研　究　報　告　書

「マルチソースデータ高度利用のための統計的データ融合」

研究期間:　平成 20 年10月～平成 24 年3月
研　究　者:　星野　崇宏

1．研究のねらい

製品開発やマーケティング戦略のために得られる様々なデータは複数の情報源から得られるマルチソースデータであることがほとんどである（図1）。マルチソースデータからは各データ間をまたがる変数間の関連を見ることができず、これは実務上では非常に大きな問題となる。

本研究では、各データセットで測定対象となっている個々人の背景情報を積極的に測定し、その情報を利用することでマルチソースデータからシングルソースデータをシミュレートし補完する統計的なデータ融合手法を開発することを目的とする。具体的には、近年統計学において発展の目覚ましいセミパラメトリックな手法を用いたデータ融合手法を開発する。また、これまで注目されなかった問題として、各データセット間での対象者の異質性が重要である。例えば購買データ上には購入頻度の高い顧客が多いのが通常である。従って各データセットで得られる対象者の違いによるバイアスを除去する必要がある。加えて、先行研究では重要視されていなかった「背景情報＝共変量情報」の積極的な探索と利用を行うことで、これまで提案されてきた欠測を補完する方法の様々な問題点を克服する。さらに、単に数理的な手法の開発にとどまらず、実際の製品開発とマーケティング戦略に関する実験調査への応用可能性について実データを積極的に用いて検証する。具体的には、本研究に協力頂ける企業が有する大規模な調査パネルに対する市場調査を実施し、企業が取得している Web 閲覧データと融合させる、あるいは POS 等の実績データなどと融合させることでマーケティング分野での実際の予測を行い、一定の汎用性が得られるかを検討する。

データ融合が対象とするマルチソースデータと同様のデータ構造、解析ニーズは社会科学一般において存在する。そこで、経済学・社会学・教育学分野においてデータ融合の応用研究を行う。例えば repeated cross-sectional dataset から panel data によってのみ得られる情報を抽出する、疑似パネルデータ解析はデータ融合が対象とするデータ構造と基本的に同じものである。そこで、既存の疑似パネルデータ解析では可能ではなかった、各データセットに対する対象者の割り当てが無作為でない場合にも利用可能な方法を、データ融合の方法論を応用し開発する。

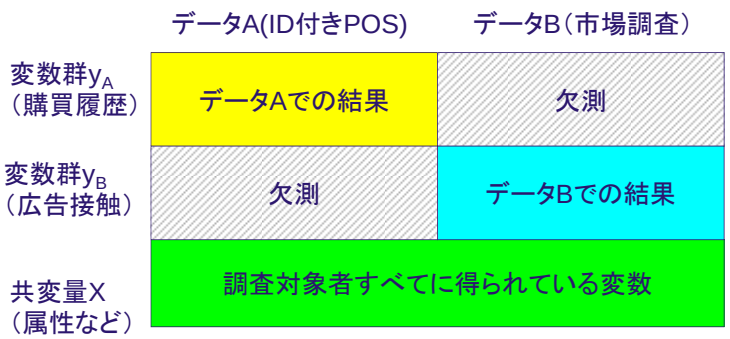

図1：もっともシンプルなマルチソースデータセットの形式

2．研究成果

得られた研究成果は大別すると以下の5つである。

（1）データ融合についての先行研究での前提条件の解明とその緩和

これまでデータ融合の具体的方法として利用されてきたのは「マッチング」「潜在変数モデル」「回帰的モデル」である。まず先行研究で提案されている方法を調べ、これらがすべて以下の仮定を置いていることを確認した。

【1】ランダムな欠測(Missing at random)

どのデータで観測されるかが、共変量に依存しており、アウトカムには依存しない。

【2】条件付き独立(conditional independence)

共変量を所与としたアウトカムは各データセットで独立である。

しかし、実は上記の条件は厳しすぎるため、現実のデータセットに適用できるかどうかは疑問である。そこで、データ構造を欠測データセットの考え方のもとに整理した結果、それぞれ特定のパラメトリックモデル、たとえば各データへの所属インディケータの背後に連続量を仮定するプロビットモデルを考え、また各アウトカムが正規分布に従うと仮定した場合には

【1‘】ランダムでない欠測

どのデータで観測されるかは、基準となるデータセットでのアウトカムには依存してもよい。

【2‘】連関の許容

各データセットでのアウトカムが連関することを許容する。

という形でそれぞれの条件を大幅に緩和してよいことが分かった。

さて、アウトカムのモデルには特定の分布仮定を置くことは消費者行動理論や計量経済学などマーケティングサイエンスの基礎となる諸研究から正当化することはできる。一方、どのデータでその対象者が観測されるか（あるいは欠測となるか）については特定の根拠をもとにパラメトリックなモデルを仮定することは難しい。そこで本研究では対象者のデータセット割り当て（欠測）モデルについてはパラメトリックな仮定を置かず、アウトカムの周辺同時分布についてはパラメトリックな仮定を置くセミパラメトリックモデルを利用することとした。このようなモデルで最も効率的であり、かつ欠測データの精度の高い予測も可能にするモデルとして、ディリクレ過程混合分布を用いたセミパラメトリックベイズ推定法を開発した。シミュレーションの結果からは既存の手法よりも大幅に推定の誤差を減少させることがわかった。また製品の購入とその製品についてのインターネットサイトの閲覧の関係を調べる広告効果測定の実データにおいても、既存法では連関指標の推定において明らかな過小推定が生じてしまうが、提案手法は真値を復元することがわかった（図2）。

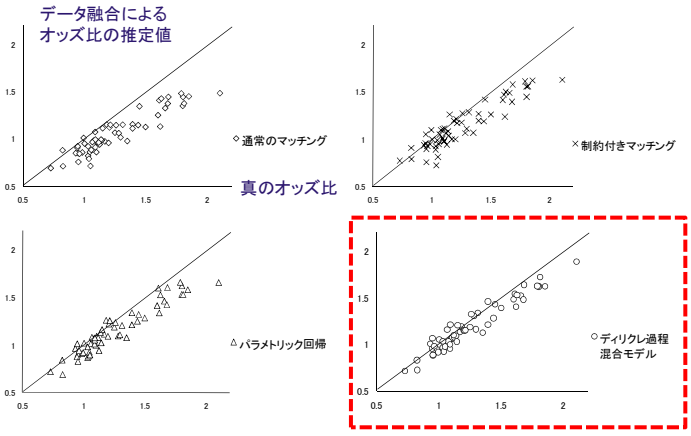

図2：既存法と提案手法（赤枠）の比較

（2）感度分析法の開発

上記の方法論によりデータ融合の仮定は大幅に緩和されたが、実務では各データセットでのアウトカムが高い関連を有する場合も多く、基準となるデータセットがどれであるか不明な場合もある。このような場合には観測されている共変量だけではなく、観測されない未知の共変量によって複数のアウトカムが説明されると考えるのが自然である。但し、潜在変数を利用したモデルでは識別性が無くなってしまう（データのみから母数が決定されない）。一方、潜在変数が観測確率に与える影響に関する母数のみ値を固定すれば、それ以外の母数については推定を行うことができるため、その母数の値を一定の範囲で変化させながら推定を行うことで、アウトカム間の連関がどの範囲に変動するか（そしてその信頼区間はどの程度か）を調べる感度分析法を開発した。この方法を用いることで、欠測を例えば最大 50%説明する未知の共変量が存在したとしても、どの程度の範囲に 95%信頼区間が収まるかを推定することが可能になる（図3）。

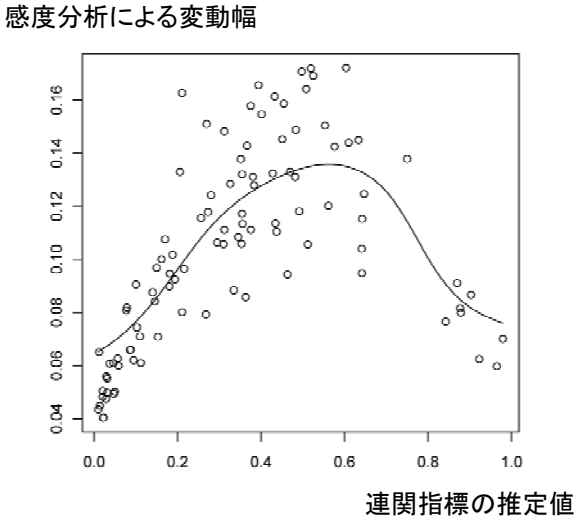

図3：感度分析法による 95%信頼区間の変動幅の推定

（3）マーケティング実務におけるニーズを満たす応用：属性情報の付加

マーケティングの実務において必要とされるデータ融合の応用を複数実施した。一例として、フリークエントショッパーズプログラム(FSP)データにおける属性情報の付加に関する応用を行った。FSP データでは同一の消費者に対して複数時点での購買履歴がひも付けされるという点で非常に豊かなデータであるが、属性情報は一般に性別や年齢、住所など会員登録時に得られる最低限の情報にとどまる。一方、企業がマーケティング戦略を策定する際には、収入や家族構成、職種、居住形態やライフスタイルなどの情報を利用し適切なセグメント設定を行うことや、細かな属性別の市場規模の予測を行うことが適切な広告戦略や価格戦略につながるが、このデータではそれを可能にするに十分な属性情報がない。従って実務上では FSP データに対する属性情報の付加には大きなニーズがある。そこで ID 付きPOS データのうち、市場調査も行うパネル型調査データから、「FSP データ」と、「細かな属性情報と比較的粗い購買履歴情報を測定する市場調査データ」の2つのデータセットにあえ

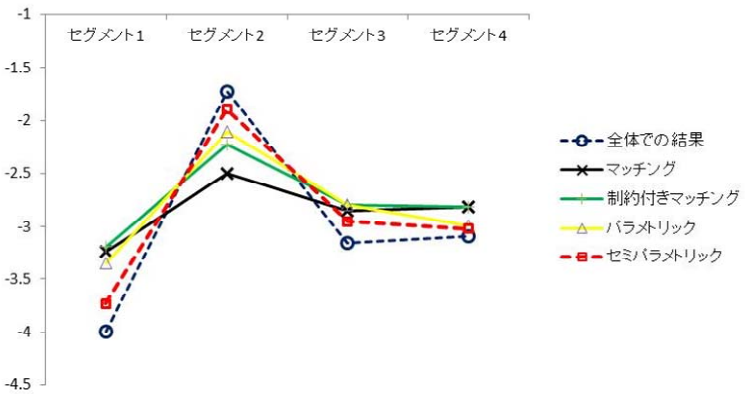

図4：データ融合を用いた価格弾力性の復元

て分離し、その2つのデータを融合させ、FSP データ上の対象者に対する意味のあるセグメンテーションを行い、セグメント別のロイヤルティ係数や価格弾力性の推定を行った。結果としては提案手法はパネル調査データから得られる真のデータ構造を正しく復元することができた（図4）。

（4）疑似パネルデータセットでの因果効果推定

経済学や社会学、心理学などでは同一対象者を追跡調査するパネル調査研究を行うことで、個人差を除去した「特定の独立変数の従属変数への因果効果」を推定することが多い。実際には標本の代表性を担保しながら、ドロップアウトのないパネル調査を実施することは難しいため、パネル調査ではなく、同じ標本抽出法を利用して抽出された、対象者の異なる２時点の調査データ（repeated cross-sectional data）から解析を行わざるを得ないことがある。このような研究関心を一般的に疑似パネルデータ解析と呼ぶが、経済学等で利用されてきた手法はマッチング及び線形の回帰モデルを用いるものであり、（1）で述べたような観点からはデータ融合の考え方を用いることでより頑健かつ効率的な推定法を得ることが可能であると考えられる。

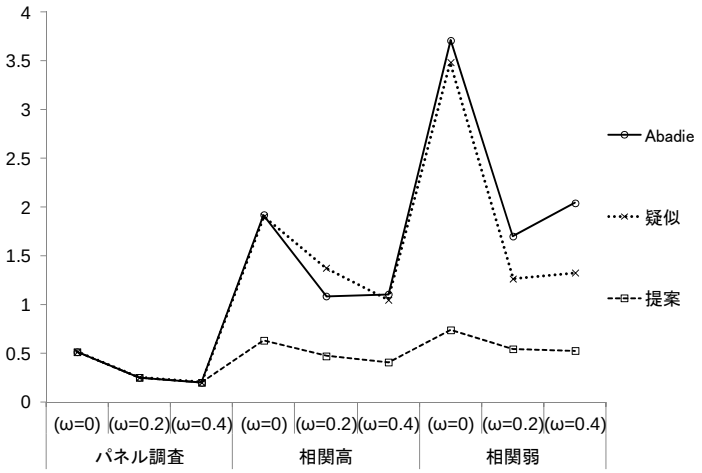

図5：様々な設定での二乗誤差の手法間比較

そこで、本研究で提案したセミパラメトリックなデータ融合法を用いることで、対象者の異なる repeated cross-sectional data を用いて因果効果を推定する方法を開発したところ、既存の手法（一時点のデータによる Abadie 法や既存の疑似パネルの解析法)よりも二乗誤差が小さい推定法を得ることが出来た（図5）。

（5）マルチレベルデータでのデータ融合

社会科学や疫学においては学校-教室-生徒、地域-病院-患者、のような階層データが得られることが普通である。ここで、学校への教育費の投入や教員の質、病院の設備など「上位のユニット(level-2 unit)の効果」と生徒の家庭環境や患者の健康状態など「下位のユニット(level-1 unit)の効果」、そしてその相互作用がどのように結果に影響を与えうるかを調べるためにマルチレベル分析が用いられるが、例えば「地方政府からの教育費投入の程度によって能力別学級の効果はどのように異なるか」といった効果の推定は教育政策や医療政策において非常に重要である。しかしどの「上位ユニットへの下位ユニットの割り当て」はランダムに行えないので、通常は上記の効果と「下位ユニットの性質」が交絡して正しい推定ができない。

そこでデータ融合の考え方を用いたマルチレベルにおける新しい因果効果の推定法を開発した。具体的には、既存の方法では上位ユニットごとへの割り当てをモデリングする必要があるが、上位ユニットは通常は極めて多く、実用的ではない。そこで、上位ユニットの特徴を説明する変数についての割り当てをモデリングするセミパラメトリックな inverse probability weighting

estimator を開発し、既存の方法よりも頑健かつ安定した推定値を与えることを確認した。

3，今後の展開

本研究では当初の予想を超えて数理的な手法論開発に重点的に取り組む必要が生じたため、マーケティングにおける応用研究としてはインターネットのサイト閲覧についてのパネル調査データ、購買履歴に関する調査データ、消費者に対する市場調査データのデータ融合のみとなった。今後はスマートフォンなどのモバイル端末からのセンサーデータや、消費者生成メディアなど様々なデータとのデータ融合についての応用を行う予定である。

また社会科学等他分野への応用という観点からは、教育学や疫学等の調査データ、心理学や行動経済学での実験データ等に対しての応用が可能であるため、これに取り組みたい。

共変量選択法の開発については、クロスバリデーションなど開発された手法から自然に導出される方法を用いることが可能であることは分かった。しかし、本来は「データ取得や調査設計段階でどのような共変量を測定すれば良いかについて一般的な示唆を与える」ことが実務的な観点からは要請されている。現状の共変量選択はあくまで事後的な選択法でしかなく、この点については実務的な要請に答える段階にはない。従って、今後より多くの応用例を積み重ねることで、どのように収集された（どのように偏った）データセットのどのようなアウトカム間の関連を調べる際には、どのような共変量を事前に測定しておくことが重要かについての経験則を得ることが、より広範囲の実務家に利用されるためには必要であると考えられる。

4，自己評価

数理的な方法論の開発と言う観点では、「既存のパラメトリックな手法をセミパラメトリックモデルで代替する」という当初目標にとどまらず、「既存の手法の仮定を大幅に緩和することが可能である」ことを発見できたため、非常に大きな進展があったと評価できると考える。また、疑似パネルデータ解析やマルチレベル分析において、今回開発した方法を応用することで既存の方法論よりも頑健かつ効率の良い因果効果推定の方法を開発することができたという点で社会科学への応用研究と言う点で当初目標を達成したと考えられる。

一方、実務に応用できるパッケージ化という観点では、4の今後の展開で記載したように事前の共変量選択に対する示唆を得ることが出来なかったという点が残念であるが、これも4で記載したように、様々な応用例の積み重ねを行うことで今後一般則を見出して行きたい。

5，研究総括の見解

マルチソースデータからシングルソースデータをシミュレートし補完する手法であり、断片的な大量データから各種の統計的推定を安定に行うために不可欠な技術である。必ず開発していかなければならない技術であり、コア的研究といえる。研究成果を具体的な調査活動につなげることを意識して研究を進めることを期待していた。

数理的な方法論の開発と言う観点では、「既存のパラメトリックな手法をセミパラメトリックモデルで代替する」という当初目標にとどまらず、「既存の手法の仮定を大幅に緩和することが可能である」ことを発見できたため、非常に大きな進展があったと評価する。また、疑似パネルデータ解析やマルチレベル分析において、今回開発した方法を応用することで既存の方法論よりも頑健かつ効率の良い因果効果推定の方法を開発することができたという点で社会科学への応用

研究と言う点で当初目標を達成していると評価する。この分野に大きく貢献していると考える。

実務に応用できるパッケージ化という観点では、事前の共変量選択に対する示唆を得ることが出来なかったという点が残念であるが、今後、様々な応用例の積み重ねを行うことで一般則を見出していくことを期待する。

6, 主な研究成果リスト

(1)論文(原著論文)発表

1. Kei Miyazaki and Takahiro Hoshino, "A Bayesian Semiparametric Item Response Model with Dirichlet Process Priors", Psychometrika, 74　375-393. 2009.

2. Kei Miyazaki, Takahiro Hoshino, Shin-ichi Mayekawa and Kazuo Shigemasu. "A New Concurrent Calibration Method for Nonequivalent Group Design under Nonrandom Assignment", Psychometrika, 74　1-19. 2009.

3. 星野崇宏 "調査不能がある場合の標本調査におけるセミパラメトリック推定と感度分析:日本人の国民性調査データへの適用" 統計数理, 第 58 巻 1 号　3-23.

4. 猪狩良介・星野崇宏 (印刷中) "非集計 Web アクセスデータを用いたサイト普及モデル:多時点・複数サイトの階層ベイズモデリング" マーケティング・サイエンス, 2012.

5. Tetsuro Kobayashi and Takahiro Hoshino"Propensity Score Adjustment for Internet Panel Surveys of Voting Behavior: A Case in Japan",Japanese Journal of Electoral Studies. 27, 104-117, 2011

6.Yusuke Takahashi, Robert W. Brent, and Takahiro Hoshino (in press) "Conscientiousness mediates the relation between perceived parental socialization and self-rated health"

7. Shiro Ojima, Naoko Nakamura, Hiroko Matsuba-Kurita, Takahiro Hoshino and Hiroko Hagiwara (2011)"Age and amount of exposure to a foreign language during childhood: Behavioral and ERP data on the semantic comprehension of spoken English by Japanese children",Neuroscience Research. 2011 Jun;70(2):197-205.

8. Takehiro Nagai, Takahiro Hoshino and Keiji Uchikawa (2011) "Statistical Significance Testing with Mahalanobis Distance for Thresholds Estimated from Constant Stimuli Method" Seeing and Perceiving, 24, 91-124.

9. Shiro Ojima, Naoko Nakamura, Hiroko Matsuba-Kurita, Takahiro Hoshino and Hiroko Hagiwara (2011). "Neural correlates of foreign-language learning in childhood: A 3-year longitudinal ERP study", Journal of Cognitive Neuroscience. 23, 183-199.

10. Atsunobu Suzuki, Takahiro Hoshino, and Kazuo Shigemasu (2010) "Happiness is unique: A latent structure of emotion recognition traits revealed by statistical model comparison" Personality and Individual Differences, 48. 196-201.

(2)特許出願

なし

(3)その他の成果(主要な学会発表、受賞、著作物等)

【著作】

・星野崇宏「調査観察データの統計科学：因果推論／選択バイアス／データ融合」岩波書店

【学会発表】

・Takahiro Hoshino, "Causal Inference Framework for Latent Variable Modeling" Invited Talk, The 76th Annual and the 17th International Meetings of the Psychometric Society, Hong-Kong Institute of Education, Hong-Kong, 2011.

・星野崇宏「反実仮想モデルを用いた統計的因果推論について」、第 13 回情報論的学習理論ワークショップ(IBIS2010),東京大学,2010

・猪狩良介、星野崇宏「非集計 Web アクセスデータを用いたサイト普及モデル－多時点複数サイトの階層ベイズモデリグ」,第 86 回日本マーケティングサイエンス学会研究大会、電通ホール 2009

・太田悠太・星野崇宏「広告効果に影響を及ぼす CM 内容：クリエイティブの科学的管理に向けて」,日本消費者行動研究学会第 41 回消費者行動研究コンファレンス、関西学院大学,2010

・直井映里砂・星野崇宏「ネガティブ情報を積極的に提示すべきか？：企業イメージや購買意図への影響」日本消費者行動研究学会第 41 回消費者行動研究コンファレンス、関西学院大学,2010

・猪狩良介、星野崇宏「サイト属性と訪問経路情報を用いた Web 閲覧行動モデル」第 88 回日本マーケティングサイエンス学会研究大会、電通ホール, 2010

・太田悠太、星野崇宏「プロスペクト理論を考慮した同時購買行動での価格プロモーション戦略」第 90 回日本マーケティングサイエンス学会研究大会、株式会社電通 電通ホール,2011

・Miyazaki K., Hoshino T, & Shigemasu K. A Bayesian Equating Method for Nonequivalent Group Design under Nonrandom Assignment via Data Augmentation Algorithm. The 75th Annual Meeting of the Psychometric Society. (The University of Georgia. Athens, Georgia, USA., 2010

・Miyazaki, K., Hoshino, T, & Shigemasu, K. Determining the direction of the path using a Bayesian semiparametric model. 19th International Conference on Computational Statistics. (Conservatoire National des Arts et Métiers, Paris, France, 2010

・宮崎慧. 星野崇宏. 複数商品購買行動のための階層ベイズグレンジャー因果性分析 日本マーケティングサイエンス学会第 88 回研究大会, 電通ホール, 2010

・宮崎慧. 星野崇宏. 動的階層ベイズモデルを用いた複数商品カテゴリー購買とブランド購買の同時分析 2011 年度統計関連学会連合大会, 九州大学, 2011

・宮崎慧. 階層ベイズ動的ポアソンモデルによる複数商品購買行動の分析 行動計量学会第 39 回大会（岡山理科大学, 2011 年 9 月）・星野崇宏「観察研究・調査データからの因果効果の推定」ICPSR 国内利用協議会統計セミナー、 関西学院大学, 2009

・星野崇宏「Web 調査の偏りの補正: 行動経済学における調査研究への適用」, 関西大学ソシオネットワーク戦略研究機構 第 5 回 RISS 経済政策特別講義、 関西大学, 2010

・星野崇宏「標本調査法への統一的なアプローチと新展開」, 2010 年度統計関連学会連合大会、早稲田大学, 2010

・星野崇宏「課題研究：教育調査の在り方を問い直す－量的研究の課題と展望－：統計学の観点から見た量的研究の課題と今後」, 日本教育社会学会 第 62 回大会、 関西大学,2010

・星野崇宏「欠測データと因果効果の推定」、ICPSR 国内利用協議会統計セミナー、

立教大学, 2010

【学会賞】

・星野崇宏　日本行動計量学会　出版賞