統計数理(2007)
第 55 巻 第 1 号 47–68

特集「統計データの可視化」
［原著論文］

Textile Plot 環境

熊坂 夏彦[1] ・柴田 里程[2]

（受付 2006 年 8 月 1 日；改訂 2007 年 1 月 12 日）

要　　旨

本論文では，汎用な視覚化手法 Textile Plot を用いることによって，高次元データの的確な認識を可能にする一つの体系的な環境の設計と実装について述べる．Textile Plot は，平行座標プロットを基本としているが，各軸の位置と尺度を適切に変換し，同時にデータのさまざまな属性をプロットと有機的に結びつけることで，汎用なヴィジュアリゼーションを可能にする．Textile Plot 環境は，Data，Parallel Coordinate，Visual Analogue，Textile Plot の 4 つのオブジェクトの流れで構成され，ユーザの視覚的操作を体系的に処理する．Data オブジェクトは座標ベクトルの集まりである Parallel Coordinate オブジェクトに変換され，Parallel Coordinate オブジェクトは Textile Plot とは相似であるがその抽象表現である Visual Analogue オブジェクトに変換される．さらに Visual Analogue オブジェクトは現実のディスプレイ，すなわちサイズや解像度に依存しない Textile Plot オブジェクトへと変換される．ユーザは最終的にこの Textile Plot オブジェクトをさまざまなインタフェイスを通じて眺めることになる．このようにデータからの Textile Plot の生成を一連のオブジェクトの変遷とみなすことで，対話的な操作を各オブジェクトに適切に振り分け，同時にそのログを保存することも可能となる．またこの環境設計は特定のデータや視覚化手法に依存しない汎用なものであるが，実際のデータを理解するうえで必要十分な環境であることも，いくつかの実データを用いて明らかにする．

キーワード： 平行座標プロット，最適な位置と尺度の変換，データの属性記述，視覚的操作，Information Visualisation.

1. 概要

計算機およびネットワークの発展に伴って取得されるデータは多様化し大規模化しているが，単に蓄積されるだけで有効に利用されていないケースも多い．その一つの理由として，大規模データ全体の様子を探る汎用で適切な視覚化手法の不足が挙げられる．Textile Plot（Kumasaka and Shibata, 2006a, b）は，大規模データのなかでも，特に変量数，つまり計測項目や調査項目が膨大で，従来の手法ではその様子を伺い知ることのできないような高次元データの的確な認識をサポートするために考案された．Textile Plot は平行座標プロット（Inselberg, 1985; Wegman, 1990）を基本にしているが，各軸の位置と尺度をどの観測パスも可能な限り水平になるよう選択することで数値データだけでなく類別データや順序付き類別データまで平行座標上に汎用的にしかも効果的に表示することができる．またデータタイプなどデータのもつ様々な属性を適切に処理し表示するだけでなく，データの重複度をプロットする円の面積に対応させるなど，

[1] 慶應義塾大学大学院 理工学研究科：〒223–8522 神奈川県横浜市港北区日吉 3–14–1
[2] 慶應義塾大学 理工学部数理科学科：〒223–8522 神奈川県横浜市港北区日吉 3–14–1

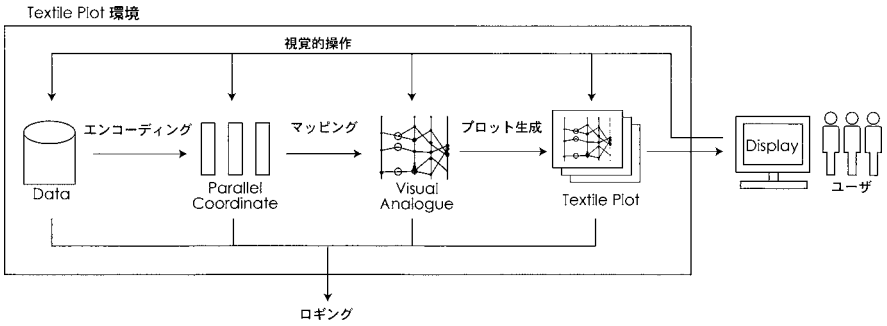

図 1. Textile Plot 環境.

データの意味にまで踏み込んだ表示が特徴である.

しかし Textile Plot をただ眺めているだけではデータのすべてを理解できるとは限らない. データを対話的に様々な角度から眺めることで初めて Textile Plot の威力を発揮させることができる. そのための環境が Textile Plot 環境 である. この環境は Information Visualisation のための一つの標準的なモデル (Card et al., 1999) にもとづいているが, ユーザからの指示が環境を構成する一連のオブジェクトに振り分けることまで考えているという点が大きく異なる.

図 1 は Textile Plot 環境を構成するオブジェクトとそれに対するユーザの視覚的操作を表している. 右向きの矢印は以下のような Data からの Textile Plot へ至る 4 つのオブジェクト間の変遷を表し, 上部の下向きの矢印は各オブジェクトに対するユーザの視覚的操作を表している. そして, 各段階で生成されたオブジェクトに対する視覚的操作は各オブジェクトの適切な変容を促し, 同時にそれはログとして出力される.

Data. Textile Plot 環境の最終目標はデータからもっとも適切なモデルを創造する快適な環境を与えることであり, Data オブジェクトの本体は最も一般的なデータ形式の一つである関係形式となっている. また Textile Plot を生成するのに必要十分な属性情報も保持している.

Parallel Coordinate. 関係形式の列ベクトルであるデータベクトルそれぞれの位置と尺度を一つの基準で定めた座標ベクトルの集まりである. この段階で類別データも数値化される. また Data オブジェクトのすべての属性情報は Parallel Coordinate オブジェクトに継承される.

Visual Analogue. Textile Plot オブジェクトと相似であるが, 具体的な描画ではなくその抽象表現である. このような中間段階をおくことによって, 単なる見栄えの違いは Textile Plot オブジェクトに担わせ, しかもフォーカスやハイライトなど Textile Plot の一部だけに注目するときなどは, Parallel Coordinate オブジェクトまで戻らずに, このオブジェクトから出発して再表示することが可能となる.

Textile Plot. Visual Analogue オブジェクトを具体的な色や幅を持つ点や線で表示したものが Textile Plot オブジェクトである. ただし実際に画面に表示されるプロットそのものではなく, 物理的な画面あるいは印刷面の制約によってその一部しか表示できないかもしれないことを前提としたオブジェクトである. このように設計することで, たとえば画面表示の段階で縦横のスクロールバーを設置することで, きわめて多数の変量の Textile Plot をスクロールしながら眺めるといったことも可能となる. つまり, 同一の Textile Plot オブジェクトをさまざまな外部環境を駆使して眺めることを可能にしている.

このように, データから Textile Plot の生成までをオブジェクトの変遷としモデル化すれば,

・ディスプレイを通したユーザの指示がどのオブジェクトに対するものか明確に区分することができ，各オブジェクトに対する必要十分な操作の定義が可能である，
・オブジェクトの変遷をそのままデータ解析のログとして保存することが可能である，
・実際のディスプレイを外部環境とみなすことで，データの規模にとらわれない汎用な環境設計が可能である，
・このモデルは Textile Plot に限らず，さまざまなヴィジュアリゼーションの環境構築に適応可能な汎用な設計である．

2. Textile Plot 環境を構成するオブジェクト

以下では，図 1 で示した流れに沿って Textile Plot 環境を構成するオブジェクトの説明を与える．なお概要ですでに述べたように，この流れで最終的に生成される Textile Plot オブジェクトは，コンピュータのディスプレイあるいは紙面といった物理的媒体を通してはじめて人間に伝わる．具体的には，縦横のスクロールバーを用いて，画面に収まる範囲を順にスクロールして，眺めるといった既存の手法を活用することになるが，これらは Textile Plot 環境が実装される外部環境に大きく依存し変化するのでここでは議論しないことにする．また，複数枚の Textile Plot を一枚のグラフィクスとしてまとめて表示したい場合なども，外部環境に頼ることになる．

2.1 Data オブジェクト

Data オブジェクトの本体はデータを保存する上でもっとも汎用な形式の一つである関係形式である．関係形式は行列に形式が似ているが，行と列はまったく異なる意味を持っている．行単位でみれば，各行は観測個体それぞれに関する属性なり計測値なりを表していると考えられるが，列単位で見れば，各列は変量の実現値の並びであるデータベクトルであると考えられる．つまり，行単位なら個体空間の点，列単位なら変量空間の点の集まりとみなすことができる．ただし，与えられた関係形式の各列は必ずしも同等に扱えるとは限らない．第一正規形を満たしていたとしても基数系を構成していたり，順序などの意味的な関連性があったりするからである．また個体の識別を与える「ID ベクトル」や各観測の重複数を表す「内訳ベクトル」は，データそのものというよりは各個体の属性とみなす方が自然であり，このような属性ベクトルはデータベクトルからは除外し，Textile Plot にその意味が適切に反映されるよう配慮する．

Data オブジェクトは関係形式データだけでなく，二つの重要な属性を保持している必要がある．一つは全体を説明する Title 属性であり，この属性は Visual Analogue オブジェクトまで Title 属性として継承される．もう一つは，各観測(行)が「何について」の観測なのかを表す Target Object 属性である．このような属性は，データベースの世界では，あまり表立って取り上げられないが，データを理解する上では，この属性は極めて重要な役割を果たす．Target Object 属性は本来の意味での母集団(Population)が何なのかを説明する属性と考えてもよいが，母集団という用語を使った場合，集団という言葉が各観測の対象が何かを表現する言葉としてはあまり適切でない．これが Target Object という用語を導入した理由である．

関係形式を構成する列ベクトルは，計測値などを表すデータベクトルと，各個体の属性を表す属性ベクトルからなることは先に述べた通りであるが，さらにデータベクトルは単なるベクトルというだけでなく，多くの属性や背景情報をもつ．すでに Kumasaka and Shibata (2006a, b) で述べたように，Textile Plot ではデータのさまざまな属性情報を座標軸に表示するため数値データであれば単位や値の範囲，類別データであれば水準名など，様々な属性情報を必要とする．そのような属性を以下に挙げる．

データタイプ. データベクトルは「計量値」,「序数値」,「計数値」,「順序付き類別値」,「順序なし類別値」,「論理値」に分類される．計量値，序数値，計数値はすべて数値型に属するが，データからその区別をすることは難しいため，あらかじめデータタイプとして与えておく必要がある．計量値は計測や観測を行った結果生じた値であり，通常は連続的な実数をとる．序数値は実験順序や観測時間などあらかじめ設定された値をとる．計数値は個数をカウントした結果であり非負整数値をとる．一方，非数値型は順序の存在する順序付き類別値と順序なし類別値にわかれる．さらに，順序なし類別値はその特殊な場合である論理値とも区別される．論理値はすでに定められた値のうち片方をとるという点で，単なる類別値とは区別されるべきである．さらなる分類も考えられるが，ここでは簡潔さを優先しこの 6 分類にとどめる．

取りうる値. 数値型ならば可能な数値の範囲，類別型ならば取りうる水準の集合である．データからは値として出現した範囲でしか知ることができないが，この情報があれば，果たしてどのような値が出現するのだろうか，絶対に取らない値はどのような値なのかなどを判断できる．

欠損および無効属性. 欠損はさまざまな理由でおきるため，それを統一的に扱うには，データベクトルの欠損の位置と種類を属性として持っている必要がある．また無効属性は，値が観測されたものの信頼性に問題があるような場合に，その値を表面的に欠損として扱う属性である．欠損同様にさまざまな種類が考えられるため，データベクトルにおける無効な値の位置と種類を属性として持っている必要がある．

単位. 数値型の場合，値の大きさを判断するには単位が必要である．数値データであれば，常識的な値の範囲があらかじめわかっているような場合もあるが，先入観によって思わぬ誤解をしたままデータ解析をすることを避けるためにも必ず表示するべき属性である．またモデル化の段階，そしてモデルを検証する段階でも，その物理量の意味や役割を解釈するために必要不可欠な属性となる．

2.2 Parallel Coordinate オブジェクト

Parallel Coordinate オブジェクトの本体は，Data オブジェクトを Kumasaka and Shibata (2006a, b)で述べたように適切な位置と尺度によって変換して得られた座標行列 $\mathbf{Y}=(\boldsymbol{\xi},\boldsymbol{y}_1,\ldots,\boldsymbol{y}_p)$ である．ここで $\boldsymbol{\xi}$ は各観測のパスをなるべく水平にするための基準位置となる基準位置ベクトル (Ideal Coordinate Vector)である．また，座標ベクトル(Coordinate Vector) $\boldsymbol{y}_j,\ j=1,\ldots,p$ は Data オブジェクトの本体である関係形式における各列ベクトルに対応している．

Parallel Coordinate オブジェクトは，Title 属性および Target Object 属性を Data オブジェクトから継承する．さらに，座標ベクトル $\boldsymbol{y}_j,\ j=1,\ldots,p$ それぞれも対応するデータベクトルが持つすべての属性を継承し，位置，尺度のパラメータ $\{\alpha_j,\boldsymbol{\beta}_j\},\ j=1,\ldots,p$ を新たな属性として保持する．

なお属性ベクトルである「内訳ベクトル」は，各観測の重みとして用いる．すなわち Kumasaka and Shibata (2006a, b)における二乗和 $S^2=\sum_{j=1}^{p}\|\boldsymbol{y}_j-\boldsymbol{\xi}\|^2_{\boldsymbol{w}_j}$ を観測パスの水平性の基準とするならば，重みベクトル $\boldsymbol{w}_j$ の各要素がこの内訳値ベクトルの要素倍される．内訳値は計数値としばしば混同されやすいが，内訳値はその総数があらかじめ定められており，各値はその内訳を個数または割合として表している点で計数値とは大きく異なる．従って内訳値ベクトルは，集約される前のデータにおける各観測の度数として扱う必要がある．

2.3 Visual Analogue オブジェクト

Parallel Coordinate オブジェクトを一足飛びに Textile Plot に変換することも可能であるが，Textile Plot の見栄えを変えるためだけに，Parallel Coordinate オブジェクトから Textile Plot オブジェクトを構成しなおすのは非効率的である．そこで，Textile Plot と相似な抽象表現で

ある Visual Analogue オブジェクトを導入し，まず Parallel Coordinate オブジェクトを Visual Analogue オブジェクトへマッピングする．Visual Analogue における色，形，大きさといった属性は抽象的に定義されるもので，具体的にどの色を割り当てるか，どの大きさで表示するかといったことは，次節で述べる Textile Plot オブジェクトの生成段階における各パラメータの設定によって定まる．Visual Analogue は以下に述べる Warp (経糸) と Weft (緯糸) という二つの要素からなる．Warp は座標ベクトルとその属性を表現し，Weft は各観測個体を表現する．

2.3.1 Warp

Warp は，前節の Parallel Coordinate オブジェクトの持つ座標ベクトルとその属性の抽象表現である．さまざまな型をもつ高次元データの全貌を一度につかむためには，各次元をなるべく簡潔にしかもその特徴が一目で理解できる表示が必要である．そのようなデータタイプごとの Warp の表現の例示が図 2 である．

これらデータタイプ別の Warp のデザインについては，すでに Kumasaka and Shibata (2006a, b) で詳しく述べているが，ここでの大きな違いは離散値データをさらに計数値と序数値に分離した点にある．これは，モデル化の段階で，計数値データと序数値データの扱いがおおきく異なるからで，たとえば，前者をポアソン分布でモデル化するのは自然でも，後者はそうではない．なお，Warp はその属性として軸のラベルと単位を持つ．

Textile Plot において，各座標ベクトルの要素は Warp における座標点を定める．これをノードと呼ぶ．すでに Wills (1996) で述べられているように，平行座標プロットにおいて重複するデータを効果的に視覚化するには座標点の重複度に比例した面積をもつ円によって表現するのがよい．Textile Plot ではすべてのデータタイプに関してその精度も考慮すれば，座標が重複することもあると想定し，たとえ数値データの場合でも，その面積が座標の重複度に比例した円によって座標点を表す．すなわち座標ベクトル $\boldsymbol{y}_j$ に対応するノードは，重複を除いた座標ベクトルの要素 y_{ij} の集合を $\{\tilde{y}_{ij};\ 1 \leq i \leq n_j\}$, $\tilde{y}_{ij}$ の重複度を l_{ij} としたとき

$$\{(\tilde{y}_{ij}, l_{ij});\ 1 \leq i \leq n_j\}$$

で表される．ここで，n_j は j 軸上で重複をのぞいた座標点の個数に相当する．

ノードは次節で述べるハイライトの段階で，ベース色，ハイライト色，シャドー色の三色に

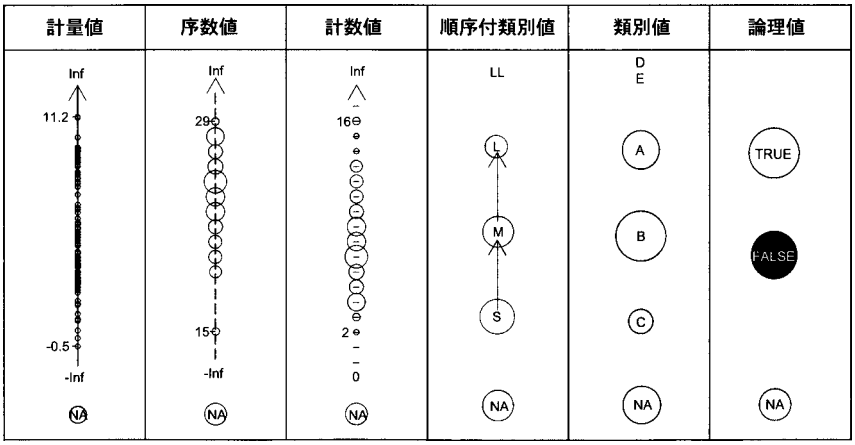

図 2. データタイプ別の Warp の表現.

よって，ハイライトされているノードとそうでないノードを区別し表現する．また観測個体数が非常に多い場合にノードの表示が込み入って認識できなくなることを防ぐために，ノードの透過率をあげて表示することも効果的である．

2.3.2 Weft

Weft は，各個体ごとに，Warp のノードを線分で結んでできる折れ線である．ノードと同様に，Weft に関しても重複度が存在するが，それは部分的な，つまり隣り合う Warp のノード間での重複度である．このような重複度を考慮した Weft の描画に関しては，平行座標プロットに対して Matthias (2003) で議論されているが，その重複度に比例した幅を持つ線分を用いることで，類別データだけでなく，重複度の多い数値データの場合でもユーザに的確な認識を与えることができる．

Weft はノードと同様，次節で述べるハイライトの段階で，ベース色，ハイライト色，シャドー色の三色によって，ハイライトされている Weft とそうでない Weft を区別し表現する．また観測個体数が非常に多い場合に Weft の表示が込み入って認識できなくなることを防ぐために，Weft の透過率をあげて表示することも効果的である．

図 3 は，コペンハーゲンのある地域で 1960 年から 1968 年に渡り賃貸住宅の居住者 1680 人について意識調査を行った結果 (Cox and Snell, 1981) である．管理に関する意見がどの程度聞き入れられているか，その住宅に満足しているかどうか，居住している住宅のタイプ，隣人と接触があるかどうかといった 4 つの項目について調査されている．左図は Weft の幅を一様にした場合で，右図は幅を重複度に比例させたものである．このデータの場合，右図のほうが，左図よりもはるかにデータの様子を的確に表していると言える．この図において左端中央に表示されている「住居者」はこのデータにおける Target Object 属性である．Weft には Target Object 属性が座標ベクトルから継承されている．

2.3.3 Warp の順序

Visual Analogue オブジェクトにおける Weft を具体的に定義するには，まず Warp の順序が定まっている必要がある．平行座標プロットでは，軸の順序について明確な基準が存在しないが，Wegman (1990) はすべての隣り合う変量間の関係を $\lceil (p+1)/2 \rceil$ 枚の平行座標プロットで表示することを提案している．しかし，これでは高次元データ全体の視覚化には程遠い．Warp の適切な順序の選択は，高次元データを理解するうえできわめて重要である．

Kumasaka and Shibata (2006b) では座標ベクトルと基準位置ベクトルの差の二乗和による基準と，階層型クラスタリングのなかで特に平行座標プロットと親和性のある順序付き末端最小

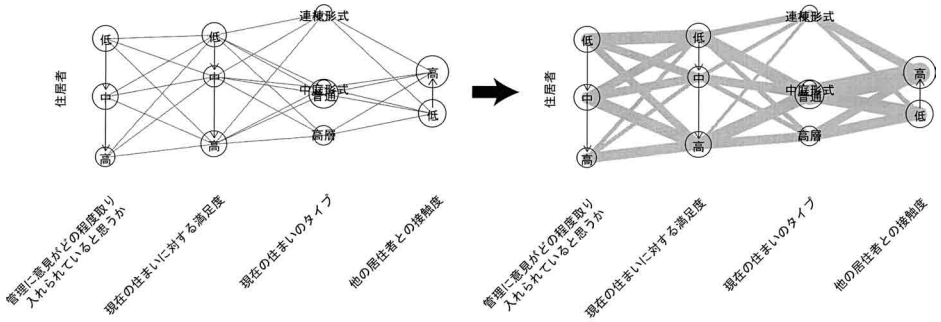

図 3. Weft の太さによる頻度の表示.

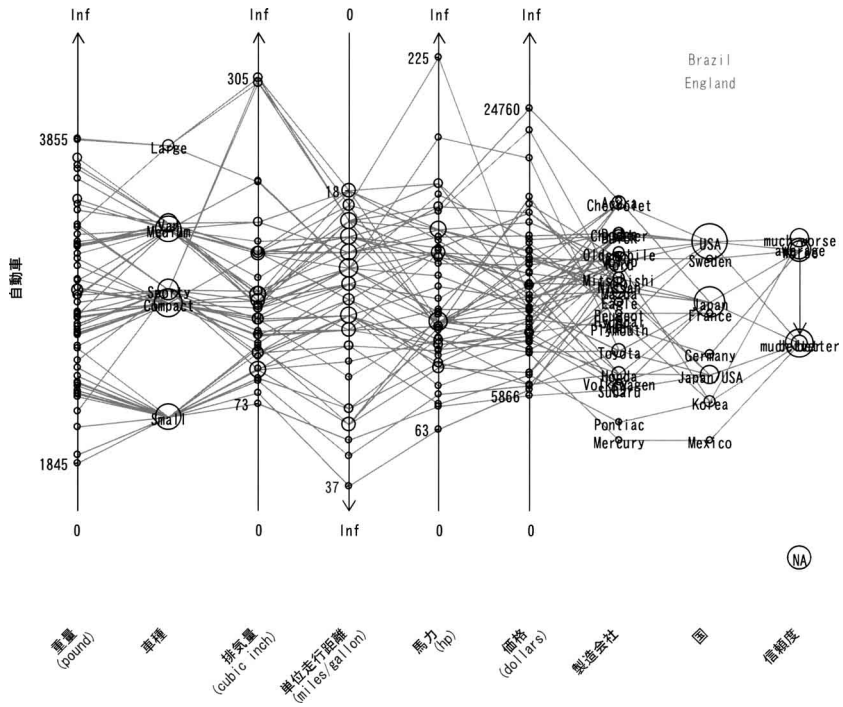

図 4. 自動車データ.

距離法(Hurley, 2004)を用いた座標ベクトルのクラスタリング基準の 2 つを提案している．どちらの基準も座標ベクトルが基本となっているため，Textile Plot 環境において Warp の順序は，Parallel Coordinate オブジェクトから Visual Analogue へマッピングされる段階で決定される．ここでは，その置換を σ で表すことにすると座標ベクトルは $\boldsymbol{y}_{\sigma(1)}, \boldsymbol{y}_{\sigma(2)}, \ldots, \boldsymbol{y}_{\sigma(p)}$ の順に並べかえられ Visual Analogue へマッピングされることになる．ここで，もし座標軸の順序が階層型クラスタリングによって決定されている場合は，クラスタ木を Warp の属性として持つ．もちろん，当初から変量間に自然な順序が存在する場合はそのような基準は不要であり，部分的にしか自然な順序が存在しない場合には，制約つきでその基準を用いることになる．

一例として，アメリカでの 1990 年 4 月の消費者レポートから取得された自動車データ(S-PLUS (株式会社数理システム，2006))の Textile Plot を掲げる．図 4 において，Warp は平均ベクトルとの差の二乗和が小さいものから順に左から並んでいる．この図からは，各自動車の差異が「重量」,「車種」,「排気量」などに現れることが直ちに読み取れる．また「単位走行距離」の座標軸の方向が，「排気量」や「馬力」と逆に向いていることも直感的に納得いく結果といえる．また最も右側に位置する自動車の「信頼度」からは，“much worse” から “much better” までの 5 水準からなる順序付類別データであるが，決して水準は等間隔ではなく，観測個体が信頼性の高い自動車と低い自動車の二群に分かれていることが読み取れる．

2.4 Textile Plot オブジェクト

Textile Plot オブジェクトは Visual Analogue にもとづいて生成される．図 5 で示す通り，Textile Plot は Title Region，Cluster Region，Plot Region，Weft Label Region，Warp Label Region の 5 つの領域からなる．Title Region にはデータから継承されたタイトル属性が表示される．Cluster Region はクラスタリングによって軸の順序が選択された場合に表示される．Warp Label Region には Warp の属性である各座標軸のラベルと単位が表示され，Weft Label Region には Weft の属性である Target Object 属性が表示される．Plot Region には Visual Analogue オブジェクトにおける Warp，Weft が表示される．このように領域を分割することで，領域ごとに異なるプロットを容易に実現することができる．なお，図 5 で用いた例は，前節で掲げた自動車データの Textile Plot である．ここでは，Warp の順序は順序付き末端最小距離法によるクラスタリング結果によって並べ替えられている．

表 1 は各領域の持つパラメータを表している．これらはいずれも Textile Plot の見栄えを変える役割を果たす．括弧内の数字は図 5 中の番号に対応する．すべてのサイズを指定するパラメータはすべて実スケール (point，pixel，cm など) で与えられる．なお Textile Plot オブジェクトの特徴は横幅をあらかじめ定めないことにある．これは，あらかじめ横幅を定めた上で，その中に収まるようにプロットしようとすると変量数が膨大なデータの場合には，すべての Visual Analogue を表示しきれないからである．したがって Plot Region の横幅は Warp の配置間隔と与えられた変量数によって決定される．また，Warp の配置間隔にあわせて Warp のラベルとクラスタ木が配置される．各領域の幅や高さのパラメータを 0 に設定することは当然その領域を非表示にすることになる．

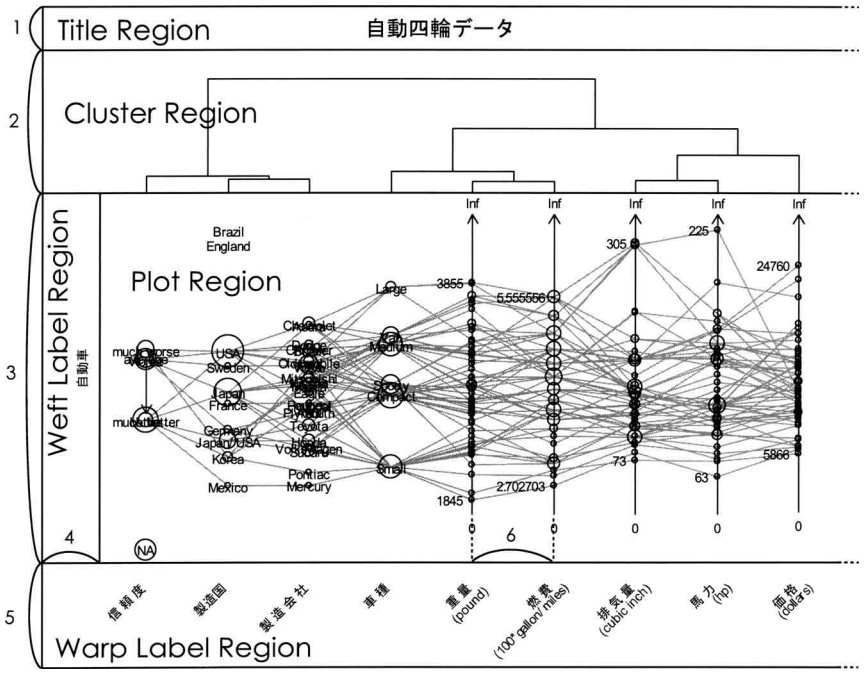

図 5. 領域の構成.

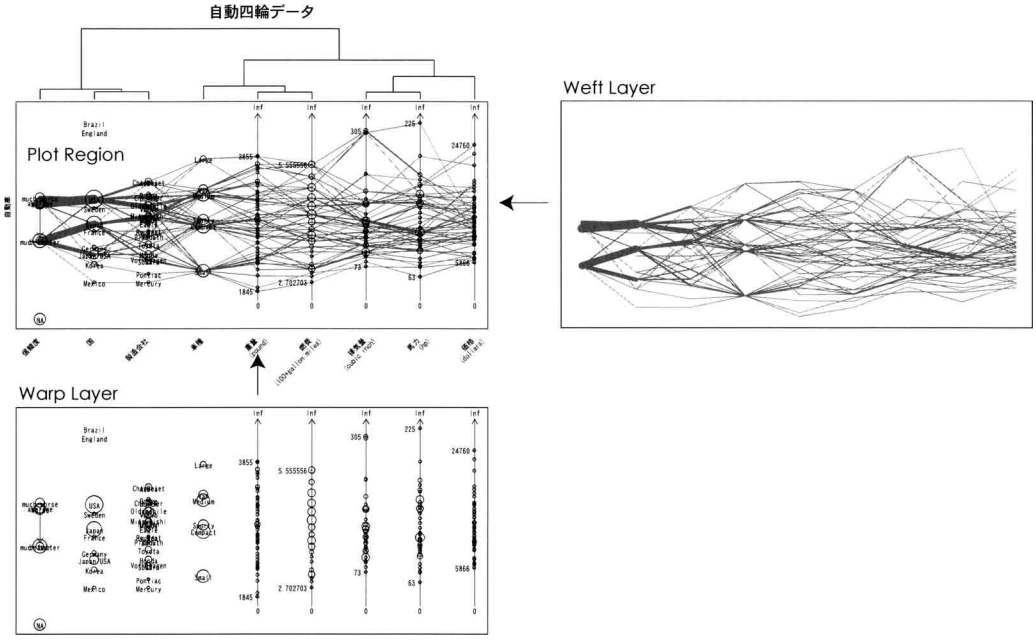

図 6. Plot Region におけるレイヤー構成.

Plot Region ではレイヤーの概念を導入し，図 6 で示すように Warp Layer，Weft Layer が重なり合って構成される．それぞれのレイヤーには Visual Analogue で定義した Warp，Weft がそれぞれ表示される．領域の幅や高さと同様にレイヤーの透過率を 0 にすることはレイヤーの非表示に相当する．領域と同様に，各レイヤーにも表 2 のようなパラメータが存在する．なおここに挙げたレイヤーの透過率は，レイヤー全体としての透過率であり，ノードと Weft 自体の透過率とは異なる．

3. オブジェクトに対する視覚的操作の振り分け

高次元データをさまざまな角度から眺めるために，古くからブラッシングやリンキング，座標軸の回転など，動的で対話的なディスプレイの研究が盛んに行われてきた（Cleveland and McGill, 1988）．Textile Plot 環境でも，人間の理解をさまたげない形の対話的な操作がスムーズに実行される必要があることはいうまでもない．最近では，リレーショナルデータベースのクエリによる操作ですら，視覚的な指示によって実現しようと試みられているからである（Presser, 2004）．本設計では，図 1 で示したようにデータからプロット生成までをオブジェクトの変遷とみなし，対話的な操作がどのオブジェクトに対するものかを区分し振り分けることで，体系的なソフトウェア構築を可能にしている．

ここではユーザがディスプレイ上でクリックやドラッグなどを行うことで与えられる操作指示という意味で，このような対話的な操作を視覚的操作とよんでいる．しかしユーザからの指示の具体的な受け取り方は，実装の環境に大きく依存するので外部環境にまかせ，Textile Plot 環境は受け取った視覚的操作を適切なオブジェクトに振り分け実行する．以下では，振り分けるオブジェクトごとにその操作を説明しよう．

表 1. 各領域のパラメータ.

Title Region	height	Title Region の高さ (1)
	color.background	Title Region の背景色
	font	フォントの種類
	font.size	フォントのサイズ
	font.color	フォントの色
Cluster Region	height	Cluster Region の高さ (2)
	color.background	Title Region の背景色
	color.cluster	クラスタ木の色
Plot Region	height	Plot Region の高さ (3)
	interval.warp	Warp の配置間隔 (6)
	color.background	Plot Region の背景色
	font	フォントの種類
	font.size	フォントのサイズ
	permutation.layer	レイヤーの順序
Weft Label Region	width	Weft Label Region の幅 (4)
	color.background	Weft Label Region の背景色
	font	フォントの種類
	font.size	フォントのサイズ
	font.color	フォントの色
Warp Label Region	height	Warp Label Region の幅 (5)
	color.background	Warp Label Region の背景色
	font	フォントの種類
	font.size	フォントのサイズ
	font.color.label	ラベルに使用されるフォントの色
	font.color.unit	単位に使用されるフォントの色

表 2. 各レイヤーのパラメータ.

Warp Layer	digits	目盛りの桁数
	characters	水準名の文字数
	color.warp	Warp を描画する際に使用されるデフォルトの色
	color.zerofrequency	頻度が 0 の水準のラベルの色
	color.logical	論理値データのラベルの色
	alpha.layer	レイヤーの透過率
	scale.radius	ノードの半径の倍率
	color.base	ノードを描画する際に使用されるデフォルトの色
	color.highlight	ハイライトされたノードに使用される色
	color.shadow	ハイライトの際に，もとのノードに使用される色
	alpha.node	ノードの透過率
Weft Layer	scale.width	Weft の幅の倍率
	color.base	Weft を描画する際に使用されるデフォルトの色
	color.highlight	ハイライトされた Weft に使用される色
	color.shadow	ハイライトの際に，もとの Weft に使用される色
	alpha.layer	レイヤーの透過率
	alpha.weft	Weft の透過率

3.1 Data オブジェクトに振り分けられる操作

Textile Plot 環境の目的は，Textile Plot を通してデータをさまざまな角度から眺めることで，データの均質な部分集合を抽出し，そこにゆるやかな変化をみいだすことにある．ひとたび，データに内在するゆるやかな変化を発見したならば，それをモデルとして記述することで，データの背後に存在する現象に迫ることができる．そのために必要なデータの変容を引き起こす操作は以下に挙げられる．

3.1.1 データの部分抽出

関係形式で保存されたデータの部分抽出は，変量あるいは観測個体のうち注目する部分を指

定することによって実現される．

実際，変量の部分抽出は，Plot Region において Warp の一部を選択するか，Cluster Region においてクラスタ木の枝を選択することで可能である．また観測の部分抽出は，ある Warp 上で値の範囲を指定することで可能である．図 7 は，Kumasaka and Shibata（2006a,b）で掲げたアヤメのデータの Textile Plot において，がく片の幅を除く変量の部分抽出（上段）と，種を Versicolor と Virginica の二つに制限した場合の観測個体の部分抽出（下段）を表している．

図 8 は，先に掲げた自動車データの「製造国」,「製造会社」,「信頼度」,「価格」を部分抽出し Textile Plot にしたものである．この図から信頼度と製造国，製造会社の関係は整緯（Parallel Weft）に近い関係であるのに対して，価格は結節（Knot）になっていることが読み取れる．Kumasaka and Shibata（2006b）において述べられているように，結節は与えられたデータベクトルの集まりの中で，あるデータベクトルが孤立している場合に生成される．それに対して，整緯は二つのデータベクトル間の完全な線形関係を意味している．この図からは，価格だけが他の変量か

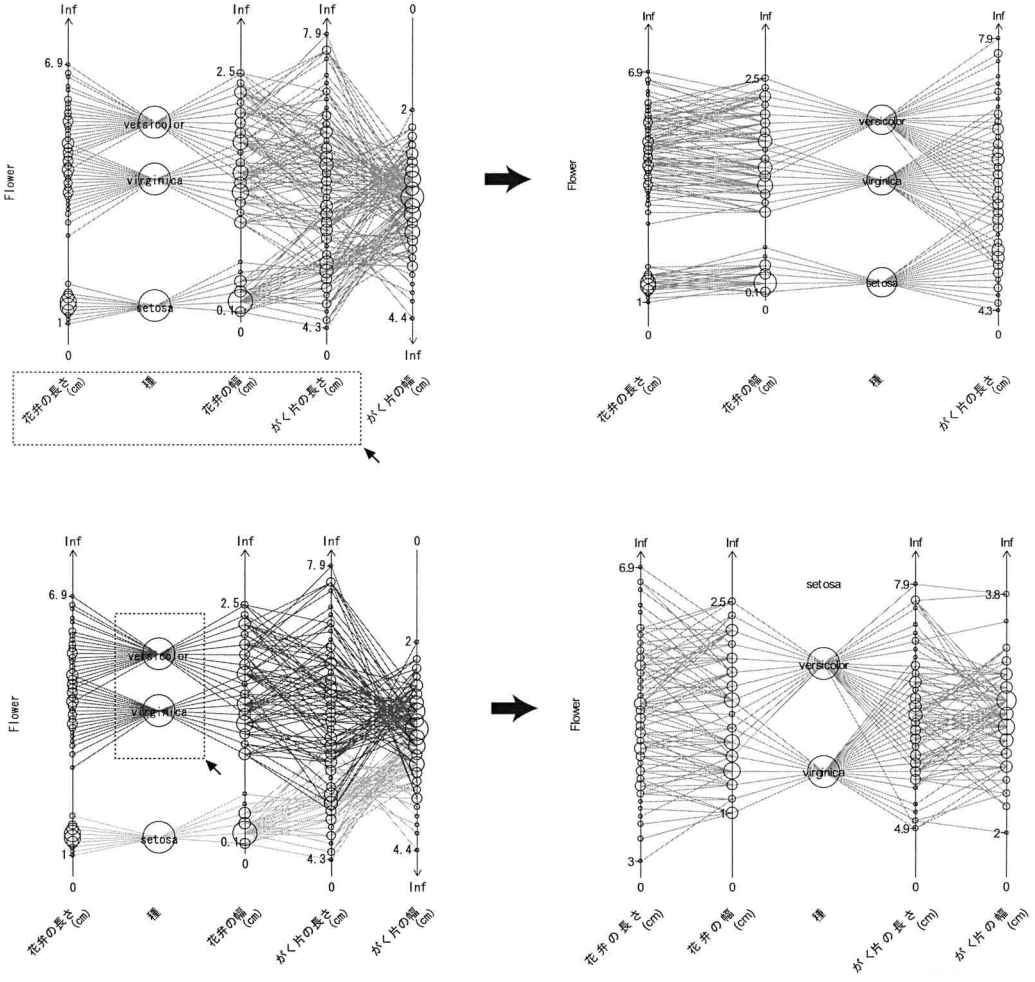

図 7. 観測項目の部分抽出（上段）と観測個体の部分抽出（下段）．

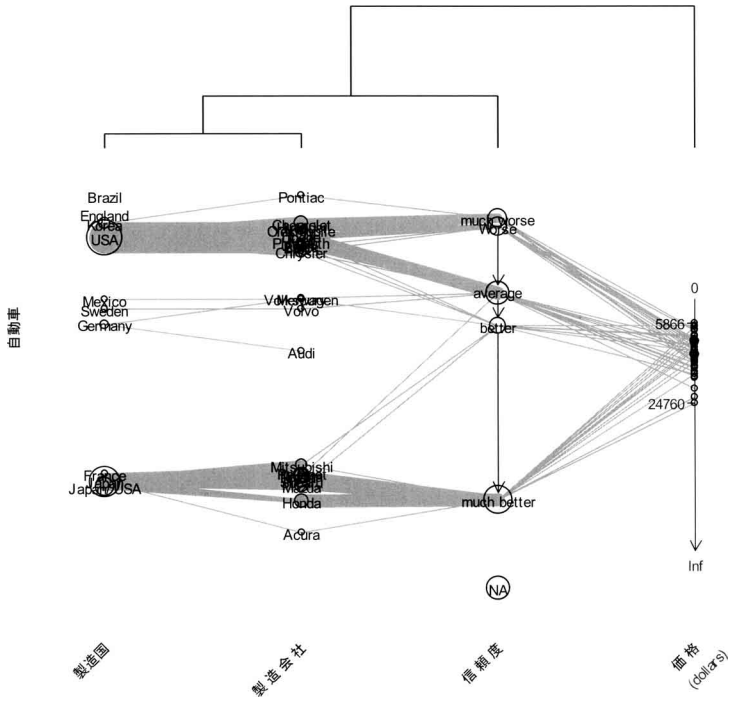

図 8. 自動車の価格と信頼性の関係.

ら孤立し，他の変量は水準間のマッピングが近いことが読み取れる．このように，一部の変量だけに注目した Textile Plot を描きなおすと，さらに明確な関係が見えてきたり，あるいは，一部の観測に制限した Textile Plot を描くことによっても，最初に描いた Textile Plot ではよくわからなかった側面が明らかになることが多い.

3.1.2 データの置換

データの値の置換は，諸刃の剣である．これを安易に許せばデータの改竄につながる．しかし，場合によっては，どうしても必要になることもある．たとえば，後に例で掲げる「鮎の釣獲尾数データ」の場合，出漁しなかった日の釣獲尾数は欠損値となっている．しかしこれを 0 と読み替えて，どのような Textile Plot となるか試してみる価値は十分にある．本環境では，このような Textile Plot 上のノードとして表される値の置換だけを許し，それをログとして記録する.

3.1.3 データの変換

ここで述べる変換とは次節で述べる軸の変換ではなく，データベクトル自体の変換をさす．データベクトルの変換には大きく分けて，一変量の変換と，多変量の変換がある．一変量の変換はデータの線形性がより自然に表現される変換であり，例えば対数変換，逆数変換などがあげられる．一例を挙げると，図 4 において「単位走行距離」の軸が他の計数値データと逆の方向性を持っていたが，ここで「単位走行距離」ではなくその逆変換である「燃費」をもちいれば，「車体の重量」や「馬力」との線形関係がより自然に表現されることが図 9 から読み取れる.

多変量の変換は，ある変量群に演算を加え新たな変量を生成する場合である．例としては，

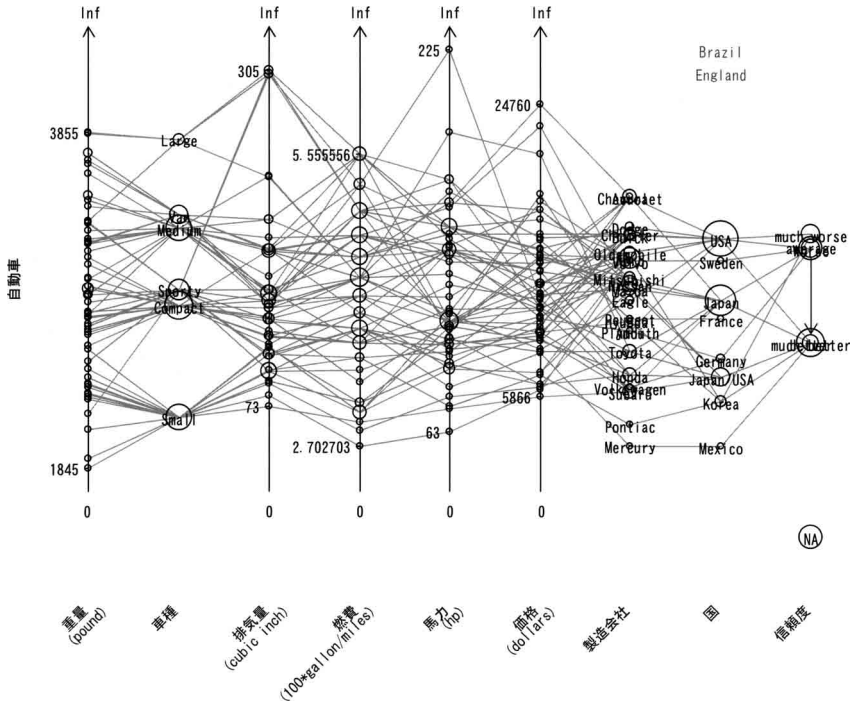

図 9. 燃費を含む自動四輪データ.

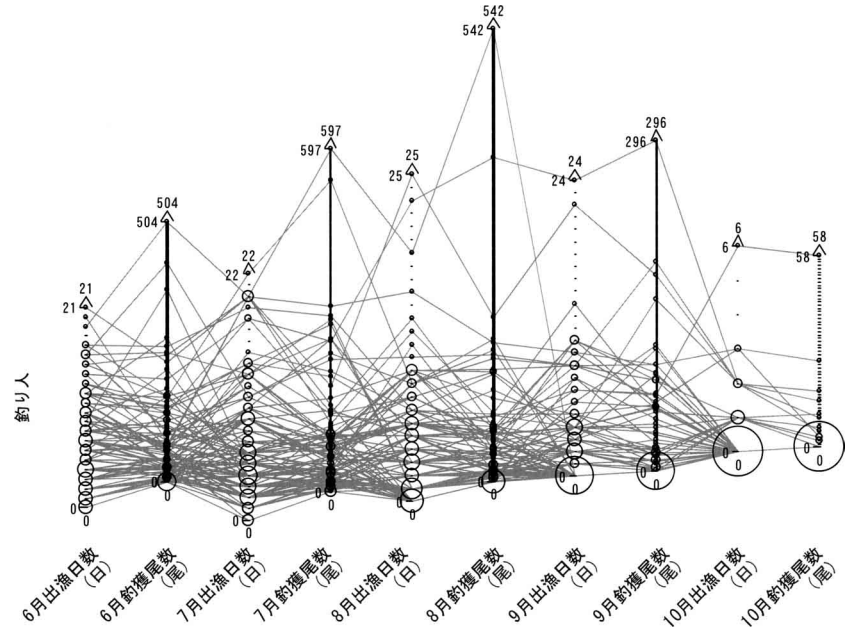

図 10. 鮎の釣果.

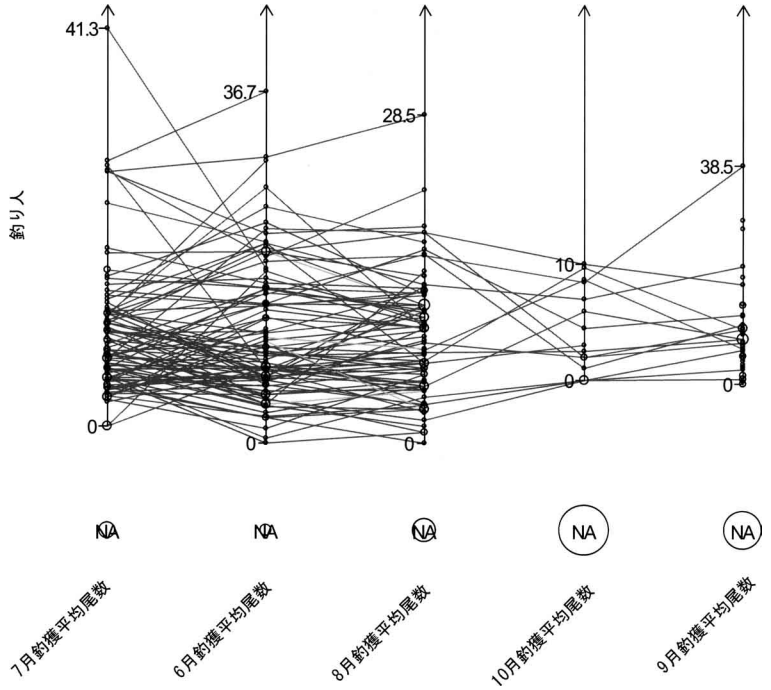

図 11. 鮎の一日平均釣果.

日付のように年月日が複数の変量に分かれているものを通日に直すような場合や，ある変量を別の変量で基準化するような場合が考えられる．その一例として，栃木県水産試験場が那珂川水系で 1993 年に実施した鮎の釣獲尾数データ(北田 他, 2001)の Textile Plot を掲げる．図 10 からわかる通り，このデータは各月の出漁日数と釣獲尾数が対になった 6 月から 10 月にかけての調査データであり，Target Object は釣り人である．

この Textile Plot からデータを理解することは難しい．明らかに釣り人の技術によって単位日数あたりの釣獲尾数は大きく異なるはずで，出漁日数と釣獲尾数の関係が線形になるとは考えにくい．実際に図 10 において，観測個体のパスが多数交差していることからもそれが読み取れる．そこで，各月の釣獲尾数を出漁日数で基準化し，単位日数あたりの釣獲尾数に変換した Textile Plot を眺めてみる(図 11)．いくつかの外れ値を除けば，観測のパスが水平に近いことが見て取れる．この Textile Plot からは，出漁した月と釣り人の効果の積による重み付ポアッソンモデルが示唆される．このように，変量群に演算を加え新たな変量を生成する操作は，その意味まで十分考慮して行えば，きわめて有効な操作である．

3.1.4 データの属性の変更

データに対する視点は，データの取得，データ解析の段階を経て，常に変化する．当然それに伴って，データの属性もユーザのデータに対する視点の変化によって変化する．そのような場合でもデータの属性をより適切なものに更新することで，より意味のある Textile Plot を得ることができる．

最も単純な例をあげるならば，順序付類別データと順序なしの類別データの扱いである．このデータタイプの違いを区別することは時として非常に難しく，順序類別データの各水準の軸

上での単調性よりも，各水準を取るグループの均質性に注目するのであれば，順序なしの類別データとして扱うことが自然である場合もある．

3.2 Parallel Coordinate オブジェクトに振り分けられる操作

Data オブジェクト自体は変化しないが，位置と尺度を選択する基準を変更したり，欠損値の扱いを変更したりすることは，Parallel Coordinate オブジェクト に対する操作となる．

3.2.1 水平性の基準の選択

最小化問題において，最小解が重複する場合や，大域的最小解がいくつかの局所最小解とほとんど変わらない場合には，異なる解，すなわち異なる水平性の基準で Textile Plot を眺めることも有効である．これは等質性分析や主成分分析などの言葉で言い換えるならば，第二主成分軸に関する個体のばらつきを眺めることに相当する．また最小化問題において二乗和を絶対値和に変更することで，異なる基準で位置と尺度の選択をすることなども選択の一つである．

3.2.2 属性ベクトルの反映の仕方の選択

Kumasaka and Shibata (2006a, b)では各観測をそろえる目標となる基準ベクトル $\boldsymbol{\xi}$ を自由に動かしたが，場合によっては $\boldsymbol{\xi}$ の動かす範囲を制限したほうがよい．たとえば，ある属性ベクトルによって観測が群に分けられるとき，同じ群に属する観測のパスはなるべく同一水平線上にそろえて表示したいといった場合がこれにあたる．具体的には，群に分ける指標の属性ベクトル $\boldsymbol{x}_{-r}, \boldsymbol{x}_{-r+1}, \ldots, \boldsymbol{x}_0$ が与えられたとき，線形空間

$$V = \{\alpha_0 \mathbf{1} + \sum_{k=-r}^{0} \mathbf{X}_k \boldsymbol{\beta}_k\} \subset \mathbb{R}^n$$

に $\boldsymbol{\xi}$ を制限することが考えられる．ただし，$\mathbf{X}_k$ は $\boldsymbol{x}_k$ をコーディングしたデータ行列である．このときの基準ベクトルの解は，二乗誤差を最小にするならば

$$\hat{\boldsymbol{\xi}} = \mathbf{X}(\mathbf{X}^T \mathrm{diag}(\boldsymbol{w})\mathbf{X})^{-1}\mathbf{X}^T \sum_{j=1}^{p} \boldsymbol{w}_j \cdot \boldsymbol{y}_j$$

で与えられる．ここに $\mathbf{X} = (\mathbf{1}, \mathbf{X}_{-r}, \mathbf{X}_{-r+1}, \ldots, \mathbf{X}_0)$, $\boldsymbol{w} = \sum_{j=1}^{p} \boldsymbol{w}_j$ である．また演算子「$\cdot$」はベクトルの要素積を表す．

3.2.3 欠損情報の処理の仕方の選択

欠損情報は単に観測しなかった欠損(unobservable)，もともと値がありえない条件での測定なり調査なりであった場合の欠測(never happen)，測定器の測定限界を超えたことによる欠損(happening)，回答拒否による欠損(refused to answer)，明らかに誤りの回答か値(sic value)などさまざまであり，これらタイプの異なる欠損が混在することもある．その扱いも，欠損を無視する，欠損を一つの水準として考える，各欠損を個別の水準として考えるなど，理由によってさまざまな選択がある．

3.3 Visual Analogue オブジェクトに振り分けられる操作

以下に述べるフォーカス，Weft のハイライト，Warp の移動は，どれも Parallel Coordinate オブジェクトは変化しないが，Visual Analogue オブジェクトが再構成される点で Visual Analogue オブジェクトに対する操作となる．

3.3.1 フォーカス

変量数が膨大な場合や，結節の周辺を詳細に調べるには，Textile Plot の一部をフォーカスする必要が生じる．これは，Visual Analogue の指定された部分に属する Warp，Weft を部分

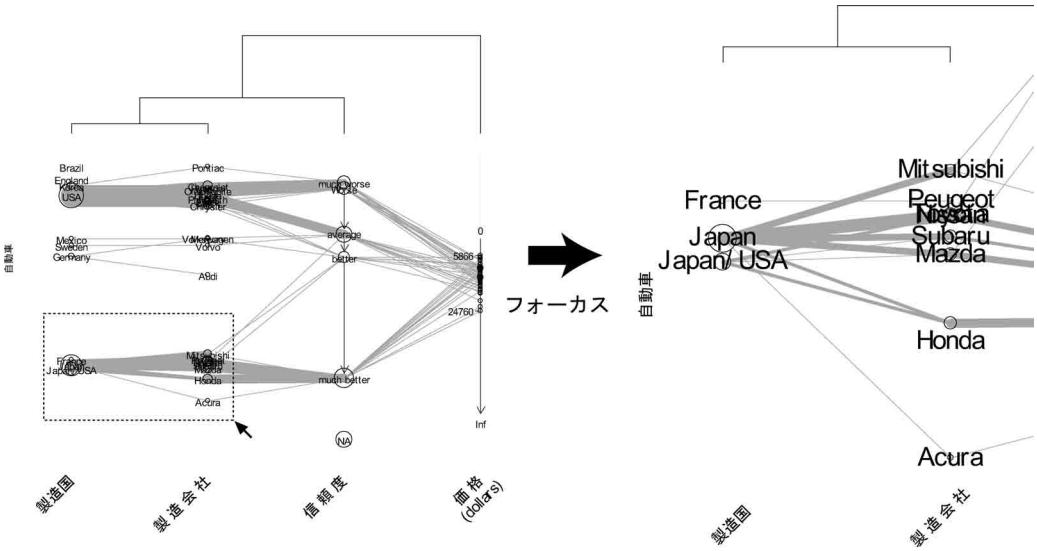

図 12. プロット領域のフォーカス.

的に再構成することにほかならない．当然，最終的に各領域に描かれる Visual Analogue の各要素は互いにリンクしており，プロット領域のフォーカスに伴って，対応するラベル領域とクラスタ領域も対応する部分だけが再描画される．

図 12 は図 8 において信頼度が高いとされる製造国と製造会社をフォーカスしたものである．この図から，日本の製造会社のうち特にホンダの自動車が 1990 年において非常に信頼度が高かったことを直ちに読み取ることができる．

3.3.2 Weft のハイライト

Textile Plot 環境におけるハイライトは，Parallel Coordinate オブジェクトにおける観測の部分集合を新たに Visual Analogue として構成し，重ね描きすることであり，Visual Analogue の変更が必要になる．Textile Plot におけるハイライトの対象は Weft である．Weft Layer において，矩形ラバーバンドに含まれる Weft が選択されることで，関連する Weft とノードがハイライトされる．ラバーバンドは一般的な集合演算 AND，OR，XOR，NOT によって二次元平面のあらゆる部分集合に含まれる Weft をハイライトすることができる．またこのラバーバンドをプロット上で動的に移動させることで，Weft のブラッシングも可能である．

図 13 は，先に掲げた賃貸住宅の調査データで，住居に満足していない住居者の観測をハイライトした結果である．この図からは連棟形式の住居に住んでいるほとんどの人がその住居に満足していないことが見て取れる．さらに，住居に満足していない住居者のうち連棟形式の住居に住んでいる観測をハイライトすることで(図 14)，他の住居者との接触が高く管理に意見が取り入れられない住居ほど，人は住みにくいと感じていることがわかる．

3.3.3 Warp の位置の変更

座標ベクトルの分散基準やクラスタリングによる軸の順序付けも，Visual Analogue に対する指示の一つでユーザに効果的な表示を提供するが，すべての場合においてよい結果をもたらすとは限らない．どのような距離法を用いるかの選択も可能とする必要がある．さらにユーザ

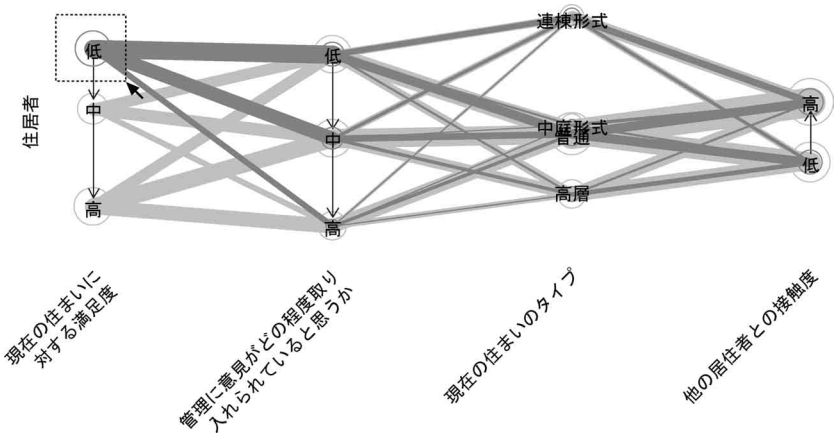

図 13. 住宅の満足度が低い Weft のハイライト.

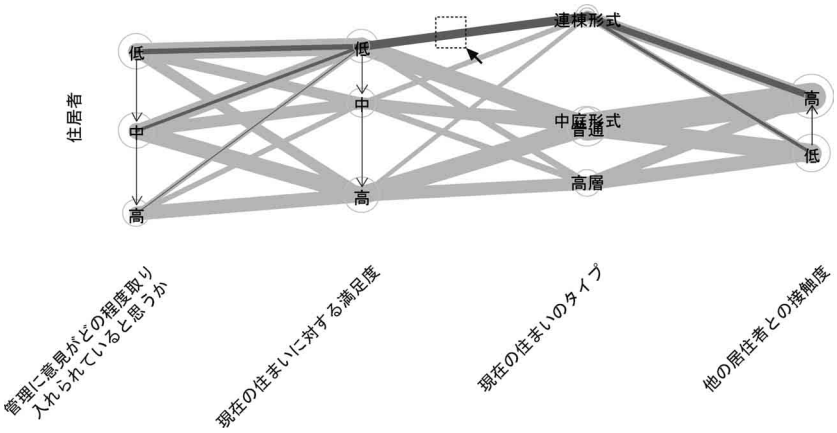

図 14. 住宅の満足度が低く連棟形式に居住している Weft のハイライト.

は Warp レイヤーにおいて，Warp の移動として軸の順序を自由に指定することができる必要がある．当然軸の並べ替えによって Visual Analogue オブジェクトは再構成される．図 15 は Kumasaka and Shibata (2006a,b) で掲げたアヤメデータの Textile Plot において，「花弁の長さ」に対応する Warp を「花弁の幅」と「がく片の長さ」の間に挿入した例を表している．

3.4 Textile Plot オブジェクトに振り分けられる操作

Textile Plot オブジェクトに振り分けられる視覚的操作は，2.4 節で掲げた領域およびレイヤーのパラメータの変更である．Textile Plot の見栄えの違いは Textile Plot オブジェクトのパラメータの値の違いに反映され，Visual Analogue オブジェクトは変化しない．

4. Textile Plot 環境 の実装

Textile Plot 環境 を実現するためには，データに関する属性情報をどのように扱うかが一つの鍵となる．既に Textile Plot を実現するエンコーディングの段階で，各データベクトルのデータ

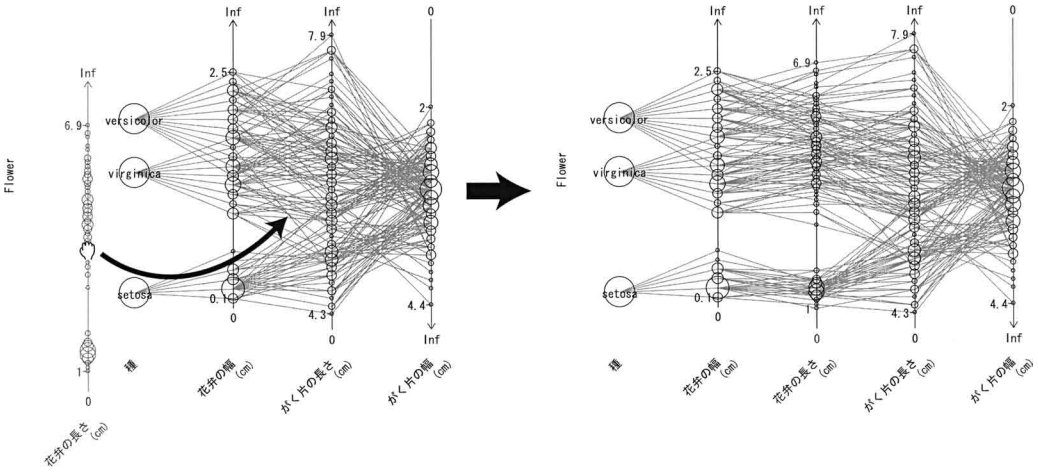

図 15. Warp の移動.

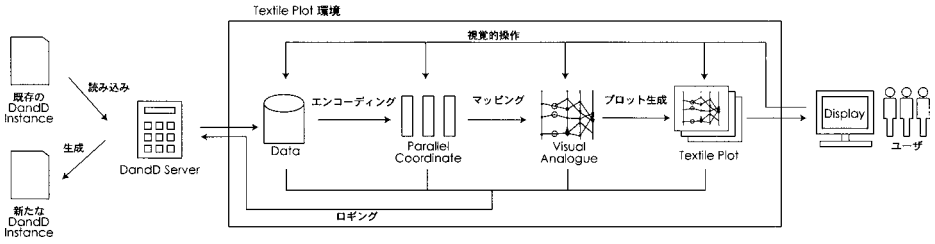

図 16. DandD を用いた Textile Plot 環境の実装.

タイプ属性が必要となるが，実際にそのような属性情報をデータベースやデータファイルそのものから得ることはできない．従ってあらかじめ，なんらかの形でデータの属性をデータと併せて保持しておく必要があるが，Metadata のように実際の Relational DataBase Management System に依存しているような場合は汎用性に乏しいといえる．またそのような記述が，データの変容まで統一的に記述できるかは疑問である．

DandD Project (2006) は，データとその属性を DandD インスタンスと呼ばれる XML (Extensible Markup Language) 文書にすることで，データと属性情報の汎用な記述を可能にした．さらにデータの変容やデータ解析の各段階において DandD インスタンスを生成することで，データの取得からモデル化までの道筋を記録することができる．また DandD では，インターデータベース (Shibata, 2004) の概念を導入することで，ネットワーク上に散在する異種のデータを，目的に沿って自由に利用することを可能にした．

図 16 は，DandD サーバクライエントシステム (Yokouchi and Shibata, 2004) を用いた Textile Plot 環境の実装を表している．ユーザは DandD サーバに対して DOM (W3C, 2006) の命令を送るだけで，Textile Plot を生成するために必要なデータとその属性情報をサーバから取得することができる．また，視覚的操作の内容にあわせて，今度はサーバー上に展開された DandD インスタンスの一部を更新し書き出せば，新たに DandD インスタンスが生成され，その蓄積

が視覚的操作の蓄積に相当する.

ただし，DandD クライエントとして一から Textile Plot を実装するのは多大な時間を要すので，今回は R Project (2006) が提供しているパブリックドメインのデータ解析ソフトウェア R 上に Textile Plot を実装した.

4.1 R 上での実装

以上のような Textile Plot 環境を検証するため，DandDR (熊坂・横内, 2003) の一部として R 上に試験的に実装をおこなった (図 17). R 上での実装の利点は，R が Windows，Linux，MacOS をはじめ様々な OS に移植されており多様なプラットフォームで利用可能な点にある．また R の豊富なグラフィクス関数は，Textile Plot 環境の外部環境として必要な要件を満たしているといえる．さらに，DandD クライエントプログラム DandDR がすでに実装されていることも，試験的な実装には好都合であった.

DandDR は DandD サーバと R のインタフェイスの一つであり，DandD インスタンスの主要な構造を `dad` オブジェクトとして R 上に展開する．そこで Textile Plot 環境を，このオブジェクトの `plot` メソッドとして実装した．なお，最適化問題を解く部分は C 言語の助けをかりている．一方，ユーザからの視覚的操作は `locator` 関数によって検知し，Data に対する変容とログは DOM 文に置き換え DandD サーバに送られる.

しかし，R 上での実装には制約も多く，ユーザインタフェイスの貧弱さのみならず，グラフィカルデバイスの能力がプラットフォームに大きく依存することも実装を制約する．また大規模データを扱うには，すべてのオブジェクトをメモリ上に展開する現在の R に依存した実装ではすぐに限界が訪れてしまう.

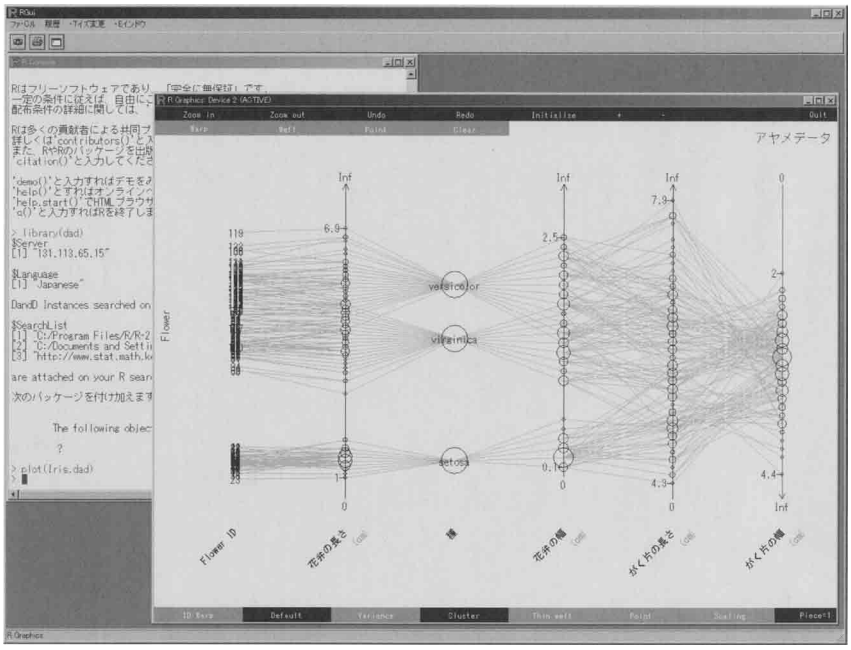

図 17. R 上での Textile Plot 環境.

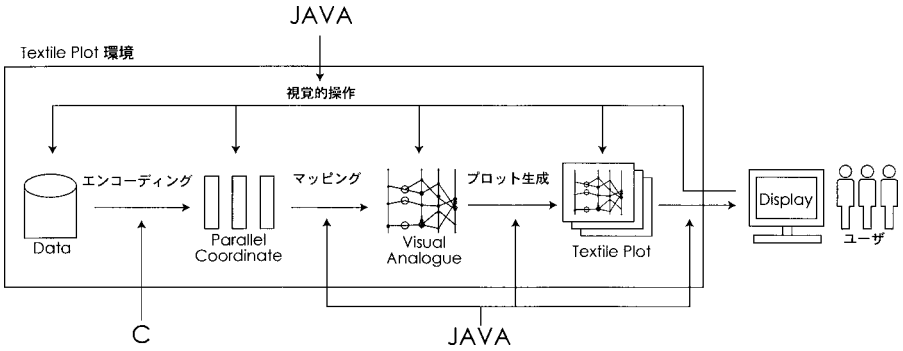

図 18. DandD+C+Java による Textile Plot 環境の実現.

5. まとめと今後の課題

Textile Plot は，平行座標プロットを基本としているが，各軸の位置と尺度を適切に変換し，同時にデータのさまざまな属性をプロットと有機的に結びつけることで，高次元データの的確な認識をサポートする汎用なヴィジュアリゼーションを可能にする．

本論文ではこの Textile Plot を用いて，与えられたデータをあらゆる角度から眺めることのできる Textile Plot 環境の設計と実装について述べた．この Textile Plot 環境では，データからの Textile Plot の生成を 4 つのオブジェクトの変遷とみなすことで，対話的な視覚的操作を各オブジェクトに適切に振り分け，結果としてプロット生成までの過程における冗長性を排除しながら，同時に適切なログを保存することも可能とした．またこの環境設計は特定のデータや視覚化手法に依存しない非常に汎用なものである一方，表示範囲の限られる現実のディスプレイを想定したフォーカスやハイライトなどの対話的なヴィジュアリゼーションもその環境設計に取り入れることで，実際のデータを理解するうえで必要十分な環境であることをいくつかの実データを用いて示した．

Textile Plot 環境の実装に関しては，R 上での実装にはさまざまな制約が伴うことが判明したので，現在は OS や特定のソフトウェアに依存しない独自な実装を計画している．その基本言語としては，アンチエイリアスや透過率など高度なグラフィクス環境とユーザインタフェイスを汎用に提供している JAVA 言語を採用し，高速化を図る部分は C 言語を併用することが望ましい(図 18)．このように独自な実装により，高次元大規模データの様子を探るための環境というだけではなく，さまざまな場面で，さまざまな利用が促進されるに違いない．

また Textile Plot 環境の今後の発展として，出力されたログの高度利用が挙げられる．ログはデータ解析の記録であり，このログを利用すればモデル化までの道筋が明確に把握でき，多くの人間と認識を共有することが可能になる．認識の共有は同時に問題意識も啓発し，モデルの再評価，再構築，現場へのフィードバックとその果たす役割は大きい．また同時に，データの変容の記録は，データに対してどのような処理がおこなわれたかの監査にも利用でき，データ解析の正当性を保証することにもつながる．

謝　辞

本研究は，21 世紀 COE プログラム「統合数理科学：現象解明を通した数学の発展」の支援による．

参 考 文 献

Card, S. K., Mackinlay, J. D. and Shneiderman, B. (1999). *Information Visualization*, Morgan Kaufmann Pub., San Francisco, California.

Cox, D. R. and Snell, E. J. (1981). *Applied Statistics*, Chapman and Hall, New York.

Cleveland, W. S. and McGill, M. E. (1988). *Dynamic Graphics for Statistics*, Wadsworth & Brooks/Cole Advanced Books & Software, California.

DandD Project (2006). Dand Project Home Page, http://www.stat.math.keio.ac.jp/DandD/.

Hurley, C. (2004). Clustering visualizations of multidimensional data, *Journal of Compuatational and Graphical Statistics*, **13**, 788–806.

Inselberg, A. (1985). The plane with parallel coordinates, *The Visual Computer*, **1**, 69–91.

株式会社数理システム (2006). S-PLUS, http://www.msi.co.jp/.

北田修一，神保雅一，田中昌一，宮川雅巳，三輪哲久 (2001).『データサンプリング』，共立出版，東京.

Kumasaka, N. and Shibata, R. (2006a). Implementation of Textile Plot, *Proceedings in COMPSTAT 2006*, Physica-Verlag, Rome.

Kumasaka, N. and Shibata, R. (2006b). High dimensional data visualisation: The Textile Plot, *Computational Statistics and Data Analysis* (submitted).

熊坂夏彦，横内大介 (2003). DandD とデータ解析ソフトウェア R, 2003 年度統計関連学会連合大会報告集，368–371.

Matthias, S. (2003). Visualizing categorical data arising in the health sciences using hammock plots, American Statistical Association, CD-ROM.

Presser, C. (2004). A database schema for constructing visual queries, *Proceedings of the International Conference on Modeling, Simulation and Visualization Methods*, 135–142, CSREA Press, Las Vegas, Nevada.

R. Project (2006). R Project Home Page, http://www.r-project.org/.

Shibata, R. (2004). Interdatabase and DandD, *Proceedings in COMPSTAT 2004*, Physica-Verlag, Prague.

W3C (2006). Document Object Model Home Page, http://www.w3.org/DOM/.

Wegman, E. (1990). Hyperdimensional data analysis using parallel coordinates, *Journal of the American Statistical Association*, **85**, 664–675.

Wills, G. J. (1996). Selection: 524,288 ways to say this is interesting, *Proceedings of the 1996 IEEE Symposium on Information Visualization*, 54–60, IEEE Computer Society, Washington.

Yokouchi, D. and Shibata, R. (2004). DandD: Client server system, *Proceedings in COMPSTAT 2004*, Physica-Verlag, Prague.

 Proceedings of the Institute of Statistical Mathematics Vol. 55, No. 1, 47–68 (2007)

The Textile Plot Environment

Natsuhiko Kumasaka[1] and Ritei Shibata[2]

[1]Graduate School of Science and Technology, Keio University
[2]Department of Mathematics, Keio University

The textile plot proposed by Kumasaka and Shibata (2006a, b) is a powerful tool for visualising high dimensional data. It is a modified parallel coordinate plot, where the locations and scales of each axis are simultaneously chosen so that all connecting lines, each of which signifies an observation, are aligned as horizontally as possible. A main theme of this paper is how to design an ideal environment for working with data through the textile plot. To meet various needs of working with data, the environment has to be as flexible as possible. A reference model for achieving this goal consists of a sequence of four objects; the data, the parallel coordinate, the visual analogue and the textile plot objects. A data object is transformed into a parallel coordinate object, which is a set of coordinate vectors. The visual analogue is an abstract representation of the textile plot produced. The textile plot object is a textile plot but constructed without any restriction in the real world like size of display or resolution. The user can view this object through various interfaces like zooming or resizing. Visual instructions given by the user are sent to one of the objects according to its own nature. A by-product of the design is to enable us to keep a log of visual instruction with the user in a systematic way. It can be used not only for auditing but also for helping the user to construct an appropriate model for the data.

Key words: Parallel coordinate plot, choice of locations and scales, data and attributes, visual operation, information visualisation.