ファクターによる人口動態に関連する行動確率計算による人口推定の試み

○李皓（静岡大学）

Sample Manuscript for SICE Symposium of Social Systems Section

* H. Lee (Shizuoka University)

概要－　人口をマクロスコープの数字ではなく，一人ひとりの意思決定に至る理由やその結果のプロセスの解明することで，複雑化する現代の人口動態の全体像を掴み，人口問題の解決に繋がる．本研究では，我々はアンケート調査の個票を用いて，性別・年齢・社会階層などの諸要因に基づく，自然動態に関連する確率算定に試みる．個票データは公開されているオープンデータを二次利用し，手法としてはカプラン・マイヤー法・離散時間ロジットモデルと機械学習を試みる．

キーワード:　人口推定，離散時間ロジットモデル，行動確率

1　はじめに

様々な社会シミュレーション手法が存在する中，人や組織のミクロモデルを構築し，その相互関係に基づいてボトムアップに社会システムを表現する手法はエージェントベースシミュレーション(以下 ABS)である．他のシミュレーション手法よりも多元的な社会モデルの記述が可能であり，様々な要素が絡み合う社会現象への理解や,政策評価などの可能性が期待されている．人口動態は社会・心理・経済など，様々な要素が絡み合う複雑な社会現象である．人口をマクロスコープの数字ではなく，一人ひとりの意思決定に至る理由やその結果のプロセスの解明することで，複雑化する現代の人口動態の全体像を掴み,人口問題の解決に繋がる．

我々は文化的・社会的・経済的・心理的・医学的な要素を学術横断的に導入することで，人口増減のみ推定する経済的なモデルではなく，社会的な価値のあり方について検討することが出来る社会モデルを設計し，様々な社会課題を多元的に検証出来るシミュレーションモデルを構築する．

本研究では，社会調査に基づくエージェントベースモデルの構築を目指し，SSM調査研究会が10年一度に行う「社会階層と社会移動」全国調査（SSM調査 2005）の個票データに対して分析を行い，婚姻や出産の行動確率の推定を行う．その上で，我々が地方都市の住民を対象に行う社会調査の質問項目について検討する．

2　SMM調査について

SSM調査は，1955年以来，10年に一度行われる大規模調査であり，これまで7回の調査を行ってきた．最新の調査は2015年に行われているが,2018年1月の現時点では調査結果が公表されていないため，我々は現在公表されているSSM2005年のデータに対して分析を行う．SSM2005の概要はTable1のようにまとめた．

SSM2005の個票データはオープンデータとして，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で公開されている．所定の手続きで申請することで，個票データを入手することができる．

メインとなるデータのファイルはタブ区切りのテキストファイルとなっており，5,742列 x 807行のデータが格納されており，ファイルサイズは約10MBである．データファイルの他，ラベルファイルや記録コードなどデータ解釈用のファイルがある．

3　データ分析

Table 1: SSM2005の概要

調査代表者	佐藤嘉倫
共同グループ名	2005 年 SSM 調査研究会
抽出方法	層化 2 段・等間隔抽出
抽出台帳	選挙人名簿，住民基本台帳
調査方法（モード）	訪問面接法 訪問留置法
調査員	民間調査機関
調査開始年月	2005, 11
母集団地域	日本全国
母集団性別	男女
母集団年齢	20 － 69
標本数	13031
回収数	5742
回収率	44.06%
調査の領域	経営・産業・労働，社会変動，教育，社会心理・社会意識，階級・階層・社会移動，コミュニケーション・情報・シンボル，余暇・スポーツ，性・世代，社会学研究法・調査法・測定法，差別問題，政治・国際関係

SSM 調査の概要は前述の通りであるが，留置調査と訪問調査が異なる問いを行っており，807 種類のデータに欠損値のないデータは存在していない．また，質問内容によって，例えば配偶者のない方に対して配偶者の職業に関する問いでは「非該当」，あるいは年収や保有資産に対する問いでは「未回答」が多く，そのままの状態で分析することは非常に困難である．

欠損値を無くしつつ，データの量と次元を確保するために，我々は留置調査と訪問調査を精査し，共通的な質問を整理した上で，利用するデータを抽出した．結果的に，居住地域名・地域のサイズ・性別・生年・兄弟姉妹の数・親の兄弟姉妹の数・学歴・支持政党・政党好感度・性差に関する意識・子育てに関する意識・社交の頻度・相談相手の数・結婚年齢・雇用状態・職種・勤務先の規模，の計 17 種類のデータを抽出した．

これらを用いて，結婚年齢の推定を行った．手法はまずカプランマイヤー法を用いて性質を調査し，その後は離散時間ロジットモデルによる年齢別属性別の婚姻ハザード確率の計算と，いくつかの機械学習の手法で分析を行い，結婚年齢を推定するためのファクターと，その精度について調査した．

カプランマイヤー法は，t時点まで未婚である人の数をn_t(t時点で結婚した場合も含む)，t時点で結婚をした人の数をd_t とすると，t時点において推定される未婚継続率S_t は式(1)で表される．

第15回社会システム部会研究会，第59回システム工学部会研究会（2018年3月14日-16日・沖縄）　PG0002/18/0000-0254 © 2018 SICE

$$\hat{S}_t = \Pi_{i=1}^{t}\left(\frac{n_i - d_i}{n_i}\right) \quad \cdots(1)$$

z v

カプランマイヤー法で算定した未婚継続率の例を Fig.1 で示す．縦軸は未婚継続率であり，横軸は年齢である．Fig.1 では，居住地域によって結婚タイミングや生涯未婚率に違いがあることが分かる．

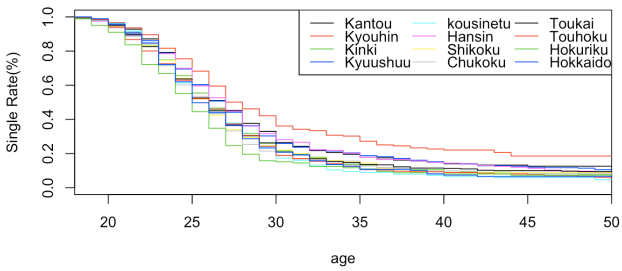

Fig. 1: カプランマイヤー法(男性・年齢・地域)

離散時間ロジットモデルとはt時点までに結婚イベントが発生していない前提に，t時点でイベントが発生する確率 P(t)を予測するモデルである．離散時間ロジットモデルは式(2)と式(3)で表される．x_kは共変量，a_kは回帰係数，n は説明変数の数である．

$$P(t) = \frac{1}{1 + \exp(-z(t))} \quad \cdots(2)$$

$$z(t) = b(t) + \sum_{k=1}^{n}(a_k x_k(t)) \quad \cdots(3)$$

評価のために，まず SMM 調査データを学習用データとテスト用のデータを分けた．その方法として，データのシリアル ID を 9 で割り，余剰が 0 にならないデータ群(N=5068)をパーソン・ピリオドデータに変換し，学習用データとした．その他のデータ群から未婚者を除いた上でテスト用データ(N=638)とした．R の GLM でロジットモデルを学習用データで回帰係数を推定した．Table.2 は計算された回帰係数の一例である．

Table 2: 離散時間ロジットモデル(26~33 歳,男性)

属性		値	属性		値
(Intercept)		−1.16	勤務先規模	1 人	−0.07
年齢		−0.04		2〜4 人	0.11
地域	京 浜	−0.32		5〜9 人	0.11
	近 畿	0.41		30〜99 人	0.04
	九 州	0.20		100〜299 人	0.12
	甲信越	0.39		300〜499 人	0.21
	阪 神	−0.02		500〜999 人	0.69
	四 国	0.35		1000 人以上	0.36
	中 国	0.48		DK･NA	−0.35
	東 海	0.34		官公庁	0.42
	東 北	0.23	勤務状態	家族従業者	0.85
	北 陸	0.43		学生	−11.63
	北海道	0.41		契約社員	0.06
生年		−0.02		経営者，役員	0.96
兄弟姉妹数		0.05		自営業主	1.23
親の兄弟数		0.03		就職前	−1.34
共産党好感度		−0.05		常勤従業者	0.76
子ども価値観		0.02		派遣社員	−0.14
				無職	−0.40
				臨時雇用	−0.12

Table 2 のパラメータを利用し，結婚イベントが発生する確率を計算できる．例えば[1980 年生まれ,30 歳の男性,京浜在住,兄弟 2 人,親の兄弟 5 人,共産党好感度 5,子ども価値観 5,勤務先規模は官公庁,勤務状態が常勤従業者]の結婚確率は以下の式で計算できる．結果として，この男性が 30 歳の間に 6.9%の確率で結婚する．

$$\frac{1}{1 + \exp(-1 * (-1.16-0.04*30-0.02*80+0.05*2+0.03*5-0.05*5+0.02*5+0.42+0.76))} = 6.9\%$$

計算した回帰係数を用いて，結婚年齢を計算する Java プログラムを作成し，テストデータを用いて結婚年齢を推定したが，確率モデルであるため，実行する度に異なる結果になる．誤差を検証するために，結婚年齢の計算を 100 回行い，その中からもっと出現回数の多い結婚年齢を使って誤差を 100 回計算し，その平均で評価した．

機械学習に関しては，R の Caret パッケージを利用し，ニューラルネットワーク，ランダムフォレスト，勾配ブースティングなどを利用し，前述のパラメータ群を採用し，学習用データでモデルを学習し，テストデータの推定と実績の差を検証した．誤差の大きさを比較するための指標として，RMSE と MAE を採用し，その結果を Table 3 でまとめた．結果として本研究で検証した手法の中では，離散時間ロジットモデルはもっとも誤差の小さい結果となった．

Table 3: 各モデルの RMSE と MAE

	RMSE	MAE
離散時間ロジットモデル	5.59047	3.15695
ニューラルネットワーク	11.29952	8.10256
ランダムフォレスト RF	10.92032	7.67486
ランダムフォレスト Ranger	10.95922	7.88942
ランダムフォレスト Rborist	10.88775	7.85608
勾配ブースティング xgbLinear	11.98776	8.46585
勾配ブースティング xgbTree	10.92858	7.97432

4 おわりに

我々は，自ら行う社会調査を設計するために，既存の大規模社会調査の個票データに対してデータ分析を行い，結婚行動に影響を与える要素について，様々な手法を用いて検証した．その結果，離散時間ロジットモデルはもっとも精度が高く，結婚の確率に影響を与える要素としては，年齢の他，居住地域，勤務先規模，勤務状態などの影響が大きいほか，生年（コーホート）や兄弟数，価値観などの影響も受けることが分かった．

結論として，社会調査にこれらの要素に関連する設問を用意する必要があることを明らかになった．

本研究は「JSPS 17H02038」の助成を受けたものである．

参考文献

1) 李皓，社会調査ベースのエージェントシミュレーションによる人口推定の構想，2017 年 9 月，計測自動制御学会 システム・情報部門 第 14 回社会システム部会研究会
2) 李皓，大規模社会調査に基づく市民モデルの設計ー SSM2005 を事例に，2017 年 11 月，計測自動制御学会 システム・情報部門学術講演会 2017