マーケティング分析における データハンドリングの注意点
―データ解析コンペティションの今昔比較を通して―

田畑　智章

データ解析コンペティションが開催されてからすでに四半世紀が経っている．当時からの模様を振り返りながら，データサイエンスの変遷について，提供されているデータ，使われている統計分析手法，統計分析を行っているプラットフォームといった観点からその特徴を明らかにしたうえで，コンペ参加者（特に学生）の取組み姿勢の違いから現在のデータハンドリングの懸念点を探り，今後に必要なスキルについて考察を行っていく．

キーワード：データ解析コンペティション，データハンドリング，QC 手法

1.　まえがき

今回，本特集としてご担当の先生より，マーケティング分析におけるデータハンドリングの注意点について記事を書いてくれないか，という依頼が来た．日本 OR 学会では半分幽霊のような存在であったので，過分なご依頼に対して引き受けるか迷ったが，データ解析コンペティション（以下「データコンペ」）にほぼ最初の頃から関わっている身として（途中結構中抜けしておりますが…）筆者なりにデータサイエンスについて見直すよいきっかけになるかと考え，思いのままに記していくことにした．乱文ご容赦されたし．

筆者のデータサイエンスとの関わりは，大学院時代に先輩である株式会社 NTT データシステム科学研究所（当時）の中川慶一郎氏にデータコンペへ誘われたことから始まる．データコンペとは，実際の企業からデータを借り，それに対して自由に分析を行い，その視点，手法の適切さ，分析結果の重要性などから評価されるプレゼン大会である．

このデータコンペに中川氏と現中央大学の生田目崇氏とともに「スキャンディーズ」というチームで参加した．当時は，木島正明先生，岡太彬訓先生，守口剛先生，豊田秀樹先生などを中心に，昼は鋭く（こっぴどく），夜は楽しく（実は呑みがメイン）とどこかのんびりとした会であったようにも思える．

さて，データコンペの HP [1] で確認すると，どうやら平成 6 年（1994 年）からのスタートとなっており，筆者はおそらく 3 回目くらいから参加させていただいていたので，データサイエンスとは四半世紀ほど関わっていることになる．この四半世紀の歳月の中で，筆者から見たデータサイエンス（そもそも 25 年前は「データサイエンス」という言葉はなかったが！）の流れを，提供されているデータ，使われている統計分析手法，統計分析を行っているプラットフォームといった観点から振り返りながら，コンペ参加者（特に学生）の取組み姿勢の違いを明らかにすることで，現在のデータハンドリングに対する懸念点を探り，今後に必要なスキルについて考察を行っていく．

2.　データコンペ草創期（1990 年代）

2.1　草創期の「データ」

まず，データコンペ草創期に提供されていた「データ」にはどのような特徴があったのか見ていきたい．マーケティングが主眼であるゆえ，依然としてアンケートデータが見られたが，社会的にデータベースが安価かつ操作が簡単化してきたため，スキャンデータを採りためたパネルデータの採用も多く見られた．特に POS データ（ID-POS ではない！）の学術的利用は研究者垂涎のまとであった．

また，モニターとなった消費者がその日何を買ったのか家で記録をしていくホームスキャンデータも題材に採用された．POS では基本的に追うことのできない，異なる店舗を跨いで購買データが採れるという点が特徴であった．しかし，この頃のホームスキャンデータは欠損やエラーが多く見られがちであったため，分析者は POS データの方を好む傾向にあったかと思われ

たばた　ともあき
東海大学経営学部経営学科
〒 108–8619　東京都港区高輪 2–3–23
tabata@tokai-u.jp

る．ただ，今後のニューリテールの時代においては，購買した商品は“どこから”買ったのかに注目されると考えられるので，重要なデータソースの一つになるであろう．

2.2　データの受け渡し

記録媒体がつたないこの時代においては，データの受け渡しは非常にシビアであった．インターネットは回線が弱く（よくてISDN），通信での受け渡しは困難である．

フロッピーディスク (FD) は1.4 MBほどしかなく，アンケートデータであれば何とか収納できたものの，POSデータは絶望的であった．もちろんUSBメモリもない時代なので，ほとんどの場合CD-Rに焼いて受け渡しを行った（DVDも当然ない）．MOという媒体も一時期出たが，なぜかあまりはやらなかった．

それに比して現在のデータの受け渡しは，クラウドサービスで共有化することができたりと，素晴らしく便利である．分析に使用するPCを固定化しなくてよいことも優秀である．

2.3　草創期に見られた分析手法

2.3.1　データマイニング

この頃，現在のデータサイエンスに相当する言葉として「データマイニング」が用いられていた．マイニング (mining) は鉱山から採掘する意であり，大量データを鉱山に見立てて宝（情報）を探す思いが込められている．鉱山からの採掘には新しい掘削機械が導入されるように，データ分析においても目新しい（必ずしも「新しい」という意味でなくあまり誰からも使われていなかったというニュアンス）分析手法が多く用いられる傾向にあったと感じる．

1990年代以前においては，「集計」が分析の中心であった．単純集計，クロス集計のほか，マーケティングではRFM分析，ABC分析なども集計であろう．こうして集計されたデータに対して検定（平均値の差の検定，分散分析，独立性の検定など）を行っていくのがオーソドックスなスタイルであったと思われる．

これに対して1990年代は多変量解析が主流になってきた．中でもマーケティング分野においては数量化III類やIV類，多次元尺度構成法 (MDS) などマッピング系か多項ロジットモデルなどの選択問題が人気であったように感じる．このあたりの理由について探ってみたい．

2.3.2　1990年代の分析ツール

いつの時代でもデータ分析はそれを行う「手段」に依存する．すなわち，与えられたデータに対して何らかの分析手法（統計手法）を用いる方針を立てても，実際はそれを何で実行するか，そのプラットフォームを決めなければならない．

この時代には当然PythonもRも一般ユーザには届いておらず（Pythonの最初のリリースは実は1991年！），分析者はもっぱらExcelを使っていた．一応，Rのオリジナルである S言語やSPSSなどはあったものの，お金のない大学院生はExcel一択であった（データコンペ参加チームにおいては，分類器にかけると「お金がありそうなグループ：立教大学，慶應義塾大学，etc.」「お金がなさそうなグループ：早稲田大学，東京理科大学，etc.」が如実に表れていたという…）．

Excelにはピボットテーブルという非常に優秀な機能があるのであるが，これは「集計」に特化したツールであり，これを用いると「前時代的」と見なされる，と多くの学生参加者は勝手ながらの劣等感を抱え込み，使っても次のメインとなる多変量解析の前座としていた感は否めない．

しかし，いざ多変量解析の段階に移ろうとしても，Excelがもっている統計分析ツールにはモデル系は重回帰分析しか入っていない．しかも，変数選択はやってくれず，何より独立変数の数を多く設定できない致命的な欠点があった（それゆえ何の工夫もなくExcelの重回帰を使ってくる分析者には冷ややかな目が注がれた）．それでも流行りの手法を使ってみたいお金のない組の分析者たちはさまざまな工夫を凝らしていく．

2.3.3　データ分析における工夫

工夫の第一歩は使用する変数の徹底的な吟味である．クロス集計や相関分析である程度関連を捉えながらも，意外と直観（これはデータを基にして考えるのではなく事象を基にして考えるという意味）で変数を決めていった．なんともアナログであるが，これは必要なスキルではないかと思うことを後述する．

合成変数を用いて変数をまとめることもあった．今でいうdimension reductionである．Excelにはこれまた強力なツールである「ソルバー」があり，これを分散最大化に用いることでExcelでも主成分分析 (PCA) は行うことができた．しかし，当時のマーケティング分野では予測問題をテーマに取り上げることが少なく，要因把握上，軸の意味が解釈しづらいPCAの直接的な活用はそこまで多く用いられなかったように思われる．それゆえPCAは立派な多変量解析の一員であるにもかかわらず，少なくともマーケティング分野の中では日の目を浴びていなかった（理論のわかりやすさに好ましさを感じていた筆者は，近年の注目度向上に

ひそかに感激を覚えている).

続いての工夫がカテゴリ化，フラグ化である．RFM などを用いて優良顧客と非優良顧客に分けることなどが典型的であろう．ダミー変数を用いるだけで，通常の（Excel の分析ツールにある）重回帰分析が数量化 I 類，II 類に化けてくれる（と思っている）．魔法である．

しかし，せっかくの量的データに対してダミー化，カテゴリ化を安易に行うのはもったいない．この工夫の優先順位は上げるべきではないだろう．とはいえ，数量データの分析で傾向があまり見つからない場合，質的データに変換してみるというのはよく行われる手口ではあるので，何をどのようにカテゴリ化，ダミー化すれば元の情報を損なわずに済むのか，考えることは大事なことであろう．

この段階でお金のない分析者は気づくわけである．Excel で PCA ができるのであれば，ユークリッド距離ではない別の距離に変換することで MDS もできるのではないか．カテゴリ化，ダミー化された変数と PCA を組み合わせれば数量化 III 類，IV 類もできるかもしれない．事実これらは十分に Excel で対応可能であるし，結果表示用のグラフは SPSS よりもきれいに作成することができた．

また，PCA で用いた「ソルバー」があれば，重回帰分析ではなく一般化線形モデルに発展できるのではないだろうか，と考えるのは自然な流れであり，マーケティング分野の中でもブランド選択，店舗選択など応用範囲の広いロジットモデルは必然的に採用率が高くなるだろう．ロジットモデルも，尤度関数を構築し，その最大化をソルバーで行えば Excel でも対応可能となる．

このようにして 1990 年代のマーケティング分野では，数量化 III 類や MDS などのマッピング，ロジットモデルなどの選択問題が比較的多く取り上げられていたのではないか，と勝手な妄想を試みた．

2.3.4　さらなる工夫：VBA

この頃の Excel のデータ取り込み容量は 1 シートに約 6 万レコード（256 項目）であったため，それを超えると扱いが大変であった（Excel の 1 シートに入りきらないデータを「大規模データ」と呼んでいた →「ビッグデータ」ではない）．そこまでに至らなくとも，当時のマシンスペックでは 2 万件もレコードがあればワークシート関数による処理は現実的に厳しかった．

そこで，そのような中でも多変量解析を行っていきたいお金がない組の分析者たちは，VBA に目を付けて自分たちで多変量解析ツール作成に挑戦していくようになる．

VBA ではデータを自由に加工できるので，多変量解析だけでなく，片平先生の「マーケティング・サイエンス」[2] や石渡先生の「パソコンによるマーケティングモデル解析」[3, 4] などから移植を行い，SPSS などには存在しないマーケティングモデルをダイレクトに扱うこともできた．

しかし，VBA を教科書どおりに使用すると Excel シートのセルをオブジェクトとして捉えてしまうので，基本的にワークシート関数で処理しているのと変わらなくなる（とてつもなく遅い）．そのため，シート上のデータには直接触れることはせず，それを配列変数にいったん格納し，配列変数に対して演算を行い，演算結果をシートオブジェクトに吐き出す形にすると，驚くほど計算速度が上がった．さらに，VBA がもっているメモリ処理には限界があるので，VBA で取得したデータ（配列変数）を C に渡し，C 上で演算を行い，結果の配列変数を VBA に返すようにすると，今までできなかった大きなデータの処理ができるようになった．

VBA で Function 関数が作成できるように，C で同様な関数群を作成して，それを VBA から呼び出せば活用しやすい．このとき，C で作成した関数群をダイナミックリンクライブラリ (DLL) ファイルとしてまとめておくと便利である．なお，関数ライブラリを作成するにあたっては，「Numerical Recipes in C」[5] や「S と統計モデル」[6] にはずいぶんとお世話になった．が，ここまで行うのは一部のマニアのみであったかもしれない．

なお，Access で VBA を操作すればもっとスムーズに扱えるのではないかと考える人もいるだろうが，VBA 上で Excel のワークシート関数が地味にいい仕事をしたりするのである．

3.　データコンペ中間期（2000 年代）

3.1　中間期の「データ」

2000 年代になってくると，社会的にもポイント付与を目的とした会員登録サービスが活性化し，それに伴いデータにも ID が付くようになった．また，これに顧客データベースを紐づけることも可能になり，「CRM (Customer Relationship Management)」を目指せる土壌ができ上がった．

もともとトランザクションデータは，提供されるとその大きさから参加者は一様に処理の苦悩を抱えていたが，優良顧客など顧客をセグメント化できる ID-POS

では，分析処理の観点からは扱いやすくなった側面もある．

その反面，顧客属性と購買行動がクロスされたデータの場合，考慮する組合せが大幅に増えたことから，分析者には想像/創造力がさらに求められるようになってきた．

3.2 中間期の分析手法

データ分析界隈ではこの頃に革命的なことが起きた．R 言語の登場である．S 言語を発展させた統計に特化した言語というべき特徴で，何よりオープンソースで，どんなマニアックな統計手法でも世界の誰かが作成・アップロードしてくれていて，かつ，無料で使えるというのはお金のない分析者たちにとっては待ち望んだ一品であった．実際は少し前にリリースされているのだが，インターネット上に存在している統計手法のソースが 2000 年代半ばにはかなり充実してきており，誰でも手軽に凝った分析ができるようになった．ただし，対応できるデータサイズはさほど大きくなく，Excel で整えたデータを R で分析という流れができていたかと思われる．

R のおかげで，かつて SPSS などでしかできなかった因子分析が手軽に無料でできるようになったため，こぞって使う分析者が多く見られたように思う．何それと思うような聞いたことのない回転名がこの頃は跋扈していた．

もう一つこれも大きなこととして Amos の登場がある．Amos はパス図を画面上に描きながら共分散構造分析ができるツールで，複雑な共分散構造分析を誰でも簡単に行えるものとした．お金のない分析者たちも，これぱかりはほかに替えのきかないものであるがゆえ，こぞって購入に走った．そのおかげで中間期の分析研究は，良くも悪くも共分散構造分析が多く見られる結果となった．

また，Twitter のサービス開始などで文字データをマーケティングでも活用する動きが出始め，ChaSen のリリースなどで形態素解析が高くないハードルで利用可能となり，テキストマイニングが起こり始めた．これにより，今まで参考的にしか参照されてこなかったアンケートの自由回答項目などを取り上げた研究が見られるようになってきた．

4. 近年のデータコンペ（2010 年代以降）

4.1 ビッグデータ時代の到来

NFC (Near Field Communication) が当たり前に使われる時代になり，世の中的にトランザクションデータは飛躍的に増した．また，インターネットの回線が ISDN から ADSL，さらに光へと進化し，それとともにインターネットサービスが大幅に増え，アクセスログを解析することが始まった．

また，スマートフォンの普及により，個人の購買行動以外の行動を購買行動に絡められるようにもなり，ますますレコメンドなどの One to One なマーケティング視点での研究が多くなっていった．

個人の行動のオープン化については，2005 年に個人情報の保護に関する法律（個人情報保護法）が施行されて以降注意が必要となっていたが，2015 年に改正個人情報保護法が成立し，企業や自治体が抱える「ビッグデータ」と名付けられた膨大なデータをコンペだけでなく，広く社会の中でも利用しやすくなった．

さらに，安価なセンサーの開発で，センシングデータが増えてそれらがオープン化し，こうしたデータを外生変数として，提供されたデータに加えて分析するチームも現れだした．

4.2 2010 年代以降の分析手法

ビッグデータ時代になると古典的な推測統計学は不利な方向に動く．大量のデータで検定を行うと有意差が出やすくなってしまい，その意義が失われてしまうからである．このような流れから，徐々に推測統計学を単体で用いるチームが減り，代わりにベイズ統計学を利用するチームが増えてきた．Stan が開発されたことも，ベイズアプローチの急速な拡大に貢献した．

また，この時代からは機械学習，深層学習などの AI が分析の主役に躍り出る．背景としてはコンピュータの進歩（特に GPU の安価化），また SaaS の出現などが挙げられよう．このあたりの潮流については本誌でも定期的に取り上げられているので，そちらを参照いただきたい．

5. データハンドリングの注意点

以上，四半世紀のデータコンペの模様を振り返って見てきたが，提供されているデータ，使われている統計分析手法，統計分析を行っているプラットフォーム，どの点をもってしても大幅に進歩していることがうかがえた．それなりの分析をするにはお金がかかる時代の人間からすると，現在のデータ分析環境は楽園のように思える．

しかしながら，人間，便利な環境になるとつい大事なことを忘れてしまっていることがある．以下の気づきについては，自分に対する戒めも含めて備忘録として残しておきたい．

5.1　その統計ツールは安心して使えますか？

現在，主たる分析環境は Python か R であろう．両方とも無料で利用でき，かつ，かなりマニアックな統計分析手法であったとしてもパッケージが存在している可能性が高い．また，利用方法も難しくなく，統計手法にしろ AI にしろ基本的には，データをパッケージに入れ結果が出てくる，というステップになっており，文系の学部学生でも扱える．ゆえに，分析者は今行おうとしている分析手法がどのようなロジックとなっているのか，理論はわからずとも結果を出すことが可能となっている．

しかし，中身のロジックを確かめずに分析を行ってよいのであろうか．筆者は 25 年ほど前に，自分で作った逆行列，固有値導出プログラムが正しく値が出ているか，市販の統計ソフトと比較検討したことがあるのだが，意外なほどソフトごとで値がばらばらであった．このような経験から筆者は基本的にはどのようなツールに対しても，もしかしたら正しい値が算出されていないかもしれないという懐疑的なスタンスをとっている．

中身のロジックを確かめずに分析を行う場合，教師つきの問題（予測や分類など）であれば，結果の精度の良さですべてが許容されることはある．しかし，教師なしの問題（クラスタリング，要因把握など）では，出された答えが妥当なのかがわからなくなってくる．

このような場合は，いくつかの統計ツールで同じ問題を解いてみてはいかがだろうか．いわゆるセカンドオピニオンである．どの統計ツールを使っても同じような結果が出てくれば，それは信頼できる結果と見なしてよいであろう．

しかし，学生であればこの機会に種々の手法の理論（アルゴリズム）を一度はしっかりと学習することをお勧めする．特に，確率論の影で忘れられがちな線形代数は，実はデータ分析の根幹に関わっているので今一度復習しておきたい．

5.2　提供されたデータは大事に管理されていますか？

草創期のデータの受け渡しについて記述した際，苦労点と現在の状況の素晴らしさについて語った．しかし，一方で気になる点もある．最近の学生（筆者のゼミ生など）を見ていると，平気でオリジナルデータを上書きしていたりする．当然あってはならないことであるが，1990 年代であれば実は物理的に上書きできないように（結果的に）なっていた．CD-R は基本的に上書きできない仕様であるし，FD も爪を折れば上書きできないようにすることができた．この点に関してはメリットかもしれない．

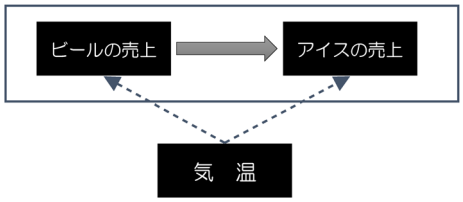

図 1　交絡が提供データになかったら

もちろんクラウド上のオリジナルデータにはアクセス管理がしっかりとなされていて上書きのリスクは少ないとは思われるが，一度それを研究室のローカルサーバに置いたりする場合，上書きされるリスクを想定して厳重に管理する必要がある．これはデータコンペに限らず，データの受け渡しがある共同研究でも同様である．

オリジナルデータだけでなく，作業用のデータであったとしても上書きはするべきではない．ファイル名に日付や時間を加えるなど別名保存を徹底するべきである．

5.3　提供されたデータがすべてだと思っていませんか？

たとえばスーパーマーケットの ID-POS が提供されたとしよう．学生がそれを分析し，ある消費者は 2 ヶ月に 1 回 X ブランドのインスタントコーヒーを購入していることを発見したと伝えてくる．オーソドックスな分析であるが，果たして本当にそうなのだろうか．

自分の行動に照らし合わせて考えてみると，すべての商材を同じ店舗から購入するケースがどれほど稀なのか理解できる．すなわち，データ提供元のスーパーからしか消費者は物を買っているわけではないのであるが，分析者は往々にして与えられたデータを中心に分析をスタートしてしまったりする．

しかし，関連系の分析であれば，たとえば交絡となっているデータが提供データの中になく，そのまま提供データのみで分析をすれば当然妙な結果となるであろう．よくあるケースで言えば，ビールとアイスの売上情報のみ保有し，背後にあるはずの気温を無視して関連性の分析を行うと，「消費者はビールで苦みを感じた口の中に甘いアイスを入れたくなる」という謎解釈が導出される（図 1）．

このような場合は，一度今もっているデータから離れて，変数間の関連を作画してみよう．このとき使えるのが昔の QC 手法 [7–11] である．中でも特性要因図（図 2），因果連鎖図（連関図）（図 3），KJ 法（図 4）などはデータ分析にかなり親和性が高いと思われる．

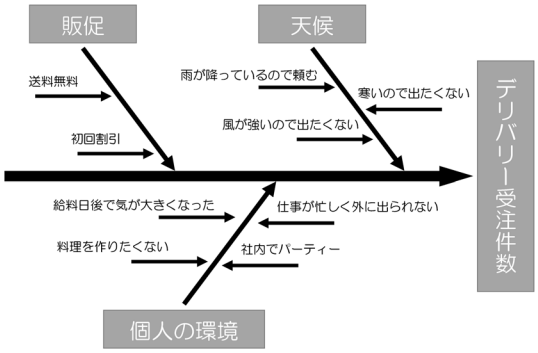

図 2　特性要因図の例

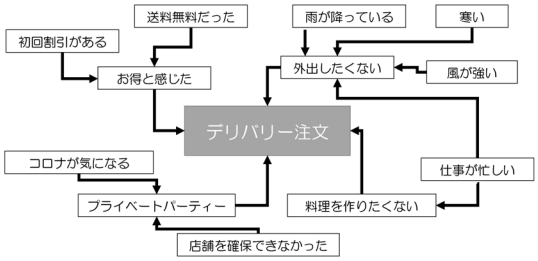

図 3　因果連鎖図（連関図）の例

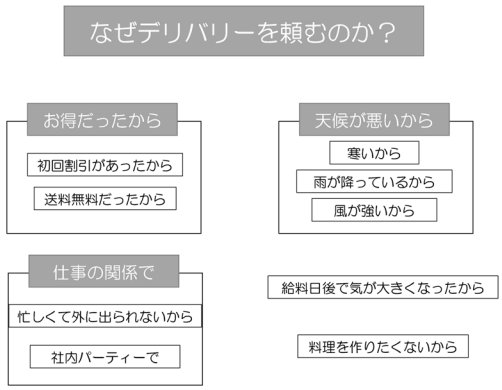

図 4　KJ 法の例

まずは所持しているデータに関係なく，今作ろうと思っているモデルをとりあえずアナログで思い描いてみる．このとき，関連系の分析では特性要因図や因果連鎖図を，クラスタリング系を行いたい場合は KJ 法を用いるとよいだろう．

描き終えた段階で，所持しているデータと見比べ，どのデータが存在していて，どのデータが足りていないのか確認しよう．

その結果，重要なデータが足りていなかった場合はどうするべきだろうか．三つほどの対処法があるであろう．

一つ目は，重要なデータの箇所を抜いた限定的なモデルで分析を行う方法である．ただしこの場合，限定的であることを強調しなければならない．

二つ目は，そのデータを潜在変数として推定してしまう方法である．因子分析や共分散構造分析が使えるかもしれない．

三つ目は，代わりとなるデータを手に入れる方法である．公開されている何らかのデータの中に使えるものがあるかもしれないし，所持しているデータの中でも，構造化されたデータには存在しなくとも，テキストデータやログデータなどの非構造化データの中には存在するかもしれない．自然言語処理の技術と組み合わせて，新たにインデックスを作成していくことも求められる．

以上のように，提供されたデータがすべてだとは思わず，自分は何を分析して明らかにしたいのかについてしっかりと考えたうえでハンドリングにのぞむべきだろう．「データドリブン」という言葉は，とにかくデータからすべてを行え，のように聞こえることもあるので注意したい．

6.　おわりに

「データサイエンス」という言葉に対する定義はさまざまあり，確定もしていないだろうが，医学や経済学，工学，社会科学など各種場面でそれぞれに行われてきたデータハンドリングに対して，方法の共有化をはかるような流れで進歩を遂げてきたことから見ても，「データサイエンス」は「方法」ではなく「方法論」なのだと思われる．方法論であれば，データと方法，状況と方法，方法と方法などの “組み合わせ” を考えることが重要となるはずで，データサイエンティストは極めて抽象度の高いメタな視点をもたないといけないのであろう．

データや分析方法だけに固執するのではなく，誰のために，何を明らかにしなければならないのか，もう一段上に立って，俯瞰して問題の構造を捉まえることが肝要かと思われる．ツールやデータがどれほど進歩しようとも，その部分については変わらないはずである．

参考文献

[1] 経営科学系研究部会連合協議会,「データ解析コンペティション事務局ホームページ」, https://jasmac-j.jimdofree.com（2022 年 6 月 18 日閲覧）
[2] 片平秀貴,『マーケティング・サイエンス』, 東京大学出版会, 1987.
[3] 石渡徳弥,『パソコンによるマーケティングモデル解析 1』, 共立出版, 1990.

[4] 石渡徳弥,『パソコンによるマーケティングモデル解析 2』, 共立出版, 1991.
[5] W. H. Press, W. T. Vetterling, S. A. Teukolsky and B. P. Flannery (丹慶勝市, 佐藤俊郎, 奥村晴彦, 小林誠訳),『Numerical Recipes in C 日本語版』, 技術評論社, 1993.
[6] J. M. チェンバース, T. J. ヘイスティ編 (柴田里程訳),『S と統計モデル―データ科学の新しい波―』, 共立出版, 1994.
[7] 飯田修平,『特性要因図作成の基礎知識と活用事例―事例付き― (シリーズ医療安全確保の考え方と手法) 』, 日本規格協会, 2018.
[8] 石川馨,『品質管理入門〈A〉』(第 3 版), 日科技連出版社, 1989.
[9] 五百井清右衛門, 平野雅章, 黒須誠治,『システム思考とシステム技術』, 白桃書房, 1997.
[10] 川喜田二郎,『発想法 改版―創造性開発のために―』, 中公新書, 2017.
[11] 川喜田二郎,『続・発想法―KJ 法の展開と応用―』, 中公新書, 1970.