統計数理（2004）
第 52 巻 第 1 号 189–195

公開講演会要旨

データに潜む知を引き出すモデルと予測

樋口 知之†

（2003 年 11 月 5 日，統計数理研究所 講堂）

1. はじめに（樋口(2002a, 2002b)）

今，インターネットに代表される IT 革命により，人間がかかわるあらゆるシステムが大きく変容しつつある．例えば物流システムが理解しやすいであろう．生産者と消費者双方が求める情報の即時的交換を可能にする総合情報サービスの登場により，生産者と消費者が直結された“中抜き”流通が普通となりつつある．実は同じことが，一般に普遍的だと思われている，研究の方法(やり方)自体にも起きつつある．データの組織的超大量取得と，そこからの当該分野における有益な情報──知識──の獲得である．もちろんこの一連の作業には，物流システムと同じく機動性と自動化が強く求められる．

バイオインフォマティクスの分野で重要な技術である，DNA マイクロアレイのデータ解析を例にとってみる．そこで重要なのは，さまざまな環境因子のもとでの超多種の遺伝子発現量から，生命システムの基礎的機能に関する，ある場合は病気を誘発する遺伝子ネットワークを自動発見することである．もはや，少数例の患者/健常者の特徴的生化学量を計測することで病因を探るといった，ある意味で過度の情報縮約のプロセスがそこにはない．

インターネットを利用したアンケートを考えてみる．インターネットを多用する集団は国民全体の平均からして明らかに偏り(いわゆるデジタルデバイドに関連した問題)がある．しかしながら，その偏りがあったとしても，サンプル数の多さ，そしてアンケート結果の即時性を考えると，今後アンケートに基づく社会調査は確実にその方向へ向かうであろう．もうそこには，面談(あるいは電話)アンケートによる少数サンプルからの適切な統計処理といった手法概念はない．

今度は話題を宇宙に向けてみる．地球に衝突する可能性のある小惑星の発見のために，毎日全天を高精度でスキャンしているシステムが日本にある．このプロジェクトをささえる基本コンセプトは，全天スキャン全データ解析そして目的物体の即時自動抽出だ．特定の方角，特定の明るさなどに絞った部分探索などしない．これら，ゲノム，社会調査，そして宇宙の分野にいたるまですべてに共通しているのは，全データの計算機記憶領域へのマップとそこからの直接情報抽出作業である．途中での過度の情報縮約のプロセスをはぶく“中抜き”がもはや先端科学領域においては普通である．

超大量のデータから当該分野において有益な情報の自動抽出を行う，言い換えれば新しい知識発見を支援するシステムの，基礎的な理論研究とその計算機への実装が，発見科学の中心的研究テーマである(北川・樋口(2000)，樋口(2002a))．実は，一般に成熟した学問と見なされ

† 統計数理研究所：〒106–8569 東京都港区南麻布 4–6–7

ている感のある “統計科学” が，理論研究の基盤，あるいは出発点となる視点の多くを “発見科学” へ提供してきたといっても過言でない．まさに今統計科学は，“モデル” と “予測” のコンセプトをもとにした枠組みで，知識発見を支援するシステム科学の一つの柱へと大きく変貌しつつある（樋口（2002a, 2002b））．

統計科学においては，現在までの情報（データ）に基づいて “モデルを構成” し（モデリング），将来に対して合理的な判断を行おうとする人間の知的活動を一般化したものを “予測” と定義する．予測の精度を高めるという目標のもとに，モデルの永続的な改良を通じて新たな知見を能動的に獲得していく（赤池・北川 編（1994, 1995），北川・樋口（1998））．つまり，モデルに基づく予測という行為から派生した，データを生成するシステムへの理解の深化機能を “発見支援” に利用するわけである．

このモデルに基づく情報処理の際，「われわれをとりまく現実の世界は複雑であり，そこから得られる情報も基本的には不完全な，そして不十分なものだ」が大前提であることを忘れてはならない（赤池・北川 編（1994, 1995））．現実世界の不確実さのなかで我々が利用するモデルは所詮その近似にすぎない．実際，複雑・大規模な超多変量現象の解析においては，比較的簡単な当該分野における理論モデルによって現象を完全に記述し，現実のデータとの差異は独立な微小観測誤差とみなせることは，ほとんどない．さらに先端科学領域においては，未知の現象はしばしば誤差限界ぎりぎりのところに発現する．このような中で，未知の部分を単純な観測誤差とみなす処理では発見の緒を見出すことはできない．先端領域における発見のためには，その未知の部分に対する我々の期待を積極的にモデルの形で表現し，能動的に情報抽出を行うことが必要である（北川・樋口（2000））．この枠組みは，ベイズ統計，あるいはベイズ的アプローチと総称され，そのさらなる特徴として対象の表現方法が階層的であることも挙げられる．これらの特徴により，異種の情報を自然に組み合わせることが可能なのである．

本講演では発見をもたらすモデルと予測の役割を，筆者が参画した五つの共同研究成果の具体的事例をあげながら以下に紹介した．

2. 具体的事例

2.1 システム解析の視点からの地球磁気圏研究（樋口（1999, 2002a））

移動体通信や GPS ナビゲーションを可能にする人工衛星の保守管理はきわめて重要な社会的課題である．人工衛星の保守の観点から考慮すべき主たる環境因子は，つきつめれば太陽から吹き出る超音速プラズマ流いわゆる太陽風の活動である（赤池・北川 編（1994, 1995），樋口（2002a））．太陽風は地球固有の内部磁場と複雑な相互作用の結果，地球磁気圏と呼ばれる磁場の領域を地球近傍の宇宙空間に形成する．これと同時に太陽風は，地球の電離層と地球磁気圏の境界面をつなぐ巨大な電流系を生み出す．極域地方に特有の自然現象オーロラは，この大規模な電流系が関連した物理現象である．太陽風のエネルギーがオーロラ現象に変換されるに至るまでの物理メカニズムはかなり複雑である．

太陽風活動が地球圏，特に地球周回の人工衛星へ及ぼす影響の予報（宇宙天気予報）には，太陽風と地球磁気圏の複雑かつダイナミックな相互作用を，精緻にかつ柔軟に記述できる定量的モデルの開発が必要である．ここでのモデルとは，太陽風を観測したデータをインプット，低高度人工衛星で（あるいは地上で）観測される物理量をアウトプットとする信号処理的センスで粗視化した理解において，その両者をつなぐブラックボックスの数理モデルである．

モデル開発研究の大きな流れの一つは，現象をある程度理想化した物理モデルにもとづく，シミュレーションを利用した演繹的なアプローチである．気象現象の背後にあるさまざまな物理法則を偏差分方程式系に表現し，それらをスーパーコンピュータで解くいわゆる数日間気象予報 — 日常我々が天気予報と呼ぶもの — がこれにあたる．もう一つの流れは，インプット，

アウトプット両方の観測データの総合(相関)解析によりルールを見つけ，それらを統合することで数理モデルを構成していく帰納的アプローチである．インプット，アウトプットの観測データについては，宇宙天気予報を最初から念頭に置いてデータの集積が設計されておれば申し分ない．しかしそうでなくとも我々の手元には，さまざまな人工衛星によって 20 数年来蓄積されてきた超大量のデータがある．これらのすべての情報を有機的に統合することによって，データベースと非線形関数を組み合わせたような機能性をもつ数理モデルの構築をめざすものもあっていいはずである．

後者のプロジェクトの端緒として，人工衛星によって観測された約 17 年分の磁場データをすべて解析し，データから大規模電流系の構造を自動的に抽出・分類する作業に取り組んだ．情報量規準 AIC (赤池・北川 編(1994, 1995))の採用により 17 年間すべてのデータファイル中の，大規模電流系の自動同定及びタイプ別への分類が行われた．なお一つのデータファイルは，約 25 分の 1 秒値データを含んでいる．地球物理学的見地からはさまざまな新しい知見が得られたことはもちろん，宇宙天気予報に関しても一定の成果が得られている．太陽風は地球の遥か太陽風上流中に常駐している別の人工衛星によって常時モニターされており，本研究で得られた大規模電流系発生パターンにこの太陽風データを統合済みで，両者間の入出力関係が数値的にすでに表現できている．現在このシステムを出発点として，データベースに基づく宇宙天気予報実現のための基礎研究が進行中である．

2.2 測定誤差の特性を考慮したサンプリング間不均一性の補正 (Kamiyama and Higuchi (2004))

鉄道線路の形状は，日々の列車荷重によってわずかずつではあるが変形し，これが蓄積すると列車走行の快適性や安全性を脅かす．そのため，線路形状は専用列車によって様々な側面から定期的に測定されている．測定結果はコンピュータで等距離間隔に離散化され保管される．レールの所定の位置からのずれを「軌道狂い」，その経時変化を「軌道狂い進み」と呼ぶ．軌道狂いは車両の乗り心地・走行安全性を損なうため，ある程度まで悪化したら保守作業によって直す必要がある．本研究の最終的な目的は，過去のデータを用いた将来の軌道狂いの予測である．これにより予防的な効率の良い保守作業が計画できるようになる．また，軌道を取り巻く多様な因子(軌道構造，列車走行速度，等)の影響を個別に定量化できれば，軌道構造を設計するときの目安になる．

線路形状の経時変化を求めるには，常に線路上の同じ位置で離散化されることが望ましい．しかし，離散化の位置を指示するパルス信号が測定車両の車輪の回転に連動しているため，測定走行ごとに測定データに対応する線路上の測定位置が変化する．本研究テーマは，この線路形状データセットが同じ線路区間に対して複数あったとき，離散化(測定)位置のずれを推定して揃えるアルゴリズムの開発である．もし，全てのデータ間の離散間隔が均一であれば，相互相関係数を用いて簡単に測定値のずれを推定できる．しかし列車走行に伴う測定なので，車輪が空転・滑走したとき，離散間隔が局所的に伸縮してしまい，この方法では “ずれ” が推定できない．

我々は，まず既知の知見，経験，情報等を数理モデル化した．このモデル化により位置合わせの問題は，基準データに対応する被補正のデータポイントを探索する問題に置き換わる．データポイントの探索問題を離散最適化問題に定式化すると，その解法は動的計画法により与えることができる．よって，目的関数に含まれる未知のパラメータが決まれば，本問題の解は一意的に求められる．実は，目的関数にベイズ統計の枠組みを適用すると，最適解が MAP 推定値(Maximum A Posteriori estimate，事後分布を最大化する推定値)と解釈できることを利用し，パラメータが尤度で評価できる．上記の時系列モデルを一般状態空間表現に変形し，非線

形フィルタリング演算によって尤度を計算した(赤池・北川 編 (1994, 1995)).

最尤パラメータ時の最適解をさまざまな観点から吟味した結果，本手法により現実的な位置合わせが可能であることが確認できた．また，これらのデータが測定されたときの列車速度(単位時間あたりに発生した車輪の回転パルスの数より計算されたもの．通常は参照できない)と比較すると，サンプリング間隔を補正した近辺で速度が不自然に低下しており，車輪が本当に滑走したと思われる．つまり提案する手法により，普通観測不可能な列車の滑走・空転が推定できる可能性が示された.

2.3 観測されない非価格プロモーション実施の有無の POS データからの統計的推測法(Sato et al. (2004))

近年，マーケティング分野においては，POS (Point of Sales)データに代表されるような超大量データから，組織的情報抽出法の実践的研究が盛んになりつつある．マーケティング分野は，統計科学と方法論・思想的に重複の多いデータマイニングの，その主たる重要なターゲットの一つでもある．マーケティング分野は，統計科学の観点からもまだ手つかずの問題が豊富に残されている，魅力的な応用分野である.

スーパーマーケットで販売される各社製品の売上には，価格プロモーションと呼ばれる値引きとそれ以外の非価格プロモーションの実施が強く影響することが知られている．非価格プロモーションとは，例えば陳列棚端のスペースに商品を積み上げるといった，値引き以外の手段により消費者の購買意欲を促すものをさす．通常，価格プロモーション実施状況の情報は POS データから簡単に取得できる．一方，非価格プロモーション実施状況は，小売店舗毎に調査により収集しない限り，データが入手できない．しかし，需要予測の精度の向上や施策の効果を踏まえてマーケティングに関する意思決定を行うためには，非価格プロモーション実施状況の情報は非常に重要である．そこで，POS データより獲得できる販売点数 PI，価格掛け率のみの情報から，過去の時点における非価格プロモーションの実施状況の推定を行うためのモデルを構築し，実際の POS データにより実証分析を試みた．この試みは，POS データには含まれないがマーケティングの観点からすると極めて有益な情報を POS データのみから推定するという，既存の POS データ解析研究領域においては大胆なものである.

分析においては，非価格プロモーションの実施の有無を潜在的な(つまり観測されない)状態として捉えて，それがマルコフ過程に従う切り換えに対応すると仮定した．さらに，そのマルコフ推移確率をシステムモデルとして捉え，非線形フィルタ・平滑化を適用することにより，非価格プロモーション実施の有無の推定を行った．提案するモデルによる判別結果を，費用など諸々の理由で現実には取得することは難しいがたまたま取得できた事実(特別陳列実施の有無)と比較することで，さまざまな模擬データへの十分な応用結果とあわせて，過去の時点における非価格プロモーションの実施状況を精度高く推定できることが分かった.

推定した情報をどう実務レベル(現場)で生かしていくか，新しい方法論の提案を含めてその具体的手続きを現在検討中である．具体的には，特別陳列実施の効率性の検証，あるいは動的な視点でのブランド診断と価格プロモーションの評価等を試みている.

2.4 マイクロマーケティング：居酒屋店日次売上高の予測 (山口 他 (2004))

従来の企業におけるマーケティング戦略は，マスマーケティングと呼ばれるものであり，これは大衆を対象としてマーケティング戦略の立案と実行をするものだった．しかし，頓に近年消費者のニーズは多様化し，マスマーケティングの手法では対応できない部分が生じてきた．このことから生産業や小売業でのマーケティングは，POS データのような大量の詳細データを利用して，大量生産大量消費を目指すマスマーケティングから個人に焦点を合わせるマイクロマーケティングへと変化している．マイクロマーケティングは，簡単にいうと地域別に詳細な

センサス・データをはじめ各種データベースから膨大な情報を的確に分析し，最適なマーケットの発見や確認，出店計画や輸送経路，ダイレクトメール戦略などに役立てていこうとする動きのことである．

1990 年代に入りバブルが崩壊しそれ以来日本は不況の波に襲われているが，外食産業もその影響を避けることはできず，97 年のピーク時には市場規模 29 兆円だったのが昨年には 25 兆円まで減ってきている．ただ全体的に不景気であっても，商業統計によると外食産業は諸々の小売業と比べて圧倒的に市場規模が大きい．また，外食産業の市場規模が減少しているのは消費者のニーズそのものがなくなってきているのではなくデフレの影響である．最近は特に安くていい物を求める消費者が増え，全体的に低価格化が進んでおり，これだけ市場規模の大きい産業では特に効率的に利益を生んでいくことが重要である．

ところが外食産業ではマイクロマーケティングの活用が乏しいという現状がある．外食産業のマーケティングには，全体のトレンドを知ることとその日その日の売上を予測することが重要である．特に後者は仕入れるものが食材で日持ちしないということから，他の産業よりも的確に予測する必要がある．それを今でも非常に古典的な方法で，それぞれの店舗の店長の勘に頼って行っていることも珍しくない．またチェーン展開している飲食店ではチェーン店全体で画一的な方法を採用し，店舗ごとの特徴は細かくは考慮されていない．

以上のことを踏まえて，外食産業に適したマイクロマーケティングの方法とそれに見合う売上予測システムを，実際のある飲食店の日次データを基に構築する．飲食店の売上は，曜日，祝日，天気，近所での催し物への人出等の様々な要因に左右される．売上時系列データをこれらの各要因成分に分解するモデルを構成し，そのモデルに基づいて将来の売上を精度よく予測することは，仕入れ，人員配置，新規出店計画等，様々なレベルにおける経営戦略に立案上有益であることは疑いようもない．

本研究では，ある大規模催事場隣接レストランの 2 年分の日次時系列データを，

- t_n： トレンド成分．店の長期的売上の動向を表す項．店が以前より流行っているかどうかを過去のトレンド成分と比べることによって把握することができる．
- W_n： 週周期成分．経験的に，売上には 1 週間単位で周期性があると考えられる．それを踏まえてこれを曜日によって，それぞれどの程度売上に影響を及ぼすかを表す項とする．祝日の効果も併せて考慮する．
- R_n： 雨効果成分．天気を晴れ，曇り，雨，大雨，雪と分け，それぞれの売上への影響がどの程度かを示す．
- q_n： イベント効果成分．店の近所で行われている催事の人出が与える売上への影響を表すもので，その人出 (X_n) を変数とする関数を考えてモデル化する．
- e_n： 残差成分．いわゆる誤差と呼ばれる項．

等に分解してそれぞれを個々にモデル化した．分解は容易な作業ではないが，各成分において当該分野における既存の知見，常識，合理的期待などを事前情報(prior information)として積極的に活用し，それらを分解モデルとあわせることにより状態空間表現，カルマンフィルターと平滑化アルゴリズムを用いて予測値を出した(赤池・北川 編 (1994, 1995)，北川・樋口 (1998)，樋口(2002b))．

完成したモデルを利用することにより，各個人経営のレストランにも，Excel 等表計算ソフトにその日の売上や天気等のデータを入力するだけで，様々な要因を考慮した次の日の売上を高精度で予測できるような使用環境プログラムを開発中である．そうして得られた店固有の情報は，仕入れや販促の計画立案に役立てられる．つまり個人ベースで使用可能なマイクロマーケティング実現支援ソフトウェアを作成中である．また，Web 上でこの作業が行えるシステム

も構築する．これは，特にチェーン展開している飲食店に有用なシステムとなるはずである．これにより，店舗を統括している組織はそれぞれの店舗のデータのみならず，次の日の売上予測値までを瞬時に把握することができる．またトレンドに注目すれば，店舗で働く店長を始め，社員の実力を評価するための大きな要素となるであろう．さらに，Web 上でこれを行うことができれば，店舗ごとの特徴の比較も容易に行うことができる．それらを事細かに分析すれば新規出店を計画するに当たって大いに役立つと考えられる．

2.5 遺伝子発現データと生物学的知識からの遺伝子ネットワークの推定（Imoto et al. (2004)）

前述したように，DNA アレイデータとは，さまざまな環境因子・条件のもとでの超多種の遺伝子発現量を同時計測し得られた超多変量データである．変数の次元は，約 2 万，サンプル数はケースによって異なるが数十から多くて 2〜3 百程度である．現在，DNA アレイデータのさまざまな解析レベルにおいて，多種多様な統計的情報処理を含めたバイオインフォマティックス技術が活躍中である．最近のバイオインフォマティックス研究の焦点の一つは，生命システムの基礎的機能に関する遺伝子ネットワークの自動発見にある．

アレイ遺伝子データから遺伝子の相互関係をモデル化する作業，つまり遺伝子ネットワークのグラフィカルモデル構成において，ベイジアンネットワークは有用な表現方法の一つである．各遺伝子発現量を確率変数として取り扱った時，ベイジアンネットワークの持つ最大の特性である連鎖法則は，超多数の確率変数の同時分布を条件付き確率分布の積へ分解することを可能にする．この分解により，ある一つの子遺伝子が(複数の)親遺伝子にどのように依存しているのかを特徴づける，子遺伝子の条件付分布に我々は注目すればいいことになる．一般にこの依存性は非線形の振舞いを示す．また遺伝子発現データが，実験による様々な人工ノイズの異常値を含むことは有名である．これらの 2 つの問題 — 非線形性と異常値処理 — に適切に対応できる，条件付分布の強健で信頼できる推定法が，ベイジアンネットワークを用いた遺伝子ネットワーク同定には鍵である．

講演時間の制約上詳細は説明できなかったが，我々が提案した手法(Imoto et al. (2004))は，以前の数値実験の中で検討された人工データの分析を通じてノイズに対して強健であることと，実際のデータセットに適用することで知識発見の観点からも有効であることが示されている．

3. 終わりに

これまでの事例に示したように，統計科学は異種・多様な情報の統合をシステマティックに行う基盤を与える．具体的には，異種・多様な情報の結合による隠れた情報の抽出や，あるいは結合させるものの自動探索を可能にする．この結果，統計モデルにもとづく汎化的推論機能を利用した未分野への挑戦が今後いろいろ期待できる．

参 考 文 献

赤池弘次，北川源四郎 編 (1994). 『時系列解析の実際 I』，朝倉書店，東京.
赤池弘次，北川源四郎 編 (1995). 『時系列解析の実際 II』，朝倉書店，東京.
樋口知之 (1999). 大規模データの発見的全探索：大規模地球電流系構造の自動同定，統計数理，47(2)，291–306.
樋口知之 (2002a). 発見の糸口をどうつかむか，日経サイエンス，5 月号，36–43.
樋口知之 (2002b). データに潜む知を引き出すモデルと予測，日本機械学会誌，105, 28–29.

Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S. and Miyano, S. (2004). Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks, *Journal of Bioinformatics and Computational Biology*, **2**(2) (to appear).

Kamiyama, M. and Higuchi, T. (2004). Adjustment of non-uniform sampling locations in spatial datasets with dynamic programming and non-linear filtering, *IEEE Signal Processing Magazine*, Special Issue on Signal Processing for Mining Information, **21**(3), 47–56.

北川源四郎，樋口知之 (1998). **予測とモデル，数理科学，36**(9), 11–18**.**

北川源四郎，樋口知之 (2000). **知識発見と自己組織型の統計モデル，**bit **別冊「発見科学とデータマイニング」，**159–168.

Sato, T., Higuchi, T. and Kitagawa, G. (2004). Statistical inference using stochastic switching models for the discrimination of unobserved display promotion from POS data *Marketing Letters*, **15**(1), 37–60.

山口類，土屋映子，樋口知之 (2004). **状態空間モデルを用いた飲食店売上の要因分解，オペレーションズ・リサーチ，49**(5), 316–324.