多変量解析のニューパラダイム

データサイエンスコンソーシアム，慶應義塾大学

柴田　里程

多変量解析

- 多変量データ
  - データ解析を必要とする場合のほとんどが多変量データ
  - 現場の期待は大きい
- 現在の多変量解析
  - がっかり
    - 多変量正規分布を前提とする線形代数の応用
    - 多種多様な手法の集まり
      - クラスタリング，主成分分析，判別分析
      - 回帰分析，分散分析，一般化線形モデル（ロジット，プロビット），正準相関分析，因子分析
  - ？？

その結果

- どの方法を使えばいいんですか？（ユーザ）
  - はっきりした答えが返ってこないので，適当な方法を使ってみるしかない
  - わかったようなわからないような結果でも我慢するしかない
  - あきらめ，がっかり

- 無茶な使い方をしている（学者）
  - じゃあ自分でやってみたら？
  - 反省はありますか．

なぜ

- 歴史的な理由
  - 計算資源の限界
    - 限られた計算資源のなかでは，線形代数程度の計算で済ますしかない
    - 2階から目薬でもやむを得ない
  - グラフィカルユーザインタフェース(GUI)
    - 高精度なディスプレイが普及してきたのは最近のこと
  - 多変量解析
    - 50年ほど前の T.W. Anderson “Multivariate Analysis”
  - 方法の科学
    - 机上でさまざまな方法を開発し提案する
    - 頭ごなし
    - 直感的な理解から乖離

多変量解析のニューパラダイム

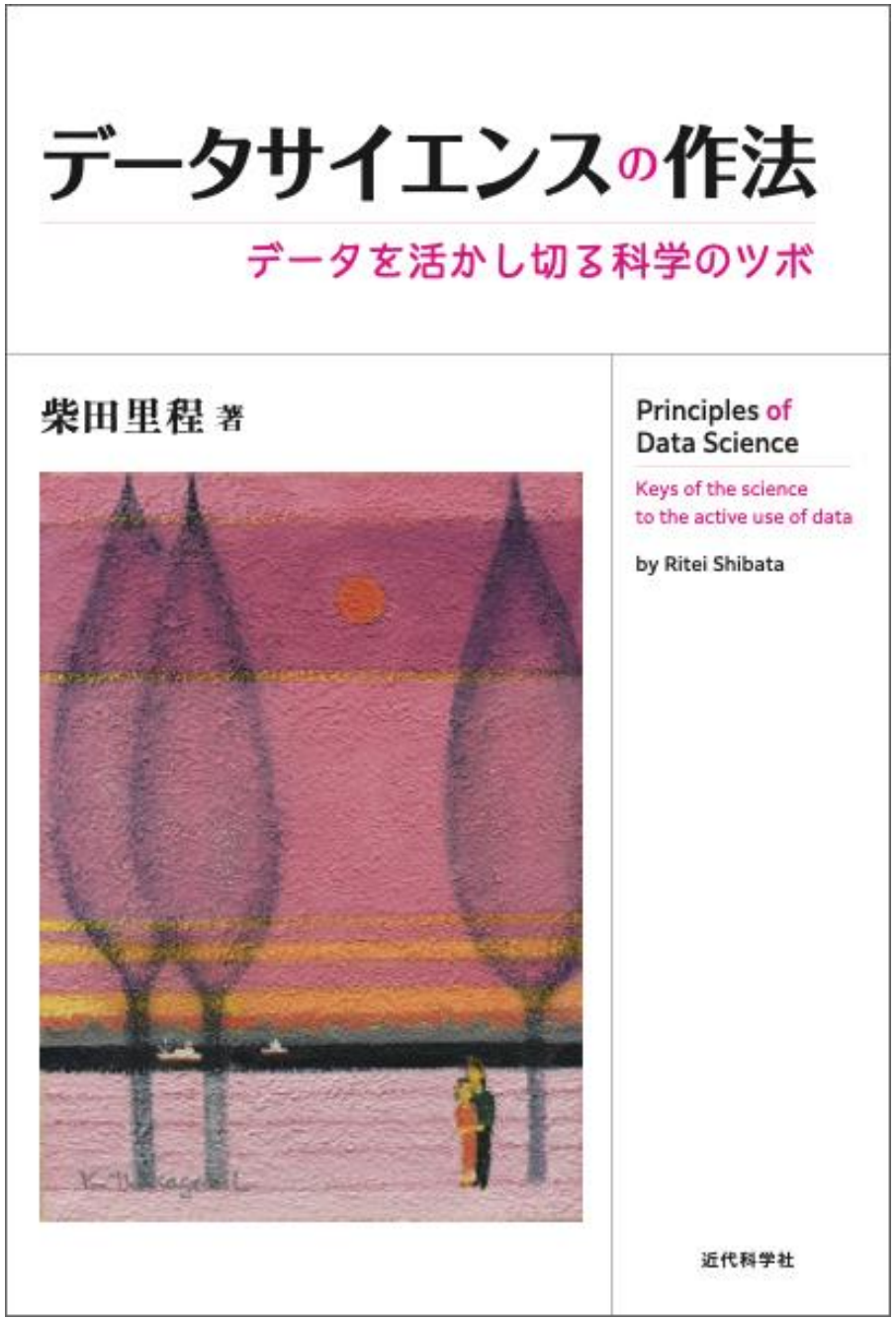
- 目標
  - 多変量データの的確な理解
    - 透明性
    - 納得感
    - 一体感
  - 直感的な確証方法の開発と普及

例

対比による非数値データの的確な扱い

数値データと非数値データの間の垣根をなくす

非数値が被説明変量になったとたん，値ではなく確率で扱う．なぜ？



第 1 章　資料，情報，データ
第 2 章　データの視覚表示
第 3 章　フィルタリング
第 4 章　型
第 5 章　データの読込み
第 6 章　射影
第 7 章　変容
第 8 章　R とその利用

TRAD による多変量解析の実践

https: //datascience.jp/TRAD.html

Windows 版，MacOS 版　無償提供

2020/12 発行

解析結果

- 半田付け実験データ
  - 実験計画の実験ミス
  - 直感的な不良要因
    - マスクに開けた穴の大きさ：小，半田：薄い，マスクの種類と厚さ：B5
  - 回帰診断
    - マスクに開けた穴の大きさが小とそれ以外では残差の様子が全く異なる
    - 記録を2分割して回帰した結果，それ以外の要因は説明力を失う
  - マスクに開けた穴の大きさが「小」のとき，不良が多発する．

- あわびデータ
  - オス，メスペアでの外れ値一件
    - 異なる種が混在？
  - 幼生とそれ以外では様子が異なる
    - 幼生：輪の数が多くの変量と関係
    - それ以外：輪の数に高さだけが大きく関係
  - 幼生でも輪の数がかなり多い個体が存在
    - 幼生かどうか
      - 性分化（生殖能力）
    - 年を経ても性分化が進まない個体が結構ある.

- 枠組み
  - 多変量データの汎用な視覚表示

    -  水平線規準によって各軸の位置と尺度を定めたテキスタイルプロット
      - データの直観的な理解

    -  ママの並行座標プロット
      - データの生の姿を見たい
      - ただし非数値変量の値の位置をどう定めるかの問題が残る

    -  すべての軸の範囲を正規化した並行座標プロット
      - バランスして実験されたかどうかのチェックなどに有効

    -  すべての軸の位置と尺度があらかじめ定められた並行座標プロット
      - 回帰モデルを当てはめた結果のように，位置と尺度がすでに定められている場合

- 多変量データの様相
  - 型
    - 数値型：計測値，計数，序数，日時，時間，記録度数
    - 非数値型：マーク，順マーク，論理
- 多変量データの変容
  - 視点
    - 変量選択
      - 主変量
      - ID
      - 補助変量
  - 視野
    - 記録選択
      - 外れ値
      - 欠損値
      - 射影
  - 視覚
    - 型の変更
    - j変量の順序変更
    - 記録の塗分け
    - 記録のハイライト

さらに，さまざまな研究課題

- 異種データの併合
  - 出元が異なるデータの併合
- データベースとの連系
  - オンラインデータ取得
- データの取引市場
  - 追加説明が不要な形でのデータの取引
  - セキュリティー環境のもとでのデータの利用