時系列データに対する説明的可視化

Explanatory Visualization for Time Series Data

牧 修平○ (東京工業大学）

##### ABSTRACT

This template is a guide to prepare manuscript for Visualization Symposium Japan. Here, please write the abstract serving as an index and as a summary of the present paper. It should be as long as approximately 150 words.

データを理解しやすくするために様々な可視化手法が使われているが、データ構造の複雑さやデータの多さなどの要因のせいで、意図する情報を全てユーザが可視化から認識することは難しい。そこで可視化に加えてデータから文章を生成し、データに含まれる有益である可能性のある情報を、前もってユーザに提供するような研究が行われている。本論文のシステムでは、多次元の時系列データに関してデータの分布や変動の分析をサポートする可視化とともに、そのデータからテンプレートベースで生成した要約文を提供する。論文中では、時系列データの類似度の計算方法や時系列データを可視化した折れ線の形からデータを探索する方法、要約を生成するテンプレートの構成について説明する。そして最後にシステム全体をSocial Progress Indexというデータを用いたユースケースを通して説明する。

to English

# はじめに

生のデータを見るだけでは分析したり理解したりするのが難しいデータに対して、その分析や理解のために可視化を行う研究はたくさん行われてきている(Bryan, Ma, and Woodring 2016; Gschwandtner and Erhart 2018; Görtler et al. 2019; Sondag et al. 2020)。 しかし可視化によるデータの分析や理解に不慣れなユーザや可視化だけでは理解が難しいデータを解析したいユーザに対しては、データの可視化だけでその目的を達成することができない可能性がある(Ynnerman, Löwgren, and Tibell 2018)。 そこで文章と可視化の二つの方法を相互的に使ってデータの説明をする。**説明的可視化**が研究されてる(Hosokawa, Arimoto, and Wakita 2020; Latif and Beck 2019)。

時系列データは日常の気温や降水量、近年では新型コロナ感謝数の推移など様々な場面で扱われているデータである。 時系列データは必ずしも同様の事柄に対して一つのデータだけではなく複数のデータが存在することがあり(新型コロナ感謝数の推移など)、このような時系列データは可視化による理解が困難になる。 しかしそのような場面で、今まで説明的可視化を使った研究は行われてきていない。 そこで本研究では[Social Progress Imperative](https://www.socialprogress.org/)が10年にわたって発表しているSocial Progress Indexのデータ(Social Progress Imperative n.d.)(以下SPIデータ)を例として、説明的可視化による時系列データの理解を手伝う手法を提案する。

2章では説明的可視化と時系列データの可視化について関連研究をまとめ、3章でシステムの設計について述べる。 4章ではシステムのユースケースを通してシステムの有効性を説明し、5章と6章で考察とまとめを行う。

# 関連研究

この章では、今まで研究されてきた説明的可視化や時系列データに対する可視化をみていく。

**説明的可視化** Latifたちは(Latif and Beck 2019)、二つの地理統計変数を理解するためにiMRというシステムを開発し説明的可視化を行っている。 地図の領域の色やその領域の上にに描かれる円の大きさを使って二つの地理統計変数を可視化するとともに、それらのデータを解析して文章を自動生成している。 **図**は各州の年間の死亡者数、ハリケーンの発生数に対して実際にiMRのシステムを使用した例である。 生成されている文章には、死亡者数の平均値や最大値・最小値など、統計的なデータをはじめ、「南部の州が他の週より死亡者数やハリケーンの発生数が多い」などデータからは読みとりづらい情報も組み込まれている。 iMRでは文章生成を行うプロセスを自身で定義し、ユーザーの要望に合わせてパラメータを変更することで柔軟な文章を生成している。 本研究ではデータからは読み取りづらい情報を文章にしてユーザーに伝えるために、この論文の文章生成プロセスを参考に文章を生成するプロセスを定める。

Hosokawaたちは(Hosokawa, Arimoto, and Wakita 2020)、多変数の地理的統計量に対して指標の階層構造や地域的な階層構造に着目して説明的可視化を行っている。 HosokawaたちはSPIデータを使用しているがある年についてのみ注目して、時系列ではない二つの階層構造が含まれるデータとしてSPIデータを扱っている。 本研究では時系列ということに注目してシステムを作る。 またHosokawaたちとはことなり、地理的階層構造についてのみに着目しており、SPIの指標の階層構造についての説明的可視化は行っていない。 **図**

**時系列データの可視化** 時系列データの情報を伝えるために可視化を使った研究はいくつか存在しており(Bryan, Ma, and Woodring 2016)、その中でJunhuaたち(Lu et al. 2020)は、二つの地理統計変数データを二変数マップと動画を用いて可視化している。 **図**では横軸を収入、縦軸を平均寿命とした例で、ドットの動きによって過去100年で世界各国の収入や平均寿命がどのように変化していったかが分かるようになっている。 この研究では、時系列データを事前に解析し時系列データの変化の傾向が変化した期間を特定して、その期間を下のビューでユーザーに伝えている。 本研究ではこの研究を参考に、タイムステップ数が10と、データが少ない時系列データに対して傾向の変化があるタイミングを特定し、可視化と文章に傾向の変化の情報を取り入れる。

# システムの設計

時系列データの分析のための可視化を表示しつつ、情報を手に入れたい地理的範囲や指標の選択に応じた情報を文章としてユーザに提供するためにシステムを設計した。 この章ではそのシステムの要件や、実際にどのように設計したのかを説明していく。 SPIデータを分析するときに、ユーザーは地理的な階層(東アジアの中の日本やヨーロッパの中のイギリスなど)の情報を含んでデータを分析したいと仮定している。 しかしSPIデータには地理的な階層の情報が含まれていないので、[国際連合](https://www.un.org/)が提供している、地理的な階層についてのデータ(UNSD 2003)を追加で扱う。

## 要件

実際にシステムを作る前に、以下のように要件を定めた。

* **R1 変化の特定** 指標の値が極端に低い国に注目するときに表示する地理的範囲が広い(表示する国数が多い)と、注目する国のデータが潰れて表示されてしまい、そのデータの分析は困難になる。 そこで10年の時系列データの中で傾向の変化があった年を特定して、データのの概要としてユーザーに提供する必要がある。
* **R2 階層的な情報** 生データをそのまま分析するとき、個々の国についての情報を分析することはできないが、その階層的な見方はなかなか難しい。 そのため、階層的な情報を提供する必要がある。
* **R3 地理的範囲の表示** 地理的階層の情報を伝える際にどの国がその範囲に含まれているかをユーザーが知っている必要がある。 しかし、その範囲に含まれる国を全てユーザーが知っていることはほとんどない。 また注目したいデータがどこの国かを確認することが難しい。 そのため、選択する地理的範囲や国位置を視覚的にユーザーに提供する必要がある。
* **R4 概要** 複数のデータに対し、時系列データの変化や地理情報、階層的な情報などを視覚的な表現によって理解することは困難である。 そこでその表示しているデータの概要を、文章によってユーザーに提供する必要がある。

## データの可視化

### 時系列の可視化

時系列データの可視化は折れ線やヒートマップを使って行うのが一般的である。 本論文では折れ線を用いてSPIデータの時間的変化を表現する。 しかし注目する地域的範囲が広いときに、ほとんどの折れ線では時間的な変動を認識することができない。 注目する国の概要やその国データと類似した国データをハイライトする。 注目する国に関してはデータの傾向が変化する点をドットを用いて表現しつつ(**R1,R4**)、色を他の表現と異なるマゼンタ色を使う。 注目する国と類似する国のデータに関してダークカーキ色を使う。

**傾向の変化する点** 時系列データの傾向の変化が会った時、その変化があった点を特定してデータの概要としてユーザーに提供したい。

まず何かしらの変化があるのか、それとも単調な動きをしているのかを判定する。 時系列データを線形回帰し一次の直線を求める。 その直線と元のデータに対し全ての年でデータの差が閾値より小さい場合には、傾向は変化していないとして判定する。 一方どこかの年でデータの差が閾値以上であれば傾向の変化があったろ判定する。

データの傾向が変化する点の求める方法は、Junhuaたちと同様にWin法(Truong, Oudre, and Vayatis 2020)を適用する。 時系列データをとして初めに、に対しての線形回帰直線の傾きを求め、に対しての線形回帰直線の傾きを求める。 次に、に対してを求める。 そして両方ともに対して、が注目する指標に対するSPIデータの平均値を元にした閾値以上であるとき、の点をデータの傾向が変化した点として捉える。 これは、ある点をを中心にして左右の直近のデータの傾きの差が大きい点を傾向の変化があった点としている。 は短い時間間隔でのデータの変化を捉えることができ、は少し長い時間間隔でのデータの変化を捉えることができる。

傾向が変化する点を抽出した後、折れ線にその情報を表現する。 できるだけ簡単な表現にするため、ドットを用いる。 傾向の変化がその指標の中でいい変化であれば白を、悪い変化であれば黒をドットの色に使って表現する。 複数傾向の変化があった点が存在する場合、その変化の大きさをドットの大きさを使って表現する。

**類似する国** 類似する国を定義する。 初めに注目する国と他の国の時系列データで相関係数が大きい国を集める。 その集合の中で、注目する国と他の国データでユークリッド距離を計算する。 その距離が注目する指標に対するSPIデータの平均値を元にした閾値未満であれば、類似する国とする。 相関係数は、二つの時系列データの上昇や現象などの大まかな変動が一致していることを条件にするために使用している。 その後時系列データを次元ベクトルと考え、ユークリッド距離が近いデータを選択することで、注目する国と値が近く似た変動をする国を抽出することができる。

### 地域の可視化

選択された地域的範囲をユーザーが難なく理解できるように、世界地図を用いてその範囲を表示する(**R3**)。 選択された地理的範囲に含まれていない国々と、データが存在しない国は灰色で表される。 注目する国については、折れ線と同様の色で表現する。 これらによって、ユーザーが注目する国と地理的範囲の国々とで地理的比較が可能になっている。

## 文章生成

データの可視化だけでは伝えきれない情報や、可視化で示したい情報を含んだデータの概要を文章で伝える(**R4**)。 文章はLatifたちやHosokawaたちと同様に、テンプレートベースで生成する。 以下では以下の情報をデータから抽出して文章にしていく。

* 選択した地理的範囲に含まれる国々が、世界からみて指標の値が低いか高いか。 例: アメリカには世界の国々の中で社会進歩指標が低い国から高い国まで幅広く存在する。
* 選択した地理的範囲の国々で、指標の値が高い、もしくは低い国が下の階層の地理的範囲で偏りがあるか(**R2**)。 例: アメリカの中では北アメリカに社会進歩指標が高い国が多い。
* 注目する国指標に対する値が、選択した地理的範囲の国々の中で低いか高いか。 例: コスタリカはアメリカの国々の中で社会進歩指標が高い。
* 注目する国データはどのような動向があったか。 例: コスタリカの社会進歩指標の値は、2011年から2020年にかけてほとんど横ばいである。
* 選択した地理的範囲の国々の中で、注目する国と似たデータの動向があった国について。 例: コスタリカと似た動向のある国は、チリとウルグアイである。

指標、注目する国、地理的範囲から以上のよう無観点でデータを分析して文章を生成した。

## インタラクション

ユーザーがより可視化からデータに関する情報を手に入れやすいようにインタラクションを導入する。

**ズーム/フィルタリング** 折れ線の表示において、注目したい国や地理的範囲のデータの変動をより詳細に観察したいときのために、ボタンによって縦軸の範囲を調整できるようにしている。選択した地理的範囲がアジアや南アメリカなど世界全体ではないときは、その選択した地理的範囲のデータだけを表す縦軸にするか世界のデータの範囲を縦軸にするかを調整することができる。 選択した国のデータだけに興味があるときは、縦軸をデータの最小値~最大値に調整することができる。 選択した地理的範囲を見ることは、世界から比べて極端に値が小さい、または大きい国しかその地理的範囲に存在しないときに、その国々のデータを確認するのに役立つ。 一方世界のデータの範囲を縦軸にすることで、世界から比べて極端に値の低い、または高い国がその選択した地理的範囲に含まれていることなどを知ることができる。

表示範囲の調整を行うズームだけでなく、表示する折れ線をフィルタリングするインタラクションを追加している。 注目する国、注目する国と類似している国、その他の国のデータそれぞれで、表示するか非表示にするか選ぶことができるようにてしている。 類似している国が多いときにそれらを非表示にすることにより、注目する国と類似していない国を目立たせることができる。 一方、類似している国に注目したいときには、類似していない国を非表示にすることができる。

**クエリ** 時系列データを見るとき、何かしら特徴的な動きから探索したいかもしれない。 その時のために、ユーザーのフリーハンドや用意した折れ線の形を入力としてデータを探索できるようにした。 このクエリでは時系列データの形に着目する。 入力されたデータの平均値が各国のデータの平均値と一緒になるように、データを平行移動させる。 その後は類似した国を見つける時と同様に、相関係数とユークリッド距離が閾値以内であればその国を出力に入れる。 このようにしてフリーハンドや用意した折れ線の形から、似た形をしたデータを探索することができる。

**関連付け** 折れ線、地理表現、文章と三つの方法でデータの情報を提供しているが、同じデータに関するものは同じ色で表現すことでデータの理解を促進できる。 例えば注目する国がどのようなデータを持っているかということと、地理的にどのような場所にある国かということをわかりやすくするために、折れ線と地理表現ではともにマゼンタを使って表現している。 また文章の中でこの国のことを述べるときにも、国名の色にマゼンタ使用している。 類似する国についても同様にダークカーキ色が使用されている。

**キー選択** ユーザーが欲しい情報を得ることができるように、データを抽出するいくつかのキーを選択できるようにした。 注目する国、地理的範囲、SPIの指標をボタンを使用して簡単に選択できるようにした。 注目する国については、地理的な位置や折れ線から選択できるようにしている。 これによってある国から地理的に近い国や、データが近い国などといった視点でデータを探索できるようになっている。

**図**

# ユースケース

この章では提案するシステムを使って四つの要件、**R1: 変化の特定, R2: 階層的な情報, R3: 地理的範囲の表示, R4: 概要**が満たされているかを確認していく。

**図**は地理的範囲を**南北アメリカ**、指標を**個人の権利**、注目する国を**アメリカ合衆国**にしている。 折れ線からはアメリカ合衆国は南北アメリカの中で最も個人の権利という指標がの値が高いこと、10年間徐々に減少していることがわかる。 またこのとき南北アメリカ内でアメリカ合衆国と類似した動向をした国はなかった。 文章の方をみていく。(**文章1**) 南北アメリカの国々は指標の値が高い国から低い国まで存在することがわかる 。 確かに折れ線を確認すると指標の高い国が多いことがすぐにわかるが、キューバが他の国と比べて極端に低い値をとっていることがわかる。 また、南北アメリカの中で北アメリカには指標の値が高い国が多いことがわかる(**R2**)。**図** このことは選択する地理的範囲を北アメリカに変更すると、実際に確認できる 次にアメリカが南北アメリカの中で指標の値が高いことや、10年間緩やかに減少していることがわかる。(**R4**)

他の国や指標についてもみていく。 **図**は地理的範囲を**世界**、指標を**情報通信へのアクセス**、注目する国を**日本**にしている。 先ほどと異なってわかる情報として、2013年に指標の値の変化の傾向が変わったことが折れ線からわかる(**R1**)。 こととき、情報通信へのアクセスが伸びることは良いことなので、その上昇率が下がることはよくない変化として黒塗りでドットに表されている。 他にはわかる情報として、似たような動向をしている国に、オーストラリアがあることがわかる。 類似していない国の表示を隠すと、日本とオーストラリアのデータは非常に似ていることが確認できる。(**図**) 次に文章の方をみていく(**文章2**)。 先ほどと同様に世界の国々の中で、オセアニアには指標の値が低い国が多く、ヨーロッパには指標の値が高い国が多いことがわかる。 国の情報として、日本は2011年から2013年にかけてこの指標が大きく上昇したことを文章からも確認できる。

最後に別のアプローチでデータを探索していく。 指標を**個人の安全**とし地理的範囲を**ヨーロッパ**とする。 まず折れ線をみると一番データの値が低い国があるのがわかる。 ホバーするとそれがロシアであることがわかり、タップしてみると注目する国が**ロシア**になり、文章として詳細な情報を得ることができる。 文章から、やはりロシアは指標の値がヨーロッパの中で非常に低いことがわかる(**文章3**)。 次にロシアと隣接している国の指標の値が気になり、ロシアの北西に位置する**フィンランド**を地図から選択してみる。 するとフィンランドはロシアとは異なり、指標の値が高いことがわかる。 また地図より、フィンランドと値が近くデータの変動が似ている国がヨーロッパの西から北に固まっていることがわかった(**R3**)。 一方、西側のベルギーという国はフィンランドと値が近くデータの変動が似ている国に含まれていない。 そこでベルギーを地図から選択すると、ベルギー2015年までは周辺諸国と同じような値をとっていたが、2016年で大きく減少していることがわかる。**図**

# 考察

この章では3章で定めたシステム設計の四つの要件をどのように満たしているのか、システムの利点や欠点、改善点などを議論していく。

## 要件に対する充足

3章でシステムが満たすべき四つの要件を定めました。

* **R1 変化の特定** データを解析することで、データの傾向の変化するタイミングを特定し、折れ線の中にドットして表現した。また文章の中では、データの変化の仕方を提供している。
* **R2 階層的な情報** 個人の権利という指標で、北アメリカの国々が南北アメリカに対して値が高い国が多いことがわかるように、文章の中で地理的階層の情報を提供できるようにした。
* **R3 地理的範囲の表示** 地図の国色を使って折れ線で表されている国のデータを関連付たことにより、ヨーロッパの西から北部の国は個人の安全の値が高いことがわかった。 また、その後ベルギーのデータが周辺の国とは似た動向をしていないことを発見した。
* **R4 概要** 文章の中に、地理的階層の情報、注目する国データの分布や変動の情報、注目する国と類似する国名を含んでユーザに提供している。

## システムの利点,問題点,改善点

次にこのシステムの利点や問題点、改善点を考察していく。

### 利点

このシステムによって、データの分析における以下のような利点がある。

* 元のデータからは見つけづらかった地理的階層に沿った情報を文章による概要から入手することができた。 これは地理的範囲を変更することで実際にデータと示し合わせることができるが、一度に複数の小さな地理的範囲の情報を手にすることができるのは有効だと考えている。
* 国名からデータを探索するだけではなく、折れ線を用いたデータの形や分布、地図を用いた国場所からデータを探索することができる。 国名からデータを探索できることはもちろんだが、分析の過程でデータの形や、あるデータと地理的に近い国のデータの関係も分析したくなることは自然であり、そのニーズに対応している。
* データ分析にデータの類似した国を折れ線や地図を通して提供することで、次の分析への手がかりになった。

### 問題点

今回のシステムに利点がある一方で、以下のような問題点があった。

* 注目する国とその国に類似する国は地図上で色を変更したが、ガンビアのような国土面積が小さい国は色から場所を見つけることができない。 現在は囲まれた範囲を色付けしているが、小さな国に対しては、国を中心とした円を表示するなど別の表現方法が必要かもしれない。
* データの地理的分布の概要をみることができない。 現状地図を注目する国や類似した国、選択した地理的範囲を示すためにしか使用していないが、地図の色をデータの値を用いて色付けする方法が有効かもしれない。 その時は、注目する国や類似した国をどのように表現するかが新たな問題となる。
* データの形からは類似していないように見える国が類似する動向をしていると判断されてしまっている。 今後パラメータを調整して、類似していないような国がしっかり類似していないと判定されるようにする必要がある。
* 自由な分析を行うときに、手がかりとなりそうな情報がユーザにほとんど与えられておらず、有益な探索と分析を行うことができない。 Jinwookたち(**seo2005rank?**)は多くの図を図をさまざまな指標でランク付けをし、ランクをつけて提供する研究を行っている。 今回の私のシステムでも、全ての図を何かしらの基準で評価し、その結果をユーザーに提供する方法が有効かもしれない。

### 改善点

時系列データを説明的可視化を行うシステムを作成し、そのシステムの利点と問題点を今まで上げた。 この説ではそれらを踏まえて、このシステムの改善点を上ていく。

* 地図上の色の使い方の再考を行う。 問題点で述べたように、現在のシステムでは注目する国とその国と類似した国の場所を示すために国の色を使用しており、データの地理的分布がわからない。 国の位置も大事だが、データの地理的分布にも興味があるためにこの地図での色の使い方を再考する必要がある。 一つの案としては、データの値に沿ったカラースケールの色を国土にあて、注目する国とその国と類似する国という意味の色を国の境界線に当てる方法である。 もう一つの案は現在の表示と、データの値に沿ったカラースケールの色で国を色付けした表示とを切り替えることができるようにすることである。 前者は両方の情報を一度にみて取れるがやや注目する国とその国と類似する国の情報を受け取りづらい。 一方後者はユーザーのニーズに沿って効果的な表示を提供することができる。 また面積の小さな国位置がわかりづらい問題に対して、その国以外の国を薄暗くするなどのフィルターを用いることで解決できる可能性がある。
* 類似している国を判定する方法の再考を行う。 現状データの動向が類似していないように見えるデータが類似していると判定されてしまっている場合がある。 その状態を改善するために、判定に用いてるパラメータを調整する必要がある。 ただし現状正しく類似していると判定されているデータの判定が覆らないようにパラメータを調整することがこんなんであるため、別の判定方法がないか調べる必要がある。
* 分析の手がかりを提供する。 今回データが類似している国とその場所をユーザーに提供することで分析の補助をしていたが、有益な情報を持っていそうなデータに対する指標を定時したり、複数の指標の関係を示したりと行ったことは行わなかった。 しかし、データの自由な分析のためにはJinwookたちのように有益な情報を持っていそうな表示をユーザーに伝える必要がある。

## 今後の課題

# まとめ

まとめ

# 参　考　文　献

Bryan, Chris, Kwan-Liu Ma, and Jonathan Woodring. 2016. “Temporal Summary Images: An Approach to Narrative Visualization via Interactive Annotation Generation and Placement.” *IEEE Transactions on Visualization and Computer Graphics* 23 (1): 511–20.

Görtler, Jochen, Marc Spicker, Christoph Schulz, Daniel Weiskopf, and Oliver Deussen. 2019. “Stippling of 2d Scalar Fields.” *IEEE Transactions on Visualization and Computer Graphics* 25 (6): 2193–2204.

Gschwandtner, T., and Oliver Erhart. 2018. “Know Your Enemy: Identifying Quality Problems of Time Series Data.” *2018 IEEE Pacific Visualization Symposium (PacificVis)*, 205–14.

Hosokawa, Natsuki, Kohei Arimoto, and Ken Wakita. 2020. “A Scalable" Exploranation" Technique for Hierarchically Indexed Table Data.” In *Proceedings of the 13th International Symposium on Visual Information Communication and Interaction*, 1–5.

Latif, Shahid, and Fabian Beck. 2019. “Interactive Map Reports Summarizing Bivariate Geographic Data.” *Visual Informatics* 3 (1): 27–37.

Lu, Junhua, Jie Wang, Hui Ye, Yuhui Gu, Zhiyu Ding, Mingliang Xu, and Wei Chen. 2020. “Illustrating Changes in Time-Series Data with Data Video.” *IEEE Computer Graphics and Applications* 40 (2): 18–31.

Social Progress Imperative. n.d. “Social Progress Imperative.” <https://www.socialprogress.org/>. Accessed: 2020-10-10. Accessed October 10, 2020. <https://www.socialprogress.org/>.

Sondag, Max, Wouter Meulemans, Christoph Schulz, Kevin Verbeek, Daniel Weiskopf, and Bettina Speckmann. 2020. “Uncertainty Treemaps.” In *2020 IEEE Pacific Visualization Symposium (PacificVis)*, 111–20. IEEE.

Truong, Charles, Laurent Oudre, and Nicolas Vayatis. 2020. “Selective Review of Offline Change Point Detection Methods.” *Signal Processing* 167: 107299.

UNSD, UNdata. 2003. “United Nations Statistics Division.”

Ynnerman, A., J. Löwgren, and L. Tibell. 2018. “Exploranation: A New Science Communication Paradigm.” *IEEE Computer Graphics and Applications* 38 (3): 13–20.