ビッグデーター時代におけるデータベース結合の目的・方法・効果

統計数理研究所　山下智志

１．セッション全体の背景・目的

多種多様なデータベースが存在する社会において、ある問題を解決する統計モデルを作成する際、複数のデータベースを用いることが一般的になってきた。
しかし、レコードの結合（名寄せ）やフィールドの類似性の評価など、モデル推計までにデータを整える作業が、単一データベースを利用するケースと比較して複雑・困難なものとなっている。このセッションではデータ結合（マッチング・リレーション）に関する問題点の整理と、結合するための方法論、それを用いた実証分析を通じて、複数データベースを用いた統計分析のあり方について議論する。

２．データ構造化の中のデータ結合

データ構造化の研究分野には、①欠損値問題（センサリングを含む）、②異常値問題（バイアス補正を含む）、③データ結合などがある。欠損値補間については、経済分野や医療分野で実用的な方法が確立されつつある。異常値問題に関しては、バイアス補正の視点では一定の成果が上がっている（季節調整など）。
データ結合については有力な方法論がなく、例えば名寄せにおいては一定の条件によりレコードごとに結合判定を行い、結合できないデータについては削除される、などの人為的な方法がとられるケースが多い。適切でないデータ結合は、新たなバイアスを生むなどの悪影響により、結合データを用いたモデルの精度を低下させる。

３．データ結合問題の類型

データ結合の問題において論点の整理が曖昧である。

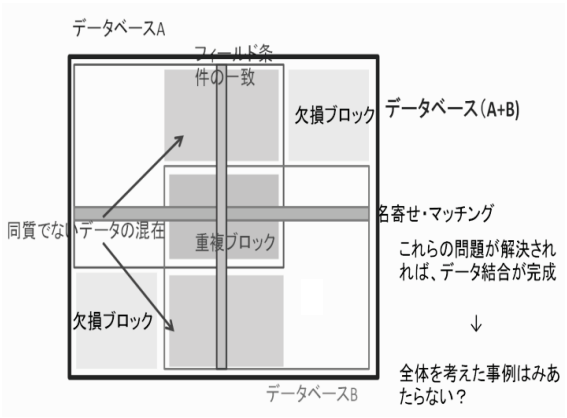

図１．データベース結合における統計的問題の種類

図１に示すように精度の異なったデータベースを結合する場合、Ⅰ名寄せマッチングに関する問題、Ⅱフィールド条件の一致性の問題、Ⅲ重複ブロックの扱い、Ⅳ欠損ブロックの扱い、Ⅴ精度の異なるデータの混在などの問題点があり、それぞれに解決策が必要である。

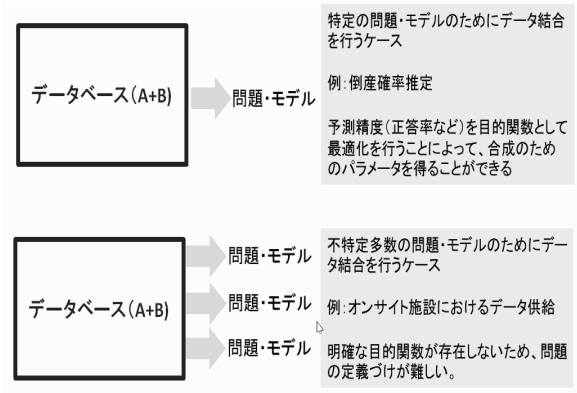

図２．データベース結合の２つのケース

また、結合する目的として、特定の問題・モデルを想定している場合と、汎用性のあるデータベースを構築する場合があり、それぞれ違う方法論が必要とされる。