Flickr の大量写真データを用いた『新たな人気スポット』の出現検出

倉田陽平

Detection of “Emerging Popular Spots” Using Flickr’s Massive Photo Data

Yohei KURATA

Abstract: Long-stored social big data will allow us to analyze not only popular spots among tourists, but also emerging/shrinking popular spots, by comparing the recent and past data of SNS-posting locations. This paper proposes several visualization techniques, which promote such comparison. The best one is a two-colored heat map, where each color represents past and recent SNS-posting locations, superimposed by an arrow mesh illustrating the difference of their volume at each location.

Keywords: 観光（tourism），写真撮影行動（photo-shooting behavior），Flickr，点分布（point distribution），ヒートマップ（heat map），母比率の検定（hypothesis testing for a proportion）

1. はじめに

ソーシャルビッグデータから観光客の行動・関心を探ろうとする取り組みが盛んになされている（たとえば Chen 2009, Shirai et al. 2012, 倉田 2013, 斎藤・横川 2013, 佐伯ほか 2015, 真田ほか 2015）．とりわけ国内では訪日外国人への関心の高さから，inbound insight[1]のような商用サービスも生まれ，観光庁[2]も取り組みを行っている．

ソーシャルビッグデータが長年蓄積してくると，人気スポットのみならず，人気が上昇／下降しつつあるスポットの検出も可能になると期待される．このような検出を行うには，過去の投稿箇所のヒートマップと直近一定期間の投稿箇所のヒートマップを並べ（図-1），目視によりその差分を見出すという素朴な方法が考えられる．しかし，二枚の地図から対応地点の色の違いを見出す作業は容易ではない．ではヒートマップを重ねれば良いかというと（図-2）．投稿量が減少した場所はなんとかわかるが，その逆はほぼ判別できない．

そこで本研究では「直近の SNS 投稿の空間分布が普段とどう異なるか」の目視判定を容易とする可視化手法を提案する．この手法は，本質的には，ある点分布とその部分集合の分布の差異を（点の総数の差を考慮した上で）可視化するものである．したがって，提案手法は「最近人気のスポット」だけではなく，「特定の時期に投稿が多い／少ない箇所の検知」や「一部の投稿者（たとえば外国人）による投稿が多い／少ない箇所」の検出などにも応用できよう．

図- 1 全期間（左）と直近（右）の投稿分布

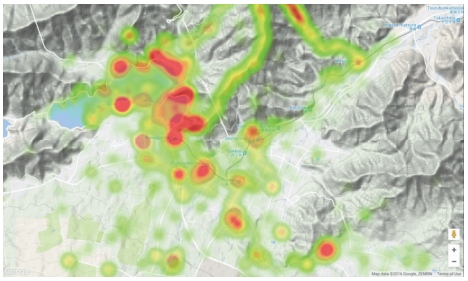

図- 2 図-1 のヒートマップの単純な重ね合わせ

なお，可視化手法の比較検討のため，本研究では富士吉田市周辺の Flickr 投稿データを用いる．ここを事例とするのは，2013 年の世界文化遺産登録にあわせ急速な観光整備が進み，近年訪日外国人が急増していることにより，SNS 投稿箇所に変

倉田陽平　〒192-0397　東京都八王子市南大沢 1-1
首都大学東京大学院都市環境科学研究科観光科学域
E-mail: ykurata@tmu.ac.jp

化が起きていると期待される．本研究では富士山駅から半径 5km の範囲で 2011 年から 2015 年に撮影された投稿写真, 合計 13,779 枚の撮影点分布を用い，そのうち 2015 年に撮影された写真 3,897 枚（28%）を「直近の写真」として議論を進める．

2. 関連研究

二種類の点分布の関係性を見る方法としては，相互最近隣距離法や相互 K 関数法（Ripley, 1981）がよく知られているが，今回の問題は片方の点分布がもう片方の点分布の部分集合であり，これらの相互点分布分析の手法には馴染まない．

Sadahiro & Masui (2004)は複数のサーフェスマップの構造的類似性を評価する手法を提案している．この手法を本問題に適用することも可能ではあるが，元来は多数の地図を一度に比較することが主眼の手法であるため，単純な二枚の地図比較では利点が失われてしまう．

3. 提案手法

本章では直近の投稿分布が普段とどう異なるかを可視化する幾つかの手法を検討する．説明のため, 知りうる過去全期間の投稿件数をn_{all}件, 直近 1 年間の投稿件数をn_{last}件とし，後者を前者で割ったものを直近比rとよぶ．

３．１　多色×モノクロヒートマップ

直近の投稿分布と普段の投稿分布は「図と地」の関係にあるという考えのもと，直近の投稿分布を多色ヒートマップで，全期間の投稿分布をモノクロヒートマップ（投稿量が多いほど無色透明，少ないほど黒色不透明のもの）で表現し，重ね合わせ表示する．これにより，普段通り投稿が多い所は多色ヒートマップの鮮やかな色で，普段より投稿が少ないところは無色透明で表現され，そも

図- 3　直近の投稿分布を示す多色ヒートマップと普段の投稿分布を示すモノクロヒートマップとの重ね合わせ

そも普段から投稿が少ないところは暗くなる．

この方法を富士吉田のデータに適用すると（図-3），たとえば河口湖駅の西側（図の中央左上側）において無色透明の部分があることから，このあたりの投稿が普段よりも少ないことがわかる．一方で，普段よりも投稿が多い箇所を見極めるのは難しいことがわかる．

３．２　二色のヒートマップの重ね合わせ

直近の投稿分布を透明～赤色半透明のヒートマップで，全期間の投稿分布を透明～青色半透明のヒートマップで表現し，重ね合わせ表示する．ただし，後者においては各点の重みを r（先述の直近比と同じ値）として補正する．この結果，直近の投稿が普段より多い箇所では赤色側が卓越し，普段より少ない箇所は青色側が卓越する．また，そもそも投稿量が少なければ透明度が上がり，投稿量が多ければ（赤，青，またはそれらの混色の）濃い色となる．

この方式を富士吉田のデータに適用すると（図-4），たとえば河口湖中央南岸や山梨県世界遺産センター（中央やや左）において赤色が卓越していることから，これらの地点での投稿が普段より多く，河口湖大橋南側で青色が卓越していることから，ここの投稿が普段より少ないとわかる．このように，この手法は普段に比べ投稿の多い／少ない地点がはっきり分かるという利点がある．しかし，各地点で投稿量が普段に比べどの程度多い／少ないか（変化量）までは把握が困難である．

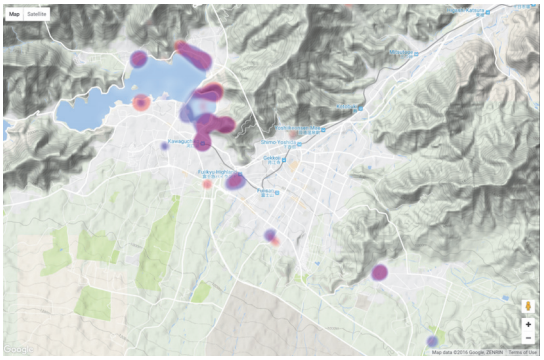

図- 4　直近の投稿分布を示す赤色ヒートマップと全期間の投稿分布を示す青色ヒートマップ（両者比較のため補正適用）の重ね合わせ

３．３　投稿の変化量のメッシュ表現

普段と比べた投稿の変化量を表現するには，「直近の期間も普段通りの投稿ペースだった」と仮定した場合の投稿量の期待値と実際の投稿量との差を可視化する，という方法が考えられる．具体的には，メッシュ分割を考え，セルc_{ij}内での

投稿総量が$n_{\mathrm{all}\,ij}$，直近の投稿量$n_{\mathrm{last}\,ij}$だとして，$n_{\mathrm{last}\,ij}-n_{\mathrm{all}\,ij}\times r$の値をもとにセルの彩色を行う．

この方法を富士吉田のデータに適用した（図-5）．ここでは無色透明の箇所は投稿量が普段通りであることを示し，赤色になるほど投稿量が普段よりも多く，青色になるほど投稿量が普段よりも少ないことが示されている．

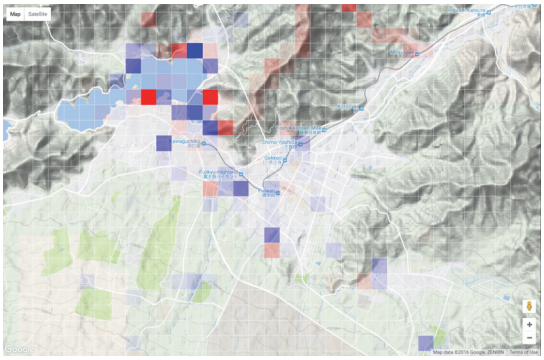

図- 5　普段に比べた投稿の変化量を示すメッシュマップ

３．４　投稿の変化量の有意性のメッシュ表現

セル A では普段 1000 件／年の投稿が直近一年で 1100 件となり，セル B では普段 100 件／年の投稿が直近一年で 200 件となったとすると，両セルとも変化量は＋100 件だが，セル B の変化の方が劇的である．このような変化こそ「流行の兆し」として検出されるべきであろう．そこで，このような劇的変化を評価するために，母比率の検定（縄田・松原 1991）を応用する．母比率の検定とは，得られた標本内で事象の起こる割合が，母集団内でその事象が起こる割合に一致するか否かを検定する手法である．今回の場合，「投稿傾向に局所的差異がない」という帰無仮説のもと，セルc_{ij}における投稿の直近比$r_{ij}=n_{\mathrm{last}\,ij}/n_{\mathrm{all}\,ij}$の期待値は全体の直近比$r$に一致し，さらに以下の検定統計量$Z$は$n_{\mathrm{all}\,ij}$が十分大きいとき（概ね 30 以上）は近似的に標準正規分布 N(0,1)に従う．

$$Z=\frac{r_{ij}-r}{\sqrt{r(1-r)/n_{\mathrm{all}\,ij}}}$$

これを利用して p 値を求め検定を行うことができる．今回は p 値に基づいてセルに彩色を行うことで，局所的な変化の劇的度合いを可視化する．

この方法を富士吉田のデータに適用した（図-6）．ここでは色の赤・青によって変化量の正負が，不透明度によって有意性（p 値の小ささ）が表現されている．３．３の結果と比べると，中央北部の山脈尾根上で赤色が濃い．これらの結果から，尾根上での投稿は，変化量自体はさほど多くないものの，普段の投稿自体が少なかったため，相対的に見れば劇的変化であることがわかる．

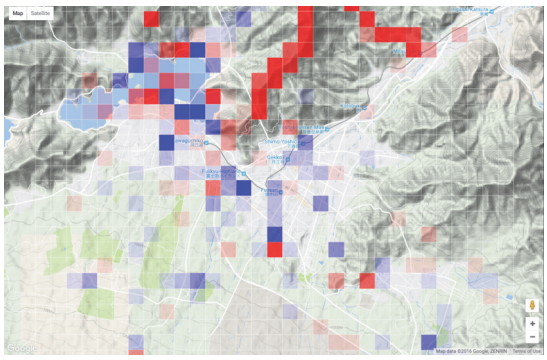

図- 6　普段に比べた投稿の変化量の有意性を示すメッシュマップ

３．５　矢印メッシュマップ，およびこれと他の表現との組み合わせ

３．３や３．４のように，変化量やその有意性をメッシュマップによって表現すると，普段や直近の投稿量の多寡の情報が抜け落ちる．そこで素朴な解決策として，普段や直近の投稿量をヒートマップで表現し，その上にメッシュマップを重ねるというものが考えられるが，実際にはこの結果は非常に見づらい．そこで，メッシュの各セルを塗りつぶす代わりに，各セルに投稿変化量を表現する以下のような「矢印記号」を配置することで，重ね合わせ可能なメッシュマップを作成する．

- 矢印の上下方向：投稿変化量の正負
- 矢印の大小：投稿変化量のボリューム
- 矢印の不透明度：投稿変化量の有意性

以上の方針で作成された矢印メッシュマップを図 7 に示す．この図では，どのあたりでどの規模の変化が起き，それが劇的なのか否かが明解に表現されている．また，この矢印メッシュマップと先の二色ヒートマップとを重ね合わせると（図 8），どの規模の変化がどの地点で起きているのかが一目瞭然になり，視認性も損なわれてはいない．

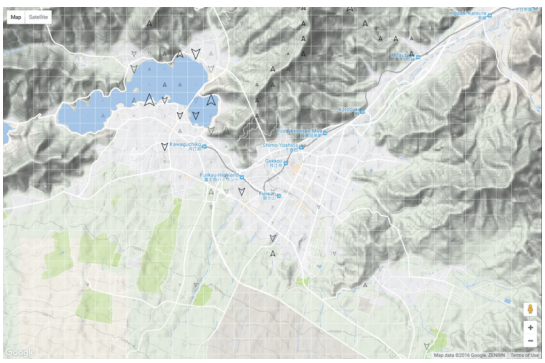

図- 7　投稿変化量の多寡とその有意性を示す矢印メッシュマップ

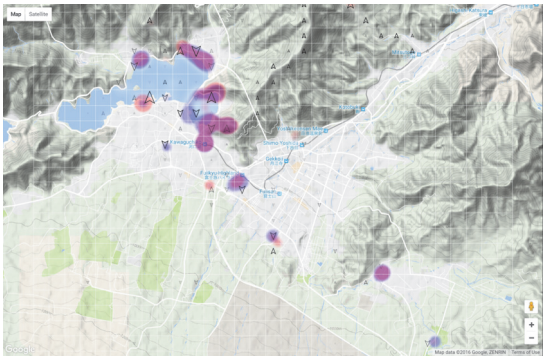

図- 8　矢印メッシュマップと二色ヒートマップとの重ね合わせ例

4. 手法の比較

今までの議論をまとめると，直近の投稿分布が普段とどう異なるかを可視化する際に考慮すべき要素は以下の五つである．

- 普段の投稿分布の視認性
- 直近の投稿分布の視認性
- 普段と投稿傾向が異なる箇所の視認性
- 投稿変化量の視認性
- 投稿変化量の劇的度合いの視認性

そして，各提案手法の各要素の達成具合を示したのが表-1 である．

表-1 で，各メッシュマップの「投稿相異箇所の視認性」が低いのは，メッシュの粒度と配置によって相異箇所の特定が困難な場合があるためである．この問題は矢印メッシュマップも抱えているが，矢印メッシュマップは重ね合わせが可能であるため，補正済み二色ヒートマップと重ね合わせればこの問題は解決できる．結果，この表からは「矢印メッシュマップと補正済み二色ヒートマップの重ね合わせ」が最も欠点のない可視化手法だと判断される．もっとも，当該手法はすべての情報を一度に盛り込んでいるため，目的に応じて知りたい要素が特定できる場合は，この表を元に別のシンプルな可視化手法についても検討すべきだろう．

表- 9　普段と直近との投稿分布の比較を促すために考慮すべき要素の，各可視化手法の達成状況

	普段の投稿量の視認性	直近の投稿量の視認性	投稿傾向相異箇所の視認性	投稿変化量の視認性	投稿変化量の劇的度の視認性
ヒートマップの単純な重ね合わせ（図-2）	△	△			
多色ヒートマップ×モノクロヒートマップ（図-3）	△	○	△		
補正済み二色ヒートマップの重ね合わせ（図-4）	○	○	○		
投稿変化量のメッシュマップ（図-5）			△	○	
投稿変化量の有意性のメッシュマップ（図-6）			△		○
矢印メッシュマップ（図-7）			△	○	○
矢印メッシュマップと二色ヒートマップの重ね合わせ（図-8）	○	○	○	○	○

4. おわりに

本研究では，Flickr の投稿分布の普段・直近比較を例に，点分布とその部分集合との分布傾向の差の判定を容易にする可視化手法を提案した．しかし，被験者実験を行っておらず，カーネル半径やメッシュサイズの影響についても今後の検討課題である．また，今回は目視判断を前提としたが，分布の相異箇所を自動抽出するようなデータマイニング手法や，それを応用した分析支援ツールについても研究を行っていくであろう．

謝辞

本研究には首都大学東京傾斜的研究費を利用した．

参考文献

倉田陽平 (2013) 観光ポテンシャルマップの信頼性向上に向けてーソースとなる投稿写真データの自動選別ルールの構築ー．地理情報システム学会研究発表大会講演論文集, 22, CD-ROM.

斎藤一・横川祥司 (2013) ツイート分析と感情語情報に基づくアプリケーション開発とその観光利用に関する研究．観光情報学会第 7 回研究発表会講演予稿集, 49-56.

佐伯圭介・遠藤雅樹・廣田雅春・倉田陽平・石川博 (2015) Twitter データを利用した訪日外国人の訪問先の言語別分析．観光と情報, 11(1), 45-56.

真田風・倉田陽平・相尚寿 (2015) 写真共有サイトに投稿された写真群を活用したテーマ別観光マップの作成．情報処理学会全国大会講演論文集, 77, DVD-ROM.

縄田和満・松原望 (1991) 仮説検定.「統計学入門」, 東京大学出版会.

Chen, W. C., Battestini, A., Gelfand, N., and Setlur, V., 2009. Visual Summaries of Popular Landmarks from Community Photo Collections. ACM Multimedia Conference.

Ripley, R. D., 1981. Spatial Statistics. New York: John Wiley.

Sadahiro, Y. and Masui, M. (2004) Analysis of Qualitative Similarity between Surfaces. Geography Analysis, 36(3), 217-233.

Shirai, M. Hirota, M., Yokoyama, S., Fukuta, N. and Ishikawa, H., 2012. Discovering Multiple HotSpots Using Geo-tagged Photographs. Advances in GIS, 490-493.

[1] http://inbound.nightley.jp/
[2] http://www.mlit.go.jp/kankocho/shisaku/kankochi/