響き合う人とデータ

統数研プロジェクト紹介

響き合う人とデータ

統数研プロジェクト紹介

大学共同利用機関法人 情報・システム研究機構
統計数理研究所
The Institute of Statistical Mathematics

もくじ contents

響き合う人とデータ

統数研プロジェクト紹介

		統計数理研究所ニュース 掲載号
第 1 回	日本人の国民性調査	No.128（2015年 5月）
第 2 回	数学協働プログラム	No.129（2015年 8月）
第 3 回	気候変動リスク情報創生プログラム	No.130（2015年11月）
第 4 回	海洋生態学と機械学習の融合による生態系評価指標の開発	No.131（2016年 2月）
第 5 回	神経系まるごとの観測データに基づく神経系の動作特性の解明	No.132（2016年 5月）
第 6 回	データ中心科学による地球環境問題解決を目指す国際ネットワーク構築	No.133（2016年 8月）
第 7 回	公的統計ミクロデータ研究コンソーシアム	No.134（2016年11月）
第 8 回	物質科学への応用に向けた位相的統計理論の構築	No.135（2017年 2月）
第 9 回	地震予測可能性を探る国際共同研究	No.136（2017年 5月）
第10回	広域撮像探査観測のビッグデータ分析による統計計算宇宙物理学	No.137（2017年 8月）
第11回	先端医療を進歩させるビッグデータ解析とデータサイエンス	No.138（2017年11月）
第12回	IR 機能の強化と異分野融合指標の開発	No.139（2018年 2月）
第13回	高度信用リスク統合データベースコンソーシアムと損失モデル	No.140（2018年 5月）
第14回	マルチモーダル生体信号データの時空間解析プロジェクト	No.141（2018年 8月）
第15回	統計思考院におけるデータサイエンス高度人材育成プログラム	No.142（2018年11月）
第16回	公的調査データの統計開示抑制	No.143（2019年 2月）
第17回	多層整数計画に基づくクリンチ／エリミネーションナンバーの計算	No.144（2019年 5月）
第18回	URA による研究力強化	No.145（2019年 8月）
第19回	基礎医学と社会医学をつなぐ離散幾何学的モデリング	No.146（2019年11月）
第20回	ものづくりデータ科学研究センターと共同研究部門	No.147（2020年 2月）

本誌は、統計数理研究所ニュースに掲載された「プロジェクト紹介」をまとめたものです。
肩書き、所属、研究成果などを含む内容は掲載当時のものです。

第1回　「日本人の国民性調査」

60年間の継続調査で蓄積した宝

統計数理研究所が手掛ける数々のプロジェクトを取り上げ、研究者たちの取り組みや、他機関との連携の様子を紹介するシリーズ。第1回は、戦後まもない時期から60年以上にわたり継続している「日本人の国民性調査」について紹介する。

◀ 前列左から中村、吉野、後列左から前田、朴

■「日本人の国民性調査」の3つの目的

国民性の解明

調査結果をとおして、日本人のものの見方や考え方とその変化を明らかにしていく。調査を継続することで、多くの興味深い知見を得ることを目指している。

調査手法の研究開発

実際の調査を行いながら、回収率の低下やインターネットの登場など調査環境の変化を見据え、今後の社会変化にも対応できる新たな統計調査手法を研究開発していく。

統計手法の研究開発

より優れた統計的方法を研究していく。シミュレーションなどの仮想データではなく、現実の調査データを用いることで、実際に役立つ統計的方法の開発を目指す。

計量的調査だからこそ浮き彫りになる実像

統計数理研究所のプロジェクトの中で、最も古くから続いているのが「日本人の国民性調査」だ。統計科学を駆使し、現代を生きる日本人の意識や行動について客観的なデータを集め、実証的に国民性を明らかにする――。そのスタンスこそが統数研ならではのものであり、思想家や評論家などによる日本人論との大きな違いだ。基本に忠実な調査の継続により、マスメディアなどによる世論調査の方法論をリードする役割も担ってきた。「基本に忠実」とは、「成人全体を母集団とし、無作為抽出により選ばれた対象者に訪問面接による聴き取りを行う手法」を指す。

スタートは終戦から8年後の1953年。日本全体が敗戦で自信を失い、考え方や行動の価値基準が揺らいでいるのではないかとの懸念から、実態を把握するために企画された。この第1次全国調査の結果、予想したほど自信喪失してはいないこと、日本人論者が描くような「典型的日本人」などごく少数に過ぎないことなどが明らかになった。計量的な調査だからこそ見えてきた「総体としての日本人」の姿だ。

民意を正しく抽出し、政策に反映させることは民主主義の基本。そのために欠かせないのが、統計的手法による社会調査だ。当初1回限りの予定だった国民性調査は、その重要性が認識され、機関研究として5年ごとに継続することになった。調査の主な目的は3つある。1つは、上に掲げたように、日本人の意識動向を示すデータの獲得。2つ目は、サンプリングや回答法など社会調査の技法を研究開発すること。3つ目が、この調査のデータを素材として実践的な統計解析法を開発することだ。

今日まで半世紀を超えて統数研を含め、国内外の多くの研究者がこの調査に携わってきた。現在の担当者は、調査科学研究センターのメンバー 5人。データ科学研究系・データ設計グループの中村隆教授をリーダーに、吉野諒三教授、前田忠彦准教授、土屋隆裕准教授、構造探索グループの朴堯星助教が携わる。直近では2013年に第13次調査を実施し、2014年10月末に結果を公表した。

時代・年齢・世代の切り口で変化を探る

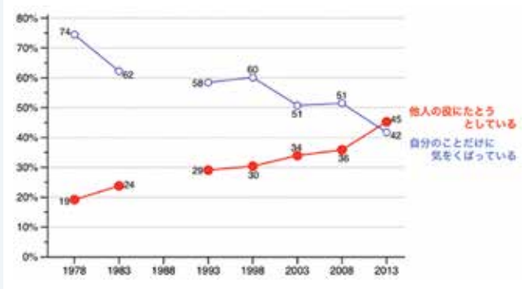
調査開始から60年の間には、いろいろな出来事があった。日本人の意識そのものは、大きく変わった面もあれば、あまり変わっていない面もある。たとえば、「たいていの人は、他人の役に立とうとしているか、それとも、自分のことだけに気をくばっているか」という設問では、1978年には「自分のことだけ」が74%で、「他人の役に」は2割に過ぎなかったのが、年を追うごとに少しずつ後者が増えて前者は減り、2013年にはついに逆転した。

「生まれかわるとしたら、男と女のどちらに生まれてきたいか」では、当初は男女共に「男に生まれかわりたい」が多数派だった。今は、9割の男性が相変わらず「男に」と答えているのに対し、女性は「女に」が上回り、7割を占めるまでになっている。

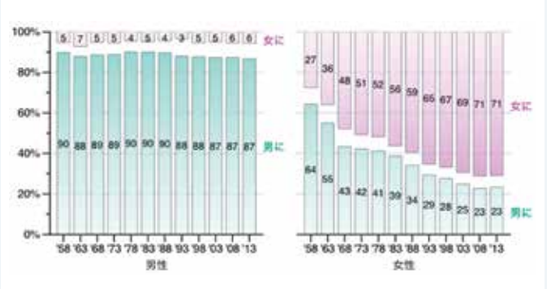
一方、統計解析法の開発は、画期的な進化を遂げた。それまで誰も突破口を見出せなかった難問に、中村の開発したモデルが風穴を開けたのである。

人々の考え方を通して社会の動きを捉えようとするとき、最も重要なことは、その意見の変化が何に起因しているかを見極めることだ。主な要因は、「時代」「年齢」「世代」の3つ。時代とは、全体の意見が同じ方向を向くような「時勢による要因」を指す。時代効果の大きい意見は流動的で、ある時点を境に異なる意見に入れ替わる可能性がある。これに対し、年齢とは時代や世代に関わりなく、加齢やライフステージにより変化していく考え方。年齢効果が大きい意見は、社会全体で見ると長期にわたり変動は少ない。そして世代とは、異なる世代間の相違のこと。ローマ時代の「軍団」を意味する「コウホート」とも呼ばれる。世代差を生むのは、若いときに受けた戦争や大災害、大事件の体験などだ。コウホート効果が大きい意見は、社会の中では世代交代によってゆるやかに変化していく。

ところが、現実にはこの3要因の影響は混ざり合っていることが多く、分離することが難しい。たとえば、若い時の意見を世代間で比較するとき、どうしても時代の違いが持ち込まれる。いわゆる「コウホート分析の識別問題」だ。中村は「これを識別できれば、過去の社会意識の変化の構造が明らかになり、将来の動向予測にも役立ちます」と話す。その手法が1982年に開発した「ベイズ型コウホートモデル」だ。



▲「他人のため」と「自分のことだけ」が逆転
たいていの人は「他人の役に立とうとしているか」あるいは「自分のことだけに気をくばっているか」を質問（#2.12）。「他人の役に」の割合は毎回少しずつ増加し、2013年には「自分のことだけ」の割合を上回った。



▲「生まれかわっても女」が主流に
「もういちど生まれかわるとしたら、男と女のどちらに生まれてきたいか」の質問（#6.2）。男性の答えには変化がなく、女性は以前「男に」が6割を超えていたものの、最近では7割が「女に」生まれかわりたいと答えている。

第1回 「日本人の国民性調査」

実際の調査結果に照らしてみよう。たとえば、「一番大切と思うもの」を自由回答で挙げてもらう質問がある。「家族」と答えた人の割合は、1958年には12%だったものの、2013年には44%に増え、「愛情・精神」「生命・健康・自分」を抑えて1位となっている。「分析の結果、『家族が大切』という回答には、コウホート効果はそれほど大きくなく、時代の影響が一番大きいことがわかりました」と中村。1973年を境に、時代要因によって世代や年齢に関わらず、社会全体にこの意識が浸透してきたという。「そして、その裏には結婚や子供の誕生により変化する加齢の様子（年齢効果）が見られるのです」とも。

日本人をより深く探究するために、1971年からはハワイやブラジルなどの日系人、欧米の人々との国際比較も始まった。特に近年は、吉野がアジア・太平洋諸国との比較を進めている。

調査の集計結果は第1次調査から最新の第13次調査まで、すべて特設サイトで公開している。データだけでなく、調査の目的やポイントの解説など一般向けのわかりやすい記事も加えた。
https://www.ism.ac.jp/kokuminsei/index.html

データの“質”と回収率アップを追究

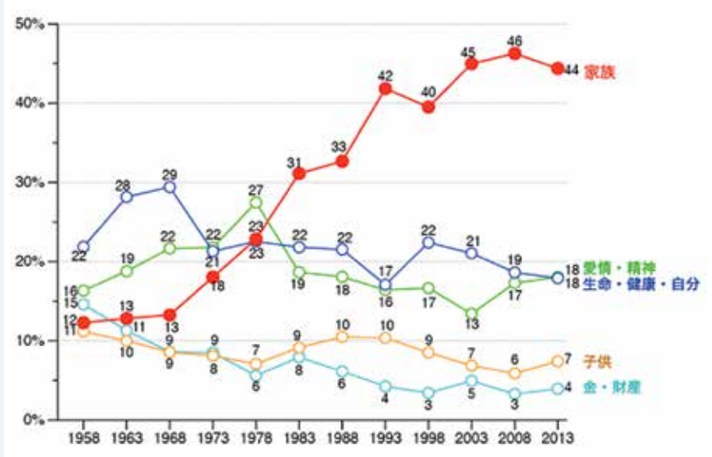
こうした統計分析をする上で、最大の強みとなっているのが、60年間にわたり蓄積されたデータの存在だ。国民全体を対象とする社会調査で、半世紀以上も継続しているものは世界中を探しても他にないという。これを支えてきたのが、統数研の研究者たちの自負と情熱だ。

分析の精度を高めるために、研究者たちはバイアスのかからない“質の高いデータ”を入手することに心を砕いてきた。第1次調査から一貫して、調査員が回答者に直接会って聞き取りをする「面接調査」の方式を踏襲しているのもそのためだ。「たとえば電話調査では『その時間に家にいる人』というふるいがかかるし、FAXやインターネットなら『受信機やパソコンを持っている人』に限定されてしまう。郵送調査にも、本人以外が代筆するリスクがある。正確な調査を実現するのは難しいのです」と吉野は指摘する。

面接の対象者を抽出するのには、「層化多段抽出法」と呼ぶサンプリング方法を用いる。手順は次のとおり。まず全国の市区町村を人口規模によって6つの層に分け、それぞれから300～400程度の町丁字を人口に比例する確率によって選ぶ。さらにそれぞれに割り当てた人数を住民基本台帳から等間隔に抜き出す。こうして選んだ人たちの自宅を調査員が訪ね、面接調査を実行する。「1988年の調査までは、役所へ行って台帳を書き写したり、全国の大学を回って学生調査員を教育したり、すべて自前でやっていました」と前田は明かす。

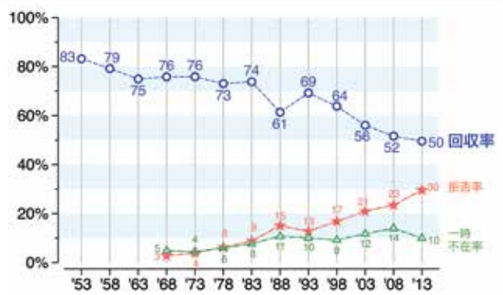
だが、近年になり、こうした方法の前に立ちはだかったのが、プライバシー保護意識の壁だ。年々回収率が低下し、2013年にはついに50%まで落ち込んだ。調査不能の理由のうち6割を占めるのが「拒否」。この状況を何とかしようと、研究者たちはさまざまなアプローチで奮闘している。たとえば土屋は、集まったデータが「調査に協力的な人」に偏る傾向があることから、これを考慮した「調査不能バイアスの補正法」を研究。また朴と土屋は、郵送調査の回収率向上策を探りつつ、郵送調査において本人確認する方法を試行し、検証を進めている。さらには、ホームページに特設サイトを設ける、調査結果を1項目ずつカードに印刷したトランプを作成して学校などに配付するといった広報活動にも力を入れている。

国民性調査の意義について、「その時代の日本人の考えを記録する最も有力な手段」と前田が表現すれば、朴も「調査の蓄積は宝物。これからますます誰もがその重要性に気づくはず」と力を込める。連綿と築き上げた“至宝”を守り通し、次世代へ受け継ぐための挑戦は続く。（広報室）

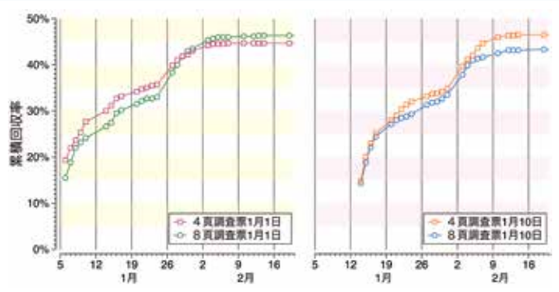


▲時代要因が押し上げる「家族が大切」
「一番大切と思うもの」を自由回答で答えてもらう質問（#2.7）。回を追うごとに「家族」の割合がふえており、コウホート分析の結果、それが「時代要因」に影響を受けた現象であることがわかった。

調査の集計結果をカード1枚につき1項目ずつ印刷したトランプを作成。子どものうちからデータに親しんでもらおうと、学校などに配付している。



1953年に83%だった調査の回収率は年を追うごとにゆるやかに下がり、2013年には50%になった。近年は調査への協力を拒否するケースが増えている。



回収率向上に向けて別の調査を対象とした手法の研究も進めている。このグラフは、郵送する調査票の枚数と回収状況を比較したもの。

第2回　「数学協働プログラム」

社会の問題解決に数学の底力を発揮

統計数理研究所が手掛ける数々のプロジェクトを紹介するシリーズの第2回。統数研が事務局となって、数学と諸科学や産業との連携拠点をネットワーク型で形成する「数学・数理科学と諸科学・産業との協働によるイノベーション創出のための研究促進プログラム（数学協働プログラム）」を取り上げる。

ワークショップの公募・審査・実施
- 重点テーマに基づくワークショップを公募し、運営委員会の審査を経て課題を採択、実施する

スタディグループの公募・審査・実施
- 諸科学・産業界から重点テーマに基づく具体的な課題の提供を受け、短期間で集中討議を行う

諸科学・産業向けチュートリアルの実施
- 諸科学・産業界側のニーズを受けて数学・数理科学のテーマを選定し、チュートリアルセミナーを開催

作業グループの設置・活動
- 数学・数理科学と諸科学の研究者からなる作業グループを設置し、勉強会などを通じて数学によって解決が期待できる課題を抽出

情報の収集と共有・発信
- 諸科学・産業界との協働による成功事例を収集・整理し、ホームページやSNSなどで共有、ワークショップや研究会などの開催情報なども提供

図1：数学協働プログラムの5つの活動。

数学と諸科学・産業との連携でイノベーションを創出

日本ではこの10年で、他分野の諸科学・産業界から数学や数学者へ向けられる眼差しが、一気に熱を帯びてきた。これらの分野でブレークスルーを起こすには、複雑化する対象の中に潜む論理構造を見出すことが必須。その課題解決に欠かせないのが、数学的アプローチだからだ。数学研究者との連携のニーズは、ますます高まっている。一方で数学自体、諸科学・産業と出会うことで、学問として発展することが期待される。

この潮流の端緒は2007年度、独立行政法人科学技術振興機構（JST）による戦略的創造研究推進事業「数学と諸分野の協働によるブレークスルーの探索」領域の設置に遡る。2011年には文部科学省と大学などの共催による「数学・数理科学と諸科学・産業との連携研究ワークショップ」なども始まった。こうした協働による研究推進の機運の高まりを受けて、2012年度から5年計画でスタートしたのが、文科省の委託事業である数学協働プログラムだ。

特徴は、「数学・数理科学によって解決が期待できる課題を発掘し、諸科学・産業と協働して問題解決に当たる」という明確なミッションを掲げている点だ。これを具現化するために、「ワークショップの公募・審査・実施」「スタディグループの公募・審査・実施」「諸科学・産業向けチュートリアルの実施」「作業グループの設置・活動」「情報の収集と共有・発信」などを実施している（図1）。

同プログラムの代表には統数研の樋口知之所長、実施責任者には伊藤聡副所長が就任。外部有識者で構成される運営委員会を設置し、関連学会や大学、諸科学、産業界の意見を運営に反映する体制を築いた。ネットワークは大学共同利用機関である統数研を核に、京都大学数理解析研究所、九州大学マス・フォア・インダストリ研究所、明治大学先端数理科学インスティテュートを含む全国8協力機関からなる（図2）。

伊藤は「これほど多くの大学同士が協力して進める事業は、他分野でもなかなかありません。数学に関する共同利用・共同研究拠点は全国に3拠点ありますが、その京都大、九州大、明治大の附置研究所が集結して協働する意義も大きい。しかも、私立の共同利用・共同研究拠点を含む大学間協力による事業はこれが初めて、かつ唯一のものです」と胸を張る。

数学者と諸科学・産業の研究者が協働してテーマを深掘り

活動の目玉の一つはスタディグループ。少人数の研究者が集まり、産業界から提示された具体的な課題について、数日から1週間かけて集中的に討議し、解決の道筋を探る試みだ。1960年代に英国オックスフォード大学で生まれたこのセッションを日本にも根づかせようと、近年、東京大学などが導入。数学協働プログラムでは、課題を諸科学分野からも募り、取り上げていく方針だ。伊藤は「数学や統計学にとっても、これまで気づいていなかった課題が提示されることを期待しています」と話す。

▲本プログラムの実施責任者を務める伊藤聡副所長。

これに対し、まだ具体的な課題とはなっていない萌芽的なテーマを探す役割はワークショップが担う。数学・数理科学へのニーズの発掘に重点を置き、若手研究者向けの奨励枠も設けた。

スタディグループとワークショップは次の6つの重点テーマについて公募し、採択する。「ビッグデータ、複雑な現象やシステムなどの構造の解明」「疎構造データからの大域構造の推論」「過去の経験的事実、人間の行動などの定式化」「計測・予測・可視化の数理」「リスク管理の数理」「最適化と制御の数理」だ。採択の審査をする運営委員会には前述の8協力機関のほか、日本数学会、日本応用数理学会、日本統計学会、産業界からも委員を招聘している。

一方、統数研が主導して諸科学・産業界へ積極的にアプローチしようというのが作業グループだ。伊藤は「テーマが外から持ち込まれるスタディグループやワークショップを“待ち”の姿勢とすれば、こちらは数学側からの“攻め”と言えます」と力を込める。

広く浅く、狭く深く。外を取り込み、内を拡張する――プログラムの全容は、じつに自由で躍動的だ。

受託機関

統計数理研究所

協力機関

北海道大学電子科学研究所附属社会創造数学研究センター
東北大学大学院理学研究科
東京大学大学院数理科学研究科
明治大学先端数理科学インスティテュート
名古屋大学大学院多元数理科学研究科
京都大学数理解析研究所
広島大学大学院理学研究科
九州大学マス・フォア・インダストリ研究所

図2：受託機関および協力機関。

一大転回点となった報告書「忘れられた科学―数学」

このプログラムをはじめ、数学と諸科学・産業との連携を目的とする取り組みが急速に進んだことには、きっかけがあった。2006年に文部科学省科学技術政策研究所が公表した報告書「忘れられた科学――数学　主要国の数学研究を取り巻く状況及び我が国の科学における数学の必要性」だ。「数学が忘れられた存在に陥っているというタイトルに、数学者たちは皆、大きな衝撃を受けました」と伊藤は振り返る。

同報告書は、日本は数学研究費の規模が主要国に比べて小さすぎること、数学研究者を継ぐ人材が足りないこと、他分野の研究者が数学の必要性を感じているにもかかわらず協力体制が整っていないことを指摘。数学研究の強力

第2回 「数学協働プログラム」

▲公募の電子申請や情報発信などのシステム管理・運用を担当する丸山直昌准教授。

連携先	鉄道総合技術研究所	宇宙航空研究開発機構
テーマ	「車輪・レール間接触問題の数値解析手法について」	「風洞試験模型変形効果データベースに基づく非線形モデリング手法の構築」
概要	列車走行時、車輪とレールの接触面にはすべり現象が起きる。これは車輪とレールの損傷の原因となることから、連携先は以前より車輪・レール間の接触解析手法を開発してきた。しかし、接触問題を展開した高次元の連立方程式を直接解法により解くところで、計算時間がかかり過ぎていた。これに対し数理側は、連立方程式に展開する以前の式をそのまま最適化問題としてシンプルにとらえ直し、前処理を施せば、内点法により高速に解ける見通しを示した。	風洞試験データの模型変形効果を正確に把握・推定するのが目的。連携先は非線形でモデリングすることを考えていた。非線形モデリングは柔軟で魅力的である半面、データ数が少ない場合は推定結果に信用がおけない。行き詰まっていたところ、統数研の教員との対話によって、非線形モデリングに使えるデータが他にも存在することが判明。連携先の研究員は、そのデータが模型変形効果の推定の助けになるとは考えていなかった。データを様々な視点から上手に使うスキルをもっている統計家ならではの長所が生きる結果となった。

図3：スタディグループの事例。

な振興の必要性を説くとともに、政府研究資金の拡充や数学と他分野の連携拠点の構築、数学研究者と産業界の共同研究の具体的な検討などの対策を提言するものだった。文部科学省は、これを受けて数学イノベーションユニットを立ち上げ、従来の枠を超えて数学を活用し、新たな価値を創出する取り組みに着手したのである。以後の急進展ぶりは、冒頭に記したとおりだ。

▲ワークショップやシンポジウムの開催は、手作りのチラシなどで積極的に広報している。

報告された状況の背景には、日本独特の数学の定義の仕方がある。海外では純粋数学と応用数学、数理科学を包括して数学と呼ぶが、日本では数学と言えば純粋数学をイメージする人が多い。丸山直昌准教授は「応用数学や統計学はとっくに諸科学・産業に活用されていましたが、これらは日本では工学と見なされる。そのせいで、実業の役に立つ数学本来の姿が、世の中から見えにくくなっていたのでしょう」と分析する。

将来に向けた課題を担う若手人材の育成にも注力

４年目を迎えた数学協働プログラムは、すでにさまざまな実績を上げている。例えば、スタディグループに関しては、鉄道総合技術研究所や宇宙航空研究開発機構との取り組みなどで、数学者の広く深い知識や、データを多角的に捉えて活用する統計学者のスキルが生きた事例が生まれている（図3）。またワークショップは、昨年度は21件を採択し、それぞれ１年間かけて実施した。

▲数理・生命科学作業グループとスタディグループ担当の藤澤洋徳教授。

▲スタディグループでは少人数の研究者が膝詰めで議論し、短期間で課題解決の方向性を見出す。

▲サイエンスアゴラでは、数学を使って動きを制御した生物模倣型ロボットを実演し、子どもたちにアピール。

た。

作業グループは材料科学と生命科学の２分野で設置し、勉強会の開催などを通じて数学を活用できる課題を抽出してきた。数理・材料科学作業グループのコーディネーターを務める松江要特任助教は「数学と材料科学の結びつきはまだ薄いのが実情ですが、企業の人と話すとニーズはすごく感じます。作業グループで両者の橋渡しをしていきたい」と手応えを示す。数理・生命科学作業グループを担当する藤澤洋徳教授は「生命科学の分野では、古くから数学・数理科学が使われてきた実績がありま す。それを基に、生命科学者の立場から数学を使う未解決の課題を挙げてもらい、将来に向けたオープンプロブレムとして提言書にまとめました」と話す。長いスパンで先を見越した取り組みであることから、作業グループのメンバーには、特に若手の研究者を起用したことも特筆すべき点だ。

▲数理・生命科学作業グループをサポートする事務局専従の風間俊裁特任助教。サイエンスアゴラでのロボット展示・実演も担当した。

▲事務局専従の松江要特任助教。数理・材料科学作業グループやイベント告知チラシ制作などを担当。

次世代を担う人材の育成に向けて、学生をターゲットとしたイベントにも注力している。2014年11月に東京の日本科学未来館で開催したJST主催のイベント「サイエンスアゴラ」では、動物の動きを模したロボットを展示し、子どもたちの人気を集めた。ロボットの制作から現地での説明までを手掛けた風間俊裁特任助教は「生きものの動きを数式に変換してからロボットを作ったのですが、子どもにはそこをしっかり説明しました。数学に興味を持ってもらうきっかけになれば」と微笑む。

伊藤はここまでの事業を踏まえて次のように展望を語る。「協働には多くの数学研究者に携わってもらいたいし、企業側にも数学との連携に興味を持つ人が増えてほしい。そのためにも、このプログラムの成果を目に見える形で残したいと思っています」。数学によるイノベーションの成功は、将来に向けて日本が世界の中で存在感を示す一助となるに違いない。（広報室）

第3回　「気候変動リスク情報創生プログラム」

気候変動予測に確率のプロとして貢献

自然災害に対する防災対策には、リスク評価が欠かせない。これに役立つ情報基盤の確立を目指し、文部科学省は「気候変動リスク情報創生プログラム」を推進している。統数研チームは確率のプロとして参加し、より効率的な「気候変動がもたらす将来予測」に貢献している。

▲上野玄太准教授

気象現象の発生確率を予測し災害のリスク評価につなげる

近年、記録的な猛暑や集中豪雨、竜巻や突風、厳寒や豪雪などの異常気象が頻発し、社会・経済に打撃を与えている。こうした被害を未然に防ぐためには、地球上のどの場所に、いつ、どのような気象現象が現れるかを、定量的に把握することが欠かせない。

自然災害に対し、防災や減災の対策に役立つこうした情報を創出するのが、2012年度から2016年度の5年にわたり文部科学省が主導する「気候変動リスク情報創生プログラム」だ。世界最高水準のスーパーコンピュータ「地球シミュレータ」を活用し、気象現象の発生確率を予測することで、災害のリスク評価につなげる。

同プログラムは現在、A～Dの領域で進められている（図1）。このうち、領域テーマCは、より詳細な予測情報の抽出や、巨大台風の上陸といったごく稀にしか起きない気象現象の想定シナリオ作成を目的とする。これは、自然災害への対策や適応策を講じるためのリスク予測や評価を手掛ける領域テーマDの基となる情報だ。

統数研が担当しているのは、領域テーマCのサブ課題「気候変動予測データの統計学的解析手法の開発」（図2）。メンバーは上野玄太准教授、中野慎也助教、鈴木香寿恵特任研究員、伊庭幸人教授の4人で、いずれもデータ同化研究開発センターに所属している。

リーダーの上野は「Dのリスク評価に活用するデータセットを作ることがミッションなので、利用者側と連携し、ニーズに沿うよう心がけています」と話す。

具体的な課題は、「高頻度事象の確率分布の創出」と「低頻度であるが影響の大きい事象の確率分布創出手法の開発」の2つ。高頻度事象とは平均気温や降水量など、一般に大規模シミュレーションでそのまま統計的予測が可能な事柄を意味する。これに対し、低頻度事象とは「台風が日本本土に来襲する」などそれほど頻繁には起こらないがリスクの高い事象を指す。前者は上野が中心となって伊庭がサポートし、後者は中野、鈴木、伊庭が担当している。

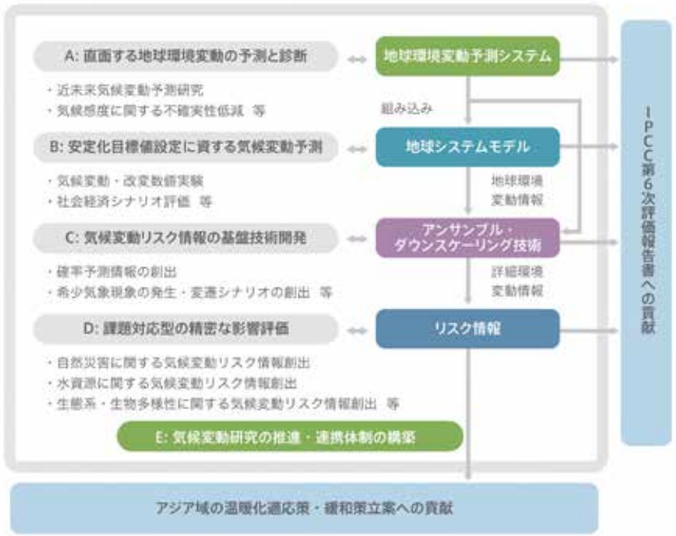


図1：気候変動リスク情報創生プログラムのスキーム。最終的にはアジア域の温暖化適応策やIPCC（気候変動に関する政府間パネル）報告書にも貢献する。

統計学的手法を駆使して効率的なシミュレーションを実施

高頻度事象について手掛けているのは、50年～100年後の気候の予測だ。

▲伊庭幸人教授

例えば、ある場所の気温がこれから上がるのか下がるのか、その年間の分布はどうなるのか、いろいろなシナリオの確率を求めたい。素材として、各国で開発された22種の「全球気候モデル（Global Climate Model,GCM）」で計算した1901年から2099年の200年分のデータが用意された。問題は、いかにこれを重みづけし、統合するか。そこに、確率のプロとしての統数研の知見が求められた。

当初、気象学を専門とする共同研究者は、「これまでの観測値の予測能力をもとにモデルを重みづけしてほしい」と統数研チームに要請。これに対し上野らは、重みをつけたうえでさらに複数のモデルの予測を統合する手法を提案して、先方を驚かせた。22種のモデルからどれとどれを採用するかの組み合わせは膨大な数となる。その中からベストなものを見つける計算時間は途方もないものになると思われたからだ。

だが、もちろん上野らには考えがあった。「まず全体の中から『当てはまりのよいモデル』を見つけ、次にこれらと類似したモデルを除きます。つまり、2段階で全体をよいモデルに集約させ、計算の効率化を図るわけです」。

最初のステップでは、クロスバリデーションと呼ぶ手法を用いてモデルの重みづけをする。データセットから1つのサンプルを取り除いてモデルを構築し、このサンプルに対する予測能力でモデルを評価する方法だ。学校の試験にたとえれば、あえて授業では取り上げていない設問を出題して理解度を測る『実力テスト』のようなものだ。

図3：スパースモデリングで似たモデルを「当てはまりのよいモデル」に集約。重要ではないモデルは自動的に「係数0」と推定される。結果として、22モデルが16モデルに集約された。

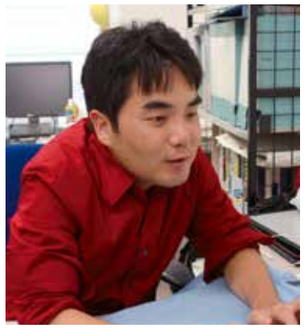
次のステップでは、スパースモデリングというデータ解析法を用いて、予測に有用なモデルを自動的に抽出した（図3）。「統計科学でいま盛んに使われている解析法を気象分野に導入してみたのです」と伊庭。スパースは「疎」の意味。多数のモデルから不要なものを間引いて疎にする

研究領域 テーマC　研究課題名「気候変動リスク情報の基盤技術開発」　領域代表 高薮 出　気象庁気象研究所 環境・応用気象研究部 部長

領域課題		サブ課題	サブ課題代表者
(i)気候変動リスクの評価の基盤となる確率予測情報の創出	a	アンサンブル予測技術と予測実験の最適化手法の開発	防災科学技術研究所 主任研究員　大楽 浩司
	b	気候変動予測データの統計学的解析手法の開発	情報・システム研究機構 統計数理研究所 准教授　上野 玄太
	c	アンサンブルデータの効率的なダウンスケーリング手法の開発	東京大学大気海洋研究所 准教授　芳村 圭
(ii)高度利活用(影響評価研究等)を支える標準的気候シナリオの整備	a	予測情報の信頼性・不確実性の定量化手法の開発	筑波大学 生命環境系 教授　植田 宏昭
	b	高解像度力学的ダウンスケーリングによる低頻度ではあるが影響の大きい気候変動事象に関する情報の創出	気象庁気象研究所 環境・応用気象研究部 部長　高薮 出
	c	雲解像大気・海洋・波浪結合モデルによる台風強度推定	名古屋大学 地球水循環研究センター 教授　坪木 和久

図2：領域テーマCの研究課題。統数研チームは気象分野の研究者たちと並んでサブ課題の1つを担当している。

第3回 「気候変動リスク情報創生プログラム」



▲中野慎也助教

手法だ。「先方は初めのうち、『全部計算したら13万年かかります』と危惧されていましたが、この方法では4分で終わりました」と上野は微笑む。

1901年から1999年までの東京の月平均気温データ1188ヵ月分から、1ヵ月ずつデータを除いた1187個のサンプルで検証した（図4）。

将来発生しうる台風の推定に確率的シミュレーションでアプローチ

低頻度事象のほうは、台風が日本に到来する頻度を見積もることをこのプロジェクトでの目標として設定した。「発生した台風が、例えば東京に影響するかどうかは非常に微妙です」と鈴木は説明する。例を挙げれば、2014年10月に関東に上陸した2つの大型台風は、ほぼ同じ軌道をたどりながら、先発の台風18号は東京23区を直撃し、後発の19号は直前で東京を避けて通過。交通機関などの厳戒態勢は肩すかしを食った。

全球気候モデル（GCM）のシミュレーションでも台風に相当するものは発生する。しかし、その個数は年間に30程度で、100年間でも3000に過ぎない。これでは、サンプル数として少なすぎる。そこで考えた手法が、GCMの結果に簡単なモデルをあてはめて、そこから疑似データを発生させることでデータを“水増し”する手法だ。乱数を導入することで、簡単なモデルでも多様な経路を生成することができる。

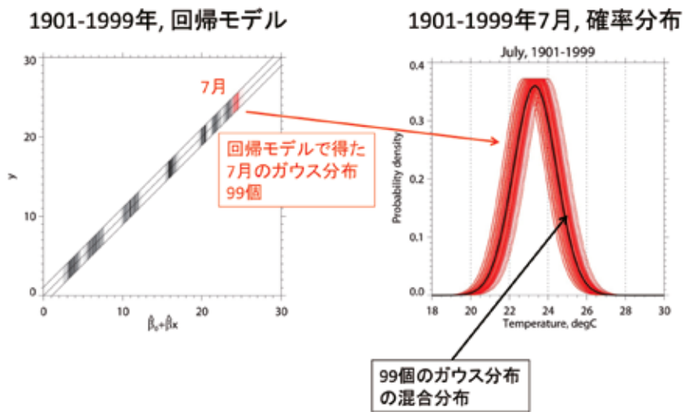


図4：スパース性を仮定した回帰モデルにクロスバリデーションを施して精度を検証する。このグラフは20世紀の7月、東京の気温の確率分布を導いた例。

乱数を用いる台風のモデルは「確率台風モデル」と呼ばれているが、まだ決定版があるわけではない。そこで、統数研チームは2種類の方法で、まず観測値に合うモデル作りから始めている。これができたら、GCMの結果にそれを適用する計画である。

中野は過去60年ほどの気象庁の実測データから台風の移動速度や強度などの確率分布を算出するモデルを構築した。（図5）。「これで、かなり実際に近い台風軌道を表せるようになりました」と中野は話す。

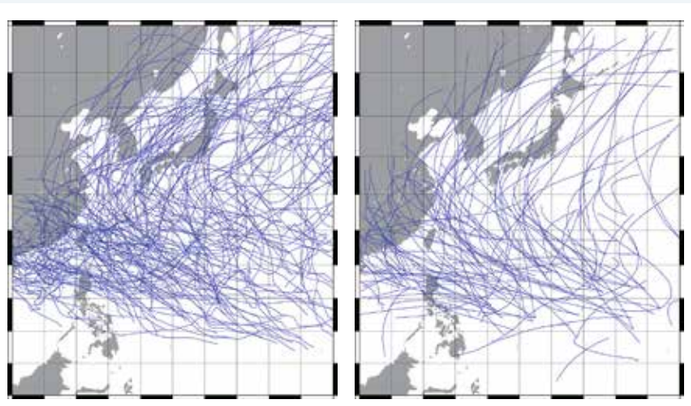


図5：左は1980年から2009年の8月の実際の台風軌道。右はモデルを使ったシミュレーション。実際に近い軌道を表せるようになった。

▲鈴木香寿恵特任研究員

一方、鈴木は台風が周囲の風に流されることに着目し、風速場（風向や風速の変化）から台風の動きを推定する方法を検討している（図6）。「この手法を確立することで、将来の気候状況をより反映した結果が導き出せる可能性があると考えています」と鈴木。

さらに、ユニークな発想で研究を進めているのが伊庭だ。「東京を直撃する台風の経路を予測したいのなら、東京から逆に解いていったらどうか、と考えました」。つまり、ターゲットとなる東京を起点として逆方向のシミュレーションを行う。モデルを単純に逆にすればすむ話ではないが、重みづけなどさまざまな手法を駆使して、正しい確率を求める方法を開発した。伊庭はこれを中野が作成した確率台風モデルに適用して検証を進めている（図7）。

統計科学とのコラボレーションで気象予測の可能性が広がる

気候変動リスク情報創生プログラムは、ほとんどが気象の専門家チームで構成されており、統計分野からは統数研チームが唯一の参加機関だ。上野は「自分たちの知っている方法を提供することで気象分野の人たちの主張の裏付けができ、議論の風通しがよりよくなればと思っています」と話す。

気象研究者との共同研究の計画もある。鈴木は「例えば私たちが東京を直撃する台風の経路を推定できれば、台風のシミュレーションを手掛ける研究者に、風や強度がどうなるか、実験を提案することができます」と期待する。

今後について、伊庭は「個人的にも気象に興味が湧いてきましたので、これからも気象の研究者たちとコミュニケーションを取っていきたいと考えています」と抱負を語る。

気象の予測はインフラの防災対策はもちろん、自治体や産業界のBCP（事業継続計画）にも重要な役割を担う。統計科学や機械学習の手法は、将来的にも気象予測に役立つ側面が多くあるだろう。このプログラムへの参画は、データ同化開発研究センターにとっても貴重な実績を刻むものになる。（広報室）

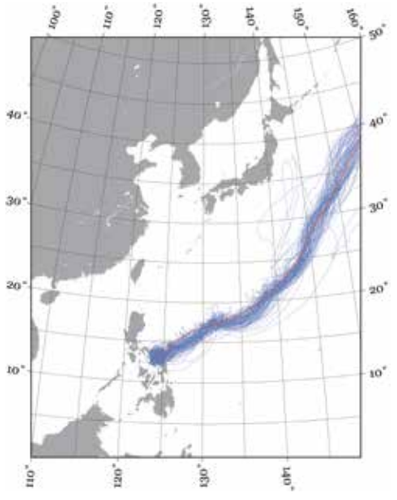


図6：気象庁発行の客観解析データより台風経路（青）を推定した。ベストトラック（赤、過去の実測データ）と比較しても再現性がよい。

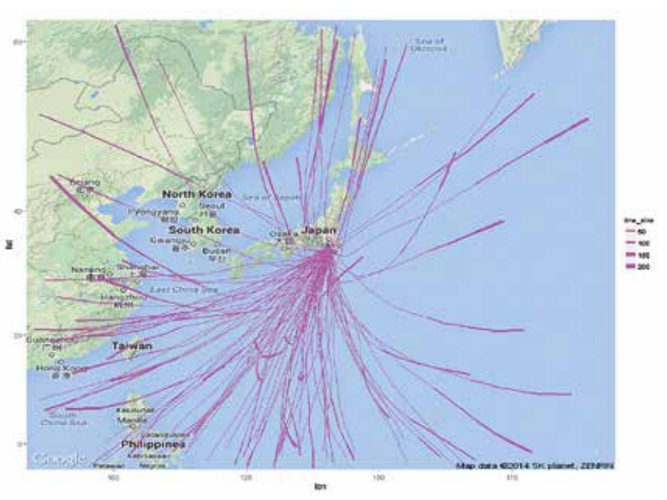


図7：東京を直撃する台風を東京起点として解いた「時間逆転シミュレーション」。図5の確率台風モデルに適用した。

第4回

「海洋生態学と機械学習の融合による生態系評価指標の開発」

水産資源の持続性を統計手法で中立的に評価

水産資源の実態は、環境保護と資源維持という対立する立場によって、評価が異なることが多い。そんななかで求められているのが、水産資源が健全な状態か、枯渇しそうな状態かを中立的に判定する評価手法だ。統数研チームは、漁獲量だけでなく体長や栄養状態などを加味したグローバルな実データを用いて、モデル化に取り組んでいる。

▲江口真透教授

世界の水産資源が枯渇する!?「2048年問題」の衝撃

世界の海で現在のような乱獲や汚染が続けば、人間が食べるシーフードは、2048年までに消滅する――。2006年、カナダの海洋生物学者ウォームらの研究グループが、アメリカの科学雑誌「サイエンス」にそんな論文を発表した。このニュースは、魚食文化の根づく日本はもちろん、世界に衝撃を与えた。

その後、この研究のデータ解析の仕方には問題があるとして、反論する意見や論文が出され、現在ではウォームのシナリオがそのまま現実となる可能性は低いという見方が主流になっている。とはいえ、この論文が契機となり、人間活動が海洋の生物多様性に与える影響に関する研究が各国で盛んになったことは確かだ。

日本でも現在、科学技術振興機構（JST）の戦略的創造研究推進事業「CREST」で「海洋生物多様性および生態系の保全・再生に資する基盤技術の創出」が進行している。統数研が参画する「海洋生態学と機械学習法の融合によるデータ不足下の生態系評価手法の開発」は、その一つとして2012年度に採択された研究だ。

水産総合研究センター中央水産研究所の岡村寛グループ長が代表を務める研究チームに、統計数理研究所からは統計的機械学習研究センターの江口真透教授、数理・推論研究系の小森理特任助教（現在、福井大学特命講師）が参加している。

水産資源の将来予測については、環境保護か資源維持かの立場の違いによって、異なる考察が出回っている。江口は「データに基づく統計手法によって中立的に推論する必要がある」と話す。

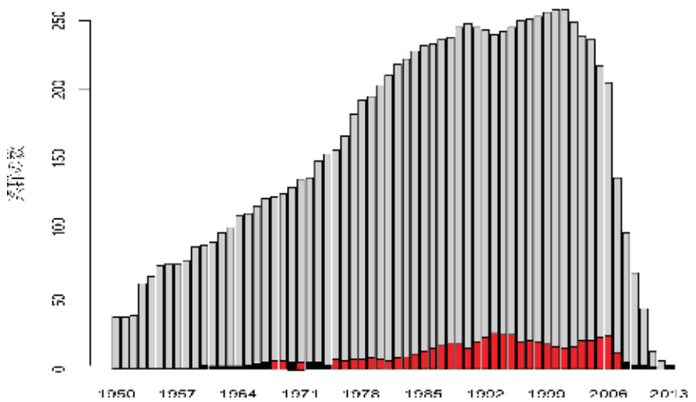


図1：枯渇した系群（赤）と枯渇していない系群（灰色）の数の年推移。

▲小森理 福井大学特命講師

「健全」か「枯渇」か。2値回帰分析に非対称ロジスティックモデルを考案

江口らはこのプロジェクトで、漁獲量と栄養段階、生息地域、最大体長などのデータから2値化したバイオマスデータを予測する統計的モデルを構築。機械学習の手法を用いて、魚などの系群ごとに生態的に「健全」か「枯渇」かを予測する「教師あり学習」を提案した。

ところが、当初からこの手法の確立に成功したわけではなかった。「最先端の機械学習を使えば解決できると思いましたが、あまり良い結果は出ませんでした。もっと根源的な問題があるとわかったのです」と江口は打ち明ける。それが、ラベルごとのサンプルの持つ不確定性の相違だ。健全な系群の分布と枯渇しそうな系群の分布とでは、前者が圧倒的に多く、後者はごくわずかなので、データ数に大きな開きがある。つまり、非対称度が高い。一般に非対称度が高いほど予測は難しくなる。

そこで、江口らが考案したのが、「2値回帰分析のための非対称ロジスティックモデル」だ。2値が非対称であっても、全世界または地域ごと、魚の種類ごとなどで、枯渇の懸念があるかないかを明らかにできる。「検証の結果、従来のモデルによる解析よりも有効なパターン認識が得られることがわかりました」と江口。適切なモデリングを行うことによって漁獲量と栄養段階、生息地域、最大体長など他の変量をうまく組み合わせれば、資源の健全さを予測できる。これによって、「ここまでなら獲っても影響がない」という量が算出できるわけだ。

このテーマでまとめた論文は、2015年9月にイギリスの生態学会誌「メソッド・イン・エコロジー・アンド・エボリューション」に掲載された。江口は「予測の精度が上がり、ウォームの論文への反論を統計的な立場で再評価することができました」と、このプロジェクトにおける成果を説明する。

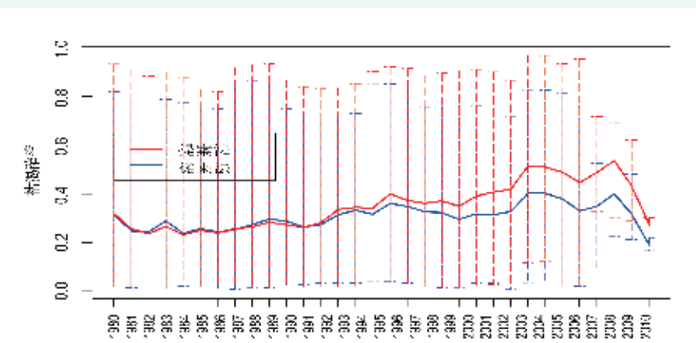


図2：枯渇した系群に対する提案法（赤）と従来法（青）による枯渇確率の比較。

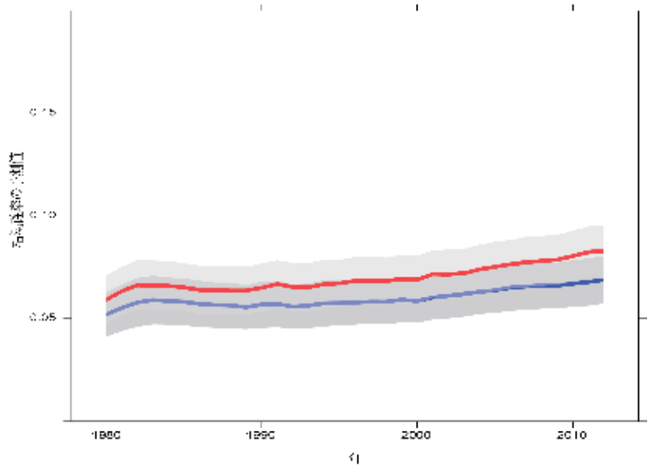


図3：提案法（赤）と従来法（青）による世界的な枯渇確率の推移の予測。

図4：水産資源評価の国際シンポジウムのポスター。

メイヤー博士の遺産として開示された生態系データ

水産資源の将来予測が研究者によって大きく異なる理由の一つは、データの少なさに起因する。陸上とは違い、海の中で生物を観測するのは容易ではない。このため、入手できるデータには限りがある。「水産資源はミッシングデータのかたまりと言っていいでしょう。観測できるのは、全体の何100万分の1程度なのですから」と江口は話す。

最も大規模かつ広域をカバーしているのは、国連食糧農業機関（FAO）の保有する漁獲量データだ。「2048年に水産資源が枯渇する」としたウォームの予測も、FAOが1950年から2003年までに集めたデータに基づく予測だった。だが、漁獲量だけで生態系の状態を把握しようとするのは無理がある。なぜなら、漁獲量は市場の需要に大きく影響され、サンプルに偏りが生じるからだ。予測精度を高めるためには、漁獲量以外のバイオマスデータが必要だった。

そこで、江口らが着目したのが、「RAM記念資源データベース」だ。精力的な研究の半ばで急逝した海洋生態学者ランソン・A・メイヤー博士の功績を記念し、学術研究のためにWeb上で公開されているものだ。ここには、世界的な範囲で200を超える系群のバイオマスを含む年次データがある。「2値回帰分析のための非対称ロジスティックモデル」には、このデータを使用した。モデル化の過程では、メイヤー博士と同じく海洋生態学者であるヒルボーン博士が来日した際に、直接会ってディスカッションをしたという。

水産資源のデータは、長いこと地域限定のクローズドなものだった。各国の経済戦略や外交に絡む重要機密として扱われていたからだ。それがここ10年ほどのトレンドでは開示されつつあるという。「今回のモデル化も、グローバルデータを入手できたからこそ実現できたのです」と江口は言う。

実データがあればこそ困難な推論を乗り越えられる

「アイデアは理論だけ考えても出てきません。現実の課題を解いていくなかから、理論的なヒントを得ることが重要。その意味で、今回はいい経験でした」と江口は振り返る。

このプロジェクトが始動する前、統数研チームは医療分野で抗がん剤の効果に関する評価を手掛けていた。遺伝子発現パターンから、患者の特性ごとに有効な投与量や頻度をモデル化するものだ。だが、モデルを考案してみるものの再現性がなく、難航していた。

そんなとき、このプロジェクトへの参加を打診してきたのが、古くからの研究仲間である岡村教授だった。「原点に戻ってみようと思い、引き受けました」と江口は明かす。抗がん剤の効く効かないも、水産資源の枯渇するしないも、統計的に見ている「ふるまい」は同じ。今回の成果を横展開できる手応えを感じているという。プロジェクトのほうでも、今後は機械学習の方法も援用してFAOデータの予測問題を進める予定だ。

大学院生の時代から江口に指導を受け、共に研究を続けている小森は「このモデルは実際のデータを見て、理論を考えるという繰り返しで到達した考え方。不確実性の多い生物の世界に対応したものなので、金融や機械故障など応用できる範囲は広いでしょう。この先も、実社会に役立つ手法を考えたい」と展望を語る。

ブームとも言えるビッグデータの波は、今回のプロジェクトがテーマとする海洋生態学の分野にも押し寄せている。それにより、調査時の気象や海域の塩分濃度、深度などのデータがリンケージされるようになった。生物にセンサを取り付けてデータを取得するバイオロギングなど、新しい調査方式も開発されつつある。江口は「生態学のなかでも統計学の役割はますます重要になるでしょう。やるべきことは、まだたくさんあります」と先を見つめている。

（広報室）

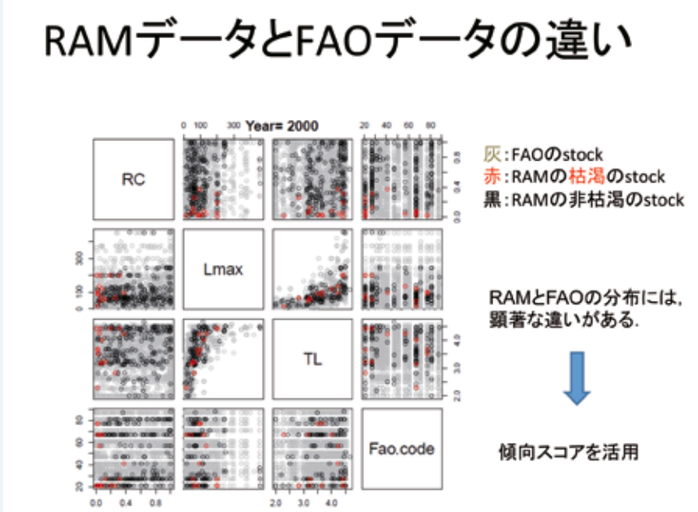


図5：今後の問題としてFAOとRAMのデータの食い違いを注目して選択バイアスの補正にチャレンジしている。

▲プロジェクトチームのメンバー。後列右から東京理科大学大学院博士課程2年の三枝祐輔さん、同修士1年の丸山智久さん、同修士1年の岡田昌之さん。「研究室では理論が中心なので、実データの解析は新鮮な体験」（三枝）、「実データを扱うのは初めてで、勉強になる」（丸山）、「研究室の研究よりも難しい」（岡田）と話す。

第5回

線虫の神経活動をまるごと観測し、神経回路の原理を解明

「神経系まるごとの観測データに基づく神経系の動作特性の解明」

単純な神経回路を持つ線虫C.エレガンスを使い、入力刺激に対する応答を観察することで、神経回路の動作原理を解き明かそうとするプロジェクト。データ駆動型科学へと急速に変貌を遂げる昨今の生物学とデータ科学の連携は、世界的にも大きな課題となっている。そんな中、両者の融合によって金字塔を打ち立てることを目指し、研究者たちの熱い取り組みが進む。

▲吉田亮准教授

生物学と数理科学の融合で創発を目指す

「柳の下の泥鰌（どじょう）」ということわざがある。ヤナギの木の下でドジョウを獲ったことがあるからといって、毎回そこにドジョウがいるとは限らない。だが、期待して再び行ってしまうのが人の性（さが）だ。ヒトだけではない。驚くことに、目も耳もないわずか1mmほどの線虫でも、以前に経験した「いい思い」を求めて同じ条件の元へ移動するという。匂いや味などの化学物質に対して誘引されたり、逆にこれを忌避したりする行動を「化学走性」と呼ぶ。感覚器で感知した情報を神経回路で処理することで引き起こされる行動だ。

線虫を使った研究ではこれまで、低濃度の塩は誘引性の刺激と考えられてきた。ところが最近になって、この化学走性は過去の記憶によって異なる結果を示すことがわかってきた。実験において、低塩濃度の環境で餌を与えられた線虫は低塩濃度の場所へ移動する一方、低塩濃度で餌を与えられずに空腹を経験した個体は低塩濃度を避けて高塩濃度の場所へ移動するという。つまり、塩自体に引き寄せられていたわけではなく、過去に餌にありついた環境と近い環境へ誘引されていたのだ。

このように、動物は「生存に適した環境を求めて移動する能力」を持っている。無数の神経が網目状に結びついた神経回路が演算を行い、「過去の経験を記憶する」「現在の状況をそれと照合する」「より好ましい環境になるように行動する」という情報処理が行われているのだ。

現在、そうした一連の情報処理のしくみを明らかにしようという研究が進んでいる。東京大学大学院理学系研究科の飯野雄一教授が代表を務める「神経系まるごとの観測データに基づく神経系の動作特性の解明」というプロジェクトだ。

この研究は2012年度に、科学技術振興機構（JST）の戦略的創造研究推進事業「CREST」の研究領域「生命動態の理解と制御のための基盤技術の創出」に採択されたもの。統計数理研究所からは、モデリング研究系の吉田亮准教授が参画している。

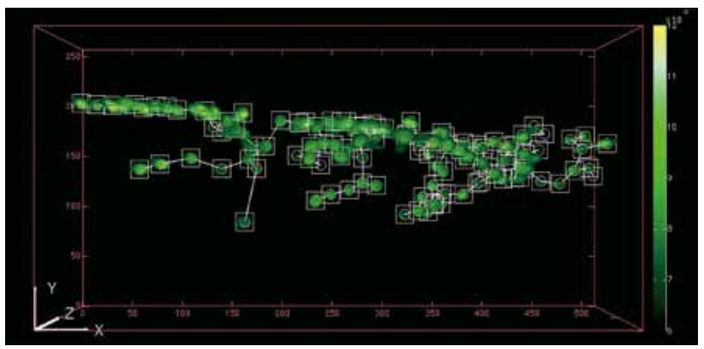


図1：神経細胞のトラッキング

統計科学者としての吉田の専門はライフサイエンス。「この分野にはまだ、データ科学の専門家が少ない。ベイジアンモデリングや機械学習のユニークな発想に基づき、革新的な取り組みをしたいと思っています」と意欲的だ。

研究チームは実験班と数理班から構成されている。実験班には飯野教授と九州大学大学院理学研究院の石原健教授、数理班には吉田と茨城大学工学部の岩崎唯史講師が名を連ねる。吉田は、データ解析班のグループリーダーを務め、機械学習や画像解析の先進技術を駆使して、バイオイメージデータに基づく神経活動の定量分析技術を開発している。

線虫の神経活動を定量化し神経系の情報処理プロセスを可視化

この研究に使われるのが、体長約1mmの透明な身体を持つ線虫「C.エレガンス（*Caenorhabditis elegans*）」だ。ノーベル賞を受賞した分子生物学者シドニー・ブレナーが、1960年代に線虫を使ったゲノムプロジェクトを立ち上げたことがきっかけとなり研究が進んだ。80年代には受精から成虫に至るすべての細胞分裂のプロセスと、302個の神経細胞のつながり方が明らかになった。現在では全ゲノム配列も解明されており、さまざまな研究分野でモデル生物として重用されている。

実験は、次の手順で行う。まず、生きたC.エレガンスを共焦点顕微鏡に設置されたチューブに流し込む。使用するC.エレガンスは神経活動の状態変化に応じてニューロンが蛍光を発するように操作されている。次に、チューブの先端から刺激物を注入。すると、刺激に反応して細胞中のカルシウムイオン濃度が変化し、その様子が光の明るさに表れる（図1）。この画像の緑色の球状物体が神経細胞を示している。このように、明るさの変化を時間経過とともに立体動画で追跡する（図2）。

吉田の役割は、画像データから神

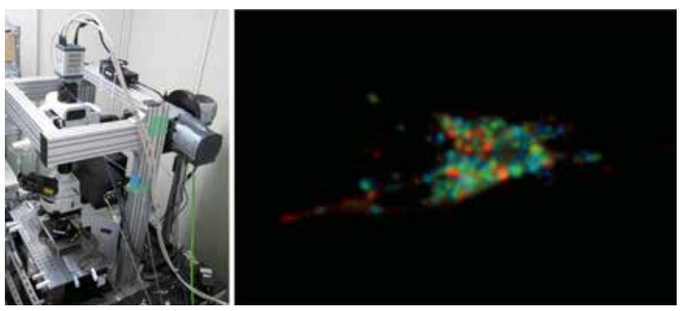


図2：4Dイメージングシステム（九州大学石原研究室）

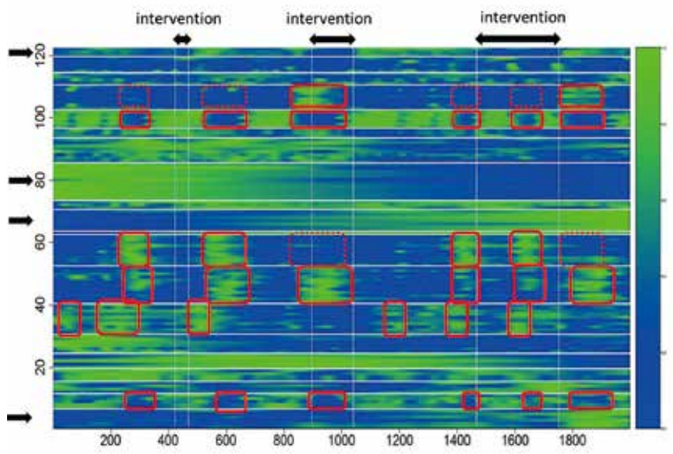


図3：定量化された神経活動の時間変化。横軸は時間、縦軸は細胞を表す。神経細胞が活性化している状態は緑色、沈静化している状態は青で表示される。

前述のように、C.エレガンスの神経細胞はすべて分かっていて、それぞれAFDニューロン、AWAニューロンというようにアルファベット3文字の名前がついている。しかし、顕微鏡で観察している細胞が、そのうちのどれに対応しているかについては手がかりがない。そこで、機械学習で名前を特定する。これによって初めて、情報処理における神経回路の役割が見えてくるわけだ。

たとえば、C.エレガンスは全般にアルコールを好み、近くに置くとそこへ向かって移動する。冒頭で挙げた「好ましい環境を求めて移動する能力」、化学走性だ。一方で、C.エレガンスは銅イオンを嫌う傾向があり、化学走性によってこれを避ける。ところが、実験でアルコールの手前に銅イオンを置くと、それを乗り越えて進む個体と、諦めてしまう個体があるという。その違いは恐らく、神経活動の回路に基づいていると見られるものの、詳しいメカニズムは明らかになっていない。

さらに、外部から刺激が与えられていないときでも、活動レベルが上がったり下がったりしている神経細胞が少なからずあることも、このプロジェクトを通じてわかってきた。「自らの活動を自発的に調整している神経細胞があるらしい」という事実は、当初は予想されていなかった。その生物学的な意味の解明はこれからだ。

現在、こうした研究を進めているチームは世界中にいくつかあり、しのぎを削っている。「このプロジェクトが成功すれば、世界最先端の研究成果を世に問うものになります。他のチームに先を越されずに完遂したい」と吉田は力を込める。

連携を加速させる金字塔へ。プロジェクトにかかる大きな期待

こうした目標を達成するには、データ科学のすぐれたアイデアが必要だ。また、生物学の側にも仮説の精度をより高めていくことが求められる。「このプロジェクトは、いわば4人の研究者の“運命共同体”です。どのピースが欠けても研究は成り立ちません」と吉田。まさに今、実験と数理が車の両輪となって、ダイナミックに前進しつつある。

吉田にとって、線虫の研究者との共同研究は初めての経験だ。連携の実感を「生物学の基礎研究ならではの伝統的なサイエンスのカルチャーがあり、それが心地よい」と表現する。

チームのデバイスの拠点は、九州大学のスーパーコンピューター。画像処理の要素技術を直列に連結して流れ作業で並行処理させる「パイプライン」を構築し、データ処理をハイスループットに行う仕組みを構築している（図4）。

チームの研究者たちの所在は、飯野教授と吉田が東京、岩崎講師が茨城、石原教授が福岡と物理的距離は遠いが、オンラインでのやりとり以外にも頻繁に顔を合わせて議論しているという。

2003年に完了したヒトゲノム計画は、生物学とデータ科学の距離を一気に縮めた。ゲノム解析で得られた膨大なデータから生物学的な意味を見出すには、バイオインフォマティクス（生命情報学）の手法を活用する以外にないからだ。それ以来、データ科学と生物学の連携の重要性はますます高まってきている。先行事例であるこのプロジェクトも、CRESTの中で大きな期待を寄せられている。「成功すれば金字塔を打ち立てることになり、連携が勢いづくことは間違いありません。ぜひともモデルケースを作りたい」。そう話す吉田の口ぶりに熱がこもった。（広報室）

障壁を乗り越えるか、諦めるか。神経回路が行動を決める

神経システムの動作原理については、線虫のように脳とは呼べない単純な神経回路を持つ動物でさえ、まだわかっていないことばかりだ。

神経細胞の活動状態を推定（定量化）し、システム全体の動的プロセスを可視化することだ。神経細胞の状態を定量化するには、物体検出やオブジェクト・トラッキングの高度な解析技術が必要となる。プロジェクト開始から約3年で画像取得から定量化までのプロセスを全自動化することに成功した（図3）。

「今、取り組んでいるのはその次の段階。ポイントは、定量化された神経活動の状態をいかに神経回路にマッピング（紐付け）するかです」と吉田は説明する。どの神経細胞が刺激に対しどう反応したか、それによって他の神経細胞にどのように影響を与えたか。そのパターンを読み解くには、活動状態を神経回路に対応付けた上で観察する必要がある。

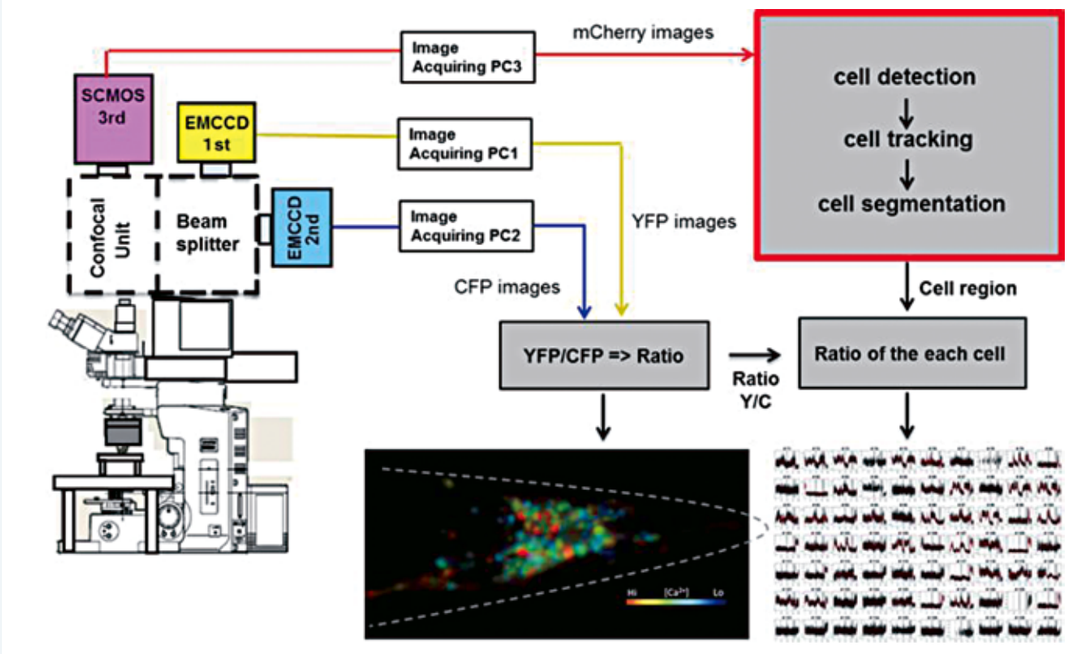


図4：画像処理パイプライン。画像データを入力すれば、細胞検出、計数、トラッキング、神経活動の定量化のプロセスが自動実行される。

第6回

「データ中心科学による地球環境問題解決を目指す国際ネットワーク構築」

データサイエンスを共通言語として国際的な人材交流をリード

地球環境の変動に起因するさまざまなリスクを低減するために、データ中心科学に期待される役割は大きい。現在、統計数理研究所では、2つの国際ネットワーク形成に向けた取り組みを進めている。アジア環境研究実践ネットワーク形成のための共同研究指針策定の準備と、欧州やシンガポールの研究機関との国際NOE形成戦略策定の準備だ。両者の概略的な内容を紹介する。

▲吉本敦教授

地球環境変動によるリスクを解決に導く国際ネットワーク

人間活動が環境へ与える影響に、国境はない。一部の国が個別に環境保護対策を進めても、他国で環境破壊行為が続けば、広域に悪影響を及ぼす懸念は避けられないからだ。この意味で、地球環境問題の解決には、国際的な枠組みによる全地球包括的な取り組みが必須となる。

統計数理研究所の「データ中心科学による地球環境問題解決を目指す国際ネットワーク構築」は、地球の環境変動に起因する諸リスクを低減させるために、データ中心科学の方法論を開発するとともに、それを実践する国際研究ネットワークを構築するプロジェクトだ。吉本敦教授、金藤浩司教授、清水邦夫特命教授、松井知子教授が主導している。

途上国自らによる森林資源管理を目指すワークショップ

吉本が手がけているのは、カンボジア、ベトナム、ネパールといった発展途上国で森林資源管理の実施に関わる人材を育成する取り組みだ。フリーソフト「R」を用いて統計解析のノウハウを身につけるワークショップを各国で継続的に開催している。2010年のカンボジアを皮切りに回を重ね、これまでに実施したワークショップは6年間で5カ国（カンボジア、ネパール、ベトナム、フィリピン、ラオス）、延べ13回に上る。1回の参加者は30人から50人程度で、合計およそ500人以上が受講した計算だ。統数研は2015年3月にカンボジア森林局森林研究所、ネパールのポカラトリブヴァン大学森林研究所と、また同年6月にはベトナムの森林開発企画研究所とMOU（覚書）を締結した。

写真1：2011年にカンボジアでワークショップを開催したときには、現地の森林を視察した。写真中央が吉本。

ワークショップを始めたのは、吉本がカンボジアの森林研究所に所属している教え子から、実情を聞いたのがきっかけだった。現地の研究者は、国連機関などが同国で実施する森林調査でデータを採取することはあっても、それを解析して目的に応じた森林管理に役立てるには至っていないという。「ローカルニーズに応じて森林資源を適切に利用しながら管理するには、途上国内に統計解析のできる人材を増やすことが重要だと考えました」と吉本は振り返る。

写真2：写真の右側は、日本に留学していたときの吉本の教え子、カンボジア森林局森林研究所所長のSokh Heng氏。

統計手法を使えば、相反する資源の利用と保全をうまく両立させることが可能になる。例えば、整数計画法を用いて、「隣接し合う土地の利用を同時期に行うことができない」という条件を加えれば、大規模な資源開発を回避することができる。また、火災や病害虫の被害が拡散するパターンを考慮して植生状態の空間配置を制御すれば、被害規模の軽減に役立つ。

だが、高額な市販ソフトを使用したのでは、途上国での普及は難しい。そこで、吉本が目を付けたのが、無料でインストールできるオープンソースの統計解析ソフト「R」だった。まずはカンボジアで、Rを用いて統計学の基礎を習得するワークショップを開催することにした。最近では、Rを使いこなす人材が増え、初期の参加者が他の研究員に使い方を教えるなどの広がりも出てきた。ワークショップ前日、自発的にレビューを行う参加者たちの熱意に、吉本は目を細めている。

統計的な管理手法は林業だけでなく農業や漁業といった他の第一次産業にも展開できることから、途上国の食糧問題を根幹でサポートすることにもつながる。将来的には各国のデータをデータベース化し、統数研とクラウドで結ぶことも考えられる。

プロジェクトの意義について、吉本は次のように語る。「統数研が取りまとめをすることで、各国のデータを通した広域のモニタリングができる。一方で、途上国側はそのデータベースを使って自国の資源を有効に利用できる。日本と途上国の双方にメリットがあるはずです」

各国のデータを共有するには、信頼関係の構築が何よりも重要。地道なワークショップの開催を通じて自然に良好な関係ができつつあるという。

旧知の教授との縁が取り持つマレーシアとの研究者間交流

一方、20年近くにわたりマレーシアのマラヤ大学と研究者間交流を続けているのが清水だ。同大学はマレーシアのトップ校で、世界でも上位に位置づけられる国立の総合大学。

第6回「データ中心科学による地球環境問題解決を目指す国際ネットワーク構築」

▲清水邦夫特命教授

清水は1997年にインドで開催された国際集会で、同大学数理科学研究所のOng Seng Huat教授と出会った。

Ong教授の手がける離散分布論の研究分野が、清水の研究対象と近かったことから、二人は意気投合。以来、清水は毎年のようにマラヤ大学を訪れるとともに、当時所属していた慶應義塾大学にOng教授を招き、離散分布論に関して共同研究をしてきた。同教授を通じてマラヤ大学の教員や学生との交流も広がった。

2014年に統数研統計思考院の特命教授となった清水は、統数研とマレーシアの大学、研究所との交流の橋渡し役を担うことになる。同年11月にクアラルンプールで開催された統計学の国際会議ICMSFMの準備にも協力し、統数研の樋口知之所長がキーノートスピーカーとして招待されることになった。会議の翌日は、統数研からの参加者とともにマラヤ大学を訪れ、研究交流の輪を広げた。

現在、清水は金藤らとともに、マレーシアにおける地すべりリスク評価手法開発のプロジェクトを進めている。このプロジェクトでは、部分的にマレーシア科学大学のグループと協働している。清水は「日本がASEAN諸国との良好な関係を築くには、支援だけではなく、一緒に研究を進めるスタンスが望ましいと思っています」と話し、このプロジェクトに期待している。

「つぶやき」の分析で都市インテリジェンスを向上

松井は、ビッグデータのための理論と方法について国際連携を進めている。連携先は、統数研がMOUを締結しているシンガポールのインフォコム研究所、スイスのETH Zurichリスク研究所、フランスのIRCICA、イギリスのロンドン大学など先進5カ国の8大学・研究所。これらの連携の母体となったのは、松井とロンドン大学のピーター・ギャレス博士が中心となり、3年にわたり開催してきたSTM（Spatial-Temporal Modeling; 時空間モデリング）のワークショップだ。

2015年2月に東京・立川の統数研で開催したワークショップには、日英のみならず世界中の同分野の研究者およそ50人が参加。5日間の会期で熱い議論を交わした。松井がギャレス博士とともに、このワークショップに参加した国立環境研究所の山形与志樹博士、村上大輔博士と行った共同研究が「ツイートを利用した都市気温の解析」だ。

▲松井知子教授

各地の気温の推移は気象庁などでも観測しているが、拠点数も時間間隔もスパース（疎）だ。そこで、この研究では、都市の各所でツイッター上に絶え間なく投稿される「暑い」「だるい」といった“つぶやき”（ヒートツイート）を利用し、既存の観測法による気象データを空間的、時間的に補間する統計・機械学習の方法を検討した。

松井らは、単純な相関ではない2種類のデータの関係を細かく決められる「コピュラモデル」を用いて、気温補間にヒートツイートが有用できることを確認。熱波で問題となるupper tail（気温は高い／気温変化は大きい／ヒートツイートは多い時に相当）では特に、気温変化がヒートツイートに関係していることなどを発見した。また、S-BLUE法の枠組みを用いて、気温データとヒートツイートを組み合わせたモデルによって、東京の中心部のヒートアイランド現象を捉えることに成功した。

地球規模の異常気象などのメカニズムの解明は、各地の防災にも直結する。今年4月、松井は同ワークショップのメンバーに声を掛け、「都市インテリジェンス研究プロジェクト」を立ち上げた。環境・エネルギーや農業の状況解析からリスク管理、セキュリティ統合、都市レジリエンスボンド設計までを俯瞰的に行うための統計数理や機械学習に基づく技術とその理論を研究開発するもの。目標は、都市レジリエンスの向上だ。

「東日本大震災をきっかけに、防災を含めて社会的に重要な課題に対し、自分のできることをしたいと考えるようになりました。STMワークショップを通し国籍を問わず気心の知れた研究仲間ができたので、彼らとともに研究を進め、統計数理が有効に使えることを社会に示していきたい」。松井はそう話し、意欲をのぞかせた。

国境に関わりなく発現する地球環境問題や異常気象。これに立ち向かう統計数理学者たちもまた、国境を越えてデータサイエンスの言葉で語り合い、ネットワークを広げている。

（広報室）

写真3：マラヤ大学にて、同大学数理科学研究所の若手教員と清水。向かって左から、Ng Choung Min、清水、井本智明（現・静岡県立大学；当時 統計数理研究所）、Sim Shin Zhuの各氏。

写真4：清水が統数研の樋口所長らとマラヤ大学を訪れたときのセミナー風景。

写真5：2015年7月13日から17日に開催されたビッグデータワークショップ「STM 2015 & CSM 2015」の様子。奥側左が松井。

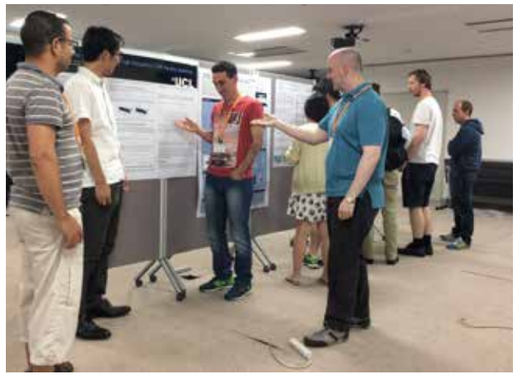


写真6：英国、日本、フランス、中国など10カ国以上から延べ55人の研究者や技術者が一堂に会した。

第7回「公的統計ミクロデータ研究コンソーシアム」

公的統計のミクロデータ活用を促す基盤整備が始動

各界の学術研究において、「ミクロデータ（集計していない調査個票情報）」を用いた実証分析は、世界の主流になりつつある。日本でも統計法の改正によって、公的機関の保有するミクロデータの二次利用が可能になった。現行の利用手続きを見直し、セキュリティーや秘匿性の課題を解決しつつ、活用を促進することを目指し、統計数理研究所を中心とする官学のコンソーシアムが発足した。

利用者ニーズを吸い上げるコンソーシアムを立ち上げ

2016年3月、一橋大学の一橋講堂（千代田区）は、およそ200人の研究者らの熱気に包まれた（写真1、2、3）。開催されていたのは、「公的統計ミクロデータ研究コンソーシアム」の設立記念シンポジウム。全国の大学や研究機関のほか、データベンダーや統計ソフトウェア開発企業など民間企業からも参加者が集まった。希望者が殺到し、当初予定していた会場を、より大人数を収容できる同講堂へ変更したことからも、このテーマへの関心の高さが伺える。

「公的統計」とは、国の行政機関や地方自治体など公的機関が作成する統計を指す。「ミクロデータ」は、その基になる個々の調査票データだ。このコンソーシアムは、公的統計ミクロデータを学術研究に利用する関係機関が共同で、データの提供側である政府と連携し、オンサイト施設を全国に開設するための課題を検討することを目的としている。

コンソーシアムを構成するのは、任意の国公私立大学や研究機関。情報・システム研究機構データサイエンス共同利用基盤施設社会データ構造化センターが事務局となり、総務省政策統括官、統計局、統計センターの協力を得て運営している。

統計数理研究所では、椿広計名誉教授を代表者とする官学融合の科研費研究として「公的統計ミクロデータの構造化と研究利用プラットフォームの形成」に取り組んでいる。目指しているのは、公的統計ミクロデータの利用効率と提供作業効率を向上させ、二次利用制度のプラットフォームを研究し、政府による実装を支援すること。コンソーシアムの設立もその一環である。

▲山下智志教授

写真3：情報・システム研究機構の北川源四郎機構長は「データ共有化と公的統計ミクロデータ」のテーマで講演を行った。

写真1：2016年3月29日に開催したシンポジウムでのパネルディスカッションの様子。左が司会の椿広計統計センター理事長、パネリストは左から一橋大学経済研究所の北村行伸所長、神戸大学大学院経済学研究科の地主敏樹研究科長、総務省統計局統計調査部の千野雅人部長、SAS Institute Japanの堀田徹哉社長。

写真2：会場には約200人の研究者などが集まった。

統数研からは、プラットフォーム設計・研究ユニットのリーダーとして岡本基リサーチ・アドミニストレーターが、またデータ構造化・標準化ユニットのリーダーとして山下智志教授が、それぞれこの研究プロジェクトに参画している（図1）。

二次利用の促進に不可欠なセキュリティー環境の整備

公的統計ミクロデータの研究利用が推進されるようになったのは、2007年に統計法が改正され、公的統計が「行政のみならず社会全体で利用すべき情報基盤」と位置づけられてからのことだ。09年に改正法が施行されて以降、学術研究など公益目的の利用であれば、調査票情報の二次的な利用が認められている。

岡本は「特に最近は、集計されたデータではなく、大規模ミクロデータに基づく研究を推進する機運が高まってきました」と話す。例えば、5年前には厚生労働省の「ナショナルデータベース（NDB）」の研究利用が開始されて話題を呼んだ。100億件を超す診療報酬明細書（レセプト）や、およそ1億7000万件に上る特定健診などの医療ビッグデータは、生活習慣病の改善や予防医学など、さまざまな医療研究に役立つと期待されている。

しかし、統計法改正から10年近くたったいまも、公的統計ミクロデータの二次利用は思うように進んでいないのが実情だ。その理由の一つに、セキュリティーの問題がある。

研究者が調査票情報の提供を受けるためには、データを操作する部屋の入退管理や保管するキャビネットの施錠、外部ネットワークからのシャットアウトなど、一定の要件を満たさなければならない。このため、共同研究室を割り当てられている若手研究者や施設の都合でセキュリティー強化が難しい組織に属する研究者などは、データ利用の申請ができなかったのである。

「人文社会科学分野の国際的なジャーナルでは、すでにミクロデータを用いた実証分析が主流で、それなしに論文が掲載されることはほぼ不可能になっている。日本は一刻も早く利用環境を整える必要があります」と岡本は指摘する。

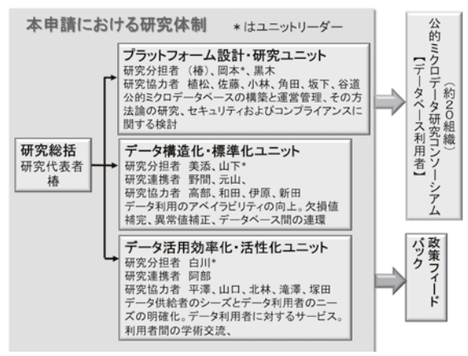


図1：研究の推進体制。

第7回「公的統計ミクロデータ研究コンソーシアム」

▲岡本基リサーチ・アドミニストレーター

リモートアクセス型拠点を増やし セキュリティーと利便性を両立

利用促進とセキュリティー確保の両立は、ミクロデータを活用するうえで避けて通れない問題だ。その解決に向けた第一歩として、2010年には機構とデータの提供元である統計センターが連携協定を結び、セキュアな環境を整えた「オンサイト解析室」が開設された（写真4、図2）。研究者は、所定の手続きを経てこの場所へ赴けば、匿名化したミクロデータの提供を受けることができる。

ただ、オンサイト解析室はこれまで、全国で統数研と一橋大学の2ヵ所にしか設置されていなかった。また、ミクロデータを研究者に直接提供する現行の方法はリスクが高く、改善が急がれる。

そこで、今回の研究で目指しているのが、リモートアクセス型のオンサイト拠点を全国に展開するとともに、セキュリティーをホスト側で一元管理するネットワークの整備だ（図3）。

現行のオンサイト解析室の設計から運営管理までを手がけてきた岡本は「こうしたネットワークを構築すれば、セキュリティーを強化しつつ、研究者にとってデータを活用しやすい環境をつくることができます」と説明する。この6年間のオンサイト解析室の運営実績から見えてきた課題を、新たなネットワークのルールづくりに反映させるつもりだという。

総務省は当面、マスタープランに賛同するコンソーシアムメンバーの大学や研究機関を対象として、10カ所程度のオンサイト拠点を設置することを目指す。手始めとして、2016年度中に4拠点を結んでリモートアクセスによるネットワークの試行運用を開始する予定だ。

欠損データの補完や構造化に 統数研の知見を生かす

一方、これまでミクロデータの二次利用が進まなかったもう一つの理由として、データの質の問題が挙げられる。集計をしていない生のミクロデータには欠損値や異常値が混在し、そのままでは分析に使えない。欠損値を補完し、異常値は補正し、データベース間の連環をさせるなど、利用可能なデータに落とし込む必要がある。このため、収集したミクロデータをそのまま提供したのでは、利用する研究者は、分析を開始する前に前処理に多大な時間を費やさなければならない。この点を解消しようというのが、山下の率いるデータ構造化・標準化ユニットの取り組みだ。

山下自身、これまでに政府からの委託を受け、公的統計ミクロデータを活用して、さまざまな分析を手がけてきた実績がある。ミクロデータを集計して分析するのではなく、ミクロデータを直接に統計モデルのパラメータ推計に用いる「非集計分析」と呼ばれる手法だ。

例えば、九州新幹線の計画時に、国土交通省の所管する「全国幹線旅客純流動調査」のミクロデータを分析して、需要予測が可能となった。この調査は、鉄道や航空機、自動車といった交通機関別の流動調査ではなく、個人に着目し、旅行の出発地からどのような交通機関を乗り継いで、どこへ到着したかを調べたものだ。

このミクロデータをもとに、ランダム効用理論を用いたロジットモデルをつくり、人が行動を決める要因である「確定効用項」を求める。すなわち、目的地に至るまでの経路と手段が、何を重視して選択されたかを推計する。そこから、九州新幹線ができて所要時間や費用が変化した場合に、どのくらいの人数が利用するかを導き出したのである（図4）。

公的統計によってはデータの欠損が多く見られるものがあり、山下は事前にその調整も手がけた。また、前述のNDBの構築にも関与し、統数研の医療統計チームと連携して欠損値補完の新たな手法を作り出した経験がある。

「一口に公的統計と言っても、所管する行政機関によって精度にバラツキがあり、また従来のデータ構造化の手法も分野によって異なるケースがあります。欠損を整えたり、手法を合併したりするのも統数研の役割です」と山下は話す。山下のユニットでは、今後、構造化に対する利用者のニーズ把握に努めるとともに、高速アルゴリズムの開発にも力を入れていくという。

図4：山下が過去に手がけたミクロデータ活用の事例。国土交通省の所管する「全国幹線旅客純流動調査」のミクロデータから、九州内の周遊状況を分析した。

日本の産業をパワーアップ 政策へのフィードバックも

データの構造化に関しては、「名寄せ（リレーション）」も大きな課題だ。一人の個人や一つの法人のデータが複数のデータベースに収載されている場合、名寄せができればデータ活用の幅は飛躍的に広がる。だが、ここには「秘匿性」というデリケートな問題が横たわる。

これから展開するリモートアクセスによるオンサイト拠点ネットワークに、秘匿性を過度に保護するルールを設定したのでは、利用者にとって使いづらいシステムになってしまう懸念がある。その半面、利用者側にも、取得したミクロデータを集計する際に、秘匿性を確保するスキルが求められる。岡本は「今回のコンソーシアムは、利用者の立場である大学にも参画してもらっていることがポイントです。利用者の意見を吸い上げて、落とし所を考え、合意を形成していきたい」と抱負を述べる。

山下は「ミクロデータに基づくエビデンスがあれば、民間プロジェクトが進めやすくなるなど、日本の産業パワーにも貢献するでしょう。また、政策へのフィードバックも期待できる。社会との二人三脚で、複雑なステークホルダーの思惑を整理するのがわれわれの使命だと思っています」と力強く語った。（広報室）

写真4：統数研内に設置されたオンサイト解析室。高いセキュリティー環境を実現している。

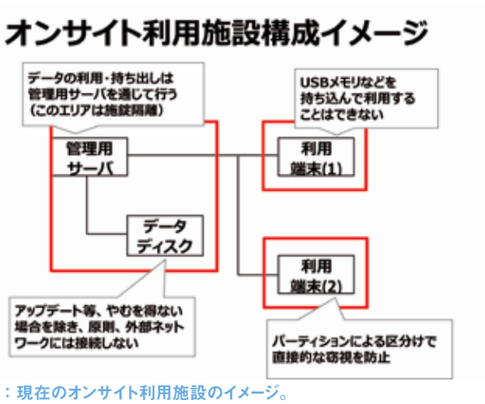


図2：現在のオンサイト利用施設のイメージ。

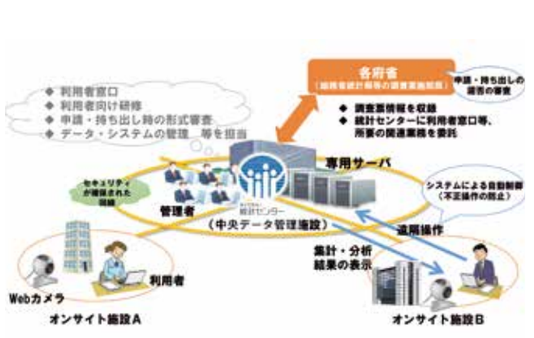


図3：全国ネットワーク整備のイメージ。

トポロジーを応用した統計理 論でソフトマターの構造解明を目 指す

ガラス、タンパク質、コロイドなど柔らかい物質、いわゆる「ソフトマター」は、その物質を構成する単位がいずれも複雑な幾何構造を形成している。新材料の開発や難病の治療法確立には、関連するソフトマターの構造解明が不可欠だ。トポロジー（位相幾何）を応用した位相的データ解析という新たな数学的手法によって、この構造を明らかにするための記述法を開発するプロジェクトで、統計数理研究所は位相的統計理論の構築を担っている。

▲福水健次教授

記述困難だった非晶質の構造を位相的に捉える新たな方法論

ガラスやタンパク質など、複雑な原子配置構造を持つソフトマター。長年にわたり、その構造を適切に記述する方法の開発が求められてきた。というのも、物の構造は、機能と密接な関連があるからだ。例えば、材料の分子構造は、その材料の持つ硬さや柔らかさや強靭さといった特性に直結する。材料の設計やネットワークの制御などを検討する際にはまず、対象とする物の形を記述し、特徴づけをすることが不可欠だ。

物の形は、三角形の集まりとして記述することで代数的な扱いが可能になる。基本構造の繰り返しからなる結晶構造を持つ物であれば、比較的容易にその形状を記述することができる。だが、ガラスのようなアモルファス（非晶質）状態の物質の場合は、一筋縄ではいかない。これらの構造は単一の大きさの構造を持たないことから、マルチスケールで考える必要があり、それを記述しようとすると情報量は膨大になってしまう。

こうした課題にまったく新しい角度から解決への道筋を照らし出したのが、トポロジーを応用した位相的データ解析の手法だ。科学技術振興機構のチーム型研究「CREST」に採択された「ソフトマター記述言語の創造に向けた位相的データ解析理論の構築」は、データの「形」を定量的に記述する理論を確立し、ソフトマターの構造解析へ応用することがテーマ。高機能ガラス材料や、疾患関連タンパク質のフォールディング構造解析といった挑戦的基礎研究を実施している。

東北大学原子分子材料科学高等研究機構の平岡裕章教授を研究代表者とし、統計数理研究所からは福水健次教授と栗木哲教授が参画している（図1）。

TDA班（東北大・GL平岡裕章）
　パーシステント逆問題理論
　ソフトマター研究
　ソフトウエア開発

表現論班（静岡大・GL浅芝秀人）
　一般化パーシステント加群理論

確率論班（九大・GL白井朋之）
　ランダムトポロジー理論

統計班（統数研・GL福水健次）
　位相的統計理論

MD班（岐阜大・GL一宮尚志）
　タンパク質フォールディング研究

図1：研究の推進体制。プロジェクトチームは純粋数学、応用数学、統計科学、確率論、物質科学など幅広い分野の研究者で構成されている。

「穴」の数をマルチスケールで数えるパーシステントホモロジー

トポロジーでは、連続的に移り合う図形は同じとみなす。たとえば、「穴が一つ」という共通の性質を持つ「取っ手のついたマグカップ」と「リング状のドーナツ」は、それぞれを引き伸ばしたり曲げたりすることで、切り貼りをしなくても移り合える。したがって、この二つの物体は、トポロジー的には同じということになる（図2）。

今世紀になって、このトポロジーを応用した新たなデータ解析手法が開発された。それが、データの位相的・幾何学的情報を抽出することによって解析を行う「位相的データ解析（Topological Data Analysis＝TDA）」だ。

図2：トポロジーは、物の形を連続変形させた際に保たれる性質に着目した数学の一分野。左のマグカップと右のドーナツは、「穴が一つ」という特性を保ったまま移り合うことから、同一とみなされる。

TDAは「パーシステントホモロジー」をキーテクノロジーとする。「ホモロジー」は、図形を点と線と三角形の集まりで記述したうえで、そこに含まれる「穴」の数を数えるもの。ただ、これだけでは複雑な構造に対応できないことから、パーシステントホモロジーではさらに穴の数だけではなく、サイズや形、階層性などの情報までも同時に見ていく。

ノイズのある有限個のサンプルから真の構造を捉えようとするとき、どのような解像度（スケール）で見るかによって、見え方は違ってくる。例えば、小さいスケールでは不連続な集合として見えてしまうし、大きいスケールでは小さいリングがつぶれてしまうなど、適切なスケールの設定が難しい。

これに対し、パーシステントホモロジーを用いれば、トポロジー的に同じとみなせる「穴」の生成と消滅の様子をマルチスケールに捉えることができる（図3）。その表現法にはバーコードと「パーシステント図」の2種類がある。

「ガラスとは何か」という命題を数学的な手法で解き明かす

プロジェクトチームはすでに、TDAを用いてガラスに含まれる階層的な幾何構造の解明に成功している。「アモルファスであるガラスの分子構造は、規則性がないように見えて実はある。ただ、物性を予測するのに何を見ればいいか、記述子がわからなかったのです」と福水は説明する。

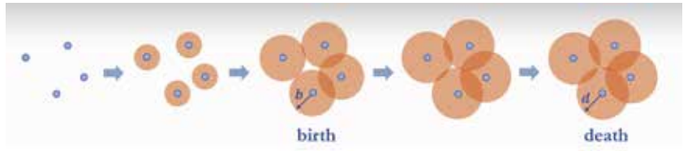


図3：パーシステントホモロジーの概念。複数の点が時間の経過とともに膨らむと、点が連結して中央に「穴」が生まれるが、この穴は、さらに膨らむと押しつぶされて消滅する。パーシステントホモロジーは、その経過を表現する。

こうした記述法の開発は、「ガラスとは何か」という基礎科学上の大問題への理解を深めるのに役立つ。同時に、産業にとっても、情報ストレージや太陽光パネルなどのガラス材料開発にも直接的に結びつく。平岡教授率いるグループの成果は、2016年6月に米国科学アカデミー紀要「Proceedings of the National Academy of Sciences」オンライン速報版に掲載され、世界中の注目を集めた。

ガラスは結晶とは異なる乱れた3次元原子配置を持っており、その構造を明らかにするには、広範囲の原子で構成される中距離構造を調べる必要がある。しかし、これまでの方法では、それぞれの原子の周辺の短距離構造までしか調べることができなかった。

中距離構造を記述する際のポイントは二つある。一つは、多くの原子からなる多体系の特徴をどのように記述するか。もう一つは、短距離から中距離までのマルチスケール性をどのように扱うか。プロジェクトはこれらの課題を、パーシステントホモロジーを用いたTDAで解決した。

具体的には、原子配置を空間内の点の集まりとみなし、そこに含まれるリ

第8回「物質科学への応用に向けた位相的統計理論の構築」

▲栗木哲教授

ングや空洞といった「穴」に着目するマルチスケールデータ解析を可能にする。これによって、シリカ（SiO_2）などのガラスについて、液体とガラス状態の内部構造の違いを幾何学的に特徴づけた（図4）。なかでも、ガラス状態において、原子配置のリング構造に階層性を持った秩序構造が存在することを明らかにしたのは、大きな発見だった（図5）。ここで得られた知見を基に、プロジェクトはガラスの硬さの起源にあたる中距離秩序構造の記述などにも成功している。

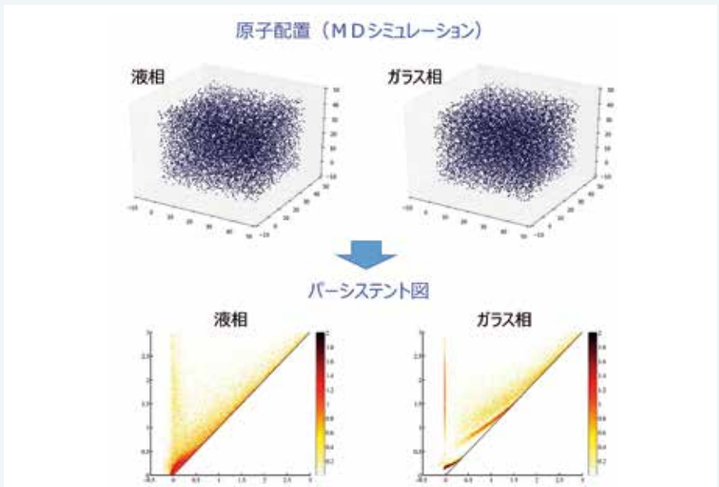


図4：SiO_2の原子配置（左）とそのパーシステント図（右）。ガラスは液体と異なり三つの曲線のような帯状領域を持つ。

図4、5のデータは中村壮伸氏（東北大）より提供。
参考：T.Nakamura et al.（2015）Nanotechnology, Y.Hiraoka et al.（2016）PNAS.

パーシステント図のデータ解析に機械学習的なアプローチを導入

今回のプロジェクトで、福水がチームリーダーを務める統計班では、ガラスの相転移の特徴づけなどに役立つデータ解析手法の確立に取り組んでいる。液状のシリカを急冷した際、液相からガラス相へ変わる瞬間の温度を特定するものだ。

データとしては、プロジェクトメンバーである中村壮伸氏（東北大）らが作成した、原子・分子の動きをコンピュータ上で再現するSiO_2分子動力学（MD）シミュレーションデータを用いる。まず、温度を変えて80セットの3次元原子配置データを取得し、これを基に計算を行い、パーシステント図を作成する。ここまでが図4の状態だ。

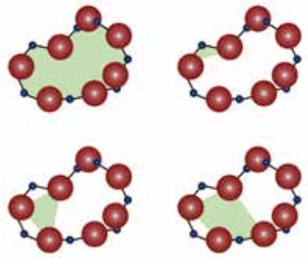


図5：図4のガラス相のパーシステント図に存在する典型的なリング構造。パーシステント図は緑色のリングを捉えたもの。赤球はO原子、青球はSi原子を表す。

従来の物理学的方法では、MDシミュレーションデータからエンタルピー曲線を描いて微分の推定を行い、その不連続点を推定する。だが、この方法では精度に限界があった。そこで今回、福水らはパーシステント図のデータ解析にあたり、カーネル法による機械学習を導入。非ベクトル的なデータをベクトル化することで、系統的データ解析を可能にした。「データセットから起こした多くのパーシステント図を記述子として、最終的に転移点を数値にまで落とし込むことを目標としました」と福水は振り返る。

データセットから得た80個のパーシステント図から、カーネル変化点検出法によって液相とガラス相の転移点を検出したのが図6だ。縦軸は変化の度合いを示し、横軸は温度を表す。35番目あたりの温度で鋭いピークを取り、これが相転移の瞬間だ。

この数値化したベクトルに対してさらにカーネル主成分分析を行い、2次元で表現したのが図7。福水は「変化点検出の結果に基づいて液相を赤、ガラス相を青で色付けしてみると、転移するポイント、つまり液体と固体の境界が明確であることがわかりました」と話す。

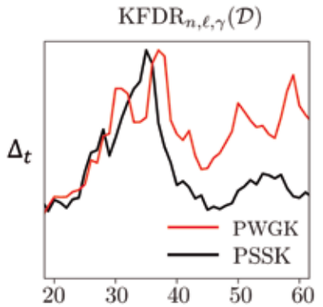


図6：カーネル法による主成分分析の結果。ピークを取るところが変化の最も大きい温度を示している。

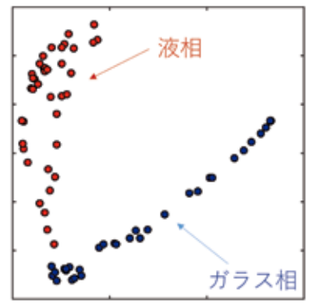


図7：図6で検出したベクトルに対し、さらにカーネル主成分分析を行い、2次元で表現することによって、液相とガラス相の変化点が明確となった。

新材料の開発や難病治療など幅広い分野に応用できる普遍的手法

機械学習のアプローチによって、SiO_2の液相とガラス相の境界がはっきりと示された。このことで、パーシステントホモロジーによるTDAが、階層的で複雑な構造を持つソフトマターの特徴を記述するのに適した手法であることが証明されたわけだ。今後はさらに、物質の電気伝導性や熱伝導率を推定する方法への展開も視野に入ってくる。

「実際に材料を作ってその性質を調べるには、多大な時間と労力とコストがかかります。TDAによってコンピュータ上のシミュレーションで望ましい物質の候補を絞ることができれば、物質探索の大幅な効率化が期待されます」。福水は、このプロジェクトの研究の意義をそのように語る。次なるテーマは、タンパク質など高分子の形状の時間変化をパーシステントホモロジーによって解明していくことなどだ。

パーシステントホモロジーと同様に応用ホモロジーの一種である期待オイラー標数法を研究してきた栗木も、プロジェクトへの展開に意欲を燃やす。これまでに、この手法を用いてイネの遺伝子相互作用の検出などに実績を上げてきた（図8）。「このプロジェクトを通じ、純粋数学や確率の研究者ともダイレクトにディスカッションをする機会を得て、刺激を受けています。自分の研究がどのような形で役立てられるか、熟考しているところです」と栗木は話す。

数学的手法の最大の特徴である「普遍性」は、実社会におけるさまざまな課題解決への可能性を秘める。このプロジェクトの成果は、物性をデータ科学的に予測するマテリアルズインフォマティックスや、ビッグデータ解析への展開も期待されている。（広報室）

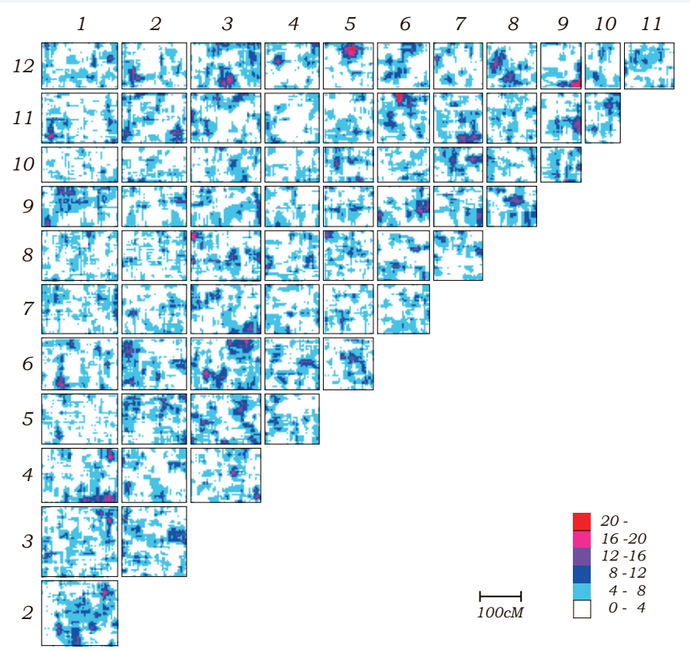


図8：栗木が期待オイラー標数法・チューブ法を用いて、「致死遺伝子」となる二つの遺伝子の相互作用を確認した例。信号の統計的有意性を評価することにより、発見の偽陽確率をコントロールできる。

第9回「地震予測可能性を探る国際共同研究」

地震活動モデルの予測精度向上を目指し、主要地震国と連携

組織的に地震予測可能性を探る国際的共同研究（CSEP）が主要地震国で連携して進められている。各国で地震活動の統計的モデルの開発を促し、確率予測の性能を評価する取り組みだ。地震大国ならではの豊富な蓄積データを活用し、確率予測の精度向上をリードする統数研グループの活動を紹介する。

▲庄建倉准教授

世界規模の共同研究プロジェクトCSEPに参画

2004年に発生したスマトラ沖地震、10年のチリ地震、翌11年の東日本大震災、そして記憶に新しい16年の熊本地震――。世界のどこかが巨大地震に見舞われ、大きな被害が出るたびに、地震予知の精度向上を求める世論が高まる。

そんななか、主要地震国が連携し、組織的に地震予測可能性を探る国際的な共同研究が進められている。それがCSEP（Collaboratory for the Studies of Earthquake Predictability）だ。2001年に南カリフォルニア地震センターがカリフォルニア州の地震活動をモデリングするために立ち上げた共同研究プロジェクトが発展したもので、現在では日本や中国、スイス、イタリア、ギリシャ、ニュージーランドなどを含む世界規模のプロジェクトとなっている。

日本では、統計数理研究所リスク解析戦略研究センターに地震予測解析プロジェクトを組織。モデリング研究系の庄 建倉准教授と尾形良彦名誉教授を中心に、情報・システム研究機構データサイエンス共同利用基盤施設社会データ構造化センターの熊澤貴雄特任助教ら、これまでに延べ約40人が参画している。

地震発生プロセスを確率的に理解する「統計地震学」

「CSEPのプロジェクトで中心的な役割を果たしているのが、統計地震学です」と庄は説明する。

「統計地震学」とは、地震学の諸問題に対処するために用いられる応用統計とは違い、地震発生のプロセスを確率的に理解するための学問であり、物理学で言えば統計力学に相当する。著名な地震学者の故・安芸敬一博士が1956年に初めて使用した言葉だ。

その後1976年に、統数研はこの分野の第一人者であったVere-Jones教授を招聘。それ以来、統数研の統計地震学研究グループは、この分野の展開に大きな役割を果たしてきた。同教授は地震の発生事象を初めてモデリングした立役者だ。「このときに用いられた点過程、すなわち突発する事象を点と捉え、その発生の確率を記述する確率過程は、現在でも統計地震学の主流になっています」と庄。

なかでも、1980年代に尾形が開発したETAS（Epidemic Type Aftershock Sequence）モデルは、地震活動を見る標準モデルとして広く普及し、さまざまな研究仮説を検証するために活用されている。引用論文は2015年の段階ですでに1000件を超えた。

図1：2017年2月にニュージーランド・ウェリントンで開催されたワークショップStatSei10には、主要地震国から多くの統計地震学研究者が集まった。

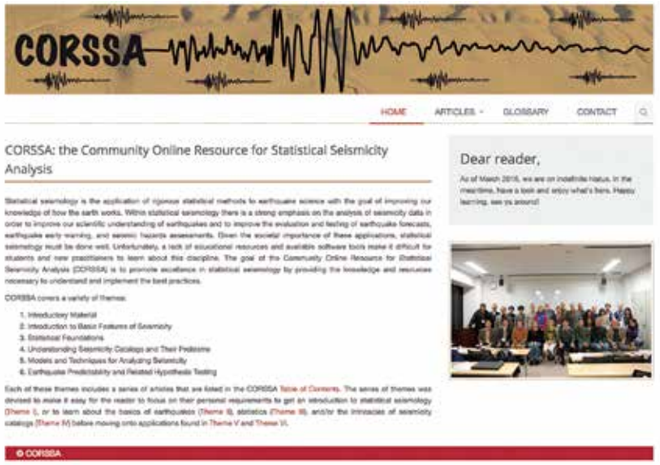


図2：庄ら有志が編集にあたったWeb版教科書「CORSSA Lectures at StatSei」のトップページ。

1998年には、中国の杭州市で第1回の統計地震学国際ワークショップStatSei（The International Workshop on Statistical Seismology）が開催された。StatSeiは、それ以降ほぼ2年ごとに世界各地で開催され、今ではこの分野における代表的な学術集会となっている。今年も2月20日～24日にニュージーランドのウェリントンで、10回目となるStatSei10が行われ、81人の研究者が議論を交わした（図1）。

StatSeiの活動を通じて、次世代の統計地震学研究者も育ちつつある。これに大きく貢献しているのが、Web版教科書「CORSSA Lectures at StatSei」だ。CORSSAはCommunity Online Resource for Statistical Seismicity Analysisの略称。2010年に庄ら有志がスイス連邦工科大学に集まり、編集委員会を組織して編集にあたったもので、2012年に公開し、現在も改訂を続けている（図2）。

地殻内部の様子を総合的に捉える確率予測で地震を予知

正確な地震予知は、世界の地震国にとっての悲願と言える。しかし、その道のりは多難だ。つい半世紀前まで、地震のメカニズムはほとんどわかっていなかったのである。地球の表面が何枚かの岩盤で構成されており、このプレートがマントルに乗って動いているという「プレート理論」でさえ、発見されたのは1960年代後半のことだった。

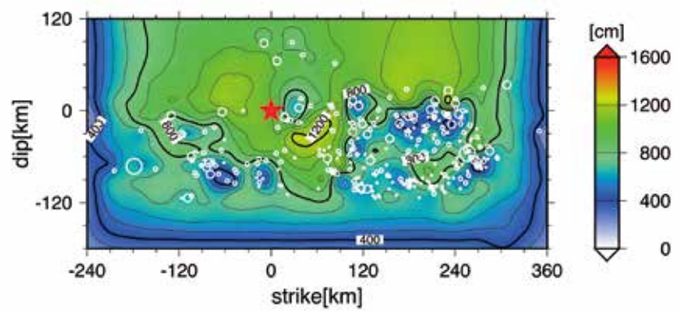


図3：2011年東北地方太平洋沖地震（東日本大震災）の余震活動（図中の白丸）をデータとして推定した、本震発生時の変位の空間分布。赤星印は気象庁による震源位置。

第9回「地震予測可能性を探る国際共同研究」

▲熊澤貴雄特任助教

それ以降、地球科学は目覚ましい発展を遂げ、物理現象の観測技術も向上。データの蓄積が進んだ。大地震が起こるたびに、その仕組みが解明されている。しかしその半面、地震現象の多様さ、複雑さも、ますます明らかになってきた。しかも、地殻内部の断層やストレス状況は、直接的に目視することはできない。データが増えるほど、地震予知の難易度が高まっているのが実情だ。これらを総合的に捉えて未来の地震を予測するには、確率予測の手法を取り入れることは必須なのである。

CSEPが推進しているのは、まさにその確率予測モデルの開発だ。その一環として、統数研では庄を中心とするリスク解析戦略研究センターの研究グループが、「時空間ETASモデル」を開発した。ETASモデルに補正を加え、過去のデータを使って将来の地震発生率を予測するものだ。

ETASモデルは地震の本震と余震の関係のみを示すもので、場所や時間は考慮していない。庄は尾形と共同で、これに断層のずれる向きを加えた（空間への拡張）。また、熊澤は特定の場所で地震の起こりやすさがどのように変化するかをモデル化した（時間への拡張）（図5）。

観測上の欠損を補い、リアルタイムかつ高精度に余震を予測

もう一つの成果は、2013年に提案した東京大学との共同研究による「本震直後の余震のリアルタイム確率予測」だ。

大きな地震が起こった後には、多くの強い余震が発生する確率が高く、被災地に追加的な被害をもたらす懸念がある。強い余震の約半数は本震後1日目に起こることから、防災上、できるだけ迅速に余震発生の確率を予測することが求められる。

ところが、本震の直後には、あまりにもおびただしい数の余震が発生するため、観測網では検出が不可能になってしまう。こうした観測上の欠損を伴うデータからは、強い余震が発生する確率について、即座に精度の高い予測を立てることは難しい。気象庁でも、余震を予測するのは本震の発生後1日以上たってからの場合が多いという。

研究グループが開発したのは、ベイズ統計理論に基づき、欠損を伴う本震後数時間の地震発生データから、実際にはどの程度の数の地震が起きていたかをリアルタイムに推定する統計学的方法だ。

研究チームは、2011年の東北地方太平洋沖地震の余震データを用いて、この予測手法の有効性を示した（図3）。また、庄は米カリフォルニア州で1992年に発生したランダース地震について、本震直後のデータを基に、その後の余震を予測。ランダース地震は、本震の約3時間後に約40km離れた場所でM6.4の余震が起こり、誘発地震の重要性が認識されるきっかけとなった地震だ。

「当面の目標は、このリアルタイム確率予測を防災上の要請に応える形で実用化に結びつけることです」と庄は話す。

標準的地震活動モデルによる異常現象の検出

一方、庄らはCSEPの参加各国からモデルを収集し、その性能を評価する取り組みも進めている。

地震予測のポイントの一つに、前兆となる異常現象の問題がある。観測によって地殻変動など何らかの異常が認められたとき、それが大地震の前兆であるかどうかの識別は容易ではない。予測の精度を高めるには、その異常現象が大地震発生の前兆かどうか、その切迫性はどのくらいかを見積もる必要がある。

観測異常現象に基づく従来の地震予測手法では、ランダムに事象が起きると仮定したポアソンモデルと比較して前兆性を検証することによって正当化する。しかし、地震活動にはクラスター（連鎖）性があることから、完全なランダムモデルによって異常現象と判定されても、実際は正常な連鎖活動の一環である場合が多い。つまり、ETASモデルのシミュレーションなどで再現できる。

そこで、研究グループでは、地震のクラスタリングモデルを用いて正常な活動による“異常”を排除することで、予測手法を改良に導く（図4）。

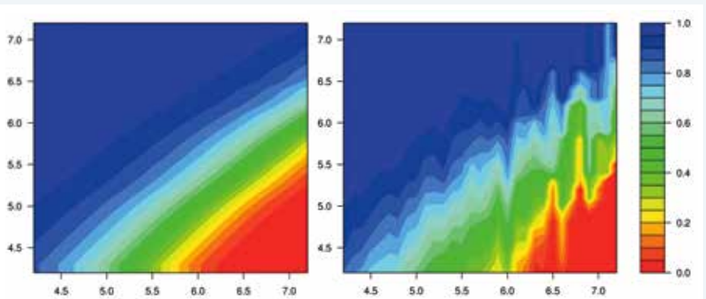


図4：標準的なクラスタリングモデル（左）と日本の地震データから得たモデル（右）による前震確率の比較（横軸はクラスタ中最初の地震のマグニチュード、縦軸は誘発された地震の最大マグニチュード）。

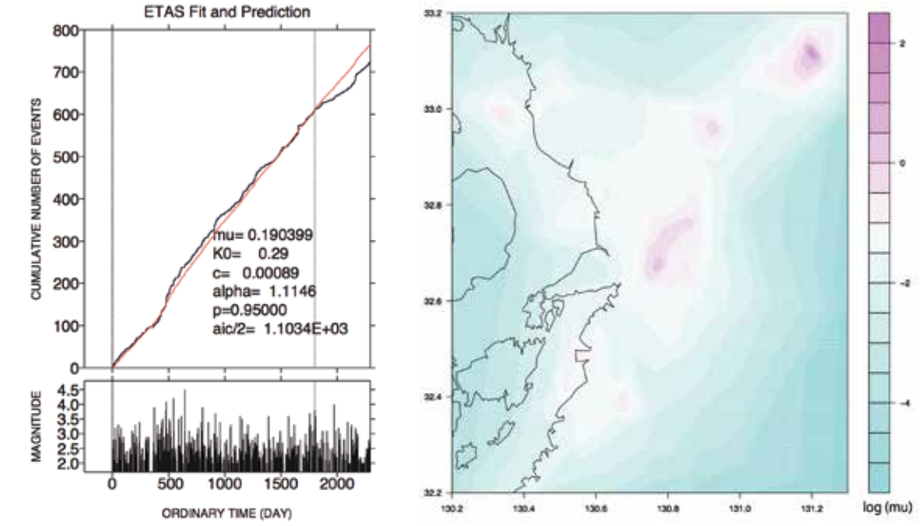


図5：熊本地震の余震域をETAS（左）と時空間ETAS（右）で解析した例。2010年から本震直前までの期間をETASモデルで解析すると、約1年前からの静穏化が有意となる。時空間ETASで同領域の2012年から本震直前までの地震活動を解析し、背景地震活動強度μはM6.5、M6.4を含む前震領域で高い値を取ることがわかった。

地震活動データを用いて地殻内物理情報を統計的に推定

確率予測を向上させる方策は他にもある。その一つが、地殻内部の変位などの物理情報を予測モデルに取り入れることだ。しかし、地表付近に設置する地震計やGPSなどの観測機器では、地中深部の情報を推定するには限界がある。

これを補うために庄らの研究グループでは、稠密な地震観測網から得た地震活動データを用いて、地殻内部の変位を統計的に推定する手法を開発している。得られた結果に、地震活動以外のデータから推定されたものを加味することで、さらに予測精度の高い確率予測を実現することを目指す。

地震国・日本ならではの観測データを生かし、予測精度の向上に貢献

統計モデルによる地震活動の計測、異常現象の定量的研究、そしてそれらに基づく地震発生確率予測モデルとその評価法の確立――。研究グループは、予測の精度向上に向けて、さまざまな角度から着実に歩みを進めている。

統計地震学は分野の歴史が浅く、研究者もまだ多くはない。そんな状況にあって、日本で研究を進めることのメリットを庄は次のように考えている。「日本は世界有数の地震国であり、予測精度の向上は切実な問題。データが取得しやすい環境を生かし、研究に役立てたいと思います」。

高精度な地震確率予測モデルの開発への社会的ニーズは高い。統数研の研究グループは、防災対策を進める行政機関からはもちろん、地震保険を扱う企業などからも大きな期待を寄せられている。（広報室）

第10回「広域撮像探査観測のビッグデータ分析による統計計算宇宙物理学」

統計手法で突発天体を探し出し、宇宙の解明を前進へ

突然出現し、短期間に急激な光度変化を示す突発天体。超新星などの突発天体を発見し、研究することで、宇宙物理学は大きく前進すると言われている。東京大学などと連携したプロジェクトで、統数研チームは突発天体をビッグデータや動画から統計学的手法によって抽出し、宇宙物理学の進展に貢献している。

天文学と統計学の融合による「統計計算宇宙物理学」の創出

ICTやセンサ技術の発達によってもたらされるようになったビッグデータ。その活用により、さまざまな分野でイノベーションが期待されている。天文学や宇宙物理学もその一つだ。

統計数理研究所の池田思朗教授、伊庭幸人教授、小山慎介准教授、森井幹雄特任助教は現在、科学技術振興機構（JST）の戦略的創造研究推進事業（CREST）のプロジェクトに参加し、宇宙の解明に向けた統計学的手法の開発を担っている。今回は池田と森井に取材を行った。

図1：ハワイのマウナケア山頂にあるすばる望遠鏡の外観。

CRESTのプロジェクトは、東京大学大学院理学系研究科およびカブリ数物連携宇宙研究機構（Kavli IPMU）の吉田直紀教授を研究代表者とする「広域撮像探査観測のビッグデータ分析による統計計算宇宙物理学」で、2014年度に採択された。Kavli IPMUと統数研のほか、筑波大学計算科学研究センター、日本電信電話（NTT）コミュニケーション科学基礎研究所が参加している。

内容は、ハワイのマウナケア山頂にある地上大型望遠鏡「すばる」で取得される膨大な画像データを機械学習と統計数理、さらには大規模コンピュータ・シミュレーションを駆使して解析し、宇宙の理解を深め、その謎にせまろうというものである。すばるで取得された画像データは公開されることが決まっているが、それに加え本プロジェクトで得られた知見や技術、データも公開され、社会へ還元される予定だ。

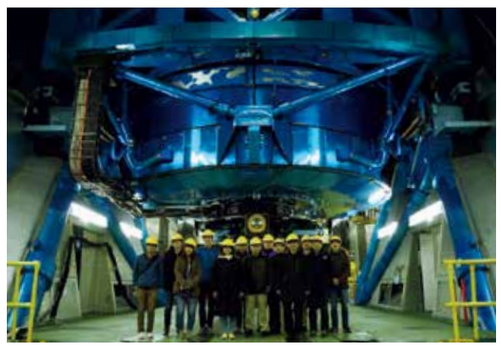


図2：すばる望遠鏡本体の前で。CRESTのチームのメンバーとともに。

さらに、このプロジェクトは、宇宙探査によるビッグデータと情報統計学を融合させた新しい学問領域として、「統計計算宇宙物理学」を創出することを目標にしている。2016年1月にKavli IPMUと統数研は、研究協力に関する協定も締結した。

超新星の遠ざかる速度から宇宙の膨張スピードを算出する

すばる望遠鏡に新たに搭載した超広視野カメラ「Hyper Suprime-Cam（HSC）」で進められる戦略枠観測プログラム（HSC-SSP）では、300夜分、25兆ピクセルに及ぶビッグデータが得られる予定である。CRESTのプロジェクトでは、このデータを用いて、宇宙の謎に迫ろうとしている（図1、2）。

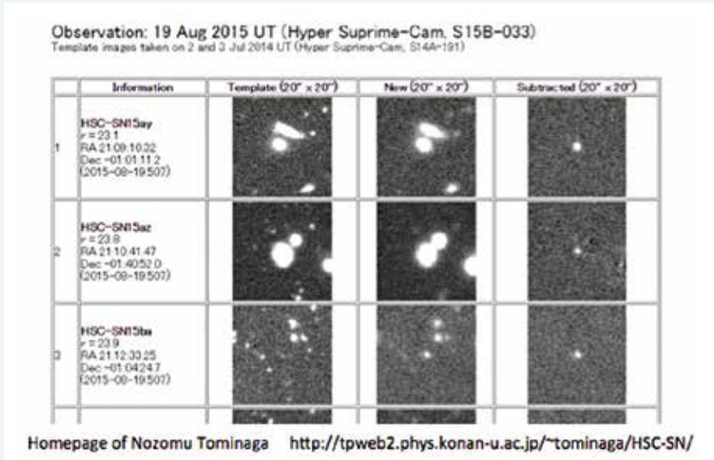


図3：機械学習の手法を用いて選びだした超新星の候補。40万個を超える物体の中から、10個の超新星を選出した。

HSCは、広い視野を持ち、従来の望遠鏡に比べて遠方の天体を写すことができる。「このカメラで同じ領域を何度も撮影すると、前回の撮影時にはなかったものが突然現れることがあります。こうした天体の中には超新星も含まれていて、一晩で超新星が100個ほども見つかる可能性もあります。こんなことは、これまでの観測ではあり得ませんでした」と池田は説明する。

超新星は、いくつかのタイプに分かれる。このうちの「Ia型超新星」は、理論的にどれも同じ明るさであることから、距離を計測する指標として利用できる。すなわち、見た目で明るいほど地球からの距離は近く、暗いほど遠い。

距離を特定したうえで、スペクトル（光の波長）を解析すれば、その星が遠ざかっていく速度を割り出すことが可能になる。これによって、宇宙の膨張するスピードが高精度に計算できるようになり、宇宙の理解が進むはずだ。

大量の超新星候補から本物を選び出す

こうした研究を進めるためには、膨大な情報量を持つHSCの画像データの中から、“本物の超新星”を見つけ出すことが大前提となる。

池田によれば、見分けのつきにくい“ニセモノ”は、本物の約1000倍の数に上るという。一晩に10個の超新星があるとすると、1万個ほどのニセモノの中から探さなければならないのだ。

従来の観測ではそれほど大量の画像データを得られなかったため、本物とニセモノの判別は人の目で行うことができた。しかし、毎晩1万個以上もの物体が写るHSCの画像では、そのようなことは不可能。コンピュータを使って本物の超新星を見つける他に手段はない。

そこで、池田と森井はNTTチームと協力し、機械学習によって超新星を判定するプログラムを開発した。使ったのは、ランダムな推測より少し良い推測をする学習機械「弱学習機械」を組み合わせて、強い学習機械を作る「ブースティング」と呼ばれる手法の一種だ。ここで用いた「AUCブースティング」は小森理客員准教授が、統数研に在籍していた当時に江口真透教授と共に開発したものだ。

まず、専門家が本物と判定した突発天体を教師データとしてマシンをトレーニング。次に、本物のサンプルを増やすため、人工天体を画像に埋め

▲森井幹雄特任助教

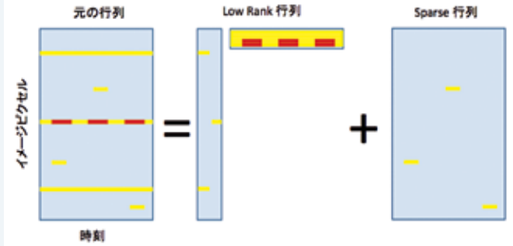


図4：行列分解の概念図。元データを時間によって変化しない「Low Rank行列」と変化する「Sparse行列」に分ける。超新星などの突発天体は、Sparse行列に抽出される。

込んだ教師データを作成し、トレーニングを重ねた。そして、ハワイの天文台にあるPC内のデータプロセスパイプラインに、3台の機械学習のマシンをインストールしておき、リアルタイムに判定を行った。

その結果、40万6121個の対象物の中から10個の突発天体を選出することに成功（図3）。「先行研究と互角の精度を確保しながら、観測したその日のうちに結果を出すという、人の目ではありえない早さでの判定を達成しました」と森井はいう。

超高視野カメラの動画から突発現象だけを抜き出す

統数研チームがCRESTのプロジェクトの中で行った別の研究の成果として、東京大学附属天文学教育研究センターとの共同研究についても聞いた。

同センターが運用する木曽観測所には、口径105cmシュミット望遠鏡がある。現在、同センターは、この望遠鏡に搭載する超広視野高速カメラ「Tomo-e Gozen（トモエゴゼン）」の開発を進めている。計画では、シュミット望遠鏡の直径9度の視野を84台のCMOSイメージセンサで覆い、1秒間に2枚の画像を撮影できる。

その特長を森井は次のように説明する。「トモエゴゼンは19等級までの暗い天体や、時間変動の速い現象も捉えられます。例えば、超新星爆発や高速電波バースト、重力波を出す天体など、短時間にピカッと光る現象です。従来の観測装置では、こうしたものを映像として写すことはできませんでした」。システムが完成すれば、高速で移動する物体の観測や、突発天体現象を発見するために宇宙空間を広く探査するのに威力を発揮する。

完成予定の2018年に先立ち、2015年11月には、8台のCMOSセンサを有する「トモエゴゼン試験機」が完成。同センターはそれ以後、試験観測を続けている。得られた観測データの処理方法の開発に関して、統数研チームも協力している。

監視カメラ映像のデモ動画にヒントを得て応用

すべてのCMOSセンサーが搭載され、高速で撮像をすれば、トモエゴゼンの出力する観測データは、1晩で30テラバイトに及ぶ。そのため、膨大なデータの中から突発天体を効率的に見つけ出すこと、ストレージへの負荷を低減するためにデータを圧縮すること、の二つが課題となる。

池田と森井が目をつけたのは、トモエゴゼンが0.5秒間隔という高頻度で撮影する動画の大部分が時間変化しないことだった。「これほど広視野の天文観測データを動画にするのは、トモエゴゼンが世界で初めてです。この動画の解析には、機械学習の手法として開発された『行列分解』が使えます」と、池田は言う。

行列分解とは、元の動画データを「低ランク行列とスパース行列の和」として考える方法だ。短時間で変化する物体はスパース行列として抜き出され、それ以外の定常的な物体は低ランク行列に集約される（図4）。

行列分解はすでに動画解析で使われようとしている。監視カメラ映像から動いている人物だけを抽出する動画は、業界では有名なデモンストレーションである。トモエゴゼンの動画はまさにうってつけだった。行列分解を用いれば、刻々と変化する突発天体と、その他の定常的な天体とを分離できる。

トモエゴゼン試験機のデータをこの方法で処理したところ、予想どおり、突発天体と、背景となる定常的な天体をきれいに分離することができたという（図5）。

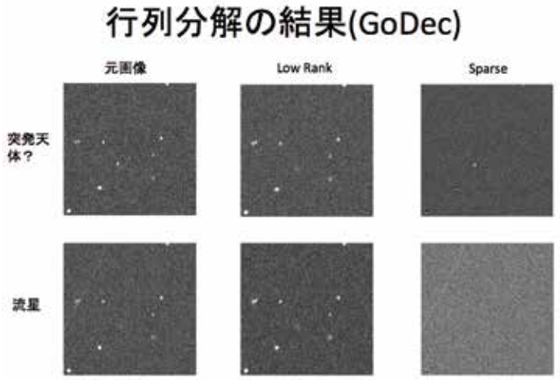


図5：実際にトモエゴゼンが撮影した元画像をLow Rank行列とSparse行列に分解し、Sparse動画から突発現象だけを抜き出した結果。

「この方法のよいところは、突発天体を効率よく抽出できるだけでなく、低ランク行列にすることで、それ以外のデータが10分の1に圧縮されることです」と、森井は説明する。ハードディスクの負荷を減らすことができるわけだ（図6）。

統計手法が天文学や宇宙物理学に新たな発見をもたらすことを期待

すばる望遠鏡やトモエゴゼンの登場が象徴するように、天文学や宇宙物理学の分野では、データの量が爆発的に増え、まさにビッグデータとなった。この流れは、これからさらに加速していくことが確実だ。一方で、コンピュータの進化と、機械学習や統計学など応用数学の進歩が同時に進み、新しいデータ科学の手法が急速に発達している。

そんな状況を見て、池田は「今後、天文学や宇宙物理学において良い成果を上げていくには、データ科学の方法を導入することが必須でしょう」と指摘する。統計学者の立場から、そのための支援をするつもりだ。

天文学出身の研究者である森井を特任助教として迎えたのも、人材育成の一環だ。森井自身も「ここで統計手法をしっかり身に着け、天文学研究に応用していきたい」と語る。

池田はまた、天文学・宇宙物理学研究者を対象とする公開セミナーも積極的に企画している。5月29日から3日間にわたり統数研で開催した研究会では、関連分野からデータ科学を共通項として幅広い話題を提供。およそ90人の若手研究者が参加し、熱い議論を交わした。

「今後、天文や宇宙物理において、統計手法によって見えてくるものが増えることを期待したい」と池田は話している。（広報室）

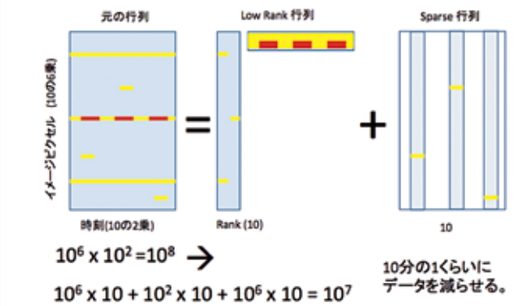


図6：データ圧縮の概念図。全体としてデータ量を10分の1に圧縮できた。

自ら企画した公開セミナーでレクチャーする池田思朗教授。セミナーには、天文学や宇宙物理学の研究者およそ90人が参加した。

第11回「先端医療を進歩させるビッグデータ解析とデータサイエンス」

先端医学研究の発展に向け統計手法で科学的エビデンスを構築

ビッグデータの利用環境整備とコンピュータの性能や技術の向上により、統計手法を活用することで臨床医学研究は大きく進化しようとしている。プレシジョンメディシン（精密医療）やネットワークメタアナリシスの研究をリードする統数研チームの取り組みを取材した。

▲先端医学研究におけるデータサイエンスの研究プロジェクトを担当する野間久史准教授

日本でも臨床研究に資する医療統計家の育成が急務に

医療の分野では、古くから統計学の重要性が認められてきた。先進諸国、なかでもアメリカでは、多数の大学が統計学科とは別に独立した「医療統計学科」を設けて人材育成に注力し、毎年多くのPhDホルダーを輩出している。

そうした人材は産官学で求められ、医薬品の開発や承認審査、環境汚染物質の健康影響評価といったリスクアセスメントなどの業務で幅広く活躍している。医療分野でのエビデンスの評価に際し、データサイエンスは欠くことのできないものとして定着しているのだ。

これに対し日本は基礎医学では世界トップクラスにあるものの、臨床医学においては一流ジャーナルの占有率がけっして高いとは言えない状況にあった。そこで不可欠となる医療統計家の重要性も十分に認識されていなかった。

ところが、製薬会社ノバルティスファーマの元社員によるデータ改ざん事件などをきっかけに、その重要性が注目を集めるようになった。これらの事件の一因として、国内の医学アカデミアに臨床試験のマネジメントや解析を専門とする統計家がほとんど配備されていない現状が浮き彫りになったからだ。

統計数理研究所データ科学研究系の野間久史准教授は「これらの事件などを境に、アカデミアの世界に臨床研究の実施体制を整備しようとする機運が高まりました」と話す。

厚生労働省は日本発の革新的医薬品・医療機器等の開発を推進するため、臨床研究中核病院の指定に乗り出した。そこでは要件の一つとして、2名以上の医療統計家の配置を定めている。

産学官共同での人材育成に統計の専門機関として協力

一方、統計家の育成支援の動きも活発化。統数研は統計の専門機関として、その取り組みに積極的に参画している。

その一つが、AMED（日本医療研究開発機構）と日本製薬工業協会が2017年度から実施している生物統計家育成支援事業への協力だ。

同事業では、東京大学と京都大学の両大学院を育成拠点として選定。製薬企業からの寄附金と国の研究資金を基として、医療統計に関する講座を新たに設置した。連携病院などでのOJT研修も行うという。こうした形での産学官共同プロジェクトは日本で初めての取り組みだ。統数研は逸見昌之准教授らを講師として派遣することとなっている。

またその他、文部科学省の健康科学領域の人材育成事業の一環として、医学部の研究者に統計科学のスキルやリテラシーをトレーニングしていく高度専門教育事業も実施している。

現在、統数研の中に健康科学研究センターを創設する計画が検討されており、関係機関との連携を強化することを目的としてコンソーシアム「健康科学研究ネットワーク」も立ち上げる予定だ。

▲スーパーコンピュータを用いた医療ビッグデータ解析を推進する大谷隆浩特任助教

ビッグデータ解析により展開が進む精密医療

統数研が直接的に取り組むプロジェクトも数多くある。一つは「個別化医療実現のためのビッグデータ解析」がテーマだ。

今、医学の世界では個別化医療への関心が再加熱している。その理由が、従来の個別化医療を大きく前進させる可能性の高い「プレシジョンメディシン（精密医療）」の登場だ。アメリカのオバマ前大統領が2015年の一般教書演説で、Precision Medicine Initiativeを発表したことから注目を集めた。

従来の医療では、患者個々人に対する治療効果や副作用の個人差を十分に予測することができず、疾患や症状において、画一的な医療が行われるのが一般的であった。だが、近年の研究で、ヒトゲノム情報などを基にこの個人差をある程度、予測することができることが明らかになってきた。

プレシジョンメディシンとは、この個人差を的確に予測した、精確な医療を行おうという試みである。ここで中心的な役割を果たすのが、ゲノム情報をはじめとする生体分子情報などの医療ビッグデータの解析だ。大規模なデータを有効に解析し、いかに「大きな治療効果が期待できるサブグループ」を同定するかが、個人の特性に応じた治療法の確立のために重要となる（図1）。

プレシジョンメディシンへの期待が高まったのは、ビッグデータの利活用に関するコンピュータ技術が大きく発展したことや、人工知能の進化やデータサイエンスの方法論の発展など技術面での進歩によるところが大きい。

「どんなにすぐれた医薬品も、万人に同じ効果を与えられるわけではないし、副作用の起こりやすさにも個人差がある。これらを正確に予測できれば、医療

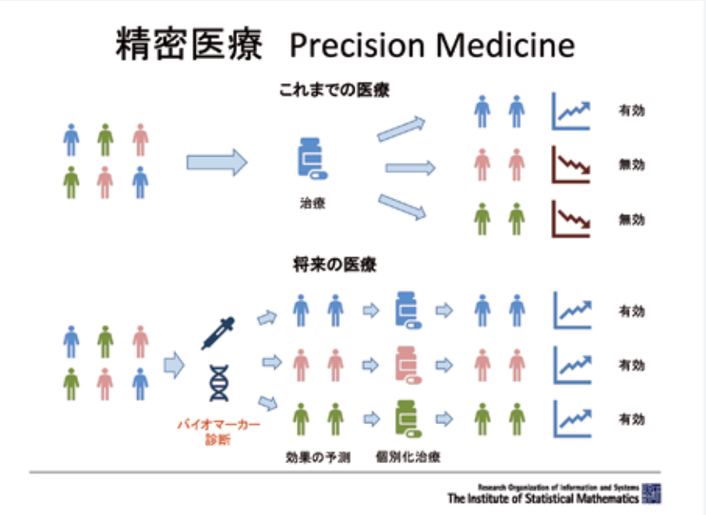
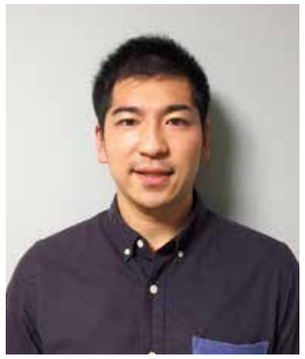


図1：精密医療の概念図。



▲最先端の数理・機械学習の方法論の開発研究に取り組む菅澤翔之助特任研究員

のベネフィットを最大化できます」と野間は説明する。

未解明の医薬品有効性評価が新たな統計手法で大きく前進

最近、統数研の野間らのグループは、名古屋大学・国立がん研究センターなどと共同で、GWAS（ゲノムワイド関連解析）やDNAマイクロアレイを用いて有望な治療効果予測マーカーを精度良くスクリーニングする方法を開発した（図2）。

例として取り上げたのは、多発性骨髄腫へのサリドマイドの有効性評価だ。サリドマイドは、催奇形性の薬害によって1960年代に発売中止になったが、2000年頃からハンセン病や多発性骨髄腫の治療薬として再承認された。

しかし、多発性骨髄腫の臨床試験データでは、無憎悪生存期間の延長には有効であるものの、生存期間についてはプラセボと有意な差が認められていなかった。今回の研究では、DNAマイクロアレイによる数万種類の遺伝子の発現パターンを解析し、生存期間が延びているサブタイプを規定する可能性がある遺伝子が抽出されたという。同様の研究を、抗がん剤や循環器疾患治療薬の臨床試験についても行い、有望な結果が得られている。

サリドマイドの場合はすでに再承認されていたが、有効性が証明できないために認可されず消えていく新薬は少なくない。今後、これらの研究が発展し、有効な患者のサブタイプが特定できれば、そうした薬の承認にもつながるだろう。治療法のなかった難病の患者が救われることも期待できる。

「数万から数百万次元に及ぶ医療ビッグデータを解析するのは臨床医だけではとても不可能であり、データサイエンスの専門家が関与しなければ成功できません。私たちは新薬開発だけでなく、これらのビッグデータを用いた予防医療などに関わる研究もしています」と野間は話す。

医薬品の「価値」を精査するネットワークメタアナリシス

統数研チームが手掛けるもう一つの大きなテーマは、ネットワークメタアナリシスの研究手法だ。従来のメタアナリシスは、対象となる治療法を直接比較した臨床試験の結果だけを統合していた。

これに対し、間接的に得られる結果も併せて分析するのが、ネットワークメタアナリシスだ（図3）。例えば、治療薬AとBを比較した試験と、BとCを比較した試験がある場合、この二つの試験結果から、A対Cの結果を間接的に比較するケースなどだ。間接比較を加えることで、統計的検出力が高くなる。

米国において、オバマ前大統領が2010年頃、総合経済政策としてComparative Effectiveness Researchに膨大な投資を行うと宣言したことは有名であり、ネットワークメタアナリシスはそのための重要なツールとなる。

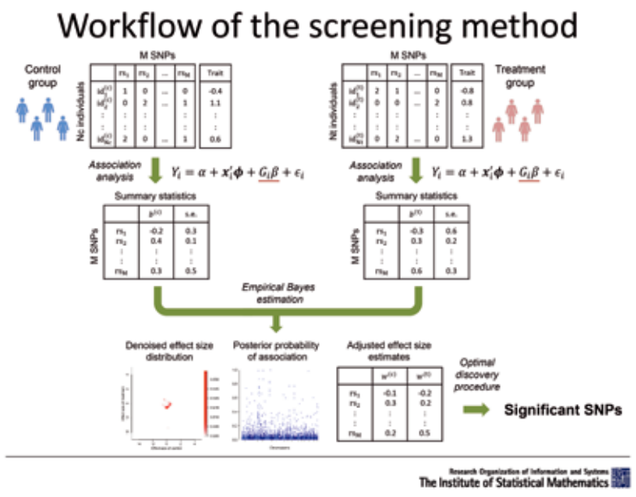


図2：新しい治療効果予測マーカーの検出のための方法論。いくつかの臨床試験の解析を通して、有望な結果が得られている。

いわゆる「コモンディジーズ」と呼ばれる一般的な疾患では多くの場合、作用の異なる複数の治療法が確立している。薬剤についても、古いものから新しいものまで多種多様で、近年はジェネリック薬品の登場もあり医療費が数倍異なるものもある。

しかし、これらの費用の差は、必ずしも実際の治療法の有効性や安全性を反映したものとは言い切れない。なぜなら、こうした複数の治療法を直接比較した科学的なエビデンスが十分に揃っていないからだ。つまり、費用対効果が十分に検証されることのないまま実臨床で用いられているのが実情。

高齢化の波が世界各国に押し寄せ、社会保障費を増大させている今、限られた医療費の効率的な分配は喫緊の課題となっている。それには、それぞれの治療法の優劣に関するエビデンスが不可欠。ネットワークメタアナリシスは過去に行われた臨床試験などのエビデンスを統合し、そのための知見を得ることのできる研究手法として大きく期待されている。

双極性障害の維持療法に有効な治療薬を比較

「ネットワークメタアナリシスで治療の有効性に関する新しいエビデンスを構築する研究において、統数研は大きな実績を上げています」と野間は胸を張る。

その一つが、九州大学・統数研・京都大学・オックスフォード大学などによる国際共同プロジェクト「双極性障害の維持治療薬の有効性と安全性についての比較研究」だ。躁病とうつ病を繰り返す双極性障害の治療には、急性期の治療に加えて間欠期にも維持療法を行うことが重要。従来、これにはリチウムが標準的な治療薬とされていたが、近年になり、抗てんかん薬や抗うつ薬などを用いる療法の有効性が示唆されていた。

そこで国際共同チームは、過去42年間に合計およそ7000人が参加して実施された33件の臨床試験の結果をネットワークメタアナリシスによって分析。17種の治療法とプラセボを比較した結果、有効性ではリチウムが総合的に最も優れていることを確認した。双極性障害の維持治療に関する研究では世界初の成果であり、その論文は国際一流ジャーナルである『ランセット・サイカイアトリー』に掲載された。統数研は、新しい統計手法であるネットワークメタアナリシスの方法論に関する研究でも成果を上げている。

国内外の大学医学部や大学病院、医療研究機関などとの連携に、積極的に取り組んできた統数研。野間は「今後も理論と実践の両輪で研究を進めていきたい」と抱負を語る。

すでにビッグデータ解析では角田達彦東京医科歯科大学教授を研究代表者とするJST CRESTの「医学・医療における臨床・全ゲノム・オミックスのビッグデータの解析に基づく疾患の原因探索・亜病態分類とリスク予測」、ネットワークメタアナリシスでは古川壽亮京都大学教授を研究代表者とするAMEDの「患者特性に応じた薬物療法・精神療法の個別化医療とその臨床試験プロトコルの開発研究」などもスタートしている。

疾病の治療や予防の可能性を大きく広げる新しいデータサイエンス。国の医療費抑制や効果的・効率的な医療のためにも、難病に苦しむ患者や家族のためにも、その活用に大きな期待がかかっている。（広報室）

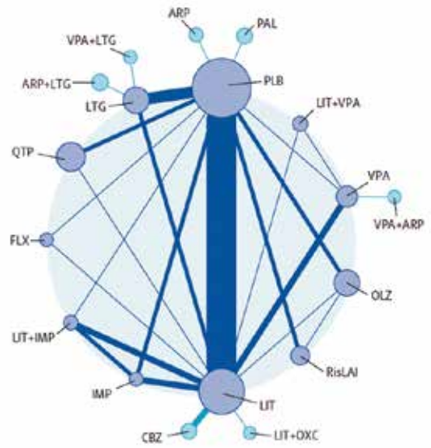


図3：双極性障害のネットワークメタアナリシス。過去に蓄積されたエビデンスを利用して、多数の薬剤の有効性・安全性を比較することができる。

第12回「IR機能の強化と異分野融合指標の開発」

萌芽型研究を支援する異分野融合の新指標を公開 へ

日本の大学や研究機関が真価を発揮し、国際社会で評価を高めるために、研究戦略への情報提供を目的とする「研究IR」が注目を集めている。統数研グループは現在、研究IRに役立つ新指標を開発中だ。萌芽型研究に資する異分野融合を公正かつ適切に評価できる指標はこれまでになく、各方面から期待が寄せられている。

▲本多啓介URA

▲濵田ひろか特任研究員

研究マネジメント改革の武器として注目が集まる「研究IR」

大学内のさまざまな情報を収集・分析し、その結果を大学経営の意思決定や計画策定に役立てるIR（Institutional Research）は、1960年代に米国で始まった取り組みだ。

日本では、国立大学が法人化された2004年頃から主に各大学に設置された評価室においてIR活動が徐々に導入されるようになった。事業成果に対する第三者評価や計画策定時のエビデンスがそれまで以上に求められるようになり、分析データの重要性が高まったことによる。

現在では、学生調査と成績を組み合わせて教育効果を測定し、教育や研究活動に生かす「教学IR」が主体であり、大学組織の活動情報を包括的に収集・分析し、評価、意思決定に活用されている。教学IRは財務分析、学生評価、教員業績評価など大学の教育の質に着目した活動で、私立大学や公立大学にも普及しつつある。

そんな中で、最近特に注目を集めているのが、研究戦略への情報提供を目的とする「研究IR」だ。

近年、日本の大学や研究機関による論文数の国際的シェアは低下傾向にあり、研究体制や研究環境の見直し、研究マネジメント改革などが喫緊の課題となっている。研究IRは、その専門家であるURA（University Research Administrator）の重要な任務の一つでもある。

新指標を開発し、大学IR支援と公募型共同利用のテーマ設定に活用

統数研では現在、URAが中心となり、大学・学術機関のIR活動を支援するツールの開発や機関のさまざまな活動を客観的に評価するための新たな指標に関する研究を行なっている。

その一つが、「IR機能強化と異分野融合指標の開発」だ。このプロジェクトの第1段階のゴールは、統計数理の知見を生かして異分野融合の進展や効果を公正かつ適切に評価する指標を開発し、これを公開して全国の大学などのIR機能強化に貢献すると同時に、統数研の公募型共同利用・共同研究におけるテーマ設定等の意思決定に有効活用することだ。

統数研の新指標開発グループは、本多啓介URAをはじめ金藤浩司教授、中野純司教授、濵田ひろか特任研究員、武井美緒特任技術専門員、水上祐治客員准教授（日本大学准教授）、総合研究大学院大学の張菱軒氏、海外から台湾・中央研究院統計科学研究所の潘建興副研究員、ノースカロライナ州立大学のStephen Porter教授の多彩な人材で構成。福岡女子大学の藤野友和准教授を代表とする日本計算機統計学会スタディーグループと協力して指標開発に当たるとともに、ユーザー側である大学評価コンソーシアムとも連携し、ニーズの把握に努めている。

2016年度からは「学術文献データ分析の新たな統計科学的アプローチ」が公募型共同利用の重点型研究に採択され、16年度には10件、17年度には15件の課題について、全国の研究者やIR実務者、URAと共に多角的な取り組みを続けてきた。この重点型研究に採択された課題では、統数研とクラリベイト・アナリティクス社（旧トムソン・ロイター社）との協力関係により、研究者・研究評価担当者に広く利用されている書誌データベースの一つであるWeb of Scienceのデータが分析や評価指標の開発に活用できる。

萌芽型研究への助成には異分野融合の適切な評価が不可欠

統数研グループが開発中の新指標の最大の特徴は、異分野融合の進展や効果を公正かつ適切に評価する『多様性指標』を用いた評価・分析を目的としていることだ。こうした指標はこれまで存在しなかった。

統数研の助成事業である公募型共同利用は、研究者の自由な発想に基づくボトムアップ型・萌芽型研究を支援するものだ。「イノベーションの源泉の一つとなるこうした研究への助成を強化するには、まず、新たな指標を開発することが必要不可欠です」と本多は断言する。

科学技術・学術政策研究所（NISTEP）では「サイエンスマップ」、米国立衛生研究所（NIH）は「相対引用率（Relative Citation Ratio：RCR）」といった指標をそれぞれ開発し、提案している。個々の論文や学術雑誌の引用数に基づく研究の「インパクト」をベースとするこれらの指標は、トップダウン型研究を助成する機関には有用と言える。だが、萌芽型研究における異分野融合などの評価・分析には向いていない。

その理由を本多は「既存の指標は学術分野間の偏りを完全には標準化できていないからです」と説明する。例えば、二つの論文の引用数がともに100件であったとしても、膨大な数の論文が世に

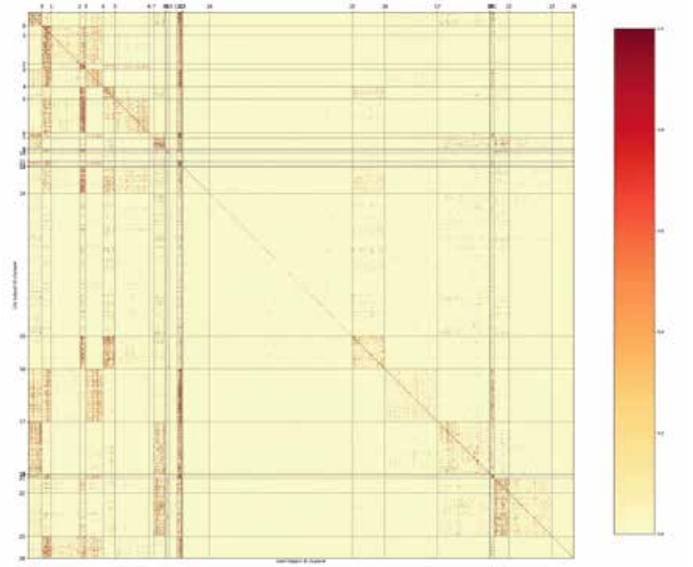


図1：確率的ブロックモデルによる引用ー被引用関係データ行列のクラスタリング結果

第12回「IR機能の強化と異分野融合指標の開発」

▲水上祐治客員准教授（日本大学准教授）

出ている分野と、ほとんど論文の出ていない分野では単純にスコアを相対比較することは不可能だ。

そこで、新指標の設計にあたっては、多様性のスコアを「論文単位で」かつ「書誌情報だけで」算出することができ、分野間の偏りを適切に補正できること、さらに、中長期的な影響を測定できることを要求仕様とした。「これを満たすには、高度な確率論や統計手法を用いる必要があります。統数研だからこそできる取り組みです」と本多は話す。

開発中の新指標は、意思決定サイクルにおける「評価」と「戦略」を担うIRへの導入を想定したものだ。例えば評価フェーズでは、自機関の研究活動の特徴や強み、弱みなどをシステムに問い合わせると、「A年からB年において、X分野とY分野の異分野融合が進みました」、あるいは「国外研究機関との共同研究が減少しました」といったレポートが出力され、自機関の状況を客観的に把握することができる。

一方、戦略フェーズでは、共同研究に向いた機関や異分野融合を促進すべき分野、ある分野で交流・連携・協力などの関係が広い研究者などを問い合わせる。すると、「C研究所がこの分野で最も広い交流・連携・協力などの関係を持ちます」といったレポートが出力され、自機関の積極的な戦略立案支援に役立てることができる。

グラフ構造化で学術分野を再構成 共起頻度を考慮して偏りを補正

統数研グループが着目したのは、「文献同士の関係性」だ。論文などの学術文献には、「引用ー被引用関係」や「共著関係」といった関係性が存在する。つまり、本質的にグラフ（ネットワーク）構造である。「異分野融合の度合いを測るには、学術文献をグラフ構造のデータとして表現し、その特徴を抽出することが重要だと考えました」（本多）。

そこでまず、1981年から2016年までのグローバルな論文の引用ー被引用関係を行列の形に変換。次に、この行列を関係性の近いものは近く、遠いものは遠く配置したブロック構造に並べ替える。「確率的ブロックモデル」と呼ばれるアルゴリズムによるクラスタリングの手法だ（図1）。

こうしてクラスタリングした一つずつのブロックを「潜在的な学術分野」とみなす。キュレーターによる従来のタグ付けではなく、引用ー被引用のデータのみに基づき、学術分野を新たに分類しなおしたわけだ。

ただ、このままでは「論文数の偏り」の問題が残っているため、補正を加える必要がある。これには、自然言語処理で使われる「自己相互情報量（Point-wise Mutual Information: PMI）」のアルゴリズムを用いた。

PMIは、同一文書内に現れる二つの単語が共起する確率を表す指標だ。英語のthisとisのように、どの文書にも頻出して現れる単語自体の出現確率を差し引いて算出する。統数研グループは、これを論文データに適用することで標準化を図った。すなわち、データの多い医学系論文などの出現確率を差し引いたうえで、別の論文との引用関係の確率を算出するように設定。これにより、満たすべき仕様をすべてクリアしたことになる（図2、図3）。

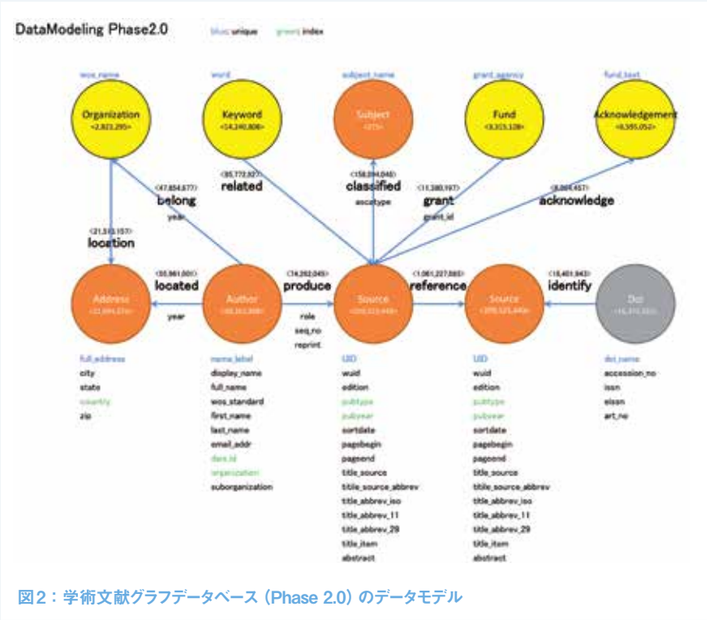


図2：学術文献グラフデータベース（Phase 2.0）のデータモデル

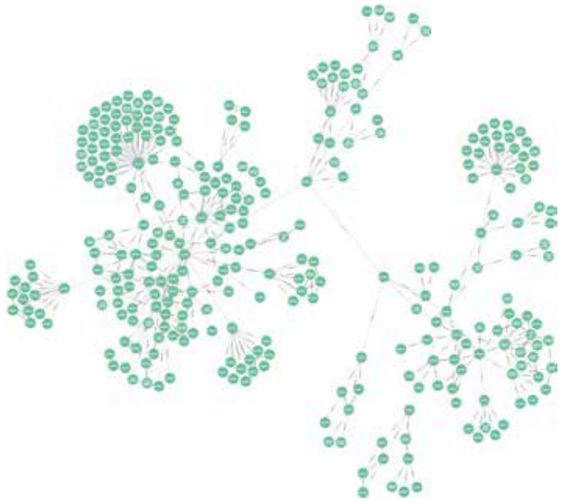


図3：学術文献グラフデータベースから切り出した論文ネットワークのサブグラフ

あとは、補正後のブロックモデルを使い、引用関係にある論文が含まれるクラスター同士の距離を比較したり、平均値を算出したりすれば、論文の影響度、つまり多様性の度合いを評価する指標が完成する。「現在はその算出方法を検討している段階です」（本多）。

公開が待たれる新指標「REDi」 多様性評価による研究の進化に期待

完成間近の新指標は、「REDi（Research Diversity Index）」と名付けられた（図4）。2018年夏に公開する予定だ。2017年12月には統数研が主催する評価関連では初の研究集会「Research Metrics Workshop 2017」を開催し、成果を発表した（図5）。また、2018年1月には、日本オラクルの主催する「Oracle Research Platform サミット@つくば」で大学・学術機関向けに発表した。

2018年度の公募型共同利用の重点テーマでは「IRのための学術文献データ分析と統計的モデル研究の深化」と、IRを強く意識した名称に変えて研究を公募した。

本多は、プロジェクトの次なるゴールを「大学評価システムへの新指標の実装」に設定している。「5件以上の大学・機関が試験的に運用することにより、新指標の有効性を検証していきます」と話す。

新指標開発に携わる水上は「アジアの大学は、欧米の指標では低く評価される傾向があるように思います。われわれの開発した指標によって正当に評価されるとしたらエキサイティングです」と先を見つめる。一方で濵田は、「数値に現れにくく既存の指標では評価できない研究のすごさや面白さを可視化することに、大きな意義を感じます」とやりがいを感じている。

まだ産声を上げて間もない研究IR。統計数理を駆使した新指標の誕生によって、日本の大学や研究機関がそれぞれの強みを活かした研究活動を進展させることが期待される。（広報室）

REDi - Research Diversity Index -

図4：研究多様性指標活用ウェブサイト用にデザインした新指標のロゴマーク

図5：Research Metrics Workshop 2017の様子

第13回

「高度信用リスク統合データベースコンソーシアムと損失モデル」

「いつか起こるかもしれない金 融危機」に備えるDBと損失率モデル

金融システムの安定には、銀行が貸付金を回収できないリスクを管理することが何よりも重要だ。その「信用リスク」計算の鍵を握るのが、デフォルト確率（PD）とデフォルト時損失率（LGD）。だが、デフォルトした企業からしか得られないLGDは、データを集めるのが困難でモデル化が遅れている。統計数理研究所は、複数の地方銀行のコンソーシアムを設立し、10年以上かけて蓄積したデータをもとに、世界で初めてのデータ統合LGDモデルを作成した。

▲山下智志副所長

信用リスクを決めるデフォルト確率と損失率

世界を揺るがせたリーマンショックから10年。当時、連日のようにマスコミを賑わせた「信用リスク」という言葉は、景気が回復したいまでは、ほとんど聞かれなくなっている。

信用リスクとはいったい何か。端的に言えば、銀行が企業に融資した資金が、返済されない可能性のことだ。統計数理研究所リスク解析戦略研究センター長でもある山下智志副所長は、「信用リスクの発生要因は二つあります」と説明する。

一つは、融資先の企業がデフォルトするかしないか。デフォルトは、法的な倒産には至っていないものの、通常の融資は困難な状態。銀行はデフォルト確率（PD）、すなわち融資先がデフォルトする確率がどのくらいあるかを管理しておく必要がある。

もう一つは、デフォルト後に債務を返済する能力があるかないか。融資先が仮にデフォルトしたとしても、担保を売却するなどして貸し付けた金額が戻ってくるならば、銀行にとって損失にはならない（図1）。従ってデフォルト時損失率（LGD）、すなわち銀行側が貸付金をどのくらい「回収できないか」を予測しておくことが非常に重要だ。

金融機関から見た企業の信用リスクはこの二つの要因によって決まり、貸出額とPDとLGDの掛け算が「損失」となる。

「信用リスクを精緻に計算することの重要性は以前から認識されていましたが、それが決定的になったのは、『バーゼル合意』が改定されて、バーゼルⅡになったときです」と山下は話す。

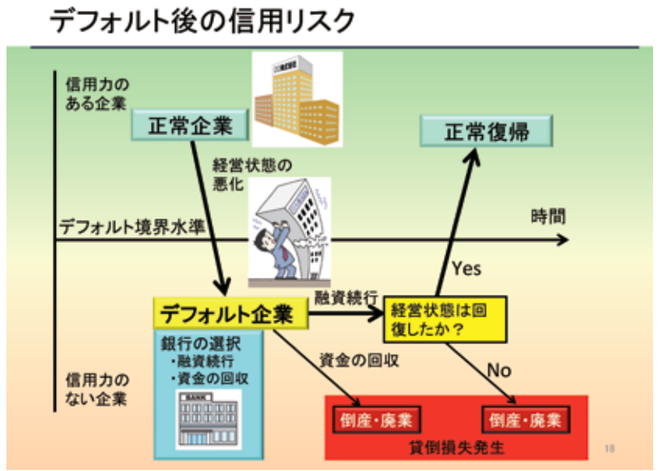


図1：デフォルトしても経営状態が回復すれば正常企業に復帰できる。貸し倒れ損失が発生するのは、デフォルト後に資金が回収できないまま倒産・廃業に至った場合と、融資を続行しても経営状態が回復せず、倒産・廃業に至った場合。

バーゼルⅡを受けて銀行にリスク計算を推奨

バーゼル合意は金融のグローバル化に伴い、金融システムの安定化と競争の公平性を維持する目的で、1988年に初めて成立した。銀行の持つ信用リスクに焦点を当て、リスクアセットの8%相当の自己資本を保有するなどの国際基準を定めたものだ。

日本はこれを受けて2007年3月からは、国際取引を行う銀行に対し、統計モデルを作成して自行の信用リスクを計算するよう強く推奨した。これをバーゼルⅡという。

こうしてバーゼルⅡの施行を前に、全国の銀行が一斉に統計モデルの作成に乗り出した。

しかし、複雑な要素が絡み合う信用リスクを精度よく算出できる統計モデルを作り上げるには、ファイナンス統計の専門家の関与が不可欠。そこで、金融庁の特別研究員を兼務しバーゼル委員会の評価担当でもあった山下を中心とする統数研チームが、銀行の課題に対応することになった。

山下は中小企業信用リスク情報データベース（Credit Risk Database）整備事業を行うCRD協会に設立時から協力しており、同協会には毎年約160万件の企業のデフォルト情報を含む決算データが蓄積されている。日本の中小企業に関する信頼できる最大のデータベースとして認知され、CRD協会が作成するモデルは銀行や信用保証協会などの政府機関などにも利用されている。

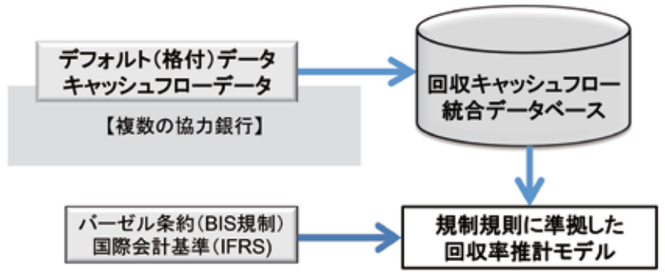


図2：高度信用リスク統合データベースコンソーシアムプロジェクトの活動。

取引先のデフォルト後データを持ち寄るコンソーシアム

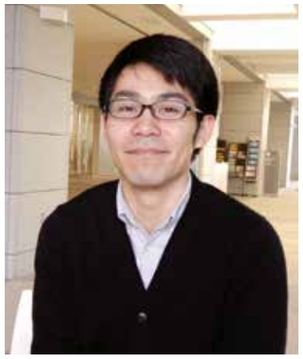
PDを計算するモデルは、CRD以外にも日本リスク・データ・バンクの提供するRDB、全国地方銀行協会のCRITSがすでに世に出ている。残る課題は、LGDを計算するモデルだ。

「PDを計算するのは比較的簡単で、CRDでも5種類ほどのモデルを提供しています。しかし、LGDをきちんと計算できるモデルは、一部の大銀行内部で用いられている簡便なモデルを除きどこにも存在しません」と山下は明かす。

なぜか。まず、データベースがない。というのは、デフォルトしていない状況で計算するPDではすべての企業が対象となるのに対し、LGDの場合、デフォルトした企業のデータしか利用できないからだ。

日本企業のPDは1%程度なので、その時点でデータ数が100分の1になってしまう。このため、一つの銀行だけでは十分なデータを得ることができない。銀行同士でデータを持ち寄ることができればいいが、LGDは各行の営業成績にも関わる秘匿性の高い情報であることから、それも難しいという。

しかし、日本にLGDのデータベース



▲田上悠太特任助教

や統計モデルがないのは問題だ。複数の銀行からデータ提供を受け、統合データベースをつくれば、分析が可能になる。この課題に取り組むことができるのは、ファイナンス統計学の専門家を擁し、しかも中立的な立場の学術機関である統数研が最もふさわしい。そこで山下は「高度信用リスク統合データベースコンソーシアムと損失モデル」のプロジェクトを立ち上げた。

地方銀行をメンバーとするこのコンソーシアムの主なミッションは、加盟銀行の取引先のLGDデータを統合したデータベースを構築することと、これを基にオリジナルのLGDモデルを作成することだ（図2、3）。CRDモデルを、各行共通モデルを前提としたレディーメイドとすれば、ここで扱うのは銀行の個別戦略を反映したモデル、いわばオーダーメイドのモデルということになる。

2007年に1行からスタートし、現在では5行がこのコンソーシアムに参加。統数研からは山下のほか、現在は田上悠太特任助教がこのプロジェクトを担当している。

銀行ごとに異なるニーズや定義のすり合わせがポイント

LGDは、データベース構築の難しさもさることながら、モデルの作成にも困難な問題がつきまとう。

「LGDは、担保や保証の定義によって、計算の仕方が大きく変わります。このためモデル作成にあたっては、高度な統計理論の知識と、金融の商習慣についての知見の両方が求められるのです」と山下は言う。

例えば、債権回収の仕方は銀行によってさまざまであり、回収にかかる期間も大きく異なる。地域密着で地元企業をじっくり支援する方針の銀行では、デフォルトしてから回収が終わるまで5年ほどかかるケースも少なくない。つまり、すぐデータを取れるデフォルトデータと違い5年後でなければデータが取れない。

また、「デフォルト後の債権回収について何を知りたいか」は銀行ごとに異なり、しかも最初はそれが明確になっていないことが多い。通常、統計に使う関数y=f(x)では、左辺yは決まっていて、右辺のf(x)をどうするかが問題になる。しかし、この場合、左辺yをどう定義するかをまず決めなければならない。そこに難しさがあるのだという（図4）。

3カ月に一度開催するコンソーシアムの会合では、加盟する銀行の担当者が統数研に集まり、その定義について2年にわたり議論を続けた。

価値が目に見えにくくともなくてはならない社会の防波堤

コンソーシアムを立ち上げてから10年がたち、LGDデータの蓄積は着実に進んでいる。すでにLGDモデルのプロトタイプも完成した。山下は「5行のLGDデータを統合したモデルは世界初だと思います。LGDに関しては、国際的に見ても日本がトップと言えるでしょう」と胸を張る。

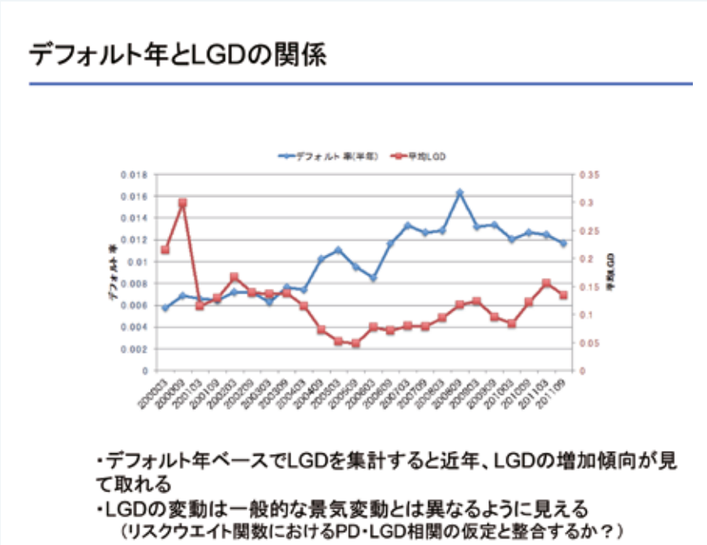


図3：データベースの集計例。

山下は信託銀行を経て統数研に入所し、一時期は金融庁の特別研究員としてバーゼル委員会の委員を務めたことがある。その経験を踏まえ、日本社会に対しこれまで一貫して、LGDの重要性を提唱してきた。

「銀行の損失は貸出額、PD、LGDの三要素で決まるにもかかわらず、日本ではPDばかりが取り沙汰され、計算の難しさもあってLGDには手を出さない地方銀行が多いのです」。山下は、そう指摘する。

大学の商学部で統計技術の重要性を感じた田上は、総合研究大学院大学で山下の門を叩き、修了後に統数研の特任助教となった。「このプロジェクトに関わってみて、改めてリスクマネジメントにはデータサイエンスや統計の視点が欠かせないと実感しています」と話す。

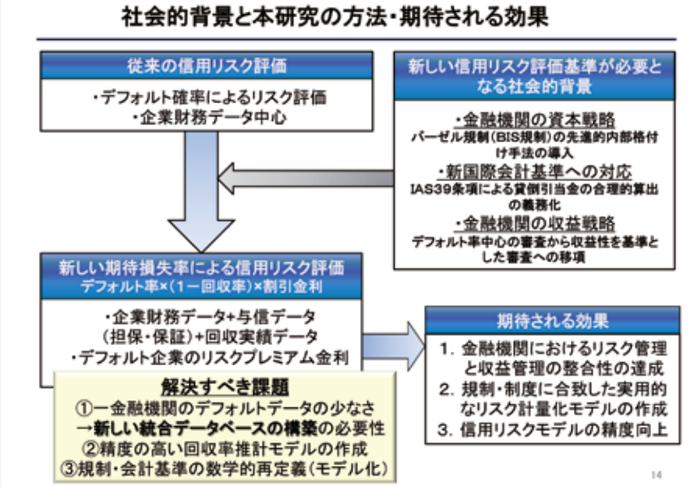


図4：金融のグローバル化などにより、新たな信用リスク評価基準が求められている。

いまは景気がよく、企業倒産が少ないことから、信用リスクへの銀行の関心も薄れがち。いつ訪れるとも知れないリスクに備えて収益を逃すことを避けたいという風潮がある。しかし、どんな経済状態のときでも、金融危機が起こる確率は一定程度存在する。何がリスクファクターになるかは起こってみなければわからないが、トリガーはいたるところ潜んでいるのだ。

リスク研究は、リスクが眼前に現れない限り価値が伝わりにくく、その意味では報われにくい。けれども、こうした研究はLGDデータがいったん途切れたら、ゼロからやりなおしになってしまう。だから諦めるわけにはいかない。

メンバーの銀行からニーズをていねいに吸い上げ、提供されたデータから得られた知見をきちんとフィードバックするなど、コンソーシアム全体のモチベーションを維持する幹事役を担うのは田上だ。「あまり目立つ分野ではないですが、世の中のためになるように頑張りたいと考えています」と思いを語る。

「企業が短期的な視点で収益を追うのは仕方がない面もあります。ただ、国全体がそれに追随すれば、もしものときにどうなるか。われわれには、そのための防波堤をつくっておく役割があると思っています」と山下も言う。

統計を使って、日本経済を見えないリスクから守り、国を支える――。統数研ならではの存在意義が、静かな光を放つ。（広報室）

第14回「マルチモーダル生体信号データの時空間解析プロジェクト」

生体のメカニズムを時空間データの解析で解き明かす

今年4月、統計数理研究所に新たなセンターが誕生した。医学・健康科学領域における先進的なデータサイエンスの研究・教育を推進することを目的とする医療健康データ科学研究センターだ。ここでは、研究プロジェクトの一つ「マルチモーダル生体信号データの時空間解析プロジェクト」の概要とともに、同センターの方向性を紹介する。

▲伊藤陽一医療健康データ科学研究センター長

統計教育・研究支援とデータサイエンスの高度化が使命

医学の世界で、科学的根拠に基づく医療の実践（EBM：Evidence-Based Medicine）が提唱されはじめたのは1990年代のこと。以降、その普及とともに、医療統計学の重要性はますます高まっている。

ところが、日本の医学アカデミアでは統計教育の体制づくりが遅れ、統計学の専門家が極端に不足しているのが実情だ。例えば、海外の医学ジャーナルに論文を投稿した日本人研究者が、レビュアーから統計解析を求められたものの対応ができないというケースも少なくない。このままでは、臨床研究における日本の情報発信力の低下にもつながるとの懸念がある。

こうした状況を受けて今年4月、統計数理研究所に新たに設立されたのが、「医療健康データ科学研究センター」だ。組織体制は9人の所内教員と、12人の客員教授、7人の客員准教授、2人の外来研究員からなる総勢30人。センター長には、医学統計の第一人者である伊藤陽一教授が迎えられた。

「現代の医学アカデミアで行われる臨床研究・臨床試験では、高度な統計解析手法を用いたデータ解析が必須なものになっています。先端医学研究の国際競争をリードするためにも、統計教育や研究支援体制の整備と、データサイエンス研究の高度化は喫緊の課題。当センターは、この2点を最大のミッションとしています」。伊藤センター長はそう現状を見据え、使命を語る。

教育面では、全国の大学の医療・健康科学分野の研究者を対象としたデータサイエンスの教育コースと公開講座を実施する。初年度の2018年度は、生物統計学基礎、疫学・公衆衛生統計、臨床研究統計、生体データ時空間解析の全4コースを開講。公開講座は前期に2講座を予定しており、すでに5月に実施した「臨床研究・疫学研究における傾向スコアを用いた統計解析」は、受け付け開始後すぐに満席となるほどの人気を博した（図1）。

図1：医療健康データ科学研究センターが5月31日に開催した公開講座は受講者の熱気に包まれた。

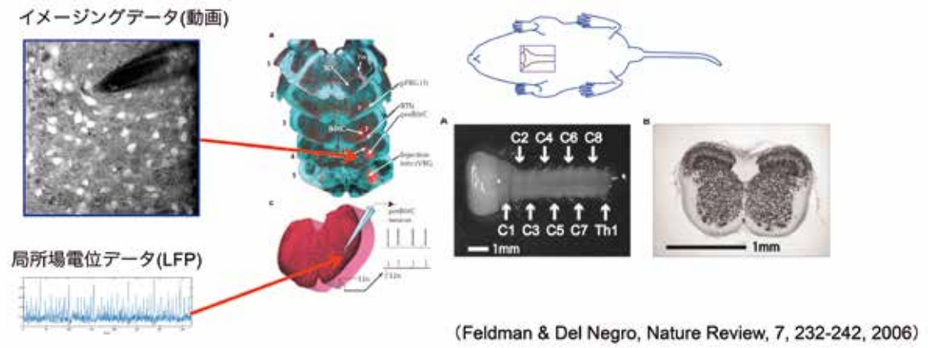


図2：脳幹の呼吸関連部位からスライス標本を作成し、イメージングデータ（動画）と局所場電位データを取得する。

呼吸を司るニューロンのネットワーク機構を解明

三分一（みわけいち）史和准教授をプロジェクトリーダーとする「マルチモーダル生体信号データの時空間解析プロジェクト」は、同センターの推進する六つの研究プロジェクトの一つ。複数の計測方法による同時計測（＝マルチモーダル）によって得られた高精度の生体信号や画像を、時系列的変動と空間的変動の両面を考慮して解析するものだ。その研究成果は、生体メカニズムの解明や生体の状態の把握と予測などに生かされる。

このプロジェクトの中で、三分一が兵庫医科大学、ドイツのゲッティンゲン大学と連携して取り組んでいるのが、「呼吸リズムを形成するニューロンネットワーク機構の解明」。動物が呼吸をするとき、脳幹のニューロンがどのような機構で動作し、呼吸を作りだしているかを明らかにする研究だ。

計測には、「カルシウムイメージング法」を用いる。マウスの脳幹を600ミクロン程度の厚さにスライスした標本に蛍光色素を添加し、ニューロンの活動による細胞内カルシウム濃度の変化をイメージングデータ（動画）で記録する方法だ。

三分一は説明する。「起きている間は呼吸を意識的にコントロールすることもできます。しかし、眠っている間を含め通常は脳幹でニューロンが自律的にネットワークをつくり、無意識的に呼吸リズムが形成されています」

だが、スライス標本の動画だけでは、細胞がいつ自発的な呼吸を引き起こす活動をしているのかを特定できない。そこで、標本に電極を置き、数十～数百個の細胞の局所場電位データを取得する。波形の中で、バーストの起こっている時点が、細胞が同期して呼吸活動を引き起こす指令を出すタイミングだ。そのときの画像を解析すれば、呼吸時のニューロンの活動の様子が分かる（図2）。

この実験のため、三分一は共同研究者で生理学が専門の兵庫医科大学尾家慶彦助教と共にゲッティンゲン大学へ赴いた。同大学は細胞種ごとに異なる蛍光特性を示す特殊な遺伝子改変マウスを保有しているからだ。

事前処理のブレ補正やノイズ除去、視覚化にも統計ノウハウ

こうして実験は無事に終わり、データは得られた。ただし、解析には、画像データ

第14回「マルチモーダル生体信号データの時空間解析プロジェクト」

▲三分一史和モデリング研究系准教授

から呼吸に関係するニューロンを検出しなければならない。「じつは、ここからがデータの事前処理の長い道のりの始まりでした」と三分一は振り返る。

まず、動画のブレ補正が必要だ。スライス標本の撮影時には、灌流液の流動やドリフトと呼ばれる顕微鏡ステージ自体の動きによって、観察面が動いてしまう。X方向とY方向、さらに回転が加わるズレが生じることから、画像マッチングの最適計算により、画像を変換しなければならない。

どの画像を補正の基準とするか、アルゴリズムをどうするかなどの判断には、本来の専門ではない画像解析のノウハウを駆使した。また、数値解析ソフトにバグが見つかり、開発元とともにその解消にあたるなど、予想外の事態にも見舞われた。「3年ほどかかって、ようやく補正の全自動化ができました」と三分一は言う。

ブレ補正の次は、ノイズ除去の作業だ。呼吸バーストの周波数帯域をフィルターとして、電位データから呼吸以外の信号（ノイズ）を除去する。同時に、画像からもニューロン以外のノイズを除去。これには、画像上に窓を移動させ、枠内のデータを平均の点に置き換えることで平滑化する「移動平均化」を組み合わせ、その他最適化プロセスにも様々な工夫を施した。これが、「時空間フィルタリング」の手法だ。

こうしてノイズを除いた画像データからニューロンを検出し、さらに統計的処理により呼吸に関連する興奮性ニューロン、グリシン抑制性ニューロン、GABA抑制性ニューロンの3種類のニューロンを特定する。「これら3種のニューロンはさらに活性化特性により複数のタイプに分類されます。それらがどのように呼吸リズムの形成をしているのか、いくつかの説はありますが、まだまだ未解明な点が多い。実験的なアプローチと統計学的なアプローチの協働として研究を行っているところです」（三分一）。

前述のように、実験に使ったマウスは細胞種ごとに異なる蛍光特性を示す。動画を1枚ずつの静止画として切り出した上で、それぞれの画像を2値化し、重ね合わせることで1つの画像に3種の分布を表現する。これら一連のプロセスは三分一がこのイメージングデータ解析のために開発した事前処理と視覚化法だ。

統数研伝統のテクニックを駆使しニューロン間の因果性を探る

解析ではまず、事前処理で得られたすべての活動性ニューロンの中から、呼吸性ニューロンを抽出する。これには、局所場電位データと個々の細胞の活動を示す細胞内カルシウム濃度変化のデータとの間で相互相関解析を行い、統計的に有意なニューロンを検出した。

そしていよいよ、検出された呼吸性ニューロン間の因果性を推定する。まずは手始めに代表的な時系列モデルである自己回帰(AR)モデルと外生変数型自己回帰（AR-X）モデルを特定のニューロンに適用し、他のニューロンからの因果性を評価した。AR-Xモデルは、あるニューロンyに対する自己回帰モデルに、他のニューロンxからの入力を考慮したモデルだ。もし、AR-Xモデルの方がARモデルより当てはまりがよければ、他のニューロンxからの影響があると解釈することができる。

「時系列モデルの当てはまりの良さの評価には、統数研の元所長である故・赤池弘次先生が考案した赤池情報量規準（AIC）を使いました。これは、統数研の伝統的なテクニックです」と三分一は言う。これにより、ターゲットとしたニューロンとその他の活動性ニューロンの2体間の因果性を推定し、視覚化することに成功した（図3）。

AR-Xモデルでは、1個ずつのニューロンの関係性を明らかにした。三分一は、今後さらに、すべてのニューロンを一度にモデル化する多変量ARモデルへと拡張するという。その過程にはまた様々な工夫とノウハウが必要であるが、モデルのパラメータを調べることで、ニューロンのネットワーク構造が解明できるのだ。

「呼吸は生体メカニズムの根幹を成すもの。この研究成果は将来、海馬における記憶の仕組みなど、脳のシステムはもちろん、地球環境システムなど、時空間解析や因果性の推定が必要な分野に幅広く応用できるはずです」と三分一は話す。

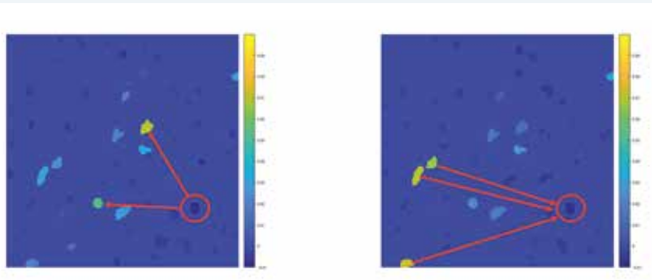
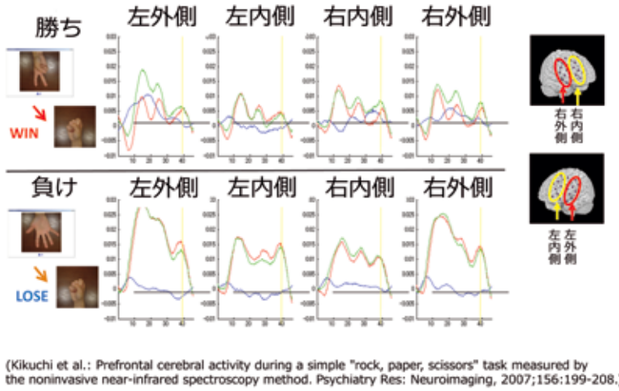


図3：左図は、グリシン抑制性ニューロン（赤丸）から他のすべての活動性ニューロンに対する影響を示す。矢印の先にある2個の活動性ニューロンが黄色寄りの色で表示され、比較的大きい影響を及ぼしている。右図は同じグリシン抑制性ニューロンに対するすべての活動性ニューロンからの影響を示す。この2枚の画像から、影響を及ぼすニューロンと、影響を受けるニューロンが異なっていることが分かる。



(Kikuchi et al.: Prefrontal cerebral activity during a simple "rock, paper, scissors" task measured by the noninvasive near-infrared spectroscopy method. Psychiatry Res: Neuroimaging, 2007;156:199-208.)

図4：「後出しじゃんけん」でコンピューター上にランダムに提示された手に勝つ手を出す課題（上）と負ける手を出す課題（下）を用いて、脳の活性化の様子を調べる。

認知システムの解明や熱中症対策システムに寄与

マルチモーダル生体信号データの時空間解析プロジェクトには、群馬大学大学院保健学研究科の菊地千一郎教授、大阪大学大学院基礎工学研究科の清野健教授、山口東京理科大学共通教育センターの木村良一准教授が客員として参加している。

菊地教授の「ヒトの認知機能と精神科領域におけるリハビリテーションの研究」は、じゃんけんテストなどさまざまな実験から得た脳波や脳組織の画像などから、精神疾患のメカニズムを明らかにするもの（図4）。

清野教授の「生体信号解析、モニタリング、予測、計測システムの開発」は、ウェアラブル生体センサが生む可能性を研究。繊維メーカーのクラボウとの産学連携プロジェクトで、インナーウェアから生体情報を取得し、熱中症リスクを管理するサイバーフィジカルシステムを構築した（図5）。

木村准教授の「電気生理実験、アルツハイマー病モデル動物を用いた認知システムに関する研究」は、海馬の信号の伝わり方と認知機能の関係性から、健常な脳とアルツハイマー病の脳の違いなどを明らかにするものだ（図6）。

いずれも、高度な統計解析が求められる研究内容であり、医療健康データ科学研究センターを通じた統数研との連携に、大きな期待がかかっている。

（広報室）

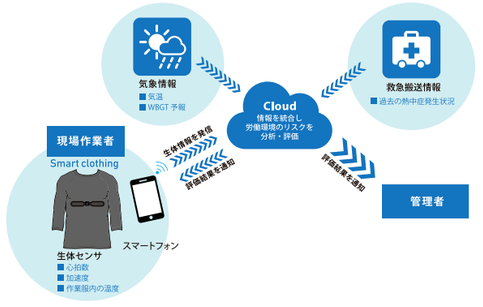


図5：サイバーフィジカルシステムの概念図。ウェアラブル生体センサによって建設現場などの作業者の心拍数や服内の温度などのデータを取得し、気象情報や救急搬送情報などと合わせてアラートを出す仕組みだ。

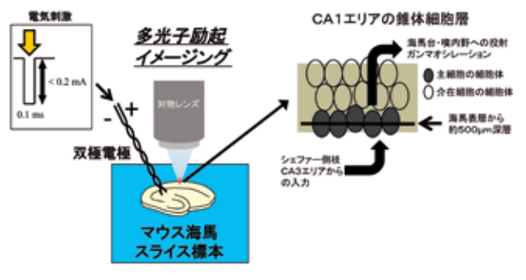


図6：海馬のスライス標本に刺激を与え、入力と出力の様子をイメージングする。

第15回「統計思考院におけるデータサイエンス高度人材育成プログラム」

実社会で活躍できる「棟梁レベル」のデータサイエンティストを育成

日本では今、データサイエンティストの圧倒的な不足が指摘されている。特に求められているのが、多くのデータサイエンティストたちを束ねる“棟梁”のような存在だ。統計数理研究所の中で人材育成に特化した組織である統計思考院は、日本社会に棟梁レベルの人材を増やすべく、裾野から頂点まで幅広い範囲をカバーする教育プログラムを開発、実施している。

▲川﨑能典統計思考院長

統計思考力を備えたT型人材を育てる“道場”

統計数理研究所の一角に、銘木の扁額が掛かった部屋がある。「統計思考院」の堂々とした揮毫は、情報・システム研究機構の北川源四郎前機構長の筆によるものだ。

統数研が「統計思考力を備えたT型人材の育成」を掲げて統計思考院を立ち上げたのは、今から7年前の2011年。ここでいう「T型人材」とは、専門分野の深い知識を縦軸とし、統計学という分野横断型の広い知識を横軸として備えた人材を意味する。統計思考院長を務める川﨑能典教授は、そのミッションを「人材育成のための教育プログラムを企画・開発し、世の中に提供すること」と説明する。

統計科学の博士の輩出については、統数研はもともと総合研究大学院大学複合科学研究科統計科学専攻での教育を担っている。統計思考院ではこれとは別に、社会人や他大学の学生などにも間口を広げ、多様な統計教育を展開している点に特色がある。

例えば、以前から統数研が社会人向けに開催していた公開講座や統計相談などの活動を統合・拡大し、現在は統計思考院の事業として継続。特に、あらゆる産業でデータサイエンスの知識が求められるようになった昨今、公開講座は大人気だ。2017年度には一般講座として「スパース推定」「ロバスト統計～外れ値への対処の仕方」「時系列解析入門」などのテーマで講座を開催し、延べ944人が受講した。

川﨑は「学生時代に理論を学ぶことも重要ですが、統計学の必要性を痛感するのは、実社会に出てデータ分析の課題に直面したときでしょう。統計学が“大人の学問”といわれる所以です」と話す。

研究者や学生はもちろん、固有分野の研究で統計学の必要性を感じた社会人などさまざまな人が集い、切磋琢磨しながら統計思考の訓練や研究をする。統計思考院は、そんな「道場」のような場だ。

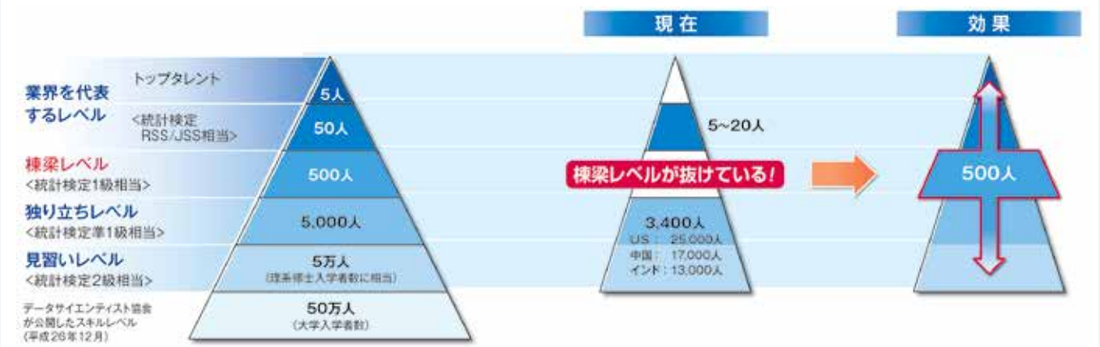


図1：左の図は、1年間に育成すべきデータサイエンティストの人数分布を表す。右側の2枚の図のように、現状では不足している棟梁レベルの人材を500人程度育成すれば、トップタレントが生まれる土壌となると同時に、大学や企業でのデータサイエンティスト育成の促進につながる効果が期待できる。

統計数理の知識を効率的に習得させる「リーディングDAT」

今、日本ではデータサイエンティストの不足が国家的な課題となっている。なかでも圧倒的に足りないのが、一般的なデータサイエンティストを指導し、統括するリーダーの存在だ。

情報・システム研究機構ビッグデータ利活用に係る専門人材育成に向けた産官学懇談会の報告書は、1年間に育成すべきデータサイエンティストの人数分布をピラミッド型で示している（図1）。

理想は、大学では基本的なデータリテラシーを身に付け、理系修士では統計検定2級程度のスキルを習得する。このレベルを「見習い」とし、その上位ランクには検定準1級相当の「独り立ちレベル」、さらにこの層を統括する1級相当の「棟梁レベル」が存在する、というものだ。

ところが報告書によれば、現在の日本は、ピラミッドの底辺に比べて棟梁レベルの層が極めて薄く、そのことが人材輩出システムのボトルネックとなっている。

そこで、統計思考院では棟梁レベル人材の戦略的育成に乗り出した。その一つが、統計数理の知識とスキルを持ったデータサイエンティスト（Leading Data Analytics Talents）の育成を目的とする「リーディングDAT」。2017年度に新たに編成したプログラムだ。

このプログラムは「リーディングDAT講座」と「リーディングDAT養成コース」で構成される（図2）。

前者は、棟梁レベルをめざすデータサイエンティストに必須の統計数理の知識を効率的に習得させることを目的とする講座だ。初年度は「実践ベイズモデリング」（L-B1）と「機械学習とデータサイエンスの現代的手法」（L-B2）の2講座を開催し、それぞれ

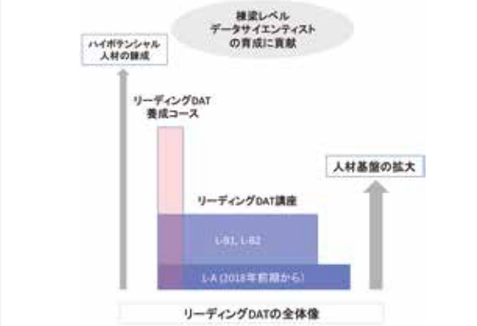
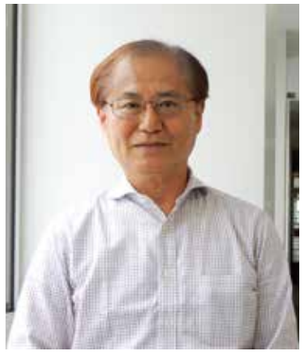


図2：リーディングDATの概念図と、2017年度の開催の様子。養成コースの修了者には樋口知之統計数理研究所長から認定証が授与された。

第15回「統計思考院におけるデータサイエンス高度人材育成プログラム」

▲野村俊一モデリング研究系助教


▲清水邦夫特命教授

▲神谷直樹特任准教授

61人が受講した。

一方、後者は少人数のハイポテンシャルな人材を対象とする。リーディングDAT講座を受講した後、その内容に関する実践的な課題が出され、1カ月かけて取り組んだ結果をレポートにまとめる。レポートの評価が合格で、かつ講評会と特別講演に出席した受講者には修了認定証が交付される。17年度は募集人数を大幅に上回る申込者の中から34人が受講し、そのうち25人が認定証を受け取った。

「短期間でさまざまなトピックを体系的に学び、自分でデータを扱う訓練ができる。社会人にとって、すぐ業務に生かせる内容となっています」と川崎は言う。L-B1を受け持ったモデリング研究系の野村俊一助教は「データ分析をなるべく実践的に教えました。レポートを読むと、概ねきちんと分析できているものの、テクニカルな部分で勘違いをしている人も。間違いやすい部分は講評会で補足しました」と話す。

高度な内容の講評や講義は統数研ならではだが、受講者のレベルは一律ではない。大学で統計学の理論を学んだ人は少ないのが実情だ。そこで、2018年度からは、3日間でデータサイエンスの基礎を学ぶ入門講座（L-A）も新設した。初心者から最高レベルまでのラインナップを揃え、一貫して教えられるのも、統計専門家を擁する統数研にしかできない取り組みだ。

データ解析・統計分析のお悩み相談「共同研究スタートアップ」

公開講座と同様に、以前から実施していた統計相談を引き継いだのが、統計思考院の実施する「共同研究スタートアップ」プログラムだ（図3）。社会人や統計学以外の研究者など、データ解析・統計分析に関わる問題で悩んでいる人々に対し、共同研究のコンサルティングを行う。清水邦夫特命教授をはじめとするベテラン教授陣が統計思考院に所属する若手研究者とペアになり、持ち込まれた課題の解決を支援する。

豊富な経験と知見を生かし、取得済みのデータに関する解析法だけでなく、データの収集法や調査の方法、実験の計画など、事前準備の段階も含め、豊富な経験と知見を生かした助言が好評を博している。清水は「年間50件ほどの相談を受け付け、そのうちの4〜5件は共同研究に結びついています」と実績を紹介する。

清水は、統計思考院の役割を戦後の大学と製造業の関係になぞらえる。「アメリカで開発されたQC活動を日本の大学が企業へ紹介し、指導しました。そして日本はものづくり大国になった。ビッグデータ時代の今、企業だけではできない統計分析を統計思考院が主導し、支援していく必要があると思います」。

さらに、この取り組みは、相談者への貢献に留まらない。数多くの多様なデータに触れ、実社会での課題解決に携わることは、特任助教などとして統計思考院へ配属された若手統計学者にとっても貴重な経験になる。「ここから巣立った人は皆、大学や研究所の良いポジションに着いています」と清水が言うとおり、有効なキャリアパスとなっている。

図3：共同研究スタートアップでは、統計思考院のベテランと若手が企業や他分野の研究者の相談を受けてアドバイスする。内容によって、共同研究に発展することも。

企業の生データを扱えるのが魅力「データ分析ハッカソン」

統計思考院では、次代を担う学生や若手社会人にも成長の場を用意している。「データ分析ハッカソン」は、チームごとにデータ分析スキルを駆使し、社会問題の解決や新規サービスの提案といった正解を決めにくい課題を競い合うイベントだ（図4）。

一般的なハッカソンとの最大の違いは、企業などが保有する生データに対するデータ分析スキルを重視することだ。例えば、「販売データを基に売上げアップにつながる新たなサービスを提案する」といったテーマでハッカソンを実施する場合、リアルなデータの確かな分析に基づいた提案でなければ意味がない。だが、企業は当然ながら、秘匿性の高い生データを外部へ出すことには慎重だ。統計思考院が生データを提供してもらえるのは、統数研への信頼はもとより、専用のクラウド計算システム上に仮想デスクトップ環境を用意し、高いセキュリティーを確保しているからだ。

データ分析ハッカソンを担当する神谷直樹特任准教授は「座学では、これほど大規模な、しかも不整合の多い、いわゆる“汚れたデータ”を扱う機会はありません。この経験を通じて参加者は自分の実力に何が足りないか分かり、次のステップへ進むことができるのです」とその意味を語る。

時代に応じて、さまざまな場面で必要とされるデータサイエンティストを育て、社会へ送り出す——。統計思考院の存在意義は、ますます重要性を増している。（広報室）

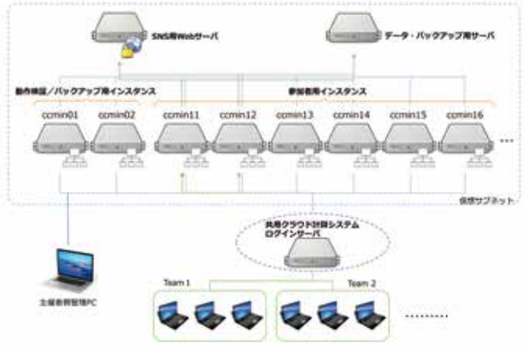


図4：2017年度のデータ分析ハッカソンには6チームが参加。メンバーはデータ補正やモデル化、事業化のアイデア出しなど役割分担をすることで、自分の得意分野に気づくことができる。オンプレミスの共用クラウド計算システム上に仮想デスクトップ環境を用意し、高いセキュリティを確保したうえで実施。

第16回「公的調査データの統計開示抑制」

調査票情報を安全に提供する持ち出し審査基準を開発

政策立案のために官庁が実施してきた公的調査のデータを学術研究にも活用しようとの動きが顕著になってきた。その大前提となるのが、調査協力者のプライバシー保護だ。統計数理研究所は、データの提供元である統計センターと連携し、研究者が全国の大学に設置されたオンサイト施設で調査票情報にアクセスして分析した結果を持ち出す際の安全性審査基準を策定。審査における経験則ルールを決定するとともに、研究者を支援するツールの開発を手掛けている。

▲南 和宏モデリング研究系准教授

提供対象の拡大で調査票情報の活用が進む

国勢調査や国民経済計算（GDP統計）、経済センサス（企業経済活動の実態調査）など、政府は長年にわたりさまざまな公的調査を実施してきた。その位置づけが「行政のための統計」から「社会の情報基盤としての統計」へ大きく転換したのは、2007年の統計法全面改正をきっかけとする。

これにより、国民が合理的な意思決定を行うための共有財産として、公的統計を活用する道が拓かれた。学術研究など公益目的に限り、調査票情報の二次利用が認められるようになったのである。

ただし、利用者である研究者には、データを操作する部屋の入退室管理や外部ネットワークからのシャットアウトなど、厳しい要件が課せられる。このため、研究室のセキュリティ環境によっては利用申請できないケースが少なくなかった。また、提供方法も、必要最低限のデータをDVDの形で渡されることから、使い勝手がよくなかった。

そこで、2010年に情報・システム研究機構とデータの提供元である統計センターが二次利用の普及・啓発に関する連携協定を結び、2012年に東京・立川の統計数理研究所内に「オンサイト施設」を開設。研究者は事前申請のうえで来訪し、調査票情報に直接アクセスできるようになった。

さらに今年度から、全国9カ所の大学などに設置されたオンサイト施設で、リモートアクセスの試行運用も始まった。今後は、施設を訪れた研究者が、シンクライアントからサーバにアクセスし、フルスペックのデータを用いて分析できるようになる。このことが、探索的・創造的な研究に資すると期待されている（図1、2）。

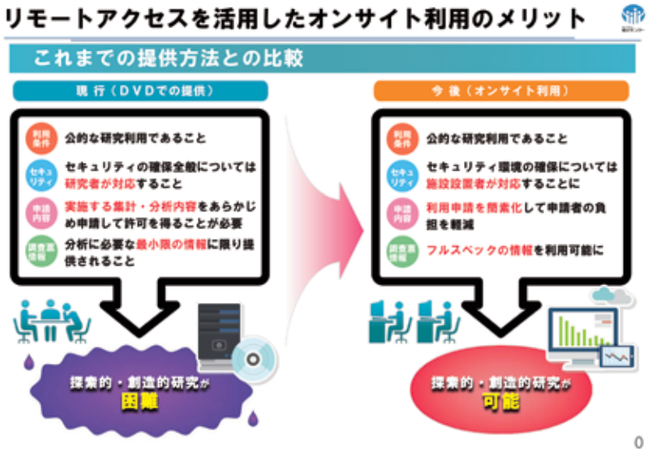


図1：公的統計情報へのリモートアクセスによるオンサイト利用では、従来に比べて利用申請が簡素化されるとともに、フルスペックのデータを活用できるようになる。

オンサイト施設利用での分析結果安全基準を作成

リモートアクセスを活用したオンサイト利用の体制を整えると同時に、より強化しなければならないのが、調査対象者の機密情報保護だ。例えば、高額商品の購買情報、家計の収入や土地などの資産額、雇用契約期間や仕事内容など、プライバシーに関わる情報が、公開された研究成果から漏洩することのないよう、慎重な取り扱いが求められる。

そこで、オンサイト施設で分析した成果物を外部に持ち出す場合には、その安全性・秘匿性のチェックを受ける仕組みを構築することとなった（図3）。統計センターの非常勤職員となり、その安全基準を作成したのが、統計数理研究所の南和宏准教授と、匿名化技術の研究で実績のあるNTTの菊池亮博士だ。

「今回の安全基準を策定するにあたり、まず二つの前提条件を決めました」と南は説明する。一つは、研究者は悪意の情報漏えいを行わない、と仮定すること。つまり、機密情報を含む分析結果が不注意によって論文に掲載され、第三者によって漏えいするケースを想定する。

もう一つは、研究者がデータの持出しを複数回行う場合、中間結果にも最終結果と同じ基準を適用すること。中間結果に対しては、論文に掲載しない前提で基準を緩和することも検討したものの、うっかり掲載してしまうミスを防止できないと判断し、この結論に達したという。

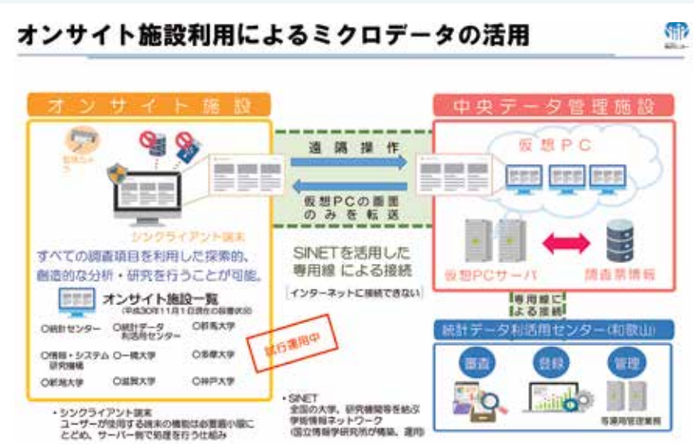


図2：オンサイトを訪れた研究者は国立情報学研究所の学術情報ネットワーク「SINET」を介して遠隔で中央データ管理施設のデータベースにアクセスする。

基準作成にあたっては、欧米諸国の統計開示抑制に関する「Eurostat基準（SDCハンドブック）」を参考にした。ここでは「安全性基準の作成は『どのような動物を入れるかわからない動物園の檻』を設計するようなもの」と汎用的な基準作成の難しさを説き、審査対象を標準的なデータ形式に限定している。南は「今回の基準でも同様に、対象を度数表や数量表、線形回帰係数および非線形回帰係数、パーセンタイル値などに絞りました」と話す。

同ハンドブックでは、実際の審査は「経験則」による一次審査、「原則ルール」による二次審査の2段階で行う。経験則は、明示的に記述されたルールで、非常に厳格なものだ。これに対し原則ルールは、データの意味や特徴を考慮し、個別の状況に柔軟に対応するもの。日本でもこの2段階審査の枠組みを採用することにした。

菊池博士は「安全性と利便性はトレードオフの関係にあり、完璧な基準をつくることは不可能。運用面でいかにバランスを取るかがポイントです」と話す。

第16回「公的調査データの統計開示抑制」

▲菊池 亮博士（工学）NTTセキュアプラットフォーム研究所

経験則での原則は、①個票データは機密情報、②客体10の原則、③占有性の原則、④グループ開示の原則の四つ。①の原則から、個票データの最小値、最大値の持ち出しは不可となる。②は、個票値の推定ができないように客体数を10以上確保するというもの。③は同様の理由で、いずれかの個票値の割合が全体の50%以上を占めてはいけないというものだ。また④の原則から、ある属性でグループ化された個票の90%以上が、別の属性に関して同一の区分に属することを禁ずる。

原則ルールでは、研究者が持ち出しを希望する分析結果を「複数の個票値を入力とする関数の出力」と捉え、関数の出力を使って個票値をどの程度特定できるのか、データの特性も加味して判定する。場合によっては経験則ではNGであっても安全であるような場合があり、そのような場合にも柔軟に対応するのが、原則ルールの目的だ。

申請者の負担を軽減するRツールを新たに開発

通常の分析結果は機密情報を含むことから、そのままでは安全性審査を通らない。このため、研究者は事前にデータの秘匿処理を行った分析結果を用意する。一方で、審査担当者は秘匿済みの分析結果を見ても、基準を満たしているか否かの判断ができない。

このため研究者には、審査用に説明資料を添付してもらうのだが、これが研究者にとって作業負担を増大させる要因となる。例えば、占有性ルールの審査で、全体の50%を占める個票値がないかどうかは、元の個票データから第一位、第二位の客体の値を抽出する必要がある（図4、5）。

また、秘匿化をするにも、集計表のうち占有性ルールを侵害する数値を持つセルだけを隠したのでは、行計や列計から復元できる可能性がある。この場合、秘匿セル変数の取りうる範囲が十分広くなるように、追加でセルを隠さなければならない（図6）。こうした処理を手計算で行うのはほぼ不可能だ。

実際に、オランダの統計局は表データの秘匿処理ツール「τ-Argus（タウ アルガス）」を開発している。南らは、日本にもこれを導入できないかと考えた。だが、調べてみると、このツールはどう考えても、分析者と審査者が同一でなければ使えないものだった。南はオランダへ赴き、直接、担当者に話を聞いた。ところが、意外なことにオランダでは安全性審査にはツールを使わず、審査担当者の経験に基づく審査が行われていた。

使えるものがないのなら、つくるしかない――。そう考え、南と菊池博士は統計ソフト“R”を使い、秘匿処理と説明文書が同時に自動で作成できる新たなツールの開発に乗り出した。

▲阿部 穂日統計センター上席統計専門職

分析者と審査者をつなぐという思いで「SDCLink」と名付けた新ツールはすでに施行運用段階に達している。「SDC」は Statistical Disclosure Control（統計開示抑制）の略。近日中にオンサイト施設に導入するとともに、完成度を高めてRのパッケージとして提供することも計画しているという。

運用状況の課題をフィードバックし利用者の利便性を高める

安全性基準の原則ルールは、裁判における判例のように、運用しながら実例を積み上げるなかででき上がっていくものだ。調査データ開示の事務を取り仕切る統計センターの阿部穂日上席統計専門職は、審査側の立場から次のように語る。「経験則ではどんなデータも基準を満たさなければ持ち出し不可とするものの、例えば消費実態調査の客体の性別や年齢のクロス集計などのように、公開しても実質的に問題ないものもあります」。

現状では、どの情報を秘匿すべきか、明確な線引きはなされていない。統計調査の所管官庁の判断は、協力者を確保するためにどうしても安全側へ傾く。しかし、そのために研究者の手間や審査時間が増大したのでは、利用促進に歯止めがかかる。阿部上席統計専門職は「統計家の研究成果を実務に反映し、審査時間の短縮を進めるなど、利用者の利便性を確保できるシステムをつくっていきたい」と話す。

このプロジェクトに携わった意義について、南は「個人のプライバシー保護は、ビッグデータ時代の最重要課題です。オンサイト利用における安全性審査プロセスの確立をビッグデータ利活用の成功例にしたい」と抱負を語る。データサイエンスの最前線を支える取り組みを統数研の知見が牽引している。

（広報室）

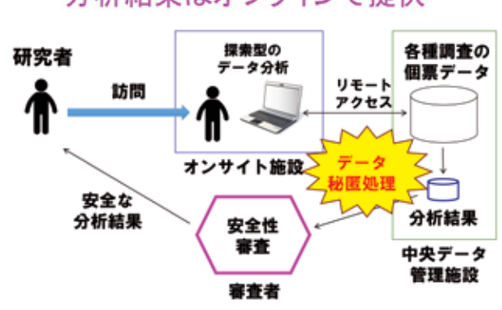


図3：オンサイト施設からのリモートアクセスによる分析結果の持ち出し手順のイメージ。研究者自身が、安全性審査を受けるまえにデータ秘匿処理を行う。

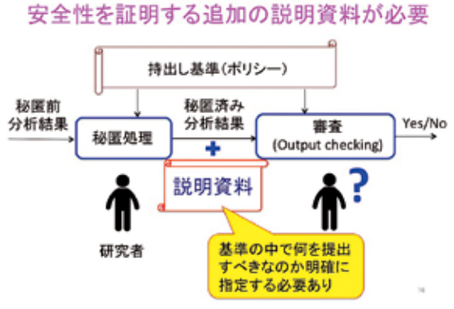


図4：データ秘匿処理を施した分析結果を審査するには、研究者自身が説明資料を揃える必要がある。

図5：占有性ルールの審査では、分析データに加えて、第一位、第二位の客体の値を抽出して提出しなければならない。

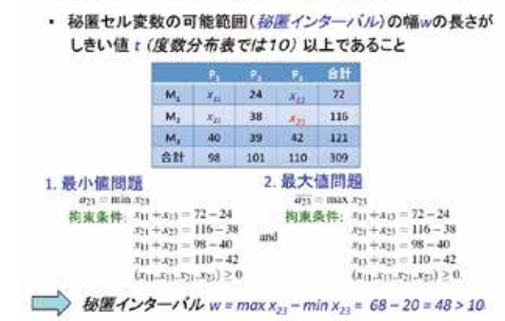


図6：秘匿化にあたっては、当該セル（この例の場合x_{23}以外に、x_{11}, x_{13}, x_{21}のセルを隠す必要がある。

第17回「多層整数計画に基づくクリンチ／エリミネーションナンバーの計算」

リーグスポーツをより面白くする勝敗数計算のアルゴリズムを確立

野球などのリーグスポーツでは、シーズン中盤を過ぎた頃から、「あと何試合勝てば優勝か」あるいは「プレーオフ進出か」がファンの最大関心事になる。それを算出するのが「クリンチナンバー」だ。ルールが複雑な競技でプレーオフのように複数チームの順位が問題となる場合、クリンチは「非線形」となり、計算量は膨大になる。統計数理研究所は、ドイツの研究機関との共同研究によって、そのアルゴリズムを確立した。

▲伊藤聡副所長

セ・リーグのCS進出を巡ってマジックナンバーに食い違い

今からちょうど10年前の秋、プロ野球界で、クライマックスシリーズ（CS）への出場が決定するマジックナンバーを巡り、ちょっとした混乱が起こった。

9月12日、マスコミは一斉に巨人の出場決定を報じた。11日の試合が終わった時点でCSマジックが「1」となったからだ。ところが、実際はその試合の行われる前の時点で、巨人はセントラル・リーグで3位以内に入ることが決定していたことが判明したのだ。

当時、日本のプロ野球セ・リーグには、「優勝マジック」と「CSマジック」の2種類のマジックナンバーがあった。前者は「他のチームが残り試合を全勝しても、自分のチームの優勝が確定する」と言える残り勝ち試合数を表す。つまり、「マジック1」ならば「あと1勝すれば優勝確定」であり、優勝が確定した時点で「マジック0」となる。マジックが「点灯」するのは、他の全チームに自力優勝の可能性がなくなった瞬間だ。

一方、CSマジックは、CSへの進出が決定するまでの最小勝ち試合数を表す。CSに出場できるのはリーグ戦の上位3チーム。

09年のケースで言えば、巨人が残り試合を全敗した場合の勝率と、他のチームが残り試合を全勝したときの勝率を比較し、巨人を上回ることのできないチームが3チーム以上になった時点で、巨人のCSマジックが0となる。

11日の試合終了時、5位の広島と6位の横浜は残り試合全勝でも巨人を上回ることができなくなり、3位のヤクルトと4位の阪神は残り試合を全勝すれば巨人をわずかに上回る状況となった。このため、巨人のCSマジックが1となったのである。

しかし、前日の10日の時点で、ヤクルトと阪神の直接対決が6試合残っており、その勝敗を加味すると、両方のチームが共に巨人を上回る可能性はなかった。巨人が4位以下になる可能性もなくなっていたことから、すでに巨人はCSマジックを0とし、CS進出を決めていたことになる。

どのメディアも、共同通信社の配信するマジックナンバー情報を基に記事を作成していたため、この食い違いが見過ごされてしまった。

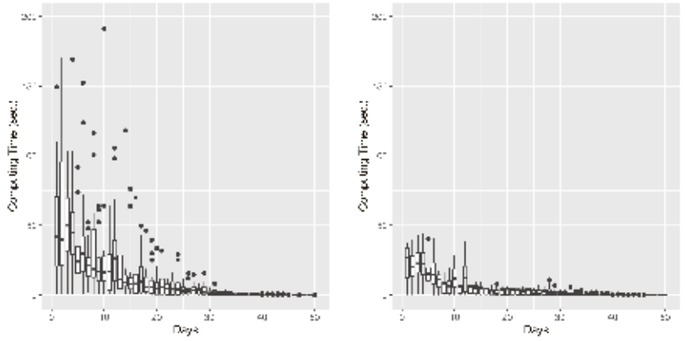


図1：Bリーグを対象としたチャンピオンシップ・トーナメント進出エリミネーションナンバーの計算時間。全18チームの計算に要した個々の時間を1日単位で箱ひげ図にプロットした。アルゴリズムを工夫することで、改良前（左）から改良後（右）には計算時間を大幅に短縮することができた。

最適化問題としてクリンチナンバーを解く手法で特許を取得

共同通信社は、この事態を受けて統計数理研究所に協力を依頼した。従来のCSマジックナンバーのように「あと何勝すればいいか」と上限を与える目安にとどまらず、最小の勝数を正確に与える新しい指標を模索してのことだった。

相談を受けた伊藤聡教授（現 統計数理研究所副所長）は同社と共同研究を進め、2010年にCS進出までの目安を計算する新たな指標「CSクリンチナンバー」を開発し、特許を取得。共同通信社は、各加盟メディアへの配信を開始した。

伊藤は「クリンチナンバーの算出が最も難しいのは、間違いなく日本のプロ野球でしょう。ルールが複雑なうえ、チーム間の勝率が同率の場合の取り扱いなどに、明文化されていないルールがあり、条件が非常に細分化されるからです」と話す。

しかも、優勝マジックでは1位になるための最小勝数だけを考えればいいのに対し、CSでは「3位以内」が条件になることから、点灯チームも比較の対象となるチームも複数となる。伊藤の開発したCSクリンチナンバーは、最適化手法を用いて計算の必要のないパターンを排除することで、短時間で計算できる方法だ。

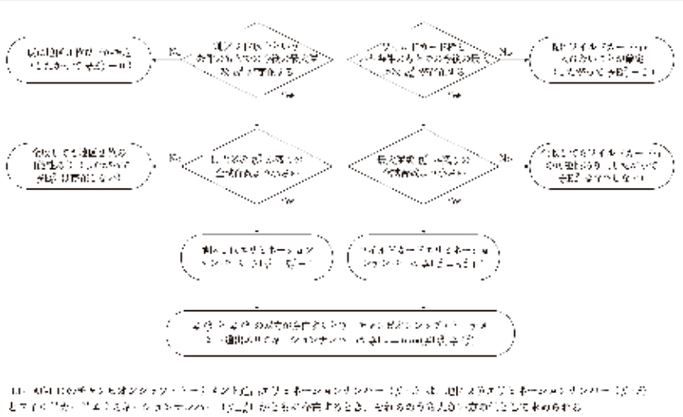


図2：整数計画法によるエリミネーションナンバーの考え方のフロー（国内男子プロバスケットボールのBリーグを対象とした例）。

クリンチナンバーの「clinch」は「決着をつける」という意味だ。ちなみにマジックナンバーは「majic word（呪文）」から派生した言葉。ビンゴゲームで、「この数字が来ますように」と祈るときの数字をマジックナンバーと呼んでいたのが、野球などのリーグスポーツにも使われるようになったという。

複雑なルールがもたらす「非線形性」の整数計画問題を解くアルゴリズム

2010年に開発したCSクリンチナンバーの計算方法は、今でも共同通信社の配信に使われている。しかし、伊藤の挑戦がそこで終わったわけではない。

「じつは開発のとき、解けない問題が一つありました。その部分を回避するために、別のさまざまな方法を組み合わせて解を求めています。計算結果は同じですが、学術的にはそれを直接解けるようにしたいと思っていたのです」と伊藤は振り返る。

そのためには「整数非線形計画問題」を解く必要があった。そこで、ドイツ・ベルリン州の研究機関「Zuse

第17回「多層整数計画に基づくクリンチ／エリミネーションナンバーの計算」

▲品野勇治客員教授

Institute Berlin（ZIB、コンラート・ツーゼ情報技術センター）」に所属する研究者で、統数研の客員でもある品野勇治教授に相談を持ちかけた。

品野は整数計画問題を含む離散最適化計算の専門家だ。2004年に「広辞苑に収載された名詞を使ってどこまで長いしりとりを作ることができるか」という「最長しりとり問題」を整数計画問題として解き、テレビ番組「トリビアの泉」に出演したことでも知られている。

世界で初めて機械式コンピューターを開発した発明家のコンラート・ツーゼの名を冠したZIBは、計算のアルゴリズムなどコンピューター技術に特化し、最先端の研究を進めている。

この研究所の大きなテーマとなっているのが、今まで解けていなかった複雑な問題について、数理計画問題としてモデル化し、それを解くソフトウェアを開発すること。鉄道、ガス輸送、医療などさまざまな分野の課題を手がけており、それらの共通する部分を一般化したソフトウェアパッケージ「SCIP Optimization Suite（略称スキップ）」を開発している。

物流の単位は1個、1箱、1ロット…と、ほとんどの場合に整数だ。従って、世の中の最適化の課題の多くは整数計画問題に帰結する。「ただし、整数計画問題を解くには膨大な量の計算が必要。この20年間でアルゴリズムが著しく進歩し、計算速度が1兆倍ぐらい上がってきたので、かなり解けるようになってきたところです」と品野は説明する。しかも、整数計画問題に非線形性が加わると、解析は飛躍的に複雑になるという。

SCIPには数年前、新たな機能が搭載された。それが、ガスパイプラインの設計に関連して必要となった混合整数非線形計画法（MINLP）ソルバーだ。ガスパイプライン設計の最適化問題は、管内の圧力と温度の関係などには非線形性があることから、これまで解くことができなかった難問の一つだ。

CSクリンチナンバーを直接解く問題もまた、非線形だ。「日本のプロ野球の場合、順位決定のルールが非常に複雑なので、数式で表すとA4の用紙に4ページ分にもなります」と伊藤は笑う。膨大な条件を満たすパラメーターの中から、ある関数を最大あるいは最小にするものを決める計算をしなくてはならない。2010年の時点では、この複雑さにコンピューターのアルゴリズムが追いついていなかった。

「MINLPソルバーは、まさにクリンチナンバーの問題を解くのにうってつけのソフトなのです。これを使ってどこまで解けるかやってみよう、ということになりました」と品野は説明する。

最初の数年間、二人はCSクリンチナンバーに適したモデルをひたすら検討し続けた。ようやくモデルが完成しても、今度は計算の段階で困難が待っていた。「1日に何問も解かなければいけないのに、1問を計算するのに数日かかることもありました」と品野は苦笑する。

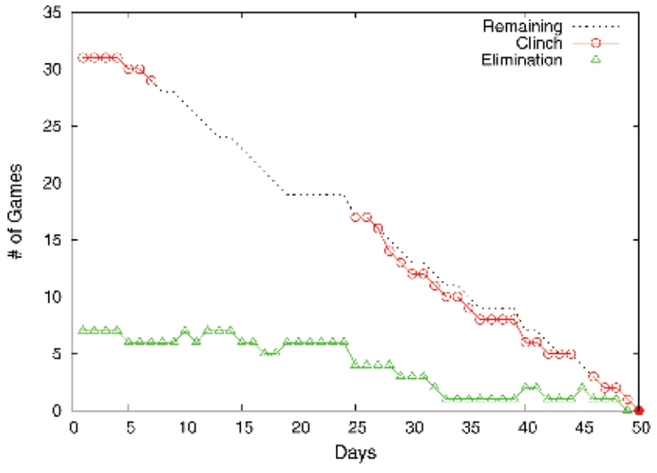


図3：Bリーグのチャンピオンシップ・トーナメント進出クリンチナンバーとエリミネーションナンバー（進出できないことが確定する試合数）の例。7日目でクリンチナンバーが消滅し、残り試合に全勝しても進出できない可能性が残っていたが、25日目に再びクリンチナンバーが復活。最終的に50日目に進出が確定したことを表している。

1問にかかる時間は最長でも1分以内に収めなければ、実用では使いものにならない。それ以後の研究は、アルゴリズムの改良に多くの時間を費やすことになった。

こうした苦労を乗り越え、今、ようやくアルゴリズムはほぼ完成したと言える段階に到達。計算時間も大幅に短縮した（図1）。

あらゆるリーグスポーツ、そして産業界の諸問題への展開を視野に

「整数計画問題は、汎用性が高いのが最大の魅力です」と品野は言う。今回のCSクリンチナンバーの解法も汎用化が可能で、世界中のほぼすべてのリーグスポーツに展開できる。

実際に、伊藤と品野は国内男子プロバスケットボール「Bリーグ」を対象としてクリンチナンバーの算出プログラムを完成させ、2018年秋に論文を発表した（図2、3）。また、統数研に留学している台湾人学生とともに、台湾のプロ野球のルールを踏まえたクリンチナンバー算出法を理論化。現地での論文発表も実現した。

さらに、伊藤は「このモデルを使って例えば、そのスポーツのシーズン中の終盤の時期、野球なら9月頃に最も盛り上がるようなルール改正を提案することもできます。スポーツをより面白くするためにも役立つでしょう」と話す。

この問題に取り組む意義はそれだけではない。整数計画化問題は、産業界が抱える最適化問題のカギを握る。その計算手法の進化に貢献することで、数理研究と産業界の連携を強化することにもつながるはずだ。

統数研とZIBは、九州大学マス・フォア・インダストリ研究所（IMI）と三者間で学術協定を結んでいる。日本とドイツの産業界を含めた連携を推進することが目的だ。2017年から毎年ワークショップを開催しており、その第2回で、伊藤と品野は整数計画に基づくクリンチナンバー、エリミネーションナンバーの計算について発表した（図4）。

伊藤は統数研の副所長として「産業界との連携を深めることは、産業発展を下支えすることにつながると同時に、研究機関にとって外部資金獲得の機会ともなる。今後も積極的な展開を目指していきます」と抱負を語った。

（広報室）

図4：統数研とZIB、九州大学IMIが2017年9月にベルリンで開催した「第2回ISM-ZIB-IMI MODAL Workshop」の参加メンバー。3者が持ち回りで幹事を務めており、今年3月に立川の統数研で第4回のワークショップを開催した。

第18回「URAによる研究力強化」

研究者に寄り添い、資金・知財を戦略的にマネジメントするURA

研究内容を理解したうえで、研究資金の調達と管理、知財の管理と活用を担うURA（ユニバーシティ・リサーチ・アドミニストレーター）。統計数理研究所では現在、3人のURAがそれぞれのバックグラウンドを生かした活動を展開している。研究者が研究に集中できるように、研究以外の面を全面的にサポートするとともに、共同研究による社会課題の解決やデータサイエンティストの育成など、産学連携の推進にも取り組む。

▲北村浩三上席URA

教員と事務職をつなぐ“第3の職種”として統数研独自のURA業務を展開

文部科学省が「研究大学強化促進事業」として、10カ年計画でURA組織の定着と展開に着手したのは2013年のことだ。

その頃すでに海外では、研究内容を理解したうえで資金調達や知財管理を一手に引き受けるURAは、教員と事務職の間をつなぐ“第3の職種”として広く認知されていた。これに対し、そうした職種を持たない日本では、研究者に研究活動以外の業務で過度の負担がのしかかっていたのが実情。研究力強化のために、URAの導入は喫緊の課題であったと言える。

これを受けて、統計数理研究所の上部組織である情報・システム研究機構にもURAステーションが設置され、北村浩三上席URA、岡本基主任URA、本多啓介主任URA（職名は2018年4月からのもの）が統数研チームに配属された。

3人のバックグラウンドはさまざまだ。北村は日本IBMに30年間勤めた後、大学連携の業務実績を生かして統数研のURAに応募し、採用された。岡本は機構本部の研究プロジェクトである新領域融合研究プロジェクトの研究員として統数研に入り、リスク解析戦略研究センターで研究・事業支援に当たった後、異動した。本多は総合研究大学院大学の統計科学専攻で博士号を取得し、民間企業を経て着任。

統数研のURAが手がける業務範囲は、研究マネジメントから、産業界との連携や大学共同利用機関としての共同利用の推進、統数研が主催するイベントの実施支援までと幅広い。企業人としての北村の実績、研究・事務の両面から統数研を知り尽くした岡本の経験、最新のデータサイエンスを習得した本多の専門知識。「それぞれが自らの経験と知見を活かし、適材適所でURA業務に取り組んでいます」と北村は話す。

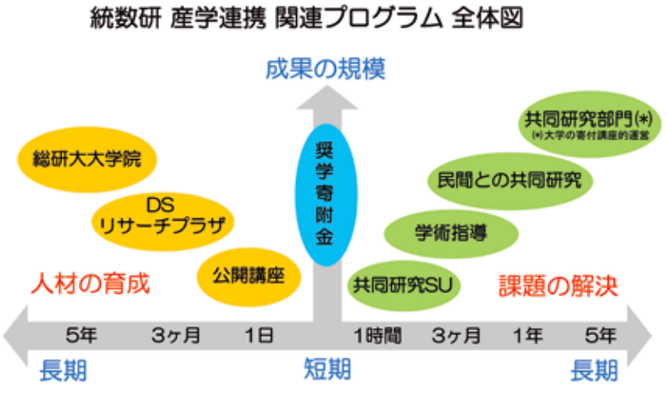


図1：統計数理研究所産学連携関連プログラム全体図。人材育成・奨学寄附金・課題の解決の3分野についてさまざまなメニューを用意している。

▲岡本基主任URA

研究内容を理解し、プロジェクトごとに的確なマネジメントを実施

国立大学法人と同様に大学共同利用機関法人に対する運営費交付金は、ここ数年、毎年減少している。それだけに、競争的資金をいかに獲得するかが重要だ。

国の科学技術政策の方向性や、情報・システム研究機構と統数研の事業方針に沿って研究戦略を立てるのが、URAの大きな役割の一つ。本多は、そのために自組織のさまざまな情報を収集・分析するIR（Institutional Research）の新指標「REDi（Research Diversity Index）」を開発し、役立てている。

こうして構築した戦略に基づき、研究プロジェクトの企画、設計、調整、申請などを支援するプレアワード業務を手がけるのは、主に岡本の担当だ。教員の競争的資金への応募に際し、研究内容や予算計画などについて相談に乗り、その獲得に向けて提案書の文言などをシェイプアップする。プロジェクトが採択された後は、適正な運営によって予算を執行するポストアワード業務を担う。

その他、新たに赴任した教員へのオリエンテーション、職務発明の窓口業務や外国人研究者招へいのための査証取得などもサポート。「教員が効率的に活動できるように支援することで、より研究に集中してもらえればと思っています」と岡本は話す。

民間資金の調達と学術成果の社会実装両面から注目される産学連携の取り組み

もう一つ、近年特に注力しているのが、産学連携の取り組みだ。民間からの資金調達は、公的資金に代わる研究財源として大きく期待されている。同時に、統数研が産業界と連携することにより、トップレベルの統計数理の学術成果を社会実装することにもつながる。

統数研では産業界の課題解決のプログラムとして、「共同研究スタートアップ」「民間との共同研究」「学術指導」「共同研究部門」の4種類のメニュー

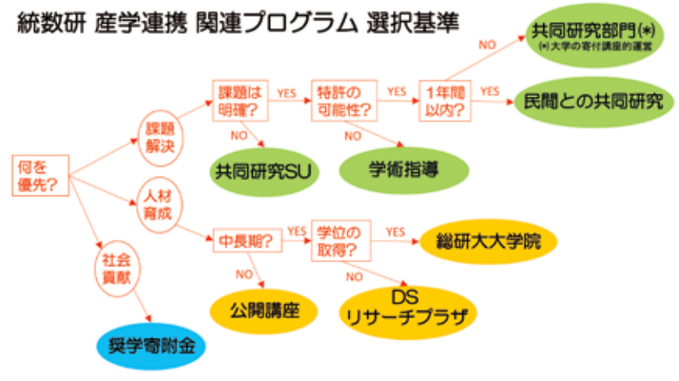


図2：統計数理研究所産学連携関連プログラム選択基準。企業は優先度や課題の明確さ、期間などによって最適なプログラムを選択できる。

第18回「URAによる研究力強化」

▲本多啓介主任URA

を用意している。このうち、学術指導と共同研究部門の2つは、北村が情報・システム研究機構に制定を提案して設けられた規定に則り、2017年に開始したものだ（図1、2）。

学術指導は、1つの案件につき3カ月から半年間程度の期間で、面談を通じて教員が企業の担当者に指導するプログラムだ。民間との共同研究は以前から実施していたものの、中には統計スキルやツールの使い方を指導するなど、コンサルティング的な内容の案件も少なくなかった。「学術指導は、それを共同研究と分別してメニュー化したものです」と北村は説明する。

また、共同研究スタートアッププログラムによる受け入れ体制も強化。統数研の教員が企業から持ち込まれた課題について相談に乗り、どのプログラムが適しているかを提案するものだ。

例えば、プロ野球クライマックスシリーズ（CS）の進出に必要な最少の勝数を正確に与える指標として共同通信社が配信している「CSクリンチナンバー」は、共同研究スタートアップから共同研究へ発展した成果だ。

案件が持ち込まれるのを待つだけでなく、積極的に新規開拓にも努めている。担当の本多は「企業での統計の使われ方を知っているので、実務者のニーズと学術研究のギャップを埋めるアドバイスをしていきたい」と思いを語る。

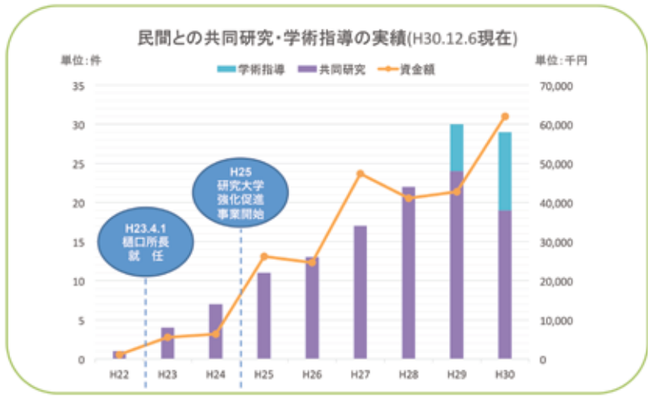


図3：民間との共同研究・学術指導の実績。URA事業を開始した2013年（H25）から共同研究・学術指導の件数と資金額が大幅に増加している。

URAは共同研究・学術指導契約の窓口として、特に契約に関する交渉を支援している。共同研究と学術指導の件数は、樋口知之前所長の所長就任とURAの活動開始を境に飛躍的に伸びている（図3）。成果に満足した企業は別案件でリピートするケースが多く、2018年の実績は12月時点で約30件に上った。

研究組織を新設する共同研究部門第1号が10月にスタート

一方、共同研究部門は、民間企業の抱える複雑で高度な課題の解決に向け、統数研が新たに開始した産学連携事業だ。大学における寄付講座的な運営により研究組織を新設するもので、部門専任の教員を新たに雇用して配置し、2〜5年をかけて本格的に企業との共同研究を行う。

統数研に籍を置く50人足らずの教員では、受諾できる共同研究、学術指導の件数には限りがある。その点、新たな教員を配置する共同研究部門ならば人的資源が確保でき、長期にわたる資金調達も可能になる。

もちろん、企業にとっても統数研の知見を存分に活用できるメリットがある。北村は「統計数理は多様性、応用性が高い研究分野であり、どの産業にとっても利用価値があります。われわれとしても独自資金を増やし、研究所として自立した経営へ近づくことになるのです」とその意義を語る。

共同研究部門の第1号となるのが、10月にスタートする三菱ケミカルとのプロジェクトだ。新たなマテリアルズ・インフォマティクスの基盤技術構築を目的として、3年間、統数研内に専用の研究室を設け、共同研究部門の看板を掲げる。

マテリアルズ・インフォマティクスとは、情報科学を駆使し、新材料や代替材料を効率的に見つけ出す取り組み。統数研のもつデータ科学による解析技術と、三菱ケミカルのもつ計算科学による予測技術を融合することで、革新的な特性を有する新材料を発見するための物質探索アルゴリズムを作り上げる。さらに、そのアルゴリズムを高分子、触媒、無機材料といった具体的な材料設計の課題に適用しながら、高度化を図っていくことが狙いだ。

この共同研究部門で構築したアルゴリズムは学術成果として積極的に発信し、マテリアルズ・インフォマティクス分野におけるオープンイノベーション・オープンサイエンスの促進に貢献する。

共同研究部門は前例のない取り組みであることから、北村は契約書の書式を定めるのに苦労した。公的機関である統数研と民間企業では、契約に関する内規に相違点が少なくない。企業人としての実務経験が豊富な北村は、「企業側の担当者と一緒に根気よく調整を続けて契約書の雛形を作り上げました」と振り返る。

データサイエンティスト育成に向けて企業や大学を多角的に支援

産業界との連携で、企業からのニーズが急速に増えているのが、データサイエンティストの人材育成だ。ビッグデータに関する深い知識を持ち、企業の保有するデータを分析して経営層へ情報提供したり、経営上の課題を抽出したりできるデータサイエンティストは、いまやどこの企業でも引く手あまただ。

これを受けて、統数研では「データサイエンス・リサーチプラザ」を設けている。統計思考院に企業の社員の席を確保し、各種プログラムを受講するなど、データサイエンスをじっくり学んでもらうものだ。

一方、大学でもデータサイエンティスト育成教育に注目し、滋賀大学をはじめデータサイエンス学部を創設する大学が増えてきた。統数研では、こうした大学の教員に対し、データサイエンスの講義設計をアドバイスするといった支援も手がけている。これは学術指導として行われており、その契約調整を行うこともURAの役割だ。

その他、国内外の企業との連携を深めていくためのワークショップ、シンポジウムなどの企画・運営支援やノウハウの提供も行う。

例えば、2015、2016年には産学連携の国際的な展開として、データ科学・計算機分野で学術界と産業界の連携のあり方をテーマとする「ISM HPCCON（High Performance Computing CONference）」を開催。2016年には他に、大型計算機・スパコンの利活用、データに基づく大学経営（IR）、次世代が求めるT型人材の育成の3分野をテーマとするコンファレンスを開催した。

また2018年には、IBMと連携し、「ICACON（IBM Cloud Academy CONference）2018」を開催。30を超える大学・機関の委員が企画するプログラムが好評を博した。

さらに、今年6月5日には統数研の設立75周年記念式典に合わせ、シンポジウムとオープンハウスを実施。100周年への次の25年に向けて、AI、IoT、ビッグデータ、セキュリティー、データサイエンスなどさまざまな分野との協働を深化させる機運を高めるとともに、統数研の研究教育活動を広く一般へ伝える機会となった（図4）。

教員と事務職をつなぎ、統数研と企業をつなぐURA。新たなシナジーを生み出す触媒として、さらなる活躍が期待されている。（広報室）

図4：統計数理研究所創立75周年記念式典と合わせて開催したオープンハウスでは、広く一般の人たちに向けたチュートリアル特別講座を実施した。

第19回

離散数学で解き明かす「細胞分化の木構造」

「基礎医学と社会医学をつなぐ離散幾何学的モデリング」

未来のイノベーションの芽を育む個人型研究「JSTさきがけ」。統計数理研究所の早水桃子助教の「基礎医学と社会医学をつなぐ離散幾何学的モデリング」は、国が推進するこのプログラムに採択された研究提案の一つだ。コンピューターサイエンスの基礎である離散数学という分野の研究からオープンソースのデータ解析ソフトウェアの開発までを手がけ、幹細胞工学など社会的に重要な領域にもつながる研究の一端を紹介する。

▲早水桃子助教

距離データに潜むグラフ構造を描出する「離散幾何学的モデリング」

早水助教が提案する「離散幾何学的モデリング」という枠組みは、有限個の点同士の非類似度が「距離」として数値化されていて、その非類似度のデータを「グラフ」（点と線からなる図）で記述したいという状況を扱うためのものだ。「複雑なデータや現象の全体像をシンプルなグラフで描出するという課題は色々な文脈で現れますが、特に生物学の研究では重要なものです」と早水。

例えば、形態学的な特徴やDNAの塩基配列などに基づく「距離」から進化の系統樹を再構築するというデータ解析は、離散幾何学的モデリングの古典的な例だ（図1）。距離の情報から系統樹というグラフ構造を知るための理論や方法は深く研究されてきた歴史があるが、生物学の世界には既存の数学的な概念や統計学的な手法だけでは太刀打ちできない課題が多数ある。例えば、生物の進化と同様に、細胞の分化もまた木構造で記述される現象だが、進化の道筋を推定する課題とはデータの種類や医学・生物学的なゴールが異なるため、細胞分化の木構造を解明する課題ならではの新しい方法論が必要だ。一方、農作物の品種改良のように、異なる種の掛け合わせで新しい種が生まれる「交雑」と呼ばれる現象は、系統樹のように単純な木構造では記述できないため、「系統ネットワーク」と呼ばれる系統樹よりも表現力の高い一般的なグラフ構造が必要になる。

早水のプロジェクトは、現代の生命科学のニーズに応えて、多様なデータや現象をモデル化するために、さまざまなグラフと距離の関係を明らかにしよう、という挑戦だ。「複雑なデータに潜むシンプルな構造を描出することは、データの可視化だけでなく、本質的に重要な生物学的発見にもつながります」と早水は語る。

現代の細胞生物学におけるデータ解析上の大きな課題

「動物の進化と、細胞の分化は、木で記述される現象という意味では似ていますが、数学的には大きな違いがあります」と早水は指摘する。図1の例では、①から⑥で示した現存種同士の距離がデータとして与えられ、それらの6点の他に●印の祖先を含むような系統樹が構築されているが、言い換えれば、6つの現存種の多様性を説明するために、●印の絶滅種が存在したと考えて木を描いたということだ。しかし、このような想像上の点を含めずに、「6種の関係性だけを描きたい」という状況もある。「細胞同士の関係性の全体像を木で説明したい」という問題意識は、まさにその一例だ。

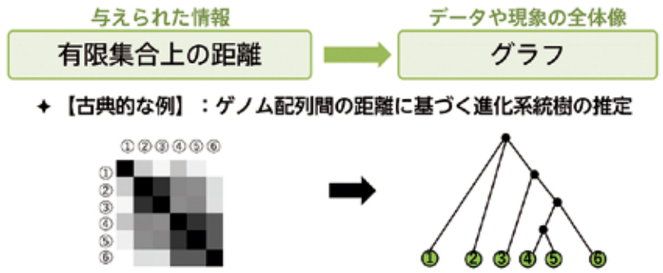


図1：「離散幾何学的モデリング」という枠組みの概要と、その典型的なデータ解析の例。①から⑥の点は現存する種、●の点は想像上の祖先を表す。

細胞の分化とは、細胞の形態や機能が変わり、別の種類の細胞に変化する不可逆的な運命決定プロセスを意味する。いったん変化すると決して元に戻らないので、生物学者たちは、細胞分化の全体像は合流のない木構造で記述できると考えている（図2）。このような細胞分化の木モデルは、細胞生物学者たちが試行錯誤を繰り返して長い歴史の中で築き上げてきたもので、人類がこれまでの実験から得た知見の集大成といえる。

ところが、近年の実験計測技術の進歩はめざましく、今日の細胞生物学は大きな転換期を迎えている。かつての技術水準では、細胞の集団の遺伝子発現の状態を大雑把に調べることしかできなかったが、「1細胞技術」と呼ばれる現代のテクノロジーを使えば、個々の細胞における各遺伝子の発現量を、細やかに、しかも網羅的に計測できる。このような新技術がもたらしたデータをうまく活用できれば、これまでの細胞生物学の常識を覆すような新発見ができるかもしれない。実際、1細胞の遺伝子発現データを活用して細胞分化の木構造の全貌を解き明かすことは、今や世界中の細胞生物学者たちが抱く一つの大きな夢になった。そこで脚光を浴びているのが、個々の細胞同士の距離を説明する木グラフを構築するという、まさに「離散幾何学的モデリング」の考え方なのである（図3）。

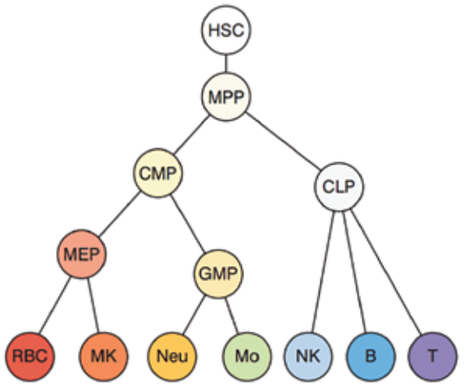


図2：血液に関係する細胞の分化の全体像を描いた木グラフ。造血幹細胞（HSC）は、どの種類の血液細胞にもなれるポテンシャルを持つが、最終的には赤血球（RBC）やTリンパ球（T）などの何らかの細胞に分化する。

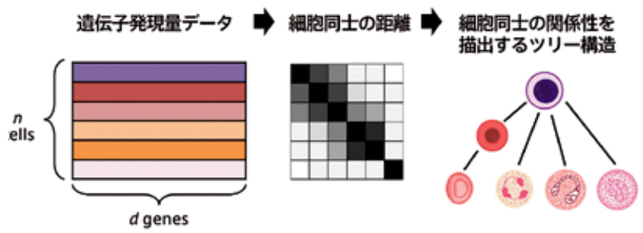


図3：細胞生物学者のニーズは、「個々の細胞から得られた詳細な遺伝子発現データを活用して、細胞分化の真の木構造を解明したい」というもの。

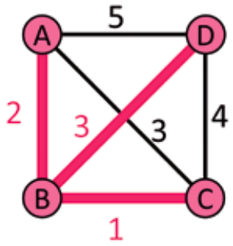


図4：4つの点からなる距離データを表す完全グラフと、その中の最小全域木。最小全域木を求めるアルゴリズムは、距離から細胞の木を構築する“経験則”として普及し始めている。

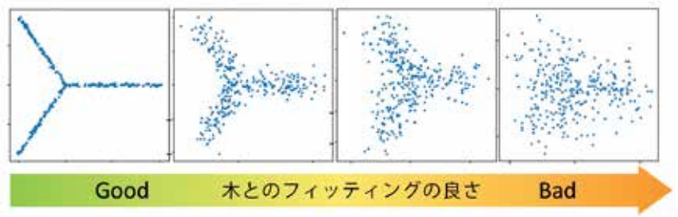


図5：最小全域木は元の距離情報をよく表わしていることもあるが、まったくフィットしていないこともある。

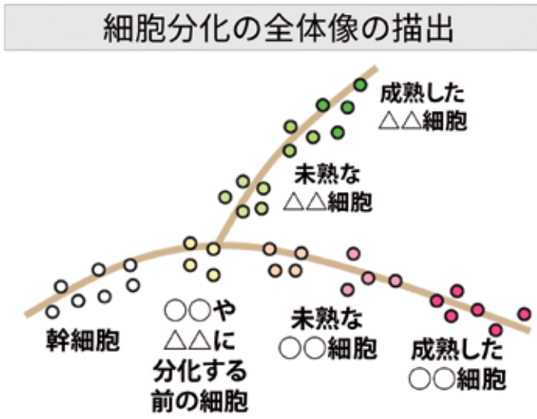


図6：プロジェクトのゴールのイメージ。細胞分化の全体像を推定して描出するような遺伝子発現データ解析ソフトウェアを開発する。

細胞分化の全体像を描き出す科学的に信頼できる方法論の開発

細胞同士の距離データから木グラフを抽出するための“経験則”として定着し始めたのは、離散数学の基本である「最小全域木問題」を解くアルゴリズムだ。これは、途中でサイクル（輪っか）ができないように距離が近いペアから順に結んでいき、すべての点を通る木が得られたら終了するという単純明快な方法だ。例えば、A～Dという4つの細胞間の距離が得られているとしよう（図4）。まずは一番短い辺BCを選び、その次に二番目に短い辺ABを選ぶ。三番目に短い辺は二つあるが、辺ACを結んでしまうと、ABCというサイクルが形成されてしまうので、辺BDのほうを選ぶ。以上で、A～Dのすべての点を通る木が得られたことになる。色を付けて強調されているのが、最小全域木だ。

「似た細胞同士を結んで木を作るという方法は直観的に自然ですが、最小全域木を求めるアルゴリズムで正しい細胞分化の木構造が得られるという理論的な保証はありません。科学的に信頼できる真に有用なデータ解析手法を創るためには、データと木モデルのフィッティングの良さを定量的に評価する尺度を確立させることが重要だと考えています」と早水は説明する（図5）。

このプロジェクトの最終的な目標は、与えられた距離情報に潜む本質的な木構造を抽出するための確かな方法論を構築し、それを実装したソフトウェアを開発し、世界中の生物学者が利用できる形で公開することだ（図6）。早水は「この研究を着想したのは、iPS細胞の分化メカニズムを研究している医学系の研究者の方からデータ解析上の相談をいただいたのがきっかけです。数学や情報の研究者だけでなく、医学を含む幅広い分野の研究者たちが使いこなせるものを作り、生命科学の研究に貢献したいと思っています」と語る。

生命科学の根底にある課題を数理科学の力で解決したい

早水はもともと医学部出身で、数理の研究者としては異色の経歴を持つ。

▲須藤功平氏（株式会社クリアコード代表取締役）

▲石井大海外来研究員（DeepFlow株式会社）

統計数理研究所で研究を始める前は、脳機能や代謝の程度を画像検査で調べる核医学分野が専門の放射線科医だった。「あまり聞き慣れない診療科かもしれませんが、放射性同位体を使って初期の認知症やがんなどの進行度合いを評価するという重要な分野で、統計学や物理学の知識は必須でした。目に見えないものをどうやって定量的に可視化するかという問題意識は、今の研究とも通じる部分があります」と早水は言う。

生命科学の知識と臨床経験に加えて、統計科学・数理科学・情報科学にわたる幅広い視野を持つ早水は、医学の研究者にとっても、数学の研究者にとっても、非常に心強い存在だ。とはいえ、異分野から転向した当初は、様々な分野との壁を感じることがしばしばあったという。数理の世界で医学部出身者はあまりにも珍しいうえに、医学の世界でも数学者とのコラボレーションに対して懐疑的な向きがあった。また、離散数学という分野自体も、統計の一部の研究者たちにとっては、あまり馴染みがないものに映ったようだ。「最近はようやく多くの方々に興味を持っていただけるようになりましたが、はじめのうちは、とにかく早く結果を出して、自分の研究の魅力や価値を伝えたいという気持ちがありました。大学院在学中に少し背伸びをして『さきがけ』に応募したのも、そういう背景があります」と早水は振り返る。

さきがけの研究タイトルに掲げたのは『基礎医学と社会医学をつなぐ』という言葉。早水は「幹細胞生物学が良い例ですが、医学の基礎研究のなかには、臨床医学や社会医学に対する影響力が大きいテーマがたくさんあります。そのような研究の突破口を、数理の力で開きたいのです」と思いを語る。

異業種連携でオープンソースのデータ解析ソフトウェアを開発

このプロジェクト紹介の取材には、早水が受け入れている統数研の石井大海外来研究員と、ソフトウェア開発を手掛ける民間企業、クリアコードの須藤功平代表が参加した。

石井は民間企業の研究開発者だが、もともとは数理論理学や計算代数が専門だった。「プログラミングの過程で離散数学やグラフ理論を使うことはありましたが、アルゴリズムを適用することの妥当性を定量的に評価するような研究は初めてなので新鮮です」と感想を述べる。専門が異なるからこそ、有意義なディスカッションができる部分があると早水も言う。

一方、クリアコードはフリーソフトウェアを開発していた須藤氏が立ち上げた会社だ。自分たちで書いたコードを独占するのではなく、誰でも自由に使える形で公開することで対価を得る開発スタイルを貫いている。

アメリカでは研究者とエンジニアが連携してソフトウェアを開発し、論文を発表するということが珍しくないが、日本ではまだあまり一般的ではない。論文の査読者や読者が内容を検証できるように、開発したデータ解析ソフトウェアを無償で公開しようと考えていた早水にとって、オープンソース・ソフトウェアを開発する企業の理念はうってつけだった。

須藤氏は「ソフトウェアを開発する技術に加えて、公開して普及させるのが、当社の得意とするところです。この強みをプロジェクトで生かしたいと思っています」と話す。

早水は「大学や研究所での研究プロジェクトに民間の意欲的なエンジニアを巻き込むことは、研究者側が積極的に動かないと難しい。私のプロジェクトが良い成功事例として広まり、日本でもこのような異業種連携がもっと一般的になればいいと思っています」と話す。医学と統計数理をつなぐ研究内容も先駆的なら、その進め方も新しい。さまざまな意味で「さきがけ」と言えるプロジェクトだ。（広報室）

第20回

「ものづくりデータ科学研究センターと共同研究部門」

データ科学を起爆剤として、従来のものづくりを刷新

ものづくりを戦略目標とするデータ科学の研究拠点として、統計数理研究所が2017年7月に設立した「ものづくりデータ科学研究センター」。2019年10月にはマテリアルズ・インフォマティクス（MI）のオープンソースプラットフォーム「XenonPy（ゼノンパイ）」を公開。また、三菱ケミカル（MCC）との共同研究部門「ISM-MCCフロンティア材料設計研究拠点」を立ち上げるなど、活発に取り組みを進めている。同研究センターの方向性や現状を紹介する。

▲吉田亮教授

データ科学と計算科学の融合で材料開発に新境地を拓く

これまで主に実験とシミュレーションによって進められてきた材料開発の分野が、データ科学の導入によって劇的に変わろうとしている。その取り組みが今、世界中で注目を集める「マテリアルズ・インフォマティクス（MI）」だ。

アメリカでは2011年に国家プロジェクト「Materials Genome Initiative（MGI）」がスタート。実験やシミュレーションのデータを利用した機械学習アルゴリズムによって、従来は10〜20年ほどかかっていた材料開発にかかる時間を半減することを目指すものだ。

これに続きヨーロッパでも2015年に「Novel Material Discovery Laboratory（NoMaD）」が、また中国では2016年に「北京マテリアルズ・ゲノム・エンジニアリング・イノベーション連盟」が設立された。

特に盛り上がりを見せているのが日本だ。物質・材料研究機構の情報統合型物質・材料研究拠点（NIMS）が以前から材料データの収集とデータベース整備に取り組んでおり、2015年7月に科学技術振興機構（JST）のイノベーションハブ構築支援事業として、情報統合型物質・材料開発イニシアティブ（MI^{2}I：“Materials research by Information Integration” Initiative）を始動。これが日本におけるMIの学術創成のきっかけとなった。現在は、統計数理研究所の吉田亮教授が物質・材料記述基盤グループのグループリーダーを務めている。

さらに、2017年には吉田がセンター長となり、統数研に「ものづくりデータ科学研究センター」が発足。データ科学の予測技術を駆使して、材料開発や特定の機能をもつ新物質の発見に取り組んでいる。

材料分野をデータ科学の応用領域の一つとして有望視する吉田は、「材料や物質の世界は物理法則や理論があるので、これらとデータ科学の組み合わせによって、どういう構造の物質をつくれば、こんな機能を発現するだろうということをある程度の精度で予測できます。実際に作って検証する段階では難しさもありますが、ワークフローは非常にシンプルです。計算科学やデータ科学によって高い到達点を目指せると信じています」と意欲を語る。

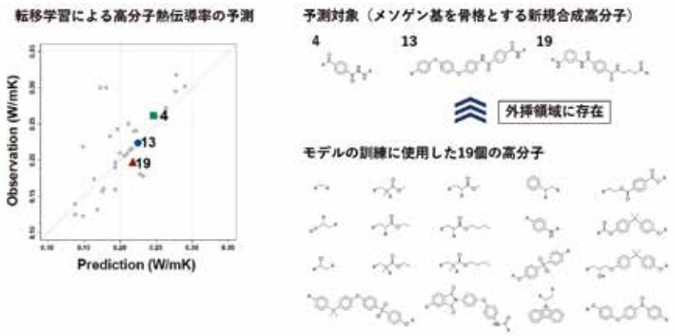


図1：転移学習による高分子物性（熱伝導率）の予測。高分子熱伝導率の19件の訓練データを用いて、XenonPy.MDLの高分子のガラス転移温度、低分子化合物の比熱容量等の訓練済みモデルを転移し、外挿性を持つ予測モデルを構築することに成功した（Yamada and Liu et al., ACS Cent Sci, 2019）。

オープンサイエンスの土壌づくりを目指すものづくりデータ科学研究センター

材料分野にデータ科学を応用し、成果を上げるためには、まず環境を整える必要がある。現状に足りないのは、オープンな研究開発の土壌だ。

材料研究のデータ量はデータ科学の他の応用領域に比べて圧倒的に少なく、MIの適用領域が限定されているのが実状。原因として、データ取得の高コスト性や、研究者のニーズの多様性によるコモンデータベース創出の難しさが挙げられる。だが、もっとも大きいのは、競合相手に対する情報秘匿の意識が高く、研究者にデータ公開に対するインセンティブが働きにくいことだ。

産業界ではデータ科学の有用性がここ2、3年で急速に認識されつつあり、成果も出始めているものの、オープンデータベースの開発は進んでいない。しかも、先端領域に近づくにつれてスモールデータの傾向はより顕著になる。コミュニティ全体でコモンデータを創出しようという動向も低調だ。

「画像認識や自然言語処理、AIなど、

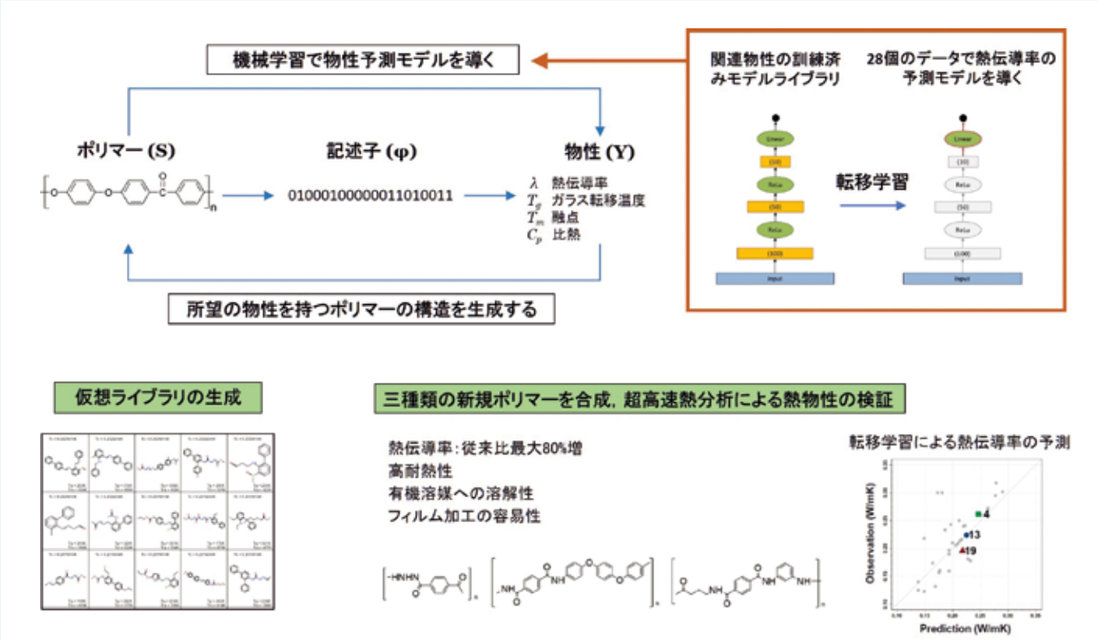
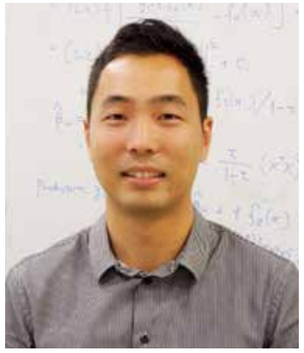


図2：三種類の高熱伝導性高分子の発見に至るワークフロー。転移学習を活用した熱伝導率の予測と分子設計の機械学習の技術が問題解決の突破口を切り拓いた（Wu et al., NPJ Comput Mater, 2019）。



▲Stephen Wu助教

データ科学をドライビングフォースにしたテクノロジーが発展している分野には、いずれもオープンデータの素地があります。材料分野にもオープンデータ、オープンサイエンスのカルチャーが育てば、飛躍的な発展の可能性があるでしょう」と吉田は指摘する。

こうしたなか、産学連携へ向けた動きも出始めている。MI^{2}Iコンソーシアムには法人会員82社、アカデミア会員10人が参加しており、双方から材料研究者、データ科学研究者、アルゴリズムの専門家など多彩な顔ぶれが集まっている。

ものづくりデータ科学研究センターが目指すのも、オープンな環境で産学連携の新たなモデルをつくることだ。「企業とタッグを組めば、理論・計算・実験それぞれの分野のトップレベル研究者と協働できる。いろいろなアイデアをもらえ、お互いに得るものが大きい」と吉田は言う。その言葉どおり、センターにはさまざまな企業や他分野のアカデミアの研究者がひっきりなしに訪れ、1日中活発なディスカッションが続く。

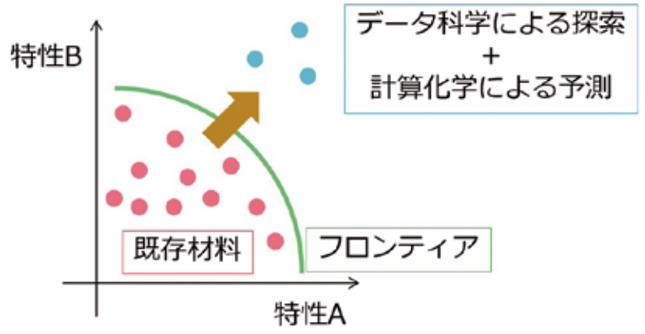


図3：三菱ケミカルとの共同研究部門による研究のイメージ。

「転移学習」により少数の訓練データから予測精度の高いモデルを導き出す

MIにおけるオープンサイエンスの一つの取組みとして、統数研のChang Liu特任助教とStephen Wu助教、吉田を中心とするものづくりデータ科学研究センターのチームとNIMSの共同研究グループが、2019年10月、物性予測タスク訓練済みモデルの包括的ライブラリ「XenonPy. MDL」を公開した。このモデルデータベースと「転移学習」と呼ばれる機械学習の技術を組み合わせることで少数の訓練データから優れた予測モデルを構築する可能性が示された。MI分野の現状を打開するものとして画期的だ。

XenonPy.MDLは低分子、高分子、無機材料の45種類の特性を対象として開発した約140,000個の機械学習の訓練済み予測モデルからなる。MIのさまざまなタスクを実行する機械学習アルゴリズムが実装されており、ユーザーはAPI経由でXenonPy. MDLの訓練済みモデルを再利用し、材料設計のさまざまなワークフローを構築できる。

このソフトウェアの鍵を握るのは、あるタスクで訓練されたモデルを他のタスクに転用する転移学習と呼ばれる解析手法だ。

機械学習は、訓練データとテストデータの類似性に基づいて予測を行うため、周辺に訓練データが全く存在しない領域では予測能力を失う。しかし、今回の共同研究では、関連するタスクで大量のデータを用いて事前学習を行い、獲得した汎用的な特徴抽出器を目標ドメインに転移すれば、目標タスクの訓練データが極めて少数だったとしても、転移されたモデルには既存データとは大きく異なるドメインの予測能力が備わっていることを明らかにした（図1）。この論文は、アメリカ化学会の論文誌ACS Central Scienceに掲載された。

「データ科学の応用領域として、MI分野は社会的・経済的にも影響力が大きく、やりがいがあります。将来的にはXenonPyをMIのオールインワンのプラットフォームに成長させることがチームの目標です」とWuは話す。

この他、“スモールデータ問題”がより顕著な「高分子MI」の分野でも、転移学習を応用することで成果を出している。ものづくりデータ科学研究センター、NIMS、東京工業大学の共同研究で、高熱伝導性高分子を設計・合成し、従来の高分子に比べて約80%の熱伝導率の向上に成功したもので、機械学習が自律的に設計した高分子が実際に合成・検証された初の事例となった（図2：Wu et al., NPJ Comput Mater, 2019）。さらに、2019年度からは第2ラウンドとしてJST-CREST熱制御領域「高分子の熱物性マテリアルズインフォマティクス」（代表：森川淳子（東工大）、機械学習・計算科学グループ代表：吉田）もスタートしている。

▲白鳥和矢三菱ケミカル主席研究員

材料メーカーとタッグを組み新規物質探索にブレークスルーを

産学連携のプロジェクトも立ち上げた。ものづくりデータ科学研究センターと三菱ケミカル（MCC）との共同研究部門「ISM-MCCフロンティア材料設計研究拠点」だ。両者の知見を持ち寄り、データ科学による解析技術と計算化学による予測技術を融合して新規物質探索を行うMIの基盤技術を構築することが目的（図3）。

データ科学に基づく物質探索では入力されたデータとの類似性から物性を予測するため、入力データに類似した範囲でのみ物質の探索が行われる。これに対し、計算化学では既存データの有無とは関係なく未だ現実には作成されていない材料、素材の性質を予測することが可能だ。

統数研の保有するデータ科学技術と、MCCの強みである計算化学技術を融合することで、既存データの範囲には含まれない革新的な特性を有する材料を見出すための物質探索アルゴリズムを構築するとともに、高分子、触媒、無機材料といった具体的な材料設計課題に適用しながらアルゴリズムの高度化を図る。

さらに、この研究部門で構築したアルゴリズムは、学術成果として積極的に発信していく。MI分野において、オープンイノベーション、オープンサイエンスを促進するチャレンジだ。

このプロジェクトに参画するMCCの白鳥和矢主席研究員は語る。「これまでの計算化学、材料分野のシミュレーションの多くは、物理法則に従って材料の性能を予測するものでした。分子など小さなサイズや短時間の現象であれば高精度に予測できる半面、物質探索においてはその計算コストが問題になることも多くあります。観点の異なるデータ科学の考え方を取り入れることで、革新的な特性を有する材料を見出すアルゴリズムが構築できると期待しています」。

同研究部門には、両者の研究員に加えて、MCCの親会社である三菱ケミカルホールディングスにおいてデジタル・トランスフォーメーションを推進する先端技術・事業開発室のデータサイエンティストも参加している。

「日本の強みであるものづくりや材料科学の技術をデータ科学と組み合わせれば、世界に勝てる戦略が見えてくるはずです」と吉田が話すように、チーム全体に従来のものづくりをMIで革新しようという気概が満ちている。（広報室）

写真：共同研究部門「ISM-MCCフロンティア材料設計研究拠点」のメンバー

編集後記

本冊子「響き合う人とデータ～統数研プロジェクト紹介」は、統計数理研究所ニュースのNo.128（2015年5月発行）からNo.147（2020年2月発行）まで5年間にわたって掲載された同名の記事20回分をまとめたものです。ここには既に終了したプロジェクトも含まれていますが、競争的外部資金による受託研究から、企業との共同研究、中期目標・中期計画に沿った独自の取組など、統計数理研究所が実施した様々なプロジェクトの一端を垣間見ていただけることと思います。冊子化にあたり最小限の編集を加えた以外、各記事の内容は統計数理研究所ニュースに掲載された当時のままとなっておりますことをご了承いただけますと幸いです。

統計数理研究所は、我が国唯一の統計数理に係る大学共同利用機関として、大規模・複雑なデータからの情報抽出・知識獲得・意思決定などデータサイエンスに関する基幹的な研究に取り組むとともに、統計的機械学習や次世代シミュレーションなどの研究領域に加え、リスク・ものづくりや医療健康など異分野融合・新分野創成に向けた領域横断型のNOE（Network Of Excellence）の形成、またリーディングDAT（Data Analytics Talents）をはじめ様々な社会需要に応じた高度人材育成に向けた統計思考力育成事業等を推進しています。

本研究所の活動については、統計数理研究所ニュース（年4回発行）の他、毎年一回発行する「要覧」や「年報」などの広報誌でご紹介しています。統計数理（年2回発行）などの学術刊行物を含め、ウェブサイトではそれぞれの最新版とバックナンバーをご覧いただけます。また、創立75周年記念式典で配布いたしました「統計数理研究所75年のあゆみ」や「研究業績紹介」もダウンロードいただけますので、ぜひご利用ください。

広報室長　伊藤　聡

響き合う人とデータ〜統数研プロジェクト紹介

編集／発行

大学共同利用機関法人 情報・システム研究機構
統計数理研究所

〒190-8562 東京都立川市緑町 10-3
Tel: 050-5533-8500（代表）
https://www.ism.ac.jp/

本誌に関するご意見・ご感想は下記までお送りください
ism_news@ism.ac.jp

2020年3月発行

大学共同利用機関法人 情報・システム研究機構
統計数理研究所
The Institute of Statistical Mathematics