

e-Stat で公開された統計情報に基づく合成人口データを用いたリアルスケール社会シミュレーションの展開

Developments of Real-Scale Social Simulations

Using Synthetic Population Data Based on Statistics Released Through e-Stat

村田 忠彦 (関西大学総合情報学部)

Tadahiko Murata, Faculty of Informatics, Kansai University

Abstract: 本報告では、発表者が、学際大規模情報基盤共同利用・共同研究拠点 (JHPCN) での共同研究成果として e-Stat の統計情報を用いて合成した日本全国の人口データを用いたシミュレーションの展開状況を報告する。2019 年から利活用を呼びかけており、今春にはコロナ禍での観光復興と感染予防を両立させるためのシミュレーションにも活用され、内閣官房、観光庁も高い関心を寄せている。匿名化と異なるアプローチでの個票データの活用方法について、報告会参加者との議論を深めたい。

1. はじめに

本稿では、著者らが学際大規模情報基盤共同利用・共同研究拠点 (JHPCN) および革新的ハイパフォーマンス・コンピューティング・インフラストラクチャー (HPCI) の共同研究・共同利用の枠組みで推進している保護レベル別合成人口データ配布システムを提案する。合成人口 (Synthetic Population) とは、公開されている統計を用いて合成した世帯単位の個票データであり、実社会を対象にした分析やリアルスケール社会シミュレーション (RSSS: Real-Scale Social Simulations) [1] に活用することができる。

近年、人間行動を含むコミュニティの動向を推定するための社会シミュレーションに注目が集まっている。最も有名な社会シミュレーションの一つは、Schelling [2] による分居モデルのシミュレーションである。Schelling のモデルでは、仮想的な空間に住む市民が、自らと同じ民族やグループの近くに住みたいという同族志向のもとに、どのように分居が構成されていくかが示されていた。分居モデルでは、必ずしも同族志向が高くなくとも、分居が構成されることが明示されている。

Schelling のモデルは社会シミュレーションの功績として非常に重要であるが、仮想空間でのシミュレーションであるため、現実社会に適用するためには、シミュレーションの観察から得られた理解を解釈し、現実社会での政策変数に当てはめる必要がある。現実社会を対象にした RSSS を行うことが可能であれば、シミュレーション結果の解釈が一層容易になることが期待できる。

RSSS を実施するためには、実際の規模の人口データが必要である。しかしながら、実際の人口データ (個票データ) は、統計法の被調査者保護の原則により、簡単に利用することはできない (日本において、調査票情報に含まれる個人情報、統計法第 39 条から第 43 条において調査票情報の保護規定が整備されていることから個人情報保護法ではなく統計法が適用される)。統計法において、オーダーメード集計、匿名データ、調査票情報の利用 (オンサイト利用) の制度により、2 次的に利用することができるようになっている。オンサイトで、匿名化された個票データを、サイトに用意された計算機を用いて処理することはできるが、大規模なシミュレーションの実施は容易ではない。そこで、本プロジェクト

トでは、個人情報情報を排除した個票データとして、公開されている統計に基づいて合成人口 (Synthetic Population) を作成している (人口合成手法の詳細については、文献[1]を参照)。

合成人口を用いたシミュレーションの事例として、小西ら[3]は、大阪府高槻市を対象に投票シミュレーションを実施し、投票率向上と投票所数最小化の 2 つの目的を最適化する投票所の配置方法を提案した。その際、高槻市の実人口の統計と選挙関心に関するアンケート結果をもとにして、投票所に割り当てられている各地域の年齢構成から投票率の推定を行なった。村田と濱口[4]は同じモデルを用いて、北海道函館市の投票シミュレーションを実施し、2016 年の参議院議員通常選挙において導入された共通投票所の有効性について検討した。この際も、函館市の統計を用いて、実際の人口構成の推計を行なった。市川と出口[5]は、東京都大島を対象に合成人口を用いた感染症シミュレーションを行ない、家庭を媒介した感染の様子を明らかにし、学校閉鎖の有効性を検証した。また、杜と村田[6]は日本の各世帯への公的年金制度の 25 年後、100 年後の都道府県別所得代替率を推計し、都道府県ごとに所得代替率に違いがあることを示した。さらに、市川[7]は埼玉県所沢市を対象にした合成人口と AED の設置情報を合わせることで、5 分以内に AED を取りに行くことができる 300m 圏内の人口カバー率を算出し、45% の住人が 300m 圏外に居住していることを示している。また、阿部と後藤[8]は、岩手県の三陸海岸沿いの 5 市町を対象に Cash for Work (CFW) のシミュレーションを行い、失業率を減らすために、CFW により身につけるべきスキルが地域によって異なっていることを示した。

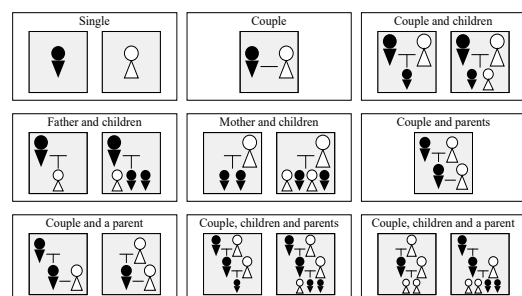
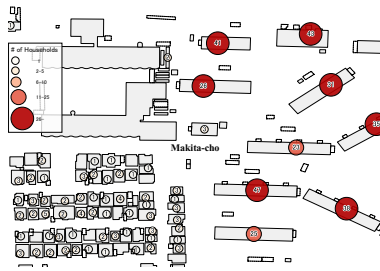


図1 合成対象の9種類の家族類型



(国土地理院基盤地図情報をもとに地図を作成)

図2 世帯の地図上への割当

表1 合成人口の合成状況

年度	2000	2005	2010	2015
市区町村	完了	完了	完了	完了
町丁目	完了	完了	完了	完了
建築物カバー率	地図 X	地図 X	42.6%	99.7%
就業状態・所得	X	X	完了	完了
ファイルサイズ	20GB	20GB	28GB	28GB

2. 日本全国の人口合成

上述のような実際の地域に居住する人々を対象としたRSSSを開発するためには、まずシミュレーションの対象となる人口合成が必要となる。各研究者が個別に行なっている人口合成の作業を軽減するため、村田ら[9-11]は日本における国勢調査(2015年)をもとにして、都道府県単位の実人口と同じ規模の人口合成を行なっている。各都道府県に居住する世帯の構成、年齢、性別[9]に加えて、現時点では所得や就労している産業分類の属性[11]が含まれている。村田らの合成人口では図1のような9種類の家族類型に対応する世帯が合成される。日本の95%の世帯がこの家族類型に該当する。また、原田と村田[10]は合成した世帯を図2のように地図上の建築物にマッピングすることにより、各世帯や世帯構成員の位置情報を考慮したシミュレーションや分析の実施を可能としている。このような合成人口データを日本全国の各市町村の町丁目単位で合成し、研究者に配布するシステムを構築した(合成人口データの例については、図4で説明する)。

村田ら[9-11]の研究では、都道府県単位の人口合成を行い、全国展開できる可能性を示していたが、著者らは学際大規模情報基盤共同利用・共同研究拠点(JHPCN)および革新的ハイパフォーマンス・コンピューティング・インフラストラクチャー(HPCI)の共同研究・共同利用の枠組みで推進するリアルスケール社会シミュレーション(RSSS: Real-Scale Social Simulations)[1]のためのプラットフォーム開発を行った。本プラットフォームでは、公開されている国勢調査の統計情報にもとづいて、合成人口(日本全国の世帯の個票データ)を合成し、表1のような形で日本全国の合成人口データの推計に成功している。

表1は2000年、2005年、2010年、2015年の国勢調査に基づいて各年度ごとに100セットの合成人口を合成した結果を示している。まず、全ての年度において、全都道府県の各市町村の全町丁目別の合成人口の100セット分の合成が完了している。続いて、建築物カバー率であるが、2015年の合成人口に対しては、日本国内の全町丁目の99.7%の町丁目

において合成世帯の建築物への割り当てが完了している。残りの0.3%は建築物がない町丁目であったため、建築物をもつ町丁目への割り当ては100%完了しているといえる。一方、2010年の建築物率への割り当てについては42.6%の町丁目の建築物への割り当てにとどまった。これは、2014年7月以前の国土地理院が提供する基盤地図情報については、7月31日以降の精度で全ての地図が用意されておらず、文献[1]の手法を全ての地域に対して得られなかったためである。2000年と2005年については、全ての地域で、必要な精度の地図を得ることができなかった。また、就業状態と所得は、2010年度と2015年度の合成人口に対して割り当てを行なっている。

最後にファイルサイズであるが、これは各年度の1セットの合成人口のデータサイズを示している。4年分100セットの合成人口は、約8TBから10TBのデータサイズとなる。

3. 保護レベル別合成データ配布システム

現在、国レベルの合成人口は、米国[12]と英国[13]で公開されているのみであり、アジア初のRSSS基盤として、日本の合成人口を公開する意義は大きく、実規模の社会シミュレーションの研究コミュニティの活性化が期待できる。また、合成人口の合成においては、統計的にデータに整合する様々な世帯の合成が考えられることから、合成データは一意に定まるものではないため、本プラットフォームでは、複数セットの合成人口を提供している。このように、本プラットフォームの合成人口は、世界的に見てもRSSSを実施するための貴重なデータであり、現実的な社会シミュレーションに取り組む研究者に必要なデータを供給することにつながる。

現在、本プラットフォームでは、図3に示すように大阪大学サイバーメディアセンターのOCTOPUSを用いて、現実と同じ規模の世帯構成をもつ人口を合成し、北海道大学情報基盤センターのハイパフォーマンスインテーククラウドを用いて、合成人口データベースを構築している。合成人口データの研究者向けの提供を行うために、北海道大学のクラウドを用いて公開できる体制を整えている。さらに、合成したデータを保存するため、東京大学情報基盤センターと理化学研究所で運用されているHPCI共用ストレージによりバックアップを行なっている。

合成人口データは、公開されている統計情報から合成されているため、個人のプライバシーに抵触する情報を含んでいないが、町丁目単位の特定のエリアにどのような世帯や年齢の市民が住んでいる可能

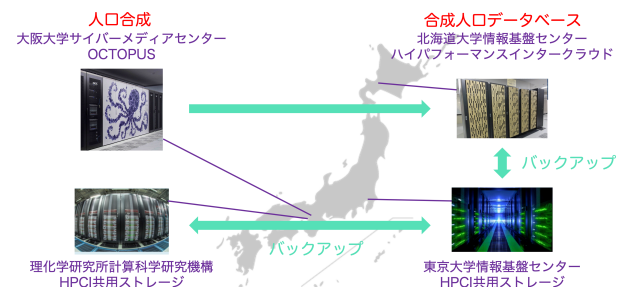


図3 合成人口データベースプロジェクト



都道府県 ID:27 大阪		市区町村 ID:27207 高槻	町丁目 ID:27207209002 霊仙寺町 2 丁目
緯度 34.876799		経度 135.575662	
世帯 ID:5		世帯類型 ID:4 女親と子供世帯	世帯人員数 2
個人 ID:7	年齢 38	性別 ID:1 女性	役割 ID:21 子供 (女性)
雇用形態 ID:10 一般労働者		産業分類 ID:90 I 卸売業・小売業	
企業規模 ID:1000 1000 人以上		所得 (月額) 300,000 円	
個人 ID:8	年齢 76	性別 ID:1 女性	役割 ID:11 妻・女親
雇用形態 ID:N/A 非就業者		産業分類 ID:N/A N/A	
企業規模 ID:N/A N/A		所得 (月額) 0 円	

※非就業者と賃金構造基本統計調査対象外の産業分類（農業、林業、漁業、公務の一部）は所得 0 円，他の就業属性 N/A

図 4 合成人口における世帯情報の例

表 2 データ保護レベル

データ 保護レベル	粒度	就業状況と所得※1	対象※2
1	都道府県	×	学部生
2	市区町村	×	学部生
3	都道府県	○	院生
4	市区町村※3	○	院生
5	町丁目	×	院生
6	位置情報	×	研究者
7	町丁目	○	研究者
8	位置情報	○	研究者

※1 研究内容により就業状況のみのデータ提供の場合がある。

※2 研究内容により，高いレベルのデータ提供も考慮する。

※3 人口 1,000 人未満の市区町村を除く。

性があるかを容易に調べられるようになるため，取り扱いを慎重にすべきであると考えている。そこで，合成人口を利用するにあたって，利用者に次の点の同意を求めている [14-16]。

- 1) 合成人口は，現実の世帯や個人の情報を含まない。
- 2) 合成人口は，世帯や個人の現実の統計と同様の特徴をもつ。
- 3) 合成人口は，合成時に使用していない統計との整合性は保証しない。
- 4) 合成人口は，統計が新たに公開された際や手法が改良された際に更新されることがある。
- 5) 合成人口を用いたシミュレーションや分析は，複数の合成人口データを用いて行わなければならない。
- 6) 合成人口を用いたシミュレーションや分析の結果は，現実の世帯や個人に結びつくような形で公開してはならない。

さらに，利用目的に応じた合成人口データを提供するため，図 4 のような合成人口データに対して，表 2 のようなデータ保護レベルを設けて，保護レベル別のデータ提供を行なう。図 4 は，実際には居住者が存在しない関西大学高槻キャンパスの校舎に，ある世帯が居住していると仮定した場合の当該世帯の情報を表している。表 2 では，データの粒度を都

道府県，市区町村，町丁目，位置情報（建物座標）として，それに就業状況と所得の情報をつけるかどうかでデータ保護レベルを設定している。利用者の範囲を広げるため，学部生であれば，都道府県と人口 1,000 人以上の市区町村単位のデータの利用を可能とする。所得データの利用については，院生または研究者の利用に限るものとする。

3. 保護レベル別合成人口利用例

本節では，1. に示した合成人口を利用した研究事例に基づいて，保護レベル別合成人口利用例を，主に，各研究者が開発したモデルと，合成人口に含まれる属性を連携させている点に焦点をあてて示す。

レベル 3 公的年金シミュレーション[6]

杜と村田[6]は，都道府県単位の人口動態を推計しながら，世帯タイプ別の所得代替率の計算を行なっている。各世帯構成員の雇用状態により年金被保険者のカテゴリ（第 1 号，第 2 号，第 3 号）ごとに年金の加入状況を計算し，世帯単位の所得代替率を推計している。

レベル 4 労働市場シミュレーション[8]

阿部と後藤[8]は，市区町村単位の各世帯の世帯構成員の年齢と性別，就業状態を用いて，獲得しているスキルを推計するとともに，世帯別の所得に連動した就職活動確率に基づいて，Cash for Work (CFW) の制度への応募をシミュレートしている。さらに，CFW により新たに獲得されるスキルを組合せることにより，労働市場における最終的な就業状態を推計している。世帯構成員が獲得しているスキルは，労働研究機構が定義している職業スキルを参考として，9 種類のスキルを上位／下位の階層構造で定義し，就業している産業や年齢・性別に応じて，各スキルの割り当ての確率の上下関係を決定した後，労働市場の実指標値とシミュレーション結果が整合するように確率を調整している。

レベル 6 投票シミュレーション[3, 4]

小西ら[3]は，位置情報が与えられた各世帯の世帯構成員の年齢をもとに選挙への関心を推計し，世帯の位置から投票所までの位置を組み合わせ，有権者個々の投票行動を推計している。選挙への関心は，明るい選挙推進協会が国政選挙の際に収集しているアンケート結果に基づいて推計している。

レベル6 AED 配置分析[7]

市川[7]は、位置情報が与えられた各世帯の世帯構成員の年齢をもとに、心停止発生確率を与えている。心停止患者が発生した場合、患者本人以外の世帯構成員がいる世帯であれば、AED を取得しに行くことができるため、単独世帯を除く世帯と市内に設置されている AED との距離を計算し、AED の人口カバー率を計算している。心停止発生確率は、総務省消防庁が公開している救急蘇生統計に基づいて推計している。

レベル7 (所得なし) 感染症シミュレーション[5]

市川と出口[5]は、町丁目単位の世帯情報をもとに世帯構成員の通学・通勤先を推定し、学校や勤務地における感染、世帯内における感染をシミュレートしている。

レベル8 (所得なし) 感染症シミュレーション[17]

倉橋と永井[17]は、合成人口データを用いることで、世帯の地理的な分布と、その地域での就業者の産業分類を考慮したモデルを構築し、観光客を受け入れた場合の新型コロナウイルスの感染シミュレーションを実施している。観光客の立寄り箇所の制限と、観光客と接する就業者への PCR 検査を実施することにより、市内での感染拡大を抑えられる可能性を提示している。

4. おわりに

本研究では、2000 年、2005 年、2010 年、2015 年の国勢調査に基づいた日本全国の人口データの合成とデータベースの整備に関して報告し、申請可能なデータ属性と、各データ属性を用いた研究事例の整理を行った。本プロジェクトでは、今後も新たな合成人口手法の考案や、新たな国勢調査結果の公開にあわせて、人口合成を行なっていく。人口合成の利用申請は本稿末尾の連絡先記載のウェブサイトで行なっている。

謝辞

本研究は、JSPS 科研費 17K03669, 20K10362 の助成、学際大規模情報基盤共同利用・共同研究拠点および革新的ハイパフォーマンクス・コンピューティング・インフラ (jh190056-MDH, jh200022-DAH), HPCI 共用ストレージ (hp190215) の支援を受けて推進いたしました。記して御礼申し上げます。

参考文献

- [1] 村田, 原田, 杉浦: リアルスケール社会シミュレーションの実現に向けて, 日本シミュレーション学会論文誌, 34 (4), pp. 58-62, 2017
- [2] T. Schelling: Dynamic models of segregation, *J. of Mathematical Sociology*, Vol. 1, pp. 143-186, 1971
- [3] 小西, 村田, 名取: 投票率上昇と投票所数削減のための投票シミュレーション, 知能と情報, 22 (2), pp. 203-210, 2010
- [4] 村田, 濱口: 投票参加モデルを用いた共通投票所の有効性の検証-函館市における平成 28 年参議院議員通常選挙の事例-, 2017 年度日本選挙学会研究会方法論部会: シミュレーションモデリングと選挙研究 (香川, 5 月 20-21 日, 2017) 6 ページ, 2017
- [5] 市川, 出口: 感染症実用シミュレーションにおける仮想都市構築法の違いによる結果への影響分析-日常生活スポット内包セル型仮想都市モデルの必要性-, 計測自動制御学会論文集, 49 (11), pp.1012-1019, 2013
- [6] 杜, 村田: 公的年金制度における 47 都道府県の所得代替率に関する考察, システム制御情報学会論文誌, 29 (9), pp.422-431, 2016
- [7] 市川: 医療分野におけるリスクマネジメント 地理情報分析と社会シミュレーション技術を用いた検討, 計測と制御, 57 (6), pp.407-412, 2018
- [8] 阿部, 後藤: 震災復興過程における地域特性を踏まえた有効な CFW (Cash for Work) のシミュレーション分析, 計測自動制御学会システム・情報部門第 22 回社会システム部会研究会, pp.22-29, 2020
- [9] T. Murata, T. Harada, D. Masui: Comparing Transition Procedures in Modified Simulated-Annealing-Based Synthetic Reconstruction Method Without Samples, *SICE JCMSI*, 10 (6), pp. 513-519, 2017
- [10] T. Harada, T. Murata: Projecting Households of Synthetic Population on Buildings Using Fundamental Geospatial Data, *SICE JCMSI*, 10 (6), pp.505-512, 2017
- [11] 杉浦, 村田, 原田: 賃金構造基本統計調査に基づく合成世帯集団の労働者への所得の割当て, システム制御情報学会論文誌, Vol.32, No.2, pp.70-79, 2019
- [12] W.D. Wheaton, J.C. Cajka, B.M. Chasteen, D.K. Wagener, P.C. Cooley, L. Ganapathi, D.J. Roberts, J.L. Allpress: Synthesized Population Databases: A US Geospatial Database for Agent-Based Models. No. MR-0010-0905 (RTI Press), 12 pages, 2009
- [13] J. Robards, C.G. Gale, D. Martin: Creating a Synthetic Spatial Microdataset for Zone Design Experiments, National Centre for Research Methods Working Paper 17/5, pp.1-40, 2017
- [14] T. Murata, T. Harada: Synthetic and Distribution Method of Japanese Synthesized Population for Real-Scale Social Simulations, Proc. of 33rd Annual Conf. of the Japanese Society for Artificial Intelligence, 3B4-E-2-05, 3 pages, 2019
- [15] T. Murata, T. Harada, M. Ichikawa, Y. Goto, L. Hao, S. Date, M. Munetomo, A. Sugiki: Distribution of Synthetic Populations of Japan for Social Scientists and Social Simulation Researchers, Proc. of Int'l Conf. on Machine Learning and Cybernetics, 4 pages, 2019
- [16] 村田, 原田: 日本の全人口合成データ配布システム, 第 35 回ファジィシステムシンポジウム講演予稿集, 4 pages, 2019
- [17] 倉橋, 永井: 観光地における COVID19 感染予防策, <http://www.u.tsukuba.ac.jp/~kurahashi.setsuya.gf/doc/slide-2020-n06-r2.pdf>, 2020

連絡先

E-mail: murata@kansai-u.ac.jp

Website: <http://www.res.kutv.kansai-u.ac.jp/~murata/rsss-distribution/>