データの価値を究める TRAD

（株）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

1　データの価値

さまざまな現場でデータの活用が叫ばれているにも関わらず一向に具体化しない理由の一つは，そのデータがどれだけの価値があるのか，どのような活用の道があるのか，その見極めが難しい点にある．そのような見極めができる人材が乏しい現状では，それを助ける適切な環境を提供できるかどうかがポイントとなる．本講演では，そのような環境の提供を目指して開発を進めてきた TRAD のこの一年間の発展を報告するとともに，その可用性を評価する．

2　TRAD

TRAD (TextilePlot, R and DandD）は，データやモデルの当てはめ結果を直感的に理解できる視覚環境で，http://datascience.jp/TRAD.html から無料でダウンロードし利用できるソフトウエアである．Java で書かれており，JRE(Java Runtime Environment) さえインストール（無料）されていれば， Windows 版，MacOS 版を選びダウンロードするだけで簡単に利用できる．

2.1　Visual Excel

データを扱う上でもっとも身近なソフトウエアはエクセルに代表される表計算ソフトに違いない．しかし，未知のデータを総体的に理解し活用するにはバリアーが高すぎる．何千列，何十万行に及ぶテーブルを眺めていても何のアイディアも浮かばず，途方に暮れるのが落ちである．これは表計算ソフトに限らず R のような高度な解析アルゴリズムを備えたソフトウエアでも同じである．よほど熟達した人でなければデータの海をさまようことになる．TRAD の TextilePlot や Parallel Coordinate Plot はこれらの問題を解決する．データを数字や文字の羅列として眺めるのではなく，平面上に並行に並べられた軸上に値を置き，それらを結んだ折れ線 (Weft) で各記録を表すことで，データの全体的な姿にかぎらず欠損値など詳細にわたることを楽しく調べ廻ることができるのが利点である．しかも最近のチューニングによってエクセルや R と同等の処理速度になった．例えば 30 列 30 万行のテーブルでも 27 秒ほどで視覚表示される．さらに R にデータを移すのも数秒である．エクセルや R でこのテーブルを読み込むだけでも，ほとんど同じ時間必要なことを考えれば，新しいデータの様子を調べるには TRAD といってもよいだろう．

2.2　機能

TRAD の開発過程で，TextilePlot や Parallel Coordinate Plot による視覚表示の各部をクリックすることで様々な情報を得られたり，Weft のハイライトや色付けができるだけでなく，次のような機能も必要なことが判明した．

2.2.1　データの変容

データベクトルの追加・削除・ 順序変更，記録の削除，型変更，水準の併合，正規化，複数の論理ベクトルからのカテゴリカルデータベクトルの生成，基数系から通算秒データベクトルの生成，データテーブルの分割．

2.2.2　属性の編集

名前・短縮名・水準名の編集，データベクトルやデータテーブルの説明の編集．