある。一方、育児にあまり関心がない妖怪既婚勤め人は、美容や禁煙、プラモデルなどに対して比較的高い関心を示す傾向にある。

6.4 まとめ

6.3 節の考察に基づき、分析結果を以下の通りにまとめる。

1. ブログ上にて“妖怪ウォッチ”に言及しているユーザについて、その大半は以下の 3 つの属性のいずれかに分類される：**主婦であるブログユーザ（約 39%）**；**既婚の勤め人（約 36%）**；**未婚のブログユーザ（約 20%）**。

2. ブログ上にて“妖怪ウォッチ”に言及している集団は、おおよそ以下の 2 種類のいずれかまたは両方に該当する：**A. ゲームやアニメに関連して“妖怪ウォッチ”に関心を示して**

いる集団；B. 育児に関連して “妖怪ウォッチ” に言及している集団。ブログ上において “妖怪ウォッチ” に言及している主婦は、A の集団に該当する傾向が強い。また、“妖怪ウォッチ” に言及している既婚の勤め人は、A と B の両集団に該当する傾向が強い。“妖怪ウォッチ” に言及している未婚のブログユーザは、B の集団に該当する傾向が強い。

3. “妖怪ウォッチ” に言及している主婦ユーザは、育児に対する関心の高さから趣味の傾向を推測できる。育児に関心の高い主婦は、複数人を対象とするような趣味や家庭的な趣味に対して興味を持つ傾向がある。育児に対する関心が低い主婦は、個人を対象とした趣味に対して興味を持つ傾向がある。なおこれらの育児に関する傾向は、同様に育児への関心が高かった “妖怪ウォッチ” に言及している既婚の勤め人においては確認されなかった。これらの傾向の違いは、両集団のユーザが持つ興味の範囲の違いに起因していると考えられる。社会に出て働いている既婚の勤め人ユーザは、主婦と比べて関心の範囲が広く、様々なことに対してブログ記事を書いていると推測できる。一方の主婦ユーザは、既婚の勤め人よりも興味の対象が限定的であり、自身の関心内における事柄を多く記事にしていると推測できる。ただしこれらは推測であるため、生データを閲覧するなどして検証する必要がある。

第7章　議論

本章では、本論文において開発した視覚的表現および分析ツールの比較分析への有用性、および本分析ツールの課題について考察する。

本論文では、複数の部分集合間における量的変量の分布を比較するための表現手法である Blade Graph を開発した。Blade Graph では、部分集合間の分布の差異が大きい部分ほど目立つ色となるように色付けを行う。これにより、集合間の差異をひと目で発見することができる。また各集合における割合によって Blade Graph の高さを正規化することにより、データ量の異なる部分集合間の分布を比較できる。本論文において実施した本分析ツールのケーススタディでは、Main Panel 上に表示された各趣味の Blade Graph を確認することにより、各集団・各趣味における詳細な分布やその差異を確認できた。

また本ケーススタディでは、質的変量であるユーザの職業属性からデータセットを分割して分析を進めた。さらに、趣味の分布からユーザをフィルタリングして部分集合を抽出することによって、さらに詳細な条件によってデータセットを分割・比較しながら分析を行った。このように本分析ツールでは、多変量データセットから着目したい部分集合を抽出し、かつそれらを比較することにより、部分集合の特徴を発見することができる。

本分析ツールでは、クラスタリングしたレコードの主成分分析を表示することにより、部分集合に属する各クラスタの分布を確認でき、またそれらの差異を把握することができる。主成分分析の結果を見ながら各主成分における影響の強い変量を確認することにより、各部分集合の傾向を推測することも可能である。本ケーススタディでは、主成分分析の結果を示した Cartesian Panel から各集団の分布を確認した上で影響の強い主成分を調査することにより、各集団の傾向を推測する一助となった。

ケーススタディを通じて、特徴的部分集合を視覚的に探索して比較できる本分析ツールならではの知見を得ることができた。6.4 節にて述べたケーススタディのまとめにおいて、一番目の結果については他の分析ツールを用いても発見可能な知見である。一方で二番目と三番目の知見は、任意の部分集合を抽出して視覚的に比較できる本分析ツールの表現および機能によって得られたものである。部分集合間を視覚的に比較することにより、多変量データセット全体を概観するだけでは発見できないような、データセットを構成する個々の部分集合ごとの特徴を発見できた。

本分析ツールの課題としては、質的変量における比較表現の改良が挙げられる。本ツールのプロトタイプ版 [45] では、各部分集合における質的変量の積み上げ棒グラフを半透明にした上で重畳表示していた。しかし、複数の部分集合を同時に表現すると個々の分布が見づらく比較が困難であった。またカテゴリの区別に色相を利用していたため、量的変量の表現に割

り当てられていた色相と混同してしまうという欠点があった。そのため、本ツールでは図 5.2 のような表現を採用し、比較のための補助的な情報はマウスオーバ時にテキスト形式によって提供するように変更した。本ツールにおける質的変量の表現は、個々の部分集合における質的変量の割合を確認することに特化している。しかし、部分集合間での質的変量の比較という観点では視覚的な支援が行えていないため、質的変量の表現については更なる改良の余地があると考えられる。

第8章　結論

多変量データセットの分析における、特徴的な部分集合の探索および比較の支援を目的とした研究を行った。研究目的を達成するためのアプローチとして、多変量データセットに含まれる特徴的な部分集合をインタラクティブに探索でき、かつデータ分布を視覚的に比較できる表現や機能を備える視覚的分析ツールを開発した。

まずデータ間の差異の探索を支援するため、データ分布の比較タスクに特化して設計した視覚的表現である Blade Graph を開発した。Blade Graph では $L^*a^*b^*$ 色空間を用いることにより知覚的均等性を考慮した上で、差の大きい部分が目立つような色付けを採用している。さらに、Blade Graph は異なるレコード数の部分集合間の分布を比較できるよう、割合を用いた高さの正規化を行う。以上により、データ分布の比較および特徴的部位の発見を支援している。

本論文では、Blade Graph を量的変量の表現として採用した多変量データセットの視覚的分析ツールを開発した。開発した分析ツールは、特徴的部分集合を探索する起点や補助となる情報として、多変量データセットが持つ変量間の関係性及びレコード間の関係性を可視化して提示する。本分析ツールはこれらの表現に加えて、部分集合を比較するためのインタラクティブな分析機能を備えている。

開発した視覚的分析ツールのケーススタディとして、ソーシャルネットワークサービスの一つであるブログの解析データから集団の特徴を分析した。本ケーススタディでは、特定の事象に対する興味を持つブログユーザに対して、さらに細かな部分集合に細分化してそれぞれの特徴を調査した。部分集合間の差異を発見しながら調査を進めることにより、各部分集合における特徴の顕在化に成功した。

今回はブログを対象とした分析を行ったが、本分析ツールはショッピングの購買履歴データや気象データ、医療データを始めとした、その他の多変量データセットにも適用可能である。今後は多種多様な多変量データセットの分析を行うことにより、開発した視覚的表現および分析ツールの有用性を追求したい。

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。研究チームのゼミや研究相談を通じて認識の誤りや改善点などを何度もご指摘いただくことにより、研究を順調に進めることができました。先生の丁寧なご指導のおかげで研究が順調に進み、無事に修士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。

田中二郎教授、志築文太郎准教授、高橋伸准教授、嵯峨智准教授、Simona Vasilache 助教には、研究室のゼミを通じて多くのご指摘を頂きました。時には厳しいご意見を頂いたこともありましたが、それ以上に多くの助言を頂けたことを感謝しております。

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、日々のゼミにおいて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。

また株式会社富士通研究所からは、本研究のケーススタディにて使用したデータを提供して頂きました。研究を進めるにあたって、富士通研究所の皆様と何度も打ち合わせを行いながら、貴重なご意見を頂くことができました。誠に有難うございます。

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] Stanley Smith Stevens, “On the Theory of Scales of Measurement”, *Science*, Vol. 103, No. 2684, pp. 677–680, 1946.

[2] James J. Thomas and Kristin A. Cook, “A visual analytics agenda”, *IEEE Computer Graphics and Applications*, Vol. 26, No. 1, pp. 10–13, 2006.

[3] Alexander Lex, Hans-Jörg Schulz, Marc Streit, Christian Partl, and Dieter Schmalstieg, “VisBricks: Multiform Visualization of Large, Inhomogeneous Data”, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 17, No. 12, pp. 2291–2300, 2011.

[4] Ramana Rao and Stuart 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*, pp. 318–322, 1994.

[5] Manuel Freire, Catherine Plaisant, Ben Shneiderman, and Jen Golbeck, “ManyNets: an interface for multiple network analysis and visualization”, In *Proceedings of the 28th international conference on Human factors in computing systems*, pp. 213–222, 2010.

[6] Haruka Suematsu, Sayaka Yagi, Takayuki Itoh, Yosuke Motohashi, Kenji Aoki, and Satoshi Morinaga, “A Heatmap-Based Time-Varying Multi-Variate Data Visualization Unifying Numeric and Categorical Variables”, In *Proceedings of 18th International Conference on Information Visualisation*, pp. 84–87, 2014.

[7] Daniel B. Carr, Richard J. Littlefield, Wesley 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.

[8] Niklas Elmqvist, Pierre Dragicevic, and Jean-Daniel 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.

[9] Hiroaki Kobayashi, Kazuo Misue, and Jiro Tanaka, “Colored Mosaic Matrix: Visualization Technique for High-Dimensional Data”, In *Proceedings of 17th International Conference on Information Visualisation*, pp. 373–383, 2013.

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

[11] Heike Hofmann, Arno P.J.M. Siebes, and Adalbert 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*, pp. 227–235, 2000.

[12] Alfred Inselberg, “The plane with parallel coordinates”, *The Visual Computer*, Vol. 1, No. 4, pp. 69–91, 1985.

[13] Alfred Inselberg and Bernard Dimsdale, “Parallel coordinates: a tool for visualizing multi-dimensional geometry”, *Springer*, 1987.

[14] Julian Heinrich, John Stasko, and Daniel Weiskopf, “The Parallel Coordinates Matrix”, In *Proceedings of EuroVis 2012 Short Papers*, Vol. 31, No. 3, pp. 37–41, 2012.

[15] Fabian Bendix, Robert Kosara, and Helwig Hauser, “Parallel Sets: Visual Analysis of Categorical Data”, *IEEE Symposium on Information Visualization*, pp. 133–140, 2005.

[16] Zhao Geng, ZhenMin Peng, Robert S. Laramee, Rick Walker, and Jonathan 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.

[17] Gennady Andrienko and Natalia Andrienko, “Parallel coordinates for exploring properties of subsets”, In *Proceedings of Second International Conference on Coordinated and Multiple Views in Exploratory Visualization*, pp, 93–104, 2004.

[18] Alexander Lex, Marc Streit, Christian Partl, Karl Kashofer, and Dieter Schmalstieg, “Comparative analysis of multidimensional, quantitative data”, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 16, No. 6, pp. 1027–1035, 2010.

[19] Lenka Nováková and Olga Štěpánková, “Multidimensional clusters in RadViz”, *SMO’06 Proceedings of the 6th WSEAS International Conference on Simulation, Modelling and Optimization*, pp. 470–475, 2006.

[20] John Sharko, Georges Grinstein, and Kenneth 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.

[21] Mike Sips, Boris Neubert, John P. Lewis, and Pat Hanrahan, “Selecting good views of high-dimensional data using class consistency”, *IEEE-VGTC Symposium on Visualization*, Vol. 28, No. 3, pp. 831–838, 2009.

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

[23] Micheline Elias, Marie-Aude Aufaure, and Anastasia Bezerianos, “Storytelling in Visual Analytics Tools for Business Intelligence”, *Human-Computer Interaction – INTERACT 2013*, pp. 280–297, 2013.

[24] Christophe Viau, Michael J. McGuffin, Yves Chiricota, and Igor 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.

[25] Jean-François Im, Michael J. McGuffin, and Rock Leung, “GPLOM: The Generalized Plot Matrix for Visualizing Multidimensional Multivariate Data”, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 19, No. 12, pp. 2606–2614, 2013.

[26] Samuel Gratzl, Nils Gehlenborg, Alexander Lex, Hanspeter Pfister, and Marc Streit, “Domino: Extracting, Comparing, and Manipulating Subsets across Multiple Tabular Datasets”, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 20, No. 12, pp. 2023–2032, 2014.

[27] Abish Malik, Ross Maciejewski, Niklas Elmqvist, Yun Jang, David S. Ebert, and Whitney Huang, “A Correlative Analysis Process in a Visual Analytics Environment”, *IEEE Conference on Visual Analytics Science and Technology*, pp. 33–42, 2012.

[28] Johannes Kehrer, Harald Piringer, Wolfgang Berger, and Eduard M. Gröller, “A Model for Structure-Based Comparison of Many Categories in Small-Multiple Displays”, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 19, No. 12, pp. 2287–2296, 2013.

[29] Tuan Pham, Rob Hess, Crystal Ju, Eugene Zhang, and Ronald Metoyer, “Visualization of Diversity in Large Multivariate Data Sets”, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 16, No. 6, pp. 1053–1062, 2010.

[30] Johanna Schmidt, Eduard M. Gröller, and Stefan Bruckner, “VAICo: Visual Analysis for Image Comparison”, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 19, No. 12, pp. 2090–2099, 2013.

[31] Ben Shneiderman, “Extreme Visualization: Squeezing a Billion Records into a Million Pixels”, In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 3–12, 2008.

[32] Ying-Huey Fua, Matthew O. Ward, and Elke A. Rundensteiner, “Hierarchical Parallel Coordinates for Exploration of Large Datasets”, In *Proceedings of the conference on Visualization ’99*, pp. 43–50, 1999.

[33] David Feng, Lester Kwock, Yueh Lee, and Russell M. Taylor II, "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.

[34] Daniel A. Keim, Florian Mansmann, Jörn Schneidewind, and Hartmut Ziegler, "Challenges in Visual Data Analysis", In *Proceedings of 10th International Conference on Information Visualisation*, pp. 9–16, 2006.

[35] Nathan Yau, "Visualize This: The FlowingData Guide to Design, Visualization, and Statistics", *Wiley Publishing, Inc.*, 2011.

[36] Jock Mackinlay, "Automating the Design of Graphical Presentations of Relational Information", *ACM Transactions on Graphics*, Vol. 5, No. 2, pp. 110–141, 1986.

[37] Waqas Javed, Bryan McDonnel, and Niklas Elmqvist, "Graphical perception of multiple time series", *IEEE Transactions on Visualization and Computer Graphics*, Vol. 16, No. 6, pp. 927–934, 2010.

[38] Wikipedia contributors, "Lab color space", Wikipedia, The Free Encyclopedia, *http://en.wikipedia.org/w/index.php?title=Lab_color_space&oldid=638454467*, accessed January 3, 2015.

[39] Wikipedia contributors, "Illuminant D65", Wikipedia, The Free Encyclopedia, *http://en.wikipedia.org/w/index.php?title=Illuminant_D65&oldid=635892229*, accessed January 3, 2015.

[40] Wikipedia contributors, "SRGB", Wikipedia, The Free Encyclopedia, *http://en.wikipedia.org/w/index.php?title=SRGB&oldid=639636200*, accessed January 3, 2015.

[41] Michael Gleicher, Danielle Albers, Rick Walker, Ilir Jusufi, Charles D. Hansen, and Jonathan C. Roberts, "Visual Comparison for Information Visualization", *IEEE Transactions on Visualization and Computer Graphics*, Vol. 10, No. 4, pp. 289–309, 2008.

[42] Robert McGill, John W. Tukey, and Wayne A. Larsen, "Variations of Box Plots", In *Journal of The American Statistician*, Vol. 32, No. 1, pp. 12–16, 1978.

[43] Jerry L. Hintze and Ray D. Nelson, "Violin Plots: A Box Plot-Density Trace Synergism", In *Journal of The American Statistician*, Vol. 52, No. 2, pp. 181–184, 1998.

[44] David Arthur and Sergei Vassilvitskii, "k-means++: The Advantages of Careful Seeding", In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035, 2007.

[45] Hiroaki Kobayashi, Tadanobu Furukawa, and Kazuo Misue, “Parallel Box: Visually Comparable Representation for Multivariate Data Analysis”, In *Proceedings of 18th International Conference on Information Visualisation*, pp. 183–188, 2013.