討論：
統計的データリンケージの可能性と方法

統計数理研究所
椿　広計

討論の視点

- データベース群のリンケージの価値可視化
  - どのような範囲のリンケージに価値があるか
- データリンケージの方法の明確化
- その方法の問題点
- 我々は何を行うべきなのか

異なるデータベース結合の価値
目的＋制御＋環境の網羅化の価値可視化

- **適切なデータ環境**
  - 政策目的となる変数ベクトルの実現値：Y
  - 政策変数ベクトルの目標値：Y_T
  - 政策変数ベクトルの制御変数ベクトル：x
  - 政策環境ベクトル：z
- **適切な機会経済損失評価**：$¥L \fallingdotseq ¥(Y-Y_T)' M(Y-Y_T)$
  - オンライン品質工学の目指したコト
- **適切な統計モデル**：$Y=f(x,z)+\varepsilon$
  - ε：誤差ベクトル：$E[\varepsilon]=0$, $Cov[\varepsilon]=\Sigma$
- **環境zにおける政策制御xの期待機会損失の可視化**
  - $¥Risk=(f(x,z)-Y_T)' M(f(x,z)-Y_T)+ tr \Sigma M$
    - 第1項は政策制御の失敗で大：行政の責任
    - 第2項は、適切なデータベース構築の失敗で大：情報専門家の責任
      - ここでモデルの偏りと推計誤差は当面無視
- 租庸調**報**
  - 第2項に基づく社会の機会損失が大きくならないように、
    市民は情報も税の一形態として社会に提出しなければならない
  - 社会には正しいデータリンケージに基づく情報データベースを構築する責務

リンケージ技法とその範囲

- データベースのリンケージ技法
  - Exact Matchingか統計的マッチングか？
  - マイクロデータか集計データか？
    - 原理的には前者が良いことは明らかなのだが
      - コストの問題
      - 情報を扱う専門家の倫理問題
        - » 集団最適化とモラル人格への配慮とのバランス
          - 個人・法人にとって価値ある、ないしは損失が発生さする可能性のある情報が漏えいするリスクへの配慮
          - 情報はなぜ秘匿すべきかの合理的根拠
- データオーガニゼーション構築の範囲
  - 経済産業政策、社会福祉政策の決定要因と特に環境要因は、何なのか？
  - 官が収集すべき情報、現場が収集すべき情報
    - 社会資産として結合すべき情報は何なのか、その方法は？
      - 誰がやれば、安心か？

今日あるいは今後討論したいこと

- 誰がどのようにやるべきか？
- どのような情報に対しては、どのように対処すべきか？
  - データ収集の方法は？
    - 非開示→統計的処理を施した開示を前提に依頼→一部秘匿を前提に開示依頼→完全開示を依頼
  - データオーガニゼーション構築の方法は？
    - イクザクトマッチングによる完備データベース
    - 統計マッチングなどによる仮想完備データベース
    - 集計地域・業界別マクロ完全データベース

開発と利用

- 官がやるか、官以外がやるか？
- 行政・業務が設計するか、学が設計するか？
- 両者が協力するとして、どのように協力するか？
- 実証を行う、世界水準の人文・社会科学研究者とどのように連携するか？
- データ科学推進のための研究者倫理は確立しているのか？