

〈研究ノート〉

インド統計資料を用いたデータセットの構築

児 玉 寿 子

目 次

- | | |
|----------------|------------|
| I. 研究の背景と目的 | IV. 欠損値の推定 |
| II. インドの統計の概観 | V. ま と め |
| III. データセットの構築 | |

I. 研究の背景と目的

インドはアジアのみならず、世界の発展途上国の中でも、最も国家統計の資料が豊富な国の一つであり、17世紀から体系的な資料が収集されてきている（杉原 1997：4）。また、インドはアジアで最も長い人口統計の歴史を有し、近い将来に中国を追い越して世界最大の人口を抱えることと推測されていることから、多くの研究者の分析対象となってきた。さらに、税収管理の観点からは、英領インド時代より、農業や森林についても詳細な統計が収集されてきた経緯がある。昨今では、これらの豊富な統計データが各省庁のウェブサイトからも入手できるようになっており、研究材料のみならず、インドでのビジネス展開を目論む企業にも広く利用されている。

著者は、インドの国勢調査を中心に、農業、森林、経済、貧困などの多岐にわたる豊富なデータを包括したデータセットを構築し、その解析を通して、地域による貧困、環境、産業などの関係性を明らかにしたいと考えている。そのため、帰納法を採用することで、先に仮説を設定せず、多様な変数を用いた多変量解析を用いて、分析結果から仮説を導き出そうとしている。より多様な社会経済事象を分析に反映させるには、多くの変数を用いることが重要である。変数を欠いてしまうことは、より大きな影響のある変数を除外するリスクを高める。よって、欠損値がある変数についても、除外するよりも、可能な限り妥当な値を推定することで、解析に利用できる完全な変数を確保することが重要である。多くの既往の研究では、人口動態（国勢調査）、農業・農地、貧困などの特定分野に限定した変数を用いた研究が中心であり、本研究ほど多様な変数を用いた分析は実施されていない。

本研究ノートでは、まず、上記研究において利用したインドの統計資料を概観する。その後、データセット構築において生じた欠損値を丁寧に特定し、その推定過程を示すことにより、欠損値を推定の妥当性を検証したい。

II. インドの統計の概観

1. インドにおける統計の発展

ムガル帝国、諸藩王国らは、土地台帳をはじめ、穀物の生産高などをとりまとめてきた。イギリスによる殖民支配以前のこれらの支配者らは、主に資源の把握と課税目的に系統的な記録作業を行ってきたが、公式な統計としての発達はイギリス殖民支配が始まって以降であるといわれている (Baines 1918: 417)。

英領インド帝国政府は、殖民支配を開始直後から、財政移転及び海上輸送による貿易の記録に注意を払ってきたが、行政統計は行政区の発展に伴って断続的に (時には時期を遡って) 収集されてきた。19世紀半ばには管区や町、市などの行政区が整備され、鉄道網の整備が進んだ。インド帝国政府は、53万7000以上の村落について、地租及び税收状況を記録・改訂するために、単独あるいは複数を単位に会計官を配属した。第一次情報源である会計官らは、村落で読み書きのできるほぼ唯一の人間であり、地租に関する記録業務に加えて、行政側が求めるあらゆる情報の収集に携わった (Baines 1918: 419)。村落で集められた情報は小区域 (territorial subdivision) または領域 (circle) に、小区域・領域から県 (district) に、県から管区 (province) のレベルで取りまとめられた¹⁾。インド帝国政府の施政は、財務、外務・法務、内務、鉄道、歳入・農業、教育などの省を通してなされており、各種統計も管轄省が収集していたが、各管区レベルでの各省にかかる業務区分は全国的に統一されていなかった。

系統だった行政統計が普及して以降、気温や降雨などのデータとその分析結果が洪水制御や農業発展に及ぼすインパクトが明確になり、国内で統計の重要性への認識が進んだ。1931年にはインド人数学者であるプラサンタ・チャンドラ・マハラノビス (Prasanta Chandra Mahalanobis) 教授によりインド統計研究所 (Indian Statistical Institute: ISI) が設立され²⁾、インド国内で本格的な統計教育が開始された。なお、今日の ISI では、統計とそれに関連する自然科学・社会科学などの研究・教育を始め、数学、物理や応用経済などが盛んである。1951年には、各省で実施される統計業務の調整や管理を目的に、中央統計機関 (Central Statistical Organization) が設立された。

2. インドの統計資料の概要 (ウェブサイトから入手可能な資料)

本章では、本研究で利用した統計資料を含め、ウェブサイトから入手可能であるインドの統計を簡潔に紹介する。

1) Baines によると、当時267箇所あった県がインド帝国の統計上の地理的・行政的な基礎単位であった (郵便や通関などの中央政府が管轄する事項は除く) (Baines 1918: 419)。

2) 設立当初は管区大学 (Presidency College) の統計研究所内に設置され、1959年より国立の研究所となっている。

まず、インド政府で統計資料を包括的に管理する立場にあるのが連邦統計プログラム実施省 (Ministry of Statistics and Programme Implementation) である。同省のウェブサイト³⁾からは、基本統計である「インド基本統計 (Statistical Abstract of India)」が入手可能であり、社会、経済、農林水産業、エネルギー、通信、運輸、貿易、労働などに関する統計を収集することができる。また、ミレニアム開発目標 (MDG) に関する到達状況を示した MDG レポートや、国民勘定、州別の環境報告書なども、同省のウェブに利用登録することで自由にダウンロードすることができる。

(1) 国勢調査

国家統計の代表である国勢調査からは、人口動態にかかる様々な情報が把握可能である。国勢調査を管轄するのは連邦計画省 (Planning Commission) の国勢調査部 (Census Bureau) であり、調査は10年に一度実施されている。インドの国勢調査はアジアで最も長い調査経験を誇る。第1回の国勢調査は、1871-72年に実施されている。第2回調査は1881年に実施され、以降、10年毎に1のつく年に実施されている⁴⁾。最新の国勢調査は2001年の第14回調査である。第1回調査は、まだインド帝国が成立する前の藩王国と東インド会社の時期であり、調査対象地はインド全土を網羅しておらず、また調査の時期にもばらつきがあった (Baines 1918: 421)。国勢調査として時期を定めて実施したのは1881年の第2回調査からである。当時のインドでは英国の領土拡張政策に対して国民の不満が高まり、1857年にセポイの乱が起きた。衰退に向かっていたムガル帝国は国民を結集させる力を失っており、英国はインド支配の核であった東インド会社を廃止してインドを直接統治する準備を進めていた。インド帝国が成立したのは、第1回国勢調査から6年経た1877年であるが、調査は英国のインド支配をより安定したものにするための基礎情報の収集と、支配体系の強化の必要性が背景にあったと思われる (久保田 2007: 71)。

度重なる調査を経て、今日の国勢調査では、人口学的指標 (年齢別・性別・居住地域・宗教・言語・教育水準別の人口)、社会的指標 (カーストや部族の別)、地理的指標 (出生地や人口移動の有無と目的)、経済的指標 (職業、土地所有、雇用状態)、生活水準指標 (住居の状態、基礎的社会インフラへのアクセス) などが収集されている。また、村レベルでは、村落記録簿 (Village Directory) が作成され、学校や病院へのアクセス、農地面積などが取りまとめられる。調査員によって収集された各世帯の基礎情報 (Primary Census Abstract) は、村落 (Village, Panchayat)、地区 (Block, *Mandal*, *Taluk*, *Tehsi* など)、県 (District)、州 (State) のレベルで集約される。そのため、世帯の基礎情報については、都市部では区 (Ward) レベル、農村地域では村落 (*Panchayat*) 単位の集計まで入手することができる。1991年調査結果以降、集計結果は報告書の形態に加えて、デジタル・データとしての提供が開始された。ウェブサイトでは、州レベルの統計と県の基礎情報の集計までを入手でき、それより詳細なデータについては、各州

3) <http://www.mospi.gov.in/> (2009年10月9日アクセス)

4) 1941年以降には、調査の基準日を3月1日に定めている。これは、インドが雨季に入る前の気候が比較的安定した時期であるためである (久保田 2007: 71)。

にある国勢調査部の支部で Microsoft Excel のスプレッドシートとして CD の形態で購入可能である⁵⁾。

(2) 家計調査 (国家標本調査)

家計に関するデータについては、連邦統計プログラム実施省の管轄下にある国家標本調査機関 (National Sample Survey Organization: NSSO) が全国標本調査 (National Sample Survey: NSS) として家計調査を実施している。NSS は、1950年の第1ラウンド調査を初めに、毎年実施されている。調査ラウンドによって取り上げる中心トピック及びサンプル数は異なっているが、世帯の消費・所得・雇用・失業など計測される主な項目である。毎年実施する調査であるため、作業量の限界から対象サンプル数は少なくサンプリングによるバイアスの可能性が高まる。しかし、ほぼ5回に1度のラウンドでは10万以上のサンプル数をとっており、多量サンプル調査は信頼度が高いといわれている (黒崎 2005: 55)。最近のものでは、1999-2000年の第55ラウンド、2004-2005年の第60ラウンドなどがそれに該当する。

統計プログラム実施省のウェブサイトからは、NSS の第48ラウンド (1993年) 以降の各ラウンドの報告書がダウンロード可能である⁶⁾。NSS を利用することで、世帯レベルの食糧品の消費などミクロの分析が可能である。なお、本研究においては、ミクロレベルの分析よりもむしろマクロの横断分析を目的とするため、NSS の調査結果は利用していない。

(3) 経 済 統 計

経済や産業、金融、証券、物価、社会基盤などに関しては、連邦財務省 (Ministry of Finance) が毎年出版する「経済白書 (Economic Survey)」が主要なものとしてあげられる。その他、連邦統計プログラム実施省による「産業年鑑 (Annual Survey of Industries)」, 財務省 (Ministry of Finance) の「インド公共財政統計 (Indian Public Finance Statistics)」なども各省のウェブサイトから入手可能である。

また、1977年より統計プログラム実施省の下部組織である中央統計機関 (Central Statistical Organization) が、経済情勢調査 (Economic Census) を実施している。この調査は、インド経済の主軸である農業分野及び非農業分野の経済情報を取りまとめ、計画に反映させることを目的とするものである。初回調査はラクシャドウィープ連邦直轄領を除く全国で実施され、事業体別の経済活動及び特徴、農村・都市別の事業体の情報などが収集され、州レベルで取りまとめられた。また、主要な経済活動や規模別の事業体数なども集計表として取りまとめられた。1980年の第2回、1990年の第3回、1998年の第4回と実施され、2009年9月時点で最新のものは2005年の第5回調査である。

(4) 農 業

農業や森林については、英領インド帝国の以前より、各地の藩王国によって綿密な情報が収集

5) 各州に設置されている Census Bureau では、設置されている州の詳細なデータに限らず、他州のデータを入手することも可能である。

6) http://www.mospi.gov.in/mospi_nssso_rept_pubn.htm (2009年10月10日アクセス)

されてきた。土地に基づいて徴収される地租が、特に19世紀から20世紀始めの時期にかけて英領インドの財政収入において重要なものであったためである。また、国勢調査と同様に、土地の所有状況を把握することは、税収集体制を確立することでもあり、イギリスの植民地支配の基盤となっていたためでもある。そのため、早い時期から土地税の納税者を確立する土地査定の作業、農業生産の動向が把握されるようになり、農業統計が作成された。作物毎の作付面積に関しては、1884-85年からは、「インド農業統計 (Agricultural Statistics of India)」が発表され、生産量については1891-92年から「インド主要穀物の耕作面積と収量の推定 (Estimates of Area and Yield of the Principal Crops of India)」が取りまとめられている (柳澤 1997)。

現在は、連邦農業省 (Ministry of Agriculture) の農業情勢調査局 (Agricultural Census Division) が、農地所有、作物別の作付面積、収量、灌漑面積など農業に関する統計を取りまとめており、州レベルの統計が同省のウェブサイトから入手可能である。また、連邦水利資源省 (Ministry of Water Resource) も水源管理の観点から、水源別の灌漑や作物別の灌漑統計などを取りまとめている。

(5) 森 林

農業とならんで、森林は歴史的にインド支配層にとって税収を得る、また材木や林産物による収益を得ることのできる有用な資源・資産とみなされてきた。インドで最初の統一国家を形成したマウリヤ朝時代には、チャンドラ・グプタ王の支配期より、巨大な王朝の財政と国民生活を支えるために森林を区分し、保全し、税を徴収する制度が導入されている (Jha 1994)。近代的な手法による森林地の調査は、1800-1860年にかけてインド南部のマラバル丘陵地域での実施されたものが最初であるといわれている (Press Information Bureau 2008)。

イギリスによる支配体制が強まった以降には、1864年に帝国森林サービス (Imperial Forest Service) が設立された。帝国森林サービスは森林行政を担う森林官制度であり、これによって森林データの収集方法が統一され、系統的な森林調査が開始された。1906年にはインド北部のDheradun にインド森林調査研究所 (Indian Forest Research Institute) が設立され、また各州においても森林官教育が進んだ。

1981年には、森林調査に特化した組織であるインド森林調査所 (Forest Survey of India: FSI) が連邦環境森林省 (Ministry of Environment and Forest) の管轄下に設立され、森林調査を担っている。FSI は各州の森林局で収集されたデータを基に、1987年から2年毎に森林概況報告 (State of Forest Report: SFR) を出版しており、2009年9月時点で最新の報告書は2007年のものである。なお全ての報告書はウェブサイトから入手可能である⁷⁾。同所による森林統計とは別に、農業省経済統計局 (The Directorate of Economics and Statistics, Ministry of Agriculture: DESMOA)⁸⁾ が土地利用統計の観点から、森林地域にかかる統計を収集・公表している。

7) <http://www.fsi.nic.in/> (2009年10月10日アクセス)

8) <http://dacnet.nic.in/eands/Default.htm> (2009年10月10日アクセス)

(6) 貧 困

インド政府は、1951年に第1次五カ年計画を開始して以降、貧困削減を開発計画のなかで重要な柱として位置づけてきた。1962年には、必要最低支出額にもとづいた最初の公式貧困ラインが計画委員会によって設定された。その後、計画委員会の専門家グループの勧告に基づいて同基準の算出方法が見直された結果、1993年には生存に必要なカロリーを接種するための必要最低食糧支出額と、非食糧支出額の合計値からなる貧困ラインが設定された。

貧困ラインの設定は以下の通りである。まず、NSSの家計調査結果を基準に、都市・農村別に生命維持のために必要なカロリー値を定める。都市部での標準カロリーは1人1日当たり2,100カロリーで、農村部では同2,400カロリーである。これらのカロリーを満たすための食品の組み合わせが「食料バスケット」として計算される。食料バスケットに含まれるのは米、肉類、野菜類、豆類などの食料であり、摂取状況は地域によって異なるものであるが、全国統一の基準が設定されている。次に、この食料バスケットを満たすために必要な食料購入費用を算出した額が貧困ライン（単位はルピー）である。

各州における、世帯の所得が貧困ラインに満たない世帯の全世帯に占める比率が、計画委員会が公表する州別貧困率となっている。計画委員会は、過去1973-1974年、1977-1978年、1983年、1987-1988年、1993-1994年、1999-2000年に州別の貧困率を公表しており、最新のものは2004-05年の値である。なお、年次の異なる貧困率を比較分析するにあたっては、貧困ライン設定の基礎になっている消費額の設定に留意する必要がある。NSSの家計支出の調査では調査の実施ラウンドによって、1週間の支出を振り返る（7-day recall period）、1ヶ月の支出を振り返る（30-day recall period）、1年間を振り返る（365-day recall period）方法がとられており、Recall periodによって回答者の応える支出額に差異が生じ、それが貧困率に影響を与えているためである（Press Information Bureau 2007, 黒崎 2005: 55）。1993-1994年までのラウンドでは、30-day recall period が採用されていたが、1999-2000年には1週間と30日間の両期間が併用された。さらに2004-2005年には、全品目について30日間で聞く方法（Uniform Recall Period: URP）と、衣類・靴・教育・医療・耐久財については1年間、その他の品目については30日間で聞く期間混合の手法（Mixed Recall Period: MRP）の2種類が採用された。たとえば、2004-2005年の貧困ライン設定の基になっているのは、第61ラウンドのNSS調査であり、URPとMRPの両方で貧困率が設定されている。2004-2005年のURPに基づく貧困率は、同方法の消費設定で算出された1993-1994年の貧困率と比較可能である。他方、1999-2000年の貧困率とは、2004-2005年のMRPに基づく貧困率である程度比較可能である（Press Information Bureau 2007）。

インドは巨大な貧困人口を抱え、また複雑な社会構造をもっていること、さらにNSSや国勢調査等の統計調査の歴史が長く、農村のパネルデータが整っていることから、貧困に関する数多くの研究・分析がなされている。本研究では用いていないが、国立応用経済研究所（National Council of Applied Economic Research: NCAER）⁹⁾と、国際半乾燥地熱帯作物研究所

9) <http://www.ncaer.org> (2009年10月10日アクセス)

(International Crops Research Institute for the Semi-Arid Tropics: ICRISAT)¹⁰⁾ は数年間にわたるパネルデータを収集しており、ICRISAT についてはウェブサイトから入手申請することができる。NCAER のデータは、農村に居住する4,118世帯から1968-1969年、1969-1970年、1970-1971年にかけて収集したものである。ICRISAT のデータは、1975-1976年から1984-1985年の10年間にかけて（一部データは1980-1981年以降の5年間若しくは1981-1982年以降の4年間）、4州5県の9村について、1村につき40世帯を対象に行った世帯調査から構成されている。特定の調査世帯の家計状況を追跡したパネルデータは、世帯の受けたインパクトや消費の変動など、貧困ライン以下に転落するリスク（脆弱性）などを分析するうえで有益である。

3. インド統計資料の課題

このように、インドでは長年にわたり多様な統計資料が収集されてきており、政策立案や研究において広く利用されてきた。しかしながら、インドの統計制度は行政による統計が主流であり、その質の低下が指摘されている。行政統計に過剰に依存していたところが、統計実施の質の管理体制が低落し、特に1990年の経済自由化以降にはデータ収集システムなどが機能不全に陥り、統計の質の低下が著しく進んだと指摘されている（岡部 2003: 55）。

インド政府は、2000年1月に特設の国家統計評議会（National Statistical Commission: NSC）を設け、農業、工業、通商、サービス、インフラ、社会経済、金融、物価、国民経済計算など多岐の分野を網羅して、各分野の統計の現状と課題を検討した（法政大学日本統計研究所 2003）。岡部（2003）では、同評議会による報告書で指摘するインド統計の課題を、① インドの統計が行政統計に依存しており、行政統計が機能不全になっている、② 州政府と中央政府の調整を図る制度が弱体化している、③ 中央の各省庁と中央統計機関の間の調整を図る制度が弱体化している、と集約している。

同評議会が特に重視したのは、複数の省庁によって出された重複データの比較照合分析である。たとえば、森林分野では、1987年に State of Forest Report (SFR) が出版される以前には、National Remote Sensing Agency (NRSA) が1972-1975年、1980-1982年のランドサットイメージに基づいて森林被覆率を計測した。しかし、インド森林調査所が1981-1983年のランドサットイメージを元に、新たな計測を行ったところ、NRSA による推定との間に大きな格差が生じ、両機関が調整して計測地を統一する事態が生じたという。平行して、農業省経済統計局 (DEMOSA) も村落土地記録を元に森林面積を公表していたが、森林調査所と経済統計局では定義及び計測方法が異なっていたことから、データは食い違ったものだった。さらに、漁業に関する統計では、州政府の担当部局が収集したデータを基に農業省が統計を編纂するが、漁業輸出額が漁業産出額を上回る矛盾が生じている。

このように、インドの統計資料の信憑性、正確性、信頼性に課題を抱えているが、インド政府

10) <http://www.icrisat.org/gt-mpi/knowledgeBase/Databases/vls.asp> (2009年10月10日アクセス)

は、2006年に常設の国家統計評議会を設立し、統計の質の回復を目指している。新たな評議会は、統計に関する技術課題の検討、省庁間の調整、定義やデータ収集方法などの統一などを担うこととなっている¹¹⁾。

なお、本研究では、利用したデータについて、明らかに誤りと判断できるデータ以外はその質を問わず、現時点での最良のデータであるとみなしている。

III. データセットの構築

1. データの範囲

インドは28の州 (State) と7つの連邦直轄領 (Union Territory) から構成される連邦国家である。州の下には、県 (District または *Zilla* などという)、区域 (Block, *Mandal*, *Taluk*, *Tehsil* などという)、村 (Village, *Panchayat*) の構造をとっている。28の州のうち、3州は州の分割によって2000年に新たに設けられたものである。これらの州は、(1) マディヤ・プラデシュ (Madhya Pradesh: MP と略す) 州の東南部に位置していたチャティスガル (Chhattisgarh) 州、(2) ウッタル・プラデシュ (Uttar Pradesh: UP と略す) 州の北西部に位置していたウッタランチャル (Uttaranchal) 州、(3) ビハール (Bihar) 州の南部に位置し28番目の州となったジャールカンド (Jharkhand) 州である。

各州及び連邦直轄地には、国勢調査局などの出先機関が設置されているほか、州・連邦直轄地政府の管轄部局が各種統計データを取りまとめている。本研究で利用した国勢調査や森林統計などは、35の州・連邦直轄地全てについての統計を備えている。他方で、農業生産や貿易や経済などについては、連邦直轄地の統計を示していないものや、特定の連邦直轄地のデータのみ示すことが多い。これは、州に比べて連邦直轄地が人口、面積、経済などの面で規模が小さいために除外されたり、データが得られていない場合があるためである。

上記の理由から、本研究ではデータの比較的そろいやすい28の州に限定するものとする。インドの多くの州では面積が広大で人口も多く、一つの州であっても地理的、社会的に多様性に富む。州レベルのデータでは、州内の多様な現状を明確に映し出すことに限界があるため、より詳細な県レベルのデータを分析に用いることが望ましい。他方、インドの県は全国で県数が590以上あり、国勢調査によるデータは入手可能であるが、それ以外のデータは白書や年鑑では公表されていないなかったり、入手が困難または不可能である。また、入手可能な州があっても、欠損となる州が多い。そのため、より多くの変数を用いてインド全域を概観するために、本研究においては州レベルデータを用いた分析を行うものとする。

2. データの種類

本研究は、インドの社会、経済、産業、環境、教育等にかかる非常に広範囲のデータを用いて

11) http://mospi.gov.in/national_statistical_commission_header1.htm (2009年10月10日アクセス)

統計分析を行った点と、データセットに生じた欠損値を丁寧に復活させた作業に特徴がある。多くの既往の研究では、人口動態（国勢調査）、農業・農地、貧困などの特定分野に限定した変数を用いた研究が中心であり、本研究ほど多様な変数を用いた分析は実施されていない。

包括的なデータセットを用いた分析によって地域の特性を明らかにし、中でもインドの自然環境、生計、貧困のありようを解明することが統計分析を行うことが本研究の目的である。そのた

表 1. 変数の概要

出 展	変数の種類	年 度	変数の数
Planning Commission	都市・農村別の貧困ライン、貧困率	1999	12
Census of India	都市農村人口、性別人口、指定カースト・部族人口、職業別人口、識字・教育水準、人口移動、家屋用材の種類、衛生設備の有無	2001	352
Ministry of Social Justice and Empowerment	乳児死亡率	2001	3
Office of the Registrar General of India	安全な水へのアクセス	2001	3
“Estimates of GDP and other Aggregates at Current Price - 1999-2000 Series”, Ministry of Statistics and Programme Implementation	産業別の州内総生産	1999-2000	23
“Fifth Economic Census”, Ministry of Statistics and Programme Implementation	農業事業所数と就労者数	1998	8
Agricultural census division, Ministry of Agriculture	土地所有面積、農家規模、負債	2000-01	22
“Water Sector at a Glance- Water Data Complete Book 2005”, Central Water Commission	灌漑、耕作地、荒廃地、電気の有無	2005	48
“Measuring Milestones Proceedings of the National Workshop on JFM 2006”, Ministry of Environment and Forest	森林管理組合の組合員数、管理面積	2006	15
“State of Forest Report 2001”, Ministry of Environment and Forest	森林面積（質、区分）、樹木面積、指定部族認定地域、高地区分	2001	40
“State of Forest Report 1999”, Ministry of Environment and Forest	植林地面積、森林がある村、森林管理組合数	1999	11
“State of Forest Report 2003”, Ministry of Environment and Forest	林木蓄積量	2003	4
“Forest and Wildlife Statistics”, Ministry of Environment and Forest	荒廃地、森林関連予算、不法占拠、不法伐採、森林局部署、森林官数	2004	92
“17th Indian Livestock Census All Summary Report”, Ministry of Agriculture	家畜の種別頭数	2003	23
India Meteorological Department for normal rainfall data	平均年間降雨量	NA [※]	1
計			657

※ 出展年不明

め、分析に用いたデータは、2001年の国勢調査による人口動態、教育水準、職業、人口移動、家の状況、水へのアクセスなどを中心に、森林統計、農業、産業、経済、乳幼児死亡率、荒廃地面積、灌漑普及状況など、州の経済社会状況を示すと考えられるデータを可能な限り多く盛り込むことに努めた。国勢調査結果は、ウェブサイトからダウンロードできるデータをほぼ全て盛り込んだ。また、森林統計についても、森林概況報告（SFR）から入手可能なデータを全て用いた。インドはまだ国民の7割以上が農村地域に居住し、農業及び農業関連産業からの収入に依存して生活しているほか、農作物の出来高が国家経済に与える影響も大きい。さらに農作物の耕作高は、降雨や気温（自然条件）及び灌漑整備状況（人為的条件）に大きく影響を受けている。そのため、土地所有、灌漑整備などの農業関連変数に特に配慮した。本研究では、国勢調査の調査年である2001年にあわせて、必要なデータの収集を行った。国勢調査以外の資料については、最新のデータが入手可能であっても2001年またはそれに最も近いものを採用することで、データ年次の整合性に務めた。

前章に示したとおり、インドでは各種統計資料が主に州別に整備されており、過去10年程度の資料が政府機関のウェブサイトからダウンロードして利用可能である。本研究では、全ての統計資料を政府機関のウェブサイトより収集し、Microsoft Excel のスプレッドシートに入力した。収集された変数は、合計657であった。データの数値は、出展元に記載されている通り利用することを原則としたが、数値が確実に誤っていると判断できた4変数¹²⁾については筆者が修正を加えた。使用したデータの出展と変数の種類、数の内訳は表1の通りである。

3. データの欠損状況

収集した657の変数のうち、506変数は欠損値のない完全な変数で、残りの151変数は欠損値を含む変数であった。欠損値の種類は大きく分類して2とおりある。まず、欠損値に関係するなんらかのデータを利用して計算することで、推定ができる欠損値である。たとえば、前述したように、インドでは2000年に3つの州が分割されて、3つの州が新設された。そのため、分割する前の州では統計データがあるが、分割によって2州になって以降の両州のデータは多くが欠如している。A+B+C=Dのように、いくつかのカテゴリー変数とその合計変数の組み合わせで、一部だけ欠損となっているケースも同様に関連データを基に推定できる欠損値である。次にあるのが、推定作業に利用可能なデータが無い、一般的な欠損値である。

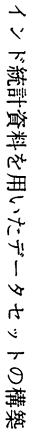
IV. 欠損値の推定

1. 欠損値推定の流れ

欠損値を可能な限り妥当な方法で推定するために、本研究では、図1のように5つの段階を経

12) 単位がパーセントであるにもかかわらず100%を超過しているものや、桁がずれているものなどである。

March 2010



て推定値を算出した。まず、(1) 欠損値には関連する何らかのデータを用いた按分での推定、次に(2) ゼロを推定投入可能な変数の推定、続いて(3) 論理式による計算での推定、(4) 多欠損ケースの変数の棄却、最後に(5) 回帰分析による推定を行った。変数の欠損ケースによっては、ある段階の推定作業で全ての欠損ケースが回復できる(単回推定)が、変数の欠損ケースがその段階の推定作業だけでは回復できない場合には、次の段階に送り推定を続けることで回復させる(多回推定)こととした。

その結果、当初は欠損値があった変数が151であったが、最終的には5つの段階で推定することのできなかった17変数のみを棄却し、640の変数にてデータセットを作成することができた。本章では、各段階での推定過程の詳細を述べる。

2. 欠損値の推定方法

(1) 関連データを用いた按分による推定

欠損値推定の最初の段階では、まず、当該変数に関連する元データがある欠損値についての推定を行った。このような欠損値を含む変数は、主に荒廃地や森林面積などに関する変数であり、56変数が該当する。その内訳としては、2000年に3州が分割された影響で欠損が生じたものが51変数、分割には関係がない州で森林管理組合数があるにもかかわらず構成の内訳が欠損であるものが5変数であった。

按分過程は大きく分けて州の分割による欠損値の按分と、州の分割とは無関係の欠損の按分で、以下のように行った。分割された州での推定では、たとえば、貧困層人口に関する変数では、ビハール州がビハール州とジャールカンド州に分割される以前のビハール州の貧困層人口が明らかであった。そのため、国政調査で判明しているビハール州とジャールカンド州の人口比率(州分割後)を用いてそれぞれの州に割り当てることで貧困層人口を算出した。荒廃地に関する変数で欠損が生じたものについては、分割された州の荒廃地状況が同様である仮定し、分割後の両州の総面積の比率を按分に用いて荒廃地面積を算出した。同様に、農村の数など、農村に関する変数は農村人口比率を、貧困層人口など人口全般にかかるものは人口比率を用いた。また、貧困率や貧困ラインについては、両州が同様の貧困状態であると仮定し、両州に同一の数値(元のデータ)をあてはめた。森林のある村落数については、分割された州であったため、農村人口比率や森林面積の比率を用いて推定を試みた。しかし、たとえば、分割されたウッタル・プラデシュ州は、分割後には、農村人口が多く森林が少ないウッタル・プラデシュ州と、農村人口が少なく森林が多いウッタランチャル州に分割されたため、農村人口と森林面積のいずれを用いた按分でも数値が大きくぶれてしまい、州の現状を明確に示すことができない。よって、インド全体の農村人口に占めるウッタル・プラデシュ州の農村人口の比率に、インド全体の森林面積に占めるウッタル・プラデシュ州の森林面積の比率をかけたものと、インド全体の農村人口に占めるウッタランチャル州の農村人口に、インド全体の森林面積に占めるウッタランチャル州の森林面積をかけたものの比率を計算した。その結果、両州の比率は0.923(ウッタル・プラデシュ州)と0.077

表 2. 按分により推定された変数の内訳

按分に使った比率	比 率 の 詳 細	変数の種類	変数の数
農村人口比率	- 比率 Bihar=78.0% : Jharkhand=22.0% - 比率 MP=72.7% : Chhattisgarh=27.3% - 比率 UP=95.4% : Uttaranchal=4.6%	村の人口, 村の数	3
人口比率	- Bihar=75.5% : Jharkhand=24.5% - MP=74.3% : Chhattisgarh=25.7% - UP=95.1% : Uttaranchal=4.9%	貧困層人口	3
面積比率	- Bihar=54.2% : Jharkhand=45.8% - MP=69.5% : Chhattisgarh=30.5% - UP=81.8% : Uttaranchal=18.2%	荒廃地面積	24
森林樹木面積の比率	- Bihar=27% : Jharkhand=73% - MP=58% : Chhattisgarh=42% - UP=47% : Uttaranchal=53%	植林面積, 森林野生動物予算, 外国資金受入額	9
2004年の森林組合数の比率	- Bihar=12.80% : Jharkhand=87.20% - MP=66.56% : Chhattisgarh=33.44% - UP=16.73% : Uttaranchal=83.27%	1999年の森林管理組合数, 管理林面積	2 (2) [※]
欠損値のある州以外の組合員の男女比率	男性平均=58.1% : 女性平均41.9%	男女別の森林管理組合員数	2
2004年の社会的属性別の森林組合世帯数 (SC, ST, OF)	- Assam: SC=7.2%, ST=55.1%, OF=37.7% - Jharkhand: SC=18.1%, ST=25.9%, OF=56.0% - Sikkim: SC=20.5%, ST=25.5%, OF=54.0%	2006年の社会的属性別の森林管理組合員数	3
元となるデータの数値を欠損州に適用	元となるデータの数値を欠損州に適用	貧困率, 貧困ライン	9
分割した州がインド全体の農村人口と森林面積に占める割合の比率	(インド全体の農村人口に占める UP 州の農村人口×インド全体の森林面積に占める UP 州の森林面積) 対 (インド全体の農村人口に占める Uttaranchal 州の農村人口×印森林面積に占める Uttaranchal 州の森林面積)=0.923対0.077。 - Bihar 州対 Jharkhand 州=0.473対0.527 - MP 州対 Chhattisgarh 州=0.785対0.215	森林のある村落	1
計			56 (2) [※]
次の段階での推定作業を必要とする変数			97 (2) [※]

※ ()内は第1段階で全ての欠損ケースを回復できなかった変数

(ウッタランチャル州) となった。この比率を森林のある村落数の元の数値にかけることで、分割後の両州の値を推定した。分割されたビハール州とマディヤ・プラデシュ州についても同様の作業を行った。上記作業の結果、51の変数の州分割で生じた欠損値が推定され、全変数が完全に回復された。

州の分割と無関係で生じた欠損値は、森林管理組合の内訳に関するものである。たとえば、トリプラ (Tripura) 州, ジャム & カシミール (Jammu & Kashmir) 州では全組合員数がわかっていて男女別組合員数が欠損であったため、欠損が無い州の男女比率平均によって男女別の組合員

数を算出した。社会的属性別（指定カースト世帯（SC と略す）、指定部族世帯（ST と略す）、その他のカースト世帯（OF と略す））の森林管理組合員数が不明なシッキム（Sikkim）州やアッサム（Assam）州、ジャールカンド州については、本研究で利用していない2004年の森林管理組合の属性別組合員の家族数が入手できたため、欠損のある州の社会的属性の比率を割り当てて各組合員数を算出した。その結果、5つの変数で、関連データのある欠損値が推定され、3変数については完全に回復された。

上記の按分を行った結果、56あった変数のうちで、54の変数は欠損値が全て推定されて完全な変数に転換することができた。2つの変数については、関連データのない欠損ケースが含まれていたため、按分作業を経ても完全化することができなかったものである。2変数は、関連データを利用した推定ができなかった95の変数と共に、2段階目の推定作業を行うこととした。本段階で推定された変数と按分の詳細は表2の通りである。

(2) ゼロの推定投入による推定

第2段階では、按分作業で完全な変数とならなかった2変数と、利用可能な関連データがなく按分推定ができなかった95変数の計97変数についての推定作業を行った。ここで行ったのは、欠損値のなかでゼロを投入できるものを推定することである。ゼロの推定方法は以下の2種類である。たとえば、水路灌漑面積、ため池灌漑面積、井戸灌漑面積、その他の灌漑面積の合計値が全灌漑面積で、井戸灌漑面積のみが欠損値であった場合には、全灌漑面積から水路、ため池、その他の灌漑面積を引いた値であるゼロを推定値として投入した。この方法により、15の変数で、ゼロの投入可能な欠損値が推定され、9変数については完全に回復された。

表3. ゼロの投入により推定された変数の内訳

ゼロ推定の方法	推定の詳細	変数の種類	変数の数
論理式による計算でゼロを推定	複数のカテゴリー変数とその合計変数の値から、欠損値を算出	灌漑種別面積、国立公園予算、不法占拠、サークル・レンジ・ビート数	15 (6) [※]
事実に基づいてゼロを推定	- 森林管理組合制度が未開始 - 該当する事業が対象外だった - 該当する階級が設けられていない	階級別の森林局員数、指定部族貧困事業予算、1999年年の森林組合数	7 (2) [※]
計			22 (8) [※]
次の段階での推定作業を必要とする変数			83 (8) [※]

※ ()内は第2段階で全ての欠損ケースを回復できなかった変数

次に、確認可能な事実に基づいてゼロを投入できる変数の特定を行った。たとえば、政府が実施している指定部族農村貧困層支援事業に対する2000-01年の支出額が欠損であった。欠損値のある州では指定部族人口が少なく、かつ政府が公表する同事業の実施対象州リストに掲載されていないことが確認できたため、ゼロを推定投入した。1999年の森林組合数の欠損については、該当する州では森林管理組合による森林管理制度が州政府によって認可されていないことを確認し、ゼロを投入した。森林局の階級別職員数については、各州森林局のウェブサイトから欠損する階

級が儲けられていないことを確認しゼロを推定投入した。この過程で、7の変数でゼロの投入可能な欠損値が推定され、5変数については完全に回復された。

上記のとおりゼロの推定投入を行った結果、97あった変数のうちで、22の変数は欠損値を推定することができた。そのうちで、14変数は欠損値が全て推定されて完全な変数に転換することができた。8つの変数については、ゼロの投入では推定できない欠損ケースが含まれていたため、本段階の推定作業を経ても完全化することができなかったものである。8変数は、ゼロを推定投入できなかった75の変数と共に、3段階目の推定作業を行うこととした。本段階で推定された変数とゼロ推定投入の詳細は表3の通りである。

(3) 論理式を用いた計算による推定

第3段階では、ゼロの推定投入で完全な変数とならなかった8変数と、ゼロの推定投入ができなかった75変数の計83変数についての推定作業を行った。ここで行ったのは、欠損値のなかで、ゼロの推定以外の論理的な計算による推定値の算出である。論理的な推定にあたっては、欠損のある変数に直接関連する別の変数を用いた。たとえば、1000人当たりの耕作面積については、数値が判明している耕作面積を州人口で割って算出した。また、灌漑種類別面積では、欠損値はインド北東に位置する州とシッキム州であった。これらの州は地理的に隣接し環境が類似していると仮定して、同じ北東に位置する州で欠損のない州の総面積に対する灌漑の比率の平均値から算出した。

表4. 論理式による計算で推定された変数の内訳

推定の方法	推定の詳細	変数の種類	変数の数
比率による算出	- 耕作面積÷州の全人口×1000=1000人あたりの耕作面積 - 灌漑種類別面積では、北東州とシッキムが類似の環境と仮定して、欠損のない北東州の州面積に対する灌漑の割合の平均をかけて算出 - 森林官全数に占めるフォレスターとレンジャーの比率の平均を欠損のある州のフォレスターにかけてレンジャーを推定した	一人当たり耕作面積、灌漑種類、フォレスター数、レンジャー数	11
引き算による算出	穀物灌漑面積－米灌漑面積＝小麦灌漑面積	小麦灌漑面積	1 (1) [※]
計			12 (1) [※]
次の段階での推定作業を必要とする変数			73 (1) [※]

※ ()内は第3段階で全ての欠損ケースを回復できなかった変数

上記の論理式による計算を行った結果、83あった変数のうちで、11の変数は欠損値を推定することができた。そのうちで、10変数は欠損値が全て推定されて完全な変数に転換することができた。1変数については、論理式による計算では推定できない欠損ケースが含まれていたため、本段階の推定作業を経ても完全化することができなかったものである。この変数は、論理式で推定できなかった72の変数と共に、4段階目の推定作業を行うこととした。本段階で推定された変数と論理式の詳細は表4の通りである。

(4) 多欠損変数の棄却

第4段階では、第3段階までの推定作業で完全な変数とならなかった73のうちで、依然として欠損ケースが7以上ある多欠損変数の棄却を行った¹³⁾。これは、第3段階までは論理的な考察に基づいて推定値を算出してきたものであるが、次の第5段階で回帰分析を行うにあたり、回帰分析による推定ケースが変数に占める比率を25%未満にとどめることを目的としたものである。この作業により、28の州のうちで7州以上が欠損ケースである13の変数が棄却された。欠損値が6ケース未満であった60の変数については、第5段階目の推定作業を行うこととした。本段階で棄却された変数は表5の通りである。

表5. 棄却された多欠損変数

変 数 の 種 類	変数の数
湖での養殖面積、汽水域での養殖面積、村落林面積、村当たりの村落林面積、不法伐採件数、不法伐採木材量、不法伐採被害額	13
計	13
次の段階での推定作業を必要とする変数	60

(5) 回帰分析による推定

第5段階では、欠損ケースが6つ未満の60変数について、回帰分析による推定作業を行った。

最初に、全657変数を対象に相関分析を行った。これにより、相関係数が高い変数を持つ欠損値のある変数のみを回帰分析の対象とした。回帰分析で、独立変数によって欠損ケースを持つ従属変数の変動の50%を説明できるよう、相関係数は0.707以上を基準に設定した。相関係数が0.707に満たない相関関係しか見出せないものが4変数あり、これらは棄却された。残る56の変

表6. 回帰分析による推定値の用い方

推定の方法	推定の詳細	変数の種類	変数の数
回帰分析の推定値が正の値であった場合	相関変数が高い独立変数を用いた回帰分析の推定値をそのまま適用	農業事業体従事者数、農家規模別耕作面積・農家数、種目別灌漑作付割合、不法伐採被害額	34
回帰分析の推定値が負の値であった場合	負の推定値を示した欠損ケースにつき、同様の負の推定値をとった州の実数の平均値を投入。若しくは平均値を用いた按分により推定値を算出	スラム人口、小麦灌漑面積、砂糖黍灌漑面積、支援を受けていない森林管理組合数	8
その他	回帰分析で推定した他の変数を利用して算出。	スラム、農地面積の平均、負債農家数、全穀物灌漑、全作物灌漑	14
計			56
最終的に推定値を投入できず棄却した変数			4

13) 多欠損変数の棄却をより早い段階で行わなかったのは、州の分割で生じた欠損ケースやゼロの推定投入可能な欠損ケースなど、論理的に推定可能な欠損値を多数持つ変数がほとんどであったため、各推定段階を経ても欠損が多い変数に限定して棄却する方法を採用したためである。

数については、統計分析ソフト SPSS の回帰分析プログラムから推定値を抽出することとした。

回帰分析においては、大きく分けて3つのケースが生じた。まず、回帰分析で抽出された正の値の推定値を適応可能な場合、次に回帰分析の推定値が負の値であったため論理式による計算を加えたもの、最後に回帰分析で推定された他の変数から推定が可能な場合である（表6）。

〈回帰分析の結果が正の推定値である場合〉

まず、相関係数が高い独立変数を用いた回帰分析を行い、そこで抽出された推定値をそのまま適用させるものである。この方法で推定できた変数は34変数であった。欠損値のある従属変数の種類、回帰分析に利用した独立変数、両変数の相関係数を表7に示す。

表7. 回帰分析で正の推定値となった場合の従属変数と相関係数

	従 属 変 数	独 立 変 数	相 関 係 数
1	2001年の全体の乳児死亡率（1,000出生あたり）（人）	2001年の都市の乳児死亡率（1,000出生あたり）（人）	.897
2	2001年の農村の乳児死亡率（1,000出生あたり）（人）	2001年の都市の乳児死亡率（1,000出生あたり）（人）	.837
3	1998年の農村の農業事業体従事者数（人）	1998年の農村の農業事業体数（件）	.988
4	1998年の都市の農業事業体従事者数（人）	1998年の都市の農業事業体数（件）	.983
5	1998年の農村の非農業事業体従事者数（人）	1998年の農村の非農業事業体数（件）	.985
6	1998年の都市の非農業事業体従事者数（人）	1998年の都市の非農業事業体数（件）	.990
7	2000年の零細農民の農地面積計（ha）	2001年の男性の家内工業従事者数（人）	.978
8	2000年の小規模農民の農地面積計（ha）	2001年の女性の労働者数（人）	.956
9	2000年の中小規模農民の農地面積計（ha）	2001年の純耕作面積（平方 Km）	.959
10	2000年の中規模農民の農地面積計（ha）	2001年の純耕作面積（平方 Km）	.915
11	2000年の大規模農民の農地面積計（ha）	2001年の石屋根の世帯数（件）	.832
12	2000年の零細農家数（件）	2001年の男性の周辺農業労働者数（人）	.944
13	2000年の小規模農家数（件）	2001年の女性の労働者数（人）	.957
14	2000年の中小農家数（件）	2001年の純耕作面積（平方 Km）	.915
15	2000年の中規模農家数（件）	2001年の純耕作面積（平方 Km）	.927
16	2000年の大規模農家数（件）	2001年の石屋根の世帯数（件）	.828
17	2000年の農村世帯数（件）：男性世帯主世帯との回帰分析の推定を投入	2001年の男性世帯主の世帯数（人）	.986
18	2000年の農家世帯数（件）	2000年の女児出生数（件）	.990
19	2000年の負債のある農家数（件）	2001年のヒンドゥ教の男性世帯主世帯数（件）	.968
20	2000年の米の灌漑作付面積（平方 Km）	2002年の穀物生産量（千トン）	.806
21	2000年のその他の作物の灌漑作付面積（平方 Km）	2000-01年の電気・ガス・水道の GSDP での額（ルピー）	.804
22	2000年の小麦の灌漑作付割合（％）	2002年の穀物耕作面積（千 ha）	.773
23	2000年の砂糖黍の灌漑作付割合（％）	1999年の都市貧困ライン（ルピー）	.861
24	2000年の綿花の灌漑作付割合（％）	2001年の総灌漑面積（平方 Km）	.783
25	2000年の全作物の灌漑作付割合（％）	2002年の穀物生産量（千トン）	.724
26	2000年の水路灌漑面積（千 ha）	2001年の純灌漑面積（平方 Km）	.930
27	2000年の全灌漑面積（千 ha）	2001年の純灌漑面積（平方 Km）	.988
28	2001-02年の森林・野生動物への第9次計画支出額	2000-01年の森林・野生動物への第9次計画支出額	-.914
29	2002年5月から2004年までに撤去された不法占拠面積（ha）	2001年のその他の水源でトイレがある世帯数（件）	.741
30	3年間の不法伐採の価値の総額（千万ルピー）	2001-02年の歳入（千万ルピー）	.827
31	2001年までの年額の不法伐採（千万ルピー）	2001年の指定部族認定県の面積（平方 Km）	.842
32	2004年のテリトリー区分のビート数（個）	2001-02年の歳入（千万ルピー）	.899
33	2004年の全ビート数（個）	2001-02年の歳入（千万ルピー）	.911
34	2006年の外国支援を受けた森林組合数（個）	2003年の合計の旧来の羊（頭）	.904

〈回帰分析の結果が負の推定値である場合〉

相関関係が0.9以上の高い相関を示しながらも、変数の種類が人口や面積といった実数であるために、その規模による影響が強く表れた変数には、抽出された推定値が負の値をとるものが生じた。たとえば、スラムの人口については、6州（アルナチャル・プラデシュ（Arunachal Pradesh）、ヒマチャル・プラデシュ（Himachal Pradesh）、マニプール（Manipur）、ミゾラム（Mizoram）、ナガランド（Nagaland）、シッキム）においてスラム総人口、男性スラム人口、女性スラム人口の3変数が全て欠損であった。これらの3変数は、いずれも都市の総人口との相関関係が高く、相関係数はそれぞれ、0.880、0.875、0.885を示した。しかし、都市の総人口を独立変数として回帰分析を行った結果、3変数の推定値は、全てで負の値をとった。同様に、これら3変数の値が欠損でない州においても、5州（ゴア（Goa）、ジャム&カシミール、メガラヤ（Meghalaya）、トリプラ、ウッタランチャル）では回帰分析の推定値が全て負の値となった。推定値が負の値となったのは、これらの州は人口規模が小さく、都市化の遅れていたことが影響したためだと考えられる。そのため、まず男性スラム人口の値が欠損値ではない州のうちで推定値が負の値となった5州について、州の男性都市人口に占める男性スラム人口の割合の平均を算出した。その結果、平均比率は9.153%であった。欠損値を持つ州も推定値が負の値であったため、スラム比率が5州に類似すると仮定し、欠損のある各州の男性都市人口に9.153%をかけることで推定値を算出した。女性スラム人口についても同様の推定作業を行い、スラム総人口については男女スラム人口の合算値をあてた。このような方法で推定できた変数は8変数であった。欠損値のある従属変数の種類、回帰分析に利用した独立変数、両変数の相関係数を表8に示す。

表 8. 回帰分析で負の推定値となった場合の従属変数と相関係数

	従 属 変 数	独 立 変 数	相 関 係 数
1	2001年の男性のスラム人口（人）	2001年の都市人口（人）	.875
2	2001年の女性のスラム人口（人）	2001年の都市人口（人）	.885
3	2000年の小麦の灌漑作付面積（平方 Km）	2001年の総灌漑面積（平方 Km）	.927
4	2000年のその他の穀物の灌漑作付面積（平方 Km）	2001年の純耕作面積（平方 Km）	.928
5	2000年の砂糖黍の灌漑作付面積（平方 Km）	2001年の総灌漑面積（平方 Km）	.831
6	2000年のため池灌漑面積（千 ha）	2001年の女性の主家内工業従事者数（人）	.804
7	2000年の井戸灌漑面積（千 ha）	2001年の純灌漑面積（平方 Km）	.983
8	2006年の何のプログラム支援も受けていない森林管理組合数（個）	2006年の森林管理組合数（個）	.911

〈関連する変数の推定値を用いて推定する場合〉

相関関係が高い独立変数を持つ従属変数であっても、以下の変数については回帰分析の推定値を用いず、関係する変数の回帰分析による推定値を利用して欠損値を算出した。これらの変数は、すでに欠損値を推定値できた変数の合算変数であるためである。この方法で推定できた変数は14変数であった。欠損値のある従属変数の種類、回帰分析に利用した独立変数、両変数の相関係数

を表 9 に示す。

表 9. 関連変数の推定値を用いて推定した独立変数

	従 属 変 数	独 立 変 数
1	2001年のスラム人口（人）	スラム男性人口スラム女性人口.
2	2000年の零細農家の平均農地面積（ha）	零細農家の農地面積÷零細農家数.
3	2000年の小規模農家の平均農地面積（ha）	小規模農家の農地面積÷小規模農家数
4	2000年の中小規模農家の平均農地面積（ha）	中小規模の農地面積÷中小規模農家数
5	2000年の中規模農家の平均農地面積（ha）	中規模の農地面積÷中規模農家数
6	2000年の大規模農家の平均農地面積（ha）	大規模の農地面積÷大規模農家数
7	2000年の全体の平均農地面積（ha）	全農地÷全農家数
8	2000年の農地面積計（ha）：	零細農地＋小規模農地＋中小規模農地＋中規模農地＋大規模農地.
9	2000年の農家数計（件）	零細農家数＋小規模農家数＋中小規模農家数＋中規模農家数＋大規模農家数
10	2000年の負債のある農家比率（％）	負債のある農家数÷全農家数×100
11	2000年の全穀物の灌漑作付面積（平方 Km）	米の灌漑作付面積＋小麦の灌漑作付面積＋その他の穀物の灌漑作付面積
12	2000年の全作物の灌漑作付面積（平方 Km）	穀物の灌漑作付面積＋砂糖黍の灌漑作付面積＋その他作物の灌漑作付面積
13	2000年のその他の灌漑面積（千 ha）	全灌漑面積－（水路灌漑＋ため池灌漑＋井戸灌漑）
14	2004年のファンクション区分のビート数（個）	全ビート数－テリトリー区分のビート数

3. 欠損値の推定結果

これら 5 段階の推定作業を終えたところ、欠損値を含む151の変数のうち、134の変数で欠損値が推定されて完全な変数に転換することができた。その結果、全657変数のうち、棄却された変数は全体の2.6％に相当する17変数にとどまり、97.4％に相当する640変数によってデータセットが整えられた。

V. ま と め

本稿では、インドの統計発展の概略と本調査で利用した統計資料の概要から、インドでは多様な統計資料が整備されており、ウェブサイトから入手できる状況であること、統計制度が行政統計に依存しているがために、統計の質的向上が課題であることを示した。そのうえで、各種統計資料に基づく包括的なデータ（657変数）を用いて、データセットの作成を行った。本論では、その過程で生じた多くの欠損値（151変数）の 5 段階での推定作業の詳細を示した。(1) 関連する何らかのデータを用いた按分での推定、(2) ゼロを推定投入可能な変数の推定、(3) 論理式による計算での推定、(4) 多欠損ケースの変数の棄却、(5) 回帰分析による推定の 5 段階の作業を行った

結果、欠損値を推定できた134の変数と欠損値のなかった506変数の計640変数から構成されるデータセットを作成した。

欠損値の推定作業は、一つの方法に限定されるものではない。変数や欠損値の性質、欠損が生じた背景など多様な現状に応じて、最も欠損値の現状を示すと考えられ、かつ誰が行っても同じ結果を得ることができる推定方法を採用することが重要である。本稿では、欠損のある変数を可能な限り妥当と考えられる方法を用いて推定し、変数を分析可能な状態に回復させることで、より多くの変数を用いた分析を行う準備を行ったものである。先行研究における欠損値の推定過程を踏まえたうえで十分にその妥当性を検証することはできていないが、著者の研究目標であるより多くの変数を包括した多変量解析の準備は十分に整えられたと考えられる。今後は、整備されたデータセットを用いて、クラスター分析、多変量回帰分析などを行うことで、多様な変数から説明される州の特性を明らかにし、仮説の構築を進めたい。

引用文献

- 岡部純一「統計制度の欠陥に関するインド統計評議会（2000-2001年）の定量的研究」『アルテス リベラレス（岩手大学人文社会科学部紀要）』第72号，岩手大学，2003年6月，pp. 55-78
- 久保田義喜「インド人口センサス・マイクロ版」『図書の譜：明治大学図書館紀要』11巻，明治大学，2007年4月，pp. 66-78
- 黒崎 卓「インドにおける貧困問題の現状と対策」『インドの経済問題と今後の効果的な対印経済協力の方策』，財団法人 国際金融情報センター，2005年3月，pp. 51-60
- 杉原薫「英領インドの貿易統計について」『アジア長期貿易経済統計データベースプロジェクト・ニュースレター』6号，大阪大学，1997年7月，pp. 4-7
- 法政大学統計研究所『統計研究参考資料 No. 80 インド統計制度の現状と課題（上）——2001年国家統計評議会報告から——』，法政大学統計研究所，2003年2月
- 柳澤 悠『インドにおける農業統計と農業生産推計』，一橋大学経済研究所中核的拠点形成プロジェクト ディスカッションペーパー No. D97-27，一橋大学経済研究所，1998年2月
- Jha, L. K., *"India's Forest Policies"*, Ashish Publishing House, 1994
- Press Information Bureau, *"Poverty Estimates for 2004-05"*, Press Information Bureau, Government of India, 2007
- Press Information Bureau, *"Forest Survey in India - Historical Perspective"*, 10 June, 2008
- Baines, Athelstane, *"The History and Development of Statistics in British India"* in Koren, John ed. *"The History of Statistics"*, Burt Franklin, 1918 (reprinted in 1970)

（受付 2009年10月22日）