労働分野におけるオルタナティブデータの活用

森脇　大輔
（株式会社サイバーエージェントリサーチサイエンティスト）

1　さまざまなオルタナティブデータ

労働分野におけるオルタナティブデータ活用は決して新しいトピックではない。10年前，GoogleのチーフエコノミストであるヴァリアンがGoogle検索データによる失業予測論文を発表したが，同じような試みはそのさらに前から行われてきた（Choi and Varian 2012；Askitas and Zimmerman 2009；Suhoy 2009；D'Amuri and Marcucci 2010）。最近でも新型コロナによる労働市場への影響分析がトップジャーナルに刊行されるなど検索データによる労働市場分析は今なお活発な研究分野である（Caperna et al. 2022）。

近年ではデータソースの多様化が進んでいる。もっとも活発なのはインターネットを通じて公開される求人情報を収集したデータセットを用いた研究であろう（Forsythe et al. 2020；Ghoshsamaddar, Marchetti and Sevcenko 2021）。求人データはサンプルサイズが非常に大きいことや求人情報の内容がリッチであること，実際に働いた人のレビューがあること，日ごとに入れ替えがあるといった公的統計にはない情報がある。新型コロナによる労働市場への子細な影響を分析したり，企業文化が従業員に与える影響などのユニークな分析が可能になっている。また，OECD（2021）は求人情報に掲載される数万のスキルを公的機関が定義した61の大まかなカテゴリに分類という試みを行っている。これは，オルタナティブデータを公的統計と併用するために不可欠の取組であり注目される。

さらに，Zhang et al.（2021）は求人データを用いた企業・スキル別の需要予測モデルを提案している。こうしたモデルを利用することで，労働者のリスキリングや企業の採用戦略の効率があがり，労働市場の機能が向上する可能性がある。

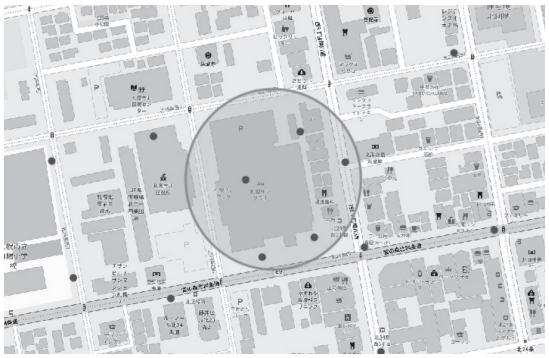
わが国においてもFukui, Kikuchi and Goalist（2020）が，求人データを活用してパンデミック下の労働市場分析を行っている。LinkedInなどのプラットフォームはデータ収集に止まらず推薦を行うことで求職行動に対する介入が実施できる点で労働経済学研究の新たな地平を開いている（Shi et al. 2020）。また，Upworkのようなクラウドソーシングプラットフォームのデータは，ギグエコノミーやリモートワークといった新たな労働現象の分析に用いられている（Huang et al. 2020；Ozimek 2020）。

2　位置情報を用いた失業率予測

ウェブデータの外に目を向けると，携帯電話の位置情報も活用されている。Bonato et al.（2020）は携帯電話から得た人流データをもとに，ロックダウンによりイタリアの地域的労働市場が分割されていったことを示した。ここでは，オルタナティブデータを用いた労働市場分析の例として，Moriwaki（2020）及びそれを拡張したToda, Moriwaki and Ota（2021）について解説する。研究のモチベーションは『労働力調査』による失業率の公表が1カ月程度の遅れをもって公表されることから，より高頻度データを用いてリアルタイムに予測したいというものである。足下の経済状況をリアルタイムに捕捉したいという動機は，景気分析を担当する内閣府や日本銀行において特に強く，実務的な要請に基づくものである。

スマートフォンには地図アプリなどユーザーの位置情報を取得するものがあり，ユーザーの同意を得てサービスの改善や統計解析に利用されている。位置情報のログデータは（緯度，経度，時刻）のタプルからなり，ユーザーの行動が軌跡として得られる。こうした移動データは別のデータと照らし合わせたり計算を行うことで解釈が可能になる（Renso, Spaccapietra and Zimányi 2013）。もっとも単純なのは店舗や公園といった地点情報データを重ね合わせることで移動データを来訪データに変換することである。Moriwaki（2020）は，わが国の失業者が失業給付の受給のために必ず公共職業安定所（ハローワーク）に出向く必要があるという制度に着目し，移動データからハローワークへの来訪をカウントし，機械学習によって失業者数や失業

図１　位置情報のイメージ



注：筆者作成。背景は OpenStreetMap を用いている。

率の予測モデルを構築している。図１は人工的に作成した位置情報のイメージである。黒い点が位置情報，円が覆っているのがハローワークである。円に入った点の数を数えることで来訪者数をカウントできる。

失業率は失業者数と労働人口によって計算できるので，まず失業者数予測モデルを考える。位置情報は随時データベースに追加されるため，１時間ごと，日次など任意の粒度で集計できる。また，全国には 500 以上のハローワークがある。ここでは日次・ハローワーク分の粒度の細かいデータセットを用意した。全国レベルで集計された『労働力調査』の失業率や失業者数は月次で公表されるため，予測に使う説明変数である来訪者数データと予測対象である失業者数データと頻度が異なる。日次データと月次データのように頻度が異なるデータを Mixed-Frequency（混合頻度）と表現する。オルタナティブデータの活用においては混合頻度をどう考慮するかがポイントとなる。もっとも単純な解決策は高頻度データを集計して予測対象データと同じ頻度にしてしまうというものである。日次と月次の関係であれば，１カ月間の数字をすべて足しあげたり平均したりすれば月次データになる。しかし，

表１　アルゴリズム

準　備：	日次ハローワーク別の来訪者数データ，月次の失業者数データ
手順１：	日次データが月末まで足りない分をARIMAモデルで欠損値補完し，集計して月次データにする
手順２：	各ハローワークのうち過去の失業者数と相関係数が0.3以下の来訪データを除外する
手順３：	残ったデータを用いて失業者数を予測するランダムフォレストモデルを訓練する

この手法は高頻度データが貯まり切る月末まで予測ができないという問題がある。Moriwaki（2020）ではデータが貯まっている分だけ活用して月中でも予測するため，時系列予測モデルを用いて月末までの系列を予測し集計して月次データに変換している。さらに，全国のハローワークのデータの中には全国レベルの失業者数と相関が乏しいものもあるため，過去データを用いてそうしたデータを除外する。最終的に残ったデータを用いてランダムフォレストによる予測モデルを訓練している。全体のアルゴリズムは表１に示す。

こうして組み上がった予測モデルは随時追加される位置情報データを用いて任意のタイミングで予測を行うことができる。図３は『労働力調査』と位置情報の関係を示している。データ収集のリアルタイム性から位置情報ベースの予測は『労働力調査』の１～２カ月先の予測ができる。

機械学習においては深層学習の性能が飛躍的に向上しているが，経済データのようなデータポイントの少ないタスクにおいては，経験上 LASSO やランダムフォレストなどのシンプルな手法が安定した性能を出す。今回は，LASSO とランダムフォレストを比較し性能の高さから後者を採用している。

得られたモデルの予測性能を確認するため，１カ月先予測，２カ月先予測で評価した。N カ月先予測では，モデルのパラメータを決めるための学習データと性能を評価するテストデータを分ける際に，学習データの

図３　位置情報の速報性

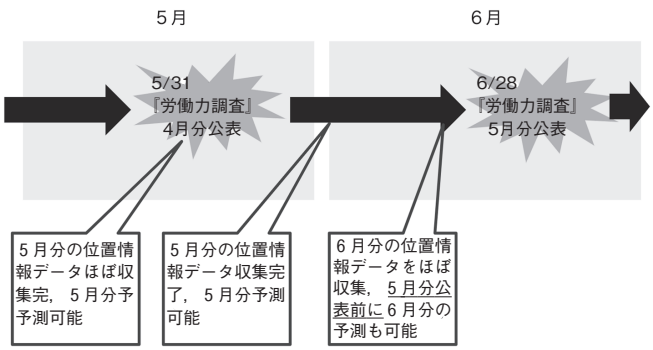


表２　予測性能の評価

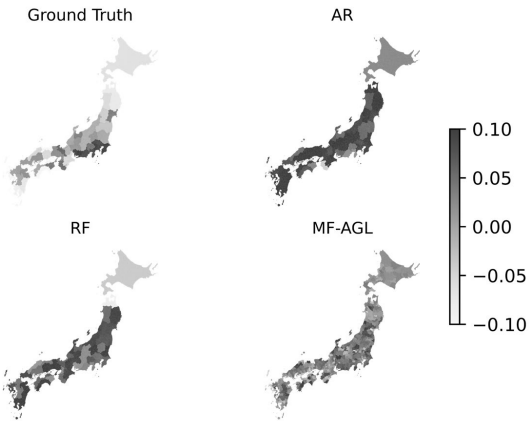
	平方二乗誤差	絶対誤差	予測時点
位置情報	0.083	0.067	31 日前
位置情報（欠損値補完あり）	0.084	0.067	34 日前
ARIMA	0.102	0.086	34 日前

N カ月先のデータをテストデータにする。これによって，モデルが未来の情報を学習してしまう情報の漏洩を許さない評価ができる。

表２に，位置情報モデルの性能を示す。指標として予測二乗誤差と予測絶対誤差を用いる。比較のため ARIMA モデルの性能も掲載する。公平を期すために ARIMA モデルは R の auto.arima によって最適な次数を決定した。位置情報を用いた予測は欠損値補完がある場合でも ARIMA モデルを上回る。位置情報モデルは，ARIMA モデルと違いラグ変数を一切考慮していないので位置情報のみの予測力を示している。この結果は，位置情報に失業者数を予測するための情報が十分含まれていることを示している。

Moriwaki（2020）を拡張した Toda, Moriwaki and Ota（2021）では，同様の位置情報を用いてより時空間的に粒度の細かい予測モデルを提案している。この論文は，ハローワークの来訪者数と新規求職者数が相関しているということを起点に，都道府県別しか公表されていないデータをより粒度の細かい来訪者数データで分割するリカレントニューラルネットワークのモデルを提案している。図４にさまざまなモデルによる新規求職申込者数の前年比予測を示す。詳細は省略するが，提案モデル（MF-AGL）はもともとの職業安定統計（Ground Truth）や時系列モデル（AR），ランダムフォレスト（RF）と違い，細かい地域の状態を予測していることがわかる。

図４　新規求職者数の前年比



３　オルタナティブデータを用いる上での注意点

これまで，Moriwaki（2020）を例にオルタナティブデータを用いた研究について紹介したが，より一般的な注意点について解説する。大前提として，オルタナティブデータ活用においてはプライバシーの問題を常に意識する必要がある。法律やルールを遵守し，データ提供者に不利益にならないよう細心の注意を払うべきである。

オルタナティブデータ活用は，分析者が知りたい経済事象と世の中に溢れるデータとの関係性を検討するところからはじまる。この過程は，アイデア勝負であり日々さまざまなデータに触れることが重要である。労働市場分析にあまり使われていない衛星画像データや決済データなども考えようによっては使える可能性がある。データ活用においてはデータマイニング分野に一日の長があるため，経済経営系のジャーナルだけでなく KDD や WSDM，Web Conference，EC といった国際学会の動向も注視する必要がある。

分析したい経済指標と活用したいデータ，それを関連づけるロジックが定まったら，本当にそのロジックが妥当なのか検証する必要がある。Moriwaki（2020）のような予測モデルの構築の場合は，学習データとテストデータを切り分けて外的妥当性を見ていく必要がある。

オルタナティブデータでは，公的統計における調査と比較してデータクリーニングに時間をかける必要がある。例えば，表記揺れの修正や，データ元のサービスにおける新規機能の追加によるデータ仕様の変更への対応，データの欠損値補完，一部の特殊なユーザーの除外などである。データパイプラインの大元である実際のロギングの実装やデータベースの生データについてしっかり理解することが必要であり，最低限のコーディング知識や SQL 言語を習得し，エンジニアやデータサイエンティストと協業することが大事である。

組み上がったモデルは論文として発表するだけではなく，オルタナティブデータ由来の指標として活用することが重要である。一つの例として予測指標をダッシュボードとしてインターネット上で公開することが考えられる。東京財団政策研究所による GDP ナウキャスティング（https://www.tkfd.or.jp/research/detail.php?id=3852）は好例である。また，指標自体を公開し活用を促すことも大事である。Google 検索データは Google トレンドというかたちで無料で公表されていることが今日までの膨大な研究の蓄積を生み出し

た。METI POS小売販売額指標はダッシュボード機能も含めて先進的な取り組みである（https://www.meti.go.jp/statistics/bigdata-statistics/bigdata_pj_2019/pos_gfk_intage_renewal.html）。また，組み上がったモデルをオープンソースソフトウェア（OSS）として公開することも重要である。公開することで追試や研究の発展を促し透明性を担保する。

機械学習の分野ではコンセプトドリフトという概念がある。これは予測モデルが陳腐化し，開発時の予測性能が維持できなくなることをいう。景気変動やテクノロジーの変化といった経済環境の変化に加えてオルタナティブデータ側の変化にも注意する必要がある。位置情報を例にすると，アップルやGoogleの開発するスマートフォンOSの仕様変更により，位置情報の取得に対するユーザーの許諾がとりづらくなったり，逆にスマホや人工衛星の機能向上によって位置情報の精度が高まったりといった変化が起きている。

さらに，データが継続的に供給されるかという点にも注意する必要がある。例えば，新型コロナの発生によってプラットフォーマーが人流データを無料で提供しているが，これがいつまで続くかはこれらの企業の意思決定にかかっている（https://covid19.apple.com/mobility，https://www.google.com/covid19/mobility/）。企業データを継続的に供給させるようなインセンティブの仕組みも検討する必要がある。

データが一部企業が提供するアプリケーションやサービスに依存している場合，その企業のシェアの増減によってユーザーの属性が変化することにも留意が必要である。

ただし，こうした不確実性は公的統計にも存在する。『毎月勤労統計』や『建設工事受注動態統計』で噴出した統計の品質問題を引き合いに出すまでもなく，すべての指標，統計は継続的にその品質を維持する努力を払う必要がある。行政記録情報も含めあらゆるデータを正確な意思決定につなげるための不断の取組が期待される。

参考文献

Askitas, N. and Zimmermann, K. F.（2009）“Google Econometrics and Unemployment Forecasting,” Applied Economics Quarterly, Vol. 55, No. 2, pp. 107–120.

Bonato, P., Cintia, P., Fabbri, F., Fadda, D., Giannotti, F., Lopalco, P. L., Mazzilli, S., Nanni, M., Pappalardo, L., Pedreschi, D., Penone, F., Rinzivillo, S., Rossetti, G., Savarese, M. and Tavoschi, L.（2020）“Mobile Phone Data Analytics Against the COVID-19 Epidemics in Italy: Flow Diversity and Local Job Markets during the National Lockdown,” arXiv: 2004. 11278 [cs, stat].

Caperna, G., Colagrossi, M., Geraci, A. and Mazzarella, G.（2022）A Babel of Web-searches: Googling Unemployment during the Pandemic,” *Labour Economics*, Vol. 74, 102097.

Choi, H. and Varian, H.（2012）“Predicting the Present with Google Trends,” *Economic Record*, Vol. 88, Issue. S1.

D’Amuri, F. and Marcucci, J.（2010）“‘Google It!’ Forecasting the US Unemployment Rate with A Google Job Search Index,” FEEM Working Paper, No. 31.

Forsythe, E., Kahn, L. B., Lange, F. and Wiczer, D. G.（2020）“Labor Demand in the Time of COVID-19: Evidence from Vacancy Postings and UI Claims,” *Journal of Public Economics*, Vol. 189, 104238.

Fukui, M., Kikuchi, S. and Goalist, Co.（2020）“Job Creation during the COVID-19 Pandemic in Japan,” CREPE Discussion Papers, No. 73.

Ghoshsamaddar, S., Marchetti, A. and Sevcenko, V.（2021）“Who Captures the Value from Organizational Culture? Evidence from Glassdoor Reviews and the Universe of Online Job Postings from Burning Glass Technologies,” INSEAD Working Paper No. 2021/58/STR.

Huang, N., Burtch, G., Hong, Y. and Pavlou, P. A.（2020）“Unemployment and Worker Participation in the Gig Economy: Evidence from an Online Labor Market,” *Information Systems Research*, Vol. 31, No. 2, pp. 431–448.

Moriwaki, D.（2020）“Nowcasting Unemployment Rates with Smartphone GPS Data,” in K. Tserpes, C. Renso and S. Matwin（eds.）*Multiple-Aspect Analysis of Semantic Trajectories*（pp. 21–33）, Springer International Publishing.

OECD（2021）“Speaking the Same Language: A Machine Learning Approach to Classify Skills in Burning Glass Technologies Data” OECD Social, Employment and Migration Working Papers No. 263.

Ozimek, A.（2020）“When Work Goes Remote,” SSRN Scholarly Paper 3777324.

Renso, C., Spaccapietra, S. and Zimányi, E.（eds.）（2013）*Mobility Data: Modeling, Management, and Understanding*, Cambridge University Press.

Shi, B., Yang, J., Guo, F. and He, Q.（2020）“Salience and Market-aware Skill Extraction for Job Targeting,” Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2871–2879.

Suhoy, T.（2009）“Query Indices and a 2008 Downturn: Israeli Data,” Bank of Israel Discussion Papers.

Toda, T., Moriwaki, D. and Ota, K.（2021）“Aggregate Learning for Mixed Frequency Data,” arXiv preprint arXiv: 2105. 09579.

Zhang, Q., Zhu, H., Sun, Y., Liu, H., Zhuang, F. and Xiong, H.（2021）“Talent Demand Forecasting with Attentive Neural Sequential Model,” Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 3906–3916.

もりわき・だいすけ　株式会社サイバーエージェントAI Lab研究員。最近の主な論文に“A Real World Implementation of Unbiased Lift-based Bidding,” IEEE Big Data 2021（早川裕太・松井暉・宗政一舟・齋藤優太・芝田将との共著）。計量経済学専攻。