jh250064

合成人口プロジェクト：エージェントへの生活行動時間割当て

村田忠彦（大阪大学）

概要

本研究課題では，居住者の時間単位の日常生活を統計的に合成するため，生活行動時間の割り当てを行う．すでに本プロジェクトでは，居住地ベースの統計である国勢調査を用いた夜間人口分布と，従業地ベースの統計である経済センサスを用いた昼間人口分布の合成を行なってきたが，リアルスケール社会シミュレーション（RSSS）を実現するためには，個々人の活動がどの時間帯に行われているかを推定する必要がある．日本在住者を対象とした生活時間について全数調査は行われておらず，抽出調査のみが行われているが，代表的なものとしては，NHK の「国民生活時間調査」と，公的統計として「社会生活基本調査」が実施されており，それぞれ，7,200 人と約 20 万人が対象となっている．本プロジェクトは，これまで公的統計を対象としてデータ合成を行っており，今回も「社会生活基本調査」を用いて，都道府県別の生活時間の割り当てを行うことにより，RSSS プラットフォームを構築する．

1. 共同研究に関する情報

⑴ 共同利用・共同研究を実施している拠点名

北海道大学 情報基盤センター

東京大学 情報基盤センター

大阪大学 D3 センター

(2) 課題分野

データ科学・データ利活用課題分野

(3) 参加研究者一覧と役割分担

総括・人口合成アルゴリズム

村田忠彦・原田拓弥・李　皓・
堀上駿太・松村優大

人口合成プログラム配備

伊達　進

合成人口 DB 構築・運用・バックアップ

棟朝雅晴，杉木章義，塙　敏博

合成人口 DB インタフェース

市川　学，後藤裕介

2. 研究の目的と意義

本研究課題は，2019 年度から継続的に採択されており，2021 年度から「合成人口プロジェクト」(以下，本プロジェクト) と名付けて，各年度の課題を推進している．本プロジェクトでは，実際の地域を対象にした実規模の世帯構成をもつ人口データを用いた社会シミュレーションであるリアルスケール社会シミュレーション（RSSS: Real Scale Social Simulation）の基盤を構築するため，国勢調査をもとに日本全体の市区町村ごとの世帯構成員を含む人口データを合成するとともに，合成された人口データを RSSS で活用するためのデータベースの作成に取り組んでいる．国レベルの合成人口データを先行して合成し，公開してきたのは，米国，英国およびベルギーであり，本プロジェクトは，アジア唯一の RSSS 基盤として，日本の合成人口データを RSSS 研究を推進する研究者に提供してきた．

本共同利用・共同研究拠点の計算資源によ

り合成された人口データを用いて次のような研究が展開されている．() に研究期間と研究代表者名を示す．●は現在進行中の研究を示す．関西大学研究拠点形成支援事業（20/4-22/3, 村田忠彦), 科学研究費基盤研究 C（20/4-23/3，村田忠彦)，内閣官房 COVID-19 AI・シミュレーションプロジェクト（20/7-23/10，村田忠彦)，進化計算学会コンペティション（21/12，後藤裕介)，科学研究費基盤研究 A（22/4-23/3，諏訪晴彦)，●科学研究費基盤研究 C（22/4-25/3，後藤裕介)，●科学研究費若手研究（22/4-25/3，原田拓弥)，●JST 未来社会創造事業本格研究（23/4-28/3，貝原俊也)，●科学研究費基盤研究 A（24/4-28/3，大澤幸生)．

今年度は，社会生活基本調査に整合する生活行動時間の割当ての実現に向けて，合成人口データの個々の世帯構成員に対して生活行動時間を割り当てるアルゴリズムの開発に取り組む．

3. 当拠点の公募型共同研究として実施した意義

本プロジェクトは，2017 年と 2018 年に大阪大学サーバーメディアセンター公募型利用制度「若手・女性研究者支援萌芽枠」で開始し，2019 年度から JHPCN で合成人口に関する共同研究を実施している．研究代表者の村田は，2020 年 4 月から 22 年 3 月に，当時の所属である関西大学の研究拠点形成支援事業により，経済学・社会学・情報学・医療の分野への合成人口データの利活用の研究を推進した．さらに，2020 年 7 月から 2023 年 10 月まで，内閣官房 COVID-19 AI・シミュレーションプロジェクトに合成人口データを提供した．さらに 2020 年 11 月から JST 未来社会創造事業探索研究で合成人口データを用いた研究を行った．2021 年 12 月には進化計算学会において社会シミュレーションの最適化コンペティションを実施し，査読論文として成果を報告した[1-5]．JST 未来社会創造事業の探索研究として採択された研究課題は，2023 年 4 月から本格研究に採択され，自治体と共同して取り組むデジタル社会実験プロジェクトを展開している．

本プロジェクトにより合成された人口データは，これまで本プロジェクトの共同研究者が所属する岩手県立大学，芝浦工業大学，静岡大学，大阪大学以外に，北海道大学，東北大学，筑波大学，東京大学，東京女子大学，東京工業大学，中央大学，京都大学，奈良先端科学技術大学院大学，大阪工業大学，大阪医科薬科大学，神戸大学，富山県立大学，滋賀県立大学，東京理科大学，早稲田大学，慶應大学，青山学院大学，国際医療福祉大学，創価大学，関西大学，国立保健医療科学院，国立情報学研究所，防災科学技術研究所，聖路加国際病院から問合せをうけ，合成人口データの配布を行なってきた．今年度はさらに，宮城大学，東京都立産業技術大学院大学，武蔵野大学，産業技術総合研究所の研究者にも提供した．これらの研究機関では，戦略的イノベーション創造プログラム (SIP3)，科学研究費基盤研究 A，同基盤研究 B，同基盤研究 C，JST 未来社会創造事業などの研究プロジェクトが展開されており，様々な規模の研究支援を行っている．

本年も本プロジェクトが提供した合成人口データを用いた多数の研究が実施され，計測自動制御学会社会システム部会で発表された 69 件の発表のうち，優秀賞２件（岡本貴大, 原田 拓弥, LLM 駆動型社会シミュレーションに向けた合成人口データへの個性の割り当て，堀上 駿太，村田 忠彦，合成人口データにおける全自治体の就業者への従業地割当て)，学生賞３件（Bosong Cheng, Tadahiko Murata，EV オーナー分布を考慮した EVCS 配置最適化，辻 優樹，原田 拓弥，深層学習を用いた建物の利用用途判別と判別結果を用いた合成人口における世帯割り当て，酒井和, 坂田顕庸, 高橋真吾，市民なりき

出典：国土交通省
基盤地図情報

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）
所得（月額）300,000 円			産業分類 ID:90 I 卸売業・小売業
雇用形態 ID:10 一般労働者			企業規模 ID:1000 1000 人以上
個人 ID:8	年齢 76	性別 ID:1 女性	役割 ID:11 妻・女親
所得（月額）0 円			産業分類 ID:N/A N/A
雇用形態 ID:N/A 非就業者			企業規模 ID:N/A N/A

図１：合成人口データの世帯情報（大阪府高槻市霊山寺町のサンプル）

りゲームに向けた 市民の行動特性に基づくロールプレイング向けペルソナの生成）は，いずれも合成人口データを用いた研究であり，リアルスケール社会シミュレーションを実施する研究者コミュニティにおいて，合成人口データを用いることが標準となり，高い評価を受けている．

4. 前年度までに得られた研究成果の概要

本プロジェクトは 2019 年度から継続して採択されている．**2019 年度**は，大阪大学の計算機を用いて，図１のような日本全国の個票データを国勢調査に基づいて合成した．図１は大阪府高槻市にある建築物に居住する世帯の一例である．本プロジェクトでは，図１に色分けされた属性を含めたデータの提供を行っている．さらに，北海道大学のインタークラウドシステムを用いたデータベースを維持すると共に 2020 年 2 月から HPCI 共用ストレージにてバックアップを行っている．

2020 年度は以下の研究により，合成人口データの精度を向上した．

・合成世帯の建物へのマッピング
・施設世帯を含めた人口合成
・人口動態に関する研究
・シミュレーションプラットフォーム開発
・救急医療の環境整備に関する研究
・新型コロナウイルス感染症に関する研究
・地震時の避難行動に関する研究
・ベーシックインカムに関する研究

2021 年度は，次の２つの課題に取り組んだ．

・就業者の従業地の割当てに関する研究
・建造物の利用用途推定に関する研究

従業地割当てアルゴリズムの研究は **2021 年度計測自動制御学会第 27 回社会システム部会優秀賞を受賞**し，2022 年度には IEEE Transactions on Computational Social Systems に採択され，AAAS Eurek Alert!や，芝浦工業大学，関西大学からプレスリリースされた．まず，国勢調査から得られる常住地による従業市区町村別就業者数の分布比率に基づいて各従業地（市区町村）に割当てる人数を決定する．表２の表 A より，就業者の常住地（市区）別，産業分類別，従業地（市区町村）別の構成比が得られ，これに比例した人数の合成人口データを無作為に抽出することで従業地（市区町村）を割当てる．これが公開されていない市町村においては，表 B および表 C を用いて段階的に割当てる手法を提案した．次に，表 D を用いて，割当てた従業地（市区町村）において，経済センサス-基礎調査を用いてさらに詳細な小地域への割当てを行った．

建造物の利用用途推定については，空中写真から住宅の利用用途を判別するアルゴリズムを用いて，合成人口データの適切な割り当てを試みた．首都圏と中部圏のデータを用いて U-Net を訓練し，近畿圏のデータで評価

表２：割当てる属性と使用する統計表の出典

割当てる属性	使用する統計名	公開されている対象	表題／境界名称	表
従業地（市区町村）	平成27年国勢調査 / 従業地・通学地による人口・就業状態等集計（人口，就業者の産業（大分類）・職業（大分類）など）	21大都市（東京都区部と政令指定都市）とその区，県庁所在市，人口20万以上の市	常住地による従業市区町村，産業（大分類）別15歳以上就業者数	A
		全市区町村	常住地又は従業地（9区分）による雇用者（3区分），産業（大分類），男女別15歳以上就業者数	B
			常住地による従業・通学市区町村，男女別15歳以上就業者数及び15歳以上通学者数（15歳未満通学者を含む通学者－特掲）	C
従業地（小地域）	平成26年経済センサス-基礎調査 / 町丁・大字別集計	全市区町村	経営組織（２区分），産業（中分類）・従業者規模（６区分）別全事業所数及び男女別従業者数－市区町村，町丁・大字	D
従業地（小地域の図形中心点座標）			平成26年経済センサス-基礎調査 町丁・大字別境界データ	－

表３：従業地における従業者の統計表

使用する統計名	公開されている対象	表題/境界名称	統計表
平成 27 年国勢調査/従業地・通学地による人口・就業状態等集計	21 大都市（東京都特別区部および政令指定都市）とその各区，県庁所在市，人口 20 万以上の市	従業地による常住市区町村，産業（大分類）別 15 歳以上就業者数	E
	全市区町村	常住地又は従業地（9 区分）による雇用者（3 区分），産業（大分類），男女別 15 歳以上就業者数	B

したところ，近畿圏の建物（面積 25 m^2 以上）の内，82.8 %の用途を正しく判別できた．

2022 年度には，mdx 上で合成人口 DB インタフェースを構築するため，608vCPU で仮想ディスク 100GB の環境を確保し，北海道大学で運用している合成人口データをコンテナ化することにより，mdx の Kubernetes 上に移動させ，保護レベル別合成人口データベースを構築した．

2023 年度には，2020 年度に実施された国勢調査に基づく合成人口データの合成に取り組んだ．2020 年調査結果のうち，人口合成に必要な統計表でもっとも遅いものは 2023 年 3 月に公開される．2020 年度から一部の統計の公開フォーマットが変更されたため，2023 年度での合成には至らなかったが，JHPCN 事務局の支援により，2024 年 3 月から総務省の連携先の検討を行い，東京大学空間情報科学研究センター（CSIS）の仲介で，統計センター情報システム部の担当者との意見交換を開始することができている．まだ，2020 年度の人口合成には至っていないが，公的統計の公表方法に関する継続性の観点での議論を進めている．

2024 年度には，従業地属性を追加する研究

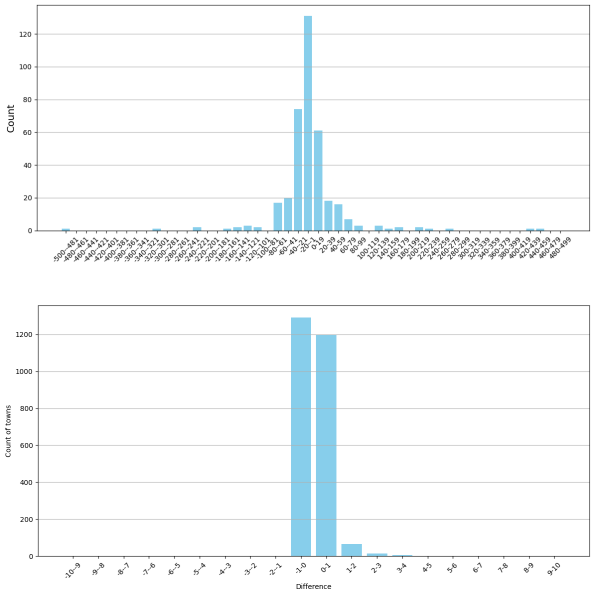


図２：従業者数の一致度別小地域数

を実施し，2021 年度の従業地割当てアルゴリズムの全国の自治体への拡張を行った．従来手法を用いて，従業地自治体での就業者数の割り当てを行うと図２上のように小地域単位で，経済センサスの就業者数との乖離が目立つことがわかった．就業者が超過・不足している小地域を調整するアルゴリズムを考案して実装したところ，図２下のように誤差の少ない割当てが可能となった．一方，従業地自治体の産業分類の割合が保持されるた

め，小地域の就業者数の合致度を高めることに限界があることが判明した．その点は，2026 年度の研究課題として取り組むことにしている．

5. 今年度の研究成果の詳細

生活行動時間に関する国内の代表的統計として NHK による「国民生活時間調査」と政府による「社会生活基本調査」がある．NHK による調査は 1960 年から 5 年ごとに行われている．2020 年の調査では，全国 150 地点から無作為抽出で選ばれた 10 歳以上の 7200 人を対象に実施され，4247 人が回答している．一方，政府の調査は 1976 年から 5 年ごとに行われ，2016 年の調査では，10 歳以上の世帯構成員 20 万人以上を対象とし，20 万人が回答している．対象者数が，政府の方が多く，地域による生活行動時間の違いも計測できることから，本研究では，政府統計である社会生活基本調査をもとに，生活行動時間の合成に取り組む．

中間報告で示したように，次のような 3 つの評価関数を用いて，探索的に統計に沿った生活行動時間の生成を行った．

・評価関数 1 ：連続した睡眠を 1 回として，睡眠回数 s を計測し，$s-1$ をペナルティとする（睡眠回数を 1 回とする）．同様に通勤回数 c を計測し，$c-2$ をペナルティとする（通勤回数を往復の 2 回とする）．仕事の回数 w に対しては，$w-1$ をペナルティとする（仕事の回数は 1 回とする）．これらの 3 つのペナルティの総和の最小化を行う．

・評価関数 2 ：別の種類の行動に切り替えた回数を *change* とし，対象人口 N 人の *change* の総和を最小化する．

・評価関数 3 ：合成した 20 種類の生活行動の 1 日の行動者率との二乗誤差を最小化する．これにより，全員が睡眠することや，仕事の割合等の統計とのずれが最小化される．

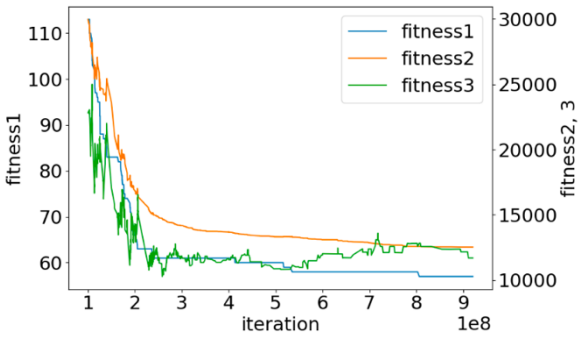
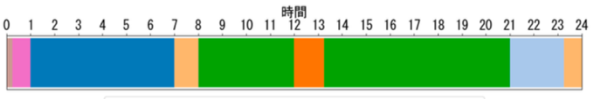
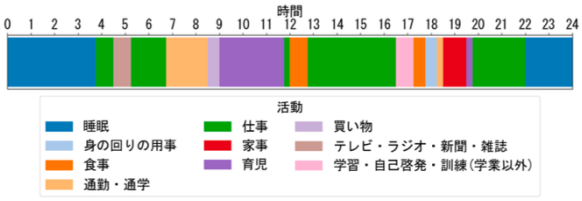


図 3 ：評価関数値の最小化の推移



(a) ある個人の 1 日のタイムスケジュール 1



(b) ある個人の 1 日のタイムスケジュール 2

図 4 ：合成された生活行動時間

上記の 3 つの評価関数に対して，辞書式最小化を行った．すなわち，評価関数 1 の値が小さな解を優先した解の探索を行なった．シミュレーティッドアニーリング法を用いた解探索においては，初期解として生成された N 人の生活行動時間に基づいて，評価関数を計算し，評価関数 1 から優先的に最小化を行なった．初期解は，15 分ごとの 20 種類の生活行動者率が統計に一致するように合成している．隣接解の生成手法としては，評価関数 1 のペナルティの大きな個人を対象にして，ペナルティの原因と思われる行動を同時刻の他の行動との入替えによって行なった．

図 3 に評価関数値の最小化の推移を示す．評価関数 1 の最小化が優先されるため，評価関数 1 が小さくなる際に，他の評価関数が悪化することが観察されている．図 4 に合成された生活行動時間を示す．

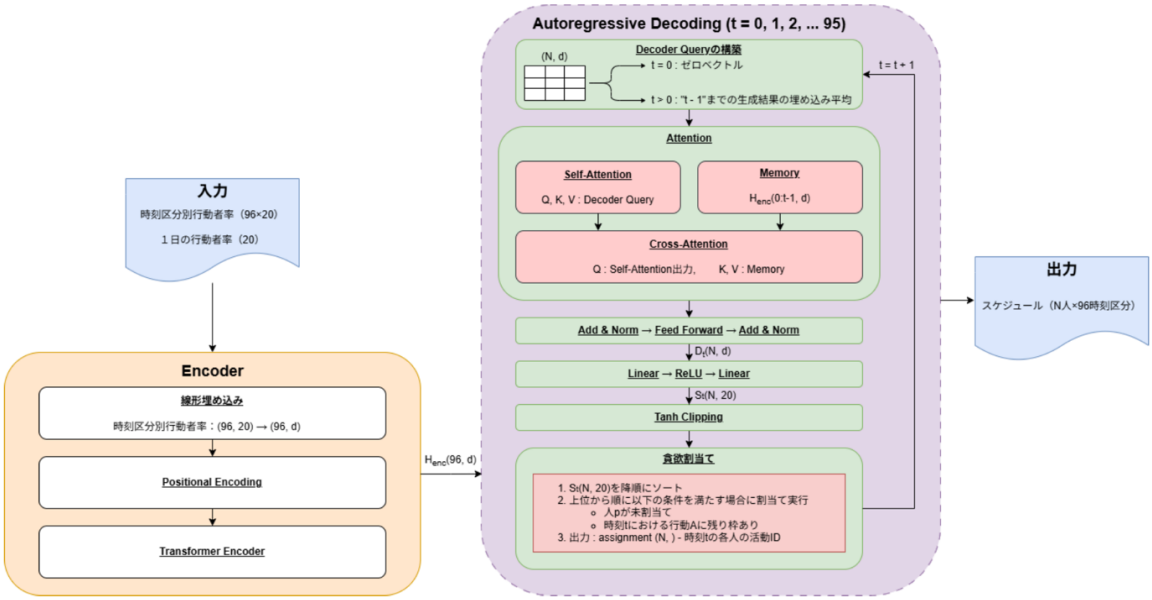
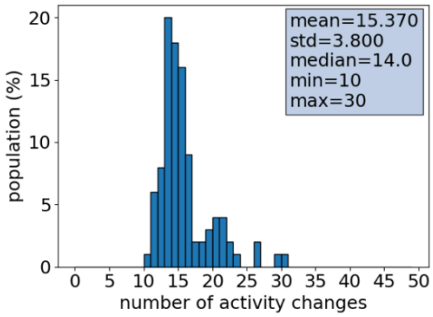
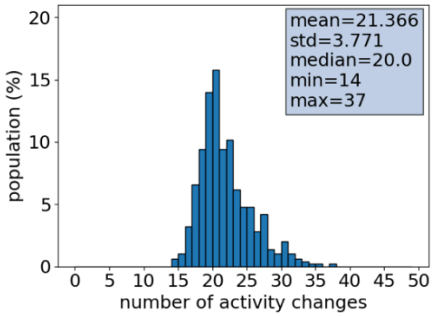


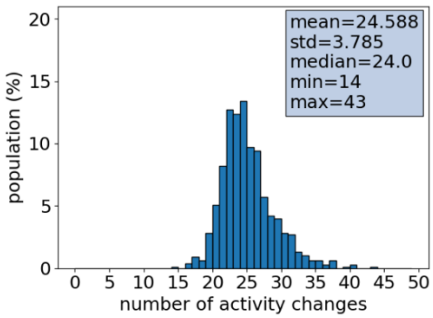
図６：Transformer Encoder-Decoder Architecture



(a) 100 人に対する生成



(b) 500 人に対する生成



(c) 1,000 人に対する生成

図５：辞書式探索手法による行動切替回数分布

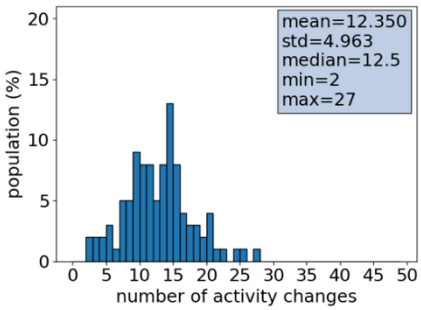
表４：人数別の生活行動生成時間

人数（人）	時間
100	1,661 秒　（27.7 分）
500	53,937 秒　（899.0 分）
1,000	325,835 秒（5,430.6 分）

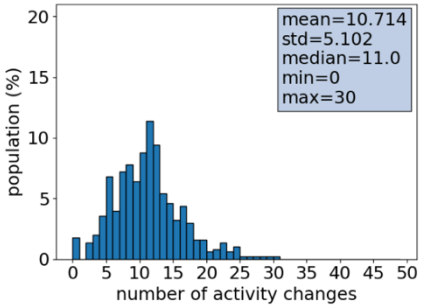
本アルゴリズムで，統計上の行動割合にある程度即した生活行動時間が集合としては合成されているものの，合成された個々の生活行動時間を考慮すると，行動の切り替えが頻繁に行われているような，現実的ではないと思われる生活行動時間が合成されている点が課題となっていた．図５に示すように合成対象人数が切り替え回数が多くなる結果となっていた．また，辞書式探索手法では，対象とするスケジュールの人数の増加に伴い，表４に示す計算時間と誤差が増加する結果になっていた．

そこで下半期では，Transformer の Encoder-Decoder アーキテクチャと強化学習を組合せた機械学習手法によるスケジュール生成モデルの構築を行った．具体的には，図６のような構造のモデルを用いて，スケジュール生成モデルを構築し，方策勾配法を用いて学習を行った．学習後のモデルから生活

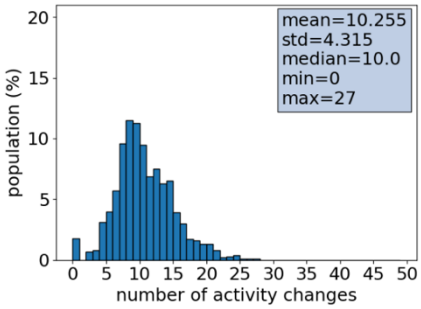
行動を生成したところ，図６のような切替回数の分布をもつ行動が生成できた．図５と図６を比較すると，人数が多い場合にも，切替回数が増加しない生活行動データが生成できていることがわかる．また，表５の一人当たりの生活行動生成時間では，辞書式探索手法(SA)では，生成人数が多くなるにつれて，計算時間が著しく増加しているが，機械学習手法では，一人当たりの生成時間がほぼ均一になっていることがわかる．



(a) 100 人に対する生成



(b) 500 人に対する生成



(c) 1,000 人に対する生成

図６：機械学習手法による行動切替回数分布

表５：一人当たりの生活行動生成時間

時間（秒）	100	500	1,000
SA	16.61	107.87	325.83
ML	0.32	0.35	0.36

なお，学習にかかる時間は 100 人，500 人，1000 人のモデルのそれぞれで，49 分，296 分 448 分であった．表４と比較すると，100 人については，機械学習モデルの方が，学習時間を含めた計算時間は長くなっているが，人数が増えるにつれて，機械学習モデルの方が高速になっていることがわかる．一方，人数分の学習にかかる時間が線形に増加しているため，大人数の生成を行うためのアルゴリズムを考案する必要がある．

6. 進捗状況の自己評価と今後の展望

本研究課題では，生活行動データの生成に取り組んだ．上半期では，SA 法を用いた辞書式探索手法に取り組み，行動の切替回数と計算時間に課題があることを確認した．下半期では，行動の切替回数と生成時間に問題のないアルゴリズムを構築することができた．大人数の生活行動時間の生成にあたっては，生成に不確実性を導入することにより，多様な生活行動時間が生成できると考えている．

今後の展望として，今回は，47 都道府県の男女の生活行動時間統計に基づいた学習を行ったのち，滋賀県，男性，一般労働者の統計を入力として，生活行動時間の生成を行った．異なる都道府県，性別，労働形態の生成を行うことにより，それぞれの特徴をもった生活行動時間が生成できているかどうかを検証していく．

※7. 研究業績はウェブ入力です