別紙

報告書

1　目的と方法

1.1　労働力フローデータの構築

本研究の目的は「労働力調査」の調査票情報を用いて、1980 年代以降のバブル期、就職氷河期、コロナ危機など、異なった経済状況下で労働力フローがどのように推移したのかを分析し、非正規雇用の増加、および、現在の人手不足をもたらす経済構造の解明の手がかりとすることである。

ただし、「労働力調査」の調査票情報を集計して、労働力フローデータを構築する際には、先行研究において十分考慮されていない課題が 2 点ある。本研究では、これらの課題に対処する手法を開発したうえで、上記目的に必要なデータを構築した。課題の第 1 点は、欠測の問題である。すなわち、労働力フローの情報を得るためには、2ヶ月連続調査された標本を用いる必要があるが、その一部にどちらか一方の月の状態が観測されていない標本が存在する。先行研究の多くは、公表フローデータを利用しているため、このような欠測値への対応が不十分であった。第 2 点は、ストックデータとの整合性の問題である。「労働力調査」はローテーションで標本入替があるため、2ヶ月連続調査される標本の規模は全体の約 2 分の 1 となる。このため、フロー値はストック値を集計する標本の半数のみで集計されることになる。そのため、フローデータから得られるストック値は、2 倍することで、おおよそ全標本から得られるストックデータと一致するが、厳密な対応関係となっていない。これは、労働力移動状況を示すフローデータを時系列的に累積した結果としてのストックを見たい場合の障害となる。

1.2　MICE による欠測値の補完

本研究では、労働力状態を、就業、失業、非労働力の状態に区分した 3 変量分析と、さらに就業を、正規雇用、非正規雇用、自営業その他の 3 状態に区分した 5 変量分析を行ったが、以下では簡単化のために、3 変量の場合について説明する。

第 1 の欠測値の問題への本研究の対処法は以下のようである。

「労働力調査」の 15 歳以上人 (L) の労働力状態は次のいずれかに区分される。

就業 (E)
失業 (U)
非労働力 (N)
不詳 (M)

2ヶ月連続調査される標本については、どちらか一方の月にしか観測されない標本があり、それらは欠測事由により、次のように区分される。

転入 (I)
「転出」(O)
「追加」(A)
前月 14 歳で今月 15 歳の者（以下、新 15 歳）(B)
「その他（死亡）」(D)

さらに、2ヶ月連続して調査される回答者のうち，1ヶ月目の観測しかない主体が存在しており、このような主体を

「脱落」(H)
として定義する。

以上の、M、I、O、A、H、B、D が欠測値となる。これらのうち、「転入」および「転出・その他」については公表されるフロー集計表データより得られるが、その他は調査票情報によってしかわからない。

ただし、B と D は人口動態に基づく欠測であり、労働移動の対象ではないため除外する。すると、欠測状態（$\Psi = M, I, O, A, H$）を含めたフロー表は次のようになる。

$$\begin{bmatrix} x^{E\to E} & x^{U\to E} & x^{N\to E} & x^{\Psi\to E} \\ x^{E\to U} & x^{U\to U} & x^{N\to U} & x^{\Psi\to U} \\ x^{E\to N} & x^{U\to N} & x^{N\to N} & x^{\Psi\to N} \\ x^{E\to \Psi} & x^{U\to \Psi} & x^{N\to \Psi} & 0 \end{bmatrix} \tag{1}$$

ここで、$x^{i\to j}$ は状態 i から j へのフロー量である。

欠測値の補完方法として，多重代入法を利用する。多重代入法では，欠測のパターンが単調（monotone）である場合に適用できる方法が多いが、労働力フローデータの欠測発生のパターンは，明らかに単調性を保証しない．単調ではない欠測パターンに対して適用可能な代入法に、連鎖方程式による多重代入法、multivariate imputation using chained equations(MICE) がある。本研究では、一月ごとのデータに対して MICE による補完を行った。補完方程式は多項ロジットモデルである。

t 期の個人 i の観測された労働力状態を $\mathbb{S}_{it} = \{S_{it}\} = \{E, U, N\}$, 対応するインデックス関数（ダミー変数）を I^S とすると、

$$\begin{aligned} \Pr(S_{it}|S_{it-1}, \mathbb{Z}_i, \beta) &= \frac{e^{(I^S_{it-1}, \mathbb{Z}_i)\beta^\Theta}}{e^{(I^S_{it-1}, \mathbb{Z}_i)\beta^E} + e^{(I^S_{it-1}, \mathbb{Z}_i)\beta^E} + e^{(I^S_{it-1}, \mathbb{Z}_i)\beta^E}}, \\ \Pr(S_{it-1}|S_{it}, \mathbb{Z}_i, \mathfrak{b}) &= \frac{e^{(I^S_{it}, \mathbb{Z}_i)\mathfrak{b}^\Theta}}{e^{(I^S_{it}, \mathbb{Z}_i)\mathfrak{b}^E} + e^{(I^S_{it}, \mathbb{Z}_i)\mathfrak{b}^U} + e^{(I^S_{it}, \mathbb{Z}_i)\mathfrak{b}^N}}, \\ \Theta &= E, U, N \end{aligned} \tag{2}$$

ここで、$\mathbb{Z}$ は欠測のないその他の説明変数，$\beta = \{\beta^E, \beta^U, \beta^N\}, \mathfrak{b} = \{\mathfrak{b}^E, \mathfrak{b}^U, \mathfrak{b}^N\}$ は係数ベクトルである（識別のために，たとえば $\beta^N, \mathfrak{b}^N = 0$ とする）。

初期値として、monotone inputation によって、(2) に基づく条件付き確率分布から、欠測値の補完値とパラメータの推定値を得て、それらを初期値 ($k = 0$) とする。ついで、(2) に基づく条件付き確率分布から、$k = 1, 2, ...$ につき、

$$\beta^{(k)} \sim \Pr(\beta|S^{(k-1)}_{it}, S^{(k-1)}_{it-1}, \mathbb{Z}_i), \quad \mathfrak{b}^{(k)} \sim \Pr(\mathfrak{b}|S^{(k-1)}_{it}, S^{(k-1)}_{it-1}, \mathbb{Z}_i),$$

$$S^{(k)}_{it} \sim \Pr(S_i|S^{(k-1)}_{it-1}, \mathbb{Z}_i, \beta^{(k)}), \quad S^{(k)}_{it-1} \sim \Pr(S_i|S^{(k-1)}_{it}, \mathbb{Z}_i, \beta^{(k)}),$$

として、推定された確率に従って欠測値に E, U, N を割り当てる。パラメータの推定値と補完値を繰り返し得る。所定の繰り返し回数を burn-in として、収束したと考えられる段階で最終の結果とする。このような、補完データセットを複数作成し、（Rubin のルールなどで）統合する。$\mathbb{Z}$ には、労働者の性別、t 時点での年齢、配偶者の有無を用いた。また、欠測値となった理由は労働力状態に依存すると考えられることから、今期の労働力状態のロジット式の補助変数として「転入」、「追加」のインデックス変数を、前期の労働力状態のロジット式の補助変数として「転出」、「脱落」のインデックス変数を加えた。他の欠測の理由のベースは「不詳」である。ため，これがベースとな

る。本研究では、欠測理由は説明変数として重要である。なぜならば、補完のベースとなる完全データと、とくに転出転入などの移動を伴う欠測データとは、フローの傾向が異なる可能性があるためである。

1.3 IPF によるマージン調整

第 2 のストックデータとの整合性の問題については、反復比例一致法、iterative proportional fitting（IPF）法を用いる。IPF 法は、元の分布（行列）に関するマージン（行和および列和）の制約下の Kullback-Leibler 距離の最小化問題の解であるため、行・列の周辺分布が既知なときの（対数線形化のうえでの）最尤法となる。MICE による補完済み行列では、状態 B と D に関するフローは除かれているため、各月についてみた前月と今月の 15 歳以上人口は等しくなり、3×3 行列のフロー行列を形成する。このフロー行列のマージンが一致すべきストック値は公表ストック値となるが、ただし、それらから新 15 歳人口と死亡人口を除いたものでなければならない。公表ストック値には、2ヶ月連続調査以外の標本も含まれているから、全標本に関する新 15 歳と死亡のストック人口はわからない。

そのため、前月から今月にかけての公表ストック値に基づく各労働力状態人口の純増は、各々の労働力状態の「新 15 歳」人口と「死亡」人口の差に等しいことを利用して、

$$\begin{aligned} \overline{X^E}_t - \overline{X^E}_{t-1} &= \theta_t^E [x_t^{B\to E} - x_t^{E\to D}] \\ \overline{X^U}_t - \overline{X^U}_{t-1} &= \theta_t^U [x_t^{B\to U} - x_t^{U\to D}] \\ \overline{X^N}_t - \overline{X^N}_{t-1} &= \theta_t^N [x_t^{B\to N} - x_t^{N\to D}] \end{aligned} \tag{3}$$

このように定義された $\theta^S (S = E, U, N)$ を、二分の一標本からの集計値を全数に復元するための乗率として用いる。ここで、$\overline{X^S}$ が整合性を図る対象となる公表ストック値を意味する。

2 経過と結果

2.1 MICE に関する計算時間と収束の検討

まず、(2) を、月次で推定することから開始した。初めの問題は、月ごとに非線形推定を繰り返し行うため、補完系列を得るためには、日単位で多くの時間がかかることであった。そのため、数十回で設定していた burin-in の回数を 10 回まで減らし、MICE による復元データセットを 5 セットとした。通常は 20 セットほどが適切とされる。その他、分析に用いた Stata コードの改善によって、ある程度の時間短縮が可能となった。定常分布への到達の判断、補完データセット数の増加については、検討課題である。

収束に関しては、当初、最尤法の収束計算は Newton-Raphson(NR) 法によっていたが、一部の月で収束が困難な（スタック状態となる）ケースが見られた。そのため、次のようなアルゴリズム

により、収束方法や定式化の変更を収束するまで繰り返した。

[Step A]　backstep 付き BFGS 法で最大反復回数 1000

↓

[Step B]　NR 法にヘッセ行列の特異性を回避するための数値的補強を追加し最大反復回数 200

↓

[Step C]　補完式の多項ロジットモデルの説明変数から転入と転出のインデックス変数を除外し、backstep 付き BFGS 法で最大反復回数 1000

↓

[Step D]　さらに、追加と脱落のインデックス変数を除外し、
backstep 付き BFGS 法で最大反復回数 1000

1983 年から 2024 年のほとんどの月のいずれの補完データでも Step A で収束した (96.6%)。残りのうち、Step B で 1.2%、Step D で 2.2%が収束し、収束しなかったケースはなかった。*

Step D は、欠測理由に関するインデックスを全て落とさないと、収束しないことを意味する。Step D に至る月については、追加や脱落の規模が大きいなどの不連続性が発生していることが、収束を困難にする主要な理由の 1 つと推測されるが、今後の検討が必要である。

2.2　IPF の実行

このようにして作成した補完系列について、IPF 法により、ストックデータと整合的なフロー系列を作成する。MICE により補完されたフロー行列を

$$\Theta = \begin{bmatrix} \chi^{E\to E} & \chi^{U\to E} & \chi^{N\to E} \\ \chi^{E\to U} & \chi^{U\to U} & \chi^{N\to U} \\ \chi^{E\to N} & \chi^{U\to N} & \chi^{N\to N} \end{bmatrix} \tag{4}$$

とする。マージンの一致は、

$$\begin{aligned} \chi_t^{E\to E} + \chi_t^{U\to E} + \chi_t^{N\to E} &= \overline{X^E}_t - \theta_t^E x_t^{B\to E} \\ \chi_t^{E\to U} + \chi_t^{U\to U} + \chi_t^{N\to U} &= \overline{X^U}_t - \theta_t^U x_t^{B\to U} \\ \chi_t^{E\to N} + \chi_t^{U\to N} + \chi_t^{N\to N} &= \overline{X^N}_t - \theta_t^N x_t^{B\to N} \\ \chi_t^{E\to E} + \chi_t^{E\to U} + \chi_t^{E\to N} &= \overline{X^E}_{t-1} - \theta_t^E x_{t-1}^{E\to D} \\ \chi_t^{U\to E} + \chi_t^{U\to U} + \chi_t^{U\to N} &= \overline{X^U}_{t-1} - \theta_t^U x_{t-1}^{U\to D} \\ \chi_t^{N\to E} + \chi_t^{N\to U} + \chi_t^{N\to N} &= \overline{X^N}_{t-1} - \theta_t^N x_{t-1}^{N\to D} \end{aligned} \tag{5}$$

という条件である。行列 (4) はこの条件を満たしていない。しかし、対角要素が正である対角行列 $\mathbb{A}, \mathbb{B}$ を用いて、行和と列和が条件 (5) を満たす行列

$$\Omega = \mathbb{A}\Theta\mathbb{B}$$

が一意に決まる。IPF は，与えられた Θ_t とマージン（(5) の右辺）の下で，反復計算により $\mathbb{A}, \mathbb{B}$ を求め，Ω を得るアルゴリズムである。

* 5 変量ケースでは、稀ではあるが、Step D でも未収束のケースがみられた。

2.3　補完結果

就業から失業へのフローを例に、補完の状況を見たものが図 1 である。[†]5 つの補完系列の平均値および最大値と最小値が、補完前の原系列とともに示されている。補完系列のばらつきは小さく、原系列からの乖離も小さい。補完により原系列の時系列的変動の傾向はほとんど影響を受けていない。補完の程度は「良好」と判断できるが、欠測値補完が原系列とは異なったフローの性質を付け

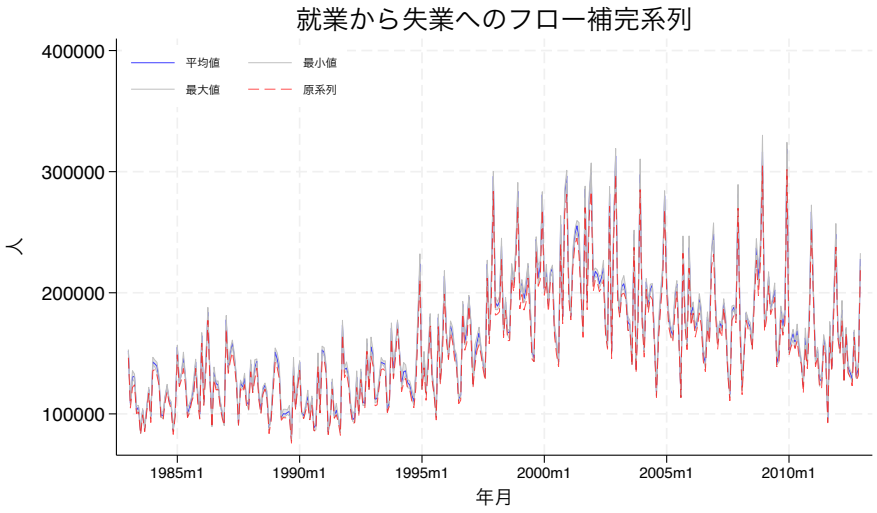

図 1

加えることはなかった。

2.4　フロー累積系列の調整過程の比較

さらに、フロー系列を累積してストック系列を算出した場合、観測されるストック系列（公表データ）とどのように乖離するのかをみる。失業ストックを例として、1983 年 1 月の公表失業者数を基準 $\overline{X_0^U}$ として、失業ストックへのネットのフローを累積していく。まず、フローとして原系列を用いた場合の失業ストックは、

$$\begin{aligned} cum\_X_1^U &= \overline{X_0^U} + (x_1^{E\to U} + x_1^{N\to U}) - (x_1^{U\to E} + x_1^{U\to N}) \\ &\vdots \\ cum\_X_t^U &= cum\_X_{t-1}^U + (x_t^{E\to U} + x_t^{N\to U}) - (x_t^{U\to E} + x_t^{U\to N}) \end{aligned} \tag{6}$$

となる。原系列に代えて、補完値の平均系列を用いた累積値 $mice\_X_t^U$ も、(6) と同様に求めることができる。両者は図 2 において「原系列の累積値」、「平均補完系列の累積値」として示されている。原系列を累積したストック系列 cum_X^U の問題は、一見して明らかなように、暫時減少していき、負値になってしまうことである。これは、原系列に沿うように補完された $mice\_X^U$ も同じである。当然、同図に示した失業の公表ストック値とは大きくずれている。[‡]これは、フロー数を求

[†]本研究のフロー推定はすべて 1983 年 1 月から 2024 年 12 月まで行ったが、事情により本報告書のグラフは 2012 年までとしている。

[‡]ここでの、フロー系列は、全体の約半数の 2 期間連続調査標本に基づき集計しているので、公表値と同じ基準で見ると乖離はさらに広がる。

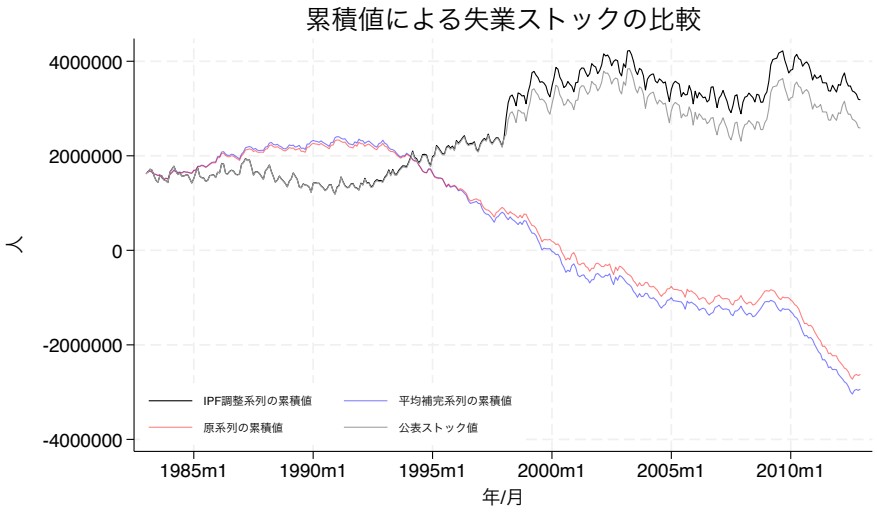

図 2

める標本について、失業からのネットフロー（ネットで流出）が、長期的にみて実際よりもかなり過大であることを示している。図には示さないが、同様の現象は就業ストックからのネットフロー（流出）についてもみられる。その一方で、非労働力ストックへのネットフロー（流入）は過大となっており、これらの偏りが公表値とはかけ離れた累積値の原因となっている。

図 2 には、IPF によって調整されたネットフロー系列によって、(6) と同様に累積値 ipf_X^U を求めた IPF 調整系列も示している。この IPF 調整系列と公表系列との差が (3) で定義された新 15 歳と死亡によるネットの人口変動による失業ストック変化分推定値であり、これを加えると定義的に IPF 調整系列は公表ストック系列と一致する。

2.5　結果の評価と課題

以上より判明したことは、まず、2 期連続調査標本によるネットのフロー系列は、全標本に基づく公表ストックに基づくネット変動から、一定方向への偏り（就業、失業からのネットの流出と非労働力へのネットの流入が過大）があることである。その原因は、2 期連続標本に発生している欠測状態が、観測状態とは性質の異なったフロー（つまり状態間移動）の性質を有することであると考えられる。そのため、上述の通り、MICE による欠測値補完に、欠測理由のインデックス変数を説明変数として加えることで、この性質を捉えた欠測値の推定を行った補完を期待した。

しかし、図 2 に見る通り、ストック系列と整合的なフロー系列の調整過程においては、欠測値補完の役割は小さく、ほとんどがマージン調整によっている。この理由は、先にみたとおり、MICE による補完系列がほぼ原フロー系列と並行して推移するためである。これは、欠測値補完が「良好」であることを意味するものの、欠測補完によってネットフローの偏りを是正しようとする目的には適ってない。今後、欠測理由のインデックス変数の補完推計式 (2) への導入の定式化についての検討が必要である。

2.6　ストック量の要因分解

以上のように、ストック公表値と整合性を持った労働力フローが構築できると、各年月の労働力状態ストック人口が、どのような状態間のフローの累積によって時系列的に変化したかを示すことができる。たとえば、以上で得られた IPF 調整フロー系列を用いて、1983 年以降の失業人口が、過去の就業、失業、非労働力間の異動の累積結果として変動する様子を示す。図 3 には、就業と失

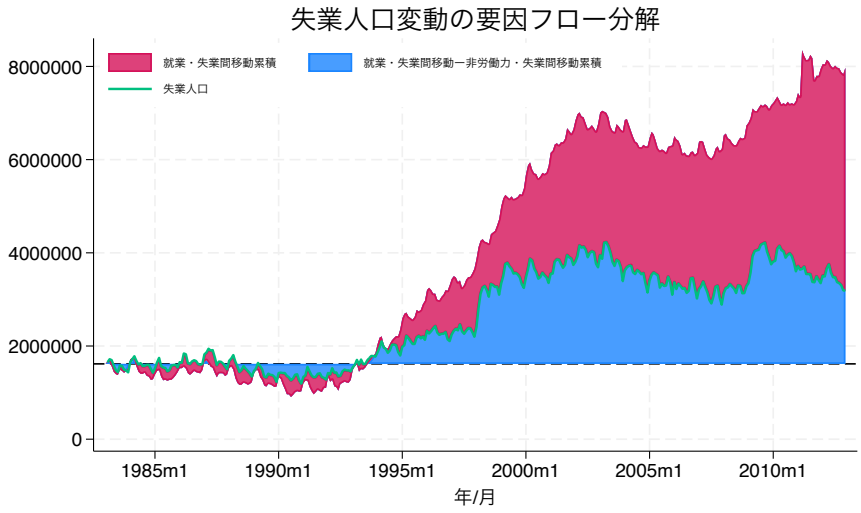

図 3

業間の IPF 調整ネットフロー量の累積値、

$$ipf\_X_t^{E\leftrightarrow U} = X_0^U + \sum_{s=1}^{t}(x_s^{E\to U} + x_s^{U\to E})$$

および、$ipf\_X_t^{E\leftrightarrow U}$ に就業と非労働力間の IPF 調整ネットフロー量の累積値を加えた

$$ipf\_X_t^{E\leftrightarrow U} + ipf\_X_t^{N\leftrightarrow U} = ipf\_X_t^{E\leftrightarrow U} + X_0^U + \sum_{s=1}^{t}(x_s^{N\to U} + x_s^{U\to N})$$

が示されている。前者が赤色の面、後者が青色の面で、後者は負値で累積していくため、就業から失業へのネットの流入フローが失業ストックを増加させる一方で、失業から非労働力へのネットの流失が失業ストックを減少させている様子がわかる。５変量の場合に同様な分析を行えば、非正規雇用人口数の増加が、どのような状態からの流入や流出に依存して変動しているかを量的に分解してみることが可能となる。

2.7　推移確率と限界効果

労働力フロー分析では、通常、各月のフローとストックの比率を状態間遷移確率と捉え、フロー行列をマルコフ遷移行列と解釈する。たとえば、今月に失業する確率は、前月が就業状態であった場合には $(x_t^{E\to U}/X_{t-1}^E)$ であり、前月が非労働力状態であった場合には $(x_t^{N\to U}/X_{t-1}^N)$ である。図 4 には、これらの遷移確率とともに、就業から失業する確率が、非労働力から失業する確率に比し

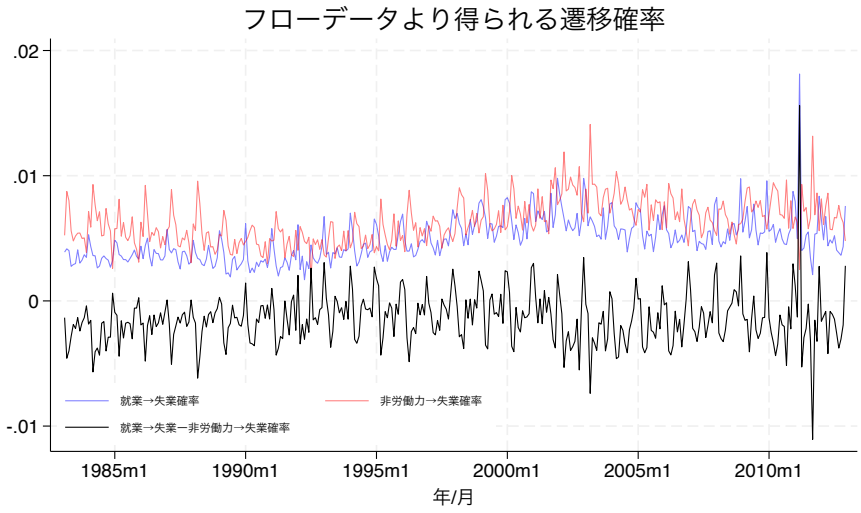

図 4

てどれほど高いかを $(x_t^{E\to U}/X_{t-1}^E)-(x_t^{N\to U}/X_{t-1}^N)$ によって示している。多くの年月で前者は後者を下回っているため、この差は負である。しかし、これは平均的に見た遷移確率であるため、すべての人口について遷移確率が等しいという前提がある。実際には、性別、年齢、地域などで遷移確率は異なると考えられる。これを考慮する場合には、以上と同様にして、人口属性別に観測ストックにマージンが等しい調整遷移確率行列を求めなければならない。たとえば、男女別に遷移確率を推定することで性差は考慮できるが、性別ごとに異なる年齢階層が含まれている。年齢階層別に遷移確率を推定すれば、各階層に男女が含まれる。そのため、性差と年齢差を考慮するためには、それらの積の数の「性 $\times$ 年齢階層」区分が必要になる。さらに地域を考慮すれば、より細密な区分が必要となり、実行できても結果の比較は複雑となり、詳細区分内の標本サイズの減少による推定効率の低下が懸念される。さらに、このようにして求めた遷移確率は平均確率であるが、実際の経済主体が直面するのは、現実の変数分布の下で生じる限界確率であるという問題もある。

これらの点は、調査票情報を用いたミクロデータ推定によって考慮可能である。欠測値補完によって得られたデータ基づいて、補完の誤差を考慮して標準誤差を修正した多項ロジット推定を行うことで、性別、年齢階層別、地域別などのインデックス変数を説明変数とした場合の限界効果（インデックス変数の差分効果）とそれらの統計的有意性を示すことが可能である。図 5 には、遷移確率による就業から失業する確率と非労働力から失業する確率の差（図 4 に同じ）と、それに対応るる概念である、非労働力をベースとした就業から失業への平均限界効果を示している。限界確率の変動は平均遷移確率の変動よりも大きく、また、その値はほぼ正の領域を推移するなど異なった性質を示すことがわかる。同様な分析を 5 変量のケースで行うことで、非正規雇用への労働移動の源泉についての数量的分析が可能となる。

2.8 今後の課題

今回の研究では、年・月単位のデータによる欠測値補完のプログラム実行に際し、非常に多くの非線形推定と代入計算を繰り返す必要性から、多大な時間がかかることが判明した。その短縮化が研究開始時における課題となり、試行錯誤のために多くの時間が必要であった。また、最尤法が収

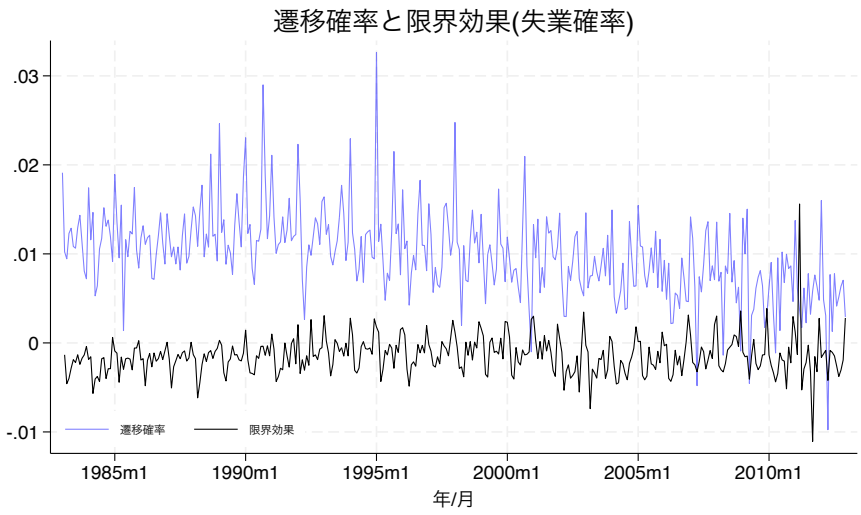

図 5

束しないケースについて、原因の検討や収束計算法の変更などにも時間を要し、以上に報告した通り、研究の方向性についてはほぼ確立できたものの、最終的な成果を得るには時間が不足した。計算時間と収束の問題については、おおよその解決を見られたので、次回に活かしたい。また、本研究の過程で、2 期間連続調査標本の集計に際して、公表フローデータと同じ集計結果を再現するには、「1/2 集計用乗率」を用いる必要があることがわかったため、総務省統計局より提供を受けて必要な集計に用いたい。なお、とくに５変量の場合について、フローデータの集計に際して、予想以上に当該標本数が少ないケースが多く見られ、持ち出しが不可能な推定系列が多く生じた。今後は、その最終的な持ち出し方法について、補完系列の集計化、区分統合や代替値の利用などの検討が必要である。