別表５
⑶

主　論　文　要　旨

No.1

報告番号	甲　乙　第　　号	氏　名	光廣　正基

主 論 文 題 名：

Latent Variable Modeling for Statistical Data Combination Problem

（統計的データ融合問題に対する潜在変数モデリング）

（内容の要旨）

マーケティングや経済学，社会科学などのさまざまな分野で観察データを分析する場合，興味関心のある変数が同時に観測された“（擬似的な）シングルソースデータ”を取得することは難しいが，代わりに個人やユニットが異なるマルチソースデータを取得することが可能である．それぞれのデータセットでのみ観測される変数間の関係を調べるためには，これらのマルチソースデータをシングルソースデータとして統合する必要がある．統計的データ融合は，マルチソースデータをシングルソースデータとして統合する方法であり，経済学でもデータコンビネーションという方法が提案されている．これらは，マルチソースデータをシングルソースデータとみなしたとき，ユニットごとに発生する欠測値を予測する方法でもある．

他にも経済学では，擬似パネルという方法が提案されている．政策の意思決定などに活かすため，個人の経済行動を調査するパネル調査を継続的に実施する場合，パネルの脱落や大きな費用により，時系列のパネル調査の継続は難しい．そこで，擬似パネルによりクロスセクションデータからユニットごとに推論を行う方法は有効とされている．さらに，政府統計のひとつである家計調査では，時代の変化と伴に回答者の偏りなどで実態を捉えにくくなっているため，消費動向指数といった指標の推計改善が必要とされている．このような課題に対して，企業が保有する購買履歴データを使い，共通属性を共変量としてデータ融合を実施し，国民全体の総購買額などの集計値を補正する研究が行われている．

近年のデータ活用の拡大に伴い，あらゆるマルチソースデータを融合するニーズは増えているため，本研究では，既存のデータ融合法の問題点を整理し，それらを解決するための新たな手法を提案する．欠測データのメカニズムに着目し，潜在変数を活用することで，欠損値の予測精度の改善を目的とし，各問題に応じて４つの新たなデータ融合法の開発を行う．各手法の特徴により，これまで既存のデータ融合法を適用することが不向きなマルチソースデータが取得できた場合でも適用でき，より多様なデータ分析が可能となる．第１章では，導入として各分野で共通しているデータ融合問題について述べ，マルチソースデータを統合したときの欠測値の予測を課題とする．

主　論　文　要　旨

第２章では，データ融合のアプローチを大きく３つに分類し，統計的マッチング，潜在変数モデリング，回帰ベース統計モデルの観点から既存手法の特徴を比較した．いずれの手法もランダムな欠測を前提としているため，回答者や調査モードの違いによる選択バイアスを持つようなランダムでない欠測には対応していない．ランダムでない欠測が得られる状況下で，欠測値の予測精度がランダムな欠測ほど高くないことを確認するため，シミュレーション例を紹介し，予測結果を比較した．

第３章では，カーネル正準相関分析と統計的マッチングを組み合わせたデータ融合法を提案した．統計的マッチングは，マーケティングや経済学での実際のデータ分析で使用されることが多く，調べたい変数と関連がある各ユニット間の共変量の距離に関して，片方のデータセットに存在するユニットに，別のもう片方のデータセットに存在する類似したユニットを割り当てる手法である．ペアとなった相手の観測値を欠測値に代入することで欠測値が予測されるが，マルチソースデータの共変量が高次元な場合，共変量の値の組み合わせが複雑になったり，注目している変数間の非線形な関係を扱ったりすることが難しいため，より予測精度の高い結果を取得することは難しい．そこで，共変量を低次元に縮約した後に統計的マッチングを適用する２段階の手法が有効とされており，多変量解析で使用されるカーネル正準相関分析に着目した．提案手法は，共変量とアウトカム変数間の関係性を保持した共通の低次元空間にある正準変量を推定し，その正準変量に対して統計的マッチングを適用する．シミュレーションデータと実データを用いて，カーネル正準相関分析とカーネルマッチングを組み合わせた提案手法とマハラノビスマッチングの予測値の二乗誤差を比較し，提案手法の有用性を示した．

第４章では，誤差最小化による制約付きマッチングによるデータ融合法を提案した．統計的マッチングを適用する際，共変量の分布がマッチング前後で異なる場合があり，特にランダムでない欠測のときは影響を受けやすい．マッチングの前後でモーメント情報が大きく変化しないように，平均値や分散が近づくような制約を追加した多対多マッチングという既存手法もあるが，加えた制約に合わせて最適化するマッチングアルゴリズムが必要となる．そこで，多変量解析の因子分析などで用いられる行列分解で推定されるウエイトを応用することによって，多対１マッチングに共変量の平均値が近づくような制約を追加し，制限したいモーメント情報の数が増えても行列演算により同じ更新式で推定可能とした．適用例では，制約の強さの違いにより，マッチング前後の共変量の平均値の差における変化の様子を確認した．

第５章では，ガウス過程による潜在変数を用いたデータ融合法を提案した．既存の潜在変数モデリングによるデータ融合は，ランダムな欠測データに適用することを前提と

主　論　文　要　旨

し,相関の希薄化が起きやすい統計的マッチングよりも欠損値の予測精度はよいものの,パラメトリックなパラメータ推定が必要であり，モデルの仮定から逸脱すると欠損値の予測が難しい．そこで，近年機械学習で応用されているガウス過程に着目し，ガウス過程による潜在変数を用いた統一的な枠組みで離散変数と連続変数を持つような複雑なデータに対するノンパラメトリックモデルを提案した．シミュレーションでは，提案手法と既存手法であるマハラノビスマッチング，潜在変数モデルと比較し，欠損値の予測精度が既存手法よりも改善した.

第 6 章では，ランダムでない欠測データに対するガウス過程を用いたデータ融合法を提案した．第 5 章で提案したデータ融合法は，ランダムな欠測を前提としているため，ランダムでない欠測も扱えるように拡張した．データ融合が必要とされるマルチソースデータは，回答者の違いや調査方法により，選択バイアスを持ちやすく，偏ったデータに対して既存のデータ融合法を適用した場合，欠測値を予測するには第 2 章で紹介したようにランダムな欠測データに対して適用した場合と比較して予測精度は高くない．そこで，欠測するかどうかの確率に離散選択モデル，今回はプロビットモデルを仮定し，このモデルに対してもガウス過程による潜在変数を用いることで，観測値におけるモデリングと統一的なフレームワークでモデルを構築した．シミュレーションでは，第 5 章で比較した手法に加え，ランダムでない欠測に対応した潜在変数モデルと本章で提案したガウス過程によるデータ融合モデルの予測結果を比較し，提案手法による予測精度の改善が示せた．適用例では，企業財務データと企業ブランド調査データを統合することを目的とし，提案手法と既存手法の欠損値の予測結果を比較した．企業戦略によって情報の開示内容が異なることで，シングルソースデータの入手が難しい企業財務データに着目し，さらに調査対象企業しか得られないブランド調査データを用いて，データ融合を行った．これらのデータからシングルソースデータを作成すれば，これまで欠測が発生することで分析できなかった内容も実施可能となる．適用例の結果では，既存手法やランダムな欠測を全体とした手法と比較し，提案手法による予測値の改善が見られ，ランダムでない欠測に対応させたデータ融合モデルの有用性が示された.

以上を踏まえ，第 7 章では各章の手法の概要を述べ，カーネルの計算コストの改善やモーメント情報による制約の追加など，実務で提案手法を活用するためのさらなる課題も挙げた．本論文で提案した各手法のシミュレーション研究や適用例の結果から得られた示唆は，データ融合の問題解決に貢献し，あらゆるデータ分析の現場において実データへの適用が期待される．経済学においても，冒頭に述べた時系列のパネル調査のデータ分析に役立ち，実態把握のための指標の補正にも応用できると考えられる.