

〔研究ノート〕

統計的データ解析のための動的グラフィクス

衛藤 俊寿*, 清田 雄二*², 平木 功明*³

*日本文理大学経営経済部経営経済学科

*²日本文理大学工学部情報メディア学科*³日本文理大学医療専門学校臨床工学科

Dynamic Graphics for Statistical Data Analysis

Toshihisa ETO*, Yuji KIYOTA*², Noriaki HIRAKI*³

*Department of Business and Economics, School of Business and Economics, Nippon Bunri University

*²Department of Media Technologies, School of Engineering, Nippon Bunri University*³Department of Clinical Engineering, Nippon Bunri University Medical College

1. はじめに

コンピューター科学の目ざましい進歩とともに、近年の統計学は大きく発展し、また「計算機統計学 (Computational Statistics)」と呼ばれる新しい学問領域を生み出すほどに変貌を遂げてきた。この領域では、統計学とコンピューターの結びつきがとくに重要であるが、両者の架け橋の役目を果たしているのが統計ソフトウェアである。統計ソフトウェアの備えるべき性質や機能及びその構成要素に対する認識は、その時代のコンピューター科学の動向によって大きく変化するため、統計ソフトウェアの指向、ひいては統計学の進展の方向もコンピューター科学の動向に大きく左右されることとなった。

利用者が諸種の統計的方法を駆使してデータ解析を容易に実行できるように、扱いやすいGUI (Graphical User Interface) 環境が提供される一方で、ディスプレイ上で直接操作とそれに呼応した瞬時のグラフィカル表示の変化・フィードバックを積極的にデータ解析に利用し、それを動的なグラフィカル解析法として位置づけようとする研究が進められている (Cleveland & McGill,

1988^[1]を参照)。「動的要素」を加味したこの表現法は一般に「動的グラフィクス」と呼称されている。動的グラフィクスを実現するための統計ソフトウェアとしていくつかのシステムが提案されている (FisherKeller et al. 1974^[2]; McDonald, 1982^[3], Donoho et al. 1985^[4]; Stuetzle, 1987^[5]; Becker et al. 1988^[6]; Held et al. 1989^[7]; SAS Institute, 1990^[8]; Best & Morganstein, 1991^[9]) が、その実現には特殊な資源と表現法が必要とされてきたことから、現状では依然として「特殊なソフトウェア」という感が利用者に強く残っている。しかし、急激なハードウェアの進歩や低価格化とさらにそれを利用するためのソフトウェア環境 (OS: Operating System を含む) の進歩を考え合わせると、動的な特性を利用した統計的接近法は今後の統計ソフトウェアの標準となることが予想される。

2. 動的グラフィクスの機能とソフトウェア

FisherKeller et al. (1974)^[2]の PRIM-9には、射影 (Projection), 回転 (Rotation), 孤立化 (Isolation), 隠蔽 (Masking) の4種の基本操作が組み込まれている (因みに、PRIM-9の“PRIM”の名称は4種の基本操

作の頭文字をとったものである)。PRIM-9に組み込まれた動的グラフィクスの考え方、及び4種の基本操作はその後の統計ソフトウェアの基礎をなしている。最近の動的グラフィクス・ソフトウェアではPRIM-9のもつ機能を包含して、同定、削除、連結、ジッターリング、ブラッシング、回転、アニメーション、スケーリング、パラメータ・コントロールといった機能が付加されている。動的グラフィクスのもつ基本的な機能を表1に示す。なお、機能の詳細については、Becker, Cleveland & Wilks (1987)^[10]を参照されたい。

動的グラフィクスの動的機能を操作するためには、利用者とシステムの間で対話のできるシステム環境が必要とされる。現在の動的グラフィクスにはGUIが採用されており、利用者はマウスやタッチパッドを使うことによってアイコンやメニューを容易に選択・操作できる。

パソコン用の動的グラフィクスに注目すると、多くのソフトウェアがGUIの完成度の高さからApple Computer社のMacintosh上で開発されているようである。例えば、MacSpin (Donoho et al., 1985^[4]), JMP (Held et al., 1989^[7]), Data Desk, StatView II, SuperANOVA, Exstatix, Fastatなどのソフトウェアが提供されており、基本的な動的機能とMacintosh独自の統一された操作法が用意されている (Best & Morganstein, 1991^[9])。

メインフレーム、ワークステーション、パソコンの各プラットフォームで動作する代表的に動的グラフィクス専用のソフトウェアを表2に示す。

表1. 動的グラフィクスの機能

機能	概略
同定	ラベル探知と位置探知の二つの命令系がある。ラベル探知では、散布図上のデータ点のラベルを知ることができる。また、位置探知では、ラベル探知とは逆に探したいラベルを指定することで散布図上の対応するデータ点が強調される。
削除	データ点を直接に操作、例えば、ポインタでデータ点に触れることによって、そのデータ点を削除することができる。その後、グラフは自動的に再縮尺されて表示される。
連結	異なる散布図行列上の対応するデータ点を高輝度を利用して視覚的に結びつけることができる。とくに、McDonald (1982) ^[3] では、高輝度になる連結を用いて散布図のデータ点に対応するすべての散布図上のデータ点を容易に識別する視覚的な方法が提案されている。
ジッターリング	カテゴリカル変数に対して任意の区間 (通常、 $-0.25 \sim 0.25$) の乱数を与えることによってデータ点の重複が緩和される。この機能を利用することでデータの重複による誤解を防ぐことができる。
ブラッシング	ブラッシュと呼ばれる矩形をマウスやタッチパッドで移動させることによって、様々な動的機能が実現される。ブラッシュには、散布図上のデータ点に対する4個の基本操作 (高輝度、背景遮蔽高輝度、削除、ラベル) と3個の描写モード (一時、継続、取消) が付加されている (Becker & Cleveland, 1987 ^[10])。
回転	3次元散布図上で奥行きのある雲状のデータ点を回転させることができる。とくに、3変量構造を探索するとき、あるいは一つの変数の他の2変数への局所的な依存性を探索したいときに有用である。利用者はマウスまたはアイコンによって回転の方向を制御する。2次元平面の画面上で3次元空間を想定するために、遠近法を用いて奥行きを強調することもできる。
アニメーション	あるパラメータに依存して表現される一つの静的グラフィクスが、パラメータの新しい値に対して変化する。
スケーリング	グラフの縦横比 (縦軸の長さ ÷ 横軸の長さ) を変えて表示させることができる。グラフの縦横比を変えることは、視覚的に解読されるグラフの情報量を変化させる。
パラメータ・コントロール	グラフィカル表示に影響する任意のパラメータ (例えば、スケーリングでのグラフの高さと広さ) を速やかに変えることができる。

3. 動的グラフィクスの有用性と課題

ここでは、著者らが試作を試みるシステムを通して、動的グラフィクスの有用性と課題をまずデータ解析過程への適用という側面から考察し、次に動的グラフィクス・ソフトウェアを開発する際の課題について言及する。

動的グラフィクスは様々な長所を有していることが多くの論文・成書を通して紹介されている（例えば、Cleveland & McGill, 1988^[11]）。ここでは、統計的データ解析、とくに回帰解析の実践に応用の効く動的グラフィクス・ソフトウェアを意図して、動的グラフィクスにおける複数のグラフ表示（view）間における多重リンクについて検討する。

試作システムでは、散布図行列（Becker & Cleveland, 1985^[11], 1987^[12]を参照）の対角セルに卓近なデータ省察用グラフィクス、例えばヒストグラム、ボックス・プロット、確率プロッタなどが選択でき、それらのグラフィクスは他の非対角セルの散布図と連動する。したがって、動的機能を併用することによって、散布図行列上で外れ値やデータの分布形状を瞬時に視察することができる。また、散布図行列で表示される変数の一つが応答変数である場合には、その応答変数と他の変数の組み合わせに該当するセルの散布図上に回帰直線や平滑化プロットを任意に表示できる。したがって、一つのグラフィクス上で説明変数間の相関を調べることができると同時に、各説明変数と応答変数の回帰関係や各説明変数に対する応答変数の傾向までも瞬時に視察することがで

きる。回帰診断において、回帰モデルの背後にある仮定を点検したい場合には、規準化残差プロットの視察によって大きな誤差をもつ観測値を調べることができる。それと同時に、同定機能を用いることで個々の残差が回帰直線を示した散布図のどの観測値に対応しているのかを調べることができる。残差プロットの視察によって外れ値が発見された場合には、削除機能を用いてその観測値を簡単に削除することができ、観測値が削除されると瞬時に回帰直線は削除後のデータにあてはめ直される。このような解析過程は従来の統計ソフトウェアを利用した回帰診断に比して非常に短時間で実行できるため、動的グラフィクスを用いることでデータにあてはまりのよい回帰直線を素早く簡単に探索することができる。さらに、複数の view 間の多重リンクの機能の実現例として散布図行列上の多重リンクを作成し、より柔軟な点の認識を可能とする。

動的グラフィクスは統計的データ解析の実践に容易に適用でき、しかも有用であるが、いくつかの問題点が存在する。まず、動的グラフィクスの特徴としては、静的グラフィクスの点の認識面での利用とグラフ表示のターンオーバーの短縮化の2点しか十分に利用できていない。多重リンクの場合も同様である。もちろんこのことによるメリットは単に静的グラフィクスの強化以上の相乗効果を生み出しており、利用者インターフェイスの向上という側面からは成功しているが、グラフィクスを利用する側面からは grand tour (Asimov, 1985^[13]) のように点の移動や画面変化といった動的グラフィクスの特徴を積極的に利用しているとは言えない。

次に、多重リンクの表示は1リンクで認識された点の

表2. 動的グラフィクス専用ソフトウェア

名称	開発者（開発団体）	コンピューター	特徴
PRIM-9	Fisher, M. A., Friedman, J. H. & Tukey, J. M. (Stanford Linear Accelerator Center)	IBM 360/91	・ 3次元表示と回転 ・ 部分集合の取り出しと表示
Plot Windows	Stuetzle, W.	Symbolics 3600	・ ウィンドウ表示と散布図 ・ Lisp コマンド入力可能
ORION I	McDonald, J. (Stanford Linear Accelerator Center)	Orion I	・ 3次元散布図 ・ 射影追跡
MacSpin	Donoho, A. W., Donoho, D. L. & Gasko, M. (D2 Software Inc.)	Macintosh	・ PRIM-9 のパソコン版
JMP	Held, G., Lehman, A. & Sall, J. (SAS Institute Inc.)	Macintosh	・ 3次元表示と回転 ・ 散布図行列 ・ SAS システムの機能を補完

集合のなかで他リンクで認識された点がどのような構造を持つかを点全体の配置のなかでとらえようと試みたものである。しかし、実際には動的な表示のなかでは、このことは必ずしも成功していない。つまり、高輝度された点の和集合としての認識が強調されすぎている。このことは、動的グラフィクス、すなわち時間とともに変化するグラフィクスについて、人がどのようにグラフィクスを認知するかについての検討の重要性を示唆している。動的グラフィクスにおいても、Tufté (1983)^[14]やTukey (1990)^[15]らが静的グラフィクスにおいて行ったような表示と認識に関する議論が必要であろう。

一般に、現在比較的容易に入手可能な動的グラフィクスは、単に静的グラフィクスの強化として捉えられ、データ解析のための真の新しいツールとして確立されず、実地のデータ解析の場でもそれほど適用されていないようである。とくに、日本では統計分野における動的グラフィクスの研究・開発がそれほど活発でなく、日本製統計ソフトウェアに動的グラフィクスと呼べるものは殆どないのが現状である。また、現状では本格的に動的グラフィクスを利用するためには、コンピューターと統計学に関する知識が必要とされ、利用者はコンピューターの構造とソフトウェアに実行させる統計的方法の両方を理解していなければならない。動的グラフィクスの機能は表現方法が高度で複雑になればなるほど、要求される統計学的知識とコンピューターに関する技術も高度になってくるのは事実であるが、本来、統計的データ解析にかかわる利用者は動的グラフィクスの見方やその統計的な解釈・意義に集中すべきであり、解析プロセスの途中でその操作のために思考が妨げられないようにシステムは構成されるべきである。その観点から、本試作システムの教育現場へのツール適用を考え、データ解析や統計教育に利用できるか否かを検証することを計画している。動的グラフィクスは、データを俯瞰して客観的に観察でき、かつ、データの傾向やデータに潜む要因を直観的に把握しやすいため、教育現場においてデータ解析プロセスや統計教育が効率的・効果的に実施できると考える。

4. 結びに代えて

本報告では、現状での動的グラフィクスの機能を整理し、その有用性と問題点を考察した。ここでは、動的グラフィクスをデータ解析に適用することで、従来の静的なグラフィクスでは得ることのできない隠れた情報を容易に描出することができることを示唆すると共に、動的

グラフィクスを実際のデータ解析の現場に適用させる際のいくつかの問題点を提示した。

動的グラフィクスの現状を踏まえると、より利用者志向のインターフェイスを備えた動的グラフィクスの登場が期待される。例えば、Becker et al. (1988)^[6]やCarr & Nicholson (1988)^[16]では、3次元散布図の立体視つまり仮想現実 (Virtual Reality: VR) をデータ解析に積極的に利用しようとする研究も報告されている。また、Nummi et al. (1989)^[17]では、動的グラフィクスをデータ解析用ツールとしてだけでなく、統計教育用ツールとして利用しようとする研究が報告されている。本研究においても教育現場への適用を計画しており、今後は統計教育の研究領域での発展も期待できる。

今後の動的グラフィクスが「健全に」発展していくためには、コンピューター科学、統計的データ解析とくにグラフィカル表現法、認知心理学、統計教育といった諸種の研究領域の動向と問題を十分に把握したうえで、動的グラフィクス手法を開発していくことが必要であろう。また同時に、早急に広範な利用者を獲得できるように、動的グラフィクスに関する基盤を整備することが非常に重要である。このことにより、提案された手法の種々の角度からの検討や動的グラフィクス特有の見方や解釈を普及させることが可能となる。

参考文献

- [1] Cleveland, W.S. & McGill, M.E. (1988). *Dynamic Graphics for Statistics*. Wadsworth.
- [2] Fisherkeller, M.A., Friedman, J.H. & Tukey, J. W. (1974). PRIM9: An interactive multidimensional data display and analysis system. SLAC-PUB-1400, Stanford Linear Acceleration Center, Stanford.
- [3] McDonald, J.A. (1982). *Interactive graphics for data analysis*. Technical Report Orion II, Stanford, CA: Stanford University, Department of Statistics.
- [4] Donoho, A.W., Donoho, D.L. & Gasko, M. (1985). *MacSpin: Graphical Data Analysis Software*. D2 Software.
- [5] Stuetzle, W. (1987). Plot windows. *J. Amer. Statist. Assoc.*, 82, 466-475.
- [6] Becker, R.A., Cleveland, W.S. & Weil, G. (1988). The use of brushing and rotation for data analysis, *Dynamic Graphics for Statistics*, 247-275.

- Wadsworth.
- [7] Held, G., Lehman, A. & Sall, J. (1989). Advance in graphical data analysis form SAS institute. Statistical Software Newsletter.
- [8] SAS Institute, Inc. (1990). SAS/INSIGHT User's Guide, version 6 edition. SAS institute Inc.
- [9] Best, A.M. & Morganstein, D. (1991). Statistical programs designed for the Macintosh : Data Desk, Exstatix, Fastat, JMP, StatView II, and SuperANOVA. Amer. Statist. , 45 (4) , 318-338.
- [10] Becker, R.A., Cleveland, W.S., & Wilks, A.R. (1987). Dynamic Graphics for Data Analysis. Statist. Sci. 2 (4) 355-383.
- [11] Becker, R.A. & Cleveland, W.S. (1985). Brushing a scatterplot matrices : High interaction graphical methods for data analysis, AT&T Bell Laboratories Statistical Research Report, Murray Hill, New Jersey : AT&T Bell Laboratories.
- [12] Becker, R.A. & Cleveland, W.S. (1987). Brushing scatterplots. Technometrics, 29, 127-142.
- [13] Asimov, D. (1985). The grond tour : a tool for viewing multidimensional data. SIAM J.Stat. Compt. , 6 , 128-143.
- [14] Tufte, E.R. (1983). The visual display of quantitative information. Graphics Press.
- [15] Tukey, J.W. (1990). Data-based graphics : visual display in the decades to come. Statistical Science, 5 , 327-339.
- [16] Carr, D.B. & Nicholson, W.L. (1988). Expor 4 : A program for exploring four deimensional data using stereo-ray-graphics, dimensional constraints, rotation and masking. Dynamic Graphics for Statistics, 309-329. Wadsworth.
- [17] Numimi, T, Nurhonen, M. & Putanen, S. (1989). Dynamic illustration of regression diagnostics, Invited Paper of 47th Session of ISI in Paris.

