DEIM2022 J23-1(day2 p41)

PCPと言語表現を組み合わせた多変量データ分析支援システムの拡張

能條　太悟†　　稲澤　朋也††　　富井　尚志†††

† 横浜国立大学大学院環境情報学府情報環境専攻　〒240–8501 横浜市保土ヶ谷区常盤台 79–7
†† 横浜国立大学理工学部数物・電子情報系学科　〒240–8501 横浜市保土ヶ谷区常盤台 79–5
††† 横浜国立大学大学院環境情報研究院　〒240–8501 横浜市保土ヶ谷区常盤台 79–7
E-mail: †nojo-daigo-gf@ynu.jp, ††inazawa-tomoya-xy@ynu.jp, †††tommy@ynu.ac.jp

あらまし　我々の先行研究では，多変量データを可視化し，その状態を SQL に類似した言語表現により保存・再現可能なシステムを提案してきた．このシステムの特徴として，多変量データを平行座標プロット（PCP：Parallel Coordinates Plot）を用いて可視化し，PCP 上でデータに対してアドホックなクエリの実行が可能である．データ操作の過程において，データ分析者はデータ操作や可視化の状態を SQL ライクな言語表現により保存・再現することが可能である．これにより，解析過程における試行錯誤のプロセスを支援する．このシステムは主にデータ可視化部とデータ操作部から構成される．我々はこのシステムを $(PC)^2$DV （Parallel Coordinates Plot Commutative Data Visualizer）と名付けた．本稿では，$(PC)^2$DV の機能拡張を行う．１つ目に，データロード部を追加する．これは，システムの外部にあるデータソースに ODBC 接続し，可視化対象のデータを取得する役割を持つ．２つ目に，SQL に存在する GROUP BY 句および集約演算の操作を追加する．可視化の対象であるデータは一般に粒度が細かく，データ一件では意味をなさないことがある．そこで，特定の単位でデータをグループ化し分析を行うことは有用であると考えられる．また，拡張機能を活用し有用な情報を得ることのできるデータ解析例を示し，本システムを用いたデータ分析が有用であることを示す．

キーワード　情報可視化，データ操作言語，多変量データ分析，平行座標プロット

1 はじめに

近年，センサ技術の発達により大量のデータを容易に取得できるようになった．それと同時に，ストレージ技術の発達および大容量化，低価格化により，取得した大量のデータの蓄積・保存が可能になった．この一例として，スマートフォンやウェアラブルデバイスなどの普及により，日常生活をデータとして記録する「ライフログ」を個人でも簡単に取得し蓄積することが可能になった．これらのデータを利用し，様々な目的のために分析を行うことが考えられる．しかしながら，取得したデータは一般に複数の属性からなるデータ，すなわち多変量データであり，かつデータ件数も膨大である．膨大な件数の多変量データから有用な情報を抽出することは，一般には容易であるとは言えない．そのため，多変量データの分析にはデータ可視化などを用いた分析者への技術的な支援が必要である．

多変量データを可視化する手法の一つとして，平行座標プロット (PCP：Parallel Coordinates Plot) [1,2] がある．PCP とは，多変量データの各属性をそれぞれ一つの軸として描き，データ一件をその軸上の点を結んだ一本の線として表す手法である．PCP の特徴として，一つのグラフでデータ全体を見ることや，データ全体のクラスタの偏りや属性間の相関を俯瞰的に把握することが可能である．

ここで，PCP 上では選択・結合・射影といった関係代数演算のようなデータ操作を GUI 上で表現することが可能である．このことは，多変量データを n-項関係として捉えると有効なデータ操作であると考えられる．この特徴に注目し，我々の先行研究では PCP を用いて多変量データを可視化し，その状態を保存・再現することが可能な分析支援システムを提案してきた [3–5]．このシステムは，主にデータ可視化部とデータ操作部から構成される．我々は，この分析支援システムを $(PC)^2$DV （Parallel Coordinates Plot Commutative Data Visualizer）と名付けた．$(PC)^2$DV は，多変量データに対してアドホックなクエリ実行が可能な，GUI を有するミドルウェアである．$(PC)^2$DV では PCP で可視化された多変量データに対し，PCP 上で選択・射影・結合といった基本的な関係代数演算によるデータ操作がアドホックなクエリとして実行可能である．分析者は，データ操作をすることやその操作結果を任意の手法で可視化することによりデータ分析を進める．その過程において，分析者はデータ操作やデータ可視化の状態を，我々が定義した，SQL に類似した言語表現である $(PC)^2$L（Parallel Coordinates Plot Commutative Language）の形式で保存することができる．また，過去の状態に戻りたい場合は保存をした $(PC)^2$L を 入力することでその状態が再現され，そこから再び分析を行うことができる．これにより，SQL を熟知する分析者に対して試行錯誤を伴うデータ分析の支援が可能となった．

本研究では，より汎用的に多変量データ分析の支援をするために，先行研究の $(PC)^2$DV に２つの機能拡張を行う．１つ目に，$(PC)^2$DV にデータロード部を追加する．これは，システムの外部にあるデータソースに ODBC 接続をし，可視化およ

び分析対象のデータを取得する役割を持つ．これにより，データソースの形式にかかわらず対象のデータを取得し，$(PC)^2$DV での分析をすることが容易になる．２つ目に，$(PC)^2$L の文法拡張として，SQL における GROUP BY 句および集約演算の操作を追加する．ここで可視化および分析対象としているデータは一般に粒度が細かく，データ一件のみでは意味をなさないことがある．そのため，特定の単位でデータをグループ化し，各グループに対して代表値を求める操作は有用であると考えられる．

本稿では，拡張機能を活用し有用な情報を得ることのできるデータ解析例を示し，$(PC)^2$DV を用いたデータ分析が有用であることを示す．

2 関連研究

2.1 PCP とデータ操作

PCP は 1985 年，Inselberg によって初めて概念が定義された [1]．それ以降，PCP に関する様々な議論がなされている．Johansson らによれば，PCP の研究カテゴリーは次の 4 つに分類される [2]：

（1） PCP の軸のレイアウトの評価
（2） PCP の乱雑さ（clutter）の削減方法の比較
（3） PCP の実用性の提示
（4） PCP と他のデータ分析手法との比較

上記のように，PCP の見せ方に関して議論がなされているものがほとんどであり，PCP の操作過程に着目した議論はされていない．また，PCP の可視化と関係代数演算のような演算をかけ合わせるような議論はされていない．

また，Bouali らは，対話型遺伝的アルゴリズムを使用することで，可視化手法の推薦を行うシステムを構築した [6]．これは，データや利用者の要求に応じてより適切な可視化手法（散布図行列や PCP など）の選択を支援するものである．我々の提案する $(PC)^2$DV は関係代数演算における選択・射影・結合が表現可能な可視化システムであるため，タプルが 1 つの線で明示され，詳細に参照・分析可能である PCP が適切である．

一方で，インタラクティブに操作をしながら PCP による分析を支援するシステムの提案もされている．Itoh らは，属性軸間の相関に基づいてインタラクティブに次元削減を行い，PCP から所望する情報の発見を支援するシステムを構築した [7]．Zhou らは，エントロピーの概念を導入することで，PCP の属性軸の整列順序をクラスタに基づいて決定する手法を提案した [8]．

また，多変量データを可視化するその他の手法として，複数の散布図を表示する散布図行列があげられる [9]．散布図行列は，属性同士の相関を直感的に把握できるが，散布図数が属性数の 2 乗に比例して増加する．そのため，属性数が多いデータでは，非常に大きい画面空間を使う必要がある．

2.2 データ解析支援

データやシステムの操作過程を管理する研究（Provenance）が行われている [10]．特にデータやシステム，プログラミングコードなどの操作過程や操作の意図を保存することは，複雑なデータ処理を支援するために重要なことであるといわれている．さらに，分析結果データの操作過程や操作の意図を示すことは，SQL のような関係代数演算をサポートする問合せ言語で記述することが有効であるともいわれている．この点において，$(PC)^2$L を用いて，$(PC)^2$DV のデータの操作過程の状態を保存することは有効な手段であるといえる．

また，データやシステムの操作過程を保存することでユーザの支援を行う手法の提案がなされている．Waldner らは，PC のアプリケーションの閲覧履歴や操作履歴を保存し，それらを時系列が理解できるように可視化することで，ユーザが過去に行った情報探索の詳細を再現する支援を行った [11]．Mindek らは，画像データと分析過程に利用する他のデータソースのデータを同時に表示し，分析者の文脈を保存したスナップショットを保存することで，シミュレーションデータの可視化や文書分析の支援を行った [12]．Gratzl らは，PCP やヒートマップ，散布図行列など様々な可視化手法を組み合わせて複数のデータソースから得られたデータとその解析過程を可視化し，データ解析の支援を行った [13]．これらの手法と比較して我々の手法は，「可視化システムのデータ解析過程を可視化して見せる」のではなく，「SQL に類似した言語を用いてデータ解析の途中結果を保存し，問合せ言語として一般的な SQL に親しみのあるデータ解析者を支援する」ものであり，立場が異なる．また，言語を用いて操作過程を保存することで，言語の一部を書き換えるだけで容易にデータ解析の改善をすることができる．その点でこれらの研究と比較して優位性をもつ．

また，大量のデータを対象とし，インタラクティブにデータ可視化を行う研究については多くの事例が見られる [14]．中でも，関係データベーススキーマに基づくデータに対し，GUI 上でクエリの記述や複数の可視化の連携を可能にし，データ解析を支援する研究も複数行われている．Derthick らは，データオブジェクトを可視化しつつ，インタラクティブに GUI でクエリが表現可能な環境を構築した [15]．North らは，データの可視化と，表示した複数の可視化間の連携をユーザーが自由に変更可能なインターフェースの構築を行った [16]．杉渕らは，クエリフローモデルによる直感的かつ段階的なクエリが構築可能な GUI を機能として備えた，可視化フレームワークを実装した [17]．これらの研究は，可視化とクエリを GUI 上で連携させることで，インタラクティブなデータ解析を支援する点では，我々と立場が同じと言える．その一方で，これらの研究は，「データベースに習熟していないデータ解析者を支援する」点を重視している．本研究は，「データ解析過程と可換な SQL に類似した言語により，データベースや SQL に習熟した解析者を支援する」ことを目的としており，これらの研究とは立場が異なる．

3 $(PC)^2$DV の概要と使用例

我々は，先行研究 [3–5] において多変量データの分析支援シ

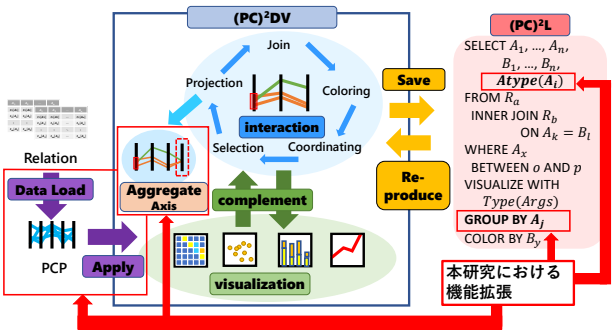

図 1　(PC)²DV の概要図

ステム（(PC)²DV ）を提案してきた．本章では，(PC)²DV の概要を説明する．

3.1　(PC)²DV の概要

(PC)²DV の概要図を図 1 に示す．なお，本稿で行った機能拡張の詳細は 4 章で述べる．(PC)²DV では，(PC)²L という SQL に類似した言語によりデータ操作を行った状態を保存・再現することが可能である．このシステムでは，データ分析者が以下のようなデータ操作手順によりデータ分析を行い，データ分析者に対して支援を行うことを想定する．

（1）　1つのリレーションを PCP により可視化する．
（2）　PCP を補完する形で，任意のグラフの描画を行う．
（3）　可視化結果をもとに，PCP 上でインタラクション（データ操作）を行う．その際，データ操作結果をリアルタイムに（2）で表示したグラフに反映する．
（4）　データ分析者が所望するときに，(2)，(3）の解析過程のスナップショットを (PC)²L で保存する．
（5）　（2）から（4）を繰り返す．その際，過去のスナップショットに戻る必要がある場合は該当する (PC)²L を入力し，システム上にデータ解析過程を再出力する．
（6）　データ分析者が所望の可視化結果を獲得する．

3.2　(PC)²DV の表示例

(PC)²DV の画面の例を図 2 に示す．実装システムは，先行研究 [3–5] と同様に，環境を問わず利用できるようにするため，Web ブラウザを通して多くの端末から利用できるように構築した．開発言語は，サーバサイドの処理に PHP，クライアント側の処理に HTML，CSS，JavaScript を使用した．

図 2 の A：PCP View では，指定したリレーションのデータを可視化した PCP が表示される．ここでは先行研究 [3] で定義したデータに対するインタラクションのうち，選択（Selection），色分け（Coloring），軸配置（Coordinating）の操作が利用可能である．PCP の各軸上を上下にドラッグし範囲選択をすることで，範囲内に含まれる線のみが PCP に表示される．これにより選択（Projection）の操作が可能である．PCP の各軸名をクリックすることで，その属性を基準にした色分け（Coloring）の操作が可能である．なお，色分けの色の順番は，指定した軸

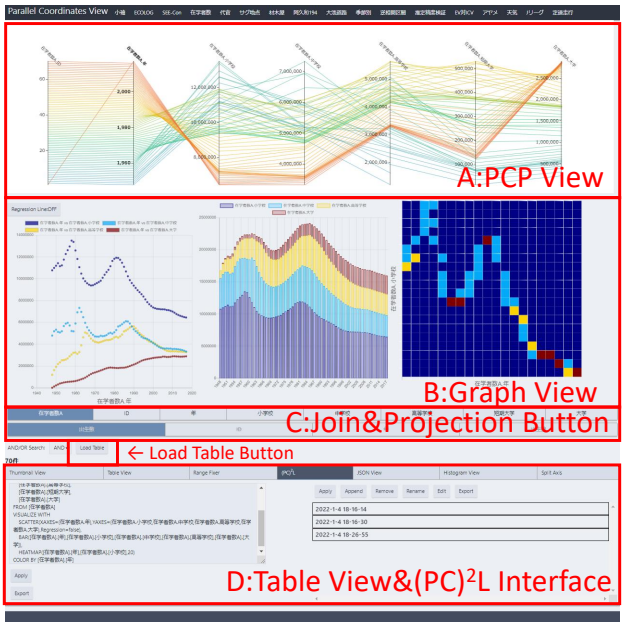

図 2　(PC)²DV の表示例

の属性値が数値の場合は昇順に「青，緑，黄，赤」がグラデーションとなるよう配色され，文字列の場合は固定色で配色される．PCP の各軸を左右にドラッグ&ドロップすることで軸の配置を変更することができる．これにより軸配置（Coordinating）の操作が可能である．

図 2 中の B：Graph View では，(PC)²L で指定したグラフが表示される．グラフ上には A，C 上で行ったインタラクションを反映したデータセットが表示される．

図 2 の C：Join & Projection Button では，横一行がリレーション一つに対応するトグルボタンが表示される．このボタンでは，先行研究 [3] で定義したデータに対するインタラクションのうち，結合（Join）と射影（Projection）の操作が利用可能である．各行の最左部には，のリレーション名が表示されたボタンが配置されている．これをクリックすることで，結合条件を選択する画面が表示される．そこで指定した結合条件に応じた結合（Join）の操作を実行することが可能である．なお，既に Join されているリレーション名が書かれたボタンを再度クリックすることで，Join を解除することができる．リレーション名が書かれたボタンより右側には，リレーションの属性名が書かれたボタンが配置されている．これをクリックすることでその属性に対応した PCP の軸の表示/非表示を切り替えることができる．これにより射影（Projection）の操作が可能となる．

図 2 の D：Table View & (PC)²LInterface では，A，C 上で行ったインタラクションを反映したデータセットのテーブル表示と，(PC)²L の入出力を受け付けるユーザインターフェースを持つ．(PC)²DV のデータ可視化状態を表す (PC)²L の出力や，テキストフィールドでの (PC)²L の入力と編集，テキストファイル形式での (PC)²L の出力が可能である．

4 (PC)^{2}DV への機能拡張

本章では，3 章で述べた (PC)^{2}DV および (PC)^{2}L に対して行った機能拡張について述べる．本稿における機能拡張は主に以下の二点である．

（1） (PC)^{2}DV におけるデータロード部の追加

（2） GROUP BY 句および集約演算に対応した (PC)^{2}L の文法拡張

4.1 機能拡張１：データロード部

3 章で述べたこれまでの (PC)^{2}DV は，主にデータ可視化部とデータ操作部により構成されている．データ可視化部とは，PCP やグラフといったデータを何らかの手法で可視化する部分を指す．また，データ操作部とは，GUI や (PC)^{2}L を用いてデータに対する選択，射影，結合などの演算により，可視化するデータを操作する部分を指す．

本稿ではこれらに加え，データロード部を (PC)^{2}DV に追加する．データロード部とは，(PC)^{2}DV の外部のデータソースにあるデータを取得する（ロードする）機能をもつ．ロードしたデータは 1 つのリレーションとして PCP により可視化される．すなわち，データロード部は 3.1 節で述べたデータ操作手順の（1）を担う部分である．外部のデータソースへのアクセスには ODBC（Open DataBase Connectivity）による接続を行う．ODBC を用いることで，データソースの形式が ODBC 接続が可能な形式であればデータの取得が可能となり，(PC)^{2}DV による分析の対象にすることができる．

実装システムでは，図 2 中の Load Table Button を押すことで，接続に必要な情報を入力する画面が表示される（図 3）．ここで必要な情報は以下のように記述する．

「＜ DSN String ＞；[＜ table name ＞；] ＜ query ＞」

（[] 内は任意）

＜ DSN String ＞は，DSN（Data Source Name）を用いたデータ接続のための文字列である．＜ table name ＞は，ロードしたリレーションに任意のリレーション名を設定するための文字列である．例えば＜ table name ＞に「tablename=Table1」と記述することでロードしたリレーションのリレーション名を Table1 に設定することができる．リレーション名の指定がない場合，リレーション名はそのデータをロードした順に R1, R2, ... となる．＜ query ＞は，接続したデータソースに対する問合せのための SQL 文である．＜ DSN String ＞で指定したデータソースに対して＜ query ＞に記述した SQL 文の問合せ結果が，＜ table name ＞で指定した名前をリレーション名とする 1 つのリレーションとして取得され，PCP により可視化される．

なお，図 3 で分析者が入力する文字列は (PC)^{2}L ではない．そのうえ，データを (PC)^{2}DV にロードする操作はデータ操作を始まる前に行われるため，試行錯誤の段階に含まれない．これらのことから，データをロードする段階で入力する文字列は (PC)^{2}DV における (PC)^{2}L によるデータ操作の保存・再現の対象外とした．

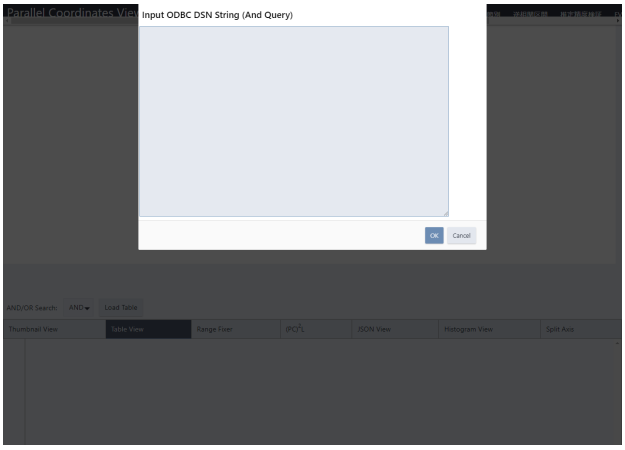

図 3 ODBC 接続に必要な情報を入力する画面

4.2 機能拡張２：GROUP BY 句および集約演算

3 章で述べたこれまでの (PC)^{2}DV は，選択・結合・射影といった関係代数演算のようなデータ操作が PCP 上で表現可能であることに着目してきた．このことは，多変量データを n-項関係として捉えると有効なデータ操作であると考えられる．

一方で，(PC)^{2}DV では可視化や分析対象としてライフログ等の大量の多変量データを扱うことが想定される．このようなデータは一般に粒度が細かく，データ一件のみでは意味をなさないということがある．そのような細かいデータを特定の単位でグループ化し，集約することで，データ分析に活用することができると考えられる．そのため，特定の単位でデータをグループ化し，各グループに対して代表値を求める操作は有用であると考えられる．

以上のことから，(PC)^{2}L に対する文法拡張を行った．この拡張によって，SQL における GROUP BY 句および集約演算を (PC)^{2}DV および (PC)^{2}L 上で表現し，新たに GROUP BY 句および集約演算を用いたデータ操作を可能とする．

以下では，GROUP BY 句および集約演算を表現し保存するために (PC)^{2}L に行った文法の拡張について述べる．この文法拡張は主に以下の二点である．

- 受理可能な集約演算の定義
- GROUP BY 句の追加

本稿では，(PC)^{2}L において集約演算の追加を行った．現在 (PC)^{2}DV 上で分析者が利用可能な集約演算は以下のとおりである．なお，これら以外の集約演算については今後の課題とする．

- 合計（SUM）
- 平均（AVG）
- 最大値（MAX）
- 最小値（MIN）
- データ件数（COUNT）
- 中央値（MED）
- 四分位数（QUARTILE）

これらの集約演算は，SQL のように (PC)^{2}L の SELECT 句

表 1　リレーション Temperature の属性

属性名	説明
Year	年
Month	月
Day	日
Max_Temperature	最高気温（°C）
Min_Temperature	最低気温（°C）

に「＜集約演算名＞ (＜ axis ＞)」と記述することで利用可能である．なお，＜ axis ＞は集約演算を行う対象の属性名である．集約演算の結果を表す軸は，SELECT 句に列挙した順に PCP 上に表示される．例えば，SUM(A_i) と記述すれば，属性 A_i の値の合計が求められ，SUM(A_i) を表す軸が PCP に追加され表示される．

また，(PC)^{2}L において GROUP BY 句の追加を行った．これにより，指定した属性の値を基準にデータをグループ化し，グループごとに集約演算を適用することができる．データをグループ化する操作は，SQL のように (PC)^{2}L の GROUP BY 句に基準となる属性名を記述することで利用可能である．例えば，「GROUP BY A_i」と記述することで A_i の属性値ごとにデータがグループ化される．その状態からさらに SELECT 句に MAX(A_j) と記述することで，A_i の属性値で分かれた各グループに対して属性 A_j の最大値が求められ，軸が追加される形で PCP に描画される．また，GROUP BY 句には複数の属性が指定可能である．例えば，GROUP BY 句に $A_1, A_2, \cdots, A_n$ と記述することで，$A_1, A_2, \cdots, A_n$ の属性値がすべて一致するデータごとにグループ化される．

5　(PC)^{2}DV を用いたデータ解析例

本章では，(PC)^{2}DV を用いたデータ解析例を示す．解析例 1 では，4 章で述べた拡張機能を活用した簡単なデータ解析例を示す．解析例 2 では，(PC)^{2}DV を用いた分析により有用な知見を得ることができるデータ解析例を示す．

5.1　データ解析例 1：気象データの解析例

日本の各官庁は近年，日本の経済や産業，気象などに関する様々なデータを，だれでも取得可能なオープンデータとして公開している．その一例として，気象庁が地域ごとの気温や降水量，最大風速などの気象データを公開していることが挙げられる [1]．解析例 1 では，解析対象のデータとして気象庁から取得した日ごとの最高気温と最低気温のデータを利用する．本節で扱うリレーション Temperature を表 1 に示す．これは神奈川県横浜市における日ごとの最高気温と最低気温のデータである．対象期間は 2015 年 1 月 1 日から 2021 年 12 月 31 日までであり，データ件数は 2557 件である．このデータは，Microsoft Access を用いて一つのテーブルに格納されている．すなわち，(PC)^{2}DV の外部に存在するデータである．

1：https://www.data.jma.go.jp/gmd/risk/obsdl/index.php（参照：2022-01-06）

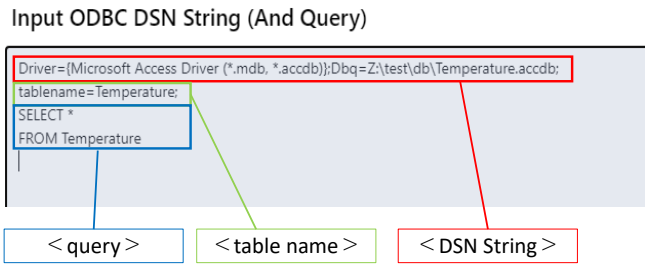

図 4　解析例 1：データロード時に必要な情報を入力

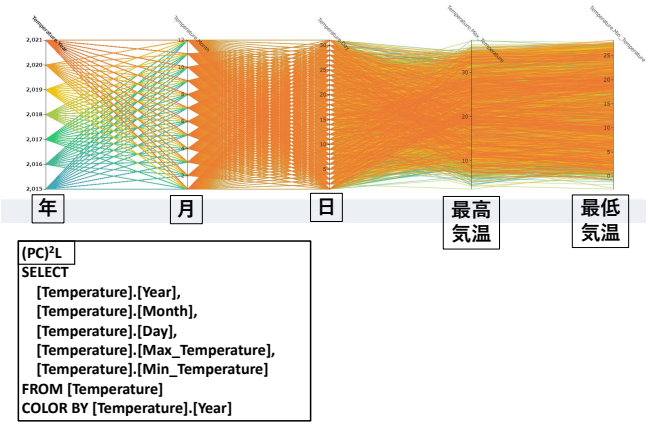

図 5　解析例 1：ロードしたリレーションを PCP で表示

まず，リレーション Temperature を (PC)^{2}DV にロードする．このとき，4.1 節で述べたデータロード部を用いる．Load Table Button をクリックし，図 3 の画面に必要な情報を記述する．その様子を図 4 に示す．文字列の各部分は 4.1 節で示した記法の各部分に対応している．ここでは，指定したファイルパスにある accdb ファイルに接続し，そのファイルの Temperature テーブルをリレーション Temperature として取得している．この記述によりデータを取得し，データが PCP として図 5 のように表示される．なお，PCP の線の色は［Year］（年）を基準として色分け（Coloring）がされている．

図 5 を見ると，PCP では線の数が多いと描画が複雑になることがわかる．この状態で，各年ごとの最高気温と最低気温を PCP 上で確認することは困難であるといえる．

また，各月ごとの最高気温と最低気温を確認するために［Month］（月）を基準として色分けを行った様子を図 6 に示す．図 5 と比較すると，［Max_Temperature］（最高気温）や［Min_Temperature］（最低気温）の軸まわりにグラデーションが現れたようにも見えるが，この状態で何らかの情報を見出すことは依然として困難である．

ここで，各年，各月ごとにデータをグループ化し，グループごとに最高気温と最低気温の平均値を表示する．この操作により，複雑な PCP の線がグループごとに集約される．すなわ

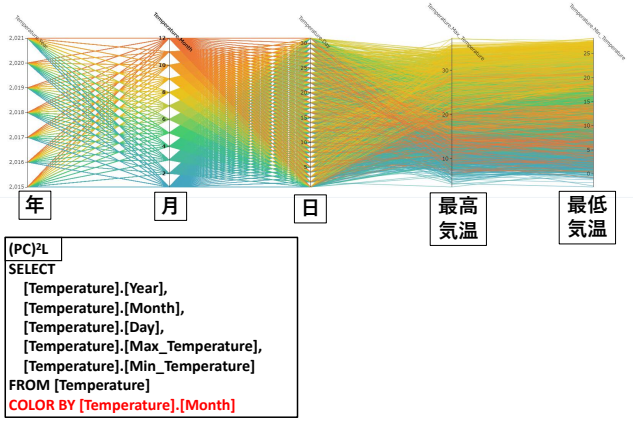

図 6　解析例 1：PCP を［Month］で色分け

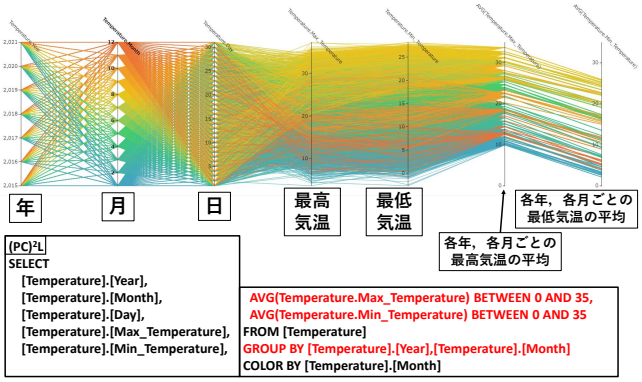

図 7　解析例 1：各年，各月ごとの集約演算結果を PCP に表示

ち，線がまとめられて表示されるため，理解が容易になることが期待される．このとき，4.2 節で述べた GROUP BY 句および集約演算を用いる．その結果を図 7 に示す．各年，各月ごとの最高気温の平均値［AVG(Temperature.Max_Temperature)］と最低気温の平均値［AVG(Temperature.Min_Temperature)］が軸として追加され，PCP 上に表示された．なお，集約演算の結果を表す二軸は軸の表示範囲を BETWEEN 句により 0 から 35 の範囲となるように統一した．これにより，最高気温と最低気温の大小関係を直感的に把握することが可能となる．図中の［AVG(Temperature.Max_Temperature)］や［AVG(Temperature.Min_Temperature)］に着目すると，線の色に関して軸の下部から上部への青→緑→黄という遷移と，軸の上部から下部への黄→橙→赤という遷移があることがわかる．これは，月ごとに色分けされた PCP において，1 月から 9 月あたりへかけて気温が上がることと 9 月から 12 月あたりへかかて気温が下がることを示している．すなわち，月ごとに寒暖差があり，推移していくという一般的な知識が情報として可視化された．

本解析例では，外部のデータソースに存在するデータをロードし，可視化した PCP に対し適切にグループ化し集約演算を行った．このことにより，単にデータを表示しただけの乱雑な状態の PCP から，拡張機能を用いて情報を可視化することができたといえる．

表 2　リレーション Solar_Generate の属性

属性名	説明
Year	年
Month	月
Day	日
Type	データの種類（predict（予測発電量）または measured（実測電力量））
DailyGenerate	一日の予測または実測発電量の合計（kWh）
ME	一日の予測発電量に対する実測発電量の平均誤差
RSME	一日の予測発電量と実測発電量の平均二乗誤差
CloudShape	その日最も通報された雲の種類
CloudCover	CloudShape の雲の，その日の雲量の平均

5.2　データ解析例 2：太陽光発電電力量データの解析例

太陽光発電とは，光エネルギーである太陽光から電気エネルギーを作り出す発電方式である．再生可能エネルギーの一種に分類されており，地球温暖化の原因となる温室効果ガスを発電時に排出しない発電方式として着目されている．

ここで，太陽光発電の発電電力量を考える．ある一日において，日射量や天気予報などを参考に前日にあらかじめ導出した予測の発電電力量（以下，予測電力量）と，当日に取得できる実測の発電電力量（以下，実測電力量）があるとする．このとき，予測電力量と実測電力量に差が生じる場合がある．この原因の一つとして，雲の影響による発電電力量の増減があると推測される．例えば，小さな雲が一時的に太陽光パネルを遮ることで発電電力量が減少し，大きな雲の切れ間から太陽光が太陽光パネルに差し込むことで発電電力量が増加すると考えられる．この推測をデータ解析により示すことを解析例 2 の目的とする．すなわち，どのような雲が太陽光発電の予測の誤差に対してどのような影響をもたらすのか，ということを $(PC)^2$DV を用いたデータ解析により明らかにする．

本節で扱うリレーション Solar_Generate を表 2 に示す．これはある建物一棟における太陽光発電の日ごとの予測発電量と実測発電量，その日の代表的な雲の種類とその雲量を表すデータである．対象日は 2020 年 1 月 1 日から 2020 年 12 月 31 日までの平日であり，データ件数は欠損日を除き 240 日の日ごとの予測発電量と実測発電量で合計 480 件である．なお，雲の情報については航空気象情報である定時飛行場実況気象通報式（METAR）を用いる．航空気象情報とは航空機の運航のために用いられる気象情報であり，各空港ごとに定められた時刻に METAR として観測結果が通報される．雲の種類は同時刻に最大三種類通報され，それぞれの雲量が 1〜8 の整数で通報される．その日最も通報された雲をその日の代表の雲とし，その雲量の平均をデータとして保存した．

まず，解析例 1 と同様に，リレーション Solar_Generate を $(PC)^2$DV にロードする．図 8 に示した記述により取得したデータが PCP として表示された（図 9）．

次に，予測発電量と実測発電量の差が大きい日に着目するため，［ME］（平均誤差）でデータを選択することを考える．ここでの［ME］は予測発電量に対する実測発電量の平均誤差であ

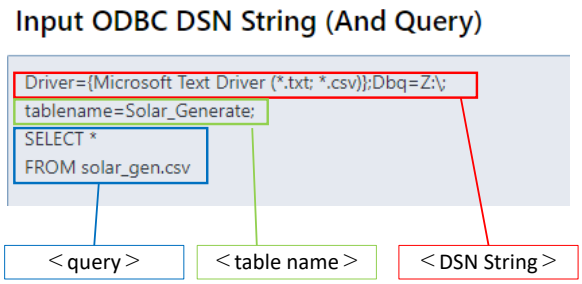

図 8　解析例２：データロード部に必要な情報を入力

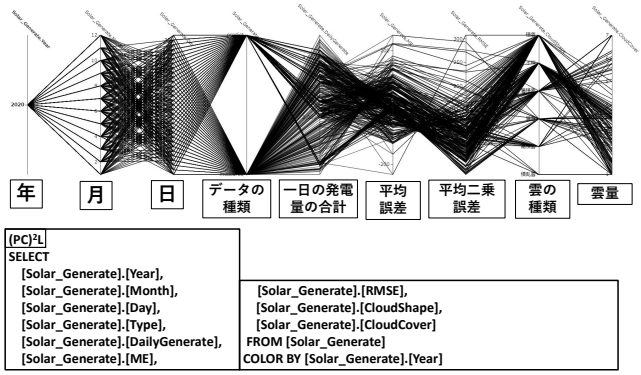

図 9　解析例２：ロードしたリレーションを PCP で表示

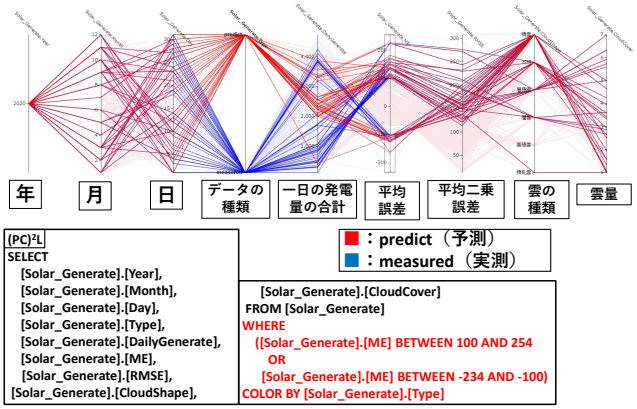

図 10　解析例２：[ME] の値でデータを選択

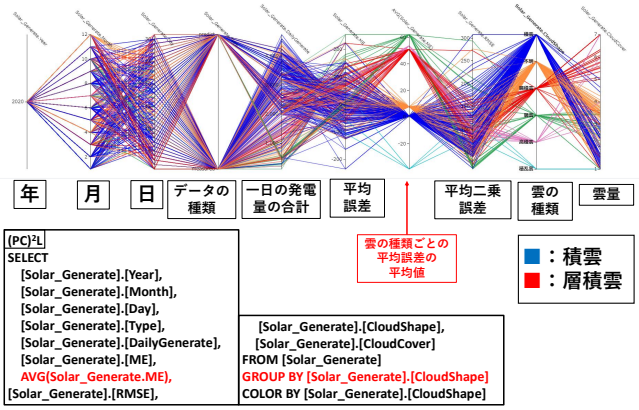

図 11　解析例２：雲の種類ごとの平均誤差の平均値を表示

る．そのため，値が正であることは予測発電量が実測発電量より大きいことを表し，値が負であることは予測発電量が実測発電量より小さいことを表す．[ME] が 100 以上または-100 以下であるデータのみを［Type］（データの種類）で色分けをし表示した様子を図 10 に示す．[DailyGenerate]（一日の予測または実測発電量の合計）の軸を見ると，確かに予測発電量（図中赤線）に対して実測発電量（図中青線）が大きく外れる場合があることがわかる．また，[CloudShape]（雲の種類）に着目すると，選択されたデータが存在する雲の種類とそうでない雲の種類があることがわかる．すなわち，平均誤差が大きいデータが存在する雲の種類とそうでない雲の種類があることがわかる．

ここで，雲の種類ごとの太陽光発電への影響の違いを分析するために，雲の種類ごとに平均誤差の平均値を求めることを考える．[ME] に対する選択を解除し，集約演算および GROUP BY 句を用いて雲の種類ごとに平均誤差の平均値を求め，[CloudShape] で色分けをした様子を図 11 に示す．この図から，雲の種類によって平均誤差の平均値が異なり，雲の種類ごとに太陽光発電の発電電力量予測への影響が異なることが推測される．

雲の種類ごとの平均誤差の平均値を求めたデータに対して，さらに［CloudShape］でデータを選択し，雲の種類ごとに予測誤差の傾向について分析を行う．ここでは，図 10 における選択したデータが存在する雲のうち，青空に浮かぶ塊状の白い雲である積雲と曇りの日に空に見られる灰色の雲である層積雲という異なる 2 種類の雲について，それぞれ分析を行う．

［CloudShape］で「積雲」を選択した様子を図 12 に，「層積雲」を選択した様子を図 13 に示す．図 12 の［ME］に着目すると，積雲が出ているときのデータの平均誤差は正負方向に幅広く存在していることがわかる．このことから，積雲は太陽光発電の実測発電量が予測発電量に対して大小の両方向に幅広く影響を及ぼすことがあるといえる．すなわち，積雲が出る場合の太陽光発電の発電電力量の予測は容易ではないということがわかる．一方で図 13 の［ME］に着目すると，層積雲が出ているときのデータの平均誤差は-70 から 200 程度の間にデータが存在し，その中でも正の値が比較的多いことがわかる．平均誤差の平均値が 50 付近であることも踏まえると，層積雲は太陽光発電の実測発電量が予測発電量に対して小さくなるように影響を及ぼす傾向にあるといえる．すなわち，層積雲が出る場合は太陽光発電の発電電力量が予測よりも少なくなることを想定して予測値を修正することで，予測誤差が減少することが期待される．

以上より，太陽光発電の発電電力量に影響を及ぼす，異なる二種類の雲について，それぞれで異なる傾向で影響を与える，という知見を分析により得ることができた．

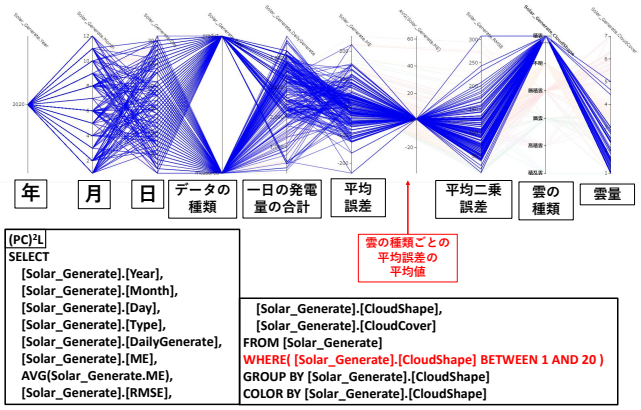

図 12　解析例２：[CloudShape] で「積雲」を選択

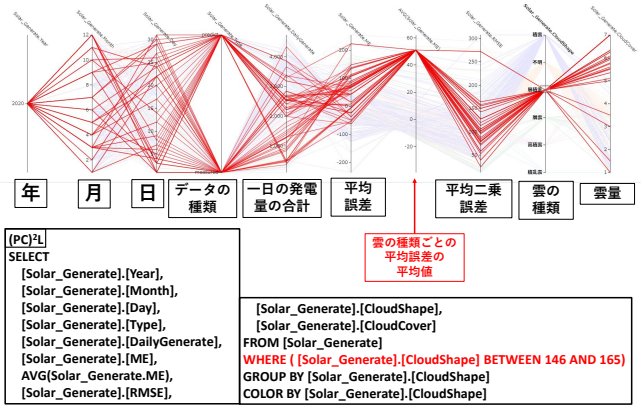

図 13　解析例２：[CloudShape] で「層積雲」を選択

6　まとめと今後の課題

本研究では，多変量データを PCP を用いて可視化し，その状態を (PC)^{2}L により保存・再現可能な，多変量データ分析支援システムである (PC)^{2}DV に機能拡張を行った．また，拡張機能を活用することで情報を得ることのできるデータ解析例を示し，(PC)^{2}DV を用いたデータ分析および機能拡張が有用であることを示した．

今後の課題として，(PC)^{2}DV 上でデータの更新，挿入，削除のような操作を実行可能する，といったさらなる拡張により，データ分析支援の範囲を拡大していくことが挙げられる．また，取得元が異なる複数のデータを (PC)^{2}DV を用いて組み合わせたデータ分析を行い，有用な知見を示すことが挙げられる．このことにより，(PC)^{2}DV により多様なデータ分析を行うことが可能であることを示していく．

謝辞　本研究の一部は横浜国立大学令和 3 年度学長戦略経費の支援による．

文　献

[1] Alfred Inselberg. The plane with parallel coordinates. *The Visual Computer*, Vol. 1, No. 2, pp. 69–91, 1985.

[2] Jimmy Johansson and Camilla Forsell. Evaluation of parallel coordinates: Overview, categorization and guidelines for future research. *IEEE Trans. on Visualization and Computer Graphics (TVCG)*, Vol. 22, No. 1, pp. 579–588, 2016.

[3] 濱崎裕太, 植村智明, 富井尚志. 多変量データを SPJ 質問により統合する平行座標プロット型情報可視化システムと操作言語. 情報処理学会論文誌データベース（TOD）, Vol. 12, No. 4, pp. 27–39, October 2019.

[4] 植村智明, 吉田顕策, 吉瀬雄大, 富井尚志. 試行錯誤を許容するデータ解析支援システムと電気自動車の走行ログ解析. 情報処理学会論文誌データベース（TOD）, Vol. 13, No. 4, pp. 13–26, October 2020.

[5] 植村智明, 能條太悟, 吉瀬雄大, 富井尚志. 解析者の興味に基づく道路区間集計が可能な EV 推定消費エネルギーデータ解析システムの構築と応用. 情報処理学会論文誌データベース（TOD）, Vol. 14, No. 4, pp. 70–85, October 2021.

[6] Fatma Bouali, Abdelheq Guettala, and Gilles Venturini. VizAssist: An interactive user assistant for visual data mining. *The Visual Computer: Int'l Journal of Computer Graphics*, Vol. 32, No. 11, pp. 1447–1463, 2016.

[7] Takayuki Itoh, Ashnil Kumar, Karsten Klein, and Jinman Kim. High-dimensional data visualization by interactive construction of low-dimensional parallel coordinate plots. *Journal of Visual Languages & Computing*, Vol. 43, pp. 1–13, 2017.

[8] Z. Zhou, Z. Ye, J. Yu, and W. Chen. Cluster-aware arrangement of the parallel coordinate plots. *Journal of Visual Languages & Computing*, Vol. 46, pp. 43–52, 2018.

[9] G. Grinstein, M. Trutschl, and U. Cvek. High dimensional visualizations. In *In Proceedings of KDD Workshop on Visual Data Mining*, 2001.

[10] Melanie Herschel, Ralf Diestelkämper, and Houssem Ben Lahmar. A survey on provenance: What for? what form? what from? *The VLDB Journal*, Vol. 26, No. 6, pp. 881–906, Dec 2017.

[11] Manuela Waldner, Stefan Bruckner, and Ivan Viola. Graphical histories of information foraging. *Proc. of the 8th Nordic Conf. on Human-Computer Interaction: Fun, Fast, Foundational(NordiCHI '14)*, pp. 295–304, 2014.

[12] Peter Mindek, Stefan Bruckner, and M. Eduard Gröller. Contextual snapshots: Enriched visualization with interactive spatial annotations. *Proc. of the 29th Spring Conf. on Computer Graphics(SCCG '13)*, pp. 49–56, 2013.

[13] S. Gratzl, N. Gehlenborg, A. Lex, H. Pfister, and M. Streit. Domino: Extracting, comparing, and manipulating subsets across multiple tabular datasets. *IEEE Trans. on Visualization and Computer Graphics(TVCG)*, Vol. 20, No. 12, pp. 2023–2032, Dec 2014.

[14] P. Godfrey, J. Gryz, and P. Lasek. Interactive visualization of large data sets. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28, No. 8, pp. 2142–2157, 2016.

[15] Mark Derthick, John Kolojejchick, and Steven F. Roth. An interactive visual query environment for exploring data. In *Proceedings of the 10th Annual ACM Symposium on User Interface Software and Technology*, UIST '97, pp. 189–198, 1997.

[16] Chris North and Ben Shneiderman. Snap-together visualization: A user interface for coordinating visualizations via relational schemata. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '00, pp. 128–135, 2000.

[17] 杉渕剛史, 田中譲. 関係データベースモデルに基づくデータベース可視化フレームワークの提案と実装. 電子情報通信学会論文誌. D, 情報・システム = The IEICE transactions on information and systems (Japanese edition), Vol. 90, No. 3, pp. 918–932, mar 2007.