博士課程進学率に関する因果モデルの構築
統計的因果探索アルゴリズム “LiNGAM” による試行的分析

高山 正行 [A*]，小柴 等 [A]，前田 高志 ニコラス [A,B,C]，
三内 顕義 [A,B,D]，清水 昌平 [A,B,E]，星野 利彦 [A]

Causal Model of the Rate of Doctoral Course Enrollment in Japan:
Trial Analysis by Statistical Causal Discovery Algorithm “LiNGAM”

Masayuki TAKAYAMA, Hitoshi KOSHIBA, Takashi Nicholas MAEDA,
Akiyoshi SANNAI, Shohei SHIMIZU and Toshihiko HOSHINO

投稿区分: 研究論文

A 文部科学省 科学技術・学術政策研究所
B 国立研究開発法人 理化学研究所 革新知能統合研究センター
C 国立大学法人 東京大学
D 国立研究開発法人 科学技術振興機構 「さきがけ」研究員
E 国立大学法人 滋賀大学

* 文部科学省 科学技術・学術政策研究所 第 1 調査研究グループ
100-0013 東京都千代田区霞ヶ関 3 丁目 2-2
E-mail: masayuki-takayama@mext.go.jp
corresponding author(代表著者, 問合せ先著者)

Abstruct

Causal Model of the Rate of Doctoral Course Enrollment in Japan:

Trial Analysis by Statistical Causal Discovery Algorithm “LiNGAM”

Keywords: STI policy, statistical causal inference, statistical causal discovery, LiNGAM, HR policy in STI, EBPM

In the context of strengthening the research capacity and supporting young researchers in Japan, increasing Ph.D. students is one of the most important political goals. Although many kinds of political variables such as financial support on Ph.D. students, research environment, and career paths after they graduate, are related to this goal, it is still an open question how the government should control these factors quantitatively, due to the lack of the causal graph in this problem.

In this paper, we estimate quantitatively the causal effects on the rate of doctoral course enrollment by the variables from the open statistical data using “LiNGAM”, a statistical causal discovery method. We also discuss the interpretation of the results with the domain knowledge of science, technology, and innovation(STI) policy, and compare this data-driven approach with the traditional statistical causal inference method in the domain of STI policy.

要旨

博士課程進学率に関する因果モデルの構築
統計的因果探索アルゴリズム “LiNGAM” による試行的分析

キーワード：科学技術政策，統計的因果推論，統計的因果探索，LiNGAM，博士人材政策，EBPM

近年の科学技術・イノベーション政策の重要課題である “研究力強化・若手研究者支援” の文脈において博士課程進学率の向上が目標の一つとして掲げられ，様々な政策的手段の議論やニーズの調査もなされてきた。しかしながら，実際に博士課程進学率の向上に寄与したかどうかの検証、そして今後博士課程進学率の向上に真に有効な政策的要因とその資源配分の適切なポートフォリオの把握のためには，各要因がどのように関与しているかという因果関係の有無も含め，統計データに基づいた定量的な分析・理解が重要である。

そこで本稿では，特に NISTEP による修士課程修了者向けのアンケート調査結果 [加藤 09, 治部 21a] を念頭に，公開されている統計データを対象として統計的因果探索アルゴリズム “LiNGAM” を適用し，博士課程進学率に関する因果グラフをデータ駆動的に推定した。

その結果，統計的因果探索によるデータ分析の結果も NISTEP アンケート調査と整合することが確認できた。加えて，例えば「大学研究本務者一人あたりの基盤的経費」が「研究時間割合」に正の影響を与えている可能性が統計的に高く示されるなど，従来の研究力強化の文脈では議論されなかった因果関係が示唆された。これらは，共分散構造分析によって領域知識のみに基づき考案した因果グラフと比較しても，遜色ないモデル適合度を示した。

本稿の結果を一つの示唆として，今後遅延時間を考慮した分析や分野別の議論，未観測共通要因の特定等の発展的な分析を通じて因果グラフ・因果モデルが定量的に決定されることで，EBPM(Evidence Based Policy Making) の観点から若手研究者支援政策への貢献が期待される。

本文

1　はじめに

近年の我が国の科学技術・イノベーション政策（STI 政策）の代表的課題として，研究力強化が挙げられる。この政策課題においては，様々な指標に基づき，我が国の研究力の現状と解決されるべき問題が議論されており，中でも若手研究者支援はこの文脈における中核的な課題である。たとえば 2020 年 1 月には「研究力強化・若手研究者支援総合パッケージ」が決定されており，また 2021 年度からの第 6 期科学技術・イノベーション基本計画でも，若手研究者支援に関する目標設定がなされている。

特に「研究力強化・若手研究者支援総合パッケージ」では，研究力強化を狙いとして，若手研究者を中心に修士課程（または博士前期課程）学生から中堅・シニア研究者までの幅広い・切れ目のない支援を資金面や環境面といった複数の側面から実行することとし，その目標設定や指標の年限は第 6 期科学技術・イノベーション基本計画の最終年度である 2025 年度に定めている。

一方これらの目標には，現状では実現可能性が高いとは言い難いものもある。たとえば博士（後期）課程[1)]進学率については V 字回復が目標とされているところ，これまで年々減少傾向が続いている。また，「大学本務教員の 40 歳未満割合を 2025 年度までに 3 割以上」という目標についてもこれまで単調減少となっており，さらに確率遷移モデルによる推計・シミュレーション [高山 21a] でも，達成にあたっては抜本的な政策改革が必要であるとしている。これらの目標達成に向けて予算・制度の集中的・抜本的な改革が期待されるところであるが，2021 年度からの 5 年間という短い期間と我が国の資源が限られていることを踏まえると，それら施策の効果を定性・定量の両側面から適切に見極める必要がある。また，目標設定の妥当性の評価は次期の科学技術・イノベーション基本計画を策定する上でも重要である。そのためには，改めて各種政策要素間の因果関係について，有識者の経験的知見のみにとどまらず，実際の統計データに基づく定量的な理解を深め，より精度の高い予測を目指していくことが重要である。

そこで本稿では若手研究者支援政策における各種政策要素の間の因果関係を統計的側面から究明すること目標とし，その第一歩として，まずは試行的に指標の一つでもある我が国の博士課程進学率について，統計的因果探索の手法 LiNGAM (Linear Non-Gaussian Acyclic Model) を用いた因果グラフの推定を行った。

本稿では，まずその計算方法について述べ，得られた結果について若手研究者支援に関する政策研究の領域知識に基づく定性的な解釈と，従来の統計的因果推論 (特に共分散構造分析) を用いた比較・評価について報告する。

[1)] 本稿では，単純に「博士課程」という用語で統一することとする。

2　背景〜博士課程進学率に関する政策的議論と因果推論

前述の通り博士課程進学については我が国の STI 政策においても重要な要素として位置づけられている。もう少し具体的に述べると，たとえば第 6 期科学技術・イノベーション基本計画においては，「知のフロンティアを開拓する多様で卓越した研究成果を生み出すため、研究者が、一人ひとりに内在する多様性に富む問題意識に基づき、その能力をいかんなく発揮し、課題解決へのあくなき挑戦を続けられる環境の実現」のために，「まず優秀な若者が、将来の活躍の展望を描ける状況の下で、『知』の担い手として、博士後期課程に進学するというキャリアパスを充実させる」と言及されており，博士課程進学を STI 政策上の「知のフロンティアの開拓」という側面における本質的な問題として論じている。また，この施策の方向性として，「博士後期課程学生の処遇の向上とキャリアパスの拡大」，「大学等において若手研究者が活躍できる環境の整備」について，具体的な記述とともに盛り込まれている。さらに，これらに先だって調査・分析結果もいくつか報告されている。

本節ではまず博士課程進学に関する先行研究を簡単にレビューし，明らかになっていない定量的な因果関係と，本稿で採用する統計的因果探索のアプローチを統計的因果推論の一般論のレビューとともに説明する。その後，それらに基づいて本稿の手法（データセットの構成方法と計算条件）について説明する。

2.1　博士課程進学に関するこれまでの調査研究

博士課程進学率に関する調査・分析は，これまでも進学率の定量的な要因分析，修士課程学生向けのアンケート調査，仮説に基づいたシナリオ構築と政策の具体的方向性の深堀といった様々なアプローチで行われてきた。ここではそれらをレビューするとともに，今後，政策検討に直接的に繋げていく上で必要な観点について述べる。

■博士課程進学率に関する定量分析　博士課程進学率の増加・減少について定量的な議論を試みた研究の例としては，浦田らによる研究 [浦田 05] が挙げられる。ここでは，1980 年から 2004 年までの範囲で，博士課程進学率を従属変数に，そして入学定員比率，家計所得，博士卒無業者，修士卒無業者数を説明変数にとることで重回帰分析を試みており，入学定員比率と正の相関，博士卒無業者率と負の相関が有意にみられたとしている。また当該研究は男女別の分析とも比較して行っていることが特徴であり，家計所得との正の相関が女子の場合のみ有意であること，一方で男子では修士卒無業者率が有意であることもわかっている。

この取組は，同様の手法で 1977 年から 1999 年までを分析した浦田による研究 [浦田 01] とともに, 博士課程進学率の増減を重回帰分析で定量的に論じた重要な研究の一つといえる。一方，本稿の時点から浦田らによる研究 [浦田 01, 浦田 05] を参照するにあたっては，時代背景・政策研究のリソースの進展度合いを踏まえ，以下の点に注意する必要がある。

- この研究当時において進学のための経済的支援はまだ盛んには行われていなかったという時代背景もあり，経済的支援の影響は考慮されていないこと
- 直接介入が可能でない変数によって要因分析がなされていること
- 統計的因果推論の枠組みの確立にあたって長らく様々な議論があり[2)]，さらにはそのためのソフトウェアも汎用的なものが開発されたのが近年であって，当時は政策研究への応用に至ることが困難であったと考えられること

これらの点を踏まえると，改めて現在のトレンドを踏まえて候補となる政策要因を再検討し，また近年，理論・ツールの構築がともに進んできた統計的因果推論の枠組みを用いて，介入可能と考えられる変数を含んだ因果推論に取り組むことが，今まで以上に有意義な政策提言を行う上では益々重要である。

■修士課程修了者に対するアンケート調査　大学院学生の実態・意識調査等を念頭に，科学技術・学術政策研究所 (NISTEP) において過去 2 回ほど修士課程修了者を対象とした大規模なアンケート調査が実施されている [加藤 09, 治部 21a]。当該調査の質問項目には博士課程進学の検討に関するものも設定されており，本研究の目的に合致する。2009 年の加藤らの調査報告 [加藤 09] によれば博士課程進学者・博士課程進学を真剣に検討したことのある就職者（つまり博士課程に進学しなかった者）が進学を検討する際に重要と考えられる要件として，経済的支援，民間企業等での博士課程修了者の雇用の増加と処遇の改善，アカデミックポストへの就職機会の拡充や任期等の待遇改善，研究環境の充実が上位となった。また，近年 治部らによって行われた調査報告 [治部 21a] でも類似した傾向の報告があることから，当事者（修士課程学生）にとってのこれらの要件の重要性自体は 10 年以上の時を経ても定性的には変化していないことがわかる。さらに治部らはこの調査結果をもとに「博士離れ」の要因として，継続性を保証されない外部資金による不安定な有期雇用の増加，研究者市場及び我が国の労働市場における低い流動性，が示唆される，としている [治部 21b]。

これらは博士課程進学の要因をホットトピックに基づき，また修士課程学生の当事者意識へのミクロスコピックなアプローチで定性的に明らかにしているが，一方それらの要因によるマクロスコピックかつ定量的な因果関係の議論には至っていない。

■仮説に基づいたシナリオ構築とアンケート・ヒアリングによる深堀　野村総合研究所による調査研究 [野村総研 10] ではアンケート調査に基づき，博士課程進学を促進するにあたって，特に就職に関する学生の進学前の不安払拭を (1) 博士課程における人材育成，(2) 博士課程学生の就職活動，(3) 企業による博士課程学生の採用，の 3 段階のシナリオから構成し，ヒアリング調査をもとに各段階での課題解決の方向性を議論している。

この研究はノンアカデミックなキャリアパスのみに論点を絞ることで，また海外の好事例との比

[2)] 例えば，本稿では詳しく追究しないが，潜在的結果変数の枠組みに基づく Rubin 流のアプローチと，構造的因果モデルでに基づく Pearl 流のアプローチは双璧をなし，時に対立軸で捉えられることもありつつ，相補的なものである [高橋 22]。

較から提言に繋げており，新たな可能性提案という意味では重要なものである。その一方で，ここではあえて議論の対象外としているものの，たとえば経済的支援といった他の要因との一体的なロジック構築には至っていない。また，ヒアリングや事例調査に基づいた事実関係の整理がロジックモデルに紐づく形でなされているものの，実際の因果関係そのものの検証には至っていない。

■本課題に関する統計的因果推論の重要性　ここまでのレビューから伺えるとおり，博士課程進学率に関する政策課題において，各種要因間の関係性は必ずしも明らかではなく，特に定量的な評価の観点では，介入可能な要因から博士課程進学率に至るまでの様々な因果関係・影響評価が未だ不足している。特に一般的な因果関係の評価は，定石的には要因を絞りいくつものステップを踏んで，重要な要因の漏れがないか，また未知の因果関係の繋がりがないか，精緻に検証することが必要である。

しかし，このような政策課題では議論の俎上に上がる要素が非常に多く，複雑な因果関係の存在も考えられるため，従来の因果推論アプローチでは定量的な影響評価まで行うことは困難であると予想される。

そこで我々は，近年発達してきた統計的因果探索のアプローチにより，データ駆動的に博士課程進学率と各種関連要因の間の因果関係の候補を導き出すとともに，定量的な影響評価に繋げていくというアプローチを採用した。

2.2　統計的因果推論

政策立案等において，関連する要素とそれらの間の関係性（因果関係）を見極めるにあたっては，通常はランダム化比較実験 (RCT) によるアプローチが有効である。他方，その実験にかかる費用・時間の問題，場合によっては倫理的な問題もあり，実現困難な場合も多い。そのような場合にも因果関係の同定を可能とするために，統計的因果推論のフレームワークがこれまで整えられてきた。

■因果推論と因果探索　因果推論，特に統計的因果推論とはデータから因果効果を統計的に推定する手法である。多くの場合は，目的変数とそれに関連する要素を集め，背景知識等を元に因果関係を仮説として設定し，データからその妥当性を定量的に検証する形で因果効果を推定することが多い。代表的な手法としては，たとえば，回帰分析や操作変数法，差分の差分法，共分散構造分析などが挙げられる。特に操作変数法は，その活用による功績が 2021 年のノーベル経済学賞を受けたことで再度注目を浴びている。

STI 政策の文脈において本稿と同じく統計的因果推論を用い，データからのモデル構築を行っているものについてこれらの手法を念頭に具体例をいくつか挙げると，たとえば，回帰分析により分野融合と研究のインパクトの関係を評価したもの [Okamura19]，共分散構造分析により共同研究・受託研究活動の分析をしたもの [中山 10]，操作変数法により科学技術の状況に関する質問項目間の関係性について定量分析を試みたもの [福澤 15]，差分の差分法で環境規制と経済的効果について定量分析を試みたもの [枝村 16]，などが挙げられる。

しかし，これらの手法にも限界はある。たとえば，共分散構造分析ではあらかじめ変数間の因果

関係（因果パス）を仮定するような手続きが必要だったり，その他の手法も 1. 多くの変数が扱えない，2. 特定の変数に関する比較的単純な因果関係しか扱えない，3. 適用に当たって強い仮定をおく必要がある，などの制約が存在する。あらかじめ，何らかのモデル，そしてその前提としての変数間の因果関係が想定されている場合などでは，これらの手法は強力かつ有効だが，具体の関係性が見えていない状況で，関係性自体を探索したい場合には使いづらい。したがって現在想定する，要因の抜けや隠れた（大きな）関係性まで議論するには適していない。

さて，この因果推論の中でも，因果関係自体をデータから自動的に見つけ出す（推測する）手法として因果探索が存在する。因果探索はこれまで述べてきた何らかの背景知識・仮定を前提とするその他の手法と比較して背景知識の必要性が低く，より多くの要素間の複雑な関係性も探索できるメリットがある。統計的因果探索の代表的な手法には（因果）ベイジアンネットワーク（Bayesian Network，BN），LiNGAM が挙げられる。これらの内容・関係性について図 1 にまとめた。

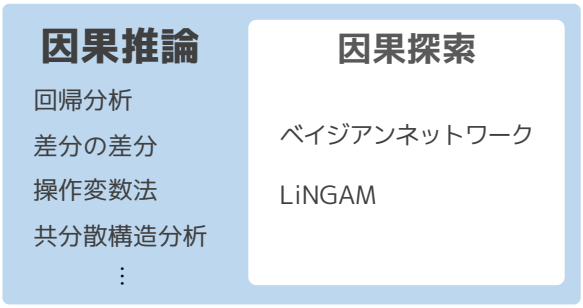

図 1　因果推論と因果探索の関係

■統計的因果探索の性質　統計的因果探索というアプローチにも様々なバリエーションが存在するが，ここでは代表的な手法である BN，および LiNGAM に議論を絞って説明する。

BN[Pearl 85] はベイズの定理に基づいて因果関係を確率的に評価する比較的古典的な手法で，これを用いた EBPM の取組としては，たとえば地域健康政策の支援における活用 [鳥海 18] が挙げられる。ただし BN は 2 値変数のように変数が離散的で取りうる値が少ない場合やカテゴリーによる分岐を考える際は強力だが，連続変数が扱い難い。そのため，アンケートなどで取得したカテゴリー変数などには適用できても，予算や人数のような間隔尺度以上のデータには適用しづらいという課題もある。

こうした連続変数としてみなされる変数についての因果関係を探索する手法として，独立成分分析によるアプローチを出発点として LiNGAM[Shimizu06] という手法が開発され，因果探索の可能性も広がりを見せている。

ここまでの説明をもとに BN と LiNGAM の違いについて，表 1 に簡単にまとめた。いずれの因果探索の手法を使うにあたっても，各手法の仮定[3)]を満たすかどうか等，妥当性の考察は必須であ

[3)] たとえば，BN も LiNGAM も基本的には有向非巡回性（各変数間の関係がある場合は全て向きを持ち，因果関係が巡回しない）が仮定される。

表 1　BN と LiNGAM

手法	変数の種別	アプローチ
BN	主に離散	ベイズ
LiNGAM	主に連続	独立成分分析

※BN: Bayesian Network

るが，扱う要因・変数の性質に応じて適切な因果探索手法を選択し，よいモデルが構築できるならば EBPM を推進する上でも大きな補助として期待できる。

後述するとおり，本稿の研究対象である若手研究者支援政策に関する各種要因の多くは連続的な数とみなす方が扱いやすいことから，因果探索手法としては LiNGAM ベースのアプローチが妥当であり，これを採用する。

■LiNGAM における仮定とアルゴリズムの性質　LiNGAM においては，以下の構造方程式モデル Structure Equation Model, SEM) に基づき係数を定めることで，因果関係を決定していく。

$$x_i = \sum_{i \neq j} b_{ij} x_j + e_i \tag{1}$$

x_i（ただし，$i = 1, 2, ..., n$）は因果モデルで考慮する n 個の変数，b_{ij} は x_i に対し x_j が及ぼす直接的な影響の大きさに対応する係数行列 $B = [\boldsymbol{b}_{ij}]_{n \times n}$ の成分，そして e_i は変数 x_i の誤差変数（外生変数）である。オリジナルの LiNGAM では，各変数同士の関係が全て線形であること（この仮定は既に式 (1) を立てる段階で加味されている），各変数の誤差項の分布の非ガウス性（正規分布とならないこと），異なる変数の誤差項同士の独立性[4]，そして因果グラフの非巡回性を仮定する。これらの仮定の下，x_1〜x_n までの各変数についてのデータセットから係数行列 B を計算し，因果グラフを一意に識別することが可能となる[5]。

推定アルゴリズムとしては，まず独立成分分析によるアプローチ [Shimizu06] が提案された。その後，回帰分析と誤差項の独立性評価を繰り返し，誤差項同士の相関の大きさを最小化するように因果的順序と係数行列を決定する “DirectLiNGAM” という手法が提案された [Shimizu11]。また，

- 時系列データで時間差を伴う効果がある場合の，ベクトル自己回帰モデルを用いた分析手法 [Hyvarinen10]
- ブートストラップ法による再標本化と DirectLiNGAM の実行を繰り返すことによる，因果関係の有無・係数値の統計的信頼性の評価法 [Komatsu10, Thamvitayakul12]
- 未観測共通要因がある場合でも因果グラフの全体像を推定する手法 [Maeda20]

等にも派生するなど，LiNGAM はその仮定を緩めつつ適用可能範囲を拡大している。これらの派

[4] つまり未観測共通要因が存在せず，数学的には $\mathrm{cov}\left(f(e_i), g(e_j)\right) = 0$（ただし，$f, g$ は任意の有界な関数）が $i \neq j$ で成り立つこと。

[5] この数学的なフレームワーク・証明については，清水らによる教科書 [清水 17] を参照されたい。

生形も含め，Python の LiNGAM パッケージ[6)]は GitHub 上で公開されており，データセットを正しく構築しさえすれば，tutorial に沿って jupyter lab 等で所定のコードを入力して実行するだけで，結果が出力される[7)]。

3 研究手法

本稿では，GitHub 上の LiNGAM パッケージのうち DirectLiNGAM[Shimizu11, Hyvarinen13] を用いて因果探索を行う。ただしその前提として，1. 博士課程進学率に関する政策課題の中から政策要因を変数として選定すること，2. さらにそれらの性質に基づいて計算条件を設定すること，が必要となる。

3.1 変数の選定とデータセットの構成

2.1 節 でもレビューした通り，博士課程進学率に関連する政策指標・手段には様々なものが考えられる。しかしながら，統計的因果探索の手法をは本政策課題において適用するにあたり，それらすべてを盛り込んで因果探索を行うことは現実的ではない[8)]。そこで本稿では，特に博士課程進学率に関する議論の最新性という観点から，NISTEP の調査結果 [加藤 09, 治部 21a] において，博士課程への進学検討における重要な要素として上位となっていた，「経済的支援」，「民間企業等での博士課程修了者の雇用の増加と処遇の改善」，「アカデミックポストへの就職機会の拡充や任期等の待遇改善」，「研究環境の充実」にフォーカスした。さらにこの中で，定量化や統計の取得自体が難しいもの[9)]は今回の分析では除外し，統計調査で定量化がある程度なされている項目のみ，公開されている統計データから対応する変数を選定した。具体的には (A) 経済的支援，(B) アカデミアへのキャリアパス，(C) 研究環境の 3 つの観点から変数を選んだ。

(A) 経済的支援に関する変数

国による経済的支援としては，博士課程進学前に受給が決まるものとして日本学術振興会の特別研究員 DC1 がまず挙げられ，これに採択されると月額 20 万円の給与と一定額程度の研究費が支給される（2021 年時点）。他にもグローバル COE やリーディング大学院，卓越大学院といった文部科学省の事業が挙げられるが，プログラムによっても，支給金額をはじめとする内容の差異が存在

6) https://github.com/cdt15/lingam

7) ただし高山ら [高山 21b] にレビューされている通り，この手法を誤用せずに，因果推論に向けた効果的な仮説生成に繋げるには，手法の仮定に関する正しい理解と，使用するにあたっての仮定の妥当性の考察，変数の性質に関する正しい理解とデータセットの適切な構築が必要となる。

8) 特に，変数の数がデータ点数を超えてしまう場合は注意が必要である。まず，少なくとも独立成分分析のアルゴリズムでは計算ができない（これは，連立方程式を解くにあたって変数の数だけ方程式が立っていないと解が一意に定まらないのと同様の理屈である）。DirectLiNGAM では，そのアルゴリズムの性質上計算そのものは実行されるが，結果の精度は大きく低下する恐れがある。

9) たとえば，「民間企業等での博士課程修了者の雇用の増加と処遇の改善」は，民間企業等での博士課程の雇用について，業界の多様性を考えると，学校基本調査の統計情報のみで議論することは現状困難であると考えられる。また，処遇改善についても，数値化にあたっては様々な議論がありうる上，たとえば博士課程修了者のみの所得の情報を毎年度分抽出することも現状困難である。

する。このように，経済的支援の性質も様々であることから本来は別々に考慮すべきだが，データ点数の限界を踏まえ変数を絞る必要があるため，本稿では以下の通り，

- 博士進学の前年度の DC1 採択者数
- グローバル COE プログラム・博士課程教育リーディングプログラム・卓越大学院プログラムの年度ごとの予算額合計[10)]

と大くくり化した上で変数として採用した。なお，「研究力強化・若手研究者支援総合パッケージ」や第 6 期科学技術イノベーション基本計画において「生活費相当程度の経済的支援の受給者数」が政策目標となっており，国による経済的支援プログラムはその月額が DC1 に準ずる，あるいはそれ以下の金額となっている例がほとんどである。他国との比較を踏まえると，経済的支援の充実に関する議論は受給者数だけでなく一人あたりの金額でも見るべきという議論もありうるところであり，将来的にこれを変数とすることもありうるが，本稿ではそこまでは追究しない。

(B) アカデミアへのキャリアパスに関する変数

キャリアパス全体では高山らのシミュレーション [高山 21a] のように，アカデミアの中でもフェーズに応じたポストの変化，および民間企業-アカデミア間の転職等の各人材流動も考慮する必要がある。しかしここでは，議論を簡略化し，特に修士課程学生にとって，“博士課程進学の検討においては博士課程修了後の最初のキャリアパスが主要な検討材料になる” と仮定し，学校基本調査の結果に基づいて，

- 博士課程修了直後の大学教員としての就職割合
- 博士課程修了直後のポストドクターとしての就職割合

を計算し，変数とした。ただし，学校基本調査におけるポスドク就職者数は 2011 年度以前については調査・公開されておらず，この欠損値は 2012 年度以降の平均値で補完した。

(C) 研究環境に関する変数

研究環境については，実験室の設備の共用等も含めて複数の観点があり，定量化が困難な部分もあるが，まずは NISTEP が定点調査として実施している項目[11)] を参考に，

- 大学における研究本務者一人当たりの基盤的経費[12)]
- 研究時間割合

---

10) 予算額には，大学院生に経済的支援として支給される金額以外にも事務費を含むと考えられるが，本稿ではこのことについても一旦棚上げしておく。

11) https://doi.org/10.15108/nr189

12) NISTEP 定点調査において研究環境の要素の一つとして問うているのは，基盤的経費の満足度であり，これに直接対応するのは一見，基盤的経費全体であるように解釈できる。一方このアンケート結果においては満足度は年々単調減少であるのに対し，基盤的経費全体はここ数年ずっとほぼ横ばいになっている。このアンケート項目により対応させるため，アンケート回答者目線から見た基盤的経費は全体の値ではなく配分の末受け取る金額に関するものであると考えられること，また大学研究本務者一人当たりであれば年々概ね単調減少であり定点調査の満足度とも相関していると考えられることから，一人当たりの経費を変数として採用することとした。

表 2　本稿でデータセットの構成にあたり採用した変数と，それぞれに課した事前知識

変数（単位）	変数名	事前知識
博士課程進学率	x_0	
前年度 DC1 採択者数（人）	x_1	外生変数
国全体の 基盤的経費（億円）	x_2	外生変数
大学研究 本務者数（人）	x_3	
一人当たりの 基盤的経費（億円）	x_4	$x_2 \div x_3$
研究時間割合	x_5	
博士修了直後の 大学教員就職割合	x_6	
博士修了直後の ポスドク就職割合	x_7	
DC1 以外の 経済的支援（億円）	x_8	外生変数

を採用した。大学の基盤的経費は文部科学省にて多用される議論に準じ，（国立大学法人運営費交付金等予算額 + 私立大学等経常費補助金）として計算し，この国全体での合計額と，総務省の科学技術研究調査に基づく大学の研究本務者数も変数に加え，一人当たりの値として算出した[13)]。また，研究時間割合については “大学等におけるフルタイム換算データに関する調査（FTE 調査）” で公表されているうち，大学等教員の値を用いた。ただし FTE 調査は現状 5 年ごとの調査のため，各年度の値は線形補完で埋めて処理した。

以上を踏まえ，表 2 には，データセットの構成にあたり採用した変数をまとめている。データの点数は，2007 年度〜2019 年度の年度単位で 13 点である。

3.2　計算に関する諸条件

本稿で用いている DirectLiNGAM のアルゴリズムでは，定義式や各学問領域の知見などに基づいた因果関係についての事前知識 (prior knowledge) を拘束条件とした計算も可能となっている。

[13)] なお，この基盤的経費はそのまま全て大学における研究に充てられるわけではなく，教育に関する費用や各種人件費等を含む。また学部・分野によっても配分額は異なる可能性がある。厳密に研究環境と対応させるにあたっては，これらの事実に注意する必要があるが，本稿ではそこまでは追究しない。

本稿でもこの手法を採用し，各変数に関する事前知識も，表 2 に併記している。x_1（前年度 DC1 採択者数）・x_2（国全体の基盤的経費）・x_8（DC1 以外の経済的支援）については，政府の予算額として定まるため，政府の意思決定により定まる外生変数であるとし，他の変数からは何も影響を受けないものとした。また，x_4（一人当たりの基盤的経費）については，その定義式を踏まえ，x_2（国全体の基盤的経費）と x_3（大学研究本務者数）に及ぼす影響はないものし，その他の変数からの影響がないものとした。

なお通常の LiNGAM では，和の形での構造的因果モデルを仮定するため，変数間の関係は線形で表現される（式 (1) を参照）が，今回は x_2〜x_4 のように積の関係が定義として入っており，既に線形結合で表現できないモデルとなっている。そこで本稿では，以下のような積の構造的因果モデルを導入し，LiNGAM を適用した。

$$x_i = \prod_{i \neq j} x_j^{b_{ij}} e^{e_i} \tag{2}$$

多変量解析においてもしばしば，変数間の弾力性（変化率の比が一定）が期待される場合に積のモデルを導入し，依存性を指数で評価することがある[14]が，式 (2) もこの考え方と同様である。また，この辺々について（自然）対数をとることで，

$$\log x_i = \sum_{i \neq j} b_{ij} \log x_j + e_i \tag{3}$$

と対数同士の和の形になる。つまり技術的には，わざわざ新たに積の構造的因果モデルに対応した因果探索アルゴリズムを構築しなくとも，データセットをすべて対数変換し，そのまま LiNGAM を適用すればよい。ただし，式 (3) を用いる場合は，以下の点に注意が必要である。

- 実数値として分析するためには，全ての変数が常に正の値をとることが条件となる。
- 式 (2)・(3) に見られるように，通常の LiNGAM とは異なり，誤差変数は変数の対数値について考え，非ガウス性を仮定する（ゆえに，誤差変数について厳密に議論しようとすると，やや複雑なものになる）。
- 和の構造的因果モデルと同様，非巡回性と外生変数の独立性（未観測共通要因が存在しないか，あっても影響は小さい）を仮定した分析になる。
- 通常の和の構造的因果モデルにおける分析では各変数の寄与（係数）を対等な条件で比較するため，各変数について平均値を引いて標準偏差で除する「標準化」を施すが，積の構造的因果モデルで現れる因果関係は式 (2) の通り指数の形であり，対数をとってから平均値を引く操作[15]を行えば，各変数からの寄与の比較という意味では十分である。

4 計算結果

[14] このモデルは重回帰を行う際のパラメーターの線形性という仮定を犯すものではない [高橋 22]。

[15] これは，対数をとる前に変数ごとに相乗平均で除して無次元化してから対数をとるという操作と等価。

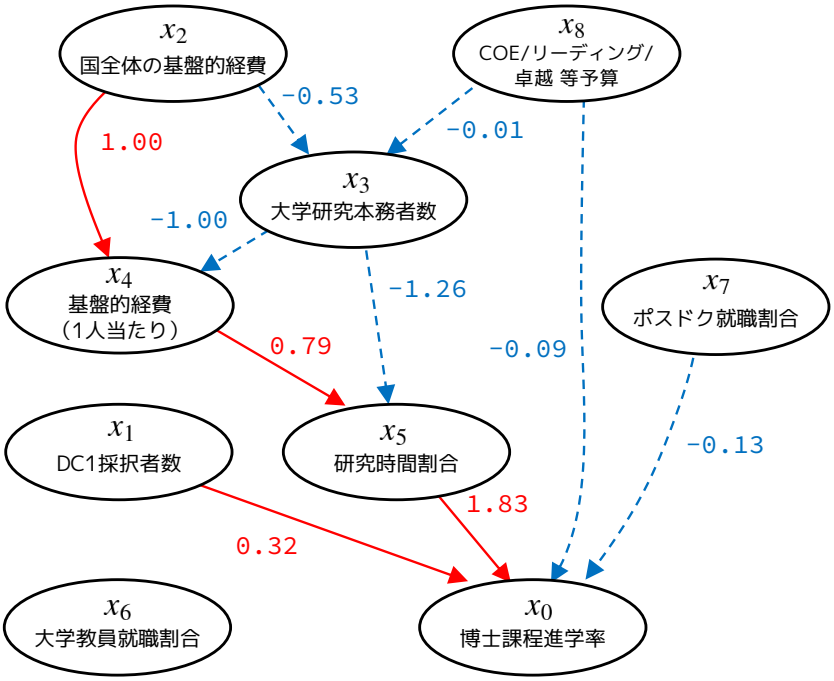

図 2　表 2 に基づき構成されるデータセット全体に対する DirectLiNGAM で計算された因果グラフと各因果効果。有向辺について，実線（赤）が正，破線（青）が負。積の構造的因果モデルでの計算であるため，因果効果は全て指数の形であることに注意。

■データセット全体に対する計算結果　まず，表 2 に基づき構成したデータセット全体に対する DirectLiNGAM での計算結果を，図 2 に示す。この計算結果において特徴的な点は，以下の通りである。

- 「国全体の基盤的経費 (x_2) → 一人当たり基盤的経費 (x_4)」と「大学研究本務者数 (x_3) → 一人当たり基盤的経費 (x_4)」の係数がそれぞれ 1.00 と −1.00 になっており，$x_4 = x_2 \div x_3$ の辺々について対数をとった結果と一致する。
- 博士課程進学率に直接，正の方向に寄与しているものは，「前年度 DC1 採択者数 (x_1)」と「研究時間割合 (x_5)」。負の方向に寄与しているものは「博士修了直後のポスドク就職割合 (x_7)」，「DC1 以外の経済的支援（COE /リーディング/卓越 等）(x_8)」，となった[16]。
- 大学研究本務者数 (x_3) および一人当たり基盤的経費 (x_4) から研究時間割合 (x_5) への影響が示唆される。
- 国全体の基盤的経費 (x_2) が大学研究本務者数 (x_3) に負の影響を与えていることが示唆される。

[16] ただし，この計算結果はあくまで博士課程進学率の増減について現状のデータ及び仮定の下で統計的に見た因果関係を示唆するにすぎず，この結果を持って直ちに経済的支援やアカデミックポストの整備が博士課程進学率に負の影響を及ぼすと結論づけるものではない。調査報告果 [加藤 09, 治部 21a] でも示されている通り，現場のニーズとして存在することを踏まえると，博士進学率の向上という目標達成に関わらず改善されることが望ましい。

表 3　ブートストラップ法によりサンプリング回数 4,000 回，DirectLiNGAM で計算した時の，各パスの因果係数が非ゼロとなる場合の符号と確率（一部）*

順位	直接的な因果関係			係数の符号	確率
1	x_2（国全体の基盤的経費）	→	x_3（大学研究本務者数）	−	94.0%
2	x_1（DC1 採択者数）	→	x_0（博士課程進学率）	+	92.9%
3	x_8（COE /リーディング/卓越 等予算）	→	x_3（大学研究本務者数）	−	92.1%
4	x_3（大学研究本務者数）	→	x_5（研究時間割合）	−	84.0%
5	x_4（一人当たり基盤的経費）	→	x_5（研究時間割合）	+	81.8%
6	x_8（COE /リーディング/卓越 等予算）	→	x_0（博士課程進学率）	−	78.6%
7	x_3（大学研究本務者数）	→	x_6（大学教員就職割合）	−	71.7%
8	x_1（DC1 採択者数）	→	x_6（大学教員就職割合）	+	62.8%
9	x_2（国全体の基盤的経費）	→	x_7（ポスドク就職割合）	−	58.9%
10	x_3（大学研究本務者数）	→	x_7（ポスドク就職割合）	−	58.3%
11	x_8（COE /リーディング/卓越 等予算）	→	x_5（研究時間割合）	+	57.8%
12	x_8（COE /リーディング/卓越 等予算）	→	x_7（ポスドク就職割合）	+	51.8%
13	x_2（国全体の基盤的経費）	→	x_6（大学教員就職割合）	−	50.9%

* 50.0% 以上のものについて，確率上位のものから順に表示。また実際の計算では，この表に掲載されているパス以外にも，$x_2 \to x_4$ と $x_3 \to x_4$ の非ゼロの確率が必ず 100% として出力される。これは $x_4 = x_3 \div x_2$ という関係にある各変数を加え，$x_2 \to x_4$ と $x_3 \to x_4$ というパスが存在することを事前知識として課して因果探索を行っていることによるものであり，本質的に意味を持たないため，この表からは除外している。

■ブートストラップ法による評価　データセット全体を用いて推定された因果グラフおよび各因果関係についての確からしさの評価の方法として，DirectLiNGAM とブートストラップ法[17]を組み合わせた手法 [Komatsu10] がある。本稿においては，サンプリング回数 4,000 回で試行したときに，直接的な因果関係が表れる割合であるブートストラップ確率が 50.0% 以上のものについて，表 3 のような結果を得た。先述のデータセット全体に対する DirectLiNGAM の結果で現れた因果関係は x_5（研究時間割合）$\to x_0$（博士課程進学率）[18]と x_7（ポスドク就職割合）$\to x_0$（博士課程進学率）[19]以外は，いずれもブートストラップ確率が 50.0% 以上であり，表 3 に含まれている。

図 3 には，代表的に x_4（研究者一人あたりの基盤的経費）$\to x_5$（研究時間割合）に関する係数の計算結果についてヒストグラムでプロットした。ここでは 0.8 付近のピークに加え，0.0 付近にも細いピーク[20]が存在する構造となっている。この 0.0 付近の細いピークを無視すれば，特に因果係数が正の

[17] データセット全体に対して再標本化をサンプリング回数分繰り返して統計解析を行う方法 [汪 92]。DirectLiNGAM では復元抽出による再標本化を行う。ただし，ブートストラップの再標本化はランダムであり，毎回厳密に同じ結果が得られるわけではないことには注意を要する。

[18] x_5（研究時間割合）$\to x_0$（博士課程進学率）は 14 位で確率 46.7%。

[19] x_7（ポスドク就職割合）$\to x_0$（博士課程進学率）は 17 位で確率 44.5%。

[20] ただし 0.0 付近の細いピークは，この $x_4 \to x_5$ にのみ現れる特徴ではなく，因果係数がゼロとなる確率（つまり，100%− 因果係数が非ゼロとなる確率 (%)）そのものであり，因果関係にない可能性が高いパスほど顕著になる。

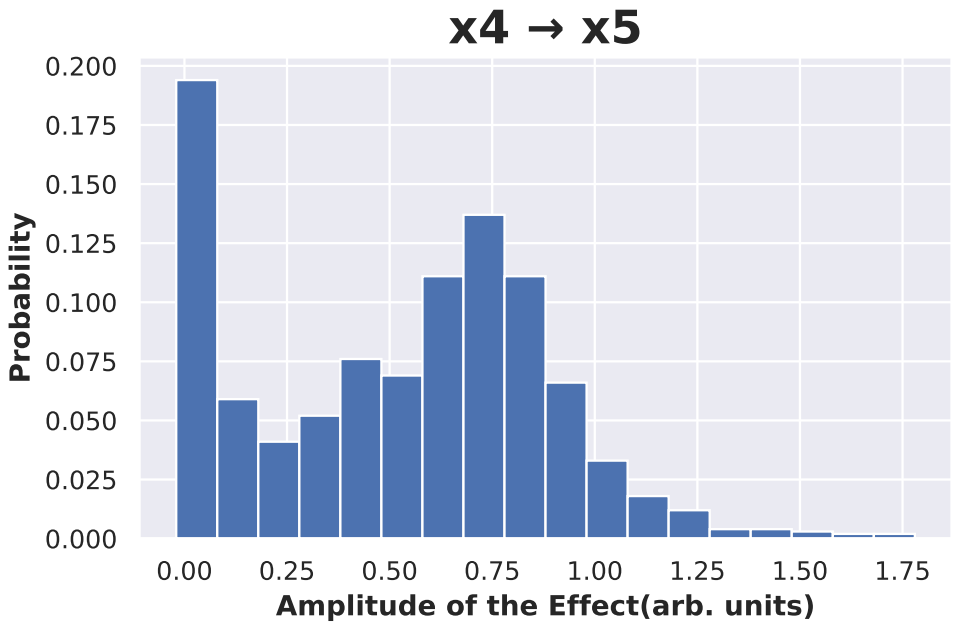

図 3　表 3 のうち x_4（研究者一人あたりの基盤的経費）→ x_5（研究時間割合）について係数の計算結果についてヒストグラムで表示したもの。

領域では，0.8 付近のピークは右側よりも左側にテールをひいた非対称な構造になるというのが特徴となっている。

5　考察

ここでは，LiNGAM での因果探索の結果をもとに，政策研究の領域知識も加えた定性的な因果関係の考察，他の統計的因果推論手法 (特に共分散構造分析) から得られる結果との比較を行う。

5.1　LiNGAM による因果探索結果の解釈

特にデータセット全体に対する DirectLiNGAM の結果に基づいて，各種政策要素間の因果関係を定性的に考察する。

■x_2（**国全体の基盤的経費**）→ x_3（**大学研究本務者数**）　**【係数**: −0.53**】**　特に負の効果となっている点は，ブートストラップ確率は 94.0% と極めて高い一方で，説明が非常に困難であると言える。国全体の基盤的経費については 2013 年度までの減少傾向以降はほぼ横ばいであるのに対し，大学研究本務者数は増加傾向にあり，一定程度の逆相関があるように見える。しかしながら，総務省の “科学技術研究調査” によれば，大学研究本務者数の増加は保健分野での増加が主たる寄与と見られ，特定の分野での研究本務者数の増加が基盤的経費の減少に起因するとは考え難い。以上の観点から，この因果探索結果を基に因果関係を深堀するにあたっては，このパスは棄却した上で進めることが有効とも考えられる。

■x_3（大学研究本務者数） → x_5（研究時間割合）【係数: −1.26】　大学研究本務者数が増えることで研究時間が減っている，という結果であり，この結果を定性的に解釈をすると，例えば大学研究本務者が増えることで会議や事務的な調整等の学内事務に関する業務が増大し，その分一人当たりの研究時間が減ってしまっているという可能性が考えられる。

そして表 3 の通り この直接的な因果関係の大きさを表す係数が非ゼロとなるブートストラップ確率が 84.0% と非常に高いことも踏まえると，現時点でこの因果関係は，未観測共通要因の存在可能性も含め，棄却できない。むしろこの因果関係の有無を明らかにするにはまず，研究時間に関する議論について更なる深堀が期待される。

■x_4（研究者一人あたりの基盤的経費） → x_5（研究時間割合）【係数: 0.79】　この $x_4 \to x_5$ が示す直接的な因果関係は，「研究本務者一人当たりの基盤的経費を増やす（減らす）」ことで「研究時間割合が増える（減る）」ということである。特にこの 2 変数は，若手研究者支援や研究力向上における重要課題としてこれまでばらばらに議論されてきた[21]が，一方でこれらの要素間の関係は，不思議と議論されてこなかった。そのため，この 2 変数間の直接的な因果関係を示す係数が非ゼロとなるブートストラップ確率が表 3 の通り 81.8% と非常に高い値が示されたことは，非常に興味深い結果である。

直観的には，たとえば大学の研究者が使える基盤的経費を減らすと，その分競争的資金の確保が必要となり，その申請書の作成等に追われ，研究時間割合が減る，という説明が考えられる。しかし，この研究時間割合の引用元の FTE 調査では，競争的資金等の申請に係る文書等の作成時間は研究時間に含むものとしている。そのためこの説明は必ずしも成り立つとは言えず，また FTE 調査でも競争的資金等の申請に係る文書等の作成時間を詳らかにしたのは最新の 2018 年度調査のみであることから，この文書等の作成時間を除いた研究時間割合がどのように変化しているのか確認することができない。一方で，計算結果において一定の統計的信頼度が出ていることや，図 3 の構造も鑑みると，引き続き何らかの未観測共通要因の存在も視野に入れつつ，説明可能なロジックを模索する価値はあると考えられる。そのためには，先述の通り研究時間に関する議論のさらなる深堀が期待される。

■x_1（DC1 採択者数） → x_0（博士課程進学率）【係数: 0.32】　DC1 採択者数は，本研究における変数の選定の観点であった「経済的支援」に対応するものの一つであり，このパスに関する計算結果は，修士課程修了者向けのアンケート調査結果 [加藤 09, 治部 21a] と整合する。DC1 の応募資格においても，特に応募時点で修士課程在籍の場合については「採用年度の 4 月に博士課程後期等に進学する予定」であることが期待されているが，DC1 に実際に採択されたことで実際に博士進学する，逆に DC1 に採択されなかったことで博士進学を断念するという可能性も想定されることから，このパスに関する結果は直観的な解釈とも整合する。

[21] 特に研究時間の減少については，大学教員の学内事務の増大とセットで課題として議論されており，政策的にはこれまで，リサーチアドミニストレーター (URA; University Research Administrator) の利活用やバイアウト制の導入がその解決策として挙げられてきた。

■x_8（COE /リーディング/卓越等予算） → x_0（博士課程進学率） 【係数: −0.09】 x_8 も，本研究における変数の選定の観点であった「経済的支援」に対応するもう一つの変数であるが，このパスの計算結果は DC1 の場合とは対照的に，係数の絶対値は 0.09 と小さいものの，符号が負となっており，修士課程修了者向けのアンケート調査結果 [加藤 09, 治部 21a] からの期待とは逆の結果となっているため，この点は現段階では直観的な説明は困難である。しかし，DC1 との経済的支援のスキームの違いをよく把握しておくことは，今後の議論において重要になる可能性もある。DC1 を始めとする日本学術振興会の特別研究員制度では，全大学の大学院生が対象となり，直接日本学術振興会が申請内容から大学院生を選定する一方，グローバル COE プログラムや博士課程教育リーディングプログラム等は，大学拠点・プログラム単位で選定がなされ，応募可能な大学院生は，選定された大学拠点に所属する者に限られ，選定も大学側の裁量により行い，さらには経済的支援の在り方もプログラムによって様々である。このように，同じ国家予算による経済的支援であってても，経済的支援のターゲット選定の方法・範囲が異なる点を踏まえ，今後因果探索をさらに高度化・深堀していくにあたって，注意深い議論を要する。

■x_7（ポスドク就職割合） → x_0（博士課程進学率） 【係数: −0.13】 ポスドクの就職割合は，本研究における変数の選定の観点であった「アカデミアへのキャリアパス」に対応するものの一つであるが，このパスの計算結果は，修士課程修了者向けのアンケート調査結果 [加藤 09, 治部 21a] からの期待とは逆になっている。一方，アカデミアとしてのキャリアパスのうちポスドクについては様々な議論がある。例えば治部らによる修士課程修了者向けの調査 [治部 21a] では自由記述において，「博士課程への入り口を増やす（経済的補助や研究費の拡充）よりも博士後期課程の出口を増やす（ポスドク）」という記述があるなど，ポスドクのポストが増えることが重要とする声もあれば，同時に「ポスドクの雇用条件の改善」も求められている。さらに，川村らによる最近の博士課程修了者向けの調査 [川村 22] では自由記述において，職業としての安定性や経済的困窮、社会保障に関する課題といった，いわゆる「ポスドク問題」が多く指摘されるなど，ポスドクの待遇状況次第では必ずしもこのポストの拡充が博士課程進学のインセンティブにはならない可能性も示唆されている。このように，ポスドクのポストと博士課程進学率の関係については，肯定的・否定的な見解が両方あり得るため，本計算結果の妥当性を検証することは困難である。この因果について，より精緻な議論を行う上では，ポスドクの雇用環境に関する変数とセットでの因果推論が有効であると期待される。

■x_5（研究時間割合） → x_0（博士課程進学率） 【係数: 1.83】 研究時間割合は本研究における変数の選定の観点であった「研究環境」に対応するものの一つであり，このパスに関する計算結果は，特に NISTEP の定点調査，および修士課程修了者向けのアンケート調査結果 [加藤 09, 治部 21a] と整合するものといえる。この解釈として直観的には「大学の教員が，大学院生に研究そのものの指導および議論の時間を十分に確保できるほど，安定して研究成果を創出しやすくなることから，修士課程学生が安心して博士課程学生に進学できるようになる」といったことが考えられる。しかし，本研究で採用した値は「大学等教員」に関する研究時間割合であり，ここでいう教員は教授，准教授，

講師及び助教と幅広く，それぞれ研究活動における役割も異なると考えられることから，一概にまとめてしまって良いかなど議論の余地がある。今後この因果関係を詳細に議論するにあたっては，実際には大学院生の研究活動に対する関わり方や度合いがそれぞれ異なる可能性を踏まえて進めていく必要がある。また，FTE 調査の「研究時間」において大学院生の研究指導に充てられている時間がどれだけを占めているのか，といった詳細を明らかにしていくことが今後期待される。

なお，本研究では研究環境の要素として「基盤的経費」「研究時間割合」をピックアップしたが，そもそも研究環境に関する指標についての統一的見解は存在せず，「基盤的経費」と「研究時間割合」の因果関係の存在可能性もあるように，相互に独立した変数のみで研究環境を定量的に議論することが困難である可能性もある。また，「研究環境」という言葉について，研究成果創出の文脈での意味と，博士課程進学の文脈での意味が完全に一致しているかどうかも定かではなく，これらの点についての今後の深堀が期待される。

5.2 共分散構造分析との比較

DirectLiNGAM で導かれたこの因果グラフの妥当性・新規性について，ここまでは政策研究の領域知識に基づいた定性的な解釈から論じたが，ここでは従来型の統計的因果推論とも定量的な比較を行う。統計的因果推論の典型的な手法は 2.2 節に記載の通りだが，ここでは共分散構造分析を採用する。共分散構造分析では，SEM に基づく回帰分析を行うとともに，統計的なモデル適合度の評価や潜在変数（構成概念）の導入による因子分析[22)]も行う。

ただし先述の通り，共分散構造分析は因果探索ではなく，あくまで事前に因果グラフを仮定した上で，その妥当性を評価するものである。本研究では，この変数群に基づき考えられる仮説（因果グラフ）をいくつか設定し，DirectLiNGAM で得られた因果グラフと共に共分散構造分析にかけて簡易的にモデル適合度の比較を行う。なお本研究における共分散構造分析の実施には，Python の semopy[23)] ライブラリ [Igolkina20] を用い，データセットについては表 2 に示した変数群を用いた。また，本研究での DirectLiNGAM との比較という観点から，積の構造的因果モデルに基づいて分析を行っている。

■政策研究の領域知識に基づいて仮説として描いた因果グラフ　本研究では，先述のデータセット全体に対する DirectLiNGAM の計算結果を含む計 5 種類の因果グラフを仮説として設定し，共分散構造分析にかけた。各因果グラフの性質について以下で説明する。なお，モデル 1〜4 の因果グラフは図 4 にまとめて示している。いずれのモデルでも，一人あたり基盤的経費の導出の関係式が入っているが，semopy では係数を固定した解析も可能であるため，ここではその関係を固定条件として導入して共分散構造分析を行っている。

モデル 0　政策研究の領域知識に基づいて立てられた仮説と比較し，DirectLiNGAM によりデータ駆動的に生成された仮説の質を評価するため，データセット全体に対する DirectLiNGAM

22) ただし本研究では，潜在変数の導入までは行わない。

23) `https://semopy.com/` version 2.3.9.

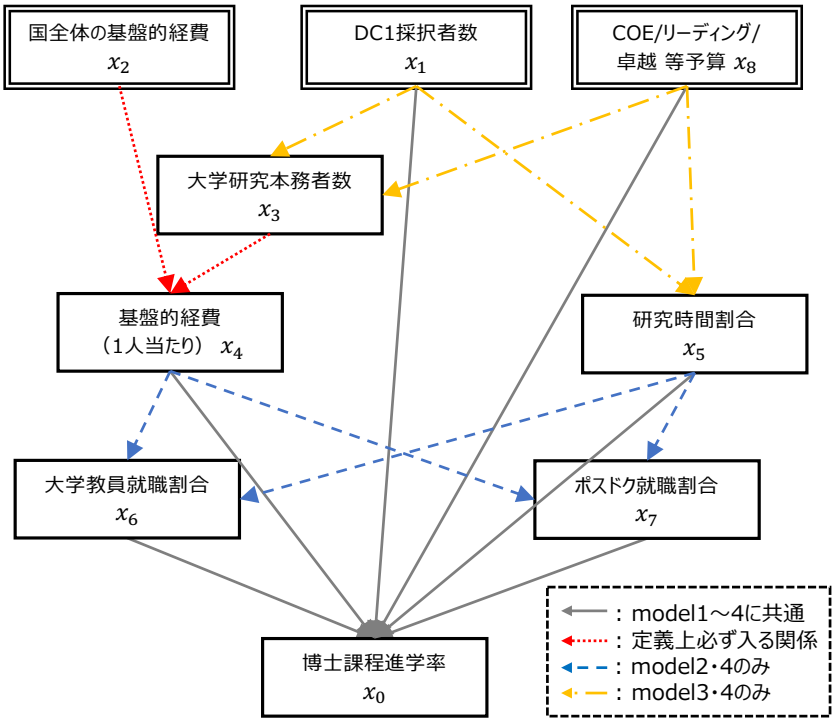

図 4　共分散構造分析にあたって考案したモデル 1～4 の因果グラフ。二重線枠の変数は外生変数。

の計算結果として得られた因果グラフをモデル 0 としている。

モデル 1　DirectLiNGAM で因果探索を行うにあたって設定した変数群の選定の考え方に準じ，「経済的支援」(x_1・x_8)・「研究環境」(x_4・x_5)・「アカデミックポストへの就職率」(x_6・x_7) が全て博士課程進学率 (x_0) への直接の要因であると仮定する。なお，これらの要因の変数は全て互いに独立であるとする。

モデル 2　大学におけるアカデミックポストへの就職率は，大学におけるポストそのものが整うこともちろん重要な要因だが，博士課程修了者がアカデミアに進むことを目指すという観点からは，やはりポストドクターや大学教員になった際の研究環境が正の影響を及ぼす可能性が想像される。そこで，x_4・$x_5 \to x_6$・x_7 を仮定し，モデル 1 に加えた。

モデル 3　経済的支援が影響を及ぼしうるのは，その採択までのプロセスや採択者の性質を踏まえると，厳密には博士課程進学率だけでない。まずこれらの採択者はルール上必然的に博士課程に進学することになるのと，大学研究本務者は博士課程学生を含むという性質を踏まえると，経済的支援の拡大は，大学研究本務者数を増大させる方向で，直接の影響を及ぼす[24)]。また，これらの経済的支援は，単に競争的資金と同様に申請書を書き，採択されたら終わりというわけではなく，たとえば DC1 であれば研究費の使用に関わる書類事務が発生したり，

[24)] なお，厳密にはこれと同様の理由で，$x_0 \to x_3$ というパスが考えられる。しかしながら，このパスを導入することで非巡回の仮定を満たせなくなること，および本研究のターゲットが博士課程進学率の向上に向けた因果関係の解明に重点を置いていることから，ここではこのパスの導入は一旦見送ることとした。

リーディング大学院等では講義受講やインターンシップ等の研究以外の義務が発生することから，研究時間割合に負の影響が出る可能性も考えられる[25)]。そこで，x_1・$x_8 \rightarrow x_3$・x_5 を仮定し，モデル 1 に加えた。

モデル 4　モデル 2・3 の両方を考慮した，複雑な因果グラフ。つまり，モデル 1 に x_4・$x_5 \rightarrow x_6$・x_7 と x_1・$x_8 \rightarrow x_3$・x_5 を加えている。

■共分散構造分析の結果比較　先述の各モデルに対し，共分散構造分析にかけることで得られるモデル適合度のうち，赤池情報量基準 (Akaike's Information Criterion, AIC)[赤池 73] とベイズ情報量基準 (Bayesian Information Criterion, BIC) を採用し，得られた各パスの係数値・標準誤差と共に表 4 に示している。なお，AIC・BIC は相対的な値であり，小さい値であるほどモデル適合度が高い。

まずモデル適合度の観点では、AIC・BIC のどちらにおいても，DirectLiNGAM により導かれたモデル 0 は，潜在変数なしの場合ではモデル 1・2 よりもモデル適合度が高く，モデル 2・3・4 のように新たなパスを加えることでモデル 1 よりも改善されていき，モデル 4 になるとモデル 0 よりもモデル適合度は高くなっていることがわかる。

次に係数の計算結果を見てみると，大雑把に以下の特徴が見てとれる。

- モデル 0 については，全体的に DirectLiNGAM の出力と異なる計算結果となっている。
  — この不整合の原因は，各分析の過程において解を決定するにあたり最適化される指標が異なることである。共分散構造分析では，あらかじめ全ての因果的順序を仮定しているので，因果的順序自体を最適化することをアルゴリズムに組み込んでおらず，純粋に最小 2 乗法に従って残差の 2 乗和が最小となるような解を探索する。一方で DirectLiNGAM は 2.2 節でも述べた通り，回帰分析と誤差項同士の独立性評価を繰り返し，誤差項同士の相関が最小化されるような因果的順序が選択され，それに従って，冗長な有向辺の枝刈りを行いつつ係数が決定される。この違いにより，DirectLiNGAM で得られた結果はしばしば共分散構造分析の結果とは異なる係数が出力されることがある。
- 特にモデル 0〜3 については，を除き，各係数の計算結果の絶対値が 0.1 を下回る結果となっており，さらに標準誤差が非常に大きく，係数の符号の決定はモデル 0〜3 のどのパスの係数においても困難である。モデル 4 においてはごく一部のパスで，標準誤差が係数の絶対値に対して小さくなるなど改善はみられているものの，因果グラフ全体で見ると決定精度は十分とは言えず，係数の符号について議論も難しい。
  — 係数の絶対値の小ささは，特にモデル 0〜3 については，x_0（博士課程進学率）に大きく影響する要因がこのモデルにおいては存在しない可能性を示唆している。これは，時間遅れの効果の考慮や未観測共通要因の特定により改善する可能性も考えられる。しかしそれ以前に，

[25)] もちろん，これらの付随的に発生する義務に係る時間が FTE 調査内でどのように整理されているかは，別途検証の余地がある。

表 4　共分散構造分析の結果比較。係数における ± は標準誤差を表す。ただし，$x_2 \to x_4$ と $x_3 \to x_4$ の係数はそれぞれ 1.00 と −1.00 と固定した上で共分散構造分析を行っている。

計算結果\モデル	モデル 0	モデル 1	モデル 2	モデル 3	モデル 4
AIC	−84.91	−43.41	−61.80	−79.99	−102.84
BIC	−78.13	−38.89	−53.89	−72.08	−91.54
$x_1 \to x_0$ の係数	−0.00 ± 4.43	−0.00 ± 7.58	−0.00 ± 7.56	−0.00 ± 4.62	−0.19 ± 3.53
$x_1 \to x_3$ の係数	—	—	—	0.01 ± 21.03	0.02 ± 19.99
$x_1 \to x_5$ の係数	—	—	—	−0.01 ± 6.02	−0.19 ± 4.70
$x_2 \to x_3$ の係数	−0.00 ± 138.30	—	—	—	—
$x_2 \to x_4$ の係数	1.00	1.00	1.00	1.00	1.00
$x_3 \to x_4$ の係数	−1.00	−1.00	−1.00	−1.00	−1.00
$x_3 \to x_5$ の係数	−0.00 ± 29.71	—	—	—	—
$x_4 \to x_0$ の係数	—	0.00 ± 34.14	0.00 ± 67.90	0.00 ± 0.06	0.18 ± 0.05
$x_4 \to x_5$ の係数	0.00 ± 29.71	—	—	—	—
$x_4 \to x_6$ の係数	—	—	0.00 ± 246.70	—	0.00 ± 0.21
$x_4 \to x_7$ の係数	—	—	0.00 ± 48.00	—	0.07 ± 0.04
$x_5 \to x_0$ の係数	0.00 ± 0.20	0.00 ± 33.28	0.00 ± 54.26	0.00 ± 0.20	0.25 ± 0.21
$x_5 \to x_6$ の係数	—	—	0.00 ± 181.84	—	0.00 ± 0.91
$x_5 \to x_7$ の係数	—	—	0.00 ± 35.38	—	0.13 ± 0.18
$x_6 \to x_0$ の係数	—	0.00 ± 26.20	0.00 ± 0.08	0.00 ± 26.89	0.03 ± 0.06
$x_7 \to x_0$ の係数	0.00 ± 8.09	0.00 ± 7.88	0.00 ± 0.39	0.00 ± 8.16	0.02 ± 0.33
$x_8 \to x_0$ の係数	0.01 ± 2.10	0.02 ± 3.74	0.01 ± 3.77	0.01 ± 2.18	0.81 ± 1.51
$x_8 \to x_3$ の係数	−0.05 ± 11.34	—	—	−0.05 ± 8.96	−0.05 ± 8.51
$x_8 \to x_5$ の係数	—	—	—	0.02 ± 2.56	0.46 ± 2.00

決定精度の悪さそのものも大きな問題であり，この原因としては，回帰分析を適用するにあたって，データ点数が変数の数に対して不足していることが考えられる。

ここまで考察したように，本研究において構築したデータセットに基づいて，これらの統計的処理のみによって因果関係に関する結論を導くことは困難である。ただし少なくとも，LiNGAM を用いることで，少ない事前知識のみでデータ駆動的に，モデル適合度が一定程度高い因果グラフを出力することが期待され，さらに領域知識に基づいて人力で因果グラフを描くのみでは見つけられなかったような因果のパスが可能性として見つけられる可能性は十分ある。他方，因果探索を行わない場合には，

- 本研究では，モデル 3・4 のように比較的モデル適合度の高い因果グラフを人力でも設定することはできたものの，属人的な部分が大きく，仮に領域知識がある程度蓄積されていても，

それらに基づいて初めからモデル適合度の高い因果グラフを設定できるという保証はない点
- 網羅的に共分散構造分析のみでモデル適合度を比較・評価しようとすると，変数の数の増大に伴って試行回数が莫大[26)]になってしまい，現実的ではない点
- 一方で既存の領域知識のみに頼って効率的に探索しようとすると，領域知識が固定観念化し，枠の外にある変数を見逃してしまう可能性も高まることから，統計的に正しくかつ政策科学的にも有用な新しい因果関係の発見にたどり着けない可能性もある点

などの課題が挙げられる。これらを踏まえると，研究成果の新規性・因果推論の効率性の両面で，LiNGAM と従来の因果推論のアプローチとの併用により，強力な効果を発揮しうることが改めて示唆される。

6 おわりに

本稿では，研究力強化・若手研究者支援に関する EBPM に向けて，統計的因果探索手法である LiNGAM を用い，公開されている統計データのみをもとに博士課程進学率に関する因果関係の推定を行うとともに，その結果について簡単に考察した。4・5 章を通じ，ここまで DirectLiNGAM を用いた計算の結果と新たな因果関係の存在可能性，その統計的信頼性とこれらに基づいた定性的解釈，および共分散構造分析による因果推論との比較について述べたが，博士課程進学率に関する因果グラフを精度よく決定する上では，以下の点が課題となる。

- 変数 9 つに対して，データ点数が 13 と変数の数と同程度となっている点。
  — 本来であれば，LiNGAM の仮定の一つである誤差変数の非ガウス性のチェックが必要であるが，さすがにこのデータ点数では誤差変数の正規分布からのずれを正しく評価することが困難である。また，共分散構造分析を行うにあたっても，係数決定精度にこのデータ点数の不足が影響している可能性がある。
- 研究時間割合や基盤的経費の現実の配分や因果関係は，分野ごとに異なる可能性。
  — 特に基盤的経費については，研究に関連する利用がどの程度のものかは不明であるのと，研究時間割合・基盤的経費は分野だけでなく，地域・大学によっても事情が異なる点があり，厳密には注意深い議論を要する。
- この分析では，その年度中に影響が生じることを暗に仮定していたが，実際には各要因の変化による影響は年単位で遅れて現れる可能性。
  — 例えば，研究環境の改善や経済的支援の増大の効果が即時 博士課程進学率の改善につながるとは考えにくく，それぞれがどの程度の時間遅れを伴って効果が現れるか，という議論も必要になる。

今後，博士課程進学率に関してより正確に因果関係を突き詰めていく上では，本稿の結果をもっ

[26)] 変数の数 n に対し，領域知識等で一切縛らなければ，任意の 2 変数の順列それぞれについて，直接的影響あり/なしの 2 通りの検証が必要であるため，$2^{n(n-1)}$ 通りについての評価が必要となる。

てそのまま因果関係・係数を断定することはせず，あくまで示唆とし，たとえば治部らによる調査 [治部 21a] の個票データを用いたサンプルサイズ・統計的信頼性の問題克服，BN により作成した因果グラフとの比較，分野に応じた因果関係の相違点の抽出，VAR-LiNGAM[Hyvarinen10] を用いた遅延効果込みでの因果探索，RCD (Repetitive Causal Discovery) を用いた未観測共通要因の存在箇所の特定 [Maeda20] 等を通じた，より厳密な因果関係の解明が期待される。

参考文献

[Hyvarinen10] A. Hyvärinen and K. Zhang and S. Shimizu and P. O. Hoyer： Estimation of a structural vector autoregression model using non-gaussianity. Journal of Machine Learning Research, 11:1709–1731, 2010 https://www.jmlr.org/papers/volume11/hyvarinen10a/hyvarinen10a.pdf

[Hyvarinen13] A. Hyvärinen and S. M. Smith： Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. Journal of Machine Learning Research, 14:111–152, 2013. https://jmlr.org/papers/v14/hyvarinen13a.html

[Igolkina20] Anna A. Igolkina and Georgy Meshcheryakov：semopy: A Python Package for Structural Equation Modeling, Structural Equation Modeling. A Multidisciplinary Journal, Vol.27, Issue 6, pp.952–963, 2020. https://doi.org/10.1080/10705511.2019.1704289

[Komatsu10] Yusuke Komatsu, Shohei Shimizu, and Hidetoshi Shimodaira： Assessing statistical reliability of LiNGAM via multiscale bootstrap. In Proc. International Conference on Artificial Neural Networks (ICANN2010), Thessaloniki, Greece, pp.309–314, 2010. https://doi.org/10.1007/978-3-642-15825-4_40

[Maeda20] T. N. Maeda and S. Shimizu： RCD: Repetitive causal discovery of linear non-Gaussian acyclic models with latent confounders. In JMLR Workshop and Conference Proceedings, AISTATS2020 (Proc. 23rd International Conference on Artificial Intelligence and Statistics), Palermo, Sicily, Italy., 735—745, 2020. http://proceedings.mlr.press/v108/maeda20a/maeda20a.pdf

[Okamura19] Keisuke OKAMURA： Interdisciplinarity revisited: evidence for research impact and dynamism. Palgrave Commun, Vol.5, No.141, 2019. https://doi.org/10.1057/s41599-019-0352-4

[Pearl 85] Judea Pearl.： Bayesian Networks: a Model of Self-Activated Memory for Evidential Reasoning. Proceedings, Cognitive Science Society, 329–334, 1985. https://ftp.cs.ucla.edu/pub/stat_ser/r43-1985.pdf

[Shimizu06] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen：A linear non-gaussian acyclic model for causal discovery. Journal of Machine Learning Research, 7:2003-2030, 2006. https://www.cs.helsinki.fi/group/neuroinf/lingam/JMLR06.pdf

[Shimizu11] S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O.

Hoyer and K. Bollen： DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. Journal of Machine Learning Research, 12(Apr): 1225-–1248, 2011. https://dl.acm.org/doi/10.5555/1953048.2021040

[Thamvitayakul12] K. Thamvitayakul, S. Shimizu, T. Ueno, T. Washio and T. Tashiro： Bootstrap confidence intervals in DirectLiNGAM. In Proc., 2012 IEEE 12th International Conference on Data Mining Workshops (ICDMW2012), Brussels, Belgium, pp.659–668, 2012. https://doi.org/10.1109/ICDMW.2012.134

[赤池 73] Akaike H.： Information theory and an extension of the maximum likelihood principle. Proceedings of the 2nd International Symposium on Information Theory, 267–281, 1973.

[枝村 16] 枝村 一磨： 環境規制と経済的効果-製造事業所の VOC 排出に関する自主的取組に注目した定量分析. NISTEP DISCUSSION PAPER, No.133, 2016. http://hdl.handle.net/11035/3132

[浦田 05] 山野井敦徳, 藤村正司, 浦田広朗： 日本の大学教員市場再考–現在・過去・未来. 広島大学高等教育研究開発センター, 第 5 章:47–54, 2005

[浦田 01] 浦田広朗： 1990 年代における大学院拡大. 麗澤学際ジャーナル, 第 9 巻第 2 号:52–75, 2001

[汪 92] 汪 金芳; 大内 俊二; 景 平; 田栗 正章.：ブートストラップ法-最近までの発展と今後の展望-. 行動計量学, 19 巻第 2 号 50–81, 1992. doi:10.2333/jbhmk.19.2_50

[高橋 22] 高橋将宜：統計的因果推論の理論と実装. 共立出版, 2022

[加藤 09] 加藤真紀, 角田英之： 日本の理工系修士学生の進路決定に関する意識調査. 文部科学省科学技術政策研究所 調査資料 (Research Material), No.165, 2009. http://hdl.handle.net/11035/895

[川村 22] 川村真理, 星野利彦： 博士人材追跡調査-第 4 次報告書-. 文部科学省科学技術・学術政策研究所 調査資料 (Research Material), No.317, 2022. https://doi.org/10.15108/rm317

[治部 21a] 治部眞里, 星野利彦：修士課程（6 年制学科を含む）在籍者を起点とした追跡調査（2020 年度修了（卒業）者及び修了（卒業）予定者に関する報告）. 文部科学省 科学技術・学術政策研究所 調査資料 (Research Material), No.310, 2021. https://doi.org/10.15108/rm310

[治部 21b] 治部眞里, 星野利彦： 博士離れの要因についての一考察. 文部科学省 科学技術・学術政策研究所 STI Horizon, Vol.7, No.2, 2021. https://doi.org/10.15108/stih.00260

[清水 17] 清水昌平：統計的因果探索. 講談社 機械学習プロフェッショナルシリーズ, 2017.

[高山 21a] 高山正行, 星野利彦：博士人材の年齢別人材流動モデルと試行的な将来予測. NISTEP Discussion Paper, No.193, Feb 2021. https://doi.org/10.15108/dp193

[高山 21b] 高山正行, 小柴等, 前田高志ニコラス, 三内顕義, 清水昌平, 星野利彦：EBPM と統計的因果探索・数理モデルの利活用. 研究・イノベーション学会　第 36 回年次学術大会 (予稿集)., 公演番号 2G02, 2021.

[鳥海 18] 鳥海 航, 生方 裕一, 久野 譜也, 岡田 幸彦： 地域健康政策へのベイジアンネットワーク

の応用．統計数理，Vol.66, No.2, pp.267–278，2018. https://www.ism.ac.jp/editsec/toukei/pdf/66-2-267.pdf

[中山 10] 中山 保夫, 細野 光章, 長谷川 光一, 永田 晃也：産学連携データ・ベースを活用した国立大学の共同研究・受託研究活動の分析．文部科学省 科学技術・学術政策研究所 調査資料 (Research Material), No.183, 2010. http://hdl.handle.net/11035/887

[野村総研 10] 野村総合研究所：博士課程進学の環境を改善するためのノンアカデミック・キャリアパスに関する調査最終報告書. 野村総合研究所, 2010

[福澤 15] 福澤 尚美, 伊神 正貫： 科学技術の状況の俯瞰的可視化に向けて―NISTEP 定点調査 2011～2014 のパネルデータを用いた質問項目間の関係性についての定量分析―．NISTEP DISCUSSION PAPER, No.128, 2015. http://hdl.handle.net/11035/3112

付録 A　実行環境等

本解析を実行するに当たっては，オンライン環境である Google Colaboratory [27)]を用いた。

本稿掲載の解析実行時における当該環境の Python のバージョンは 3.7.12。その他，LiNGAM パッケージのバージョンは 1.5.5。関連する主要なパッケージのバージョンは以下の通りである。

graphviz (0.10.1)	joblib (1.1.0)	numpy (1.21.5)	pandas (1.3.5)
patsy (0.5.2)	python-dateutil (2.8.2)	pytz (2018.9)	scikit-learn (1.0.2)
scipy (1.4.1)	six (1.15.0)	statsmodels (0.10.2)	threadpoolctl (3.1.0)

各種アルゴリズムの更改により，使用するバージョンやその組み合わせによって，解析結果が変化する可能性があることに留意を要する。

付録 B　データ及びコード

本稿で解析に用いたデータ及びコードの一部は以下から取得，確認できる。

データ

DOI: https://www.doi.org/10.15108/data_doctoral_2022_0317

コード

https://colab.research.google.com/drive/1bQ5u-yy_1qBSb4Uf8lx2D03FS6FbxMNd?usp=sharing

27) https://colab.research.google.com/

付録 C　参考情報等

C.1　利益相反

本研究の実施について，外部機関等からの資金提供は受けておらず，その点での利益相反はない。

C.2　倫理条項

本研究で用いているデータは公開情報かつ統計値であるため，個人情報等を含まず，その点での倫理的問題を有さない。

C.3　著者の貢献

研究の立案，原稿執筆等，主要な部分は責任著者である高山が実施した。その他，具体の分担等は以下の通り。

研究の企画・構想　高山正行，小柴等，前田高志ニコラス，三内顕義，清水昌平，星野利彦
分析手法等検討・確認　高山正行，清水昌平，前田高志ニコラス，三内顕義
データ収集・解析　高山正行
解釈・考察　高山正行，清水昌平，星野利彦，前田高志ニコラス，三内顕義，小柴等
論文執筆　高山正行，小柴等