社会調査データの統合データクリーニングシステム開発の研究

――DCSS の開発と試用――

専修大学　　羅一等

1　目的

この研究の目的は，社会調査データのクリーニング作業を一括して行える統合システムを開発することである。システムがカバーする作業範囲は、クリーニング前のデータを読み込み、異常値を検出して修正を施し、作業経過を記録、クリーニング済みのデータを書き出すまでのプロセスである。

2　方法

本システムが具現するのは、Fellegi and Holt（1976）の原則と保田（2011）の粘土細工アプローチ（Clay Modeling Approach）に基づく作業プロセスである。このプロセスでは、異常値を検出する編集段階とそれを修正する補填段階とを区分し、編集段階では論理式で表された編集ルールを作成・適用して異常値を検出、補填段階ではそれをケース単位で修正していくという手順になる。このアプローチの特徴はクリーニング作業がケース単位で行われるという点であり、したがって本システムも、元値の表示と異常値の検出、修正値の検討と入力、作業経過の記録と閲覧などがケース単位で行われるという点が特徴となる。

3　結果

DCSS（Data Cleaning System for the Survey）という社会調査データの統合データクリーニングシステムの JAVA アプリケーションを開発し、それを用いて実際の社会調査データのクリーニング作業を行うことに成功した。この試用では、ケースの数約 6,500、変数の数約 3,000 の大規模のパネルデータを扱った。データクリーニングの作業を一つのアプリケーションでシステム化することで次の利点が得られることを確認した。（１）データクリーニングの手続きを可視化してそれを作業者間で共有でき、作業者間でコミュニケーションがとりやすくなる。（２）GUI による直感的な操作を実現することで、データクリーニングに関する専門知識や経験がなくてもすぐに作業に参加できる。（３）第三者が詳細な作業記録を簡単に閲覧でき、それを評価できる。

4　結論

統合データクリーニングシステムを利用することで、社会調査データのクリーニング作業における手続きの可視化、操作の単純化、作業内容の透明化が実現できた。このシステムを利用すれば、データクリーニングにかかる費用を減らし、質の良いデータを仕上げることが期待される。

文献

Fellegi, I. P. and Holt, D., 1976, “A Systematic Approach to Automatic Edit and Imputation,” *Journal of the American Statistical Association*, 71(353): 17-35.

保田時男, 2011, 「NFRJ-08Panel における調査票の設計−研究課題とクリーニングを視野に−」『家族社会学研究』23(1): 89-95.

謝辞

本研究の遂行にあたり、独立行政法人日本学術振興会の科学研究費助成事業特別推進研究事業「少子高齢化からみる階層構造の変容と格差生成メカニズムに関する総合的研究」（課題番号：25000001）の支援を受けました。