SSJDA－096

社会調査・データアーカイブ共同利用・共同研究プロジェクト

社会調査データの統合による二次分析研究の刷新：作業手順と今後の調査実施に向けた示唆

麦山亮太（学習院大学）
平松誠（北海道大学）
コンアラン（大阪経済大学）
那須蘭太郎（東京大学大学院）
吉田航（国立社会保障・人口問題研究所）

東京大学社会科学研究所
附属社会調査・データアーカイブ研究センター

2025 年（令和 7 年）9 月

社会調査データの統合による二次分析研究の刷新：

作業手順と今後の調査実施に向けた示唆

麦山亮太（学習院大学）
平松誠（北海道大学）
コンアラン（大阪経済大学）
那須蘭太郎（東京大学大学院）
吉田航（国立社会保障・人口問題研究所）

要旨

本稿の目的は，複数の大規模社会調査データを統合（ハーモナイゼーション）し活用する「社会調査データの合併による二次分析研究の刷新」プロジェクトにおける作業手順を詳述することにある．既存の社会調査データの二次分析は標準的な研究手法となっているが，複数データの統合は各研究者が個別に行うにとどまり，労力の浪費，再現性の欠如，分析可能な研究課題の限定といった問題があった．この問題意識から，本プロジェクトでは国内大規模社会調査データの統合作業を複数名の研究者のもとで実施した．その過程は，（1）データの選定，（2）Stata と共有フォルダによる環境構築，（3）全データの連結，（4）変数のハーモナイゼーション，という段階からなる．特に変数作成においては，担当者間の共同作業を通じて，調査間で異なる設問文や選択肢の扱いといった課題にどのように対応したかを詳細に記録した．一連の作業を通じて，共同作業による作業の透明化，正確性の向上，そして調査間の異同に関する知見の蓄積といった利点が明らかになった．さらにこれらの経験に基づき，今後の社会調査の企画・設計に向けていくつかの具体的な提言も記載した．本稿で提示する作業手順と知見は，社会調査データの利用可能性を拡大させるとともに，より頑健で再現性の高い実証研究の発展に貢献する．

1　序論

東京大学社会科学研究所附属社会調査・データアーカイブ研究センターと大阪商業大学 JGSS 研究センターによって国内の社会調査データの共有・公開がなされるようになってから 20 年余，日本の大規模社会調査データの利用をめぐる環境は飛躍的に改善した．1990 年代末より相次いで全国を対象とした無作為抽出の社会調査が実施され，それらの研究成果が公刊されるだけではなく，アーカイブへの寄託・公開を通じて様々な研究者等に利用され二次分析による研究成果の産出につながっている．今や実証研究において既存の大規模社会調査データの二次分析は標準的な方法である．

こうしたなかで求められるのは，既存のデータをより有効に活用することである．なかでも既存社会調査データを統合した分析は，サンプルサイズを増やすとともに複数時点の比較分析を可能とし，より正確により大きな問いに答えるうえで重要である．国内外でこうした実証研究例は枚挙に暇がない．たとえば社会階層研究における Citation classic となっている研究として，F. Torche による大卒層において出身階層と到達階層の関連が小さいことを示した研究（Torche 2011），R. Breen らによるヨーロッパにおいて教育機会の不平等が縮小傾向にあることを示した研究（Breen et al. 2009）が挙げられる．これらは複数の社会調査を統合したデータを利用することによってこれまで十分に明らかになっていなかった現象を明らかにした研究である．日本においても複数種類・複数時点の社会調査を合併し，変数を統合させて分析を行った事例は少なくない（Fujihara and Ishida 2016; 平尾・太郎丸 2011; Mugiyama and Toyonaga 2022; 打越 2016）．数十年におよぶ不平等の趨勢，少数集団の詳細な特徴などに関わる問いは，単一時点・一回限りの社会調査データで明らかにすることは難しく，複数の社会調査データを組み合わせることによってはじめて明らかにすることができる．

しかしながら，これまで日本における既存社会調査データを統合した分析は，各研究者が個別に行うにとどまっている．この状況は以下の 3 点で望ましくない．第 1 に，通常研究者が社会調査データの分析で用いる変数には，たとえば性別，年齢，職業，学歴，婚姻状態，所得等，共通のものが多いにもかかわらず，各研究者が個別に統合データを作成する場合，本来必要でない多くの労力を割く必要がある．第 2 に，複数の調査データの統合は変数の作成に際して通常の分析以上に労働集約的であり，個別の研究者が行う場合，誤りが生じる可能性は無視できず，かつ結果の再現性も損なわれる．どのように変数を作成するべきかについてはしばしば単一の答えがない場合もあるが，データ構築に用いるコードを利用者間で共有し，複数の目を通すことによって，少なくとも明らかな誤りについてはそのリスクを減らすべきである．第 3 に，上記の作業の負担から既存社会調査データの統合には及び腰になりやすく，その結果，データさえあれば答えられるはずの多くの重要な研究課題が未解決のまま残されている．

上記の問題意識のもとで立ち上げられたのが「社会調査データの合併による二次分析研究の刷新」プロジェクトである．本プロジェクトの目的は 2 点ある．第 1 に，社会学者が中心となって実施してきた大規模社会調査データを統合（ハーモナイゼーション harmonization）するコードを作成することである．なお本稿では，各調査で聴取された回答に対して事後的に定義を揃え比較可能とする作業のことを指してデータ統合ないしハー

モナイゼーションと呼び，互換的に用いる[1]．第 2 に，構築されたデータを活用して，日本社会における不平等の趨勢と実態に関わる様々な記述的問いに答えることである．

とくに本稿では，主に 2023 年 5 月から現在にかけて行われた第 1 の目的に係るデータ統合作業の工程について記述する．本稿の目的は 3 点である．第 1 に，データ統合の手順を詳述することによって，本統合データを利用した実証研究論文のデータ作成過程に関する技術的な詳細をたどれるようにする．第 2 に，データ統合の手順を詳述することによって，将来同様の作業を行う場合に参考となる手順の一例を提供する．第 3 に，データ統合を通じて得られた今後の社会調査企画やデータの整備方法に関する示唆を明文化し，今後の調査実施やデータ利用環境の整備に関していくつかの提言を行う．

本稿の構成は以下のとおりである．第 2 節では，本プロジェクトにおいてどの社会調査データを統合データに含めたか，またその選定基準について説明する．第 3 節では，具体的な統合データ作成の手順について詳述する．第 4 節では，一連のデータ統合作業を通じて発見した示唆を記し，今後の調査企画や設計に向けての提言を行うとともに残された課題について述べる．第 5 節では本稿の内容をまとめ結びとする．

2　使用した調査データの一覧

表 1 には本プロジェクトでデータ統合に使用した調査の一覧を記した．2023 年度は社会階層と社会移動全国調査（Social Stratification and Mobility，以下 SSM），日本版総合的社会調査（Japan General Social Survey，以下 JGSS），全国家族調査（National Family Research of Japan，以下 NFRJ），東大社研・若年壮年パネル調査（Japanese Life-Course Panel Survey，以下 JLPS），社会階層と社会意識調査（Stratification and Social Psychology，以下 SSP），教育・社会階層と社会移動調査（Education, Social Stratification and Mobility，以下 ESSM）の統合データを構築した．2024 年度はここに NFRJ2018，職業移動と経歴調査（Occupational Mobility and Career History，以下 OMCH），世代とジェンダー調査（Gender and Generation Survey，以下 GGS）を，2025 年度は JGSS-2021H/2022H をそれぞれ追加し，全 41 種の調査における約 200 の変数をハーモナイズした．

[1] 本稿では，複数データを結合することで観察数を増やすことを指して合併といい，合併することも含め，聴取された項目をデータ間で比較可能なように事後コーディングを行うことを指して統合（harmonization）と呼ぶこととする．2023 年度のプロジェクト開始当初はこれらの概念が適切に整理できていなかったため，プロジェクト名には合併という語が含まれている．しかし，実際にはたんにデータを結合するということを超えて変数を比較可能なように再コーディングするのが作業の中心であるため，その内容を適切に反映する統合という語を用いることとした．

表１　合併した調査データの一覧

調査略称[1]	調査方法	対象者年齢	対象者出生年	有効回収率[5]	回答数	データ取得先	SSJDA番号
SSM1975A[2]	訪問面接	20–69	1905–1954	68.1	2,724[m]	SSJDA	0761
SSM1985A	訪問面接	20–69	1916–1965	61.03	1,239[m]	SSJDA	0762
SSM1985B	訪問面接	20–69	1916–1965	60.79	1,234[m]	SSJDA	0762
SSM1985F	訪問面接	20–69	1916–1965	67.89	1,474[f]	SSJDA	0762
SSM1995A	訪問面接	20–69	1926–1974	65.8	2,653	SSJDA	0763
SSM1995B	訪問面接	20–69	1926–1974	67.2	2,704	SSJDA	0763
SSM2005	訪問面接+留置	20–69	1935–1984	44.1	5,742	SSJDA	0764
SSM2015	訪問面接+留置	20–79	1935–1994	50.1	7,817	SSJDA	1508
NFRJ1998[3]	訪問留置	28–77	1921–1971	66.52	6985	SSJDA	0191
NFRJ2001[3]	訪問留置	32–81	1920–1969	69.5	3475[f]	SSJDA	0400
NFRJ2003[3]	訪問留置	28–72	1926–1975	63	6302	SSJDA	0517
NFRJ2008[3]	訪問留置	28–72	1936–1980	55.35	5203	SSJDA	0817
NFRJ2018[3]	訪問留置	28–72	1946–1990	55.05	3033	SSJDA	1733
JGSS2000	訪問面接+留置	20–89	1910–1979	64.9	2,893	SSJDA	1417
JGSS2001	訪問面接+留置	20–89	1911–1980	62.4	2,790	SSJDA	1418
JGSS2002	訪問面接+留置	20–89	1912–1981	62.3	2,953	SSJDA	1419
JGSS2003	訪問面接+留置	20–89	1913–1982	51.5	3,663	SSJDA	1420
JGSS2005	訪問面接+留置	20–89	1915–1984	50.5	2,023	SSJDA	1421
JGSS2006	訪問面接+留置	20–89	1916–1985	59.8/59.8	4,254	JGSSDDS	
JGSS2008	訪問面接+留置	20–89	1918–1987	58.2/60.6	4,220	JGSSDDS	
JGSS2009LCS	訪問面接+留置	28–42	1966–1980	51.1	2,727	JGSSDDS	
JGSS2010	訪問面接+留置	20–89	1920–1989	62.2/62.1	5,003	JGSSDDS	
JGSS2012	訪問面接+留置	20–89	1922–1991	59.1/58.8	4,667	JGSSDDS	
JGSS2015	訪問面接+留置	20–89	1925–1994	52.6	2,079	JGSSDDS	
JGSS2016	訪問面接+留置	25–49	1966–1991	50.8	968	JGSSDDS	
JGSS2017	訪問面接+留置	20–89	1927–1996	55.6	744	JGSSDDS	
JGSS2018	訪問面接+留置	20–89	1928–1997	54.3	1,916	JGSSDDS	
JGSS2017G	訪問留置	20–89	1927–1996	64	860	JGSSDDS	
JGSS2018G	訪問留置	20–89	1928–1997	62.7	678	JGSSDDS	
JGSS2021H	訪問留置	20–89	1931–2000	57.5/59.2	3522	JGSSDDS	
JGSS2022H	訪問留置	20–89	1932–2001	52.1/52.7	3145	JGSSDDS	
JLPS2007[4]	郵送訪問回収	20–40	1967–1987	40.4	4,800	SSJDA	PY140, PM140
JLPS2011[4]	郵送訪問回収	24–44	1967–1987	31	963	SSJDA	PY140, PM140
JLPS2019[4]	郵送訪問回収	20–31	1988–1999	41.3	2,051	SSJDA	PY140, PM140
SSPI2010	訪問面接	20–59	1950–1989	50.4	1,763	SSJDA	1374
SSP2015	CAPI	20–64	1950–1994	43	3,575	SSJDA	1375
ESSM2013	郵送訪問回収	30–64	1948–1983	60.3	2,893	SSJDA	1594
OMCH1981	訪問面接	25–69	1912–1956	70.9	4,255[m]	SSJDA	0133
OMCH1983	訪問面接	25–69	1914–1958	83.0	2,490[f]	SSJDA	0134
OMCH1991	訪問面接	25–69	1922–1966	80.0/82.7	2,401/2,481	SSJDA	0135
GGS2004	訪問留置	18–69	1934–1986	57.9/63.0	4,265/4,809	SSJDA	1075

[1] SSM：社会階層と社会移動全国調査（Social Stratification and Mobility Survey of Japan），NFRJ：全国家族調査（National Family Research of Japan），JGSS：日本版総合的社会調査（Japan General Social Survey），JLPS：東大社研・若年壮年パネル調査（Japanese Life-Course Panel Survey），SSP：階層と社会意識全国調査（Social Stratification and Psychology Survey），ESSM：教育・社会階層・社会移動全国調査（Education,

Social Stratification, and Mobility Survey of Japan)，OMCH：職業移動と経歴調査（Occupational Mobility and Career History Survey)，GGS2004：世代とジェンダー調査（あるいは結婚と家族に関する国際比較調査, Japan Generation and Gender Survey)．

(2) B 票は威信調査．SSJDA からは提供されていないため使用しない．

(3) 家族社会学会全国家族調査委員会での NFRJ の略称は NFRJ-98, NFRJ-S01, NFRJ-03, NFRJ-08, NFRJ-18 であるが，ここでは他の調査と名称を揃えるため若干変更した．

(4) JLPS は PY***と PM***でそれぞれ別々にサンプルを提供しているが，両者はまったく同じ調査を実施しており結合して用いることが推奨されている．そこでサンプルを結合したうえで，初回サンプル（2007 年からの対象者)，追加サンプル（2011 年からの対象者)，リフレッシュサンプル（2019 年からの対象者）をそれぞれ抽出し，これらをそれぞれ JLPS2007, JLPS2011, JLPS2019 と表記している．なお JLPS はパネル調査データであるが，クロスセクションデータとみなしてサンプルに追加する．

(5) /が入っている場合はそれぞれ前が男性，後が女性の回収率を示す.

(m) 男性のみ対象の調査.

(f) 女性のみ対象の調査.

上述のデータを選出した理由は 3 点ある．第 1 に，いずれも住民基本台帳または選挙人名簿からの層化多段抽出を行った無作為抽出標本を用いている．第 2 に，性別および年齢以外の条件によって対象者の限定を行っていない．たとえば特定地域などといった限定が設けられている場合，他の調査と結合の際にウェイトなどの追加的な論点を生ずるため，今回候補からは外している．第 3 に，社会階層・不平等現象の分析にとって有益な変数をある程度含んでいる，という条件を満たすものを対象とした．もちろん，たんに条件という観点のみからみればこの条件に合致する他の調査データがないわけではなく，今回選定した調査はあくまでプロジェクト参加者の主たる問題関心をカバーするものを中心としているに過ぎない．本稿の方法にもとづけば，他の調査についても同様に変数を作成することが可能である[2].

これらのデータを合併したサンプルサイズは 136,433（男性 65,236，女性 71,197）である．対象者の年齢は 18 歳から 89 歳までと幅広い．図 1 には参考として各調査の性・年齢別サンプルサイズを示した．すべてのデータが必要な変数を揃えているわけではないため，問題関心に即した変数を収集しているデータのみを使用する場合，サンプルサイズは

[2] 研究会内で他に候補に上がった調査データは以下のとおりである：「福祉と生活に関する意識調査（SPSC 調査)，2000」「福祉と公平感に関するアンケート調査，2005」「仕事と暮らしに関する全国調査，2003」「国際化と市民の政治参加に関する世論調査，2009」「国際化と市民の政治参加に関する世論調査，2013」「国際化と市民の政治参加に関する世論調査，2017」「中高年者の生活実態に関する全国調査，2010」「SSJDA Panel Wave 1, 2021 年 2-3 月調査」「SSJDA Panel Wave 3, 2022 年 2 月調査（2022 年サンプル)」.

これより小さくなる．年齢や就業状態といった変数で対象者に限定をかける場合には，当然ながら実際の分析に用いられるサンプルサイズはこれより小さくなる．

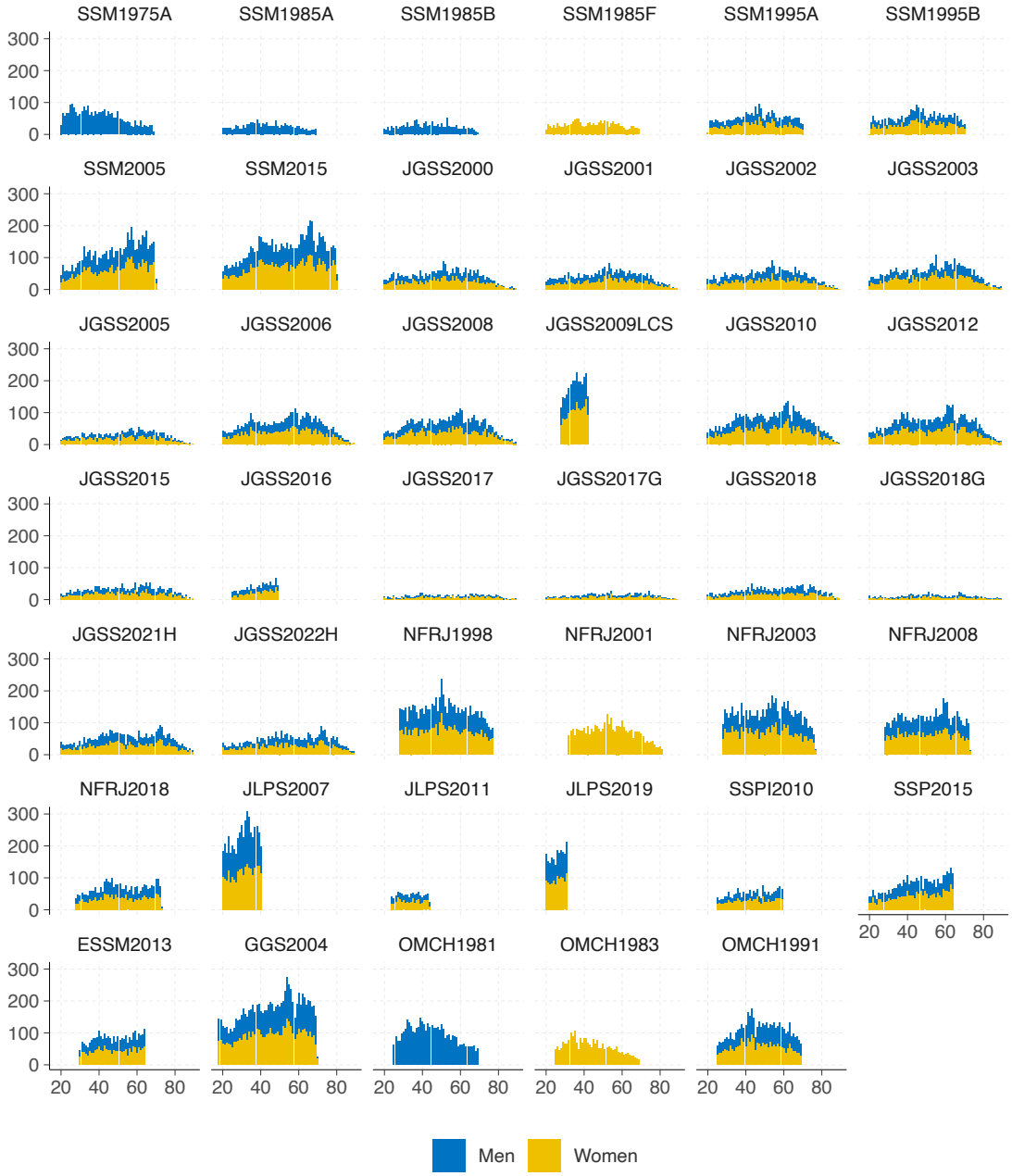

図１　調査・性・年齢別サンプルサイズ

なお本プロジェクトには，異なる調査名（たとえば SSM と JGSS など）の調査の統合だけでなく，同一調査の異なる時点間の比較が可能なようにデータを整備するというねらいもある．たとえば SSM，NFRJ など，時系列比較を念頭において設計された調査においては，調査時点間の比較可能性を担保するように調査項目が設けられている．しかし，提供

データにおいて変数名や値のコードが揃っているわけではないため，個別の研究者がハーモナイゼーションのための作業，具体的には異なる調査時点・調査間を横断した変数名の統一，値の再定義とリコードが必要となる．したがって本プロジェクトで作成したコードは，同一調査の時点間比較を行いたいという場合にも有益である．

3　作業手順

3.1　データの利用申請

表 1 のデータ取得先列に示されているように，本統合データは SSJDA からダウンロードしたデータと JGSSDDS からダウンロードしたデータから構成されている．以下では本稿執筆時点におけるデータ利用申請手順を示す．まずデータを申請するにあたり，本プロジェクトでは図 2 のとおりメンバーで共通の研究題目と研究計画を利用した．【】内には各自の具体的な研究内容を記入した．

研究題目：社会調査データの合併による二次分析研究の刷新

研究計画：本研究の目的は，社会学者が中心となって実施してきた大規模社会調査データを統合するコードを作成・共有し，構築されたデータを用いて，これまで十分に答えられなかった不平等に関わる問いに答えることにある．統合データの構築によって，必ずしも公的統計では捉えられなかった社会の実態およびその変化を従来よりも正確に記述すること，および単一の社会調査ではサンプルサイズが不足するゆえに分析することのできなかった層に焦点を当てた分析が可能となる．本研究は積極的なコードの共有を通じて誤りが生じるリスクを減らし，分析の再現性を高めることに資する．さらに，異なる社会調査データ間での結果の異同や，より効果的な調査項目や選択肢の設定の仕方を示すことも可能となる．申請者は本プロジェクトにおいて【】に関する分析を実施する予定である．

図 2　研究計画

SSJDA からのデータ利用申請手順は以下の通りである．まず，東京大学社会科学研究所附属社会調査・データアーカイブ研究センターの SSJ データアーカイブ（https://csrda.iss.u-tokyo.ac.jp/infrastructure/）の SSJDA Direct の ID を作成する．ID 作成後，SSJDA Direct にログインし，統合データに利用するデータの利用申請を行う．その際に自らの研究計画を入力する必要がある．利用承認の許可が東京大学社会科学研究所附属社会調査・データアーカイブ研究センターより得られた後，データをダウンロードすることができる．なお，JLPS や ESSM など一部のデータは，居住都道府県等に関する情報が別データとして提供されている．これらのデータの使用を希望する場合には別途利用申請が必要である．

JGSSDDS からのデータ利用申請手順は以下のとおりである．まず，大阪商業大学 JGSS 研究センター（https://jgss.daishodai.ac.jp/index.html）の JGSS データダウンロードシステム（JGSSDDS）のユーザーアカウントを登録する必要がある．アカウント作成後，JGSSDDS にログインし，データの利用申請を行う．利用承認の許可が大阪商業大学 JGSS 研究センターより得られた後，データをダウンロードすることができる．なお，居住都道府県等に関する情報が別データとして提供されている．これらのデータの使用を希望する場合には別途利用申請が必要である．

3.2　環境構築

データ統合に際してはメンバー共通で統計ソフト Stata を利用した．まずは，自分の PC に Stata をインストールする必要がある．合併データは大量の変数（列）を含むため，多くの変数（列）を有するデータを読み込むことのできるバージョンである Stata SE もしくは Stata MP を用意する必要がある．本稿執筆時点での最新バージョンは Stata 19 であるが，Stata 16 以降に開発された関数（たとえば frame 関数など）は使用していない．

本プロジェクトでは，分析に用いる Stata の do-file などを Dropbox で共有した．まずは，各自が Dropbox のデスクトップアプリケーションを PC にインストールする．その後，メンバー同士で Dropbox の共有フォルダを作成し，そこに do-file を入れた code という下位フォルダを作成した．code フォルダには変数ごとに分割された do-file と，それらをまとめて実行する _master do ファイルが置かれている．共有 Dropbox フォルダは以下のような構造からなる．

- code：最新バージョンの Stata コード一式をアップロードしている．
- code-backup：定期的に code フォルダをコピーし，バックアップをとる．
- codebook：各調査で作成されたコードブックや報告書，各調査の値ラベル（SSJDA や JGSSDDS より提供されるテキストファイルや Excel ファイル），調査票の pdf ファイルをアップロードしている．
- document：毎回の研究会の報告資料および二次分析研究会のために使用した申請書などをアップロードしている．
- literature：データ統合に関連する先行研究や論文をアップロードしている．
- manuscript：本稿を含め執筆中の論文をアップロードしている．

3.3　データの連結（append）

自身の PC のローカル環境にデータを入れるフォルダを作成して，そこに SSJDA および JGSSDDS からダウンロードしたファイルをフォルダごと入れる．この際，ダウンロードしたフォルダは名前を変えないようにする．データを入れたフォルダは図 3 のようになる．

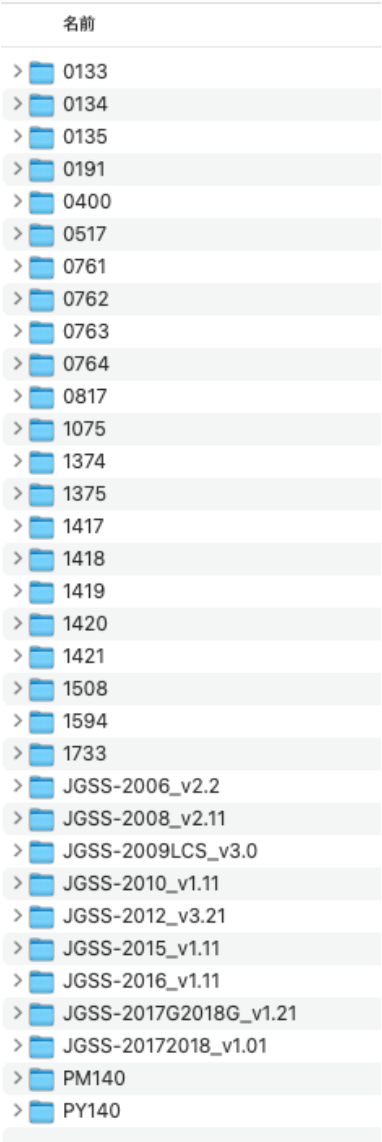

図 3　データをダウンロードしたフォルダ

データ連結の手順は以下のとおりである．まず，code/_master.do ファイルを開く．上述した通り，自身のローカル環境上に個票データの入ったフォルダを入れたうえで，当該フォルダの絶対パスをコピーし，自分の名前の入った行に貼り付ける．その後，自分の絶対パスを入力した行を実行する．たとえば以下のような行になる．

global datadir "/Users/mugi/Documents/Research/ASSD-local/data" // mugiyama

全データ合併ファイルを作る際には，以下のようにする．第 1 に，上記の自分の行に当たる行のコードを実行する．第 2 に，do "code/1_append/preparesurvey.do"および do "code/1_append/appenddata.do"の行を実行する．うまく実行できた場合には，ローカル環境

上に fulldata.dta というファイルが作成される．これが，すべての変数を連結したデータセットである（以下，これを連結データとよぶ）．ファイルサイズが 4GB 近くあるので，自身の PC の容量に注意する．この時点では，fulldata.dta はたんにすべてのデータを append コードによって行方向に結合しただけのデータにすぎない．そのため，たとえば q2 という変数は調査間で同じ調査項目を指すわけではない．

変数作成作業に使用するため，当該データがどの調査から得られたものかを示す変数 a_survey（string 変数で，SSM1995A, JLPS2007 などの値をとる）と，調査種類を示す変数 a_surveytype（string 変数で，SSM, NFRJ, JGSS, JLPS, SSP, ESSM, OMCH, GGS の 8 種類の値をとる）を作成している．表 2 にそれぞれの値と分布を示す.

表 2　a_survey（調査変数）と a_surveytype（調査種別）の度数分布表

[a] a_survey（調査変数）の分布

	N	%
ESSM2013	2,893	2.12
GGS2004	9,074	6.65
JGSS2000	2,893	2.12
JGSS2001	2,790	2.04
JGSS2002	2,953	2.16
JGSS2003	3,663	2.68
JGSS2005	2,023	1.48
JGSS2006	4,254	3.12
JGSS2008	4,220	3.09
JGSS2009LCS	2,727	2.00
JGSS2010	5,003	3.67
JGSS2012	4,667	3.42
JGSS2015	2,079	1.52
JGSS2016	968	0.71
JGSS2017	744	0.55
JGSS2017G	860	0.63
JGSS2018	1,916	1.40
JGSS2018G	678	0.50
JGSS2021H	3,522	2.58
JGSS2022H	3,145	2.31
JLPS2007	4,800	3.52
JLPS2011	963	0.71
JLPS2019	2,051	1.50
NFRJ1998	6,985	5.12
NFRJ2001	3,475	2.55
NFRJ2003	6,302	4.62
NFRJ2008	5,203	3.81
NFRJ2018	3,033	2.22
OMCH1981	4,255	3.12
OMCH1983	2,490	1.83
OMCH1991	4,879	3.58
SSM1975A	2,724	2.00
SSM1985A	1,239	0.91
SSM1985B	1,234	0.90
SSM1985F	1,474	1.08

[b] a_surveytype（調査種別）の分布

a_surveytype	N	%
ESSM	2,893	2.12
GGS	9,074	6.65
JGSS	49,105	35.99
JLPS	7,814	5.73
NFRJ	24,998	18.32
OMCH	11,624	8.52
SSM	25,587	18.75
SSP	5,338	3.91
Total	136,433	100.00

SSM1995A	2,653	1.94
SSM1995B	2,704	1.98
SSM2005	5,742	4.21
SSM2015	7,817	5.73
SSP2015	3,575	2.62
SSPI2010	1,763	1.29
Total	136,433	100.00

なお，連結データに含まれている変数はあらかじめすべて小文字に変換している．これは，変数名を小文字にするか大文字にするか各調査で統一的な基準がないことによる混乱を防ぐための処置である．また，居住地などの別途申請が必要な秘匿情報は fulldata.dta には含んでいない．当該変数を使用する場合は，自身のローカル環境で別途接合する必要がある．

ここで重要なのは，上記のワークフローにおいては，あくまで Dropbox 上で共有されているのはコード（do-file）のみであり，個票データそのものは決して Dropbox 上にアップロードしないということである．すなわち，個票データそのものは個別のローカル環境に保存されており，したがって実行結果も個別のローカル環境上に展開されることになる．

3.4　変数のハーモナイゼーション

まずデータを連結するという手順から明らかなように，変数の作成に際しては，調査ごとに統合用の変数を作ったあとにそれらを連結するというのではなく，最初にデータを連結した後に，調査をまたいだ統合変数を一つずつ作成していく，という方法で変数のハーモナイズを進める．この方法の利点は 2 つある．第 1 に，通常研究を進めるプロセスに沿っている．研究者が何らかのデータ分析を行いたいと考える場合，はじめに関心のある変数を考え，次にその変数を聴取している調査データを探すという手順を取るはずである．今回の統合作業の手順はこれを再現したものであり，理解しやすい．第 2 に，変数ごとに do-file を分割しやすく，したがって変数ごとに担当者を分けて別々に作業を進めやすい．この場合，すべての変数ができあがっていないとしても，できるところから分析を進めることが可能になる．今回の作業の手順であれば，優先度の高い変数から順にハーモナイズを進めることも可能である．

作業をはじめるにあたり，メンバーから研究計画を集めると同時に，使用予定の変数をリストアップし，少なくとも一名からリストアップされた変数についてのみ統合変数を作成することとした．仮に変数を作成したとしても，誰も使う人がいなければ徒労に終わるからである．リストアップされた変数について，それを使用する，あるいは専門に関連するメンバーに担当を割り振った．担当者は，それぞれの変数について統合のための do-file コードを作成した．

具体的な各担当者の作業は次のとおりに進められた．第 1 に，担当変数（たとえば現職の企業規模）について，どの調査に当該変数が含まれているかを確認した．その際には，作成担当者，調査ごとの各変数の有無，進捗状況が確認できる「統合変数一覧表」という Excel ファイルを作成し，すべての作業者間で共有した（図 4 参照）．この表には，「備考」欄が作成され，調査ごとの変数設計の微妙な違いや統合にあたっての注意点がまとめられた．調査データを新たに追加した場合にはこの表も更新する．具体的には，追加した調査データに自分の担当変数が含まれているかどうかを確認するとともに，共有事項などがあればこれを更新した．

	A	B	C	D	F	H	I	J	K	L	M	N	O	P
1	order（作業	変数名	変数ラベル	担当候補	進捗状況	SSM1975A	SSM1985A	SSM1985B	SSM1985F	SSM1995A	SSM1995B	SSM2005	SSM2015	JGSS2000
	1	a_survey	調査フラグ	麦山	完了	1	1	1	1	1	1	1	1	1
	2	a_surveytype	調査種フラグ（SSM, NFRJ, JGSS, JLPS, SSP）	麦山	完了	1	1	1	1	1	1	1	1	1
1	3	a_year	調査年	麦山	完了	1	1	1	1	1	1	1	1	1
2	4	a_age	年齢	麦山	完了	1	1	1	1	1	1	1	1	1
3	5	a_birthyear	出生年	麦山	完了	0	0	0	0	1	1	1	1	1
4	6	a_birthyearfy	出生年度	麦山	完了	0	0	0	0	1	1	1	1	0
5	7	a_birthmonth	出生月	麦山	完了	0	0	0	0	1	1	1	1	0
6	8	a_sex	性別	麦山	完了	1	1	1	1	1	1	1	1	1
7	9	a_currentwork	現職：就業状態	吉田	完了	1	1	1	1	1	1	1	1	1
8	10	a_currentstatus	現職：従業上の地位	吉田	完了	1	1	1	1	1	1	1	1	1

図 4　統合変数一覧表の形式（一部抜粋）

第 2 に，実際の変数作成作業を行った．担当変数ごとに分割された do-file を用意し，その do-file を編集することで異なる調査データの変数を統合した．各変数の do-file は，たとえば「a_age.do」や「a_sex.do」といったように，それとわかるように命名した．これらの do-file 名は変数名と対応しており（たとえば年齢であれば a_age，性別変数であれば a_sex），変数名は図 4 のように統合変数一覧表で共有された．

実際の作業にあたり，各変数の do-file の最上部には，（1）何の変数のコードか，（2）作成担当者の氏名，（3）最終更新日，（4）備考をコメントアウトして記載した．これにより，作業者が年度の変化によって変わっても，作業が円滑に引き継がれた．特に，do-file 内の（4）備考には，変数作成にあたって気づいたこと，共有しておくべきことなどを記載し，作業担当者以外で疑問に思うことがあればこの備考を参照することとした．場合によっては，Slack や研究会内で議論し，対応方法を検討した．

変数作成用の do-file はすべて Dropbox 上で共有した．また，_master.do には各変数の作成コードを直接実行できるようにコードを記載した（たとえば do "a_sex.do"など）．このよ

うにすることによって，_master.do に書かれたコードを実行するだけで，個別の do-file を開かなくともすべての変数を作成することができるようにしている[3].

3.5　ハーモナイゼーションにあたっての細かな対応

異なる調査データの変数を統合する作業では，調査間で異なる設問の選択肢を，どのように統合するかが問題となる．統合変数作成にあたっての基本方針は，できるだけ元のカテゴリを残すようにして統合するということである．この場合の方法は 2 つある．第 1 に，元のカテゴリをまったく変更することなく，無回答やわからないなどの選択肢のみを統合するというパターンである．たとえば支持政党に関する変数の場合，どの調査にも含まれている政党は「自民党（または自由民主党)」「共産党」のみであり，そのときどきの政党の離合集散に応じて選択肢は変化する．このような場合には，「その他の政党」などとまとめることなく，元々の選択肢に設けられていた政党名を残している．

第 2 に，可能な限り元々の情報を残しながらカテゴリの統合をはかるという場合である．たとえば，現職従業上の地位（雇用形態）に関する質問項目の場合，調査によっていわゆる非正規雇用に含まれる雇用形態として何を含めているのかが異なる一方で，一般従業員，経営者・役員，自営業主，家族従業者，内職についてはほとんどの調査で常に含まれている．このような場合には，パート，アルバイト，臨時雇用，契約社員，派遣社員，嘱託社員などをまとめて「非正規雇用」というカテゴリに統合し，その他についてはもとのカテゴリを保持している．

ただし，すべての調査で統一的なカテゴリを作成するのが難しい場合もある．このような場合は，いくつかの種類の変数を作成するという対応も行っている．典型的な例が従業員数（企業規模）である[4]．従業員数に関する調査項目で使用される選択肢は，調査ごとに異なっている．表 3 は，調査ごとに異なる現職の従業員数に関する質問の選択肢をまとめたものである．たとえば SSM と JGSS の従業員数変数を統合する場合には，JGSS の「1000～1999 人」「2000～9999 人」「1 万人以上」のカテゴリをまとめ，SSM における

---

[3] 本稿執筆時点では，表 1 に記載されたすべての調査データを利用することを前提としている．そのため，仮にいずれかの調査データが欠けている場合，Stata コードを実行しても途中でエラーとなってしまう．具体的には，Stata の replace 関数などで既存の変数を参照する場合，当該変数がない場合にはエラーとなる．これを避けるためには，変数作成に関連する generate, replace, recode の各関数の直前に capture という関数をあらかじめ入れておく（capture を入れた場合，当該コードの実行結果がエラーになる場合にはそれを無視して次のコードを実行する)，そもそものコードの構造を変える（変数ごとに do-file を分割するのではなく，調査ごとに do-file を分割する)，といった方法が考えられる．この点は今後取り組むべき課題として記しておく．

[4] 具体的には，現職の従業員数，初職の従業員数，配偶者の仕事の従業員数，父親の主な仕事の従業員数，15 歳時の父親の仕事の従業員数，母親の主な仕事の従業員数，15 歳時の母親の仕事の従業員数である．

「1000 人以上」というカテゴリに統合すれば良い．一方で，NFRJ の従業員数変数を他の調査と統合する際には，NFRJ のカテゴリに合わせる必要がある．この場合，SSM や JGSS における「1 人」「2〜4 人」「5〜9 人」のカテゴリはすべて NFRJ の「1〜9 人」というカテゴリに，「10〜29 人」「30〜99 人」のカテゴリはすべて NFRJ の「10〜99 人」というカテゴリに，といったようにまとめられてしまう．もし分析者が NFRJ を使用しない場合のであれば，NFRJ の分類に合わせて作成された統合変数は不必要にカテゴリを粗くしたものとなってしまう[5]．そこで，「a_currentfirmsize1」と「a_currentfirmsize2」という 2 つの従業員数変数を作成した．これによって，NFRJ を分析に含める場合には，NFRJ の区分に合わせた「a_currentfirmsize1」を使用し，そうでなければ詳細版である「a_currentfirmsize2」を使用するというふうに使い分けることができる．

表 3　現職の従業員数に関する統合変数のカテゴリ区分

各調査で使用されている選択肢				作成した統合変数	
SSM, SSP	JGSS	JLPS	NFRJ	a_currentfirmsize1	a_currentfirmsize2
1 人	1 人	1 人	1〜9 人	1〜9 人	1 人
2〜4 人	2〜4 人	2〜4 人			2〜4 人
5〜9 人	5〜9 人	5〜9 人			5〜9 人
10〜29 人	10〜29 人	10〜29 人	10〜99 人	10〜99 人	10〜29 人
30〜99 人	30〜99 人	30〜99 人			30〜99 人
100〜299 人	100〜299 人	100〜299 人	100〜299 人	100〜299 人	100〜299 人
300〜499 人	300〜499 人	300〜999 人	300〜499 人	300〜999 人	300〜499 人
500〜999 人	500〜999 人		500〜999 人		500〜999 人
1000 人以上	1000〜1999 人	1000 人以上	1000 人以上	1000 人以上	1000 人以上
	2000〜9999 人				
	1 万人以上				
官公庁	官公庁	官公庁	官公庁	官公庁	官公庁
わからない	わからない	わからない		わからない	わからない

第 3 に，例外的な対応を行ったケースとして，個人年収，配偶者年収，世帯年収に関する変数がある．今回対象とした調査ではこれらの年収はいずれも「100〜125 万円」などといった区間に区切った選択肢から当てはまるものを選ぶ形式で年収が聴取されていた．他方，ほとんどの調査において区間の区切りは厳密には揃っておらず，もしすべてを揃えようとすれば極めて粗くほとんど実質的に使用することのできないカテゴリとなってしまう．そこで，区間の中間値をとって連続変数に変換したうえで変数を統合した．このよう

[5] たとえば EGP カテゴリを作成したり（鹿又ほか 2008），従業員数が少ない企業に勤めている場合，経営者・役員を自営業主に変換する（2015 年社会階層と社会移動調査研究会 2018）といった処理を行う場合には，従業員数が 1 名かそれ以上か，30 名未満かそれ以上か，といった閾値が重要となる．NFRJ 等で採用されている従業員数の区分はこれらを区別できない．

に区間の中間値を取る処理は所得を連続変数として分析する場合に標準的に用いられている方法であり，許容されるであろうと判断した．

なお，一見単純にみえる出生年などの項目を統合する作業でも細心の注意が必要である．たとえば回答者の出生年の変数を作成する場合は，西暦または和暦および出生年を参照して出生年を計算するか，この変数が含まれていない調査の場合には調査実施年 – 回答者の年齢 – 1 を行って求める．単年度の調査データを用いる場合，コーディングにおけるミスはあまり生じない変数であるが，複数のデータをハーモナイズする際には，繰り返される作業で意外にコーディングミスが生じやすい．たとえば SSM2015 では出生年に関わる項目として西暦（q1_2_1），元号（q1_2_2），生年（q1_2_3）という 3 つの変数が含まれている．いずれにも値が入っているというわけではなく，回答者が西暦を選択した場合には q1_2_1 にのみ回答が記録され，回答者が元号を選択した場合には q1_2_2 と q1_2_3 のみに回答が記録されるという形式になっている．3 つの変数の中身を確かめることなく q1_2_1 のみを用いてコーディングすると，半数以上のサンプルが欠損値扱いになってしまう．

このような問題の発生を防ぐため，変数の作成担当者を年度ごとにできるだけ交代するとともに，自身の分析で変数を用いる際にはコーディング内容を確認し，クロスチェックを行うように努めた．その過程で修正点が見つかった場合には，do-file に日付とともに修正内容を簡潔に記録することで，コーディングの変更時期と内訳が明確に把握できるようにした．

3.6　追加変数

そのほか，変数の統合とは別に，統合した変数を参照してさらに追加の変数を作成するという作業も行った．実際の分析においては元々設けられたカテゴリをまとめたり値を付与することによって使用する変数を作成することはしばしばある．もしこうした変数が一般的に同一の手続きにもとづいてなされるのであれば，そうした変数についてもあらかじめ準備しておくほうが効率的であろう．このような変数をここでは追加変数とよぶ．

具体的には，以下の追加変数を作成した．本人初職・本人現職・父職・母職など職業関連変数に対して，EGP 階級分類（鹿又ほか 2008），Oesch の階級分類（Oesch 2006），SSM 職業 8 分類（1995 年 SSM 調査研究会 1995），JSEI および JSSI（Fujihara 2020），職業威信スコア（直井 1979），日本版 O-NET 数値指標（麦山・田上 2024）を付与した．また，調査年に対して消費者物価指数を付与したほか，最後に通った学校の卒中退の別を考慮した中退を考慮した学歴変数についても作成した．今後もたとえば等価世帯所得など，さらに一般的に使用頻度の高い追加変数についても作成していくことが望まれる．

4　作業を通じて得られた示唆

4.1　調査間の差異の発見とそれに対する対応

上記のように統合作業を進めていく過程でもっとも早く気づいた点は，同様な事象を問う設問であっても，質問文の文言や選択肢が調査間に相違がみられることであった．とくに態度や社会意識など主観的側面の強い設問でその傾向が強い．「主観的健康」と「階層帰属意識」の設問項目を例に取り上げて説明する．主観的健康・階層帰属意識の設問の文言と選択肢をまとめたものを表 4 に示す.

表 4 主観的健康と階層帰属意識の質問文・選択肢一覧

設問項目	調査名	実施年	問番号	質問文	選択肢
主観的健康	SSM	2005 2015	b25 dq20	あなたは，現在のご自分の健康状態をどう思いますか.	1 とてもよい 2 まあよい 3 ふつう 4 あまりよくない 5 わるい
	JLPS	2007 2011 2019	zq25 dq11 lq15	あなたは，自分の健康状態についてどのようにお感じですか.	1 とても良い 2 まあ良い 3 普通 4 あまり良くない 5 悪い
	JGSS	2000, 2001, 2002, 2003, 2005, 2008, 2009LCS, 2010, 2012, 2015, 2016, 2017, 2018, 2021H, 2022H	op5hlthz	あなたの現在の健康状態は，いかがですか.	1 良い 2 . 3 . 4 . 5 悪い
	JGSS	2006 2012 2017G 2018G	op5hlthe	あなたの現在の健康状態は，いかがですか.	1 非常に良い 2 . 3 . 4 . 5 非常に悪い
	SSP	2015	q15_2	あなたは，現在のご自分の健康状態をどう思いますか.	1 よい 2 まあよい 3 ふつう 4 あまりよくない 5 わるい
	NFJR	1998 2003 2008 2018	q22 r14 r11 r35	あなたのこの 1 年間の健康状態は，おおむね，いかがでしたか.	1 たいへん良好 2 まあ良好 3 どちらともいえない 4 やや悪い 5 たいへん悪い
5 段	JGSS	2000, 2001, 2002, 2003, 2005, 2006, 2008, 2009LCS,	op5levk	かりに現在の日本の社会全体を，以下の 5 つ	1 上 2 中の上

階階層帰属意識		2010, 2012, 2015, 2016, 2017, 2018, 2021H, 2022H		の層にわけるとすれば，あなた自身は，どれに入ると思いますか.	3 中の中 4 中の下 5 下
	SSM	1975A 1985A, B 1985F 1995A 1995B 2005 2015	a18a q13a q17a q34 q26_1 a08, b23 q46	かりに現在の日本の社会全体を 5 つの層に分けるとすれば，あなた自身はこのどれに入ると思いますか. [6]	1 上 2 中の上 3 中の下 4 下の上 5 下の下
	ESSM	2013	q48		
	SSP	2010I 2015	q13_1 q13_1	かりに現在の日本の社会全体を，このリストに書いてあるように 5 つの層に分けるとすれば，あなた自身はこのどれに入ると思いますか.	
	OMCH	1983	q36	かりに現在の日本の社会全体を，次にあげる 5 つの層に分けるとすれば，あなた自身は，どの層に含まれると思いますか.	1 上位に含まれる 2 どちらかといえば上位 3 中位 4 どちらかといえば下位 5 下位に含まれる

主観的健康（self-rated health）は多くの調査で採用されているが，設問のワーディングや選択肢の表現，測定尺度は調査ごとに異なる．SSM では「あなたは，現在のご自身の健康状態をどう思いますか.」，JGSS では「あなたの現在の健康状態はいかがですか.」と尋ねる一方，NFRJ では「あなたのこの 1 年間の健康状態は，おおむね，いかがですか」と 1 年間に限定した質問になっているなど，文言が異なる．選択肢も調査間で異なり，SSM は「1 とてもよい～5 わるい」の 5 つからなる尺度を採用している一方，JGSS では中間値については文言を示さずに「1 良い～5 悪い」とする尺度を採用し，NFRJ は「1 たいへん良好～5 たいへん悪い」という尺度を採用している．すなわち，調査間で設問の文言も選択肢の幅も異なるということである．

[6] 調査年度によって質問文に若干の相違が見られる．1975 年・1985 年・1995 年調査では，「かりに現在の日本の社会全体を，このリストにかいてあるように 5 つの層に分けるとすれば，あなた自身はこのどれに入ると思いますか.」と尋ねている．一方，2005 年・ESSM2013 年調査では，「かりに現在の日本の社会全体を 5 つの層に分けるとすれば，あなた自身はこのどれに入ると思いますか.あなたの気持ちにいちばん近い番号をひとつ選び，◯をつけてください」という質問文になっている．

階層帰属意識においても同様に違いがみられる．SSM，SSP，ESSM の選択肢は「1 上，2 中の上，3 中の下，4 下の上，5 下の下」であるが，JGSS は「1 上，2 中の上，3 中の中，4 中の下，5 下」というように，5 つの選択肢の中で「中」のカテゴリが 3 つ含まれる選択肢が採用されている．OMCH では「1 上位に含まれる，2 どちらかといえば上位，3 中位，4 どちらかといえば下位，5 下位に含まれる」のように他とかなり異なる選択肢が採用されている．以上のように，測定しようとする事象は共通しているように見えても，各調査が実際に測定している概念や分散は同一でない可能性がある．

この問題には統一の対処方法があるわけではないが，本プロジェクトでは以下のとおり対応した．まず，質問文においては，各調査間における文言の相違は見られるものの，設問の核心となるキーワードに値する表現（「（調査回答者の）健康状態」や「日本社会を 5 つの層に分ける…あなた自身はどこに入る…」など）はすべての調査において共通して盛り込まれていることから，同じような項目を測っているとみなしても大きい問題は生じないと判断した．選択肢に関しては，それぞれの尺度は異なるものの，こうした尺度の順序性を前提として，多くの先行研究はこれらを連続変数として扱っていることが多いことを踏まえ，これらの尺度を相対的な健康の良し悪しや階層的地位の高低を表すものとし，基本的には連続変数に変換するコーディングを行った．こうした変数を使用する際には，質問文の文言と選択肢ラベルの間に相違があることを十分に把握して使用する必要があり，結果を解釈する際も，こうした事情を踏まえた慎重な対応が求められる[7]．

このような設問間の質問文・選択肢の違いは，意識関連設問に限ったものではない．多様な調査データを選出して統合作業を行っているだけに，多くの項目で質問文や選択肢の相違が確認された．3.5 項で述べた従業上の地位や従業員数はその一例であるが，そのほかにも，学歴を尋ねる設問の選択肢に専門学校（専修学校専門課程）は含まれているか，別の質問として設けられているか，あるいは尋ねられていないか，きょうだい数として数えるのは対象者が 15 歳のときに生きていた者のみか，それともすでに亡くなったものも含めて回答するのか，「食事の用意」の頻度と「夕食の用意」の頻度のどちらを尋ねているか，など対応に注意を要した変数は多岐におよぶ．同じ調査であっても，調査時点によって選択肢の種類や個数，あるいは尋ね方が異なるなどの違いがみられた．

以上の点については，3.4 項で述べた通り，各変数の担当者が各調査の質問文とワーディングを精査し，どのように統合するかを各自で検討のうえ，作業の過程と方法を do-file の

[7] なお，このように値の間隔や個数が異なる変数を分析するための手法も開発されている．M. Gangl（2025）は，値の間隔や個数が異なる変数を従属変数とする際，質問種別ごとに閾値が異なることを仮定した一般化順序ロジットモデルを用いることで独立変数の係数を比較可能にする方法を提案している．他方，独立変数として用いる場合は，調査ダミーとの交互作用項を投入することによってある程度対処することが可能と考えられる（斉藤・吉田・下瀬川 2024）．

コーディングとともに記録にしておくことを基本方針としたことが各自の作業や実際の分析において有用であった．これを活用することで，最終的な分析の方針については各自で判断を行えるようにした．

4.2　共同作業の意義

第 3 節で詳述した通り，本プロジェクトでは変数ごとに 1 人の担当者を設定し，分担して統合変数を作成した．作業を進める中で，1 人の研究者がすべての変数をクリーニングする場合と比べて，こうした共同作業には時間や労力の削減以上の利点があることが実感できた．

第 1 に，作業過程の透明化につながる．本プロジェクトでは，いつ，誰が，どの変数を，どのようにクリーニングしたのか，あるいは do-file をどのように変更したのかについて，do-file の冒頭や Excel ファイルにすべて記録した．あわせて，do-file を担当者以外のメンバーが確認することもあるため，欠測値のハンドリング方法などを可能な限り統一し，可読性の高い do-file の作成も心がけた．こうした透明化は，研究者個人のプロジェクトでも原理的には可能であるが，時間的・労力的な制約から，現実的には難しいことも少なくない．データ統合作業を共同で実施したことは，こうした作業の透明化に大きく寄与した．

第 2 に，変数のクリーニング・標準化における正確性を高める．本プロジェクトでは，1 つの変数につき担当者を 1 人設定していたが，実際には，新年度での担当者の交代や，データセット追加に伴う do-file の追加・修正，自身の研究テーマで使用する変数を確認するなど，担当者以外のメンバーが do-file を確認する機会も多かった．その過程で，do-file の訂正や改善が少なからず発生した．その中には，タイプミスなど明確な誤りもあれば，標準化の方法について判断が分かれたため，メンバー間で議論した上で do-file を修正した箇所もある．かりに研究者個人でデータ統合作業を行っていた場合は，どちらのケースも認識されないまま，データ分析が行われていた可能性が高い．

これら 2 つの利点は，研究成果の再現性向上にもダイレクトに貢献する．再現性にはいくつかの種類があるが，ここではデータと分析手法が同一の場合に，同一の分析結果が示されることを担保する Verifiability への貢献が期待できる（Freese and Peterson 2017）．今後統合データの分析を進める過程で，さらなる誤りや改善点が見つかった場合でも，高度に透明化された作業プロセスは，原因の特定やその修正を容易にする．

第 3 に，各種社会調査の特徴や異同に関する知見の蓄積に資する．計量分析を用いる社会学者の多くは，特定の社会調査プロジェクトへの参加経験や，個別の社会調査データを用いて論文を執筆した経験をもつ．一方で，今回統合した社会調査（表 1）のすべてについて，その概要や調査項目に精通している人はきわめて限られている．そもそも，個々の研究者が関心をもつ調査項目について，それがどのような社会調査に，どのような聴取方法で含まれているかを網羅的に把握している者は少ないのが実情である．このプロジェク

トでは，メンバー間での議論や，統合変数一覧表（図 4 参照）の作成・共有により，調査ごとの特徴やその差異に関する知見の共有に努めた．このことは，個々の研究者の研究構想・計画を洗練させるだけでなく，今後の大規模社会調査の設計や実施における有用な示唆を得ることにもつながった．これについては次項（4.3 項）で詳しく述べる．

4.3　今後の社会調査企画・設計に向けての提言

本プロジェクトを通して，データの統合作業だけでなく，今後の社会調査企画・設計・実施についてもいくつかの示唆が得られた．以下に提言として整理するが，これらはいずれも，複数の社会調査データを統合して分析するという目的のもとでの提案であることを付言しておく．

第 1 に，調査項目の画一化である．設問のワーディングや選択肢について，特段の理由がなければ，過去の調査あるいは類似の社会調査と同一の文言を用いることが望ましい．3.5 項や 4.1 項で説明してきたように，変数のハーモナイゼーションにおいて，異なる設問文や選択肢をどのように扱うかには多くの労力を要する．もっとも情報量の少ない調査に揃えるため，本来利用できるはずの情報量を縮約してハーモナイズしなければならない場面も少なくない．表 3 で示した企業規模の例に依拠すると，かりに NFRJ の選択肢が SSM や JGSS の選択肢に準拠していれば，NFRJ を分析データに含める場合でも，より詳細な情報をもつ変数「a_currentfirmsize2」を使用できていたことになる．当然，その調査固有の目的により，他の調査とは異なる設問・選択肢を採用することはあるだろうし，紙幅の関係で同一の文言を断念することもありうる．しかし，こういった事情がない限りは，なるべく他の調査と同一の設問・選択肢を使うことが，比較可能性および測定の一貫性の面で推奨される[8]．

第 2 に，他の調査回や調査データとの統合がなるべく容易になるようにデータを公開することも求められる．JGSS のように，異なる調査年の間で，同一の項目に対応する変数名を揃えるといった工夫がその典型である．他にも，学歴や職業といった基本変数のアフターコーディングを，他の主要な社会調査に準拠して行うといった工夫も考えられる．関連して，データ公開の際には，なるべく多くのケースや変数を公開することが望ましい．調査に関わった特定のメンバーのみが利用できる変数やケースがある場合，分析結果の再現が不可能になり，公開されなかった変数を聴取する他の社会調査データとの統合も難しくなる．これは当該調査の活用を妨げる点で，調査主体や調査協力者にとっても好ましい事態とはいえない．特段の事情がある場合を除き，すべてのケース・変数について情報を公開することが統合データの構築・活用の側面からは強く推奨される．

---

[8] 他方で，他調査で使用された質問項目を用いる場合に，当該項目が著作権上の問題で（少なくとも無条件には）使用できないことがある．この場合，部分的ないし全面的に文言の変更を行わざるを得ず，調査間の比較可能性は損なわれることとなる．

第 3 に，共同研究におけるデータ申請手続きの簡素化である．1990 年代後半以降，大規模社会調査データが SSJDA などの研究センターに委託されるようになり，研究や教育を目的とする利用者が簡単な手続きを通じて自身の研究に活用できる環境が整えられた．これにより，既存の良質な調査データが有効に活用されるようになったことはもちろん，これらのデータを用いた研究の蓄積が進むことにより，当該分野における学術的貢献につながっている．本プロジェクトでは，各データを利用するために各研究者が毎年 SSJDA に 24 件，JGSSDDS に 13 件のデータ利用申請を出す必要がある．利用申請自体は比較的簡単であるものの，データの種類の多さ，また一件ずつ利用申請を行う必要があることから，少し手間がかかる作業でもあった．そしてそれは各機関においてもかなりの負担を生じさせるものと考えられる．本稿執筆時点で本プロジェクトのデータ利用に関わる研究者は 20 名なので，本プロジェクトに関連する利用申請は SSJDA においては毎年計 480 件，JGSSDDS においては 260 件におよぶ．申請を受領する機関においても膨大な数の利用申請を検討・受理の必要が生じる．研究プロジェクトの目的と研究計画が申請するデータ間・研究者間で共通している場合，データの申請手続きを一括・簡略化することによって，利用者及び管理者双方の負担の軽減につながると考えられる．

4.4　残された課題

一方で，プロジェクトを通じて明らかになった課題もある．第 1 に，作業の過程でどうしても作業の負担には不均衡が生じてしまう．表 1 に示した全 41 の調査データについて 200 程度の変数を統合する場合は，今回のように積極的かつ計量分析に習熟したメンバーが十数名程度いれば，数ヶ月程度で——むろんその後も小規模な修正は生じるものの——おおむね作業を完了できると考えられる．なかでもプロジェクトリーダーは，全体の作業計画を構築し進捗を管理するという意味で重要な役割を持つ．データ分析におけるワークフロー構築の意義を強調する J. S. Long（2009: 16）も，共同プロジェクトにおける“the enforcer”，すなわち意見の相違や判断が分かれた際に，最終決定権をもちそれを enforce する役割の重要性を指摘する．一方で，プロジェクトリーダーに責任，負担，あるいは権限が偏るリスクはあるだろう．関連して，メンバー間に生じる作業負荷の偏りも，それが不満や作業の遅れ等につながるようであれば問題である．

第 2 に，コーディングの際に一定の確率で生じる誤りを体系的に検出するしくみも構築できるかもしれない．今回のプロジェクトでは設けなかったが，1 つの変数について，コードを作成する主担当にくわえてチェック役を 1 人割り当てることで，すべての変数について，少なくとも 2 人の目を通す体制を構築できる．あわせて，タイプミスなど単純な誤りについては，生成 AI を活用することでミスの予防・検出を効率的に行える可能性もある．

第 3 に，統合データを用いて異なる社会調査データ間での結果の異同を分析することも今後取り組むべき課題といえるだろう．複数の社会調査にかかわる研究者は多かれ少なか

れ調査データ間での結果の違いや異同についての相場観を持っている．しかしながらそれが明示的に検証されることはほとんどない．統合データはこうした相場観の検証を容易し，よりよい調査方法の提案や調査の質の評価を可能にすると考えられる．この点は今後の調査実施にとって有益であり，今後取り組むべき価値があろう．

第 4 に，プロジェクトで作成したコードをいつどのような方法で公開するかも今後検討すべき課題である．分野における透明性や再現性を向上させるには，統合データ作成に用いたコードを広く公開し，こうしたデータを活用した新たな研究の促進に貢献することが望ましい．本プロジェクトでもコードの公開を前提に準備を進めている．データをどのように加工したのかという過程を公開することは，結果の再現性と信頼性，さらには将来の研究者がより短時間でより正確により進んだリサーチクエスチョンに答えることを可能にする．むろんこれは社会科学諸領域で進むオープンサイエンスの流れに即した取り組みでもある（三輪ほか 2021）．

一方で，統合データ作成に用いたコードが即時公開される場合，個々の研究者にとっては，自身でハーモナイゼーションを行うのではなく，他の研究者が公開するコードを使うこと（いわゆる「フリーライド」）が合理的な選択となるかもしれない．その結果，統合データ構築の労をとる研究者が減ることで，分野全体にとってはかえってマイナスに働く可能性がある．このジレンマを解決することは容易ではないが，統合データ作成のためのコード公開を考えるうえで，透明性・再現性の担保という金科玉条だけでなく，こうしたコードを作成する個々の研究のインセンティブをいかにして確保するかも合わせて考慮されるべきであることを指摘しておく．たとえば，統合データ構築用のコードを利用する際に引用が必要な仕組みを作ることは，インセンティブ設計の一案である．

5　結語

本稿では，「社会調査データの合併による二次分析研究の刷新」プロジェクトの一環として，複数の大規模社会調査データを統合するための具体的な作業手順を詳述した．既存データの二次分析が隆盛する一方で，その統合は少数の研究者が取り組むのみにとどまっていた．本稿は，データ統合のプロセスを可能な限り透明化し文書化することで，同様の試みを行う研究者にとっての参照点を提供することを試みた．

その作業手順は，分析対象とする調査データの選定から始まり，データ利用申請，Stata と Dropbox を用いた共同作業環境の構築，全個票データの連結，そして変数ごとのハーモナイゼーションという一連のフローから構成される．特に，調査間で選択肢が異なる変数の扱いについては，情報量を可能な限り維持しつつ比較可能性を担保するための具体的な方針を示した．こうした一連の作業を複数人のチームで分担・協働して進めることで，単独での作業に比べて労力や時間を削減できるだけでなく，コードのクロスチェックによる

正確性の向上，多様な調査に関する知見の共有といった，共同作業ならではの利点が生まれることを明らかにした．

さらに，データ統合作業を通じて得られた知見から，今後の社会調査の企画・設計やデータ公開のあり方についても具体的な提言を行った．調査項目の可能な範囲での画一化や，データ公開・申請手続きの工夫は，将来の比較研究をより容易にし，データ利用の価値を一層高めるだろう．

むろん，残された課題も多い．作業負担の偏り，コーディングエラーを体系的に検出する仕組みの構築，作成したコードの公開範囲とインセンティブ設計の問題などは，今後の検討を要する．また，本稿で構築した統合データを用いて，調査方法の違いが結果に与える影響を分析することも重要な研究課題である．

本稿，および本プロジェクトで示したのは，あくまで多数の社会調査データを協働で統合する際のワークフローの一つの事例にすぎない．しかしながらこのモデルは研究者コミュニティにおけるデータ共有と共同利用を促進し，より信頼性が高く，より大きな問いに挑む実証研究の土台となるであろう．今後，統合データの取り組みのさらなる発展，そしてそれを活用した実証研究論文のさらなる発信が期待される．

付記

本研究は東京大学社会科学研究所附属社会調査・データアーカイブ研究センター2023–2025 年度課題公募型研究会「社会調査データの合併による二次分析研究の刷新」の研究成果である．本稿は麦山・平松・コン・那須・吉田の 5 名で執筆し，麦山の他はアルファベット順に著者順を表記している．2025 年 3 月までの本プロジェクトにおけるハーモナイズ作業は，上記 5 名のほか，石橋挙（社会データ構造化センター），黒川すみれ（福岡県立大学），新田真悟（日本学術振興会特別研究員 PD），斉藤知洋（国立社会保障・人口問題研究所），佐野和子（滋賀大学），下瀬川陽（作新学院大学），苫米地なつ帆（武蔵大学），豊永耕平（近畿大学），柳下実（佛教大学）により実施された．

二次分析にあたり，東京大学社会科学研究所附属社会調査・データアーカイブ研究センターSSJ データアーカイブから「日本版 General Social Surveys＜JGSS-2000＞」「日本版 General Social Surveys＜JGSS-2001＞」「日本版 General Social Surveys＜JGSS-2002＞」「日本版 General Social Surveys＜JGSS-2003＞」「日本版 General Social Surveys＜JGSS-2005＞」（東京大学社会科学研究所）「東大社研・若年パネル調査（JLPS-Y）wave1-14, 2007-2020」「東大社研・壮年パネル調査（JLPS-M）wave1-14, 2007-2020」（東京大学社会科学研究所パネル調査プロジェクト）「家族についての全国調査，1999（第 1 回全国家族調査，NFRJ98）」「全国調査「戦後日本の家族の歩み」（略称：NFRJ-S01），2002」「家族についての全国調査，2004（第 2 回全国家族調査，NFRJ03）」「家族についての全国調査（第 3 回全国家族調査，NFRJ08），2009」「家族についての全国調査（第 4 回全国家族調査，NFRJ18），2019」

（日本家族社会学会全国家族調査委員会）「1975 年 SSM 調査，1975」「1985 年 SSM 調査，1985」「1995 年 SSM 調査，1995」「2005 年 SSM 日本調査，2005」「2015 年 SSM 日本調査，2015」（2015SSM 調査管理委員会）「SSP-I2010 調査：2010 年格差と社会意識についての全国調査（面接）, 2010」「2015 年階層と社会意識全国調査（第 1 回 SSP 調査）, 2015」（SSP プロジェクト）「教育と仕事に関する全国調査（教育・社会階層・社会移動全国調査，ESSM2013），2013」（2013 年教育・社会階層・社会移動調査研究会）「職業移動と経歴調査（第 2 回男子調査），1981」「職業移動と経歴調査（第 2 回女子調査），1983」「職業と家庭生活に関する全国調査（第 3 回職業移動と経歴調査），1991」（労働政策研究・研修機構）「結婚と家族に関する国際比較調査，2004」（福田亘孝）の個票データの提供を受けた．

日本版 General Social Surveys（JGSS）は，大阪商業大学 JGSS 研究センター（文部科学大臣認定日本版総合的社会調査共同研究拠点）が，大阪商業大学の支援を得て実施している研究プロジェクトである．JGSS-2000～2008 は学術フロンティア推進拠点，JGSS-2010～2012 は共同研究拠点の推進事業，JGSS-2015 は JSPS 科研費 JP26245060，JP15H03485，JP24243057，大阪商業大学アミューズメント産業研究所，日本経済研究センター研究奨励金 2014 年度（岩井紀子），労働問題に関する調査研究助成金 2015 年度（岩井八郎ほか），JGSS-2017/2018，JGSS-2017G/2018G は「特色ある共同研究拠点の整備の推進事業機能強化支援」，JGSS-2017/2018 は JSPS 科研費 JP17H01007 の助成を受けた．JGSS-2000～2005 は東京大学社会科学研究所と共同で，JGSS-2006～2012 は東京大学社会科学研究所の協力を得て，JGSS-2017/2018，JGSS-2017G/2018G は京都大学大学院教育学研究科教育社会学講座の協力を得て実施した．JGSS-2021H/2022H は，文部科学省特色ある共同研究拠点の整備の推進事業 JPMXP0620335833，JSPS 科研費 JP20H00089 の助成を受け，京都大学大学院教育学研究科教育社会学講座の協力を得て実施した．データの整備は，JSPS 人文学・社会科学データインフラストラクチャー構築推進事業 JPJS00218077184 の支援を得た．二次分析に当たり，JGSS データダウンロードシステムで個票データの提供を受けた．

文献

1995 年 SSM 調査研究会，1995，『SSM 産業分類・職業分類（95 年版）』1995 年 SSM 調査研究会.

2015 年社会階層と社会移動調査研究会，2018，『2015 年 SSM 調査コード・ブック』2015 年社会階層と社会移動調査研究会.

Breen, Richard, Ruud Luijkx, Walter Müller and Reinhard Pollak, 2009, “Nonpersistent Inequality in Educational Attainment: Evidence from Eight European Countries,” *American Journal of Sociology*, 114 (5): 1475–1521.

Freese, Jeremy and David Peterson, 2017, “Replication in Social Science,” *Annual Review of Sociology*, 43: 147–65.

Fujihara, Sho, 2020, “Socio-Economic Standing and Social Status in Contemporary Japan: Scale Constructions and Their Applications,” *European Sociological Review*, 36(4): 548–61.

Fujihara, Sho and Hiroshi Ishida, 2016, “The Absolute and Relative Values of Education and the Inequality of Educational Opportunity: Trends in Access to Education in Postwar Japan,” *Research in Social Stratification and Mobility*, 43: 25–37.

Gangl, Markus, 2025, “A Generalized Ordered Logit Model to Accommodate Multiple Rating Scales,” *Sociological Methods & Research*, 54(1): 315–50.

平尾一朗・太郎丸博，2011,「世代間移動レジームにおける非正規雇用の位置」『理論と方法』26(2): 355–70.

鹿又伸夫・田辺俊介・竹ノ下弘久，2008,「SSM 職業分類と階層指標：EGP 階級分類・SIOPS・ISEI への変換」前田忠彦編『2005 年 SSM 調査シリーズ 12：社会調査における測定と分析をめぐる諸問題』2005 年社会階層と社会移動調査研究会，69–94.

Long, J. Scott, 2009, *The Workflow of Data Analysis Using Stata,* Stata Press.

三輪哲・下瀬川陽・打越文弥・工藤沙季，2021,「再現性ポリシーの意義・現状・課題」『社会と調査』27: 14–22.

麦山亮太・田上皓大，2024,「SSM 職業分類への日本版 O-NET のマッチングと応用：スキルと賃金の関連を事例に」『理論と方法』39(1): 64–76.

Mugiyama, Ryota and Kohei Toyonaga, 2022, “Role of Cohort Size in Trends in Class and Occupational Returns to Education at First Job: The Case of Japan,” *European Sociological Review*, 38(2): 269–85.

直井優，1979,「職業的地位尺度の構成」富永健一編『日本の階層構造』東京大学出版会，434–72.

Oesch, Daniel, 2006, *Redrawing the Class Map: Stratification and Institutions in Britain, Germany, Sweden and Switzerland*, Palgrave Macmillan.

斉藤知洋・吉田航・下瀬川陽，2024,「異種類の国内社会調査データの統合と標準化」東京大学社会科学研究所附属社会調査・データアーカイブ研究センター2023 年度二次分析研究会課題公募型研究「社会調査データの合併による二次分析研究の刷新」成果報告会資料.

Torche, Florencia, 2011, “Is a College Degree Still the Great Equalizer? Intergenerational Mobility Across Levels of Schooling in the United States,” *American Journal of Sociology*, 117(3): 763–807.

打越文弥，2016,「学歴同類婚の世代間連鎖とその趨勢：大規模調査データの統合による計量分析」『家族社会学研究』28(2): 136–47.

Japanese Social Survey Data Harmonization Project: The Procedures and Insights for Future Surveys

Ryota Mugiyama (Gakushuin University)
Makoto Hiramatsu (Hokkaido University)
Aram Kwon (Osaka University of Economics)
Rantaro Nasu (The University of Tokyo)
Wataru Yoshida (National Institute of Population and Social Security Research)

Abstract

This paper explains the procedures of the Japanese Social Survey Data Harmonization Project. While Japanese scholars have become familiar with analyzing existing social survey data, the harmonization of multiple datasets has rarely been conducted. This has led to problems such as wasted effort, a lack of reproducibility, and a limited scope for research questions. To address these issues, this project integrated major social surveys conducted in Japan through the collaboration of several researchers. The process consisted of the following stages: (1) selecting data, (2) establishing a working environment using Stata and shared folders, (3) merging all datasets, and (4) harmonizing variables. In particular, we detail the process of variable harmonization, explaining how we addressed challenges such as different question wording and response options across surveys through collaborative work. We identified clear benefits of this collaborative approach, including increased transparency and accuracy in the process, as well as an accumulation of insights about each survey. Furthermore, we offer recommendations for future survey design. In conclusion, this paper expands the usability of social survey data and contributes to more reliable and reproducible empirical research.