

クラスタリングされたヒートマップと多重折れ線グラフによる 時差のあるデータの可視化

Visualization of Time Series Data with Time Differences Using Clustered Heatmaps and Multiple Line Graphs

遠藤 麗香

Reika Endo

法政大学情報科学研究科

E-mail: reika.endo.3g@stu.hosei.ac.jp

Abstract

There are some data in the world that affect other data, although there are time differences. Data that have time difference of several months but that are relevant include the amount of pollen dispersal and the previous year's summer temperatures. Higher summer temperatures are said to increase the amount of pollen dispersal. In this paper, we propose a method to cluster data and visualize them by combining heatmaps and line graphs, with the aim of drawing multiple types of time series data with time differences on a single screen. Our method facilitates reading and comparing not only different data of the same type but also different types of data. The number of clusters can be changed by using a slider, and which data is used for clustering can be selected with a button. Clicking on a set of heatmap bands will draw the data belonging to that cluster as a line graph above the heatmap area. The experiment results showed that the method combining one line graph with a heatmap achieved the highest accuracy and shortest response time. Furthermore, the method combining two line graphs with a heatmap was found to make it easier for participants to compare data.

1. はじめに

時差はあるが、関連があるデータが存在する。例えば数分、数時間の時差で関連があるデータには、降水量と河川の水量のデータがあり、数日、数週間の時差で関連があるデータにはコロナ感染者数と病床使用率がある。数週間、数ヶ月の時差で関連があるデータには、花粉飛散量と前年の夏の気温のデータがある [1]。花粉飛散量はその日の気温や風速、湿度も関連するが、前年の夏の気温の影響も受けると言われている。夏の気温が高いと形成される雄花の量が増え、飛散する花粉の量も増える。逆に夏の気温が低いと形成される雄花の量が少なくなり、花粉飛散量も減る。

時系列データの可視化には、折れ線グラフやヒートマップがよく使用される。折れ線グラフは各データを 1 つの折れ線として描画する。詳細な値が読み取りやすいが、

データの数が増えると折れ線の重なりが増えて読み取りにくい問題がある。軸を増やすことで単位が異なるデータの可視化も可能である。ヒートマップは各データを 1 つの帯として描画し、その帯を縦に並べて描画する可視化手法である。データが増えても重なりは発生しないが、折れ線グラフと比べて正確な値が読み取りにくい。

本研究では、時差のある複数の時系列データを 1 つの画面にまとめて描画し、読み取りや比較をしやすいヒートマップベースの可視化手法を提案する。降水量と河川水位など時差がありながらも関連があるデータをクラスタリングし、複数の折れ線グラフとヒートマップを組み合わせる可視化する。ユーザはスライダーやボタンによってクラスタリング結果を変更でき、自身のニーズに合わせた対話的な可視化を行える。

本手法の有効性を確認するため、2つの実験 1,2 を行った。実験 1 は 20 年分の日ごとの平均気温データと花粉飛散量データを用いて、ヒートマップに折れ線グラフを 1 つ組み合わせた手法と折れ線グラフのみの可視化を比較するものである。実験 1 の結果を踏まえてよりデータの比較がしやすくなるよう、表示する折れ線グラフを複数にするなどの追加実装を行い、実験 2 を行った。実験 2 は、実験 1 と同様に日ごとの平均気温データと花粉飛散量データを用いて、ヒートマップに折れ線グラフを 1 つ組み合わせた手法、ヒートマップに折れ線グラフを 2 つ組み合わせた手法、折れ線グラフのみの可視化を比較するものである。実験の結果、ヒートマップに折れ線グラフを 1 つ組み合わせた手法が最も高い精度、短い回答時間を得た。アンケート結果では、データの比較のしやすさの項目でヒートマップに折れ線グラフを 2 つ組み合わせた手法が高い評価を得た。

2. 関連研究

Endo ら [2] は時系列データのクラスタリング、ヒートマップ、折れ線グラフを組み合わせた可視化手法を提案し、ヒートマップ単体や折れ線グラフ単体で使う場合よりも読み取り精度が向上することを示した。この研究では可視化したデータの種類の数が気温データの 1 つであったが、本研究では可視化するデータの種類の数を 2 つ以上にす。複数種類のデータを用いることで、異なる種類のデータ間に関連が存在するかどうか、また関連がある場合

どのような相関を示すかを調査できる。さらに、クラスタリングにより、1つの種類のデータで類似するデータ同士が、別の種類のデータでも同様に類似する傾向があるかを検証可能とする。

Tanakaら[3]は降水量と河川水位の関係を2色塗り分けスパークライン(two-tone pseudo coloring sparklines; TPCS)[4]によって可視化した。TPCSはヒートマップのように色で値を表現するが、上部と下部を2色で塗り分ける2色塗り分け法[5]を使用することで詳細な値も表現できる。ある地点の降水量と河川水位をそれぞれTPCSで可視化し、2つのTPCSを縦に並べて1つの観測地点のデータとしている。その各観測地点のTPCSを河川の流れて沿って配置する。河川は本流と流れ込む支流も流れて沿って配置する。複数の観測地点の降水量と河川水位を河川の流れて沿って並べることで、雨が降ってから河川水位が上がっていることや支流は雨が降ってから少ない時間差で水位が上がり、本流は支流に比べて時間差が発生してから水位が上がるなどが読み取れる。

3. クラスタリングの背景

クラスタリングには、近い距離のデータを集める階層型クラスタリングと、事前にクラスタの数を決めてデータを集める非階層型クラスタリングの2種類が存在する。階層型クラスタリングは、似たデータ、つまり近い距離のデータを1つずつ階層的に連結させて木の葉の部分からボトムアップ式にクラスタを作成する手法である。距離は相関係数やピアソン相関係数を使用して算出でき、近いデータを選択する方法はワード連結法、最短距離法、群平均法などがある。非階層型クラスタリングのk-means法などと違い、クラスタ数を事前に決める必要がないが、計算量が大きくなる。階層型クラスタリングで得られるグラフは図1のようになり、このグラフをデンドログラムという。横軸が各データ、縦軸が距離を表している。似たデータはデンドログラムの下部、近い距離でクラスタを形成する。クラスタの数はデンドログラムの階層を横に切る場所によって変わる。例えば、デンドログラムを0.5の位置で切ると、6つのクラスタ{2}, {4}, {1, 5}, {7, 8}, {3, 6}, {0, 9}が生成される。また、0.7の位置で切ると、{2}, {0, 1, 3, 4, 5, 6, 7, 8, 9}の2つのクラスタが生成される。

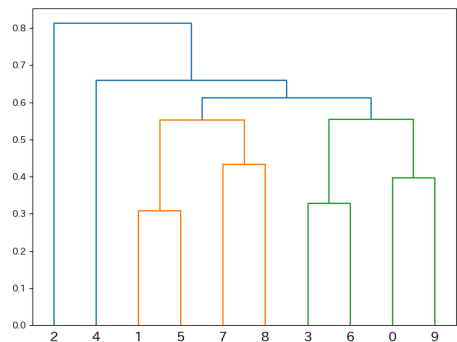


図1 階層型クラスタリングの結果のデンドログラム

4. 提案手法

本論文では、Endo[2]らの研究をベースに、時系列データをクラスタリングし、ヒートマップで可視化する。ヒートマップでは各クラスタのデータの平均値を、折れ線グラフではクラスタ内のデータの実際の値を可視化する手法を提案する。

入力データは m 個あるとし、各データは1から n 時刻を持つとする。データ群を A, B, C のように表すとする。データ群はデータの種類の数に合わせるため、データの種類が2種類であればデータ群は A, B と表す。 i 番目のデータの j 番目の時刻における値を v_{ij} や u_{ij} とすると、以下のよう表せる。

$$A = \{a_1, \dots, a_m\}$$

$$B = \{b_1, \dots, b_m\}$$

$$a_i = \{v_{i1}, \dots, v_{in}\}$$

$$b_i = \{u_{i1}, \dots, u_{in}\}$$

i 番目のデータ a_i と k 番目のデータ a_k の相関係数を $c_a(i, k)$ とすると、2つのデータ間の距離 d_{aik} は以下のように表せる。

$$d_{aik} = 1 - c_a(i, k)$$

距離 d_{aik} は、類似するデータ同士では小さくなり、類似しないデータでは大きくなる。この距離 d_{aik} を元に階層型クラスタリングを行う。データ b_i とデータ b_k の相関係数や距離も同様に求められる。

可視化結果が図2、図3、図4、図5である。各クラスタのヒートマップを表示する領域を下部に、そのクラスタに属するデータの折れ線グラフを表示する領域を上部に配置する。本研究で使用するデータは気温と花粉飛散量であり、各ヒートマップのセットの上段の帯は気温、下段の帯は花粉飛散量とする。また、上側の折れ線グラフは実線が気温の折れ線グラフ、点線が花粉量の折れ線グラフを表している。折れ線グラフの横軸は5月15日から翌年5月14日までの日にちを表し、左の縦軸は気温を、右の縦軸は花粉飛散量を表している。折れ線グラフはヒートマップの帯をクリックした時に表示されるようにする。画面の上部にクラスタ数を変えるためのスライダと気温と花粉量の2つのうちどちらのデータを元にクラスタリングするか変えるためのボタンを配置する。スライダを動かすと描画されるヒートマップの帯の数が変わるようにし、ボタンを押すと気温に着目してクラスタリングするか、花粉量をもとにクラスタリングするか変わるようにする。各ヒートマップの帯で描画される値は各クラスタ内のデータ値の平均値とする。スライダの隣にある、気温、花粉と表示されたボタンが、どちらのデータを元にクラスタリングするか変えるためのボタンである。オレンジ色で表示されているボタンが現在元になっているデータである。このボタンによりユーザは着目したい種類のデータでクラスタリングし、データを読み取れる。

図2はスライダを調整し、1番上のヒートマップをクリックした後の画面である。スライダの値は5を示しており、10本のヒートマップの帯が2本ずつセットになり5つのクラスタを形成している。最上部のクラスタは黒い長方形で囲われており、今回クリックされたことを示している。折れ線グラフ部分では6本の実線と6本の点線が表示されており、その右側に6つのデータの色と年が

表示されている。図3は画面上部の花粉ボタンをクリックし、最上部のクラスタをクリックした時の画面である。スライダは変更されていないため、クラスタ数は5のままである。ヒートマップの本数は変わっていないが、気温を元にクラスタリングした図2と色が異なっている。また、上の折れ線の本数や色も異なっている。実験1で使用する図2、図3の手法では、クラスタをクリックして描画される折れ線グラフの線の色は、クラスタをクリックするたびにランダムに選ばれるため、ユーザが見やすい色になるように折れ線の色を変更できる。

図4、図5は、表示する折れ線グラフを2つにしたものである。1つ目にクリックしたクラスタは上の折れ線グラフエリアに描画され、2つ目にクリックしたクラスタは下の折れ線グラフエリアに描画される。3つ目は上のエリアが更新され、4つ目は下のエリアが、5つ目以降も同様に上下のエリアが交互に更新される。クリックして描画される折れ線グラフの色は、図2、図3の時と異なり、ランダムではなく、カラーユニバーサルデザインを元に事前に決定した色で描画される。

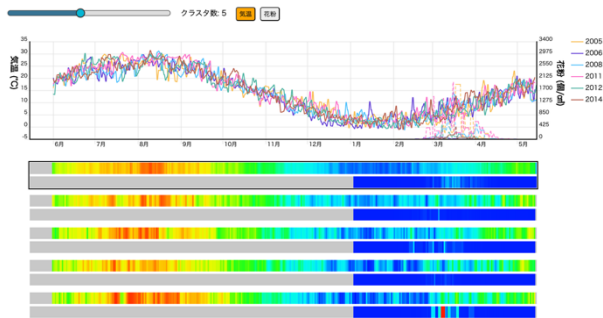


図2 スライダ調節して1番上のヒートマップをクリックした後の画面

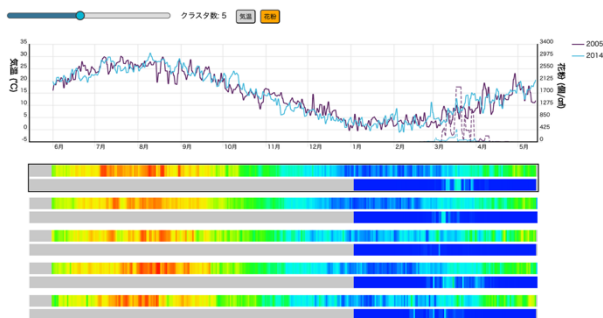


図3 花粉のボタンを押し、1番上のヒートマップをクリックした後の画面

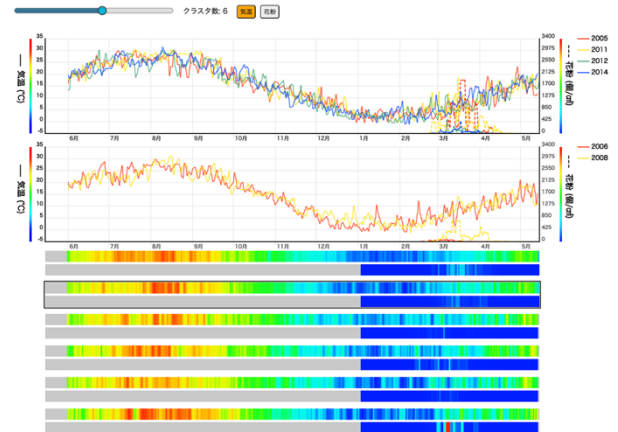


図4 スライダ調節して2つ折れ線グラフをクリックした画面

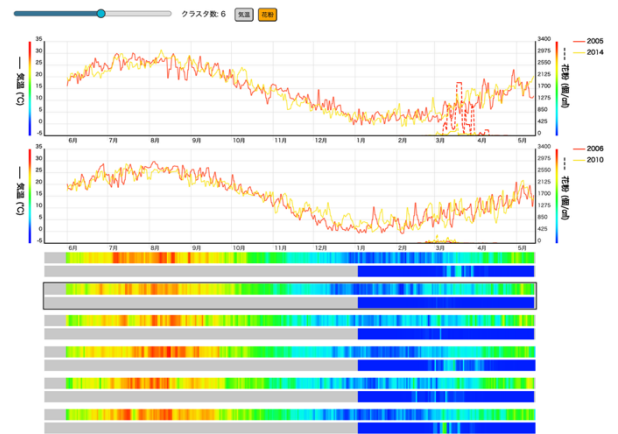


図5 花粉のボタンを押し、2つ折れ線グラフをクリックした後の画面

5. 実装

可視化する気温データは、気象庁の過去の気象データ・ダウンロードページから2005年から2024年の20年分の、東京都青梅市、八王子市、府中市の6月1日から翌年の5月14日までの日ごとの平均気温データをCSVファイルとして取得した。また、花粉飛散量のデータは東京都保険医療局の過去の花粉飛散数データページから、2005年から2024年の1月4日から5月14日までの青梅市、八王子市、府中市のスギ花粉飛散量データをCSVファイルで取得した。ファイルの読み込み、類似度計算、クラスタリング、JSONファイルへの保存にはPython、ヒートマップと折れ線グラフの可視化にはJavaScriptとJavaScriptのライブラリP5.jsを使用した。

CSVファイルを読み込んだ後、必要な情報の抽出、欠損値の補間、閏年への処理を行う。その後、データ間の相関係数から距離を求め、クラスタリングを行う。階層型クラスタリングにおけるクラスタの結合方法は群平均法を使用した。群平均法は、2つのクラスタ内のすべてのデータの組み合わせで距離を求め、その平均を新たなクラスタの距離として使用する。クラスタリング結果、

補間されたデータ、年の名前はそれぞれ JSON ファイルに保存され、可視化する JavaScript 側に渡される。JavaScript では、JSON ファイルの読み込み、ヒートマップや折れ線グラフによる可視化、スライダやボタン、ヒートマップがクリックされた時の処理を実装する。

ヒートマップの可視化では、データの各値が長方形として表され、閏年に合わせて 366 個の長方形が横に並べられる。長方形の色は各値から決定される。値がない場合は、灰色の長方形で表される。各種類のデータにおける最高値は赤色、最低値が青色で表され、その間の値は黄色、緑色、水色などの色で表される。折れ線グラフの可視化では、クリックされたクラスタ内のすべてのデータがそれぞれ折れ線として表示される。ヒートマップをクリックすると、クリックされたヒートマップのクラスタを囲む黒い長方形が描画され、同じ年のデータを同じ線の色の実線と点線で描画する。クラスタリングに使用するデータを切り替えるボタンをクリックすると、ボタンの色を変更し、現在のスライダの値のままクラスタを変更する。

6. 実験 1

6.1. 実験方法

実験 1 では、ヒートマップに折れ線グラフを 1 つ組み合わせた手法、折れ線グラフのみの 2 通りの手法で複数種類のデータを持つ時系列データを可視化し、ユーザーがデータを容易に読み取り比較できるか評価する実験を行った。折れ線グラフのみでの可視化(図 6)では、すべてのデータがそれぞれ折れ線として描画されており、各年ラベルが折れ線グラフの右側に表示されている。同じ年の気温データと花粉データはランダムに選ばれた同じ色で可視化され、気温データは実線、花粉データは点線で可視化される。折れ線グラフの横軸は 5 月 15 日から翌年 5 月 14 日までの日付を表し、左の縦軸は気温、右の縦軸は花粉量を表している。実験で使用したデータは 2005 年から 2024 年までの青梅市の日ごとの平均気温データと花粉飛散量データ、および 2015 年から 2024 年までの府中市の日ごとの平均気温データと花粉飛散量データである。

被験者は 8 名であり、16 歳から 49 歳までの男性 5 名、女性 3 名である。被験者は 2 つのグループに分けられ、片方のグループは最初にヒートマップに折れ線グラフを 1 つ組み合わせた手法を使用し、もう片方のグループは最初に折れ線グラフのみの可視化手法を使用し、指定したデータの数を答えてもらった。ただし、各被験者の最初の手法では 2005 年から 2014 年までの青梅市のデータを可視化し、2 番目の手法では 2015 年から 2024 年までの青梅市のデータを可視化した。実験前に府中市の 2015 年から 2024 年のデータを使用して各手法の説明を行った。指定したデータの数を問う質問は 2 問ずつ聞き、答えてもらった。1 問目は最も高いピークの位置が指定された範囲を満たすデータ数(E1)であり、2 問目は指定された気温と花粉量を満たすデータ数(E2)である。回答までの時間とそれぞれの回答から平均回答時間(MRT)と平均二乗誤差(MSE)を求めた。実験後に値の読み取りやすさ(Q1)、

データの比較のしやすさ(Q2)、視覚的に見やすいと感じたか(Q3)、初見でも見方、使い方は理解できたか(Q4)の 4 項目で 5 段階の評価を問うアンケートを実施した。選択肢は 5 (とても良い)、4 (良い)、3 (どちらでもない)、2 (あまり良くない)、1 (全く良くない)の 5 つである。また、改善点や意見を書いてもらうための自由回答欄を設けた。

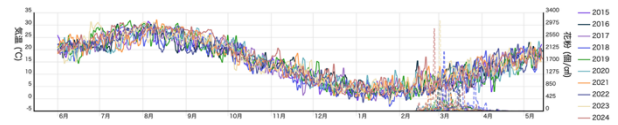


図 6 折れ線グラフのみの可視化

6.2. 実験結果

表 1 は実験 1 の結果をまとめたものである。E1 の最も高いピークの位置が指定された範囲を満たすデータ数の問題で、平均二乗誤差は、ヒートマップに折れ線グラフを 1 つ組み合わせた手法が 0.750 となり、折れ線グラフのみの 10.1 を下回った。平均回答時間は、折れ線グラフを 1 つ組み合わせた手法が 95.3 秒、折れ線グラフのみが 147 秒となった。E2 の指定された気温と花粉量を満たすデータ数の問題で、平均二乗誤差は、折れ線グラフを 1 つ組み合わせた手法が全員正答の 0.00、折れ線グラフのみでは 0.375 となった。平均回答時間は、折れ線グラフを 1 つ組み合わせた手法が 90.2 秒、折れ線グラフのみが 111 秒となった。これらの結果に t 検定を行った。E1 の MSE の p 値は $p = 0.00881 < 0.05$ 、MRT の p 値は $p = 0.0279 < 0.05$ であったため、E1 の 2 つの手法間の MSE と MRT に有意な差があると分かった。E2 の MSE の p 値は $p = 0.598 > 0.05$ 、MRT の p 値は $p = 0.175 > 0.05$ であった。E2 の MSE と MRT に有意な差は見られなかった。

表 1 実験 1 の結果

		E1	E2
ヒートマップに折れ線グラフを 1 つ組み合わせた手法	MSE	0.750	0.00
	MRT	95.3s	90.2s
折れ線グラフのみ	MSE	10.1	0.375
	MRT	147s	111s

6.3. アンケート結果

アンケートの値の読み取りやすさ(Q1)、データの比較のしやすさ(Q2)、視覚的に見やすいと感じたか(Q3)、初見でも見方、使い方は理解できたか(Q4)の評価の平均が図 7 である。各棒グラフの左側がヒートマップに折れ線グラフを 1 つ組み合わせた手法、右側が折れ線グラフのみの結果である。最初の 3 つの質問 Q1、Q2、Q3 で、折れ線グラフを 1 つ組み合わせた手法は折れ線グラフのみより高い評価を得た。しかし、Q4 の初見でも見方、使い方は理解できたかの項目では、折れ線グラフを 1 つ組み合わせた手法は折れ線グラフのみの評価を下回った。折れ線グラフを 1 つ組み合わせた手法はすべての質問で 3.5 点以上と評価されたが、折れ線グラフのみでは 3.5 点を上回ったのは Q4 のみであった。これらの評価に対して t 検定を行ったところ、Q1 の p 値は $p = 0.0000678 < 0.05$ 、Q2 では $p = 0.00830 < 0.05$ 、Q3 は $p = 0.0449 < 0.05$ 、Q4

は $p = 0.281 > 0.05$ となった。したがって、Q1, Q2, Q3 では折れ線グラフを 1 つ組み合わせた手法と折れ線グラフのみの手法の評価に有意差があったが、Q4 では有意差がないことが分かった。自由回答欄では、「1 度に載せるデータ数が多いほど折れ線グラフを 1 つ組み合わせた手法の方が見やすいと感じた。クラスタ数を増やしても一つのクラスタにデータが集まりすぎて見づらいことがあった。」、「クラスタリングを活用できなかった。色が見づらかった。」といった回答があった。

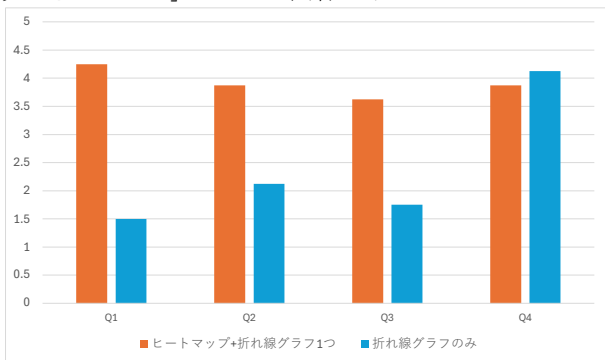


図7 実験1のアンケート結果

7. 実験2

7.1. 実験方法

表示する折れ線グラフを 2 つに増やした手法が、折れ線グラフ 1 つの時や折れ線グラフのみの手法よりデータを比較しやすいか調べるために、さらに実験を行った。実験 1 ではクリックして表示される折れ線や、折れ線グラフのみの手法の折れ線の色が毎回ランダムに選ばれていたが、色の見づらさに関して被験者による指摘が多かったため、実験 2 では 3 つの手法すべてでカラーユニバーサルデザインを元に折れ線の色を選び、表示される色を固定した。折れ線グラフのみの手法は図 6 と同様にすべてのデータがそれぞれ折れ線として描画されており、各年のラベルが折れ線グラフの右側に表示されるが、描画の度にランダムに選ばれた折れ線の色ではなく、カラーユニバーサルデザインを元に決定した折れ線の色で毎回描画される。折れ線グラフの横軸は 5 月 15 日から翌年 5 月 14 日までの日付を表し、左の縦軸は気温、右の縦軸は花粉量を表している。実験で使用したデータは 2005 年から 2024 年までの八王子市の日ごとの平均気温データと花粉飛散量データ、および 2005 年から 2014 年までの府中市の日ごとの平均気温データと花粉飛散量データである。

被験者は 6 名であり、17 歳から 50 歳までの男性 4 名、女性 2 名である。被験者には 3 通りの手法を被験者ごとに異なる順序で使用してもらい、指定した条件に当てはまるデータを探してもらった。ただし、各被験者の最初の手法では 2005 年から 2014 年までの八王子市のデータを可視化し、2 番目の手法では 2015 年から 2024 年までの八王子市のデータを可視化し、3 番目の手法では 2005 年から 2014 年までの府中市のデータを可視化した。実験前

に府中市の 2015 年から 2024 年のデータを使用して各手法を説明した。指定した条件を満たすデータをすべて答えてもらう質問を 2 問ずつ行った。1 問目は前年に比べて花粉量の合計が増えている年をすべて答えてもらい(E1)、2 問目は指定された気温と花粉量を満たす年をすべて答えてもらった(E2)。回答までの時間とそれぞれの回答から、平均回答時間(MRT)と正答率を求めた。条件に当てはまるデータをすべて答えてもらう設問であるため、採点基準は次のように設定した。解答数が正解数と同じか、正解数より少ない場合は、正解に応じて点を与える。例えば、正解が a, b の 2 点満点に対し、a と答えると 1 点となる。解答数が正解数より多い場合は、正解数以上の解答 1 つにつき、0 点まで 1 点減点する。例えば、正解が a, b に対し、a, b, c と答えると正解の 2 点から 1 点減点し、1 点となる。満点を 100% として正答率を求めた。実験後に値の読み取りやすさ(Q1)、データの比較のしやすさ(Q2)、見方、使い方の理解のしやすさ(Q3)の 3 項目で 5 段階の評価を問うアンケートを実施した。選択肢は 5 (とても良い)、4 (良い)、3 (どちらでもない)、2 (あまり良くない)、1 (全く良くない)の 5 つである。また、改善点や意見を書いてもらうための自由回答欄を設けた。

7.2. 実験結果

表 2 が実験 2 の結果をまとめたものである。E1 の前年より花粉量が増えている年を答える問題、E2 の指定された気温と花粉量を満たす年を答える問題、どちらもヒートマップに折れ線グラフを 1 つ組み合わせた手法の正答率が最も高く、2 番目にヒートマップに折れ線グラフを 2 つ組み合わせた手法、3 番目に折れ線グラフのみの手法と続いた。E1, E2 ともに平均回答時間は折れ線グラフのみが最も短く、ヒートマップに折れ線グラフを 2 つ組み合わせた手法が最も長くなった。

E1, E2 の正答率と回答時間に対して、各被験者の結果から被験者平均との差を引いた後、一元配置分散分析を行い有意差の有無を確認した。E1 の正答率の p 値は $p = 0.605 > 0.05$ となり、ヒートマップに折れ線グラフを 2 つ組み合わせた手法、折れ線グラフを 1 つ組み合わせた手法、折れ線グラフのみの 3 手法間に有意な差は見られなかった。E2 の正答率の p 値は $p = 0.191 > 0.05$ となり、E1 同様 3 手法間に有意な差があるとは言えなかった。E1 の MRT の p 値は $p = 0.00293 < 0.05$ となったため、3 つの手法の MRT に有意な差があると分かった。また、E2 の MRT の p 値は $p = 0.0475 < 0.05$ となり、有意な差が見られた。分散分析に追加で対応のある t 検定に、 p 値に検定回数をかけるボンフェローニ補正を施した多重比較を行った。有意水準は 0.05 とした。多重比較の結果が表 3 である。E1 のヒートマップに折れ線グラフを 2 つ組み合わせた手法と折れ線グラフを 1 つ組み合わせた手法の MRT の p 値は $p = 0.358$ であったため、有意な差はなかった。また、ヒートマップに折れ線グラフを 2 つ組み合わせた手法と折れ線グラフのみの p 値は $p = 0.161$ 、折れ線グラフを 1 つ組み合わせた手法と折れ線グラフのみの p 値は $p = 0.0677$ であったため、有意差は見られなかった。E2 の折れ線グラフを 2 つ組み合わせた手法と折れ線グラフを 1 つ組み合わせた手法の MRT の p 値は $p = 1.41$ 、折れ

線グラフを2つ組み合わせた手法と折れ線グラフのみのp値は $p = 0.162$, 折れ線グラフを1つ組み合わせた手法と折れ線グラフのみのp値は $p = 0.448$ であった。いずれも0.05を超えているため、それぞれの手法間に有意な差は見られなかった。

表2 実験2の結果

		E1	E2
ヒートマップ+折れ線グラフ2つ	正答率	75.0%	91.7%
	MRT	180s	110s
ヒートマップ+折れ線グラフ1つ	正答率	83.3%	100%
	MRT	143s	91.5s
折れ線グラフのみ	正答率	70.8%	86.1%
	MRT	109s	68.8s

表3 実験2の平均回答時間の多重比較結果

	E1	E2
ヒートマップ+折れ線グラフ2つとヒートマップ+折れ線グラフ1つ	0.358	1.41
ヒートマップ+折れ線グラフ2つと折れ線グラフのみ	0.161	0.162
ヒートマップ+折れ線グラフ1つと折れ線グラフのみ	0.0677	0.448

7.3. アンケート結果

アンケートのQ1からQ3の評価の平均が図8である。Q1からQ3の各項目は値の読み取りやすさ(Q1), データの比較のしやすさ(Q2), 見方, 使い方の理解のしやすさ(Q3)である。各棒グラフの左側がヒートマップに折れ線グラフを2つ組み合わせた手法, 真ん中がヒートマップに折れ線グラフを1つ組み合わせた手法, 右側が折れ線グラフのみの結果である。ヒートマップに折れ線グラフを2つ組み合わせた手法がQ1とQ2で最も高く, ヒートマップに折れ線グラフを1つ組み合わせた手法がQ3では最も高くなった。折れ線グラフのみの手法は3つの項目すべてで最も低い評価となった。これらの評価の結果に対して, 各被験者の評価結果から被験者平均との差を引いた後, 一元配置分散分析を行い, 有意差の有無を調べた。Q1のp値は $p = 0.00000125$, Q2は $p = 0.0000131$, Q3は $p = 0.353$ となり, Q1とQ2では有意差があったが, Q3では有意差があるとは言えないことが分かった。ただし, 分散分析ではどの手法間に有意な差があるか判断できないため, 対応のあるt検定に, p値に検定回数をかけるボンフェローニ補正を施した多重比較を行った。有意水準は0.05とした。多重比較の結果が表4である。Q1, Q2の両方で, ヒートマップに折れ線グラフを2つ組み合わせた手法と折れ線グラフのみの手法, ヒートマップに折れ線グラフを1つ組み合わせた手法と折れ線グラフのみの手法に有意な差があった。ヒートマップに折れ線グラフを2つ組み合わせた手法とヒートマップに折れ線グラフを1つ組み合わせた手法はQ1では $p = 1.09$, Q2は $p = 0.175$ となったため, 有意差があるとは言えない。自由回答欄では, 「折れ線グラフ1つの方が, 切り替えが

素早くできるため, 比較がしやすかった。」、「グラフの表示をスクロールなしで出来るようにして欲しい。」、「赤とオレンジが見分けづらいと感じた。」、「折れ線グラフが2つのものは比較が一番しやすく便利だった。」、「ヒートマップのみで前年との比較をすることもできた。微妙な感じの時は折れ線グラフをチェックする感じで進められた。」といった回答があった。

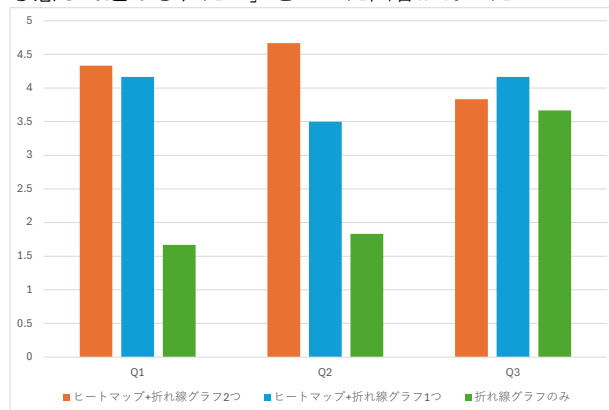


図8 実験2のアンケート結果

表4 実験2のアンケートの多重比較結果

	Q1	Q2
ヒートマップ+折れ線グラフ2つとヒートマップ+折れ線グラフ1つ	1.09	0.175
ヒートマップ+折れ線グラフ2つと折れ線グラフのみ	0.00437	0.0102
ヒートマップ+折れ線グラフ1つと折れ線グラフのみ	0.0202	0.0325

8. 議論

8.1. 実験

実験1の結果から, ヒートマップに折れ線グラフを1つ組み合わせた手法は折れ線グラフのみの可視化より2問とも平均二乗誤差が小さくなった。E1の最も高いピークの位置が指定された範囲を満たすデータ数を答える問題では, 折れ線グラフを1つ組み合わせた手法の平均二乗誤差は0.750, 折れ線グラフのみでは10.1となった。t検定の結果から, この2手法間のE1の平均二乗誤差には有意な差が見られた。折れ線グラフのみの可視化で精度が低下した理由は, 線が重なって読み取りづらくなったからだと考える。折れ線グラフを1つ組み合わせた手法を使う際は, スライドを調整し, ユーザが読み取りやすいクラス数に分けて, 線の重なりを減らしたのだと考える。E1の平均回答時間は, 折れ線グラフを1つ組み合わせた手法が95.3秒, 折れ線グラフのみが147秒となった。t検定の結果より有意な差が見られた。E1では花粉飛散量データにのみ焦点を当てたのに対し, E2では気温データと花粉量データの両方を読み取る質問を行った。折れ線グラフのみの可視化と折れ線グラフを1つ組み合わせた手法のどちらでも, E1よりE2のほうが平均二乗誤差が小さく, 平均回答時間が短くなった。この理由と

文 献

- [1] 高橋裕一, 川島茂人, 相川勝悟, "空中スギ花粉濃度に及ぼす地球温暖化の影響: 山形市とその周辺地域で得られた予測結果," *日本アレルギー学会誌*, vol. 45, no. 12, pp. 1270-1276, 1996.
- [2] R. Endo and H. Hosobe, "Visualization of Time Series Data Using Clustered Heatmaps and Line Graphs," *Proceedings of the International Symposium on Visual Information Communication and Interaction (VINCI)*, no. 25, pp. 1-8, 2024.
- [3] Y. Tanaka, G. Angeliki, E. Henry, R. Raidou, E. Groller and T. Itoh, "Visualization of Relationships between Precipitation and River Water Levels," *Proceedings of the International Conference Information Visualisation (IV)*, pp. 58-63, 2024.
- [4] 本田健悟, 斎藤隆文, 藤代一成, "二色塗り分けスパークラインを用いた COVID-19 感染データの可視化," *情報処理学会全国大会論文集*, no. 1, pp. 89-90, 2021.
- [5] T. Saito, H. N. Miyamura, M. Yamamoto, H. Saito, Y. Hoshiya and T. Kaseda, "Two-Tone Pseudo Coloring: Compact Visualization for One-Dimensional Data," *Proceedings of the IEEE Symposium on Information Visualization (InfoVis)*, pp. 173-180, 2005.