ビッグデータ連携会議 レポート 02

2022年2月

人材紹介サービスのトランザクションデータを用いた労働統計の速報指標開発

高田 悠矢 *

* 株式会社リクルート 特任研究員

人材紹介サービスのトランザクションデータを用いた労働統計の速報指標開発

2022 年 2 月

要旨

近年、政府が取り組む抜本的な統計改革においては、これまでの統計改革とは明確に異なる点がある。それは、統計作成のために企業や家計に調査票の記入を依頼し、それらを回収・集計するという既存の政府統計作成の枠組みがもたらすボトルネックに、民間企業が保有するトランザクションデータを活用する事でメスを入れようとしている点である。民間企業が保有するトランザクションデータが経済指標として活用されるケースは、① トランザクションデータの集計値そのものを代替指標として活用するケース、② 公的統計のナウキャスト指標を作成するケース、③ 既存統計の推計方法を一部変更してトランザクションデータを活用した推計を取り入れるケースの三つに分類する事ができる。我が国における労働市場のトランザクションデータの活用事例は、求人広告事業や人材紹介事業を行う企業・業界団体が独自に公表している指標が上述①の文脈で活用されるケース以外は、筆者が知る限り存在しない。本稿では、労働市場に関する指標において初となる②や③を視野に入れた検証を行った。具体的には、厚生労働省「雇用動向調査」における「転職時の賃金変動状況：転職時に賃金が明確に（1 割以上）増加した転職者の割合」について、リクルート社の保有する人材紹介サービスのトランザクションデータを活用し、確率密度比推定や共変量シフト下での教師付き学習の考え方を応用する事で、速報指標を作成できる可能性を示した。現状は、1 月～6 月分が 12 月末（令和 2 年の場合は翌年 2 月初）公表、7 月～12 月分を加えた暦年分が 8 月下旬（令和元年分は翌年 9 月末）公表というかたちでタイムラグが存在するが、1%強の速報・確報間乖離を許容する前提のもとでは、概ねリアルタイムでの公表が可能である。

キーワード：雇用動向調査、転職時の賃金変動状況、ナウキャスト、確率密度比推定、uLSIF、共変量シフト下での教師付き学習

本稿の作成にあたっては、中村英昭氏、前原庸司氏、竹内貴史氏、松井伸司氏（以上、総務省）、柏村美生氏、佐藤学氏、三好譲二氏、島昌平氏、緒方真樹子氏、林史子氏、宮村収氏、岩元洋介氏、高島優花氏（以上、株式会社リクルート）、和泉潔氏（東京大学）、田原健吾氏（日本経済研究センター）、木幡賢人氏、高瀬洋人氏（以上、TDSE 株式会社）の各氏から有益な助言を頂いた。加えて、雇用動向調査のデータを提供いただいた厚生労働省 政策統括官（統計・情報政策、政策評価担当）付 参事官付 審査解析室をここに記して感謝したい。ただし、本稿の内容と意見は筆者個人に属し、所属組織あるいは総務省の公式見解を示すものではない。また、ありうべき誤りはすべて筆者個人に属する。

目次

1. はじめに
   1.1 背景と目的
   1.2 「転職時の賃金変動状況」という指標の有用性
   1.3 「雇用動向調査」における速報性の課題
2. データについて
   2.1 リクルート社のデータのカバレッジ
   2.2 推計に利用したデータセット
3. 手法
   3.1 推計手順の全体像
   3.2 転職者人数と属性情報の推計
   3.3 リクルート社のデータに対するウエイト付け
   3.4 ラベル情報の補正
   3.5 密度比をウエイトとした分類モデルによる追加処理
   3.6 密度比をウエイトとした回帰モデルによる追加処理
4. 推計結果
   4.1 転職時に賃金が明確に（1割以上）増加した転職者の割合の推計結果
   4.2 転職者人数と属性情報の推計の結果について
   4.3 リクルート社のデータに対するウエイト付けの結果
   4.4 密度比をウエイトとした分類・回帰モデルによる追加処理の結果
5. 結びにかえて

（別紙）
（参考文献）

1. はじめに

1.1 背景と目的

近年、政府が取り組む抜本的な統計改革においては、これまでの統計改革とは明確に異なる点がある。それは、統計作成のために企業や家計に調査票の記入を依頼し、それらを回収・集計するという既存の政府統計作成の枠組みがもたらすボトルネックに、民間企業が保有するビッグデータを活用する事でメスを入れようとしている点である。

民間企業が保有するビッグデータが経済指標として活用されるケースは三つのタイプに分類する事ができる。一つ目は、① トランザクションデータの集計値そのものを代替指標として活用するケースである。従来の経済統計を補完する位置付けであり、オルタナティブデータと呼ばれる事も多い。例えば、経済産業省では「令和元年度 ビッグデータを活用した新指標開発事業（短期の生産・販売動向把握)」では、POS（Point of Sales）データを用いた小売販売に関わる統計について、試験的な公表を始めている。二つ目としては、② 公的統計のナウキャスト指標を作成するケースである。上述の経済産業省の取組みでは、POS データを用いて消費者物価指数をナウキャストしたケースや、SNS におけるつぶやき情報を用いて鉱工業生産指数をナウキャストしたケースが存在する。三つ目としては、③ 既存統計の推計方法を一部変更してトランザクションデータを活用した推計を取り入れるケースである。こちらは、POS データやウェブスクレイピングにより取得したデータを物価統計の算出に活用する取組みがあるほか、例えば、エストニアでは、政府の予算制約の都合で、従来活用していた国境に係る調査が中止になった事が取組の契機となり、携帯電話の SIM カード情報を活用する事で旅行客の消費を捕捉し、旅行収支に関する統計を作成するといったケースが存在する[1)]。

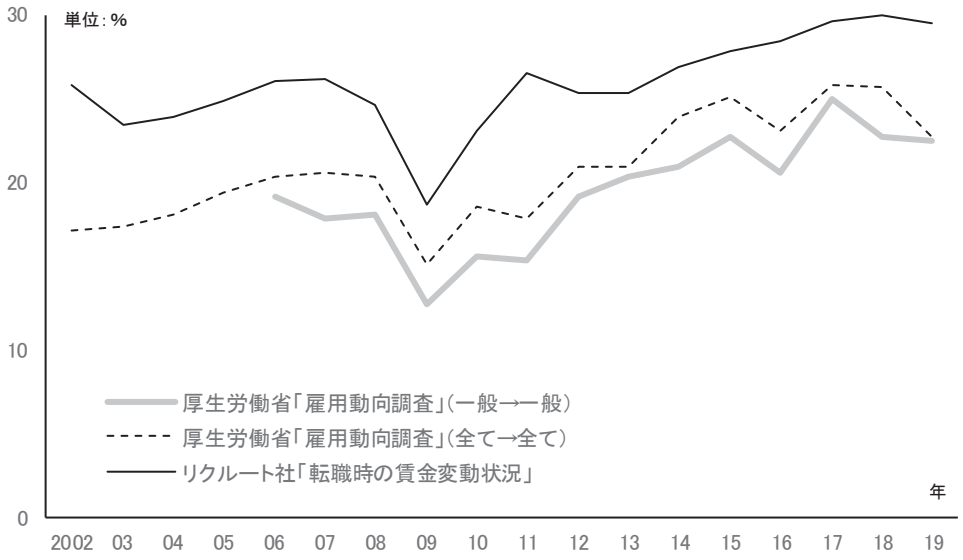
我が国における労働統計へのトランザクションデータの活用事例としては、求人広告事業や人材紹介事業を行う企業・業界団体が独自に公表している指標が、上述①の文脈で活用されるケースは存在するものの、筆者が知る限り②や③の事例は存在しない。本稿は、労働統計に関する②の初の事例という位置づけである。また、前述の SNS でのつぶやき情報で鉱工業生産指数をナウキャストするようなケースにおいては、両者の関係はあくまで相関の関係であり、SNS でのつぶやきは、鉱工業生産“そのもの”の情報ではないという意味で、捕捉対象は異なるものである一方、今回のケースでは「リクルート社の人材紹介サービスを通じて転職を行った転職者」の情報を用いて「我が国全体の転職者」の振る舞いをナウキャストする構造にあるため、カバレッジは異なるが、捕捉対象は全く同じ転職者情報である。これは、本取組みが、②のみでなく、③に対してもインプリケーションを持つという事を意味する。

今回推計対象とした指標は厚生労働省「雇用動向調査」における「転職時の賃金変動状況：転職時に賃金が明確に（1 割以上）増加した転職者の割合」（図表 1）である。単位は％であり、分子が「転職時に賃金が明確に（1 割以上）増加した転職者数」、分母が「総転職者数」である。詳細は後述するが、この指標は外部労働市場の需給が逼迫した際に、その過熱感が賃金に反映されているか否かを示す非常に重要な指標と考えられる。

---

1) Kroon and Pank(2012)、Hammer et al.(2017)を参照。

図表 1 転職時の賃金変動状況[2)]
転職時に賃金が明確に（1 割以上）増加した転職者の割合



出所：厚生労働省「雇用動向調査」、リクルート社「転職時の賃金変動状況」より筆者作成

雇用動向調査は非常に有用な統計調査である一方、速報性の面においては課題がある。本稿では、この速報性の面での課題を、リクルート社のトランザクションデータを用いて解決するための枠組みを提示する。ここで重要なのが、リクルート社のデータは、当然ながらリクルート社の人材紹介サービス経由の転職者に捕捉対象が限られるという点である。すなわち、今回の取組みは「公表までに時間を要するが、我が国全体をカバーする事ができる雇用動向調査」と「リアルタイムに取得できるが、リクルート社が運営する人材紹介サービスを通じて転職をした者のみにカバレッジが限定されるトランザクションデータ」とを組み合わせる事で、新たな価値を創出しようとする試みである。より単純な表現を用いれば「遅い全体」を「速い部分」で補完しようという取組みである。

以下、1.2 節では、転職時の賃金変動状況の指標としての有用性について述べ、1.3 節では厚生労働省「雇用動向調査」における速報性の課題について述べる。続く、2 章ではデータについて触れる。2.1 節では、リクルート社のデータのカバレッジについて、2.2 節では、今回の推計に利用したデータセットの詳細について言及する。その後、3 章では具体的な推計手法を説明し、4 章にて推計結果の考察を行う。

2) 後述の通り、今回は「一般労働者から一般労働者への転職」を対象としている（図中太実線）。「一般労働者から一般労働者への転職」の集計値は、2005 年以前は公表されていないため、図表 1 では表示していないが、今回の推計においては、2002 年分より当該属性情報を取得・利用している。

1.2「転職時の賃金変動状況」という指標の有用性

齊藤ほか（2010）は、我が国の研究者により執筆された代表的なマクロ経済学の教科書であるが、ここでは、労働市場に関する経済指標として、総務省「労働力調査」、厚生労働省「毎月勤労統計」、同「一般職業紹介状況」の三つが紹介されている。労働力調査では、就業状態別の人数を把握する事ができる。すなわち、就業者数、完全失業者数といった労働力人口に加え、非労働力人口の推移を確認する事が可能である。労働力人口に占める完全失業者の割合を示す失業率も、労働力調査に含まれる。

毎月勤労統計は、賃金動向や労働時間を把握する際に利用される事が多い。就業者数と平均賃金の積の推移は、分配側 GDP の約半分を占める雇用者報酬の推移と概ね連動する。また、就業者数と労働時間の積により総労働量を把握する事ができる。一般職業紹介状況では、求人倍率、すなわち、求職者一人当たりの求人数を知る事ができるが、この指標が十分に高い値となれば、今後、就業者数が増えていくか、あるいは賃金が上昇する可能性が高いと考える事ができる。このように、これら三つの指標を観察する事で、労働市場動向の大枠を掴む事が可能になる。

一方、これらの三指標のみでは把握できない重要な指標もある。その一つが、今回の推計対象である「転職時の賃金変動状況：転職時に賃金が明確に（1 割以上）増加した転職者の割合」という指標である。例として、求人倍率が年々高まり、売手市場化が進むような局面を想定しよう。企業は人材を獲得するため、求職者に自社の魅力を強くアピールする必要がある。その際、高い賃金水準の提示は有効な手段の一つである。すなわち、労働需給が逼迫は、賃金への圧力として機能する可能性があると言える。しかし、求人倍率が高まれば、賃金への圧力が“必ず”高まるとは限らない。企業が「賃金を上げてでも人材を確保したい」という状況、あるいは、募集ポストに紐づく賃金水準は不変でも「採用要件を下げてでも人材を確保したい」という状況となれば、需給の逼迫が（少なくとも間接的には）賃金への圧力として機能していると言える。一方で、これ以上、賃金を上げる事は難しく、採用要件も下げる事ができないが、そうした方法以外で自社の魅力を強くアピールする事で、どうにか人材を確保したいというような、需給の逼迫が賃金への圧力として機能しない状況も考えられる。転職前後の賃金変動という情報は、これらの違い、すなわち、労働需給の逼迫が、賃金への圧力として機能しているのか否かを見分ける事を可能にする。

また、求職者側の視点で見た場合、求人倍率は、求職者数を分母、求人数を分子とした指標であるため、単純に考えれば、この値が大きければ大きい程、望ましいという事になる。しかし、仮に、求人倍率が「1」を十分に超えている局面において、既存の求人と賃金以外の条件が全て同じで、既存の求人よりも賃金水準が著しく低い求人が大量に新規掲載された場合、それらの求人は求人倍率の値を上げる事には寄与するが、求職者視点で「労働市場がより魅力的になった」とは言い難いであろう。こうした局面において、転職時の賃金変動状況という指標は、賃金水準が著しく低い求人が大量に新規出稿された前後で、概ね同じ値を示すと考えられる。逆に、企業が「賃金を上げてでも人材を確保したい」という状況、あるいは、募集ポストに紐づく賃金水準は不変でも「採用要件を下げてでも人材を確保したい」という状況となれば当該指標の値は上昇すると考えられる。つまり、求人倍率に着目するような場面においては、転職時の賃金変動状況という指標を併せて観察する事で、労働市場の変化を、より明確に捉える事が可能となる。

こうした情報は、政策当局のみでなく、企業経営者・採用担当者にとっても有益であると考えられる。以下では、企業経営や採用に携わる方々を対象として行ったアンケート調査を紹介したい。この調査では、足許の採用活動の状況と、賃金制度変更の必要性に関する以下の設問（図表 2～4）に対して、「会社・事業の方向性について、最終決裁をする立場」の方々1,000 人と、「人事や採用について最終決裁をする立場」の方々802 人の合計 1,802 人から回答を得た。

図表 2 共通の設問

Q1.この1～2年の中途採用の活動状況をお知らせください。

全体として採用が難しくなっている
一部の職種等について、採用が難しくなっている
概ね採用できている
全く問題なく採用できている
採用活動は行っていない

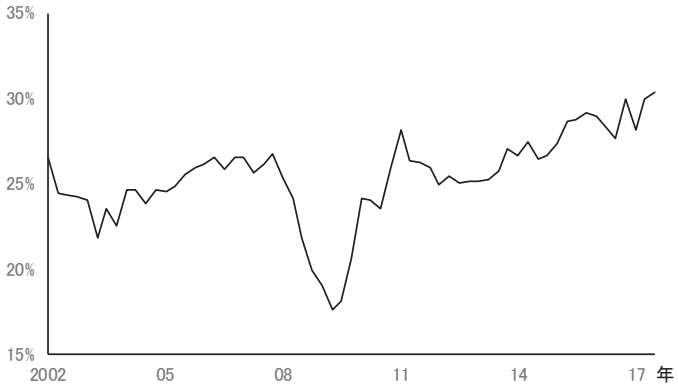
Q2.この1～2年で賃金体系の変更について何らかの検討を行いましたか。
（賃金体系の異なる子会社や事業部の設立を含む）

賃金体系の変更について検討を行う必要はない
賃金体系の変更について検討を行う必要は殆どない
賃金体系の変更について、今後検討を行う可能性がある
賃金体系の変更について検討を行ったことがある

出所：リクルート社調査

図表 3 閲覧記事

転職時に1割以上賃金が増加した転職者の割合



採用競争が過熱するなか、転職によって収入が増加するケースが増加。
2013年頃から年々高まり、直近ではリーマン危機前の水準を大きく超えている。

出所：リクルート社調査

図表 4 記事閲覧後の追加設問

Q3.この記事を見て、賃金体系の変更についての必要性を感じましたか。

賃金体系の変更について検討を行う必要性は感じない
賃金体系の変更について検討を行う必要性は殆ど感じない
→ 必要なし

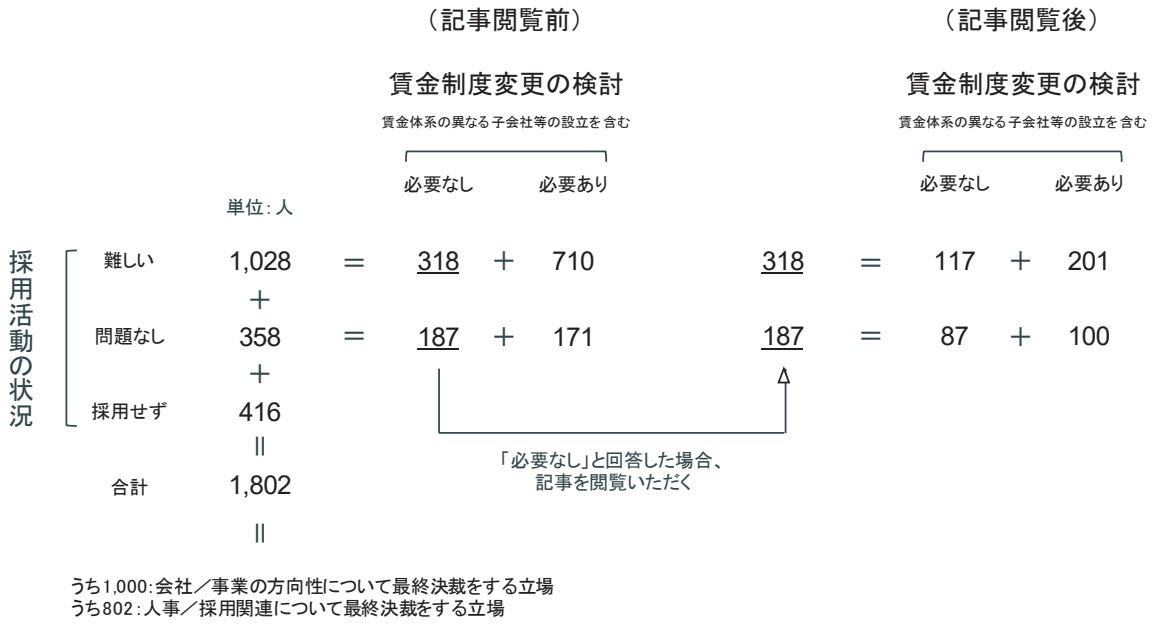
賃金体系の変更について検討を行う必要性を少し感じた
賃金体系の変更について検討を行う必要性を感じた
→ 必要あり

出所：リクルート社調査

当該アンケート調査では、一番目の設問（図表 2 の Q1）にて、ここ 1～2 年の中途採用の状況を、二番目の設問（図表 2 の Q2）にて、賃金制度の変更の検討有無を確認した。また、賃金制度の変更について「必要ない」「必要は殆どない」と回答した場合は、図表 3 に示した記事の閲覧と、追加の設問（図表 4）を用意した。なお、記事の内容は、リクルート社のデータによる「転職時の賃金変動」の時系列推移と、その簡易的な説明である。

これらの設問の回答結果を以下の図表 5 に記した。回答が得られた 1,802 名のうち、ここ 1～2 年の中途採用の状況について「採用が難しくなっている」と回答したのは 1028 名であった。このうち 318 名は「賃金体系の変更について何らかの検討を行った／行う予定がありますか？」という問いに対し、「必要ない／必要は殆どない」と答えている。そして、この「必要ない／必要は殆どない」と回答した 318 名に、前述の記事を閲覧していただいた上で「この記事をみて、賃金体系の変更についての必要性を感じましたか？」という追加設問を回答いただいたところ、201 名が「必要性を感じた／少し感じた」と回答した。すなわち、約 3 分の 2 の回答者が、記事の閲覧をきっかけとして賃金制度の変更に対するスタンスを変えた事になり、「転職時の賃金変動」という指標が企業の意思決定に対して与える影響は、決して小さくないと考えられる。

図表 5 回答結果

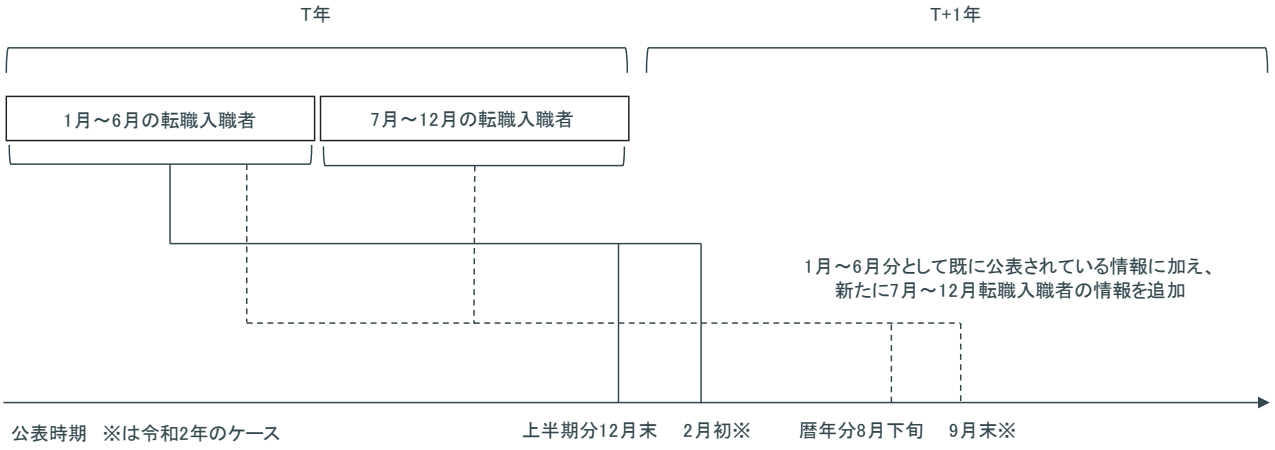


出所：リクルート社調査より筆者作成

1.3「雇用動向調査」における速報性の課題

1.2 節では、雇用動向調査「転職時の賃金変動状況」が、「労働力調査」、「毎月勤労統計」、「一般職業紹介状況」では捕捉する事のできない非常に有用な情報を持つ指標であるという事を述べた。しかし、足許の労働市場の状況を踏まえてリアルタイムに意思決定を行うような場面では、現状の雇用動向調査「転職時の賃金変動状況」は、殆ど役に立たないのが実態である。労働力調査、毎月勤労統計、一般職業紹介状況は、ともに、月次で公表される統計であり、タイムラグは概ね 2 か月程度である[3)]ため、政策判断や経営判断といった足許の労働市場の状況を踏まえたリアルタイムな意思決定に利用する事が可能である。一方、雇用動向調査は半期毎に公表される統計であり、年に二度の統計調査が実施されるが、1 月～6 月分が 12 月末（令和 2 年の場合は幾分遅れ翌年 2 月初）公表、7 月～12 月分を加えた暦年分が翌年 8 月下旬（同・9 月末）公表というかたちで、6 か月～14 か月強の大きなタイムラグが存在しており、足許の状況を踏まえた意思決定に利用する事は難しい（図表 6）。

図表 6 雇用動向調査の公表タイミングについて



出所：厚生労働省ホームページの情報をもとに筆者作成

3) 例えば、2021 年 1 月分は、毎月勤労統計速報の場合は同年 3 月 9 日、労働力調査基本集計、一般職業紹介状況はともに同年 3 月 2 日に公表されている。

2. データについて

2.1 リクルート社のデータのカバレッジ

1.3 節では雇用動向調査における速報性の課題を述べた。本取組みは、この速報性の面での課題を、リクルート社のトランザクションデータによって解決しようとするものであるが、言うまでもなく、リクルート社のデータは万能ではない。当然ながらリクルート社の人材紹介サービス経由の転職者に捕捉対象が限られるというカバレッジの問題がある。民間人材紹介サービスを通じて転職をする者は我が国全体の転職者の 6%程度に過ぎない（図表 7：A を参照）。ビジネスモデルの特性上、ある程度、求職者特性が類似していると考えられる求人広告経由の割合を加えても、4 割に満たない（図表 7：A+B を参照）。また、ハローワークと民間人材紹介では、扱う求人の特性や利用する求職者属性が異なると考えられる。我が国の労働市場全体の動向を掴むという目的で利用するには、そうしたバイアスが存在する点を認識する必要がある。

図表 7 経路別の転職者の割合

	職業安定所	16.7%
	ハローワークインターネットサービス	3.4%
A：	民営職業紹介所	6.0%
	学校	1.3%
B：	広告	30.4%
	その他	11.6%
	縁故	26.9%
	うち前の会社	6.6%
	出向	2.4%
	出向先からの復帰	1.2%
A+B：		36.4%

出所：厚生労働省「雇用動向調査」（2019 年）より筆者作成

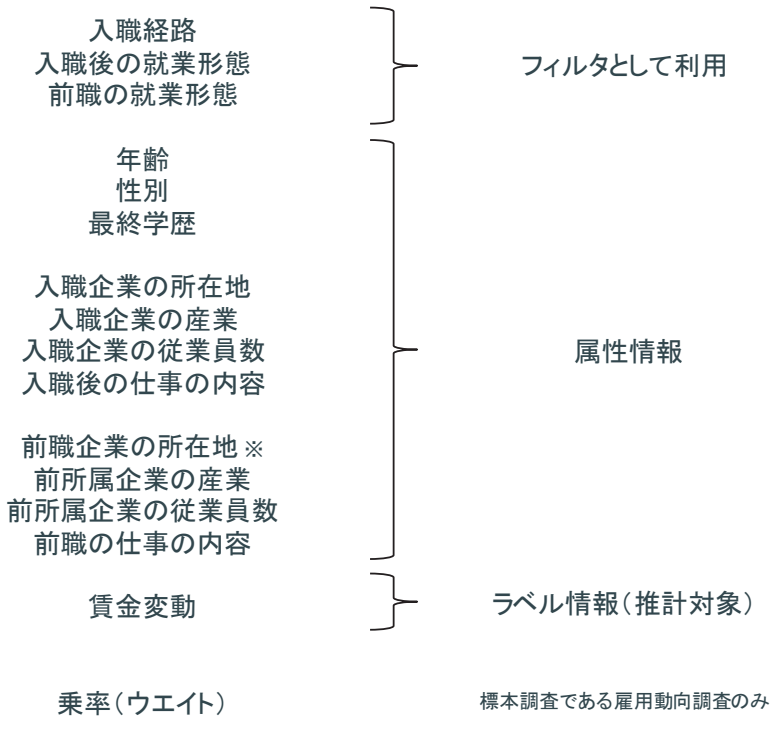
2.2 推計に利用したデータセット

この節では本取組みに用いるデータセットについて述べる。厚生労働省「雇用動向調査」では、転職者の情報は調査客体企業より報告される。標本調査であり、例えば2020年上半期調査における標本数は36,412人、ここから推計された我が国全体の転職者数は4,360,700人であった。今回はこれらの標本と乗率（ウエイトバックを行う際に乗じる値）について、2004年分より取得した。なお、本取組では、このうち一般労働者から一般労働者への転職を対象としている。

具体的に推計に利用した項目は「年齢」、「性別」、「最終学歴」、入職前後における「企業の所在地」、「産業」、「従業員数」、そして「仕事内容」といった属性情報と、ラベル情報となる「転職時の賃金変動：転職時に明確に（1割以上）賃金が増加したか否か」である。このほか、「入職経路」を3.2節で述べる推計の際にフィルタとして利用した。また、入職前後の「就業形態」については、一般労働者から一般労働者への転職に利用レコードを絞り込む際に利用した（図表8）。

なお、リクルート社のトランザクションデータについては、雇用動向調査の項目と同等と見做せる情報を集めて利用した。産業分類や仕事内容などの枠組みは当然ながら異なるものの、対応表をつくる事でリクルート社の情報を、雇用動向調査側に寄せている。また、3.6節にて詳細を後述するが、ラベル情報となる転職時の賃金変動については、リクルート社側のデータのみ、連続値（転職前後のそれぞれで単位「円」の具体的な金額情報）としても取得・利用している。

図表8 利用項目一覧



※ 入職前1年以内に仕事についた経験がない場合は入職前の居住地を記載。

出所：厚生労働省「雇用動向調査」者表を参考に筆者作成

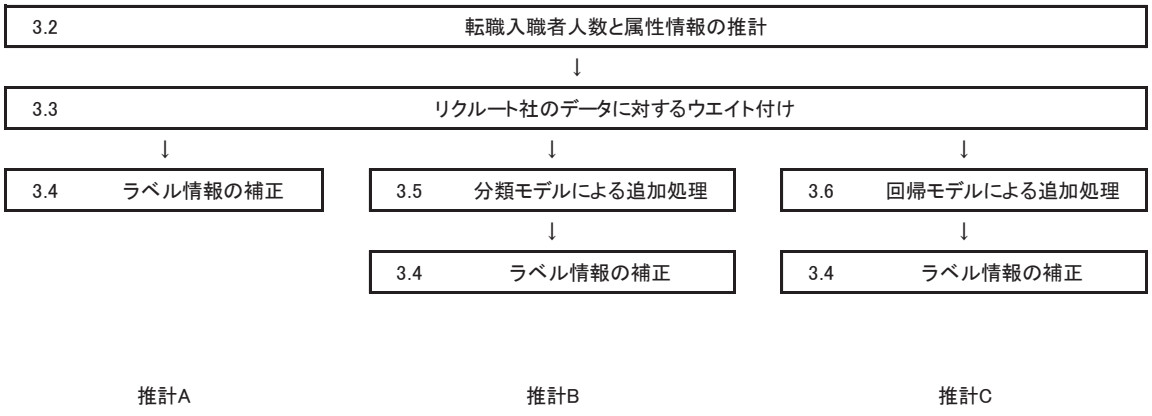
3. 手法

3.1 推計手順の全体像

3 章では推計手法について述べる。まず、3.1 節で大枠の全体像に触れ、3.2 節以降で具体的な手法を説明する。推計の枠組みについては、図表 9 に記した通り、三つの経路によって構成される。まず、共通の手順として、我が国全体のデータにおいて対象期間の「転職者の人数と属性情報の推計」（3.2 節）を行い、次に、その結果を用いて同期間の「リクルート社のデータに対するウエイト付け」（3.3 節）を行う。この段階で、3.4 節で述べる補正処理を行い、ウエイト付けされたリクルート社のデータを集計する事で、今回の目的である「転職時に賃金が増加した転職者の割合」の推計値を得た場合、その経路は図表 9 に示した「推計 A」に相当する。

以降で述べる手順は、前述の共通の手順の後に追加的に加えるものである。具体的には、前述 3.3 節のプロセスで得られた密度比によるウエイト情報を考慮して、リクルート社のデータより「転職者の属性情報」を説明変数、「転職時に賃金が 1 割以上増加したか否か」を目的変数とした分類モデルを構築し（3.5 節）、そのモデルの引数に 3.2 節で得られた属性情報を投入、得られた値を集計する事で、今回の目的である「転職時に賃金が増加した転職者の割合」の推計値を得る事ができる。この経路は、図表 9 では「推計 B」に相当する。目的変数として連続値を利用した場合、すなわち回帰問題として扱った場合（3.6 節）は、図表 9 における「推計 C」に相当する。なお、分類・回帰いずれのケースにおいても、集計の前に、3.4 節で述べる補正を行っている。

図表 9 推計の枠組みについて



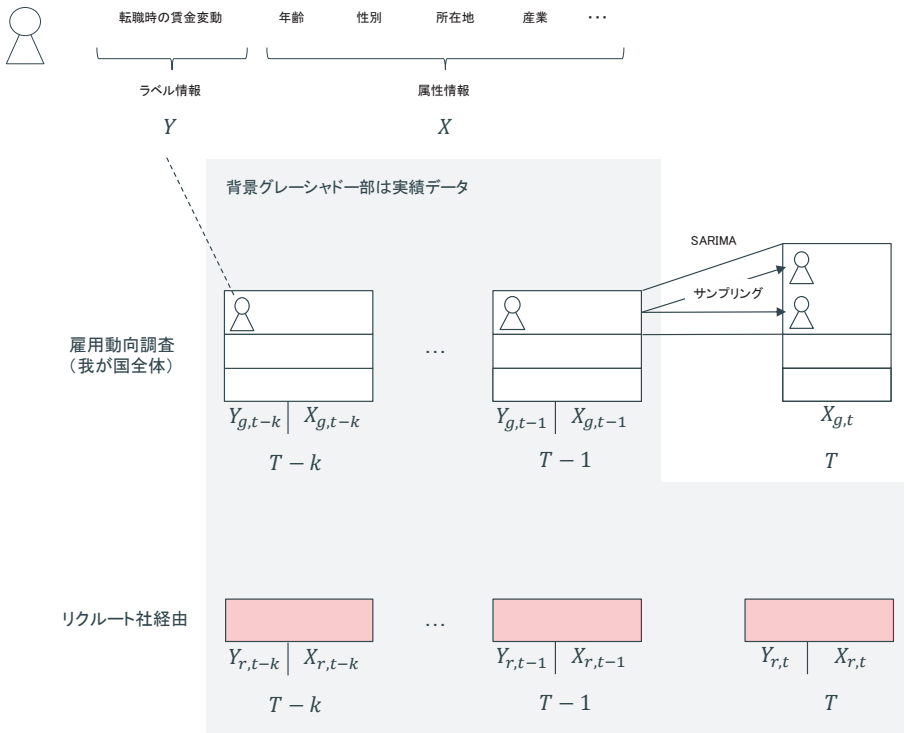
3.2 転職者人数と属性情報の推計

ここでは「推計対象期間の転職者の人数と、それらの属性情報の推計」について述べる。当該推計には、雇用動向調査側のデータのみを用いている。また、推計対象は、あくまで転職者の人数と属性情報の算出のみであり、ラベル情報は算出していない。

雇用動向調査における T 期の転職者毎の転職時の賃金変動情報（ラベル情報）を成分とするベクトルを$Y_{g,t}$、T 期の転職者毎の属性情報を成分とする行列を$X_{g,t}$とし、同様に、リクルート社データおける T 期の転職時の賃金変動情報のベクトルを$Y_{r,t}$、T 期の属性情報行列を$X_{r,t}$とすると、この節で扱う推計過程は、$X_{g,t-k}$~$X_{g,t-1}$（T 期の推計時点で T-1 期までの雇用動向調査情報が手に入る場合）あるいは、$X_{g,t-k}$~$X_{g,t-2}$（T 期の推計時点で T-2 期までの雇用動向調査情報が手に入る場合）の情報のみを用いて$X_{g,t}$を算出する部分である（図表 10）。なお、図表 10～図表 14 では「T 期の推計時点で T-1 期までの雇用動向調査情報が手に入る場合」のみを図示している。今回は「2004 年 1 月～6 月期」以降のデータを利用しており、T を「2018 年 1 月～6 月期」とした場合、T-1 期は「2017 年 7 月～12 月期」となり、時系列の始点（2004 年 1 月～6 月期）は T-28、すなわち、k=28 となる。

リクルート社のデータから得られる情報はこの過程では用いない。すなわち、この 3.2 節で述べる手順は「リクルート社のデータによる推計が難しく、過去からのトレンド情報のみに頼らざるを得ない」部分の推計過程と言える。具体的には、まず、① 転職経路毎の転職者数を SARIMA モデルによって求めた。パラメータは、AIC（赤池情報量基準）が最小となるように設定した。その後、② その直近期の転職者サンプルから、総量が①で求めた転職者数と一致するように、重複を許すかたちでランダムにサンプリングを行った。サンプリングを行う直近期としては、言葉の通りの直近期と季節性を加味した直近期の両方を試している。例えば、学習に利用可能な実績データの末端が T 年 1 月～6 月で、推計対象期間が T 年 7 月～12 月であった場合、言葉の通りの直近期 T 年 1 月～6 月となり、季節性を加味した直近期は T-1 年 7 月～12 月となる。

図表 10 転職者人数と属性情報の推計の概念図

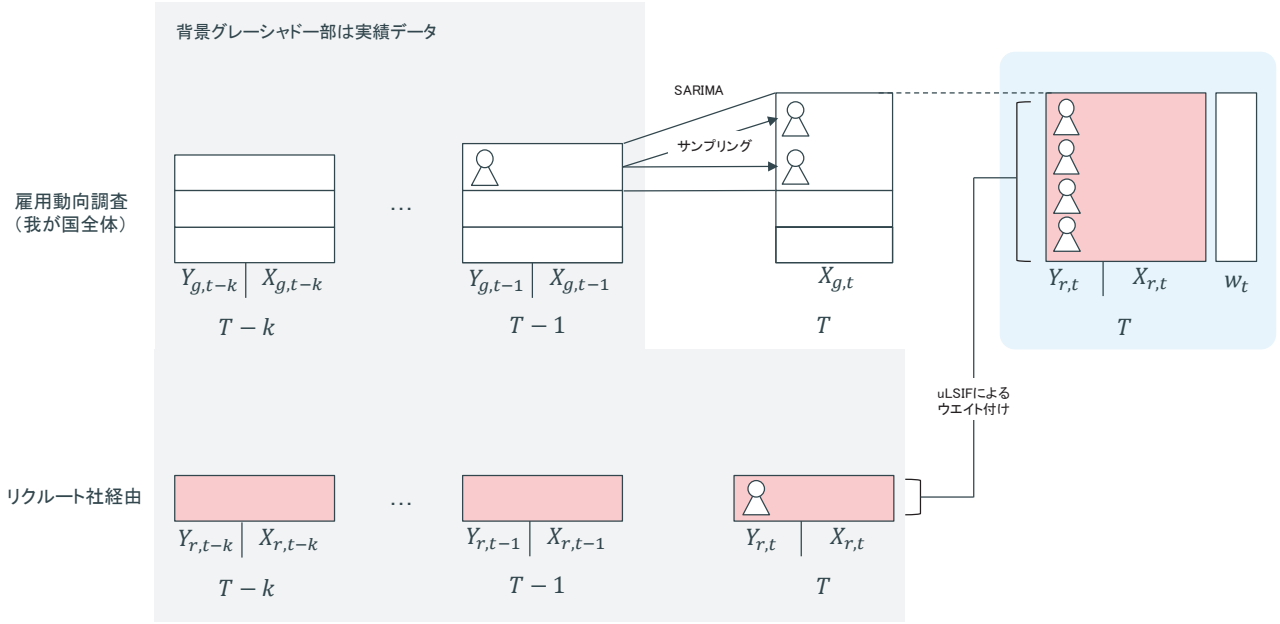


※ 「雇用動向調査（我が国全体）」は、職業安定所、広告、民間職業紹介、縁故といった転職経路毎に SARIMA による推計を行っているため、層を分けた画としている。なお、（図表では 3 層の表示であるが）実際の経路の数は 9 経路である。

3.3 リクルート社のデータに対するウエイト付け

次に、リクルート社のデータにおける属性の分布を、3.2 節で推計した我が国全体における属性分布に近づけるための処理として、リクルート社のデータレコードに対するウエイトの付与を行う（図表 11）。リクルート社のデータは、最新の「属性情報が与えられたもとでのラベル情報（転職時の賃金変動状況）」を保有している一方で、カバレッジがリクルート社経由の転職者に限られるのであった。ウエイトを付与する目的は、この最新の「属性情報が与えられたもとでのラベル情報（転職時の賃金変動状況）」を、我が国全体に“引き伸ばす”ためである。

図表 11 リクルート社のデータに対するウエイト付けの概念図



※ ここでは直観的な理解を優先するため、SARIMA によって算出された我が国全体の転職者数と、uLSIF によるウエイト付与後のリクルート社のデータの転職者数があたかも完全に一致するような画としているが、あくまで各属性の分布を近づけるためにウエイトが付与されているため、実際の総数は必ずしも一致しない。

ウエイト付与にあたっては、今回、確率密度比推定によるウエイト付けの手法を採用した。Sugiyama et al.(2009)で提案されたような、確率密度の「比」を推定する統計的機械学習の枠組みは、Support Vector Machine が、データ生成における確率分布を推定するという問題を避け、決定境界のみを学習するように、困難な問題として知られている確率密度（そのもの）の推定を回避するという考え方に従っている。密度比推定は、データの定義域を$\mathcal{D} \subset \mathbb{R}^d$とした時、確率密度$P_S(x) > 0 \ \ for \ \ all \ \ x \in \mathcal{D}$を持つ確率分布に独立に従う i.i.d.標本$\{x_i\}_{i=1}^{n}$、および、確率密度$P_T(x)$を持つ確率分布に独立に従う i.i.d.標本$\{x_j'\}_{j=1}^{n'}$から、確率密度比$w(x) = P_T(x)/P_S(x)$を推定する問題と定式化できる。

密度比推定の方法のバリエーションについて、以下の図表 12 に示した。今回は、この整理を参考に uLSIF を採用した。

図表 12 密度比推定のバリエーション

手法名	特徴
kernel density estimation; KDE	高次元データに対するノンパラメトリック密度推定は精度が良くない。
kernel mean matching: KMM	KDE の弱点を克服しようとしているが、モデル選択法がないため、カーネル幅などのチューニングパラメータを適切に設定するのが実用上難しい。
ロジスティック回帰に基づく方法	クロスバリデーションによりモデル選択を行う事ができるという特徴を持っているが、解を求めるためには非線形最適化問題を解く必要があるため、計算に時間がかかる。
Kullback-Leibler importance estimation procedure: KLIEP	
least-squares importance fitting: LSIF	クロスバリデーションによるモデル選択が可能な推定法であり、更に,ロジスティック回帰に基づく方法や KLIEP よりも計算効率が良い。しかし、数値的にやや不安定であるという弱点がある。
unconstrained LSIF: uLSIF	クロスバリデーションによるモデル選択が可能であり、更に解が解析的に求まる故に解が高速かつ安定に計算される。uLSIF が実用上最も優れた密度比推定法であると考えられる。

※ Sugiyama(2010), Härdle et al.(2004), Huang et al.(2007), Qin(1998), Cheng and Chu(2004), Bickel et al.(2007), Sugiyama et al.(2008), Kanamori et al.(2009)を参考に筆者作成。

なお、uLSIF は LSIF の一部手順を変更した手法であるため、以下では、まず、Least-squares importance fitting: LSIF による方法について紹介し、その後、今回採用した unconstrained LSIF: uLSIF について述べる。

Kanamori et al.(2009)で提案された Least-squares importance fitting: LSIF は、二乗損失のもとで、この密度比の適合行う手法である。この手法では、密度比を、以下のようにモデル化する。

$$\begin{aligned}\widehat{w}(x) &= \sum_{l=1}^{b} \alpha_l \varphi_l(x) \\ &= \boldsymbol{\varphi}(\boldsymbol{x})^{\top}\boldsymbol{\alpha}\end{aligned} \tag{1}$$

ここで、$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_b)^{\top}$はパラメータベクトルであり、$\boldsymbol{\varphi}(x): \mathbb{R}^d \to \mathbb{R}^b$ は、非負の値を取る基底関数ベクトルである。$\boldsymbol{\alpha}$は、以下の二乗誤差J_0を最小にするように学習する。

$$\begin{aligned}J_0(\boldsymbol{\alpha}) &\coloneqq \frac{1}{2}\int \big(\widehat{w}(x) - w(x)\big)^2 P_S(x)\, dx \\ &= \frac{1}{2}\int \widehat{w}(x)^2 P_S(x)\, dx - \int \widehat{w}(x) P_T(x)\, dx + \frac{1}{2}\int w(x) P_T(x)\, dx\end{aligned} \tag{2}$$

第三項は定数のため無視し、第一項・第二項のみをJと定義する。

$$J(\boldsymbol{\alpha}) \coloneqq \frac{1}{2}\int \widehat{w}(x)^2 P_S(x)\, dx - \int \widehat{w}(x) P_T(x)\, dx \tag{3}$$

Jに含まれる期待値を標本平均で近似すると、以下の式が得られる。

$$\hat{J}(\boldsymbol{\alpha}) \coloneqq \frac{1}{2n}\sum_{i=1}^{n}\hat{w}(x_i)^2 - \frac{1}{n'}\sum_{j=1}^{n'}\hat{w}(x_j)^2$$

$$= \frac{1}{2}\sum_{l,l'=1}^{b}\alpha_l\alpha_{l'}\hat{H}_{l,l'} - \sum_{l=1}^{b}\alpha_l\hat{h}_l \tag{4}$$

ここで、

$$\hat{H}_{l,l'} \coloneqq \frac{1}{n}\sum_{i=1}^{n}\varphi_l(x_i)\varphi_{l'}(x_i) - \sum_{l=1}^{b}\alpha_l\hat{h}_l$$

$$\hat{h}_l \coloneqq \frac{1}{n'}\sum_{j=1}^{n'}\varphi_l(x'_j)$$

である。密度比関数の非負性を考慮し、正則化項を加えると、以下の最適化問題が得られる。

$$\min_{\{\alpha_l\}_{l=1}^{b}}\left[\frac{1}{2}\sum_{l,l'=1}^{b}\alpha_l\alpha_{l'}\hat{H}_{l,l'} - \sum_{l=1}^{b}\alpha_l\hat{h}_l + \lambda\sum_{l=1}^{b}\alpha_l\right]$$

$$subject\ \ to\ \ \alpha_1,\alpha_2,\dots,\alpha_b \geq 0 \tag{5}$$

ここで、λは非負の正則化パラメータである。以上が、LSIF である。

続いて、今回採用した uLSIF であるが、LISF における (6)式の非負拘束を外し、正則化項を 2 次に変更する。

$$\min_{\{\alpha_l\}_{l=1}^{b}}\left[\frac{1}{2}\sum_{l,l'=1}^{b}\alpha_l\alpha_{l'}\hat{H}_{l,l'} - \sum_{l=1}^{b}\alpha_l\hat{h}_l + \frac{\lambda}{2}\sum_{l=1}^{b}\alpha_l^2\right] \tag{6}$$

(7)の解は解析的に、以下の式で与えられる。

$$\tilde{\boldsymbol{\alpha}} = (\tilde{\alpha}_1,\tilde{\alpha}_2,\dots,\tilde{\alpha}_b)^{\mathrm{T}} = \left(\hat{\boldsymbol{H}} + \lambda\boldsymbol{I}_b\right)^{-1}\hat{\boldsymbol{h}} \tag{7}$$

$$\hat{\boldsymbol{H}} \coloneqq \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\varphi}(x_i)\boldsymbol{\varphi}(x_i)^{\mathrm{T}}$$

$$\hat{\boldsymbol{h}} \coloneqq \frac{1}{n'}\sum_{j=1}^{n'}\boldsymbol{\varphi}(x_j)$$

ここで、$\boldsymbol{I}_b$はb次元の単位行列である。非負拘束を外したため、パラメータ値は負になる可能性がある。そのため、以下の補正を行う。

$$\hat{\alpha}_l = \max(0,\tilde{\alpha}_l) \quad for\ \ l = 1,2,\dots,b$$

以上が、uLSIF である。

なお、今回、基底関数は Kanamori et al.(2009)で推奨されている Gaussian Kernel

$$K_\sigma(x,x') \coloneqq \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$$

を採用した。

uLSIF による推定時に考慮する属性情報については、いくつかの組み合わせを試したが、年齢、最終学歴、前所属企業の従業員数の 3 変数の利用が最も精度が高かったため、この組み合わせを採用した。

3.4 ラベル情報の補正

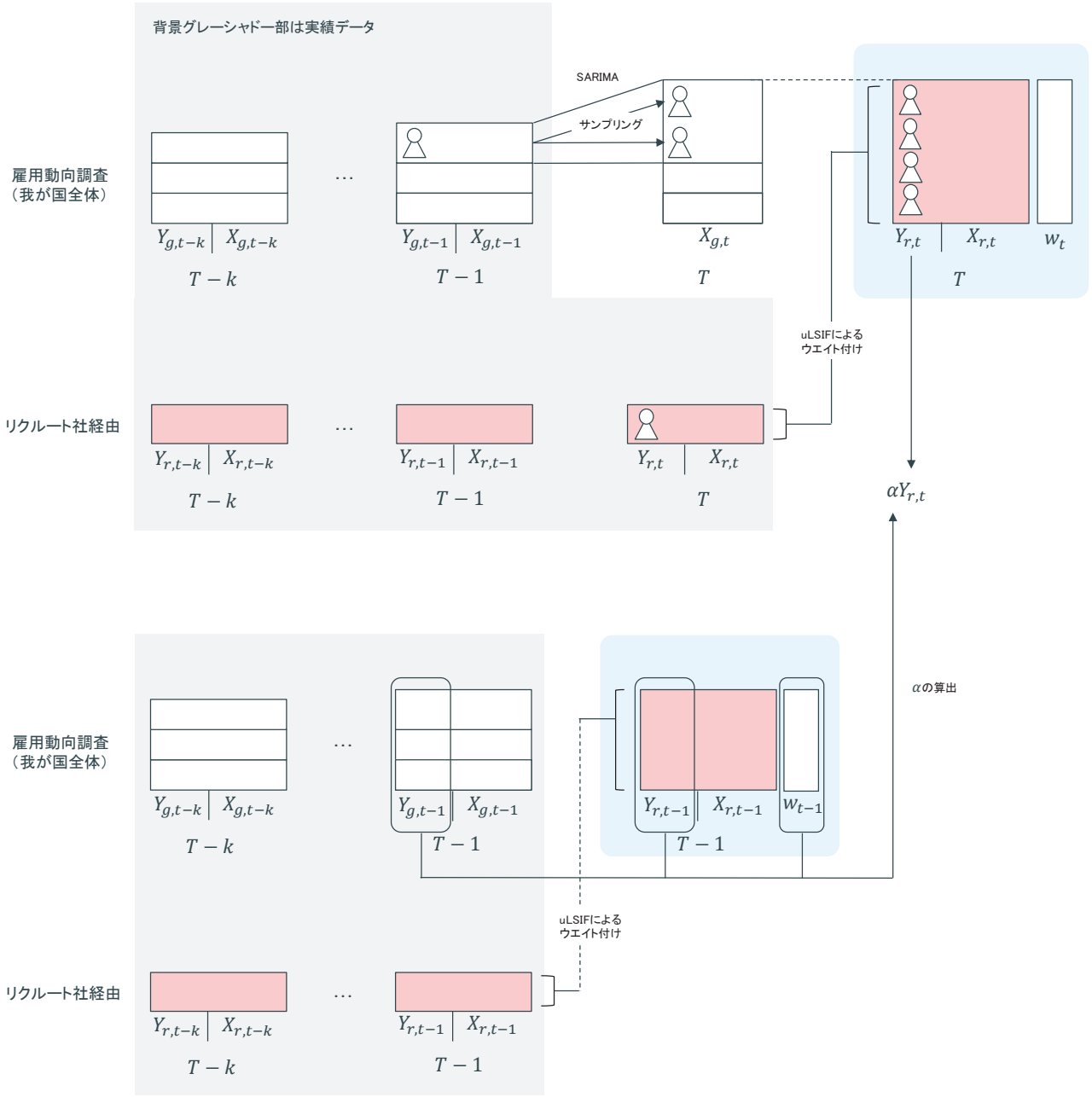
3.3節にてウエイトを付与した目的は、リクルート社のデータが保有する最新の「属性情報が与えられたもとでのラベル情報（転職時の賃金変動状況）」を、我が国全体に“引き伸ばす”ためであった。即ち、ウエイト付与により、我々は推計期間における転職者のラベル情報を得た事になる。このラベルを集計する事により「転職時に賃金が明確に（1割以上）増加した転職者の割合」が得るわけであるが、このラベルの集計の前には、以下で述べる補正を行っている。具体的には、得られたラベル情報に一律で定数を乗じる処理を施す。定数を乗じる理由は、今回のシチュエーションを「共変量シフトを一部、緩和した状況」であると捉えているためである。

共変量シフトとは「与えられた入力に対する出力の生成規則は訓練時とテスト時で変わらないが、入力（共変量）の分布が訓練時とテスト時で異なるという状況」である。今回の設定において言い換えれば「我が国全体とリクルート社のデータとの間で転職者の属性分布は異なるが、属性情報が与えられたもとでのラベル情報は同じである」となる。属性情報が十分に取得できるもとでは「属性情報が与えられたもとでのラベル情報が同じ」という仮定は、今回の状況において自然な仮定と言えよう。一方、属性情報が十分には取得できないもとでは、その限りではない。今回は、属性情報が十分には取得できていないと見做した上で、我が国全体とリクルート社のデータとの間では、共通の属性情報が与えられたもとで、ラベル情報に一定のバイアスが生じると仮定し、そのバイアスを、定数を乗じる事で緩和するという処置を考えた。この定数をαと呼ぶ。ラベル情報ベクトル$Y_{r,t}$の成分は「1」か「0」の2値変数であり、仮に定数αが「0.8」であったとした場合、$\alpha Y_{r,t}$の成分は「0.8」か「0」の2値変数となる。3.3節の手順で算出したウエイト付与後のリクルート社のT期のラベル情報ベクトルに対して、この定数αを乗じ、成分の値の平均 $\alpha Y_{r,t}{}^{\mathsf{T}} w_t / d_t$（$d_t$はラベル情報ベクトル$Y_{r,t}$とウエイト情報ベクトル$w_t$の次元）を求めたものが図表9における「推計A」となる。なお、3.5節における分類モデルを伴った手法の結果（図表9：推計B）や、3.6節における回帰モデルを伴った手法の結果（図表9：推計C）の算出にあたっては、αを乗じる対象が異なる。この点は3.5節・3.6節にて後述する。

αの値は、ウエイト付与後のリクルート社データのラベル情報で算出された「転職時に賃金が明確に（1割以上）増加した転職者の割合」（$Y_{r,t}{}^{\mathsf{T}} w_t / d_t$）を分母、実際の雇用動向調査における当該指標の実績値（$Y_{g,t}$の成分の平均値）を分子として算出した値を複数の期間Tについて平均する事で求める（なお、図表13では、概念説明のためT-1の1時点のみを図示している）。

αを算出する際の期間Tについて、今回は二つのケースを試した。一つ目は、今回の全推計期間2012年下期～2018年上期の平均を採用したケースである。αを時間に依存しない値と仮定し、時間軸の観点で可能な限りの多くの情報を集めた格好である。ただ、この算出方法は、本来その時点では取得不可能な未来の情報を使用している事になる。そこで、二つ目のケースとして、推計対象期間より前の利用可能な期間のみで平均を算出した場合を試している。

図表 13 ラベル情報の補正の概念図



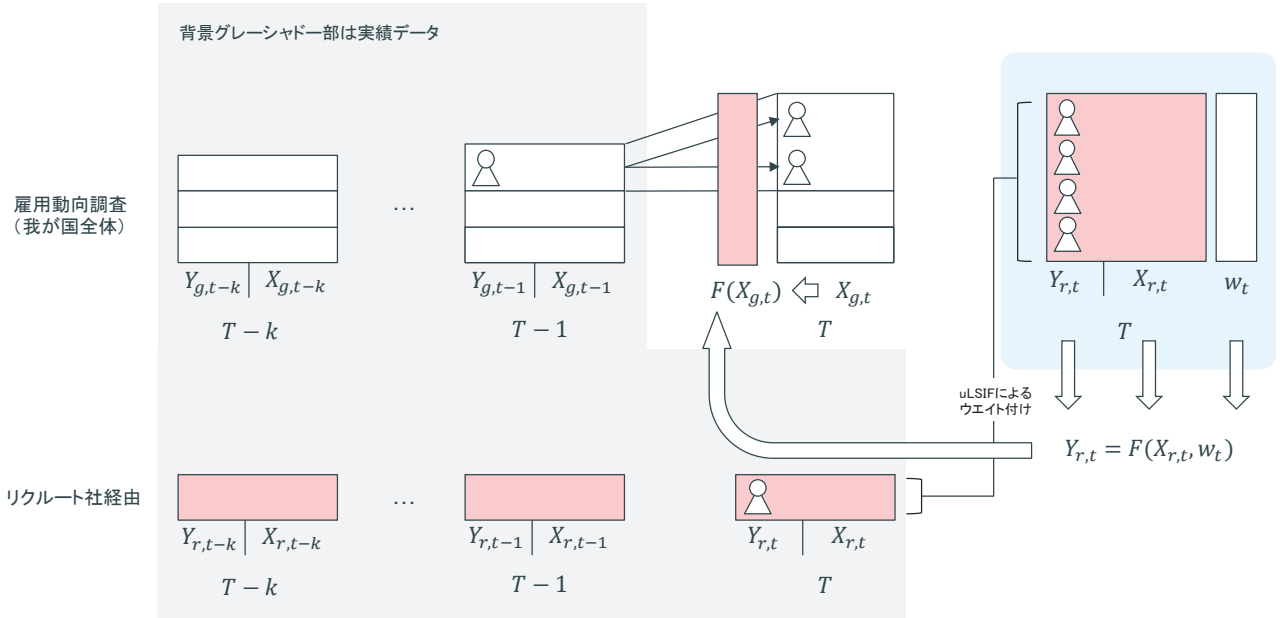
3.5 密度比をウエイトとした分類モデルによる追加処理

3.3 節における推計は、リクルート社のデータに対して、確率密度比によるウエイト付けを行う事で、そのカバレッジを我が国全体へ拡大させるためのものであった。仮に、あらゆる属性において我が国全体の分布に一致するようなウエイト付けが可能であれば、以降の処理を検討する必要はない。以降の処理は、3.3 節までの推計で生じる誤差が小さくないという前提に立った上で、追加的な処置を加える事で、誤差を縮小するための試みといえる。

カバレッジを我が国全体へ引き伸ばしたリクルート社のデータについて、ラベル情報のベクトルを$Y_{r,t}$、属性情報の行列を$X_{r,t}$、ウエイト情報をw_tと記す。また、3.2 節で求めた属性情報の行列については$X_{g,t}$とする。この表記に従うと、3.3 節までの手順と 3.4 節の補正のみで求めた推計結果（図表 9 で「推計 A」と記した手順の結果）は、$\alpha Y_{r,t}{}^{\mathrm{T}} w_t / d_t$（$d_t$はラベル情報ベクトル$Y_{r,t}$とウエイト情報ベクトル$w_t$の次元）となる。一方、3.5 節における推計（図表 9 で「推計 B」と記した手順）では、$Y_{r,t} = F(X_{r,t}, w_t)$となるような$F$を推定した上で$F(X_{g,t})$を求め、$\alpha F(X_{g,t})$ のベクトル成分（分類確率のスコアを用いる）の平均値を求める（図表 14）。Fの手法としては、ロジスティック回帰を用い、変数選択については、Stepwise のような探索は行わず、Elastic Net による正則化法を適用した。

3.4 節で述べた通り、今回の状況は「共変量シフトを一部、緩和した状況」と見做す事ができる。共変量シフト下では、最尤推定などの標準的な学習法はバイアスを持つが、Shimodaira（2000）では、このバイアスは損失関数を確率密度比によってウエイト付けする事で漸近的に打ち消す事ができる事が示されている。

図表 14 追加処理の概念図



3.6 密度比をウエイトとした回帰モデルによる追加処理

雇用動向調査においては、賃金情報は転職前後の変化のカテゴリ（例：「1 割以上 3 割未満増加」「3 割以上増加」「変わらない」）によって捕捉している。一方、リクルート社のデータでは賃金情報を連続値（転職前後のそれぞれで単位「円」の具体的な金額情報）として取得できる。この利点を活かすべく、ここでは$Y_{r,t}$を「分子を転職後の賃金、分母を転職前の賃金」とした回帰問題として扱う。転職後に賃金が 1 割増加した場合、$Y_{r,t}$の値は「1.1」となる。回帰問題としての推計値を得た後は、$F(X_{g,t})$の値を「分類モデルにおける確率スコア」に相当する 0 以上 1 以下の値に変換している。その後は、分類問題の時と同様に、$\alpha F(X_{g,t})$ のベクトル成分の平均値を求める。

なお、F を求める際には、事前に$Y_{r,t}$に対して Box-Cox 変換を施した上で、分類モデルと同様に Elastic Net による正則化法を用いた。また、$F(X_{g,t})$の「分類問題の確率スコアに相当する値」へ変換は、残差が正規分布に従うと仮定して、$F(X_{g,t})$が「1.1 を Box-Cox 変換した値」を超える確率を求める事で行った（すなわち、$F(X_{g,t})$の値が「1.1 を Box-Cox 変換した値」と一致する時、スコアは 0.5 となる）。

4. 推計結果

4.1 転職時に賃金が明確に(1 割以上)増加した転職者の割合の推計結果

4.1 節では、本稿の推計の最終目的である「転職時に賃金が明確に（1 割以上）増加した転職者の割合」の結果について示す。まず、前半部分で推計対象期間などの結果を確認するための前提について触れた後、後半にて具体的な結果を述べる。

本稿の取組みは、雇用動向調査における実績が示す時点と、現時点とのタイムラグを、リクルート社のデータによって埋める構造にあるが、現段階では、雇用動向調査の個別サンプル情報がどのタイミングで利用可能となるか確定していないため、このタイムラグの幅も確定しない。例えば、仮に現在が 2021 年 1 月 1 日としよう。リクルート社のデータはリアルタイムに取得可能なので、この段階では 2020 年 12 月 31 日までの情報が利用可能である[4)]。言い換えれば、2020 年 7 月～12 月期(当該半期を T とする)のナウキャストが可能である。この時、仮に、2020 年 1 月～6 月分の雇用動向調査の個別サンプル情報が手に入っていれば、タイムラグは 1 半期であり、T-1 期までのデータがある下で、T 期のナウキャストを行う構造にあると言える。一方、2020 年 1 月～6 月分の雇用動向調査の個別サンプル情報が推計時点では手に入らず、2019 年 7 月～12 月分までのみが利用可能というような状態であれば、T-2 期までのデータがある下で、T 期のナウキャストを行う構造と言える（図表 15)。

今回、3.2 節、3.3 節、3.5 節、3.6 節で述べた手順の推計対象期間としては「2012 年 7 月～12 月期」から「2018 年 1 月～6 月期」の 12 時点[5)]を設定しているが、3.4 節の補正処理では推計対象よりも前の期間の推計結果が必要なため、2012 年 7 月～12 月期は補正を伴う結果が得らない。すなわち、補正を伴った最終的な推計結果が得られるのは「2013 年 1 月～6 月期」以降の 11 時点となる。また、雇用動向調査のデータの始点は「2004 年 1 月～6 月期」である。例えば、「2012 年 7 月～12 月期」を T として、T-1 までの情報が入手可能という前提のもとにナウキャストする場合、「2004 年 1 月～6 月期」から「2012 年 1 月～6 月期」の情報を利用し、3.2 節以降で紹介した過程を実行する。同じ「2012 年 7 月～12 月期」を T としてナウキャストする場合でも、T-2 までの情報のみが入手可能という前提に立った場合には、「2004 年 1 月～6 月期」から「2011 年 7 月～12 月期」までの情報を利用するかたちとなる。なお、図表 15 において「サンプリングのラグ」と記載した部分であるが、3.2 節でも述べた通り、サンプリングを行う期としては、言葉の通りの直近期（ラグが 1 半期）と季節性を加味した直近期（ラグが 2 半期＝1 年）の両方を試している[6)]。例えば、推計対象期間が 2012 年 7 月～12 月であった場合、「言葉の通りの直近期」は 2012 年 1 月～6 月となり、季節性を加味した直近期は一年前の 2011 年 7 月～12 月となる。

このような前提のもと、以下、図表 16 に、本題である「転職時に賃金が明確に（1 割以上）増加した転職者の割合」の推計結果の誤差の絶対値の平均（MAE：Mean Absolute Error）を示した。ここでは「2013 年 1 月～6 月期」から「2018 年 1 月～6 月期」の 11 時点全体の平均（上段）と、「2017 年 7 月～12 月期」を除いた 10 期間平均（下段）を算出している。「2017 年 7 月～12 月期」は、今回推計対象とした雇用動向調査側の時系列が大きく変動している一方、リクルート社のデータではそれが見られず、いずれの推計結果においても実績とのかい離が非常に大きくなっているため、参考までに当該期間を除いた場合の平均も算出した格好である（「誤差の絶対値の時系列」は別紙の図表 20 参照)。

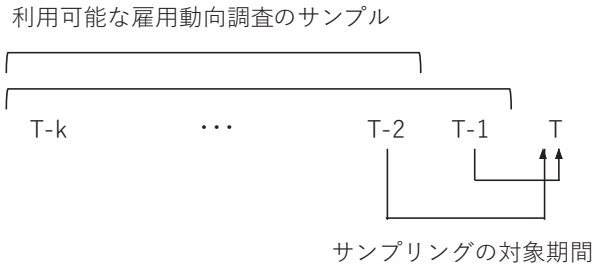
---

4) 実際には休日（年末年始の休み等）の問題もあり、1 月 1 日にデータを取得するのは実務上難しいといった話が別途存在するが、ここでは説明の明快さを優先し、そうした問題は扱わない事とする。

5) 執筆時点で取得できていた雇用動向調査標本の最新時点が 2018 年 1 月～6 月期であったため。

6) ここでは雇用動向調査標本を T-1 まで利用したケースにおいて、サンプリングのラグが 1 の（季節性を考慮しない）ケースと、サンプリングのラグが 2 の（季節性を考慮した）ケースを比較している。雇用動向調査標本を T-2 まで利用したケースについては、直近期をサンプリングした場合にラグが 2 となり、季節性が一致しているため、サンプリングのラグを 3 とした場合の検証は行っていない（行う必要がない)。

図表 15 検証の枠組みについて



図表 16 MAE: Mean Absolute Error

αの算出方法[7]：全推計期間平均

（11 期間平均）2013 年 1 月～6 月期から 2018 年 1 月～6 月期

雇用動向調査 サンプル利用期間	サンプリング の対象期間	ウエイト付けのみ	分類	回帰
T-1まで	T-1	1.69%	2.18%	1.88%
T-1まで	T-2	1.32%	1.16%	1.14%
T-2まで	T-2	1.38%	1.34%	1.38%

（10 期間平均）上記より 2017 年 7 月～12 月期を除く

雇用動向調査 サンプル利用期間	サンプリング の対象期間	ウエイト付けのみ	分類	回帰
T-1まで	T-1	1.52%	1.99%	1.74%
T-1まで	T-2	1.12%	0.89%	0.86%
T-2まで	T-2	1.19%	1.10%	1.14%

αの算出方法[8]：推計対象期間より前の期の平均

（11 期間平均）2013 年 1 月～6 月期から 2018 年 1 月～6 月期

雇用動向調査 サンプル利用期間	サンプリング の対象期間	ウエイト付けのみ	分類	回帰
T-1まで	T-1	1.87%	2.43%	2.39%
T-1まで	T-2	1.70%	1.42%	1.36%
T-2まで	T-2	1.64%	1.51%	1.59%

（10 期間平均）上記より 2017 年 7 月～12 月期を除く

雇用動向調査 サンプル利用期間	サンプリング の対象期間	ウエイト付けのみ	分類	回帰
T-1まで	T-1	1.70%	2.23%	2.26%
T-1まで	T-2	1.52%	1.14%	1.06%
T-2まで	T-2	1.47%	1.26%	1.33%

[7] 詳細は 3.4 節を参照。

[8] 詳細は 3.4 節を参照。

図表 16 における列名は、3.3 節で述べた uLSIF によるウエイト付与と 3.4 節で述べた補正処理のみで推計を行った場合の結果（図表 9：推計 A）を「ウエイト付けのみ」と記している。また、3.5 節における分類モデルを伴った手法の結果（図表 9：推計 B）は「分類」、同様に、3.6 節における回帰モデルを伴った手法の結果（図表 9：推計 C）は「回帰」とした。

図表 16 に示した結果の解釈であるが、まず、サンプリング時に季節性を考慮すべきかという点に関しては、「T-1 までの情報を得られたケース」において「季節性を考慮せず最も近い T-1 からサンプリングをした場合」と「季節性を考慮して T-2 からサンプリングをした場合」を比較すると、11 期間平均、10 期間平均、ウエイト付けのみ、分類、回帰、いずれの場合においても「季節性を考慮して T-2 からサンプリングをした場合」の方が、明確に精度が良く、季節性は考慮すべきという結論となった。

次に、密度比をウエイトとした教師付き学習による追加処理が機能しているかを確認するために「ウエイト付けのみ」と「分類」、「ウエイト付けのみ」と「回帰」をそれぞれ比較すると、「季節性を考慮せず、最も近い T-1 からサンプリングをしたケース群」を除いた 16 のケース[9]のうち、15 ケースで「分類」もしくは「回帰」が、「ウエイト付けのみ」よりも良い結果となり、3.5 節、3.6 節の追加処理が一定機能している事が確認できた。

また、「分類」と「回帰」の間の比較では、明確な差はみられなかったものの、僅かながら「回帰」の方が望ましい結果となった。

最も良い精度となった「回帰」のケースでは「α：全期間平均」、「T-1 までの情報を得られたケース」において「季節性を考慮（T-2 からサンプリング）した場合」、11 期間平均で 1.14%、10 期間平均で 0.86%、「T-2 までの情報を得られたケース」では 11 期間平均で 1.38%、10 期間平均で 1.14%、「α：推計対象より前の平均」、「T-1 までの情報を得られたケース」において、「季節性を考慮（T-2 からサンプリング）した場合」では、11 期間平均で 1.36%、10 期間平均で 1.06%、「T-2 までの情報を得られたケース」では 11 期間平均で 1.59%、10 期間平均で 1.33%であった。

[9] 図表 16 の各表の 2 行目と 3 行目を参照（1 行目が季節性を考慮しないケース）。

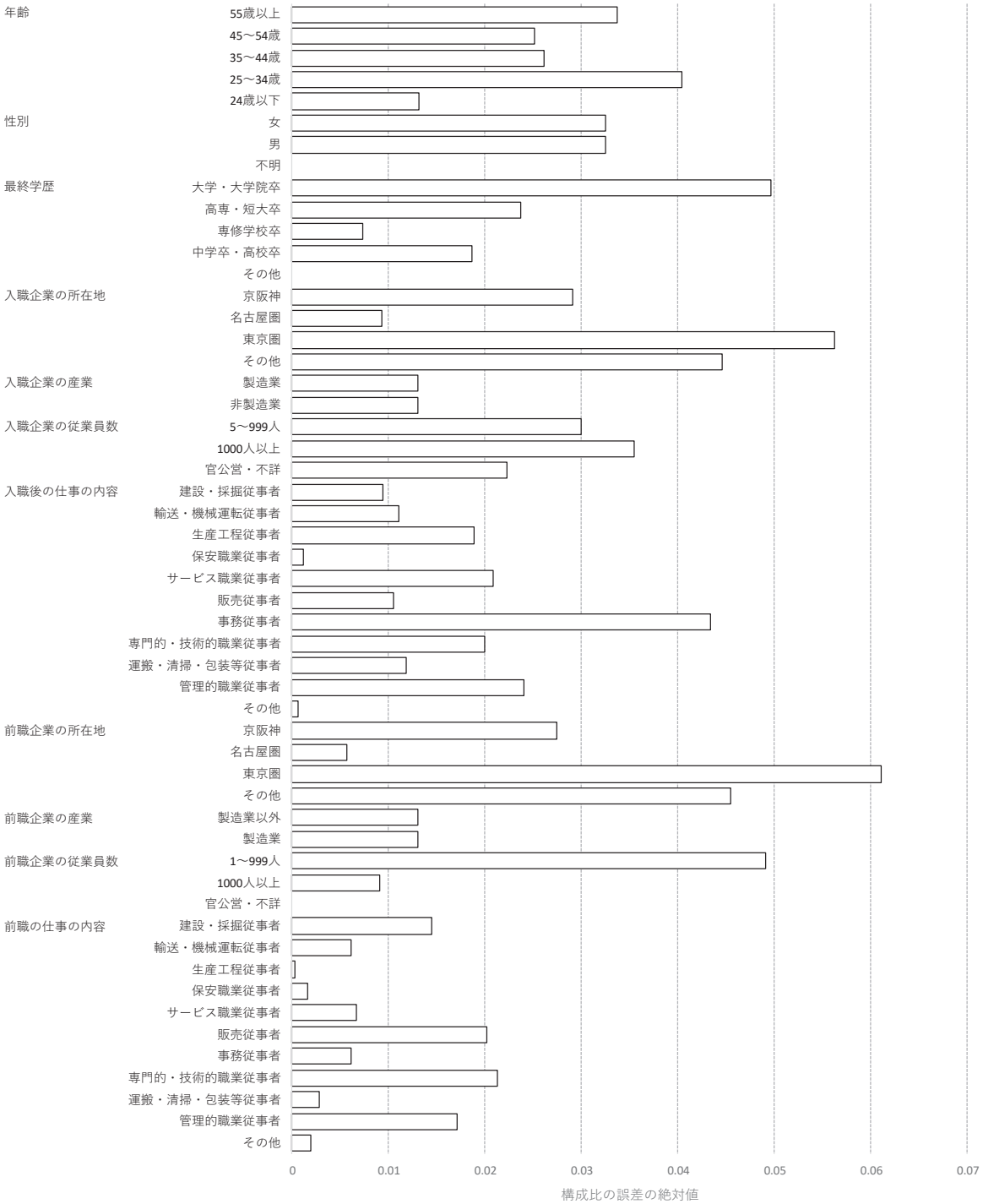
4.2 転職者人数と属性情報の推計の結果について

ここからは、推計プロセス毎の結果について補足していく。3.2 節では「推計対象期間の転職者の人数と、それらの属性情報」の算出を行った。具体的には、まず、① 転職経路毎の転職者数を SARIMA モデルによって求めた。その後、② その直近期の転職者サンプルから、総量が①で求めた者数と一致するように、重複を許すかたちでランダムにサンプリングを行った。本取組みで最終的に求めたい値は、転職時に明確に（1 割以上）賃金が増加した転職者の“割合”である。故に、転職者の総人数はこれに直接的な影響を与えない。すなわち、3.2 節のプロセスでは、①②の結果として、属性情報の“分布”が推計できているか否かが論点となる。

4.2 節では、3.2 節（上述①②）のプロセスの結果として、属性情報の分布の予実乖離を確認するため、属性項目ごとに構成比の誤差の絶対値を観察した（図表 17）。図表 17 では、縦軸が各属性項目の値、横軸が構成比の誤差の絶対値を示している。例えば、性別という属性項目には「女性」「男性」「不明」という値がある。女性の構成比について、実際は 39.1%であり、推計結果が 35.9%であったとする。この場合、誤差の絶対値は 3.2%（図表内の軸では 0.032）となる。今回の推計では、10%以上乖離してしまうような属性項目は存在しなかった。5%を超える誤差が存在した項目は「前職企業の従業員数」、「前職の所在地」、「企業の所在地」であった。当該推計においてはリクルート社のデータによる最新情報を利用しておらず、あくまで SARIMA モデルによって経路毎の人数を過去のトレンドを捉えているのみであり、改善の余地は大きい。今後の課題としては、前述の乖離が大きな属性を中心として、構成比の変化を早期に捉える事のできる情報を追加するという方向が考えられる。

図表 17 各属性項目における構成比の誤差の絶対値の一例

SARIMA モデルによる推計と実績の乖離：2017 年 7 月～12 月期と 2018 年 1 月～6 月期の平均[10)]



※1 所在地の分類）東京圏：東京、神奈川、埼玉、千葉、名古屋圏：愛知、岐阜、三重、京阪神：京都、大阪、兵庫

※2 前職の所在地において、入職前の 1 年以内に仕事についた経験がない場合は入職前の居住地を記載。

10) ここでは、雇用動向調査標本を T-1 まで利用し、季節性を考慮するためにサンプリングのラグを 2 としたケースを示している。

4.3 リクルート社のデータに対するウエイト付けの結果

3.3節では、リクルート社のデータにおける属性の分布を、3.2節で推計した雇用動向調査の標本（我が国全体の標本）における属性の分布に近づけるため、リクルート社のデータレコードに対するウエイトの付与を行った。4.3節では、この3.3節の推計について、各属性における乖離の度合いを把握するため、属性ごとに構成比の誤差の絶対値を確認する（図表18）。

3.3節でも述べた通り、今回、uLSIFによる確率密度比推定の際に用いた属性情報は「年齢」、「最終学歴」、「前職企業の従業員数」である。それらの属性項目においては、ウエイト付与前のリクルート社のデータと、雇用動向調査による実績値との間の乖離幅（図表18：灰色棒グラフ）と比べ、推計値（ウエイト付与後のリクルート社データ）と、雇用動向調査による実績値との乖離幅（図表18：白棒グラフ）が小さく、明確な改善がみられる。一方で、それ以外の属性においては、前職の所在地、企業の所在地などで幾分改善傾向がみられるものの、前職・入職後ともに仕事内容では殆ど縮小がみられず、僅かながら悪化している項目もあるという結果となった。

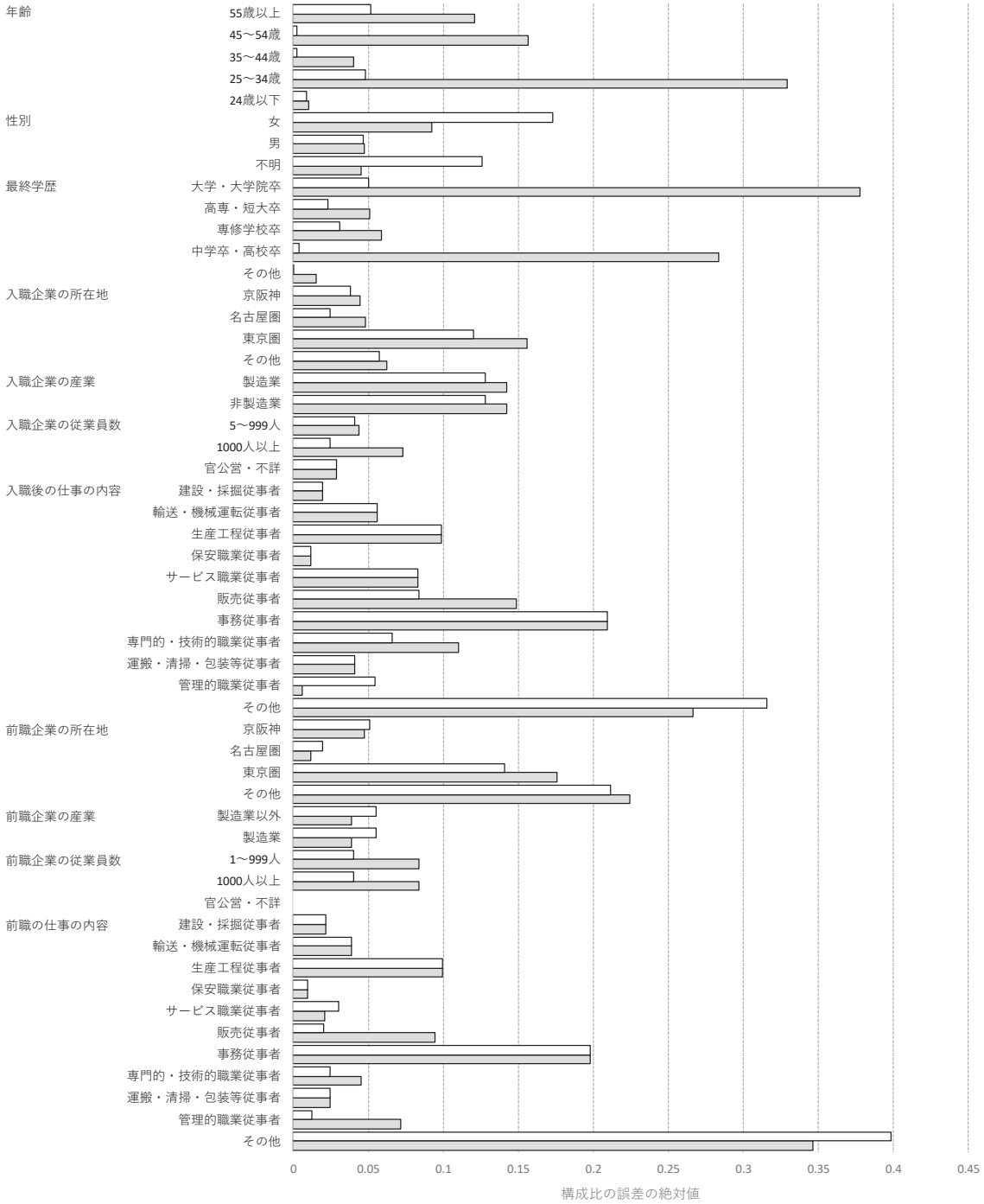
確率密度比推定によるウエイト付けの手法については、前述のとおり、様々なバリエーションが存在しており、それらを試す事で今回よりも良い推計結果が得られる可能性がある。それらの手法の比較検討については、今後の課題としたい。

図表 18 各属性項目における構成比の誤差の絶対値の一例

白棒：uLSIF でのウエイト付与による推計値（リクルート社のデータを引き伸ばしたもの）と実績値（雇用動向調査）との乖離

灰色棒：リクルート社のデータそのものと実績値（雇用動向調査）との乖離

ともに 2017 年 7 月～12 月期から 2018 年 1 月～6 月期の平均[11)]



※1 所在地の分類）東京圏：東京、神奈川、埼玉、千葉、名古屋圏：愛知、岐阜、三重、京阪神：京都、大阪、兵庫

※2 前職の所在地において、入職前の 1 年以内に仕事についた経験がない場合は入職前の居住地を記載。

18) ここでは、雇用動向調査標本を T-1 まで利用し、季節性を考慮するためにサンプリングのラグを 2 としたケースを示している。

4.4 密度比をウエイトとした分類・回帰モデルによる追加処理の結果

3.3 節 では、リクルート社のデータに対して、uLSIF によって算出した確率密度比でウエイト付けをする事で、そのカバレッジを我が国全体へ拡大させる処理を行った。その後、3.5 節、3.6 節では、3.3 節までの推計で生じる誤差が小さくないという前提に立った上で、追加的な処置を加える事で、誤差の縮小を試みた。具体的には、カバレッジを我が国全体へ引き伸ばしたリクルート社のデータについて、ラベル情報のベクトルを$Y_{r,t}$、属性情報の行列を$X_{r,t}$、ウエイト情報をw_t、3.2 節で求めた属性情報の行列を$X_{g,t}$とした際に、$Y_{r,t} = F(X_{r,t}, w_t)$となるような$F$を推定した上で$F(X_{g,t})$を求めた。

4.4 節では、$Y_{r,t} = F(X_{r,t}, w_t)$となるような$F$を求める過程における精度を確認する。なお、この評価は ROC-AUC によって行った。3.6 節で述べた通り、今回の推計においては、回帰問題の場合においても、分類問題と同様のスコアに変換しているため、AUC を用いた同一指標での比較が可能である。

以下の図表 19 では、$Y_{r,t} = F(X_{r,t}, w_t)$となるような$F$の推定について、第一列目に学習時の AUC、第二列目に検証時の AUC、第三列目に、引数をX_Gとした$F(X_G)$の AUC を示した。「分類」・「回帰」の比較としては、4.1 節で述べた結果と整合的で、僅かながら「回帰」が優れているという結果となった。もっとも、分類・回帰問わず、全体として当該プロセスの精度水準は高くなかった。今後は、決定木を用いた機械学習モデルなどの適用も検討し、精度向上を目指したい。

図表 19 ROC-AUC の比較

2013 年 1 月～6 月期～2018 年 1 月～6 月期の 11 時点平均

	学習	検証 (Source domain)	検証 (Target domain)
分類	0.61	0.58	0.56
回帰	0.60	0.60	0.58

学習 ： ウエイト付与後のリクルート社のデータで学習した際の AUC

検証（Source domain）： ウエイト付与後のリクルート社のデータで学習し、別途確保しておいた学習データとは異なる同社のデータで検証した際の AUC

検証（Target domain）： ウエイト付与後のリクルート社のデータで学習し、雇用動向調査のデータで検証した際の AUC

5 結びにかえて

以上、厚生労働省「雇用動向調査」における「転職時の賃金変動状況：転職時に賃金が明確に（1 割以上）増加した転職者の割合」について、リクルート社の保有するトランザクションデータを用いる事で、速報ベースの指標を作成できる可能性を示した。現状は、1 月～6 月分が 12 月末（令和 2 年の場合は翌年 2 月初）公表、7 月～12 月分を加えた暦年分が 8 月下旬（令和元年分は翌年 9 月末）公表というかたちでタイムラグが存在するが、1%強の速報・確報間誤差を許容する前提のもとでは、概ねリアルタイムでの公表が可能になる。1.2 節でも述べた通り、我が国の労働市場に関する代表的な月次統計としては、総務省「労働力調査」、厚生労働省「毎月勤労統計」、同「一般職業紹介状況」の三つがあるが、今回推計対象とした雇用動向調査における「転職時の賃金変動状況：転職時に賃金が明確に（1 割以上）増加した転職者の割合」という指標は「労働需給の逼迫が賃金に対する上昇圧力として、どの程度、機能しているのか」という、この三指標では捕捉できない貴重な情報を与えてくれる。現状は、調査から公表までのタイムラグの存在により、足許の労働市場の状況を踏まえたリアルタイムな意思決定局面においては、当該指標を活用する事は難しいが、本稿で提案した速報化が実現すれば、こうした局面においても活用可能となり、政策・経営上の意思決定の質向上に繋がると考えられる。

また、民間企業が保有するトランザクションデータが経済指標として活用されるケースは、① トランザクションデータの集計値そのものを代替指標として活用するケース、② 公的統計のナウキャスト指標を作成するケース、③ 既存統計の推計方法を一部変更してトランザクションデータを活用した推計を取り入れるケースの三つに分類する事ができるが、我が国における労働統計へのビッグデータの活用事例は、求人広告事業や人材紹介事業を行う企業・業界団体が独自に公表している指標が上述①の文脈で活用されるケース以外は、筆者が知る限り存在しない。労働市場は、言うまでもなく、経済分析における最重要ドメインの一つである。それにも関わらず、POS データの活用による消費・物価指標等に代表される他のドメインの進捗と比較すると、幾分遅れを取っている状況にあると言える。当該分野の発展のためには、産官学のより密な連携が必要である。本稿がその一助となる事を願っている。

別紙

図表 20 「転職時に賃金が明確に（1 割以上）増加した転職者の割合」推計における誤差の絶対値の時系列（4.1 節では MAE のみ表示）

αの算出方法[12]：全推計期間平均

	雇用動向調査サンプル利用期間	サンプリングの対象期間	2013 1月~6月	2013 7月~12月	2014 1月~6月	2014 7月~12月	2015 1月~6月	2015 7月~12月	2016 1月~6月	2016 7月~12月	2017 1月~6月	2017 7月~12月	2018 1月~6月	平均 11期	平均 10期
ウエイト付けのみ	T-1まで	T-1	1.29%	0.30%	0.55%	1.91%	0.12%	2.14%	2.66%	2.64%	2.56%	3.39%	0.98%	1.69%	1.52%
	T-1まで	T-2	0.67%	0.79%	0.26%	2.19%	0.87%	1.20%	1.54%	1.08%	0.94%	3.35%	1.65%	1.32%	1.12%
	T-2まで	T-2	0.14%	0.71%	0.49%	1.60%	1.01%	1.89%	1.59%	1.99%	1.12%	3.20%	1.42%	1.38%	1.19%
分類	T-1まで	T-1	1.98%	0.45%	0.17%	2.02%	1.64%	2.05%	3.59%	5.34%	2.42%	4.06%	0.28%	2.18%	1.99%
	T-1まで	T-2	1.01%	0.14%	0.15%	1.49%	2.74%	0.32%	1.80%	0.55%	0.14%	3.89%	0.50%	1.16%	0.89%
	T-2まで	T-2	0.26%	0.78%	0.14%	1.00%	2.93%	0.99%	2.17%	1.57%	0.91%	3.65%	0.30%	1.34%	1.10%
回帰	T-1まで	T-1	1.39%	0.79%	2.33%	2.74%	0.46%	1.49%	5.17%	1.30%	1.13%	3.31%	0.61%	1.88%	1.74%
	T-1まで	T-2	0.56%	1.83%	0.25%	0.30%	0.82%	0.73%	1.72%	0.84%	1.19%	4.01%	0.33%	1.14%	0.86%
	T-2まで	T-2	0.61%	1.41%	1.60%	0.53%	0.73%	0.43%	4.42%	0.47%	1.01%	3.74%	0.18%	1.38%	1.14%

αの算出方法[13]：推計対象期間より前の期の平均

	雇用動向調査サンプル利用期間	サンプリングの対象期間	2013 1月~6月	2013 7月~12月	2014 1月~6月	2014 7月~12月	2015 1月~6月	2015 7月~12月	2016 1月~6月	2016 7月~12月	2017 1月~6月	2017 7月~12月	2018 1月~6月	平均 11期	平均 10期
ウエイト付けのみ	T-1まで	T-1	1.76%	0.72%	0.75%	1.92%	0.23%	2.47%	2.68%	2.28%	3.09%	3.63%	1.07%	1.87%	1.70%
	T-1まで	T-2	1.54%	2.23%	1.00%	2.71%	0.85%	1.38%	1.56%	0.88%	1.23%	3.52%	1.80%	1.70%	1.52%
	T-2まで	T-2	1.40%	1.53%	0.18%	1.98%	0.99%	2.07%	1.69%	1.86%	1.44%	3.38%	1.54%	1.64%	1.47%
分類	T-1まで	T-1	1.83%	0.62%	0.78%	2.83%	1.88%	2.12%	3.82%	5.05%	3.08%	4.44%	0.30%	2.43%	2.23%
	T-1まで	T-2	0.50%	1.08%	1.10%	2.19%	3.02%	0.56%	1.67%	0.18%	0.55%	4.23%	0.55%	1.42%	1.14%
	T-2まで	T-2	0.78%	0.19%	0.58%	1.60%	3.23%	1.23%	2.12%	1.21%	1.38%	3.99%	0.32%	1.51%	1.26%
回帰	T-1まで	T-1	4.49%	1.62%	2.62%	3.19%	0.62%	1.56%	5.33%	0.82%	1.68%	3.71%	0.67%	2.39%	2.26%
	T-1まで	T-2	1.71%	1.52%	1.04%	0.82%	0.45%	1.23%	1.43%	0.33%	1.72%	4.38%	0.35%	1.36%	1.06%
	T-2まで	T-2	1.88%	1.06%	2.25%	0.62%	0.77%	0.53%	4.40%	0.07%	1.56%	4.13%	0.20%	1.59%	1.33%

平均 11 期：2013 年 1 月～6 月期から 2018 年 1 月～6 月期の平均。

平均 10 期：上記「平均 11 期」より 2017 年 7 月～12 月期を除く。

[12] 詳細は 3.4 節を参照。

[13] 詳細は 3.4 節を参照。

参考文献

Bickel, S., Brückner, M. and Scheffer, T. (2007) . Discriminative learning for differing training and test distributions, *Proceedings of the 24th International Conference on Machine Learning (ICML2007)*, 81-88

Cheng, K. F. and Chu, C. K. (2004) . Semiparametric density estimation under a two-sample density ratio model, *Bernoulli, 10*, 583-604.

Hammer, G., Kostroch, D., Quiros, G. & STA Internal Group. (2017). Big Data: Potential, Challenges, and Statistical Implications, *Staff Discussion Note,* IMF

Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004) . *Nonparametric and Semiparametric Models,* Springer, Berlin.

Huang, J., Smola, A., Gretton, A., Borgwardt, K. M. and Schölkopf, B. (2007) . Correcting sample selection bias by unlabeled data, *Advances in Neural Information Processing Systems 19*, 601-608, MIT Press, Cambridge, Massachusetts.

Kanamori, T., Hido, S. and Sugiyama, M. (2009) . A least-squares approach to direct importance estimation, *Journal of Machine Learning Research, 10*, 1391-1445.

Kroon, J. and Pank, E. (2012). Mobile positioning as a possible data source for international travel service statistics, *Conference of European Statisticians*, UNECE

Qin, J. (1998) . Inferences for case-control and semiparametric two-sample density ratio models, *Biometrika, 85*, 619-639.

Shimodaira, H. (2000) . Improving predictive inference under covariate shift by weighting the loglikelihood function, *Journal of Statistical Planning and Inference, 90*, 227-244.

Sugiyama, M. (2010). A New Approach to Machine Learning Based on Density Ratios, *Proceedings of the Institute of Statistical Mathematics, 58, no.2*, 141-155.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von B ¨unau, P. and Kawanabe, M. (2008) . Direct importance estimation for covariate shift adaptation, *Annals of the Institute of Statistical Mathematics, 60*, 699-746.

Sugiyama, M., Kanamori, T., Suzuki, T., Hido, S., Sese, J., Takeuchi, I. and Wang, L. (2009) . A density-ratio framework for statistical data processing, *IPSJ Transactions on Computer Vision and Applications, 1*, 183-208.

齊藤誠、岩本康志、 太田聰一、 柴田章久 (2010) マクロ経済学、 有斐閣