第 55 回土木計画学研究発表会・講演集

小地域の推定精度からみたビッグデータ時代における交通日誌調査設計とモデル選択

力石　真[1]・藤原　章正[2]

[1]正会員　広島大学大学院　准教授　国際協力研究科（〒739-8529 広島県東広島市鏡山 1-5-1）
E-mail: chikaraishim@hiroshima-u.ac.jp

[2]正会員　広島大学大学院　教授　国際協力研究科（〒739-8529 広島県東広島市鏡山 1-5-1）
E-mail: afujiw@hiroshima-u.ac.jp

ビッグデータの出現により交通現象の観測費用が低下し，モデルを通じた演繹的推論の一部を観測が代替／補完する可能性がある今，施策検討プロセス全体における推定誤差／予測誤差最小化の観点からビッグデータの利用，交通日誌調査の設計，モデルの選択を行う必要性は高いものと考えられる．本研究では，仮想状況を想定した簡便なシミュレーション分析を通じて，小地域推定の観点から標本サイズ，モデル精度，ビッグデータの利用の関係性について整理する．具体的には，我が国のスタンダードといえる design-based 直接推定量，欧米における交通日誌調査のスタンダードとなりつつある model-based 直接推定量の比較，及び，ビッグデータを共変量として導入した場合の推定量の精度について分析し知見を整理した．

Key Words: *travel diary, survey design, big data, small area estimation*

1.　はじめに

交通日誌調査に限らず，多くの標本調査設計は，調査費用の制約が議論の出発点にある．標本数を増やし観測密度を高めれば精度の高い推定量を得ることは可能だが，膨大な観測費用がかかる．そこで，限られた予算の中で出来る限り精度を高めるための様々な標本調査設計手法が提案されてきた[1]．一方，ビッグデータの出現により観測費用は大きく低下しており，観測精度と費用のトレードオフに主眼を置いた従来の調査設計手法だけでは，最適な調査スキームを導くことが難しくなりつつある．

多くの場合，交通日誌調査の実施目的は将来交通需要予測にある．従って，調査設計の議論はモデル選択の議論と不可分である．具体的には，例えば**図-1** に示すように，観測精度の高い大規模な交通日誌調査を行ったとしても，モデルの精度が低ければ予測精度は極めて低いものになり得るし，反対に，小規模の調査であっても適切に観測情報を利用できれば精度の高い予測を行うことができる．しかしながら，現在我が国において広く用いられている交通日誌調査の標本サイズは，地域標本の平均／集計値の推定精度に基づき設計されるものであり（小地域推定の文脈では design-based 直接推定量と呼ばれる[2]），観測精度とモデル精度の双方を踏まえた調査設計とはなっていない．プローブデータや携帯電話基地局データ，携帯アプリの経路検索履歴データといった 365 日 24 時間の観測データが，モデルを通じた演繹的推論の一部を代替／補完する可能性がある今，従来の調査設計の枠組みを拡大し，ビッグデータの利用可能性を視野に入れた上で，施策検討プロセス全体における推定誤差／予測誤差最小化の視点から調査設計，モデル選択の議論を展開する必要性は高いものと考えられる．

以上の背景を踏まえ，本研究では，小地域推定の観点から標本サイズ，モデル精度，ビッグデータの利用の関係性について整理する．ここで行う分析はシミュレーションに基づく簡便な分析であるものの，以下の 4 つの推定量の精度（平均二乗誤差）を比較している点に特徴がある．

①　日本における交通日誌調査設計のスタンダードとい

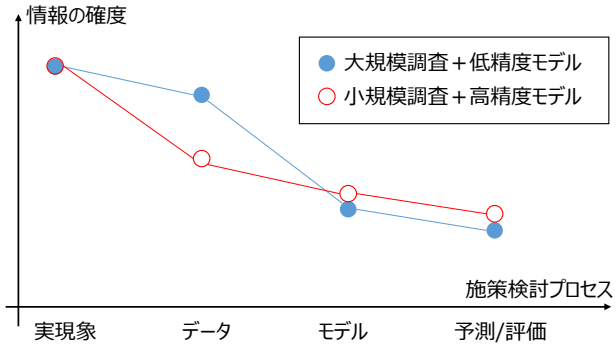

図-1　施策検討プロセスと予測精度

える（四段階推定法の利用を前提とした）design-based 直接推定量
② 欧米における需要予測手法のスタンダードとなりつつある非集計／Activity-based(AD)モデルの利用を前提とした model-based 直接推定量
③ ②の model-based 直接推定量にビッグデータを共変量を加えた model-based 直接推定量（以降，B-model-based 直接推定量と呼ぶ）
④ 小地域推定の文脈でスタンダードとなりつつある model-based 間接推定量（EBLUP (Empirical Best Linear Unbiased Prediction)推定量）

以上4つの推定量の精度比較を通じて，(1)標本サイズが比較的小さく，ビッグデータの精度が高い場合においては，ビッグデータを共変量として用いることにより推定精度は向上するものの，本来のデータ生成過程とは異なるモデル選択となる恐れが高いこと，(2) 推定量のバイアス最小化，推定量の分散最小化いずれの観点からも，多くの場合，①と②を組み合わせた推定量である EBLUP 推定量の精度が高いこと（従って，わが国のスタンダードといえる「大規模交通日誌調査＋四段階推定法」，欧米諸国のスタンダードとなりつつある「小規模交通日誌調査＋非集計/AD モデル」，いずれのやり方よりも高い推定精度を得られる可能性があること）を示す．後者の点については理論的には既に自明であるものの，交通日誌調査の設計の文脈において指摘した文献が見当たらないことから，本研究にてシミュレーション分析を通じて明示することとした．

なお，上述したように，交通日誌調査の主な目的は将来交通需要の推計にあることから，トリップの発生・集中，分布，交通手段分担に及ぶ全ての段階で精度を検証することが望ましいが，本研究では，推定量が解析的に定義可能な回帰モデルの枠組みでモデル化が可能なトリップ生成量に焦点を絞る．また，経時変化が把握可能なビッグデータの特性を踏まえると，社会構造の変化に伴う交通行動の構造変化を含む形でデータを生成し，将来時点の予測精度をもって精度検証を行うことが望ましいが，本稿では断面データを用いた精度検証に留まっている．トリップ生成量以外の行動側面，及び，時間軸の導入については今後の課題としたい．

2. 推定量の定義と精度検証方法

(1) 小地域推定からみた調査設計及びモデル選択

本研究では，地域（ゾーン）単位のトリップ生成量の推定精度を高めることを目的として，調査設計及びモデル選択を行うことを考える．標本調査の観点からは，地域ごとの特性値（本研究ではトリップ生成量）の推定精度に焦点を当てる小地域推定（Small Area Estimation）の問題として解釈できる．

小地域推定問題の観点から交通日誌調査設計の考え方を再解釈すると，各国で採用されているアプローチを比較的クリアに整理できる．具体的には，我が国で広く採用されている交通日誌調査の設計は，当該小地域内の標本データのみ用いて推定量を構成する design-based 直接推定量の精度に焦点を当て標本数を決定したものといえる[3]．また，欧米でスタンダードとなりつつある，比較的小規模なデータを用いて非集計／AD モデルを構築し，構築したモデルを通じて小地域の推定量を得る方法は，小地域推定の観点からは model-based 直接推定量として解釈できる．さらに，小地域推定の分野においては，(一般化)線形混合モデルやその時系列モデル／空間統計モデルへの拡張モデル等，推定精度を高めるための様々な方法論が提案されている（詳細は Rao and Molina[2]参照）．その代表的な推定量が EBLUP 推定量であり，後述するように，design-based 直接推定量と model-based 直接推定量の重み付け平均として定義される推定量として解釈できる．

以上を踏まえて本研究では，(1) design-based 直接推定量，(2) model-based 直接推定量, (3) EBLUP 推定量の 3 つの推定量の精度比較を行う．また，(2)については，地域レベルの集計量として利用できるビッグデータを共変量として加えた推定量についても考察する．なお，(3)の場合，地域レベルの変数を共変量として加えても推定量の精度は変わらない．

(2) 小地域推定とビッグデータ

地域ごとの特性値の推定精度向上の観点から，ビッグデータがどのような役割を果たすかについて議論した既往研究は多くはない．数少ない例外として，Marcheti et al.[4]の研究が挙げられる．Marcheti et al.は，ビッグデータの有する特徴として，自己選択バイアスの問題を指摘している．例えば ETC2.0 のデータは比較的裕福な層の行動を反映したデータといえる．また，株式会社 NTT ドコモが提供するモバイル空間統計は，ドコモ利用者の情報を母集団に拡大したものである．また，多くのビッグデータは利用者の同意のもとで得られたデータであるため，自己選択バイアスの影響を受けやすいデータといえる．

以上の特徴を有するビッグデータを小地域推定に利用する方法として，Marcheti et al.は以下の 3 つを挙げている．

1) 標本調査データから得た推定量とビッグデータから得た地域毎の特性値を比較する
2) ビッグデータから得られた地域毎の特性値を小地域推定の共変量として導入する
3) 標本調査データを用いてビッグデータが持つ自己

選択バイアスを除去する

本研究では，2)の方法を採用した場合におけるビッグデータの推定精度への影響を確認する．

(3) 推定量の定義

本研究で用いる記号は以下のとおりである．

n_d: 地域dの標本サイズ（$n = \sum_d n_d$）
i: 個人（$i = 1,2,\dots,I$）
d: 地域（$d = 1,2,\dots,D$）
Y_{id}: 地域dに居住する個人iの 1 日当たりトリップ生成量
$\bar{Y}_d$: 地域dの 1 人日当たり平均トリップ生成量
x_{jid}: j番目の個人属性（$\mathbf{x}_{\mathrm{id}} = \left(x_{1id}, x_{2id}, \dots, x_{Jid}\right)^T$）
$\bar{x}_{jd}$: 地域dにおけるj番目の個人属性の平均値（$\bar{\mathbf{x}}_{\mathrm{d}} = \left(x_{1id}, x_{2id}, \dots, x_{Jid}\right)^T$）
β_j: j番目のパラメータ（$\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_J)$）
z_{kd}: k番目の地域属性（$\mathbf{z}_{\mathrm{d}} = (z_{1i}, z_{2d}, \dots, z_{Kd})^T$）
γ_k: k番目のパラメータ（$\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_K)$）
L_d: $\bar{Y}_d$の代理指標となるビッグデータ共変量
τ: ビッグデータL_dに対応するパラメータ
η_d: 地域d固有の非観測要因を表すランダム変数

なお，推定値／予測値についてはハット記号を用いて区別する．

以下，それぞれの推定量を具体的に定義する．我が国の交通日誌調査においては非復元抽出法に基づく推定量を用いることが多いが，ここでは簡便のため復元抽出法の場合の推定量を利用する．そのため，非復元抽出の場合に比べて推定量の分散が多少過大に推計されるが，抽出率は通常数%未満であるため，ここでは無視できるものと考える．また，通常地域毎で人口規模が異なるため，地域ごとの集計量（総トリップ生成量）に着目するか，地域ごとの平均値（1 人日当たり平均トリップ生成量）に着目するかで推定量の分散は異なる．どちらの推定量に着目すべきかは文脈に依存すると考えられるが，本研究では後者の平均トリップ生成量に着目することとする．

Estimator 1: design-based 直接推定量

無作為抽出の場合，Design-based 直接推定量は以下のように定義される．

$$\hat{Y}_d = \bar{Y}_d = \frac{\sum_{i=1}^{n_d} Y_{id}}{n_d} \tag{1}$$

Estimator 2: model-based 直接推定量

以下の回帰モデルに基づく結果を推定量とする．

$$\hat{Y}_d = \hat{\boldsymbol{\beta}}\bar{\mathbf{x}}_{\mathrm{d}} + \hat{\boldsymbol{\gamma}}\mathbf{z}_{\mathrm{d}} \tag{2}$$

Estimator 3: B-model-based 直接推定量

以下の回帰モデルに基づく結果を推定量とする．

$$\hat{Y}_d = \hat{\boldsymbol{\beta}}\bar{\mathbf{x}}_{\mathrm{d}} + \hat{\boldsymbol{\gamma}}\mathbf{z}_{\mathrm{d}} + \hat{\tau}L_d \tag{3}$$

Estimator 4: EBLUP 推定量

以下の線形混合モデルに基づく結果を推定量とする．

$$\hat{Y}_d = \hat{\boldsymbol{\beta}}\bar{\mathbf{x}}_{\mathrm{d}} + \hat{\boldsymbol{\gamma}}\mathbf{z}_{\mathrm{d}} + \hat{\eta}_d \tag{4}$$

ここで$\hat{\eta}_d$の予測量は，個人レベルの誤差分散と地域レベルの誤差分散の比を$\hat{\rho}$とすると$\hat{\rho}n_d/(1+\hat{\rho}n_d) \times \left(\bar{Y}_d - \hat{\boldsymbol{\beta}}\bar{\mathbf{x}}_{\mathrm{d}} - \hat{\boldsymbol{\gamma}}\mathbf{z}_{\mathrm{d}}\right)$で与えられるため，$\hat{Y}_d$は以下のように書き換えることができる．

$$\hat{Y}_d = \frac{\hat{\rho}n_d}{1+\hat{\rho}n_d}\bar{Y}_d + \frac{1}{1+\hat{\rho}n_d}\left(\hat{\boldsymbol{\beta}}\bar{\mathbf{x}}_{\mathrm{d}} + \hat{\boldsymbol{\gamma}}\mathbf{z}_{\mathrm{d}}\right) \tag{5}$$

従って EBLUP 推定量は，Estimator 1 と Estimator 2 の重み付け平均として表現される．この式は，n_dまたは$\hat{\rho}$が小さければ，モデルを介した推定量である Estimator 2 に近い推定値を，反対に，地域毎の標本サイズが十分にあれば標本平均を推定量とする Estimator 1 に近い推定値を返すことを意味する．

(4) 平均二乗誤差の計算手順

Design-based 推定量の場合，不偏推定量であることを前提に議論が展開されることが多いため，推定量の分散にのみ焦点を当てて推定精度を議論することが多い．しかしながら，モデルを介した推定量やビッグデータの利用は，推定量にバイアスを生じさせる可能性が高い．したがって本研究では，平均二乗誤差$MSE = E\left(\hat{Y}_d - \tilde{Y}_d\right)^2$により推定精度を評価する．ここで$\tilde{Y}_d$は地域$d$の 1 人日当たり平均トリップ生成量の真値とする．広く知られているように，MSEは以下のように分散成分とバイアス成分に分解できる．

$$MSE_d = E\left(\hat{Y}_d - \tilde{Y}_d\right)^2$$

$$= \underbrace{E\left(\hat{Y}_d - E\left(\hat{Y}_d\right)\right)^2}_{\text{分散}(=\sigma_d^2)} + \underbrace{\left(E\left(\hat{Y}_d\right) - \tilde{Y}_d\right)^2}_{\text{バイアスの2 乗}(=b_d^2)} \tag{6}$$

平均二乗誤差の算出においては，以下のブートストラップ推定値を用いる．

$$\hat{\sigma}_d{}^2 = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{Y}_d - \frac{1}{B}\sum_{b=1}^{B}\hat{Y}_d\right)^2 \tag{7}$$

$$\hat{b}_d = \frac{1}{B}\sum_{b=1}^{B}\hat{Y}_d - \tilde{Y}_d \tag{8}$$

ここでBはブートストラップ反復抽出の回数であり，本研究では$B = 500$とした．本研究では，真値$\tilde{Y}_d$が既知である仮想データを作成し分析を進める．

表-1 データ生成過程

	データ生成過程（地域レベル）	データ生成過程（個人レベル）
x_{1id}：世帯自動車保有台数	$\dot{x}_{1d}$~urn[0.5,1.5]	$\dot{x}_{1d}$+rnorm[0,0.5]を生成し，0.5未満であれば$x_{1id}=0$，0.5以上1.5未満であれば$x_{1id}=1$，1.5以上2.5未満であれば$x_{1id}=2$，2.5以上であれば$x_{1id}=3$
x_{2id}：世帯収入[100万円]	$\dot{x}_{2d}$~ urn[2.0,7.0]	$x_{2id}=\dot{x}_{2d}+$ lrnorm[0,0.5]
x_{3id}：高齢者（65歳以上）ダミー	$\dot{x}_{3d}$~urn[0.1,0.4]	urn[0.0,1.0]が$\dot{x}_{3d}$未満であれば$x_{3id}=1$，そうでなければ$x_{3id}=0$
x_{4id}：子供（18歳未満）人数	$\dot{x}_{4d}$~urn[0.5,1.5]	$\dot{x}_{4d}$+rnorm[0,1.0]を生成し，0.5未満であれば$x_{4id}=0$，0.5以上1.5未満であれば$x_{4id}=1$，1.5以上2.5未満であれば$x_{4id}=2$，2.5以上3.5未満であれば$x_{4id}=3$，3.5以上4.5未満であれば$x_{4id}=4$，4.5以上であれば$x_{4id}=5$
x_{5id}：就業状態 （1: 就業，0: その他）	$\dot{x}_{5d}$~urn[0.4,0.6]	urn[0.0,1.0]が$\dot{x}_{5d}$未満であれば$x_{5id}=1$，そうでなければ$x_{5id}=0$
z_{1d}：居住地ダミー （1: CBD，0: その他）	urn[0.0,1.0]が0.9より大きければz_{1d} =1，そうでなければz_{1d} =0	—
z_{2d}：当該地域のアクセシビリティ	z_{2d}~urn[1.0,5.0]	—

※ urn$[x,y]$: 区間$[x,y]$の一様分布に従う乱数 ; rnorm$[x,y]$: 平均x，標準偏差yの正規分布に従う乱数 ; lrnorm$[x,y]$: 平均x，標準偏差yの対数正規分布に従う乱数

3. シミュレーション設定

(1) 仮想状況の設定

本研究では，以下に示す単純な仮想状況を想定してシミュレーション分析を行う．なお，以下では day-to-day の交通需要変動は存在せず，したがって 1 日の交通行動の観測情報が他の日の現象記述に利用できる状況を考える．

総人口 100 万人，地域（ゾーン）数が 100 の都市圏を考える．なお，地域毎の人口規模は，100 万人を 100 ゾーンにランダムに割り付けることで作成する．100 万人分の 1 日当たりトリップ生成量は，以下の式に基づき生成する．

$$Y_{id}=\beta_0+\beta_1x_{1id}+\beta_2x_{2id}+\beta_3x_{3id}+\beta_4x_{4id}+\beta_5x_{5id}+\gamma_1z_{1d}+\gamma_2z_{2d}+\eta_d+\varepsilon_{id} \quad (9)$$

ここで，x_{1id}は世帯自動車保有台数，x_{2id}は世帯収入[100 万円]，x_{3id}は高齢者（65 歳以上）ダミー，x_{4id}は子供（18 歳未満）人数，x_{5id}は就業状態（1: 就業，0: その他），z_{1d}は居住地ダミー（1: CBD，0: その他），z_{2d}は当該地域のアクセシビリティを想定した．x_{1id}~x_{5id}については，地域間の変動と地域内の変動を反映するため，表-1 に基づきデータを生成した．z_{1d}およびz_{2d}については地域レベルの属性であるため，地域内の変動は存在しない．なおパラメータは$(\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5,\gamma_1,\gamma_2)=(0.2,0.1,-1.0,0.2,1.0,0.3,0.1)$と設定した．$\eta_d$は平均 0 標準偏差 0.3 の正規分布，$\varepsilon_{id}$は平均 0 標準偏差 0.3 の正規分布に従い乱数を発生させ，データを生成した．生成した母集団データから，各地域の平均トリップ生成量の真値$\tilde{Y}_d$を計算する．

次に，仮想的なビッグデータを作成する．本研究では，ビッグデータは地域毎の集計量のみ利用可能であるとし，その値は以下の平均$\tilde{Y}_d$，分散σ_Lの正規分布に従って生成する．

$$L_d\sim N\left(\tilde{Y}_d,\sigma_L\right) \quad (10)$$

ここでσ_Lはビッグデータのバイアスの程度を表す標準偏差であり，$\sigma_L=0$のとき，ビッグデータは母集団の特性値を正確に反映する．この場合，調査を行わずとも今回の分析の目的である各地域の平均トリップ生成量はビッグデータにより正確に把握できる．本研究では，$\sigma_L=0.1,0.5,1.0$の 3 ケースについて考察する．

(2) 比較対象とする推定量

以上の仮想状況のもとでは，上述した 4 つの推定量は以下のように定義できる．

Estimator 1: design-based 直接推定量

$$\hat{Y}_d=\frac{\sum_{i=1}^{n_d}Y_{id}}{n_d} \quad (11)$$

Estimator 2: model-based 直接推定量

$$\hat{Y}_d=\hat{\beta}_0+\hat{\beta}_1\bar{x}_{1d}+\hat{\beta}_2\bar{x}_{2d}+\hat{\beta}_3\bar{x}_{3d}+\hat{\beta}_4\bar{x}_{4d}+\hat{\beta}_5\bar{x}_{5d}+\hat{\gamma}_1z_{1d}+\hat{\gamma}_2z_{2d} \quad (12)$$

Estimator 3: B-model-based 直接推定量

$$\hat{Y}_d=\hat{\beta}_0+\hat{\beta}_1\bar{x}_{1d}+\hat{\beta}_2\bar{x}_{2d}+\hat{\beta}_3\bar{x}_{3d}+\hat{\beta}_4\bar{x}_{4d}+\hat{\beta}_5\bar{x}_{5d}+\hat{\gamma}_1z_{1d}+\hat{\gamma}_2z_{2d}+\hat{t}L_d \quad (13)$$

Estimator 4: EBLUP 推定量

表-2　推定量のバイアス，分散，平均二乗誤差($\sigma_L = 0.1$)

標本サイズ	Estimator 1			Estimator 2			Estimator 3			Estimator 4			Estimator 5			Estimator 6			Estimator 7			Estimator 8			Estimator 9		
	bias	var	mse	bias	var	mse	bias	var	mse	bias	var	mse	bias	var	mse	bias	var	mse	bias	var	mse	bias	var	mse	bias	var	mse
100	6.47	23.1	29.6	10.8	0.91	11.7	1.96	0.66	2.62	5.50	2.62	8.12	14.1	2.06	16.1	1.14	2.38	3.52	10.3	4.20	14.5	1.12	1.35	2.47	1.04	1.24	2.29
300	3.26	20.5	23.8	10.8	0.29	11.1	1.95	0.20	2.16	2.85	2.07	4.93	14.1	0.63	14.7	1.16	0.73	1.89	6.84	3.27	10.1	1.10	0.41	1.52	1.05	0.39	1.43
500	2.54	17.4	19.9	10.8	0.17	11.0	1.97	0.12	2.09	2.09	1.73	3.82	14.1	0.37	14.5	1.15	0.43	1.58	5.24	3.17	8.40	1.10	0.24	1.34	1.05	0.23	1.28
700	1.83	15.7	17.5	10.8	0.12	10.9	1.96	0.09	2.05	1.60	1.54	3.15	14.1	0.28	14.4	1.16	0.32	1.47	4.33	3.04	7.37	1.09	0.18	1.28	1.04	0.17	1.21
1,000	1.36	13.1	14.4	10.8	0.09	10.9	1.97	0.06	2.03	1.18	1.32	2.51	14.1	0.20	14.3	1.17	0.23	1.40	3.31	2.84	6.15	1.12	0.13	1.25	1.06	0.12	1.18
3,000	0.37	6.88	7.25	10.8	0.03	10.8	1.97	0.02	1.98	0.39	0.74	1.13	14.1	0.06	14.1	1.17	0.07	1.24	1.49	1.98	3.47	1.11	0.04	1.15	1.05	0.04	1.09
5,000	0.16	4.98	5.14	10.8	0.02	10.8	1.96	0.01	1.98	0.20	0.53	0.74	14.1	0.04	14.1	1.16	0.04	1.21	0.94	1.61	2.55	1.10	0.03	1.13	1.05	0.02	1.07
7,000	0.10	3.93	4.03	10.8	0.01	10.8	1.96	0.01	1.97	0.13	0.42	0.55	14.1	0.03	14.1	1.16	0.03	1.19	0.70	1.37	2.06	1.10	0.02	1.12	1.05	0.02	1.07
10,000	0.03	2.91	2.94	10.8	0.01	10.8	1.96	0.01	1.97	0.08	0.32	0.40	14.1	0.02	14.1	1.16	0.02	1.18	0.49	1.10	1.59	1.10	0.01	1.12	1.05	0.01	1.06
30,000	0.00	1.04	1.04	10.8	0.00	10.8	1.96	0.00	1.96	0.01	0.13	0.14	14.1	0.01	14.1	1.16	0.01	1.17	0.12	0.53	0.65	1.10	0.00	1.11	1.05	0.00	1.05
50,000	0.00	0.59	0.59	10.8	0.00	10.8	1.96	0.00	1.96	0.01	0.08	0.09	14.1	0.00	14.1	1.16	0.00	1.16	0.05	0.37	0.42	1.10	0.00	1.11	1.05	0.00	1.05
70,000	0.00	0.40	0.40	10.8	0.00	10.8	1.96	0.00	1.96	0.00	0.06	0.06	14.1	0.00	14.1	1.16	0.00	1.16	0.03	0.28	0.31	1.10	0.00	1.11	1.05	0.00	1.05
100,000	0.00	0.28	0.28	10.8	0.00	10.8	1.96	0.00	1.96	0.00	0.04	0.04	14.1	0.00	14.1	1.16	0.00	1.16	0.02	0.21	0.22	1.10	0.00	1.11	1.05	0.00	1.05

※　var : $\sum_{d=1}^{100} \hat{\sigma}_d{}^2$，bias : $\sum_{d=1}^{100} \hat{b}_d{}^2$，mse : $\sum_{d=1}^{100} \hat{\sigma}_d{}^2 + \sum_{d=1}^{100} \hat{b}_d{}^2$

$$\hat{Y}_d = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_{1d} + \hat{\beta}_2 \bar{x}_{2d} + \hat{\beta}_3 \bar{x}_{3d} + \hat{\beta}_4 \bar{x}_{4d} + \hat{\beta}_5 \bar{x}_{5d} + \hat{\gamma}_1 z_{1d} + \hat{\gamma}_2 z_{2d} + \hat{\eta}_d \quad (14)$$

また，モデルの精度が低い状況を表現するために，実際のデータ生成過程からいくつか変数を除外した以下の推定量についても考察を加える．

Estimator 5: model-based 直接推定量（変数一部削除）

$$\hat{Y}_d = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_{1d} + \hat{\beta}_2 \bar{x}_{2d} \quad (15)$$

Estimator 6: B-model-based 直接推定量（変数一部削除）

$$\hat{Y}_d = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_{1d} + \hat{\beta}_2 \bar{x}_{2d} + \hat{\tau} L_d \quad (16)$$

Estimator 7: EBLUP 推定量（変数一部削除）

$$\hat{Y}_d = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_{1d} + \hat{\beta}_2 \bar{x}_{2d} + \hat{\eta}_d \quad (17)$$

さらに，ビッグデータを用いた推定量（Estimator 3）についてはさらに補助情報を削減した以下の 2 つの推定量も検証の対象とする．これは，ビッグデータ共変量は実際のデータ生成過程と整合的な変数ではなく，他の説明変数と一定程度の相関を持つ可能性があることから，関心のない説明変数を削除することで推定精度を向上できる可能性があるためである．

Estimator 8: B-model-based 直接推定量（変数x_{1id}のみ）

$$\hat{Y}_d = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_{1d} + \hat{\tau} L_d \quad (18)$$

Estimator 9: B-model-based 直接推定量（変数なし）

$$\hat{Y}_d = \hat{\beta}_0 + \hat{\tau} L_d \quad (19)$$

4. シミュレーション結果

以下のシミュレーション分析では，標本サイズを 100, 300, 500, 700, 1000, 3000, 5000, 7000, 10000, 30000, 50000, 70000, 100000 と変化させた場合の平均二乗誤差を確認する．なお，標本サイズが 0 となる地域が発生した場合は，全体平均で補完した．

表-2 に$\sigma_L = 0.1$の場合のバイアス，分散，平均二乗誤差（全地域の集計量）を示す．また，地域間の平均二乗誤差の違いを示すために，図-2～図-5 に平均二乗誤差の分布（箱ひげ図）を示す．結果から得られる主要な知見は以下のとおりである．

1) バイアス最小化という視点からは，標本サイズが大きい（3000 以上）場合は Estimator 1 が望ましく，標本サイズが小さい場合は Estimator 4 が望ましい．
2) Estimator 1 および Estimator 2 の比較から，標本サイズが 1000 を下回る場合は Estimator 2 の方が平均二乗誤差は小さいが，標本サイズが大きい場合，Estimator 1 の方が平均二乗誤差は小さくなる（ただし，需要予測に利用する場合，Estimator 1 の結果をもとにモデルを作成する必要があり，モデルの誤差が追加される点に注意する必要がある）
3) ビッグデータを共変量として導入した Estimator 3 は，標本サイズが 10000 以下の場合 Estimator 1 および Estimator 2 よりも精度の高い推定量となったが，標本サイズが 30000 を超える場合，Estimator 1 の方が精度の高い推定量となる
4) （理論的に明らかなように）Estimator 4 は Estimator 1 及び Estimator 2 よりも常に低い平均二乗誤差となる
5) ビッグデータを利用しない場合，変数を一部削除することで推定精度は低下するが（Estimator 5 及び 6）ビッグデータを共変量として導入する場合，変数を削除した推定量（Estimator 8 及び 9）の方が平均二乗誤差の観点から優れた推定量が得られる

表-3 にσ_Lを変化させた場合のバイアス，分散，平均二乗誤差（全地域の集計量）を示す．また，図-6 に$\sigma_L = 1.0$の場合の平均二乗誤差の分布（箱ひげ図）を示す．結果より，ビッグデータの精度が高い場合（$\sigma_L = 0.1$），上述したように他の説明変数を除外したモデルの推定量が望ましい一方で，ビッグデータの精度が低い場合（具体

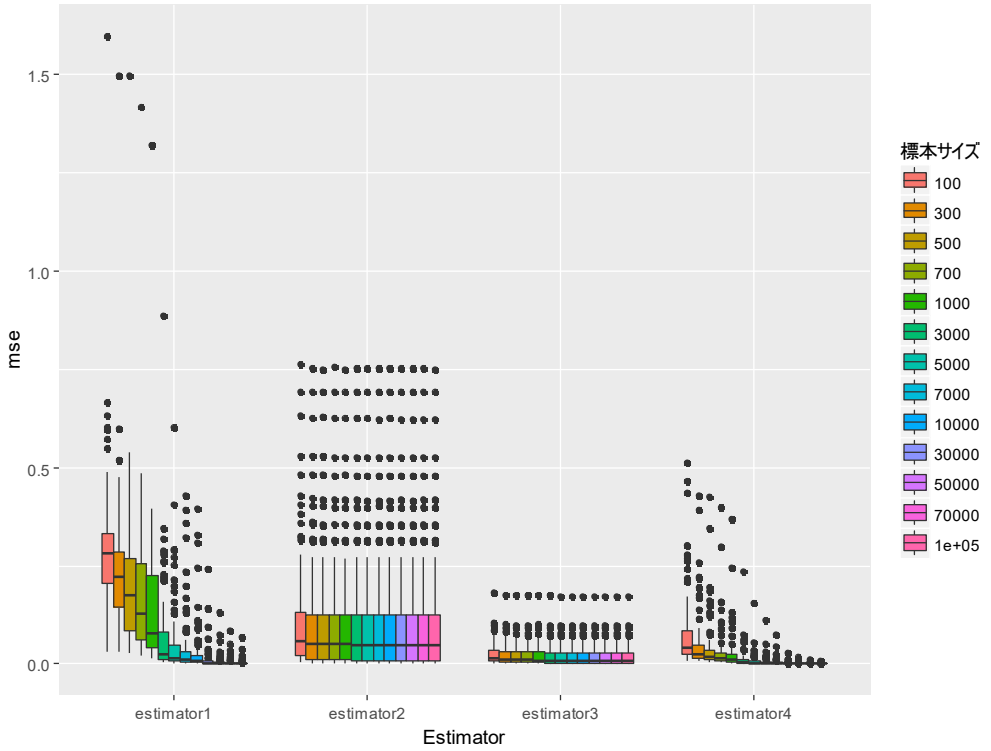

図-2　Estimator 1~4の平均二乗誤差の分布 ($\sigma_L = 0.1$)

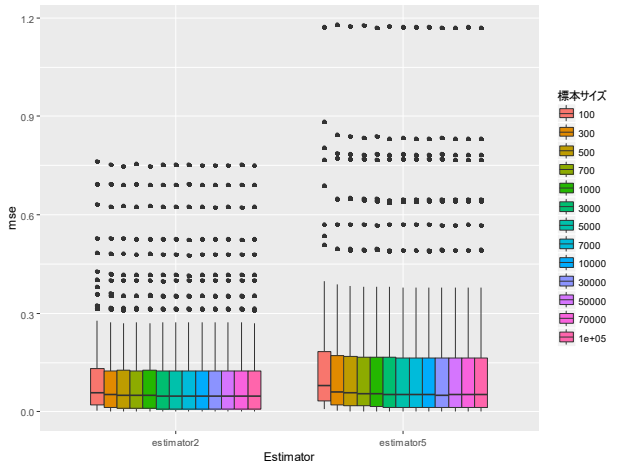

図-3　Estimator 2および5の平均二乗誤差の分布 ($\sigma_L = 0.1$)

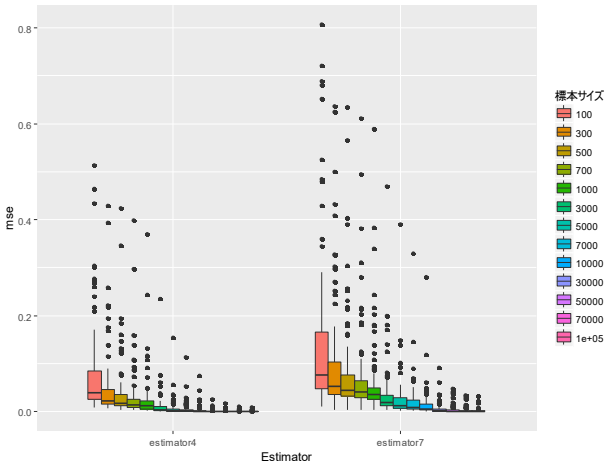

図-4　Estimator 4および7の平均二乗誤差 の分布($\sigma_L = 0.1$)

表-3　異なるビッグデータ精度下における推定量のバイアス，分散，平均二乗誤差

	σ_L=0.1												σ_L=0.5												σ_L=1.0											
標本	Estimator 3			Estimator 6			Estimator 8			Estimator 9			Estimator 3			Estimator 6			Estimator 8			Estimator 9			Estimator 3			Estimator 6			Estimator 8			Estimator 9		
サイズ	bias	var	mse	bias	var	mse	bias	var	mse	bias	var	mse	bias	var	mse	bias	var	mse	bias	var	mse	bias	var	mse	bias	var	mse	bias	var	mse	bias	var	mse	bias	var	mse
100	1.96	0.66	2.62	1.14	2.38	3.52	1.12	1.35	2.47	1.04	1.24	2.29	7.76	0.87	8.63	7.96	2.53	10.49	8.71	1.31	10.02	9.22	1.21	10.43	10.67	1.12	11.79	13.64	2.72	16.37	16.97	1.71	18.67	17.72	1.58	19.31
300	1.95	0.20	2.16	1.16	0.73	1.89	1.10	0.41	1.52	1.05	0.39	1.43	7.77	0.28	8.05	7.93	0.84	8.77	8.65	0.44	9.09	9.16	0.41	9.56	10.68	0.37	11.05	13.67	0.91	14.58	16.94	0.56	17.50	17.71	0.53	18.25
500	1.97	0.12	2.09	1.15	0.43	1.58	1.10	0.24	1.34	1.05	0.23	1.28	7.75	0.16	7.91	7.95	0.52	8.47	8.67	0.29	8.96	9.18	0.27	9.45	10.71	0.22	10.92	13.68	0.52	14.21	17.01	0.32	17.32	17.76	0.29	18.05
700	1.96	0.09	2.05	1.16	0.32	1.47	1.09	0.18	1.28	1.04	0.17	1.21	7.78	0.11	7.89	7.97	0.36	8.33	8.68	0.19	8.87	9.20	0.18	9.37	10.73	0.15	10.88	13.72	0.37	14.09	16.98	0.22	17.20	17.75	0.20	17.96
1,000	1.97	0.06	2.03	1.17	0.23	1.40	1.12	0.13	1.25	1.06	0.12	1.18	7.77	0.08	7.85	7.96	0.24	8.20	8.69	0.13	8.81	9.20	0.12	9.31	10.70	0.10	10.80	13.68	0.24	13.92	16.95	0.15	17.10	17.72	0.14	17.86
3,000	1.97	0.02	1.98	1.17	0.07	1.24	1.11	0.04	1.15	1.05	0.04	1.09	7.78	0.03	7.81	7.97	0.08	8.06	8.68	0.05	8.72	9.18	0.04	9.23	10.71	0.03	10.74	13.68	0.09	13.77	16.96	0.05	17.01	17.73	0.05	17.78
5,000	1.96	0.01	1.98	1.16	0.04	1.21	1.10	0.03	1.13	1.05	0.02	1.07	7.78	0.02	7.80	7.97	0.05	8.02	8.68	0.03	8.71	9.19	0.02	9.22	10.69	0.02	10.72	13.66	0.05	13.71	16.97	0.03	17.00	17.73	0.03	17.76
7,000	1.96	0.01	1.97	1.16	0.03	1.19	1.10	0.02	1.12	1.05	0.02	1.07	7.78	0.01	7.79	7.98	0.03	8.01	8.68	0.02	8.70	9.19	0.02	9.21	10.70	0.01	10.71	13.66	0.04	13.69	16.96	0.02	16.98	17.73	0.02	17.75
10,000	1.96	0.01	1.97	1.16	0.02	1.18	1.10	0.01	1.12	1.05	0.01	1.06	7.78	0.01	7.78	7.97	0.02	8.00	8.68	0.01	8.69	9.19	0.01	9.20	10.70	0.01	10.71	13.67	0.03	13.70	16.96	0.02	16.97	17.73	0.01	17.74
30,000	1.96	0.00	1.96	1.16	0.01	1.17	1.10	0.00	1.11	1.05	0.00	1.05	7.78	0.00	7.78	7.98	0.01	7.98	8.68	0.00	8.69	9.19	0.00	9.19	10.70	0.00	10.70	13.67	0.01	13.68	16.96	0.01	16.97	17.73	0.00	17.74
50,000	1.96	0.00	1.96	1.16	0.00	1.16	1.10	0.00	1.11	1.05	0.00	1.05	7.78	0.00	7.78	7.97	0.00	7.98	8.68	0.00	8.68	9.19	0.00	9.19	10.70	0.00	10.70	13.68	0.01	13.68	16.97	0.00	16.97	17.73	0.00	17.74
70,000	1.96	0.00	1.96	1.16	0.00	1.16	1.10	0.00	1.11	1.05	0.00	1.05	7.78	0.00	7.78	7.97	0.00	7.98	8.68	0.00	8.68	9.19	0.00	9.19	10.70	0.00	10.70	13.68	0.00	13.68	16.97	0.00	16.97	17.74	0.00	17.74
100,000	1.96	0.00	1.96	1.16	0.00	1.16	1.10	0.00	1.11	1.05	0.00	1.05	7.78	0.00	7.78	7.97	0.00	7.98	8.68	0.00	8.68	9.19	0.00	9.19	10.70	0.00	10.70	13.68	0.00	13.68	16.97	0.00	16.97	17.73	0.00	17.74

※　var : $\sum_{d=1}^{100} \hat{\sigma}_d{}^2$,　bias : $\sum_{d=1}^{100} \hat{b}_d{}^2$,　mse : $\sum_{d=1}^{100} \hat{\sigma}_d{}^2 + \sum_{d=1}^{100} \hat{b}_d{}^2$

的には$\sigma_L = 0.5$ および 1.0の場合），すべての説明変数を加えた Estimator 3 を採用することが望ましいことが確認された．従って，小地域の推定精度を高めることを目的としてビッグデータを共変量として導入したモデルを用いる場合，モデルの選択（説明変数の選択）はビッグデータの精度に影響を受けるといえる．

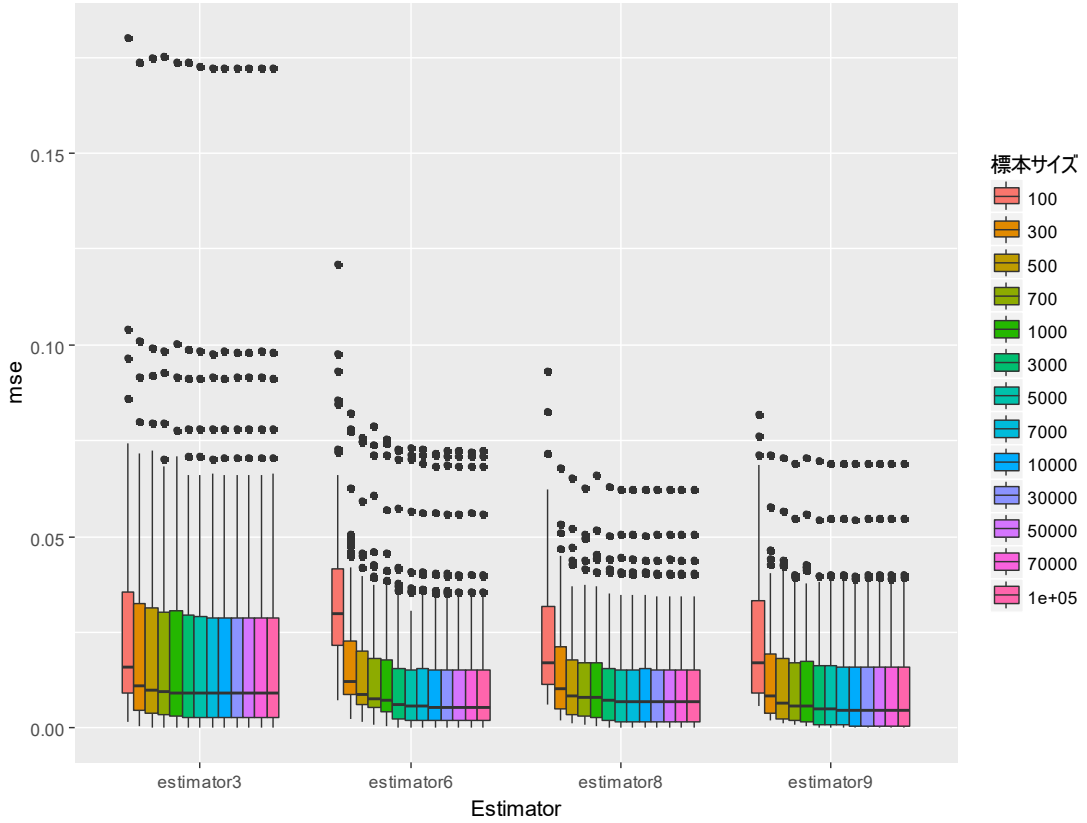

図-5　Estimator 3, 6, 8および7の平均二乗誤差 の分布($\sigma_L = 0.1$)

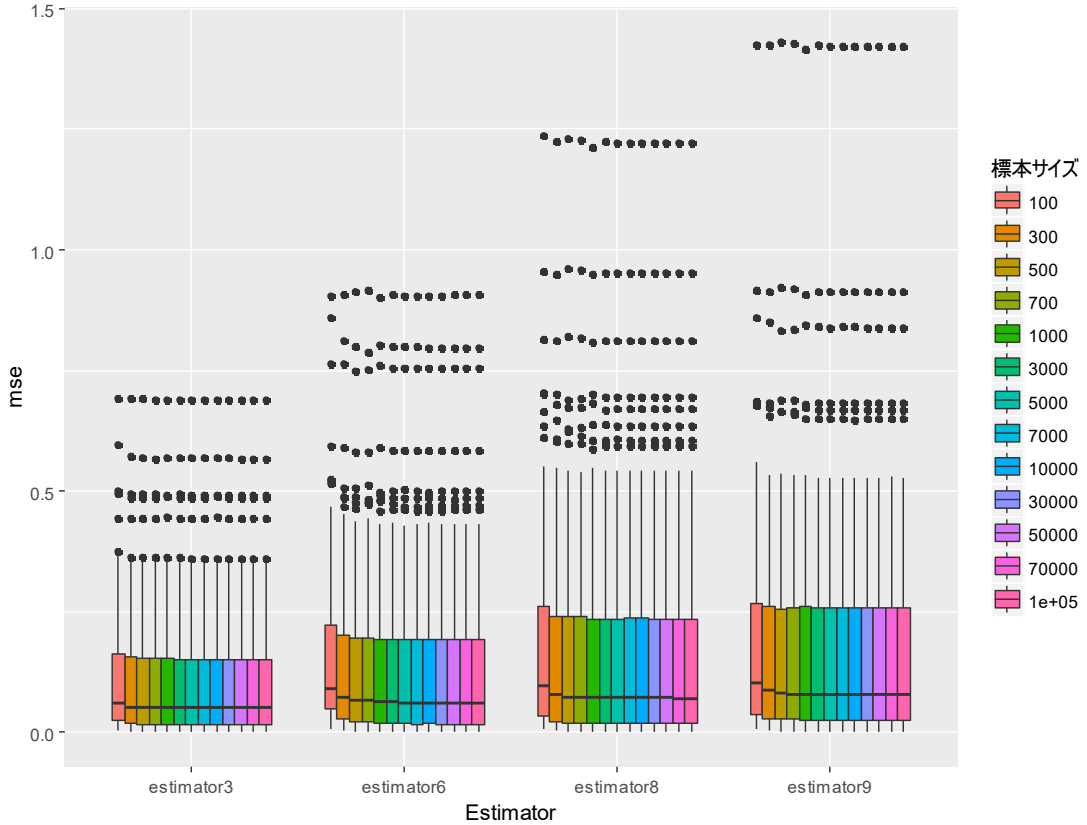

図-6　Estimator 3, 6, 8および7の平均二乗誤差 の分布($\sigma_L = 1.0$)

6. おわりに

本研究で行ったシミュレーションは，簡略化された仮想状況におけるシンプルな分析であるものの，以下に示すように，いくつかの重要な示唆が得られた．

1) 標本サイズが小さい場合においてはビッグデータを利用した推定量の精度が最も高い
2) 標本サイズが小さい場合，欧米のアプローチ（model-based 直接推定量）の採用が推定精度の観点から望ましいものの，model-based 直接推定量を採用する場合，標本サイズの増加による精度の向上は極めて小さく，標本サイズが一定以上になると我が国のアプローチ（design-based 直接推定量）の方が推定精度が高い（ただし，モデルの誤差は含まない）
3) design-based直接推定量とmodel-based直接推定量の重み付け平均である model-based 間接推定量を用いることで，推定精度を大きく向上させることができる
4) ビッグデータ共変量はデータ生成過程を無視した変数であるため，推定精度最大化の観点からモデル選択をすると，本来のデータ生成過程を反映していないモデルを選択してしまう可能性がある

1)は，データの利用が困難な途上国での解析や，標本サイズが不可避的に小さくなる細かな地域の推定量が必要となる場合にビッグデータの利用価値が特に高いことを示唆している．2)は，我が国で行われている大規模交通日誌調査を小規模交通日誌調査に置き換えようとする場合，精度の高いモデル構築ができることが前提条件となることを示唆している．3)は，EBLUP 推定量を離散選択モデル等にまで拡張したベイズの枠組みで発生・集中，分布，分担までモデル化すれば日本型アプローチ，及び，欧米型アプローチのそれぞれの利点を活かした推定が可能になることを示唆している．4)は，本稿では解析できていないものの，本来のデータ生成過程と異なるモデルとなってしまうことから，図-1 の予測の段階で更なる誤差が発生することが予想される．このことは，ビッグデータを共変量として推定量を構成する際の大きな課題となり得る．また，この誤差が大きい場合，予測の文脈においては，ビッグデータの共変量としての利用よりはむしろ，予測値の外的妥当性を確認する補完的な情報としてビッグデータを用いる方が望ましい可能性がある．この点に関する検討結果は発表時に報告する．

本研究の結果は，第一に，ビッグデータを需要予測に利用する前に，ビッグデータの精度検証を丁寧に行うことの重要性を示唆している．また，上述したように，ビッグデータの導入は推定結果にバイアスをもたらす可能性がある．バイアスを有する推定値を利用することの是非についても今一度議論を深める必要がある．

本研究の延長として，同一人物の行動を複数時点において観測するパネル交通日誌調査や，365 日連続して小標本の交通行動の観測を行う Continuous 交通日誌調査など，他の調査手法を組み合わせた場合の精度検証を行う方向性が考えられる．また，本研究ではトリップ生成にのみ焦点を当てたが，その他の行動側面についても推定精度を検証する必要がある．加えて，求められる予測精度を今一度精査することが，実務の場面における交通日誌調査の標本サイズの設計やモデル構築の文脈においては極めて重要である．この点についても仮の目標精度を定めた上でシミュレーション分析を実施することは可能であり，様々なケースを想定したシミュレーション分析を蓄積することが有用と考えている．

参考文献

1) 土屋隆裕, *概説標本調査法*: 朝倉書店, 2009.
2) Rao, J. N. K., and Molina, I., *Small Area Estimation*, 2nd Edition ed.: Wiley, 2015.
3) 計量計画研究所（編）, *総合都市交通体系調査の手引き*, 第 2 版 ed.: 計量計画研究所, 2007.
4) Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., Rinzivillo, S., Pappalardo, L., and Gabrielli, L.: Small Area Model-Based Estimators Using Big Data Sources, *Journal of Official Statistics*, 2, 2015, p. 263.

Travel Diary Survey Design and Model Selection in an Era of Big Data: A Small Area Estimation Approach

Makoto CHIKARAISHI, Akimasa FUJIWARA