自由回答データにおける代理回答バイアスの推定

上原　一輝[1]・川野　倫輝[2]・円山　琢也[3]

[1]学生会員 熊本大学工学部社会環境工学科（〒860-8555 熊本市中央区黒髪 2-39-1）
[2]学生会員 熊本大学大学院自然科学研究科社会環境工学専攻（〒860-8555 熊本市中央区黒髪 2-39-1）
[3]正会員 熊本大学准教授 くまもと水循環・減災研究教育センター（〒860-8555 熊本市中央区黒髪 2-39-1）
E-mail:takumaru@kumamoto-u.ac.jp

社会調査において，調査対象者本人以外が回答することで生じる代理回答バイアスについては，広く認識されているのにも関わらず，それを体系的に分析する方法の構築や検討は十分ではない．本研究では自由回答データに着目し代理回答バイアスの実態の一部を明らかにすることを目的とする．具体的には，2012 年熊本 PT 調査の付帯調査の自由回答の記入内容と先行研究に従って算出した代理回答確率の関係を分析する．分析の結果，代理回答される確率が大きくなるほど自由回答中の文字数が少なくなることを示した．また，代理回答される確率で特徴的な語が異なるという結果を示した．

Key Words: *proxy response, open-ended survey data, social survey, bias*

1. はじめに

パーソントリップ(PT)調査などの土木計画・交通計画で利用される社会調査の方法論が訪問型から郵送自記回答方式や Web 回答方式に変化している．調査方法の変更にともない，回収率の低下，サンプル構成の歪みなどの課題が指摘されてる．本研究では，自記式，Web 回答方式の調査の課題のうち，代理回答の問題に着目する．代理回答とは調査対象者本人以外が回答することで，それにともなう調査結果の歪みを代理回答バイアスと呼ぶ．

PT 調査は世帯単位で調査の依頼がなされ，世帯構成員すべての移動についての回答が求められる．子どもや高齢者の回答は世帯内の他者が代理回答をしている可能性が考えられる．特に Web 調査回答においては代理回答の可能性が高いことが考えられる．現実には高齢者が私事トリップを行っていても，それを代理回答された場合には，記入されない可能性が発生する．

この点は，調査結果の根幹に関わる基礎的な点であるのにもかかわらず，既存調査・研究での検討は十分とはいえない．特に代理回答を体系的に分析する方法論は未構築といえた．

この課題に対して，細谷ら[1]は，世帯内で誰が調査に回答するのかという事象を，集団意思決定型の離散選択モデルで表現し，代理回答される確率の理論式を導出している．そして，2012 年熊本 PT 調査のデータを利用して，代理回答されているとモデルから推測されるサンプル集団は，平均トリップ数が小さくなっていることを示した．すなわち，平均トリップ数という調査結果における代理回答バイアスを実証的に示すことに成功している．

ここで，代理回答バイアスは，他の調査結果にも影響を与えていることが想像され，本研究では自由回答データを取り上げる．代理回答されている場合，自由回答設問へは無記入が多いこと，回答量が少ないことが予想される．さらに回答内容についても，対象者本人が答える場合とは異なるトピックが記入されていることも予想される．

以上の背景を踏まえて，本研究では社会調査の自由回答における代理回答バイアスの実態の一部を明らかにすることを目的とする．

方法論としては，細谷ら[1]により開発された PT 調査の回答有無を集団意思決定として表現した離散選択モデルを利用する．利用するデータは 2012 年熊本 PT 調査の付帯調査として実施された「60 歳以上の方の外出に関する意識調査」における「外出時の問題点・不満点」に関する自由回答データである．

本研究では PT 調査データを利用するが，代理回答バイアスを精査することは，土木計画・交通計画で利用される調査だけでなく，すべての社会調査の結果の分析において重要と考えられる．

社会調査における代理回答の問題は広く認識され，指摘されているが，研究例は限られている[2,3,4,5]．交通調査の事例では，海外の研究で代理回答の有無を収集したデータを利用した研究例が見られる[6,7,8,9]．我が国の今

後の交通調査実施においては，代理回答の有無も尋ねることが重要であることが示唆される．

本研究で利用する細谷ら[1]の方法論の有用性は，代理回答の有無のデータが収集されていない場合も，それを推測することができる点にある．推測に基づく限界はあるが，代理回答の有無の設問がない既存の調査データにも適用可能である点が特徴となる．このような方法論を利用して代理回答を分析している研究例は，筆者に知る限り存在せず，本研究の独自性となる．

2. 熊本 PT 付帯調査と代理回答確率推測モデルの概要

(1) 熊本 PT 付帯調査データの概要

利用するデータは 2012 年熊本 PT 調査の付帯調査として実施された「60 歳以上の方の外出に関する意識調査」における「外出時の問題点・不満点」に関する自由回答データである．

この付帯調査は，PT 調査対象世帯のうち，60 歳以上の高齢者のいる世帯に無作為に配布され，8,188 人分のデータを回収している．そのうち，自由回答に回答した人は 1,733 人，自由回答に回答しなかった人は 6,455 人であった．

ここで，付帯調査回答者の代理回答確率を算出するには，付帯調査の回答者と PT 調査世帯票中の個人を対応させる必要がある．しかし，一部の付帯調査回答者のデータは，記入違い等が原因で PT 調査世帯票中の個人との対応付けが不可能である．このようなデータを除き，自由回答に回答した人の中で代理回答を算出できたのは 1,521 人分のデータであった．よって，以下の分析では，断りのない限りこの 1,521 サンプルを用いて分析を行う．

(2) 代理回答確率推測モデル

細谷ら[1]による代理回答確率モデルは以下のように定式化される．世帯iにJ_i人の世帯構成員がいるとする．このうち一人が世帯を代表して回答すると仮定する．世帯iが調査に回答する場合，j番目の構成員が回答する確率を$P_{i,j}$とする．また，無回答の場合を$j=0$とし，その確率を$P_{i,0}$とする．無回答の効用関数$V_{i,0}$，および世帯構成員jが回答を行う効用関数$V_{i,j}$を以下のように定義する．

$$V_{i,0}=\sum_k \alpha_k x_{k,i},\ V_{i,j}=\sum_l \beta_l y_{l,j} \tag{1}$$

式(1)において$x_{k,i}$は世帯属性に関する説明変数，$y_{l,j}$は個人属性に関する説明変数，α_k，β_lはそのパラメータを示す．$P_{i,0}$および$P_{i,j}$に多項ロジットモデルを適用すると選択確率は以下の式で表すことが出来る．

$$\begin{aligned}P_{i,0}&=\frac{expV_{i,0}}{expV_{i,0}+\sum_{j=1}^{J_i} expV_{i,j}}\\&=\frac{exp\sum_k \alpha_k x_{k,i}}{exp\sum_k \alpha_k x_{k,i}+\sum_{j=1}^{J_i} exp\sum_l \beta_l y_{l,j}}\end{aligned} \tag{2}$$

$$\begin{aligned}P_{i,j}&=\frac{expV_{i,j}}{expV_{i,0}+\sum_{j=1}^{J_i} expV_{i,j}}\\&=\frac{exp\sum_l \beta_l y_{l,j}}{exp\sum_k \alpha_k x_{k,i}+\sum_{j=1}^{J_i} exp\sum_l \beta_l y_{l,j}}\end{aligned} \tag{3}$$

本調査では，世帯 i の中で何番目の構成員が回答したのかは不明である．しかし本モデルより世帯が回答する確率は$\sum_j P_{i,j}$で与えられることを利用する．このとき,実現した状態を示す同時確率L^*と対数尤度Lはそれぞれ以下で表され，この式を利用してパラメータの最尤推定を行う．

$$L^*=\prod_{i\in N} P_{i,0}^{1-\delta_i}\left(\sum_{j=1}^{J_i} P_{i,j}\right)^{\delta_i} \tag{4}$$

$$\begin{aligned}L&=\ln L^*\\&=\sum_{i\in N}\left\{(1-\delta_i)\ln P_{i,0}+\delta_i \ln\sum_{j=1}^{J_i} P_{i,j}\right\}\end{aligned} \tag{5}$$

$$\delta_i=\begin{cases}1 & :\text{世帯}i\text{が回答の場合}\\0 & :\text{世帯}i\text{が無回答の場合}\end{cases} \tag{6}$$

以上のモデルを各段階に応じて適切な説明変数を考慮し，適用していく．なお，このモデルは張ら[10),11)]が提案した集団意思決定モデルにおける Max-Max モデルの 1 種と等価である．この詳細は，佐藤，円山[12)]を参照されたい．

次に，A を個人jが回答する事象，B を世帯iが回答する事象とする．また，世帯の回答確率を$\sum_{j=1}^{J_i} P_{i,j}$，個人の回答確率を$P_{i,j}$とすると，代理回答される確率P_{proxy}は条件付き確率を利用して以下で与えられる．

$$\mathrm{P(B)}=\sum_{j=1}^{J_i} P_{i,j},\ \mathrm{P(A)}=P_{i,j}$$
$$\mathrm{P}\left(\overline{\mathrm{A}}\right)=1-P(A)=1-P_{i,j} \tag{7}$$

$$\begin{aligned}\mathrm{P}\left(\overline{\mathrm{A}}\cap B\right)&=P(B)-P(A)\\&=\sum_{j=1}^{J_i} P_{i,j}-P_{i,j}\end{aligned} \tag{8}$$

$$\begin{aligned}\mathrm{P}\left(\overline{A}\middle|B\right)&=\frac{P(\overline{A}\cap B)}{P(B)}\\&=\frac{\sum_{j=1}^{J_i} P_{i,j}-P_{i,j}}{\sum_{j=1}^{J_i} P_{i,j}}\\&=P_{proxy}\end{aligned} \tag{9}$$

表-1　世帯単位のPT調査回答有無選択モデル推定結果[1)]

説明変数		パラメータ	t値	
世帯属性(α_k)				
定数項		2.62	9.57	***
単身世帯		−0.04	−0.47	
世帯人数		0.22	10.98	***
第1,2種低層住居専用地域		−0.56	−2.62	***
第1,2種中高層住居専用地域		−0.23	−3.12	***
近隣商業・商業地域		−0.06	−0.61	
準工業工業地域		−0.22	−1.52	
個人属性(β_i)				
男性	20歳代ダミー	0.09	0.32	
	30歳代ダミー	0.79	2.84	***
	40歳代ダミー	1.17	4.19	***
	50歳代ダミー	1.67	5.30	***
	60歳代ダミー	2.17	7.37	***
	70歳以上ダミー	2.24	7.38	***
女性	20歳代ダミー	1.18	4.50	***
	30歳代ダミー	1.82	7.03	***
	40歳代ダミー	2.21	8.14	***
	50歳代ダミー	2.24	8.26	***
	60歳代ダミー	2.17	7.55	***
	70歳以上ダミー	1.46	4.53	***
サンプルサイズ			13,279	
ρ^2			0.192	
修正済みρ^2			0.190	

*:10%有意, **:5%有意, ***:1%有意

α_k:回答ありだとパラメータ低い, β_i:回答ありだとパラメータ高い

本研究では，細谷らの推定結果(表-1)を使用して代理回答確率を求めていく．

3.　自由回答データへの代理回答の影響分析

2 章式(9)を用いて代理回答される確率を算出し分析を行う．

図-1 は性別で代理回答される確率ごとの人数分布を示している．代理回答される確率が高くなるにつれて女性の割合が大きくなることから，高齢者において男性より女性の方が代理回答されやすいということが読み取れる．

図-2 は年齢別で代理回答される確率ごとの人数分布を示している．代理回答される確率が高くなるにつれて高年齢の割合が大きくなることから，年齢が高い人ほど代理回答されやすいということが読み取れる．

図-3 は回収分類別で代理回答される確率を示している．Web の方が代理回答される確率が 50%以上の割合が 5 割を超えており，調査票と比較すると代理回答されやすいということが読み取れる．

図-4 は代理回答される確率別で文字数の比率を示している．代理回答される確率が高くなるにつれて 0 文字もしくは 1-10 文字，11-30 文字といった少ない文字数の割合が大きくなっていることが分かる．このことから，自由回答に記入する人において，自分で記入する人は自由回答に長文で回答する傾向にあることが読み取れる．

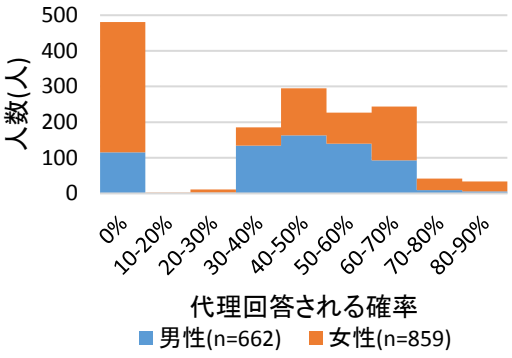

図-1　性別代理回答される確率ごとの人数分布

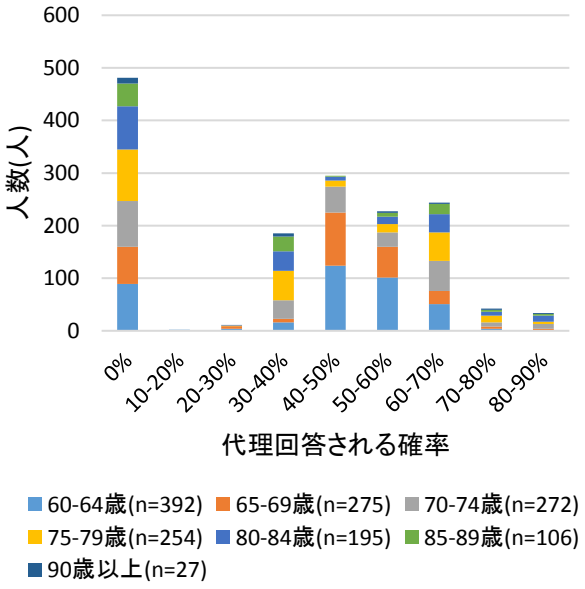

図-2　年齢別代理回答される確率ごとの人数分布

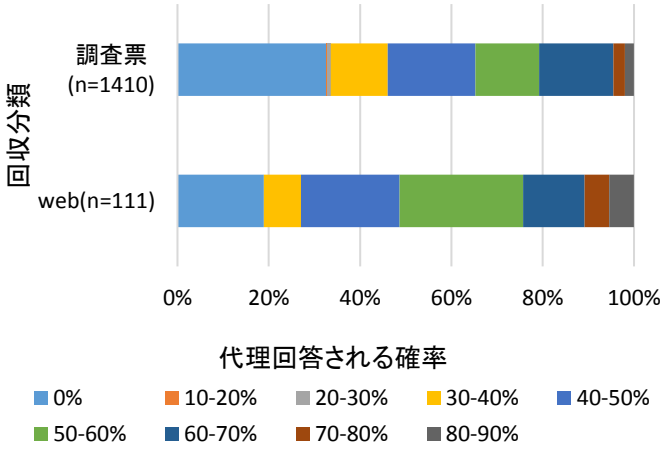

図-3　回収分類別代理回答される確率の分布

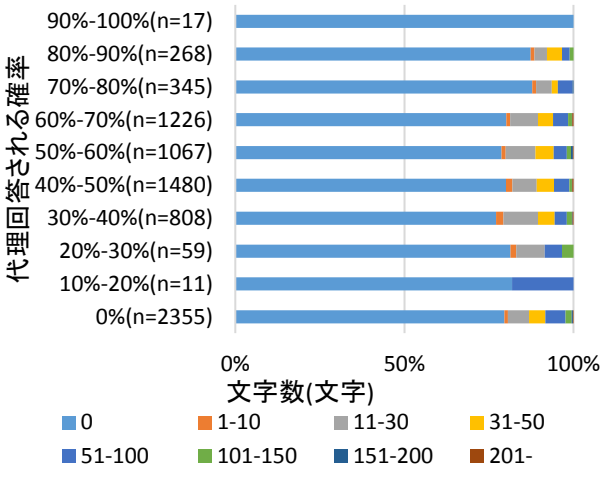

図-4　代理回答される確率別文字数比率の分布

4.　個人属性と自由回答で記入された語の対応分析

個人属性を用いた対応分析を行うことで，回答者の属性と自由回答で記入された語の関係性を見る．ここで，原点に近い要素ほど分析対象中では平均的に出現しており，原点や他の要素群からの距離が大きい要素ほど相対的に異なる出現傾向を持つ要素である．また，異なる変数間の関係は原点からの方向で解釈を行う．本研究では，「性別と年齢」「代理回答確率」に分けた分析を行う．

用いる品詞は，名詞，動詞，形容詞，形容動詞，副詞である．さらにここから，ひらがなのみで表記される名詞，動詞，形容詞，副詞，さらに非自立の形容詞を除いた．これは，例えばひらがなのみで表記される動詞であれば「ある」「いる」「する」のように，単体では解釈が困難なためである．同じく非自立の形容詞も「づらい」「がたい」のように，単体では解釈に困難なため除いている．ここから出現頻度 21 語以上の語を抽出し，さらに属性間の差がないという帰無仮説に対応したカイ二乗値が大きいものから 80 語をプロットしている．

(1)　性別と年齢

図-5 は，回答者の性別と年齢を用いて対応分析を行ったものである．

性別について見ると，第二象限は女性，第四象限は男性の回答において特徴的な語が位置している．男性の回答としては「自動車」「自家用車」が特徴的であることが分かる．原文を見ると，「今は自動車が乗れるのでいいのですがもう少し年をとったら僻地なのでたいへんだと思う(男性・60 歳代)」「自家用車にのれなくなった時バス停まで遠く不便である(男性・70 歳代)」のように用いられており，現在は自動車で移動しているが将来的に不安を抱いている人が多いということが特徴的であった．また「通行」「危険」などの自転車や自動車で通行する際の危険性に関する語も特徴的であることが分かる．原文を見ると，「自転車外出の時，道路通行が大変危険(男性・60 歳代)」のように用いられている．これに対して，女性では「バス」「バス停」などのバス利用に関する語や「回数」「少ない」「不便」などの，バスをはじめとした公共交通機関に対する否定的な語が特徴的であることが分かる．原文を見ると，「バスの回数が少なく不便なので利用していない(女性・70 歳代)」「バス停は近くにあるが運行本数が少ない(女性・70 歳代)」のように用いられている．

年齢について見ると，原点から縦軸正方向には 60-74 歳の現役と前期高齢者，第三象限には 75 歳以上の後期高齢者において特徴的であった語が位置している．60-74 歳の回答としては「運転」「駐車」「渋滞」などの自動車に関する語や，「駅」「バス」などの交通機関に関する語が特徴的である．これに対して，75 歳以上では「足」「不自由」「病院」などの語や，「段差」「階段」などの語が特徴的であり，身体的な衰えによる影響で段差や階段などが障害となっていることが推測される．原文を見ると，「道路のちょっとした段差が多い(女性・90 歳代)」のように用いられている．また，「タクシー」が特徴的な語となっている．原文を見ると，「足が悪く交通手段はタクシーですので，タクシーの安い運賃で乗れるとよいと思います(女性・80 歳代)」のように用いられており，交通手段が自動車や公共交通機関からタクシーに移行した人が多いと推測できる．

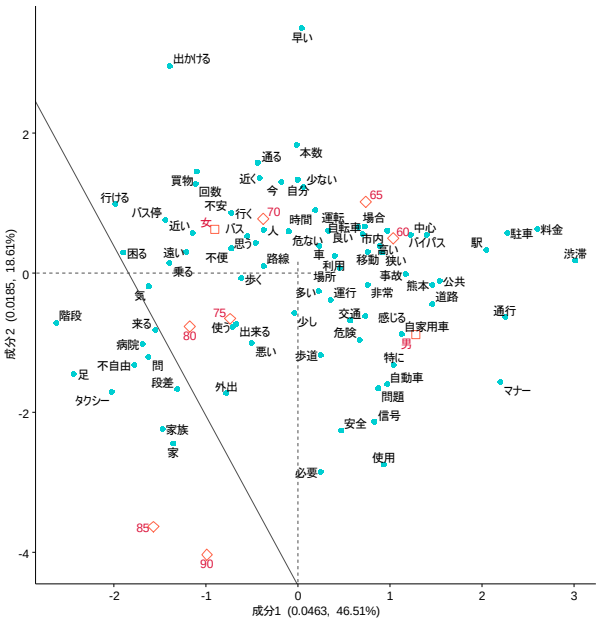

図-5　対応分析(性別×年齢)

(2) 代理回答確率

図-6 は，代理回答される確率を 0%(単身世帯)，0-20%，20-40%，40-60%，60-80%，80-100%の 6 つに分割して対応分析を行ったものである．

第二象限には 0%，0-20%，原点から縦軸負方向の縦軸近辺には 20-40%，原点から横軸正方向には 40-60%，第三象限には 60-80%，80-100%の回答者に特徴的な語が位置している．0%，0-20%の回答者の回答としては「タクシー」「階段」などの語が特徴的であった．ここで図-5 の結果から，「タクシー」「階段」は 75 歳以上の後期高齢者にも同様に特徴的な語であったことが分かる．図-2 を見ると代理回答される確率が 0%における 75 歳以上の後期高齢者の割合が大きいことが分かる．代理回答確率は式(9)を用いて求めているが，これには表-1 の推定結果を利用している．表-1 の推定結果では年齢を説明変数として用いている，したがって図-6 の結果にも年齢による影響が大きく出たことが考えられる．また，「歩く」という語も特徴的であった．原文を見ると，「歩道が歩きにくい，せまい，ななめに傾斜がひどく，歩きにくい(女性・70 歳代)」のように用いられており，自分自身が体感したことに関する回答が特徴的であった．20-40%の回答者の回答としては「自動車」「自家用車」，40-60%の回答者の回答としては「駐車」「渋滞」などの自動車に関する語が特徴的であった．図-5 の結果から，「自動車」「自家用車」は男性に特徴的な語，「駐車」「渋滞」は 60-74 歳の現役と前期高齢者に特徴的な語であることが分かっている．つまり代理回答される確率が 20-40%の回答者は男性，40-60%の回答者は 60-74 歳の回答者と同様の特徴を持つと推測できる．60-80%，80-100%の回答者の回答としては「家族」が特徴的となっている．原文を見ると，「体調が悪く，家族の運転でしか外出できない．もし体調が良くても，交通の便が悪く外出しにくい(女性・80 歳代)」のように用いられており，生活の際に家族の介護が必要である人が多いと推測できる．また「運行」も特徴的となっている．原文を見ると，「コミュニティバスの運行回数がもう少しふえたら良い (男性・70 歳代)」のように用いられており，バスなどの公共交通機関の運行に関する否定的な回答が多かった．

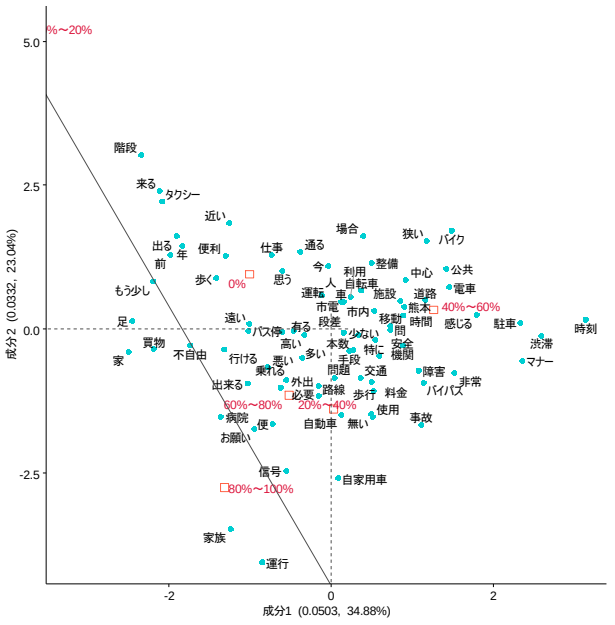

図-6　対応分析(代理回答確率)

5. おわりに

本研究では 2012 年熊本 PT 調査の付帯調査として実施された「60 歳以上の方の外出に関する意識調査」における「外出時の問題点・不満点」に関する自由回答データを対象に，細谷ら [1] により開発された PT 調査の回答有無を集団意思決定として表現した離散選択モデルを利用し，自由回答における代理回答バイアスの分析を行った．

本研究の成果を以下に示す．

(1) 性別や年齢等の属性と代理回答される確率を分析した．その中でも文字数においては，記入された文字数が多くなるにつれて代理回答されない人の割合が増えていることから，代理回答されない人は代理回答される人と比較すると自由回答に長文で回答する傾向にあるということを明らかにできた．

(2) 個人属性と自由回答で記入された語の対応分析を行った．代理回答されない回答者の回答に「歩く」が特徴的な語であったように，自由回答に自分で記入している人は「自分自身が体感した不満点」に関する回答が特徴的であった．それに比べて代理回答される確率が 60-80%，80-100%の回答者の回答に「家族」が特徴的な語であったように，代理回答される確率が上がると「他人を頼る・頼らなければならない」といった意味合いを含む回答が特徴的であることが分かった．

今後の課題を以下に示す．

(1) 2012 年熊本 PT 調査の付帯調査では，本研究で対象にした「60 歳以上の方の外出に関する意識調査」だけではなく「都市交通に関する意識調査」「住まいに関する意識調査」という付帯調査も同時に実施されているため，前述した 2 つの付帯調査においても同様の分析を行い，代理回答バイアスの実態の一部を明らかにしていく．

(2) 本研究では，対応分析を用いて回答内容と属性の関

係を分析した．この点に関しては，コーディング等を用いるなどの改良を今後の検討課題としていく．

謝辞：本研究は，JSPS 科研費 JP18H01561 の助成を受けた成果の一部です．

参考文献

1) 細谷謙太, 川野倫輝, 渡邉萌, 佐藤嘉洋, 円山琢也: 集団意思決定を考慮した世帯単位の交通調査回答行動分析, 第 57 回土木計画学研究発表会, 2018.
2) 林英夫：郵送調査法の再評価と今後の課題，行動計量学,第 37 巻,第 2 号, pp.127-145, 2010.
3) 荒牧 央: 世論調査の手法に関する現状と問題点, マス・コミュニケーション研究, 77, pp. 59-75, 2010.
4) 調査方式比較プロジェクト: 世論調査における調査方式の比較研究　　個人面接法，配付回収法，郵送法の 2008 年比較実験調査から, NHK 放送文化研究所年報, 54, pp. 105-175, 2010.
5) Seebauer, S., Fleiß,J., and Schweighart.,M.: A household is not a person: Consistency of pro-environmental behavior in adult couples and the accuracy of proxy-reports, *Environment and Behavior*, Vol. 49(6),pp. 603-637, 2017.
6) Badoe, D. & Steuart, G.:Impact of interviewing by proxy in travel survey conducted by telephone, *Journal of Advanced Transportation*, Vol.36, No.1, pp.43-62, 2002.
7) Wargelin,L. and Kostyniuk,L.: Proxy respondents in household travel surveys, in Stopher, P. and Stecher,C. (ed.) *Travel Survey Methods*, pp.201 - 212, 2006.
8) Richardson, A. J.: Proxy responses in self-completion travel diary Surveys, *Transportation Research Record*, No. 1972, pp. 1–8, 2006.
9) Verreault, H. and Morency, C.: What about proxy respondent bias over time?, Montreal, Quebec: CIRRELT: Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation, 2015.
10) 張峻屹，桑野将司，藤原章正：集団離散選択モデルの比較分析~世帯の車種選択を例に~，土木計画学研究・論文集, No.23, no.2, pp. 463-472, 2006.
11) 張峻屹，A.Borgers，H.Timmermans：集団効用関数に基づく世帯時間配分モデルの開発及び実証的分析, 土木計画学研究・論文集,Vol.19, No3, pp. 391-398. 2002.
12) 佐藤嘉洋, 円山琢也: 集団意思決定モデルを用いた益城町仮設住宅居住者の郵送調査回答行動分析, 第 57 回土木計画学研究発表会,2018.
13) 川野倫輝,佐藤嘉洋,円山琢也：トピックモデルと離散連続モデルを用いた自由記述の量的分析法,土木学会論文集 D3(土木計画学),Vol.72, No. 5, 掲載予定, 2018.

INVESTIGATING BIAS BY PROXY RESPONSE IN OPEN-ENDED SURVEY DATA

Kazuki UEHARA, Tomoki KAWANO and Takuya MARUYAMA