交通日誌データにおける無回答バイアスの修正方法†

A Correcting Method of Non-response Biases in Travel Diary Data

藤原章正††・杉恵頼寧†††・原田慎也††††
By Akimasa FUJIWARA, Yoriyasu SUGIE and Shinya HARADA

1. はじめに

交通容量の拡大を目指した時代から交通需要を管理する時代へと変遷する中で，交通主体の意思決定プロセスをより直接的に扱う交通需要予測手法として非集計アプローチが急速に普及してきた．それに伴って交通行動調査も，1960年代の都市圏規模の“恐竜時代”から90年代の施設規模の“細菌時代”へとその適用範囲が広がってきた[1]．調査デバイスに関してもアンケート調査票一辺倒であった60年代と比べ，最近では電話によるインタビューが一般化し，コンピュータネットワークを用いた応答型調査の実施もアンケート調査に迫る勢いである．対象とする交通政策の時間スケール（長期，中期，短期），空間スケール（ネットワーク全体，幹線網，地区），対象者（受益者，被害者，非利用者）に応じて，適切な交通行動調査手法を選別し適用する時代になった．

このような国際的な趨勢の中でわが国ではアンケート式の家庭訪問調査を実施する場合が少なくない．これは都市の総合交通体系の計画策定を目指したPT調査に代表されるように，共通の調査項目に関する情報を大サンプルから得ることに焦点が置かれることが多いためである．しかし調査費用は膨大であり，第3回東京都市圏PT調査に費やした費用は約7億5千万円（1票当たり1,200～1,500円）と報告されている[2]．サンプルサイズを増すことによるデータの精度と単位費用のバランスをとるような調査を設計することが重要となるが，調査方法の工夫だけでその解を見つけることは困難である．

そこで本稿ではアンケート式の交通日誌調査から得られたデータの効率的な利用方法を見出すために，データに存在する無回答バイアスの問題についてレビューし，その対応方法について検討する．ここで無回答バイアスとは回答が欠損することよって観測される交通行動や行動と要因との因果関係が歪むことをいう．

† キーワード：調査論，交通行動分析
†† 正員，工博，広島大学大学院国際協力研究科
（東広島市鏡山1-5-1, Phone&Fax: 0824-24-6921）
††† 正員，工博，広島大学大学院国際協力研究科
（東広島市鏡山1-5-1, Phone&Fax: 0824-24-6919）
†††† 学生員，広島大学大学院国際協力研究科
（東広島市鏡山1-5-1, Phone&Fax: 0824-24-6921）

表1 KONTIVにおけるアイテム無回答

調査項目	欠損率(%)	調査項目	欠損率(%)
＜個人属性＞		＜交通属性＞	
性別	3.00	目的地	15.0
年齢	4.50	目的	10.0
結婚	3.50	手段	54.7
学歴	9.40	所要時間	20.3
職業	9.50		
運転免許	9.50		

表2 広島都市圏PT調査におけるユニット無回答

市・町	人口	回収全数	有効標本数	欠損率(%)
広島市	1,042,308	83,342	72,822	12.62
呉市	225,040	18,211	16,053	11.85
大竹市	32,576	2,543	2,283	10.22
廿日市市	55,080	4,328	3,718	14.09
府中町	49,440	3,920	3,465	11.61
海田町	30,510	2,447	2,167	11.44
熊野町	26,089	2,109	1,911	9.39
坂町	13,401	1,034	951	8.03
大野町	24,405	1,938	1,751	9.65
合計	1,498,849	119,872	105,121	12.31

2. 交通行動調査データにおける無回答問題

PT調査などのアンケート調査では無回答の問題は不可避である．無回答は次の2つに分類される．

- アイテム無回答
  個人としては多くの質問項目に答えているものの，一部の質問項目について欠損がある場合
- ユニット無回答
  個人が白票で返却したり，回答を拒否したりして当該個人の回答に関する情報が一切入手できない場合

前者の例としてドイツの大規模交通行動調査であるKONTIVの報告結果を表1に示す[3]．個人属性に関する項目について欠損率は10%未満と低いものの，交通属性に関する項目ではデータの欠損率が高い．特に交通手段に関しては半数を超える回答者が無回答になっている．1日の行動を逐一思い出すことの困難さや代替の交通サービスの情報欠如などが主な原因とされている．

後者の例として広島都市圏で1987年に実施されたPT

調査におけるユニット無回答の報告結果を表2に示す[4]．全体として約12%のユニット無回答の存在が報告されており，最大で4.6%の地域間格差がみられる．ユニット無回答が特定の社会階層や地域に偏って発生する可能性があると言える．

以上の無回答データは通常コーディング段階で，例えば世帯属性等の無回答については住民台帳等の外部資料で補正したり，トリップ関連項目の無回答については電話で再調査したり，回答の前後関係から判断できる箇所の補正を行ったりする．それでもなお残った無回答については次のような対応がなされる．

(a)無回答の情報を全て無視する方法

(b)回答データに相応の重み付けをする方法

(a)は調査効率について考慮しない対応であり，通常バイアスも大きい．(b)の対応方法としては，無回答や無効票の内容は有効回収票と同質であると仮定して，適切な拡大率により処理するものである．この方法は無回答がランダムに発生する場合それほど問題は生じないが，調査への関心度や抵抗など非観測要因に起因して無回答が生じる場合，回答と無回答の間に偏りすなわちバイアスが生じ，結果が歪められる危険性がある．

また，都市の幹線交通をとらえることを主目的とする大規模調査では無回答の問題は許容誤差として無視し得ることもあるが，交通行動分析を前提とした小規模調査では重大な問題になることがある．例えば，単身世帯，共稼ぎ世帯などで有効回答率が著しく低い場合，自動車保有や休日の買物など，世帯タイプと密接に関連する交通行動を過小に見積もってしまうことが多い．このようなデータに基づいて需要予測をした場合は，交通計画全体が誤ったものになる．

3. 無回答バイアスの修正方法

(1) Imputation 法

無回答を防止するためには，まず調査段階において，サンプリング方法，調査の設計，調査の道具，調査の管理などに細心の注意を払うことが先決である．しかし一度ランダムな誤差や系統的な誤差が生じた場合はimputation法を適用しこれらの誤差を修正することが実用的であると考えられる．

Imputation法とはアイテム無回答を以下の方法で補完し，疑似完全データとして扱うものである[5),6)]．

(c)hot-deck imputation：無回答をサンプル中の回答データでそのまま置換する方法

(d)平均値 imputation:無回答を回答データから求められる平均値で代用する方法

(e)回帰 imputation:無回答アイテムを回答されている他のアイテム値で回帰推計する方法

(2) EM アルゴリズム法

(1)の imputation 法では個々のアイテム無回答を補完した後に完全データと同様の手順で分析を進めるが，EMアルゴリズム法は個々のデータを推定するのではなく，推定母数に含まれる無回答バイアスを修正し，正しい十分統計量を推定するものである．特徴は他の方法に比べて汎用性が高い点にある[7)]．

EM アルゴリズムは Dempster らによって開発された手法であり[8)]，一般的手順は以下のとおりである[9)]．

無回答のない完全データ $\mathbf{x}$ の確率変数 $\mathbf{X}$ の確率密度関数を $g_c(\mathbf{x};\Psi)$ とし，無回答を含むデータ $\mathbf{y}$ の確率変数 $\mathbf{Y}$ の確率密度関数を $g(\mathbf{y};\Psi)$ とする．Ψはパラメータ空間Ωの中の未知パラメータベクトル$\Psi=(\varphi_1,...,\varphi_d)^T$であるとすると，パラメータ$\Psi$の対数尤度関数は次式で表される．

$$\log L_c(\Psi)=\log g_c(\mathbf{x};\Psi) \tag{1}$$

$\mathbf{x}$ と $\mathbf{y}$ は各々2 つのサンプル空間 S_x と S_y から得られるとし，空間 S_x と S_y は多対1の関係にあるとすると，

$$g(\mathbf{y};\Psi)=\int_{S_x(y)} g_c(\mathbf{x};\Psi)dx \tag{2}$$

ただし，$S_x(y)$ は S_y の中の $\mathbf{x}$ に対する部分集合．

いま，Ψの初期値を$\Psi^{(0)}$とおき，尤度関数の期待値を計算する．

$$Q(\Psi;\Psi^{(0)})=E_{\Psi^{(0)}}\{\log L_c(\Psi)|\,\mathbf{y}\} \tag{3}$$

これをE (Expectation) ステップと呼ぶ．

次に，パラメータ空間Ω内のすべてのΨの中から$Q(\Psi;\Psi^{(0)})$を最大にする$\Psi^{(1)}$を見つける．

$$Q(\Psi^{(1)};\Psi^{(0)})\geq Q(\Psi;\Psi^{(0)}) \tag{4}$$

これをM (Maximization) ステップと呼ぶ．

以下，同様に式(5)のEステップと式(6)のMステップを尤度差$L(\Psi^{(k+1)})-L(\Psi^{(k)})$が小さくなるまで反復する．

$$Q(\Psi;\Psi^{(k)})=E_{\Psi^{(k)}}\{\log L_c(\Psi)|\,\mathbf{y}\} \tag{5}$$

$$Q(\Psi^{(k+1)};\Psi^{(k)})\geq Q(\Psi;\Psi^{(k)}) \quad for\ all\ \Psi\in\Omega \tag{6}$$

具体的に，交通需要モデルの無回答バイアスを回避するためにEMアルゴリズムを適用する場合の基本的な手順をまとめると以下の通りである．

1) 前節で述べた単純な imputation 法によって欠損値の初期推定値を計算し，観測値と推定値で仮の完全データを作る (式(3))．
2) 最尤法によってモデルパラメータを推定し，それを仮の完全データに適用する (式(4))．

3) モデルのパラメータなどの期待値を用いて欠損値を再推定する（式(5)）.
4) 欠損値の再推定値を用いてモデルパラメータを再推定する（式(6)）.
5) 3)と4)を収束するまで繰り返す.

4. アイテム無回答の修正

交通日誌データを仮想的に作成し分析を行う．仮想データを用いるのは各アイテム間の相関や欠損率などの条件を変化させ，EMアルゴリズム法の有効性を調べるためである.

実際の交通日誌データでは変数（アイテム）間に多様な相関関係が発生していることが予想される．そして欠損データの修正効果はこの相関関係に依存することが予想される．ここでは，1)変数間の相関の変動が大きい場合，すなわち一部の特定の変数間の相関は高いものの，他の変数間の相関が低い場合，2) 変数間の相関の変動が小さい場合，すなわち特に強い相関をもつ変数ペアは無いが，どの変数間もある程度の大きさの相関を有する場合を想定して分析を行う.

(1)変数間の相関の変動が大きい場合

① 分析データの作成

9アイテム1000ユニットの仮想データを作成する.9アイテムのうち1つは鉄道と自動車の選択結果を表す2値データであり,残り8アイテムは交通機関選択を説明する要因(費用,乗車時間,アクセス時間,待ち時間)である.

まずN(40,10)の正規乱数を1000人分発生させ,乗車時間(鉄道)t_Rとする.次に,この乗車時間と相関を持たせるために費用(鉄道)を式(7)により発生させる.

$$c_R = 10t_R + 150 + \varepsilon_{R1} \tag{7}$$

$$\varepsilon_{R1} \approx \mathrm{N}(0,10) \tag{8}$$

同様の方法でアクセス時間(鉄道)a_R,待ち時間(鉄道)w_R,乗車時間(自動車)t_C,費用(自動車)c_Cを発生させる.アクセス時間(自動車)a_C,待ち時間(自動車)w_Cはともに0である.

これらの8変数を用いて鉄道,自動車の各々の効用関数を式(9)～(10)に示す2項選択ロジットモデルの線形効用関数として設定し,機関選択結果Zを得た.

$$Z = \begin{cases} 1 : u^* > 0 \\ 0 : u^* \le 0 \end{cases} \tag{9}$$

$$\begin{aligned} u^* = &-0.005(c_R - c_C) - 0.05(t_R - t_C) \\ &-0.1(a_R - a_C) - 0.07(w_R - w_C) + (\Gamma_R - \Gamma_C) \end{aligned} \tag{10}$$

$$\left.\begin{matrix} \Gamma_R \\ \Gamma_C \end{matrix}\right\} \approx \text{擬似ガンベル分布} \tag{11}$$

なお，効用関数u^*におけるパラメータは,過去に行ったRPデータを用いた研究結果を参考に決定した.

表3　変数間の相関の変動が大きい場合の仮想データの作成方法

	鉄道の乗車時間t_R(分)	鉄道の費用c_R(円)	鉄道のｱｸｾｽ時間a_R(分)	鉄道の待ち時間w_R(分)
乱数の発生方法	$\mathrm{N}(40,10)$の正規乱数	$c_R = 10t_R + 150 + \varepsilon_{R1}$ $\varepsilon_{R1} \approx \mathrm{N}(0,10)$	$\mathrm{N}(8,2)$の正規乱数	$w_R = 0.2a_R + 1.5 + \varepsilon_{R2}$ $\varepsilon_{R2} \approx \mathrm{N}(0,1)$
平均値	40.25	552.30	8.01	3.08
標準偏差	10.21	102.11	2.05	1.07
	自動車の乗車時間t_C(分)	自動車の費用c_C(円)	機関選択結果Z(鉄道:0,自動車:1)	
乱数の発生方法	$\mathrm{N}(60,10)$の正規乱数	$c_C = 5t_R + 120 + \varepsilon_C$ $\varepsilon_C \approx \mathrm{N}(0,20)$	$Z = \begin{cases} 1 : u^* > 0 \\ 0 : u^* \le 0 \end{cases}$ $u^* = -0.005(c_R - c_C) - 0.05(t_R - t_C)$ $-0.1(a_R - a_C) - 0.07(w_R - w_C) + (\Gamma_R - \Gamma_C)$	
平均値	60.23	420.83	0.68	
標準偏差	9.96	54.75	0.47	

表4　変数間の相関の変動が大きい場合の仮想データにおける各変数間の相関係数

	鉄道の乗車時間	鉄道の費用	自動車の所要時間	自動車の費用
鉄道の乗車時間	1.000			
鉄道の費用	0.995	1.000		
自動車の乗車時間	-0.036	-0.037	1.000	
自動車の費用	-0.051	-0.052	0.925	1.000

表 3 にこのようにして作成した各変数の乱数の発生方法と発生結果（平均値，標準偏差）について総括した．この仮想データの作成方法を用いると，各誤差分布のパラメータを変化させることにより変数間の相関関係を変動させることが可能となる．表4 より，ここでは鉄道の乗車時間と費用，自動車の乗車時間と費用の 2 変数間に強い相関が表れ，それ以外は弱い相関が表れていることが明らかである．

② 分析方法

このデータを完全データと考え，選択結果を除く 8 つのアイテムの中から一部を欠損させた場合を不完全データとする．欠損方法はランダムに行うのではなく，例えば数値の大きいものから順に欠損させ，データ分布に偏りが生じるようにする．このような欠損の仕方を変化させて数種類の欠損データを作成し，一様でない無回答パターンの下でのEMアルゴリズム法の適用効果を，平均値，標準偏差の改善率，モデルパラメータを指標として測定する．

このようなデータを用いることにより，データの欠損が無視できない程度に偏って生じる場合のEMアルゴリズムの有効性を検討することができる．

ここで改善率 κ は以下の式で算出した．

$$\kappa = [1-(\varphi-\hat{\varphi})/(\varphi-\tilde{\varphi})]\times 100 \qquad (12)$$

ここで

φ ：完全データの平均値（真値）

$\tilde{\varphi}$ ：修正前データの平均値

$\hat{\varphi}$ ：修正後データの平均値

③ 変数の平均値および標準偏差に含まれるバイアスの修正結果

図 1 に 1 アイテム（鉄道の費用）のみ欠損させた場合において，修正前のバイアスを含む平均値と，EM アルゴリズム法により修正した後の平均値を，欠損率を 20%～80%の範囲で 4 段階に変えながら比較した結果を示す．欠損率が高くなればなるほど欠損データの平均値は下がるが，修正後のデータの平均値は完全データの平均値に近いことがわかる．特に欠損率が 60%までであれば，EM アルゴリズム法によりアイテムの平均値をほぼ正確に修正できることがわかる．

次に複数のアイテムが同時に欠損した場合について検討する．図 2 は 1～3 個のアイテムが同時に 20%ずつ欠損した場合の欠損アイテム数とEMアルゴリズム法による平均値の改善率との関係を示したものである． なお，2 アイテム（3 アイテム）欠損の場合の改善率は，当該アイテムと他の各アイテムとの 5 通り（計 10 通り）の組合せにおいて得られた改善率の平均値である．

1 アイテム欠損の場合に比べて 3 アイテム欠損の場合は，改善率が約 72～84％まで低下する．図 2 の中でアクセス時間（鉄道）と待ち時間（鉄道）の改善率が低いの

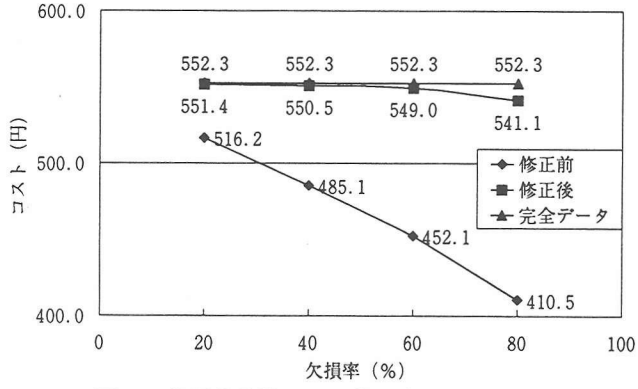

図 1　費用(鉄道)の平均値と欠損率との関係

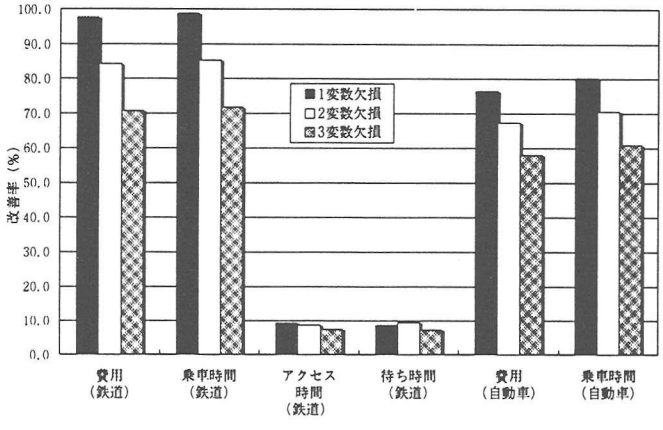

図 2　欠損アイテムの数と平均値の改善率との関係

表 5　標準偏差の変数別平均改善率
（欠損率 20%で 2 変数欠損した場合）

変数	改善率
費用(鉄道)	77.0%
乗車時間(鉄道)	79.0%
費用(自動車)	51.7%
乗車時間(自動車)	57.2%

表 6　EM アルゴリズム法による非集計ロジットモデルの推定パラメータの修正

	完全データ	20%欠損	EM 修正値	改善率(%)
費用	-0.059	-0.001	-0.058	99.0
乗車時間	-0.554	-0.081	-0.530	94.9
ｱｸｾｽ時間	-1.319	-0.260	-1.228	91.4
待ち時間	-0.187	-0.255	-0.225	45.0

は，他のアイテム間に比べてアクセス時間（鉄道）と待ち時間（鉄道）との相関が小さく設定されているためと考えられる． したがって変数間に高い相関がある場合には，EM アルゴリズム法によりアイテム無回答に伴うバイアスが修正されることが期待できる．

同様に標準偏差についてEMアルゴリズム法による修正

効果を検討した.表5に欠損率20%で2変数が欠損した場合を例として，鉄道と自動車の費用と乗車時間の標準偏差の平均改善率を示す.

平均値の場合の改善率に比べるとやや低いものの，鉄道で75～78%，自動車で50～60%の改善効果が期待できる.

④ モデルパラメータに含まれるバイアスの修正結果

交通需要予測において無回答バイアスの重大な問題は予測モデルの推定パラメータにバイアスが生じることである．そこで，交通機関選択の需要予測で頻繁に用いられる非集計ロジットモデルを事例として，不完全データとEMアルゴリズム法による修正後のデータに適用し，推定パラメータの比較を行ってアイテム無回答がモデルパラメータに及ぼす影響について分析する.

表6は費用（鉄道）のみが20%欠損した場合の各変数（アイテム）のパラメータ推定値とその改善率を示したものである．欠損した場合，待ち時間を除くパラメータ推定値は完全データの時に求められる真値よりも絶対値が小さく過小評価となる．しかしEMアルゴリズム法の適用によりこのようなバイアスは90%以上改善されており，欠損データが存在してもそのアイテムと相関の強いアイテムが存在する場合には，この修正法のバイアス修正効果が大きいことが認められた.

(2) 変数間の相関の変動が小さい場合

① 分析データの作成

前節(1)と同様の方法で,9アイテム1000ユニットの仮想データを作成する.ただし鉄道の乗車時間と費用，自動車の乗車時間と費用の4変数について，特定の変数間に強い相関関係を持たせるのではなく，どの変数間にもある程度の大きさの相関を持たせ，相関の変動幅が小さくなるようにする.

鉄道の乗車時間と費用，自動車の費用の乱数発生方法は(1)の変数間の相関の変動が大きい場合と同様であるが，自動車の乗車時間の乱数値は鉄道の乗車時間と相関を持たせるため，式(13)により作成する.

$$t_C = 1.5t_R + 10 + \varepsilon_{C1} \tag{13}$$

$$\varepsilon_{C1} \approx \mathrm{N}(0,100) \tag{14}$$

乱数の発生方法と発生結果（平均値と標準偏差），各変数間の相関マトリクスをそれぞれ表7,表8に示す.変数間の相関は0.30～0.69の範囲にあり表4と比べてその変動が小さいことが確認できる.

②変数の平均値に含まれるバイアスの修正結果

図3に1アイテム(鉄道の費用)のみ欠損させた場合において,変数間の相関の変動が大きい場合と小さい場合の改善率を比較した結果を示す．変数間の相関の変動が小さい場合は，変動が大きい場合と比べて欠損データを含む鉄道の費用が他の変数と強い相関を持たないため，改善率が低いことがわかる.

次に2アイテム(鉄道の所要時間と鉄道の費用)が同時に欠損した場合について検討する.図4はこの2アイテムが同時に欠損した場合における,アイテム間の相関の変動が大きい場合と小さい場合の各アイテムの改善率を比

表7　変数間の相関の変動が小さい場合の仮想データの作成方法

	鉄道の乗車時間t_R(分)	鉄道の費用c_R(円)	鉄道のｱｸｾｽ時間a_R(分)	鉄道の待ち時間w_R(分)
乱数の発生方法	$\mathrm{N}(40,10)$の正規乱数	$c_R = 10t_R + 150 + \varepsilon_{R1}$ $\varepsilon_{R1} \approx \mathrm{N}(0,100)$	$\mathrm{N}(8,2)$の正規乱数	$w_R = 0.5a_R + 1.5 + \varepsilon_{R2}$ $\varepsilon_{R2} \approx \mathrm{N}(0,3)$
平均値	40.25	550.45	8.03	4.06
標準偏差	10.21	141.32	2.06	3.56
	自動車の乗車時間t_C(分)	自動車の費用c_C(円)	機関選択結果Z(鉄道:0,自動車:1)	
乱数の発生方法	$t_C = 1.5t_R + 10 + \varepsilon_{C1}$ $\varepsilon_{C1} \approx \mathrm{N}(0,15)$	$c_C = 4t_R + 120 + \varepsilon_{C2}$ $\varepsilon_{C2} \approx \mathrm{N}(0,100)$	$Z = \begin{cases} 1 : u^* > 0 \\ 0 : u^* \le 0 \end{cases}$ $u^* = -0.005(c_R - c_C) - 0.05(t_R - t_C)$ $-0.1(a_R - a_C) - 0.07(w_R - w_C) + (\Gamma_R - \Gamma_C)$	
平均値	69.57	396.10	0.61	
標準偏差	21.20	132.21	0.49	

表8　変数間の相関の変動が小さい場合の仮想データにおける各変数間の相関係数

	鉄道の乗車時間	鉄道の費用	自動車の所要時間	自動車の費用
鉄道の乗車時間	1.000			
鉄道の費用	0.690	1.000		
自動車の乗車時間	0.695	0.483	1.000	
自動車の費用	0.464	0.308	0.661	1.000

較した結果を示す．相関の変動が大きい場合，相関が高い所要時間(鉄道)と費用(鉄道)が同時に欠損すると改善率は10%以下まで極端に低下する．一方，相関の変動が小さい場合，大きい場合に比べてその改善率の低下の程度は激しくなく，欠損率20%の時，1アイテム欠損と同程度の改善率（約40%）を保つことができる．

以上の結果より，各変数の分布パラメータに無視できない程度の無回答バイアスが存在する場合，EMアルゴリズム法の適用による修正が可能であるが，その効果はデータ欠損のし方や変数（アイテム）間の相関関係に依存して変動することが明らかになった．

5. ユニット無回答の修正

ユニット無回答の問題は他のデータソースを活用することによって対処することが考えられる．例えば，交通行動調査の被験者の住所，年齢，自動車保有台数などがサンプリングの段階で国勢調査データ等から別途得られている場合には，これらの個人情報を基にアイテム無回答と同様の方法でユニット無回答バイアスを修正することが理論上可能である．そこで，アイテム無回答の分析で用いた仮想データと同じ変数に性別，公共交通機関までの距離，自動車運転免許の保有状態という3つの属性変数を加えた，新たな仮想データを作り分析を進めることとする．

図5は，ユニットの欠損率と平均値の改善率との関係を，各アイテム別にグラフで表したものである．改善率は12～47%程度であり図2のアイテム無回答の場合に比べてEMアルゴリズム法による改善効果は低い．その中でアクセス時間(鉄道)に関してはやや改善効果が認められる．これは，バイアス修正のために採用した外部情報の中に「公共交通機関までの距離」というアクセス時間と相関の高いアイテムが含まれていたためであろう．換言するとユニット無回答のバイアスの修正は使用する外部情報に大きく依存することになる．

本分析で想定したユニット無回答とアイテム無回答の大きな違いは，モデルの目的変量である交通機関選択データが欠損するか否かである．すなわちユニット無回答のように目的変量が欠損する場合においては，アイテム無回答のような修正効果は期待できないことが予想される．

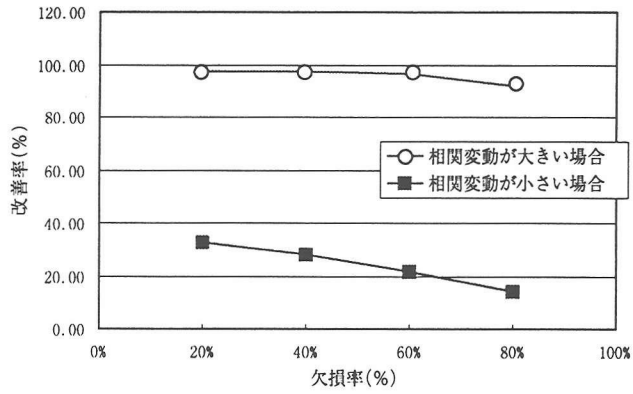

図3　変数間の相関の変動が大きい場合と小さい場合の改善率の比較（1アイテム欠損の場合）

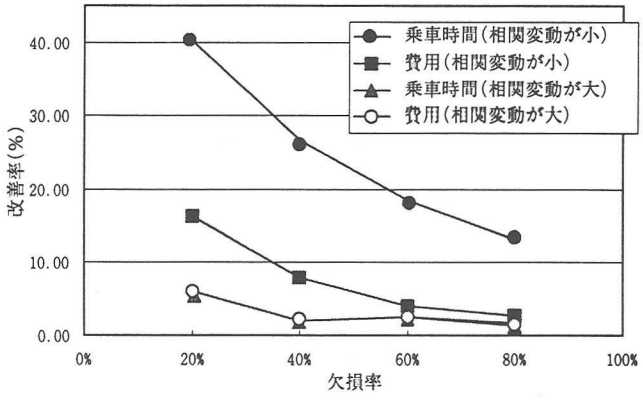

図4　変数間の相関の変動が大きい場合と小さい場合の改善率の比較（2アイテム欠損の場合）

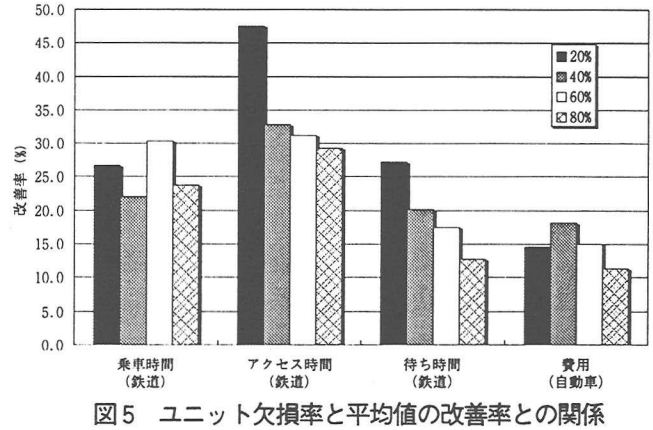

図5　ユニット欠損率と平均値の改善率との関係

6. 実データを用いたEMアルゴリズム法と平均値Imputation法の比較

(1) 変数の修正結果の比較

4.,5.の仮想データを用いた分析から，アイテム無回答については，無回答を含む各変数の分布パラメータ（平均値および標準偏差）とその変数組を用いた非集計モデルの各パラメータに含まれるバイアスを修正に対するEMアルゴリズム法の有効性について検討した．ここでは実データを用いて従来の平均値Imputation法とEMアルゴリズム法を適用した場合の違いについて調べる．

使用データは広島都市圏で1997年11月に実施されたミニPT調査から得た通勤トリップデータである．なお以下の分析に用いるデータの欠損率は全個人の全アイテムを通じて29.3%であった．

まず無回答を含む個人の交通日誌データを2種類の方法で修正して各々ゾーン別変数値を作成する．次に，集計型ロジットモデルを推定しモデル推定結果の違いを明らかにする．集計モデルの分析データは，個人の交通日誌データをゾーン単位で平均した値であるので，平均値Imputation法およびEMアルゴリズム法を用いてゾーンjアイテムdの修正後の分析データ$\hat{w}_{jd}$を各々式(15)および(16)を用いて算出する．

$$\hat{w}_{jd}=\bar{x}_{jd}=\frac{\sum_{n=1}^{N_{jd}} x_{jdn}}{N_{jd}} \tag{15}$$

$$\hat{w}_{jd}=\hat{\varphi}_{jd} \tag{16}$$

ただし，N_{jd}は回答データの数

式(15)から明らかなように平均値Imputation法で得られた修正データ$\hat{w}_{jd}$は回答データのゾーン平均値であり，2. の(a)で述べた無回答の情報を無視する方法で得たゾーン平均値と一致する．

修正データを用いて自動車と公共交通機関（JR·市電·新交通システム·バス）の間の機関別分担率を予測する集計型2項ロジットモデルを推定する．説明変数は，自動車の有無（1:保有），自由に利用できる車の台数，最寄の駅·バス停から勤務先までのエグレス距離，所要時間差（＝自動車－公共交通）の計4変数とする．

分析対象とする250Dペアの各変数の平均値と標準偏差について，2つの方法による修正値の違いを表9に示す．平均値については両修正法で大きな違いは見られないが，標準偏差については特に所要時間差で非常に大きな差が認められた．EMアルゴリズムによる修正値がより真値に近いことを考えると，この結果は平均値Imputationによる修正を行うと一部の変数のゾーン間のばらつきが失われてしまうことを意味している．

(2) 集計モデルの推定結果の比較

集計ロジットモデルの推定結果を表10に示す．モデル推定はBerkson法[10]に習い分担率比の対数変換を行い線形回帰式として推定した（式(18)）．誤差項は正規分布に従うものとする．説明変数は表9の4変数と着地都心ダミー，定数項である．

$$\ln\left(P_{car}/P_{transit}\right)=\gamma_o\, c.own+\gamma_u\, c.use+\gamma_e\, egress+\gamma_t\, d.time+\gamma_d\, downtown+const. \tag{18}$$

ここでP：分担率，*c.own*：自動車保有ダミー，*c.use*：自由利用車台数，*egress*：エグレス距離，*d.time*：所要時間差，*downtown*：着地都心ダミー，γ：未知パラメータ．

モデル推定結果を表10に示す．モデル適合度を示す重

表9　修正データの平均値と標準偏差の比較

変数	EM alg.[a]	平均値Imp.[b]	a/b-1.0
自動車保有ダミー	0.909	0.912	-0.3%
	0.063	0.062	1.6%
自由利用車台数(台)	1.080	1.088	-0.7%
	0.059	0.059	0.0%
エグレス距離(m)	546.9	534.7	2.3%
	404.7	391.7	3.3%
所要時間差(分)	-6.193	-6.062	2.2%
	20.91	8.59	143.4%

上段：平均値，下段：標準偏差

表10　集計モデルの推定結果の比較

	EM alg.	平均値Imp.
自動車保有ダミー	5.008 *	4.996 *
自由利用車台数	1.144	1.750
エグレス距離	-0.048	0.105
所要時間差	-0.010	0.010
着地都心ダミー	0.438	0.572 *
定数項	-11.836 *	-12.548 *
重相関係数	0.612	0.563
サンプル数	25	25

*：5%有意

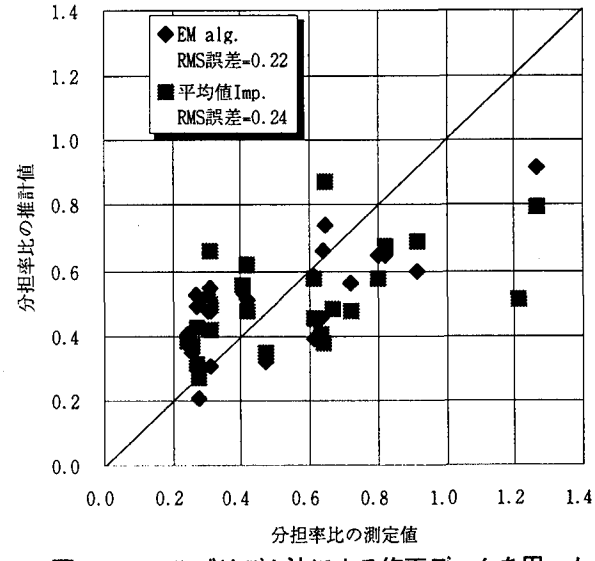

図6　EMアルゴリズム法による修正データを用いた交通機関分担モデルの予測精度

相関係数はEMアルゴリズム法による方がやや高い値を示している．説明変数のパラメータに着目すると，エグレス時間と所要時間差の符号が両者で異なる結果となった．両者ともにEMアルゴリズム法のモデルの符号が妥当である．また，図6に示すように，EMアルゴリズム法による修正データを用いた交通機関分担モデルの推計精度は平均値imputation法の場合よりも良好であることも明らかである．

これらの結果から判断するとEMアルゴリズム法による無回答バイアスの修正効果は，集計型交通需要予測モデルのモデルパラメータにも反映されるものと考えられ，

平均値Imputation法による修正を行った場合には予測結果により大きな誤差が生じることになると考えられる.

7. おわりに

本研究では仮想的な交通日誌データ用いて，アイテム無回答とユニット無回答に伴うバイアスの修正方法としてEMアルゴリズムに着目して，その特性と改善効果について検討した.1つの重要な結果として，データの分布パラメータや非集計モデルの推定パラメータに現れるアイテム無回答によるバイアスは無視できるものではない場合には，EMアルゴリズム法を適用することによって一定範囲内であれば修正可能であることが確認された．またこの効果は,各アイテム間の相関によって大きく変動するという特性があることが確認された.

また，集計型交通需要モデルを実データを用いて推定した結果，EM アルゴリズム法の適用と伝統的な平均値Imputation 法の適用では結果に大きな差が現れることが実証された.

一方，ユニット無回答に伴うバイアスの修正に関しては，あまり修正効果が認められなかった．特に交通需要モデルを前提とした調査データを扱う場合には，目的変量が欠損するため，アイテム無回答と同様の扱いでは不十分な場合があると考えられる．またユニット無回答バイアスを扱う際には，外部情報の利用が重要となると予想される．例えば昨今普及が著しいGIS を活用し交通サービス水準に関する客観値を利用するなど，関連する外部情報の収集法について検討することが必要である.

参考文献

1) D.Hartgen: Coming in the 1990s: The agency-friendly travel survey, Transportation, Vol.19, No.2, pp.79-95, 1992.
2) 日本交通政策研究会：道路交通統計のあり方，日交研シリーズ，A-244，1998.
3) A.Richardson, E.Ampt, A.Meyburg: Survey Methods for Transport Planning, Eucalyptus Press, p.314, 1995.
4) 広島都市圏交通計画協議会：昭和63年度広島都市圏パーソントリップ調査報告書－3現況集計編，p.6，1989.
5) R.Little and D.Rubin: Statistical Analysis with Missing Data, John Wiley and Sons, 1987.
6) D.Rubin: Multiple Imputation for Nonresponse in Surveys, John Wiley and Sons, 1987.
7) J.Polak, X.L.Han: Iterative Imputation Based Methods for Unit and Item Non-Response in Travel Diary Surveys, 8th Meeting of the IATBR, Austin, pp.21-25 1997.
8) A.Dempster, N.Laird and D.Rubin: Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society B, No.39, pp.1-38, 1977.
9) G.McLachlan and T.Krishnan: The EM Algorithm and Extensions, John Wiley and Sons, 1997.
10) J.Berkson: A Statistically Precise and Relatively Simple Method of Estimating the Bioassay with Quantal Response, Based on the Logistic Function, Journal of American Statistic Association, No.48, pp.565-599, 1953.

交通日誌データにおける無回答バイアスの修正方法

藤原章正，杉恵頼寧，原田慎也

本研究は交通日誌データに含まれる無回答に伴うバイアスを緩和するための方法について検討することを目的とする．アイテム無回答とユニット無回答を含む仮想データのバイアスを修正するため，本研究ではEMアルゴリズム法を採用する．仮想データを用いた分析の結果，変数間の相関が強い場合にアイテム無回答に伴う変数の平均値および分散と交通需要モデルの推定パラメータのバイアスの改善にはEMアルゴリズムの適用が有効であることが確認された．また実データを用いて交通需要モデルを推定し従来のImputation法との比較を行った結果,EMアルゴリズム法が有効であることが確認された.

A Correcting Method of Non-response Biases in Travel Diary Data

By Akimasa FUJIWARA, Yoriyasu SUGIE and Shinya HARADA

This study aims at examining the methods for relaxing non-response biases in travel diary data. EM algorithm is employed to correct the biases existing in the hypothetical data sets including item and unit non-responses. A result of the analysis shows that the EM algorithm is significantly capable to improve the biases in mean and variance of variables and model parameters in a case of high correlation among variables. Moreover, it is confirmed that the EM algorithm is more effective in correcting the non-response biases in the estimated parameters of travel demand models as compared with the conventional imputation by using the actually observed travel diary data.