データ解析環境の今後の発展

慶應義塾大学理工学部数理科学科

柴田 里程

1976 年頃からベル研究所で開発の始まった S は Splus という形で商用化され普及し, 一方でパブリックドメインソフトウエアという形での R もその普及のスピードを増してきている．本講演では，まず商用版の Splus とパブリックドメイン版ともいうべき R を対比する．その上で，Splus や R がいずれも抱える問題点の解消に向けて現在研究開発中の「データの高度利用を促進するための統合環境 DandD 」を紹介する．

１．Splus と R

いずれもおなじ S をベースにしている以上，ユーザから見た使い勝手に本質的な違いはないが，あえてあげるとすれば次のようになる．

i) Splus は Windows ユーザを意識して，Click and Go 的なカジュアルな使い方もできるように機能拡張しているのに対し，R はプラットフォームになるべく依存しない形での開発を基本としているので，コマンドベースである．

ii) Splus は S の機能をそのまま継承しているため, 組み込まれた関数の信頼性は高いが，R は 独自に作り直しているため，その過程で生まれたバグらしきものがあったり，同じ関数でも引数の定義が微妙に違うことがあるなど，どこまで信頼してよいか不明の点がある．

iii) R はその性格上， Linux の開発のようにネットワークを介して，多くのユーザの協力で次々と新しい機能や関数群が盛り込まれるため，ソフトウエア開発という観点からは，きわめて興味深く，面白いが，データ解析の安定した道具として使うには不安な面が多い．一方，Splus は商用版であるため，このような問題は少ないが, その対価として自由にならない部分も多い. つまり小回りはきかない. とくに，保守開発主体が米国なので，ユーザの意見がどれだけ即座に反映されるかが今後の課題であろう．

iv) R の最大の不便は，実行中に生成されたオブジェクトはメモリー上に保存されるので，エラー終了したときや，強制終了したとき，そのセッションで生成されたオブジェクトはすべて失われる点にある．これは，実行速度の向上には寄与しているが，本格的なデータ解析を行っているときは，本当に泣けてくる．もちろん途中でセーブを頻繁に行えばよいのではあるが.

v) Splus にも R にも共通する点として，大規模なデータや計算には向かない．これは，ユーザインターフェイスを重視し，統一性を重んじた S の設計思想がい

ずれにも受け継がれているからである．

vi) もちろん，ユーザからみた最大の違いは，コストであり，簡便さである．R ならば, いつでもどこでもインターネット経由で自由にダウンロードし, すぐ使える．これが魅力である．もちろん商用版のよさは，サポートと責任の主体がはっきりしている点であろうが, それにしても, 無料と何十万円の違いは大きい．とくに，大勢の学生，生徒に教育しなければならない教育機関では，その負担の大きさから，R で我慢せざるをえないケースも数多く見受けられる．もちろん，学生にとって自分の PC にすぐインストールできる R の魅力はおおきい．また，データ解析を専門としない研究者や実務家にとって，Splus の敷居の高さは無視できない．

このように見てくると，Splus も R も一長一短であり，小生が属する学科の３年次学生に対する,「統計科学同演習」,「データ解析同演習」でも，いちどこれまでの Splus 使用から R 使用に変更しようとしたこともあったが, その微妙な違いや予期しない問題の発生を考え，いまだに変更をためらっているのが現状である．もちろん学生には R の存在は知らせているので, 自分の PC には R を入れて使い, 演習では Linux ベースの Splus を使っている．また，研究面でいえば，どのマシンでも自由に使え，小回りのきく R を使うことがどうしても多くなってしまう．

さて，今後のことでいえば，今年でベル研究所のチェンバースのグループも解散してしまい，Omega プロジェクトもどうなるかわからない．Splus は商用版といいながらも，その背景にはチェンバースのグループの最先端の研究成果をいつも反映し，その革新性を保ってきたが，今後はどうであろうか．ユーザの要望を取り入れていくだけで，その革新性を保っていくのはなかなか難しいのではないだろうか．一方　R のほうは Ross Ihaka (New Zealand)　を中心として，Foundation を組織してさまざまな企業からの寄付によってその開発を進めていこうという方針のようであるが，昨年来日したときの様子では，ソフトウエアとしての面白さに興味の中心があり，データ解析の道具としての革新性にはあまり興味はないように見受けられた．　R の Supporting Members にはチェンバースも名前を連ねているものの, R の開発思想は S の開発思想とはかなりずれてきているように思えてならない．つまり，探索的データ解析を存分に展開するための環境を構築したいという S 開発当初の目的からはかなりずれ，さまざまな仕掛けを作ったり，自分の手法を広めるためのプラットフォームとしての R という性格がますます強まってきているように見受けられる．　このことは, R の Web ページ（http://www.r-project.org/）のタイトル”The R project for Statistical Computing” からも伺える．

ソフトウエアの寿命ということでいえば，1984 年の 「データ解析言語 S」の公表から数えてもすでに 20 年たっており，Splus にしても R にしてもその基本は変わっていない．チェンバースたちが，S を開発はじめた契機は，それまでの統計ソフトウエアが，単なる

手法の寄せ集めで，自由な探索的データ解析を展開するにはどれも不満足であるという認識をもったところにある．そのため，それまでのソフトウエアとはまったく違う発想で一からつくり直したわけであるが，そのエネルギーには感服せざるをえない，小生が S に魅力を感じ, その普及に多少なりとも努力してきたのもその新しさにあった. しかし, 20 年以上たったいま, Splus や R にも同じような不満がないとはいえないのではないだろうか. S にさまざまな機能を追加していった結果，Splus も R も総体としてはなんとも見通しの悪いソフトウエアになってしまった気がする．残念ながら世界中どこを見回しても，現在，ベル研に相当する余裕のある研究所はみあたらないので，実際にはなかなか困難ではあるが次世代の魅力的なデータ解析環境を根本的に創り始める時期にあることは間違いない．その際，当然「ネットワークの存在」と「大規模データの解析」を基本にすえた設計となるに違いない．

2. データの高度利用を促進するための統合環境 DandD

S ファミリーに代わるデータ解析環境を一から作ろうというような大それた企てではないが，ソフトウエアオリエンテッドよりもデータオリエンテッドな環境を構築しようという試みが，ここ 10 年以上にわたって実施してきた DandD (Data and Description) プロジェクトである．教科書的なデータではなく，また定型的な手法を適用するだけではすまないよう場合にも，能率よくデータ解析を行うためには，どうしてもデータ本体だけではなくその属性や背景情報の十分な記述が欠かせない．そのような情報が他人にも間違いなく伝わるためには，ある程度の抽象化と形式化が必要である．しかも，データと一体化していなければ，効率化もかなわない．

このプロジェクトの過程で，さまざまな研究の結果，大雑把にいえば

- 記述は XML ドキュメントとするが，データ自体は別のファイルなり，データベースなり Web ページなどに存在していてもよい
- DandD サーバは XML ドキュメントを解釈し，クライエントの要求に応じて必要なデータを必要な形式に整えて送り返す

を基本とすることにし，必要なソフトウエアを開発し，現在 DandD III (http://www.stat.math.keio.ac.jp/DandD/) として公開している．データ自体はネットワーク上のどこにどんな形で分散して存在していても，記述ドキュメントである DandD インスタンスさえ入手すれば，どのようなデータ解析ソフトウエアからでも DandD サーバと通信する部分だけ付加するだけで，データの統合や組織化に関する詳細をを気にすることなく，すぐデータ解析にとりかかれる．すでに R に対するアドオンは開発済みであり，ごく小さな C プログラムといくつかの R 関数からなる．また，将来的には，クライエントからのリクエストの種類を増やすことにより，基本的なデータ操作はサーバに任せれば，大規模データであっても，それをすべて自分のところにもってくる必要もなく，下計算も能力の高いサーバに任せられる．こんな夢に向かって研究開発中である．