HOW TO USE 人流データ

コロナ禍において、人流データに注目が集まっているものの、人流データを使った分析にはさまざまな制約があることは、あまり知られていない。
本稿では、人流データを使って分析を行う際の困難について述べ、その分析例として「ステイ・ホーム指数」を紹介する。

水野 貴之　Mizuno Takayuki
国立情報学研究所情報社会相関研究系准教授

著者紹介
2005 年、中央大学大学院理工学研究科博士課程修了、博士（理学）。日本学術振興会特別研究員、一橋大学経済研究所専任講師、筑波大学システム情報系准教授等を経て、2013 年より現職。同年より総合研究大学院大学複合科学研究科准教授も務める。論文：“A Stochastic Model for Order Book Dynamics in Online Product Markets,”（共著）*Evolutionary and Institutional Economics Review,* 10: 93-105, 2013.

1　はじめに

2020 年度ほど、人流データを身近に感じた年はなかったのではないか。人流データとは、人々の位置と時間が記録された、人々の移動が確認できる情報である。テレビをつければどのチャンネルでも「徹底的に人流を抑える…」「人流 8 割減…」「東京で減らない人流…」と、一般語化してしまった。

新型コロナウイルスの主な感染経路はツバなどの飛沫によるヒト・ヒト感染であるから、ヒトが動けば動くほど感染が拡大する。また、経済活動とはヒトが動くことであるから、家に籠もって動かなくなれば、観光業や飲食業は特に大きな打撃を受ける。つまり、人流を見るということは、感染状況や経済状況を知ることに繋がる。

経済活動をしたいが感染拡大が怖い、経済が耐えられるなら感染を抑え込みたい。この相反する要望のバランスをとるために、リアルタイムに感染状況と経済状況を知りたいが、潜伏期間があるために答え合わせは 2 週間後、景況感は集計に時間を要するので 1 カ月後と、どうしてもタイムラグが生じる。だから、現状を知るために、人流を計測することが大事なのである。

2　プライバシーと個人情報保護

テレビや政府の新型コロナウイルス感染症対策推進室ウェブサイト[1)]に映る人流の値が、バラバラであることにお気づきであろうか。もし、わが国が中国であれば、このような問題は発生しない。人々の位置が正確にわかるのである。感染者が発見されれば、感染者と周囲の人々の持つ携帯電話の移動履歴（携帯電話は常に最寄りの基地局と通信しているために、基地局に通信記録が残る）をチェックすることで、感染リスクの空間的な範囲を高精度で捉えて、ロックダウン等で封じ込めができる。医療における癌治療とも似ていて、癌だけを蛍光塗料で光らせて、そこだけ摘出すれば、健康な細胞への影響、つまり、経済活動への影響を最小限にとどめて、コロナを取り除ける。しかし、人流における単位は細胞ではなく、ヒトであるため、プライバシーや個人情報保護の問題が発生する。

位置情報そのものは個人情報ではないが、該当者が 1 人になるような位置情報は、他の情報と重ね合わせることによって容易に個人が特定できる恐れがあることから、プライバシーに該当し、個人情報に準じて扱う必要がある[2)]。同一時刻、同一空間に複数の人間は存在できないので、ある時刻における匿名のヒトの緯度・経度・高さ情報は保護対象である。空間を広くすれば解決できる問題でもない。ポツンと一軒家のように、周囲に誰もいなければ、該当者は 1 世帯になり保護対象である。多くの場合、余裕をもって、該当者数の閾値は 1 人ではなく、6 人や 6 世帯に設定して、ある地域が閾値以下であれば、その地域は秘匿とし、周辺地域と結合することで匿名性を担保する。リアルタイムの人流のスナップショットを得るには、この処理を毎時間行わなければいけない。

移動人数の計測になるとさらにハードルが上がる。午後 7 時、舞浜駅はディズニーランド帰りの観光客で溢れている。舞浜駅 19:00 の人口を計測することは問題ない。舞浜駅から京葉線に乗り、午後 7 時 20 分、多くの人は東京駅に到着する。舞浜駅 19:00→東京駅 19:20 の移動人数を計測することも問題ない。しかし、ここから一気に人々は散らばる。東海道線、山手線、京浜東北線、横須賀線、総武本線、中央線、中央本線、上野東京ライン、東北新幹線、上越新幹線、北陸新幹線、東海道新幹線、丸ノ内線、他にも経路はあるだろう。このように経路が別れた瞬間に、その条件、例えば、舞浜駅 19:00→東京駅 19:20→大宮駅 20:30 を満たす該当者が 6 人以下となり、移動人数を計測することはできない。時間を丸め込んで、なんとか 6 人以上にするという努力が常に必要である。

さらに、個人情報を特定してよいという許可をとっていない場合には、統計データを作る加工の過程で、分析者が 6 人以下の状態を見てしまうことにも問題がある。分析者が見ることができないように、移動履歴を暗号化したままの状態で演算処理する秘密計算を行わなければならない。このような状況なので、Go To トラベルと感染拡大の関係性を調べることは、一筋縄ではいかない。

個人情報の問題をクリアするもう一つの方法は、同意を取ることである。ただし、「あなたの携帯電話の位置情報をお客様のサービス向上のために利用することに、■同意します、

□同意しません」といった形で同意を取ればよいかというと、これでは個人情報保護法の同意要件を満たさない。利用目的を「できる限り特定しなければならない」のである。提供する位置情報がどのような事業の用に供され、どのような目的で利用されるのかが、一般的かつ合理的に想定できる程度に特定して同意を取らなければいけない。まさか、新型コロナウイルスが発生して人流データが必要になるとは、コロナ前には想像していなかったので、データは通信事業者に存在するが、移動と感染拡大の関係を統計的に調べるために使ってよいという同意は取っていない。

現状、同意のある人流データは、複数の利用目的で同意が取れ、そして、いつでも同意を簡単に取り消せるオプトアウトの仕組みが組み込まれている。同意のハードルが高く、人口に対してわずか数十万人のカバレッジである。しかも、このようなデータは、1日を跨いで移動を追跡できないように毎日IDが変化し、自宅が特定されないように住宅地ではデータが消えるなど、提供者への配慮がなされている。サンプルバイアスに目をつぶったとしても、移動と感染拡大の関係を調べることは、一筋縄ではいかないことがわかるであろう。

3　経済・感染症データサイエンティスト

さまざまな制約のある人流データを分析するのは、本当に難しい。政策担当者は、携帯電話から収集された位置情報を測るだけなので、人流データを持つデータベンダー（ここでは、通信事業者）に協力を仰ぐだけでよいと思ったかもしれない。まさか、サンプルバイアスや異なる匿名加工処理によって、データによっては、人口が20倍以上違う地域が発生しているとは思っていないだろう（菅他 2019）。そのため、テレビや政府の新型コロナウイルス感染症対策推進室ウェブサイトに映る人流の値がバラバラなのである。これを見ながら、2020年4月は日々の人流減（後の接触減）に一喜一憂していたのである。

政府の個票調査とは違い、統計を作るために集められたデータではなく日常の業務で生成されたビッグデータには、人流データに限らず、必ずイレギュラーなノイズが入る。ビッグデータを統計に活用する場合には、外れ値処理をしたり、オーバーサンプリングをしたり、ノイズが観測したい統計量に影響を与えないように、データを加工することが必須である。そのためには、生データの統計的な理解と、政策担当者の要望をデータ処理に落とし込める、経済や感染症畑のデータサイエンティストが、データベンダーと政策担当者との間に入り活躍する必要がある。

4　人出と外出状況の把握

経済のデータサイエンティストとして、制約のある人流データから人出と外出状況を把握する方法を示していく。商業統計から観測できる地域の店舗の売上総額と昼間人口とには、ほぼ比例関係がある。また、商業施設を含む不動産投資信託の予測のために、携帯電話

で捕捉した来店者数が使われたりしている。商業地や繁華街の人口をリアルタイムで観測することで、その地域の経済状況を把握し、困窮している店舗への経済対策を考えることができる。

マスメディア等では、商業地の人出ばかりを観測しているが、感染状況を把握する上で注意しなければいけないことがある。小売などの経済活動は、主に商業地で行われており、外出の自粛により被害を受ける店舗は主に商業地にある。つまり、人出の減少から、それぞれの店舗の被害がわかり、被害に沿った補償や支援が可能である。一方で、感染は、人々が多く集まる商業地で広まっている可能性があるが、人々は商業地には住んではいない。人出の多い商業地を規制したところで、人々が別の商業地に外出してしまうと、感染症対策としては不十分である。感染症対策で必要な人流に関する情報は、感染症がどこで広まっているかを知るための商業地の人出だけではなく、感染者がどこの地域にいるかを知るための居住地別の外出状況が必要である。商業地の人出の観測だけでは、どの居住地の人が出歩いているのかが不明なため、広範囲のエリア、例えば県民全員に対して自粛を要請するしかならず、それでは、十分に自粛できている地域にとっては過剰な要請になり、逆に自粛ができていない地域にとっては過小な要請になる。

商業地の人出とは、その商業地に住んでいる人を除いた各日時の人口である。商業地に住んでいる人は稀なため、人出≒人口となる。この定義を理解しておかないと、経済的な損失の見積もりに失敗する。5月8日、別府市が「観光客が減少している実態とかけ離れている」と抗議し、新型コロナウイルス感染症対策推進室ウェブサイトから「別府駅」の人出が削除された。別府駅周辺には住宅が密集しており、自粛で人口が増加したのである。各地の外出状況を知るためにも、居住者のうちの何人が残っているかを知りたいのであるが、居住地域別の人口のデータセットを構築するには許諾をとったり、居住地域のエリアを広げたりと、工夫が必要になってくる。

では、リアルタイムの人口分布だけが与えられたとき、人出と外出状況が観測可能な地域はどこであろうか。人出については、国勢調査で把握した居住者人口に比べて、夜間を除いて人口が大きく増える地域である。歌舞伎町や渋谷駅周辺などの商業地が該当する。逆に、外出状況ついては、国勢調査で把握した居住者人口に比べて、夜間を除いて人口が大きく減る地域である。外出先の目的地となる施設等の少ない住宅地が該当する。それら以外の地域は、極端な話、居住者は全員外出して、同数の別の地域の人々が訪れている可能性があり、人出や外出状況を推定することはできない。

5　ステイ・ホーム指数

われわれは、ドコモの携帯電話約 8,000 万台の基地局情報から推定された1時間ごとの人口統計である「モバイル空間統計®」の「国内人口分布統計（リアルタイム版）」[3)]を利用して、2020 年1月1日より各地の繁華街の人出と住宅街からの外出について、それぞれ、

性別・年代別に減少率を算出して公開している[4)]。

日本全国を500m四方のメッシュに区切り、コロナ前の2020年1月における各メッシュの昼間人口と居住者人口（夜間人口）を比較し、「昼間人口／居住者人口＞1.5」のメッシュを商業地、「昼間人口／居住者人口＜0.8」のメッシュを住宅地と定義した。ここで定義した商業地と住宅地では、多くの人がコロナ禍も相まって家に籠もった、関東で大雪の降った2020年3月29日の昼間人口が、居住者人口とほぼ同じとなった。つまり、居住者人口に対する日々の昼間人口の比率を観測することにより、人出や外出状況を測定することができる。

住宅地からの外出の自粛を観測するために、住宅地での「外出者数＝居住者人口－各時刻の人口」を毎時見積もり、ある日の9時から18時までの「外出の自粛率＝1－（その日のその時間帯の平均外出者数）／（平常時のその時間帯の平均外出者数）」を観測することで、ステイ・ホーム指数として定量化した（水野・大西・渡辺 2020、Mizuno, Ohnishi and Watanabe 2021）。

例えば、ステイ・ホーム指数が0.6なら、これまで外出していた100人のうちで60人が外出を自粛していることを意味する。同様に、商業地への外出の自粛を観測するために、商業地での「流入者数＝各時刻の人口－居住者人口」を毎時見積もり、ある日の9時から18時までの「商業地への外出の自粛率＝1－（その日のその時間帯の平均流入者数）／（平常時のその時間帯の平均流入者数）」を観測することで、ステイ・ホーム指数（商業地）として定量化した。

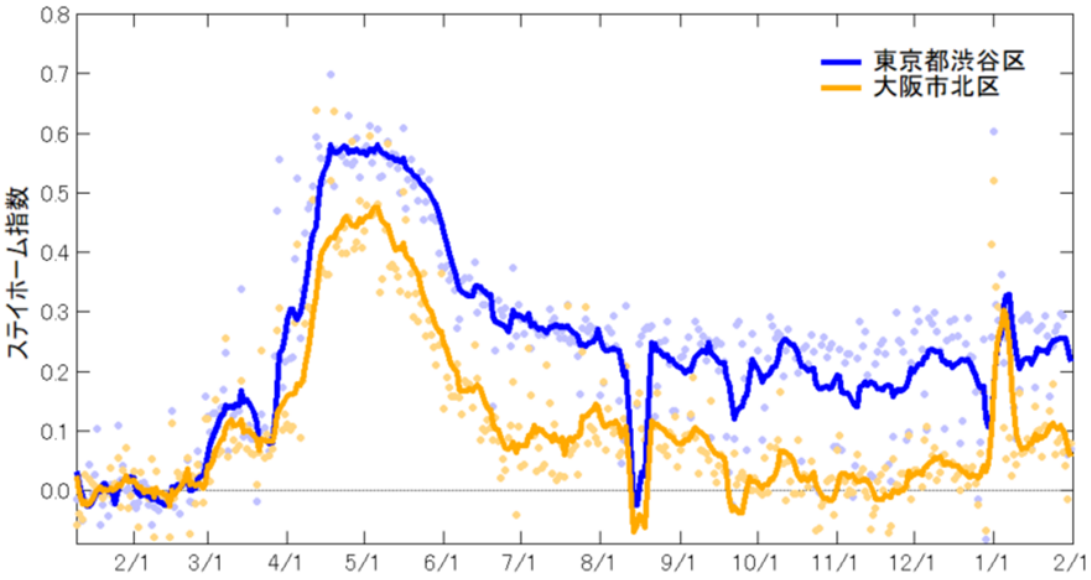
プライバシーが保護された人流データであっても、このように加工することで各地の自粛状況が定量化できる。感染率との対比で外出の自粛率が低い地域を見つけることで、自治体による地域の実状にあわせた自粛要請のサポートが可能になる。われわれは市区町村レベルでの指数を公開しているが、最小500m四方でも算出可能である。あまり狭くすると、地域差別を招く可能性があるので、バランスが重要である。

ステイ・ホーム指数を観測すると、第一波の時期では、各地で、ステイ・ホーム指数はステイ・ホーム指数（繁華街）と比べると10ポイント近く小さくなっていた。人々は繁華街を避けて外出していたのである。当時は、行き先ではなく外出自体を自粛しようという世論であったため、繁華街の人出だけを観測していたことには問題があったと言える。若者が自粛できていないとターゲットにされることが多い。しかし、世代別でステイ・ホーム指数を観測すると、中年と若者に大きな差はない。同じだけ頑張っているのである。ただ、もともとの外出数が多いため、若者は一層の努力が必要なのである。

最後に、地域差について言及しよう。同じ都内であっても、ステイ・ホーム指数は大きく異なる。最も高い指数は4月26日の品川区で、0.82と8割超えであった。一方で、その日の足立区では0.61と20ポイント近い開きがある。20ポイントも異なれば、感染や経済の被害も異なる。感染対策や補償は、エビデンスを揃えて地域の実情にあわせなければ不公平感も現れ、頑張る人のインセンティブを削いでしまう。

図１は、東京都渋谷区と大阪市北区のステイ・ホーム指数である。東京の感染拡大は早いために、都民の自粛が批判にさらされることがあるが、第一波から現在まで、常に東京の指数が高い。ここから言えることは、東京の都心は人口が多いので、大阪よりも一層の努力が必要だということである。また、大阪では秋口にゼロ付近まで指数が下がってしまったことから、テレワークの定着が十分に進んでいないことが読み取れる。他には、自粛の呼びかけが効きにくくなっていることや、政策の開始よりも世論の雰囲気などのアナウンス効果のほうが、人出を変化させていることもわかる。

図１　東京都渋谷区と大阪市北区のステイ・ホーム指数



注）実線は７日間移動平均で、点は日次指数である。

詳細にステイ・ホームの状況を観測することは、感染症対策にも経済対策にも役に立つ。これを実現するためには、データベンダーが提供した制約のあるさまざまな人流データから、精度の高い重要な情報を抽出して政策担当者に渡す、経済や感染症畑のデータサイエンティストが必要なのである。2020 年 1 月 1 日から 2021 年 3 月 31 日までのステイ・ホーム指数は、オープンデータ化、およびウェブサイトでインタラクティブに可視化してある（注４にあげたウェブサイト「外出の自粛率の見える化」）。人流と経済・感染との関係に興味のある方は、ご活用ください。

注

１）内閣官房新型コロナウイルス感染症対策推進室のウェブサイト
（https://corona.go.jp/）。
２）総務省の個人情報保護・位置情報に関するウェブサイト
（https://www.soumu.go.jp/main_sosiki/joho_tsusin/d_syohi/privacy.html）より。
３）国内人口分布統計（リアルタイム版）モバイル空間統計®のウェブサイト
（https://mobaku.jp/）。
４）水野研究室「外出の自粛率の見える化」
（http://research.nii.ac.jp/~mizuno/covid19.html）。

参考文献

菅愛子他（2019）「東京都における流動人口データの有効性の検証」総務省統計委員会担当室ワーキングペーパー、2019-WP03。
水野貴之、 大西立顕、渡辺努（2020）「流動人口ビッグデータによる外出の自粛率の見える化」『人工知能』35 巻 5 号、667-672 頁。
Mizuno, T., Ohnisi, T. and Watanabe, T.（2021）“Visualizing Social and Behavior Change due to the Outbreak of COVID-19 using Mobile Phone Location Data,”Submitted to *New Generation Computing*.