情報処理学会論文誌　数理モデル化と応用　Vol.11　No.1　1–11 (Mar. 2018)

共同利用施設における実験終了後の研究成果数予測

神辺 圭一[1,a)]　諏訪 博彦[2]　篠田 孝祐[3]　栗原 聡[3]

受付日 2017年8月28日，再受付日 2017年10月16日，
採録日 2017年10月25日

概要：大型放射光施設 “SPring-8” は，国内外の産官学に開かれた共同利用施設であり，幅広い分野の研究開発に利用されている．本施設のリソースには限りがあることから，成果に基づいた施設運用が求められる．そのため，成果が増加・減少する研究領域の把握は，施設運用の方向性を考えるために重要である．研究成果は論文として公表されるケースが大半であるが，論文化には実験後 2，3 年を要する場合が多く，即時的な把握は困難という問題がある．そこで本論文では，研究施設の運営支援に活用することを目的に，実験期終了後 3 年経過時点の成果公開状況を事前に予測するモデルを構築した．その結果，相関係数 0.937 で予測できることを確認した．

キーワード：共同利用施設，研究成果数予測，ランダムフォレスト，回帰分析，機械学習，SPring-8

Prediction of the Number of Registered Publications after Experiments in the Shared Utilization Facility

Keiichi Shinbe[1,a)]　Hirohiko Suwa[2]　Kousuke Shinoda[3]　Satoshi Kurihara[3]

Received: August 28, 2017, Revised: October 16, 2017,
Accepted: October 25, 2017

Abstract: Large synchrotron radiation facility “SPring-8” is a shared utilization facility opened to domestic and abroad researchers of industry, government and academia. It is used for research and development in the wide range of fields. Due to limitation of resources, facility operation needs to be based on research outcomes. Therefore, understanding how research area/method is blooming or declining is essential to consider the direction of facility operation. Most of the research results are published as papers, but in many cases it takes 2 or 3 years after the experiments, which makes it difficult to figure out research trends immediately. In this paper, for the purpose of utilizing to support the operation of the research facility, we make a predicting model of the number of registered publications in advance after 3 years since the end of the experimental period. As a result, our model can be predicted with a correlation coefficient of 0.937.

Keywords: shared utilization facility, prediction of the number of published papers, random forest, regression analysis, machine learning, SPring-8

1. はじめに

国費を投入して整備・運用される共同利用施設は，利用者による利用研究成果を最大化し，学術の進歩と社会経済の発展に寄与する責務がある．そのため，限られた予算と人的リソースの中で，施設側のサービスを最適化することが肝要であり，これまでの利用実績をもとに今後成長が期待される研究領域（分野・手法）を予測することは，実験設備の更新をともなう将来計画の策定といった施設運用の方向性を考えるためにも重要である．だが，共同利用施設の研究成果である論文が公表されるまでには，実験終了から年単位の時間を要することが多いため，施設内の特定の実験設備を利用して発表された研究成果が今後どの程度増加または減少するかを，即時的に把握することは困難である．そこで本論文では，国内外の産学官の研究者等に開かれた共同利用施設である SPring-8（スプリングエイト）[*1]の

[1] 電気通信大学/高輝度光科学研究センター
The University of Electro-Communications, Chofu, Tokyo 182–8585, Japan / JASRI, Sayo, Hyogo 679–5198, Japan
[2] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology, Ikoma, Nara 630–0192, Japan
[3] 電気通信大学
The University of Electro-Communications, Chofu, Tokyo 182–8585, Japan
[a)] shinbe@spring8.or.jp

[*1] http://www.spring8.or.jp/
本論文の内容は，執筆者の見解に基づいてまとめられたものであり，執筆者の属する機関の公式見解を示すものではないことを付記する．

利用統計データに対して機械学習を適用し，将来の公表論文数を予測するモデルを構築することを目的とする．本モデルを利用することで，実験終了から一定期間経過後の公表論文数を事前に予測することが可能となり，今後発展が期待される研究領域のトレンド把握や計画から実施まで時間を要する設備更新を進める際の需要予測データとして活用されることが期待される．

2. SPring-8 の概要

SPring-8 は，兵庫県南西部の播磨科学公園都市に建設され，1997 年 10 月に供用を開始した大型放射光施設である．赤外線から硬 X 線までの光を使う，国内外の産学官の研究者に広く開かれた共同利用施設として，物質科学・地球科学・生命科学・環境科学・産業利用等の幅広い分野の研究開発に活用され，年間のべ約 1 万 5 千人が来所し，2 千件以上の実験すなわち利用研究課題（以下，課題）が実施されている．

SPring-8 には，“ビームライン”[*2]と呼ばれる特性の異なる実験設備が複数設置されている．研究者は，神辺ら [1] が構築し 2005 年から運用している専用のポータルサイト “SPring-8 User Information”[*3]（以下 UI サイト）上でユーザ登録を行い，Web ベースの課題申請書に研究の目的，手法，分野，希望利用時間，用途に応じた利用希望ビームライン等の情報を記入し，提出する．その後，科学・技術・安全上の観点から審査を受け，採択されると利用が可能となる．採択率は応募時期やビームラインによって異なるが，約 6 割である．施設側は，毎年 5・6 月，11・12 月にかけて課題の公募を行っており，おおむね 10 月〜2 月頃，4〜7 月頃に実験時間（ビームタイム）を提供している．また，夏期の長期点検期間を境に前期（A 期）・後期（B 期）の 2 つに運転サイクルが分かれているため，「2011B」「2013A」のように「年 + 期番号」で実施期を識別している．なお，運転期間中は 24 時間稼働であり，研究者は “シフト” 単位（1 シフト= 8 時間）で実験を行っている．

SPring-8 で実施される実験において，利用料を免除される課題（成果非専有課題）[*4]は全体の 8 割近くを占めるが，これらの課題を実施した研究者は，期終了後 3 年以内に「成果公開認定要件を満たす研究成果」（以下，認定成果）[*5]を公表し，UI サイトの成果データベースに発表媒体等の情報を登録する義務を負う．また，「成果公開の促進に関する選定委員会からの提言」[2] に基づき，2011B 期から，期終了後 3 年以内に正当な理由なく認定成果を登録しなかった研究者に対し，新たな課題申請書の受付けを行わない措置が開始された．そのため，過去に実施された課題の成果登録状況を定期的に確認し，期日内の成果登録を促す取り組みを行うことは，施設側の重要なミッションであるといえる．さらに，利用研究成果の最大化を達成するには利用動向に応じたビームラインの再整備が不可欠であるが，成果登録数の増減を各期終了後 3 年経過してから追跡し，その後の高度化計画に反映した場合，アップグレードしたビームラインが利用可能になるまでに少なくとも 4〜5 年の時間差が生じ，急激に発展する研究領域の受け皿になることが困難になることも考えられる．そこで本論文では，期終了後 3 年経過時点の認定成果の登録件数の予測を，課題申請数や研究分野・手法のカテゴリ，課題申請者の所属分類といった説明変数を特徴量とした機械学習モデルにより行う．これにより，たとえば成果登録数の減少が予測されるビームラインに対して成果登録の推進策をあらかじめ促したり，研究領域の趨勢を指し示す成果登録数を事前に推定したりすることで，中長期的な整備投資の判断資料として用いることが可能となる．

SPring-8 では，各課題における責任者を「実験責任者」と定義し，実験責任者と共同で実験を行う研究者を「共同実験者」と呼ぶ．また，実験責任者と共同実験者の総称を「ユーザ」と規定しているため，本論文でもこれらの呼称を使用する．

3. 先行研究

共同利用施設における利用効果の分析事例として，江端ら [3] による北海道大学オープンファシリティの使用者申請データに関する統計分析ならびに論文の謝辞情報のテキストマイニング分析の事例がある．当該施設の取組効果を測定する指標として「利用装置数」の把握が重要であること，また謝辞に北大オープンファシリティの名称を記述した論文は，平均より被引用数が多い傾向が見られたことから，謝辞情報をもとに共同利用施設の効果を測ることは有効であることが示されている．また，米谷ら [4] は，日本国内の大学の研究開発投資（インプット）と論文数（アウトプット）との関係を Web of Science のデータをもとに回帰分析し，研究者数および研究費と論文数には正の相関があることを明らかにしている．

昨今，機械学習手法が普及し様々な分野の予測に用いられている [5], [6], [7]．なかでも，論文発表から特定年経過後の引用論文数を予測する先行研究としては，イギリスの医学誌 BMJ のデータベースに登録された論文について発表から 2 年後の引用論文数を予測した Lokker ら [8] の回帰モデル分析や，MEDLINE データベースの登録論文に対し機械学習アルゴリズムの 1 つである SVM を用いて 10 年後の被引用論文が閾値以上になるかを予測した Lowrence ら [9] の研究，Web of Knowledge の書誌情報および著者

[*2] http://www.spring8.or.jp/ja/about_us/whats_sp8/facilities/bl/
[*3] https://user.spring8.or.jp/
[*4] 課題の種類については，https://user.spring8.or.jp/?p=672 参照．なお，利用料免除課題においても，消耗品等の実費負担分は別途請求される．
[*5] 定義の詳細は，https://user.spring8.or.jp/?p=748 参照．

情報をもとにSVMの回帰モデルであるSVRを適用して3年後の被引用論文数を予測したMatsuiら[10]の研究がある．また関ら[11]は，LokkerらやLowrenceらの先行研究に対し，以下の問題を指摘している．

- モデル構築に用いた特徴量が汎用的ではない．
- 一定以上の被引用論文数を獲得する論文は，論文全体の中では一部であるにもかかわらず，学習に用いたデータセットの半分以上が高被引用論文で占められている．
- 検証に用いたデータの説明変数が確定した時点では，学習用データの目的変数の値はまだ確定しておらず，結果的に未来のデータから生成されたモデルを用いて予測精度の評価を行っている．

これらの先行研究では，論文数と相関のある変数の分析や被引用論文数の予測モデルに関する提案は行われてきたものの，研究機関や共同利用施設における一定期間経過後の発表論文数自体を直接予測するものはなかった．そこで本論文では，SPring-8を利用して創出された成果である論文数を機械学習モデルによって予測し，予測値と実測値とを検証することで予測精度について議論する．

4. 予測モデルの構築

本章では，構築した成果登録数の予測モデルの概要について述べる．

4.1 提案モデルの概要

成果登録数の予測モデル構築に用いる学習データとして，UIサイトのデータベースに蓄積された各種データから予測に有効と考えられる複数の統計値を特徴量として抽出し，「課題情報」「研究分野・手法情報」「ユーザ属性情報」にグループ分けした．各グループを，本論文では“データセット”と呼ぶ．課題の応募数，採択数，利用者数といった主要な統計情報は，ユーザ向けオンライン情報誌“SPring-8/SACLA利用者情報”[*6]等のWebサイトで公表されている．

4.2 データセットに含まれる特徴量の構成

各データセットに含まれる特徴量の構成について述べる．

4.2.1 共通情報

以下の情報は，成果登録数の予測に関する基本的なパラメータであるため，すべてのデータセットに共通の特徴量として含まれる．

- 期番号（期名を整数値に置換）
- ビームライン（各ビームライン名を整数値に置換）
- ビームライン種別[*7]（共用ビームライン＝1，専用ビームライン＝2に置換）

[*6] https://user.spring8.or.jp/sp8info/

4.2.2 データセットA：課題情報

施設利用を希望する研究者は，具体的な使用希望ビームラインと希望シフト数を課題申請書に記入のうえ，課題審査を受ける．課題申請が採択された場合は，実験で使用するビームラインと利用可能なシフト数が正式決定するが，施設側が提供するビームラインが課題申請時の希望とは異なる場合もある．また，実験装置の不具合やユーザの都合による実験未実施といった事態も発生しうるため，ユーザが実験で使用したシフト数の合計値は，各期終了時点で初めて確定する．

予測モデルで使用する1つめの特徴量群として，課題申請数・希望シフト数ならびに課題終了後の実施数（キャンセルされた課題を除いた件数）・実施シフト数の合計値を期・ビームライン単位で集計し，用いた．具体的な特徴量群の構成を以下に示す．

- 申請課題件数
- 申請課題共用ビームライン件数
- 申請課題専用ビームライン件数
- 申請課題希望シフト数
- 申請課題共用ビームライン希望シフト数
- 申請課題専用ビームライン希望シフト数
- 実施課題件数
- 実施課題共用ビームライン件数
- 実施課題専用ビームライン件数
- 実施課題使用シフト数
- 実施課題共用ビームライン使用シフト数
- 実施課題専用ビームライン使用シフト数

4.2.3 データセットB：研究分野・手法情報

SPring-8で実施される課題は，研究分野・手法ともに多岐にわたる．そのため課題審査は，課題申請書に記載された希望審査分野に基づき，グループ分けしたうえで行われる．認定成果の公開（研究成果の論文化）に必要な平均期間や1課題あたりの平均成果登録数は研究分野・手法ごとに傾向が異なるため，これらを特徴量群に用いた．なお，研究分野・手法および希望審査分野は，課題申請書内に選択肢（大分類・小分類）が用意されており，申請者はいずれかのカテゴリを選択する必要がある[*8]．本論文では，このうち大分類のみを特徴量に使用した．具体的な特徴量群の構成を以下に示す．

- 実施課題希望審査分野［生命科学］件数
- 実施課題希望審査分野［散乱回折］件数
- 実施課題希望審査分野［XAFS・蛍光分析］件数
- 実施課題希望審査分野［分光］件数

[*7] ビームラインは，設置者の違いによって3タイプに分類される（https://user.spring8.or.jp/?p=976）．本論文では，このうち「共用」および「専用」ビームラインの成果登録数の予測を行う．

[*8] 希望審査分野，研究分野分類，研究手法分類の一覧は，https://user.spring8.or.jp/?p=1499からダウンロード可能な課題申請書下書きファイルに記載されている．

- 実施課題希望審査分野［産業利用］件数
- 実施課題研究分野［ビームライン技術］件数
- 実施課題研究分野［素粒子・原子核科学］件数
- 実施課題研究分野［生命科学］件数
- 実施課題研究分野［医学応用］件数
- 実施課題研究分野［物質科学・材料科学］件数
- 実施課題研究分野［化学］件数
- 実施課題研究分野［地球・惑星科学］件数
- 実施課題研究分野［環境科学］件数
- 実施課題研究分野［産業利用］件数
- 実施課題研究分野［その他］件数
- 実施課題研究手法［X 線回折］件数
- 実施課題研究手法［X 線散乱］件数
- 実施課題研究手法［X 線磁気散乱］件数
- 実施課題研究手法［X 線非弾性散乱］件数
- 実施課題研究手法［X 線・軟 X 線吸収分光］件数
- 実施課題研究手法［光電子分光］件数
- 実施課題研究手法［X 線イメージング］件数
- 実施課題研究手法［X 線光学］件数
- 実施課題研究手法［特殊環境実験］件数
- 実施課題研究手法［その他］件数

4.2.4　データセット C：ユーザ属性情報

課題申請書の申請を行った実験責任者の所属分類や実験のために SPring-8 に来所したユーザののべ人数，初利用者数といった，課題審査を経て採択された課題に関する情報を特徴量に用いた．具体的な特徴量群の構成を以下に示す．

- 実施課題実験責任者分類［大学等教育機関］のべ数
- 実施課題実験責任者分類［国公立研究機関等］のべ数
- 実施課題実験責任者分類［産業界］のべ数
- 実施課題実験責任者分類［海外］のべ数
- 来所のべ数
- 共用ビームライン来所のべ数
- 専用ビームライン来所のべ数
- 来所初利用数
- 共用ビームライン初利用数
- 専用ビームライン初利用数

4.3　学習アルゴリズムの検討

機械学習モデルの構築には，統計分析ソフトウェアの R 言語[*9]および統合開発環境の RStudio[*10]を用いた．また機械学習アルゴリズムは，用途に応じた様々な手法が提案されているが，本論文では集団学習アルゴリズムの 1 つであるランダムフォレスト [12] を成果登録数の予測に用いた．ランダムフォレストは，学習・評価速度が速く，説明変数の重要度（寄与度）が算出可能といった特徴がある．そこで，同一データセットから重回帰分析とランダムフォレストによるモデルをそれぞれ構築し，予測精度の比較を行った．

[*9] https://www.r-project.org
[*10] https://www.rstudio.com

4.4　特徴量群の絞り込みとチューニング

続いて，モデルの予測精度を高めるため，データセットの組合せを変えながら，予測結果がどのように改善されるかを調べた．さらに，重要度が高く判定された特徴量を組み合わせたデータセットを抽出し，モデルを再構築することで，予測精度を最大化する特徴量群の絞り込みを行った．

5.　予測モデルの評価実験

本章では，学習アルゴリズムおよびデータセットの違いによる予測精度の評価とモデルのチューニングによる精度改善の結果について述べる．

5.1　評価実験の概要

モデルの構築に使用する学習データには，2005B～2012B 期（7 年半，15 期分）のビームライン別集計値 606 件を用いた[*11]．

まずはじめに，アルゴリズムの違いによる予測精度の差異を評価するため，データセット A・B・C および全特徴量群を連結したデータセット（A＋B＋C）に含まれる実績データを用いてランダムフォレストおよび重回帰分析モデルを構築し，10 交差検証法[*12]による予測精度の評価を行った．重回帰分析に基づくモデル式の作成においては，ステップワイズ法[*13]による変数選択を行っている．

次に，データセット A・B・C および各データセットを連結した特徴量群（A＋B，A＋C，B＋C，A＋B＋C）における 2005B～2012B 期の実績データからランダムフォレストによる予測モデルを構築し，2013A 期の成果登録数の予測値と実測値との適合度を調べた．

SPring-8 では，定期的な公募課題に加えて，年間を通じてそのつど募集を行う課題制度や，スタッフの R&D 業務の一環で行うインハウス課題等が存在するため，応募・採択課題総数といった，ある期における利用実績データが完全に確定するタイミングは，A 期は 9 月末，B 期は 3 月末頃となる．したがって，本論文における A 期のデータは毎年 10 月 1 日早朝，B 期のものは毎年 4 月 1 日早朝にデータベースから取得したものを使用した．なお，予測対象である成果登録数は，期終了後 3 年経過時点の値であるため，2013A 期のデータがすべて確定した日時は，2016 年 9 月末であった．

[*11] SPring-8 の供用開始は 1997B 期であるが，UI サイトは 2005B 期から運用が始まったため，データセットも当期からとなる．
[*12] データを 10 分割し，検証データを 1 グループずつ取り出していく．残る 9 グループを学習データとするモデルを計 10 回構築し，各モデルから目的変数を予測することで，予測精度を評価する手法である．
[*13] 回帰式を構成する変数を組み換えながら，「モデルのよさ」の判定基準である AIC（赤池情報量規準）を最も改善する変数を選択する方法である．

5.2　ランダムフォレストと重回帰分析による予測精度の評価

10 交差検証法による予測値と実際の成果登録数との適合度の評価には，相関係数および RMSE（Root Mean Squared Error）を用いた．RMSE は，次の式によって求められる値で，0 に近いほど予測値と実測値との乖離が小さいことを示す．

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

n：予測対象数，y_i：実測値（成果登録数），$\hat{y}_i$：予測値

ランダムフォレストと重回帰分析の相関係数の比較を**表 1** に，RMSE の比較を**表 2** に示す．いずれのデータセットにおいても，ランダムフォレストの方が相関係数が高く，また RMSE が小さかったことから，成果登録数の予測にはランダムフォレストが有効であることが確認された．よって，今後の分析にはランダムフォレストを用いて行う．

5.3　データセット別の予測精度の比較

次に，各データセットからランダムフォレスト予測モデルを構築し，2013A 期の成果登録数の予測を行った．相関係数および RMSE の比較を**表 3** に示す．

表 3 の結果から，データセットは単体で学習データに用いた場合よりも複数組み合わせてモデル構築した方が，予測精度は高くなることが確認された．だが，特徴量の中には予測への寄与が低いものも含まれており，これらの特徴量が予測精度の低下の原因となることも考えられる．そこで次節では，全特徴量を結合したデータセット A＋B＋C から特徴量の取捨選択を行い，予測精度の改善を行う．

表 1　相関係数の比較
Table 1　Comparison of correlation coefficients.

	A	B	C	A＋B＋C
ランダムフォレスト	0.873	0.909	0.893	0.908
重回帰分析	0.773	0.822	0.812	0.853

表 2　RMSE の比較
Table 2　Comparison of RMSE.

	A	B	C	A＋B＋C
ランダムフォレスト	6.424	5.595	5.964	5.565
重回帰分析	8.310	7.469	7.646	6.857

表 3　データセットの組合せによる相関係数と RMSE の比較（2013A 期の成果登録数の予測）
Table 3　Correlation coefficient and RMSE for each data set.

	相関係数	RMSE	特徴量数
A	0.914	6.010	15
B	0.909	6.374	28
C	0.923	5.734	13
A+B	0.935	5.349	40
A+C	0.929	5.505	25
B+C	0.930	5.497	38
A+B+C	0.934	5.309	50

5.4　特徴量の重要度に基づく説明変数の取捨選択と予測精度の改善

データセット A＋B＋C を学習データに用いたランダムフォレスト予測モデルにおける特徴量の重要度（Increased Mean Squared Error，以下 IncMSE）を求めた．IncMSE は，各特徴量がモデルにどのぐらいの影響があるかを，ランダムフォレストの学習に用いられなかったデータを利用して評価した値である．モデルへの影響度の大きい特徴量ほど，IncMSE の値は高く算出される．各特徴量の重要度の順位を**表 4** に記す．なお，表中の「種別」は，各特徴量が前述のデータセット A，B，C のいずれかまたはすべてのデータセットに含まれているかを示している．

さらに，特徴量を IncMSE の高い順に並べ替え，最上位から特徴量を 1 つずつ追加したデータセットを計 50 個作成し，ランダムフォレストで学習を行った．各データセットの予測モデルにおける，2013A 期の成果登録数の予測値と実測値の適合度（相関係数，RMSE）を**図 1** に示す．

その結果，IncMSE 上位 13 位までの特徴量を含んだ予測モデルが，相関係数・RMSE ともに最適な値を示すことが分かり，14 位以降の特徴量を加えた場合に予測精度が低下することが判明した．そこで，IncMSE 上位 13 位までの特徴量を含むモデルを，本論文では「チューニングモデル」と呼ぶことにする．全特徴量を用いた予測モデル（すなわちデータセット A＋B＋C）における相関係数は 0.934，RMSE は 5.309 であるが，対してチューニングモデルの相関係数は 0.937，RMSE は 5.157 となり，重要度に基づく特徴量の絞り込みによって予測精度の改善が確認された．チューニングモデルの特徴量の構成は，前述の表 4 の第 1 位から第 13 位（区切り線の上）までが該当する．

これらの特徴量の組合せによるチューニングモデルが，2013A 期以前の予測についても有効であるかを確認するため，過去の期の成果登録数を，予測対象期以前の学習データをもとに推定した場合の相関係数および RMSE の検証を行った（**表 5**）．相関係数の最低値は，学習期間：2005B～2011A 期・予測対象：2011B 期における 0.842 であり，また直近 2 期（2012A・2012B 期）においては 0.935・0.924 の値を示していることから，特定期の予測にオーバフィッティングしたモデルではないことが確認された．

チューニングモデルに基づく 2013A 期の成果登録数の予測値と実測値をビームラインごとに取得し，実測値の高いビームラインから並べ替えたグラフを**図 2** に示す．なお，グラフ中に具体的なビームライン名は表示していない．

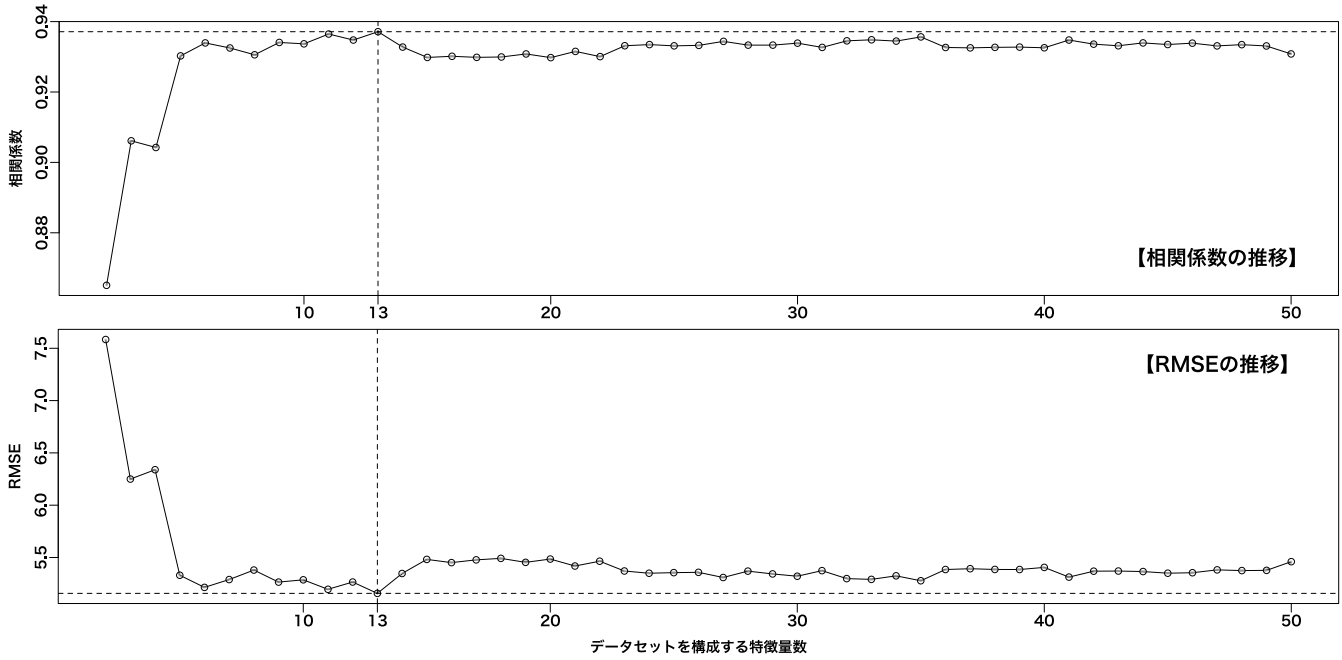

図 1　重要度（IncMSE）の高い順に特徴量を 1 つずつ加えて作成した計 50 個のデータセットをランダムフォレストの学習データに用いた場合の，各モデルにおける 2013A 期の予測値・成果登録数の適合度の推移（上図：相関係数，下図：RMSE）

Fig. 1　Fitness transition of predicted and actual value in 2013A period for each random forest model created by adding feature one by one in descending order of IncMSE (Upper figure: correlation coefficient, lower figure: RMSE).

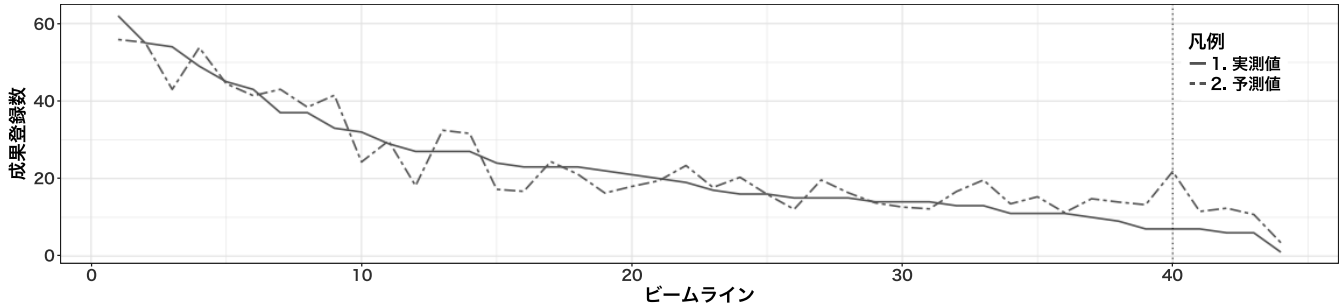

図 2　2013A 期の成果登録数の予測値と実測値
※ Y 軸方向の点線は，予測値と実測値が最も乖離したビームラインを示す

Fig. 2　Predicted and actual value of registered publications in 2013A period.

6. 考察

本章では，はじめにランダムフォレストと重回帰分析のモデル構造の違いについて定性的な側面から言及する．続いて，チューニングモデルで選択した特徴量群の構成について考察し，2013A 期の成果登録数の予測値と実測値との乖離の大きかったビームラインについて，原因を分析する．最後に，本論文の予測対象である将来の成果登録数を，どの時点の確定データから予測することが妥当であるかについて検討する．

6.1 学習アルゴリズムのモデル構造に関する考察

本論文では，ランダムフォレストと重回帰分析の各手法に対して予測精度に関する評価実験を行い，相関係数および RMSE といった定量的な観点から，ランダムフォレストの予測精度が重回帰分析よりも優れていると結論づけた．

定性的な視点で 2 つの手法を比較すると，重回帰分析は線形モデルであり各因子には従属関係がない一方，ランダムフォレストは決定木を弱学習器とした集団学習アルゴリズムであることから，因子間に暗黙的な従属関係が存在している違いがある．ランダムフォレストの回帰モデルでは，複数の決定木（回帰木）[*14]の平均値を予測結果に用いているため，決定木分析のように変数間の階層構造を可視化することは困難であるが，アルゴリズム内部に決定木が組み込まれていることはすなわち，研究活動に関わる複数

[*14] 本論文では，決定木の生成数を 1,000 とした．複数回の試行の結果，おおむね 500 以上であればモデルが安定することを確認している．

表 4　特徴量の重要度の順位
※区切り線は，重要度上位 13 位までを示す．

Table 4　Descending order list of features by IncMSE.

順位	特徴量名	種別
1	実施期	共通
2	実施課題件数	A
3	実施課題実験責任者分類［大学等教育機関］のべ数	C
4	来所のべ数	C
5	実施ビームライン	共通
6	実施課題研究分野［物質科学・材料科学］件数	B
7	申請課題件数	A
8	実施課題研究手法［光電子分光］件数	B
9	実施課題実験責任者分類［国公立研究機関等］のべ数	C
10	共用ビームライン来所のべ数	C
11	実施課題研究手法［X 線回折］件数	B
12	実施課題実験責任者分類［産業界］のべ数	C
13	実施課題研究分野［産業利用］件数	B
14	来所初利用数	C
15	実施課題研究分野［生命科学］件数	B
16	共用ビームライン初利用数	C
17	申請課題共用ビームライン件数	A
18	実施課題研究手法［X 線・軟 X 線吸収分光］件数	B
19	実施課題実験責任者分類［海外］のべ数	C
20	申請課題共用ビームライン希望シフト数	A
21	申請課題希望シフト数	A
22	実施課題研究分野［地球・惑星科学］件数	B
23	実施課題研究分野［化学］件数	B
24	実施課題共用ビームライン件数	A
25	実施課題専用ビームライン件数	A
26	実施課題希望審査分野［産業利用］件数	B
27	実施課題研究手法［X 線非弾性散乱］件数	B
28	実施課題研究手法［X 線散乱］件数	B
29	専用ビームライン来所のべ数	C
30	実施課題研究手法［X 線イメージング］件数	B
31	実施課題希望審査分野［散乱回折］件数	B
32	申請課題専用ビームライン件数	A
33	実施課題希望審査分野［XAFS・蛍光分析］件数	B
34	実施課題使用シフト数	A
35	申請課題専用ビームライン希望シフト数	A
36	実施課題専用ビームライン使用シフト数	A
37	実施課題共用ビームライン使用シフト数	A
38	実施課題希望審査分野［生命科学］件数	B
39	実施課題研究分野［ビームライン技術］件数	B
40	実施課題研究手法［特殊環境実験］件数	B
41	実施課題研究手法［その他］件数	B
42	実施課題研究分野［その他］件数	B
43	実施課題希望審査分野［分光］件数	B
44	実施課題研究分野［環境科学］件数	B
45	実施課題研究分野［医学応用］件数	B
46	実施課題研究手法［X 線光学］件数	B
47	専用ビームライン初利用数	C
48	実施課題研究手法［X 線磁気散乱］件数	B
49	ビームライン種別	共通
50	実施課題研究分野［素粒子・原子核科学］件数	B

表 5　2005B 期を起点として学習データを 1 期ずつ増やした場合の予測精度の検証
※最小学習期間は 2005B～2007B 期の 5 期分とした．

Table 5　Verification of prediction accuracy when learning data is incremented by one period starting from 2005B period.

学習期間	予測対象期	相関係数	RMSE
2005B～2007B	2008A	0.858	6.141
2005B～2008A	2008B	0.859	4.576
2005B～2008B	2009A	0.872	7.199
2005B～2009A	2009B	0.883	5.215
2005B～2009B	2010A	0.895	6.210
2005B～2010A	2010B	0.908	4.797
2005B～2010B	2011A	0.879	9.383
2005B～2011A	2011B	0.842	13.204
2005B～2011B	2012A	0.935	7.369
2005B～2012A	2012B	0.924	6.387

の因子の従属関係に基づいて成果が創出されることを示唆している．本論文では，IncMSE に基づく各特徴量の定量的な評価は行ったが，成果創出に寄与する複数の特徴量間の相互作用の解明については，今後の課題としたい．

6.2　チューニングモデルで選択した特徴量に関する考察

チューニングモデルの 13 個の特徴量群には，データセット A・B・C に由来するものがそれぞれ 2 個，4 個，5 個含まれており，モデルの構成要素にまったく用いられないデータセットは存在しなかった．また，全データセットに共通する 3 個の特徴量群のうち，「実施期」「実施ビームライン」という，予測対象の成果登録数の傾向を最も端的に象徴すると考えられる特徴量は構成要素に含まれていた一方で，ビームラインの運用形態を示す「ビームライン種別」は，重要度が低く判定され，チューニングモデルには用いられなかった．これは，「ビームライン種別」は「ビームライン」ごとに一意に決まり，同一ビームライン内や実施期ごとに変遷するパラメータではないため，「実施ビームライン」の特徴量で代替できたものと推測される．

データセット別に着目すると，データセット A に基づくものとして，「実施課題件数」「申請課題件数」がチューニングモデルの特徴量群に含まれていた．一方，シフト数（実験時間）の累計値等の特徴量は閾値以下となったが，これは研究分野・手法ごとに 1 課題あたりの平均シフト数は異なるものの，成果登録数の予測の観点においては，課題数の影響の方が相対的に強かったためと考えられる．

データセット B からは，研究分野・手法ともに 2 つのカテゴリがチューニングモデルの特徴量群に取り込まれた．モデルに用いられた研究分野である「物質科学・材料科学」と「産業利用」，研究手法の「光電子分光」と「X 線回折」は，それぞれ対応するビームライン群が大きく分かれており，成果登録数の傾向を表現するパラメータとして，重要

度が高く判定されたと考えられる．

データセットCに由来する特徴量群には，当該ビームラインを利用したユーザののべ数である「来所のべ数」に加え，「共用ビームライン来所のべ数」という共用ビームラインに限定したユーザ数の集計値が含まれていた．これは，専用ビームラインの場合，各期のユーザ層に大きな変化がない一方，共用ビームラインはユーザの流入・流出が継続的に発生しているため利用者数に変動があり，共用ビームラインの成果登録数の傾向予測に本パラメータが寄与したためと考えられる．また，「大学等教育機関」「国公立研究機関等」「産業界」といった実験責任者の所属分類に関する特徴量が複数含まれていたが，これは大学・研究機関と産業界のユーザでは前者の方が論文による成果公表への意欲が相対的に高いため，成果登録数の予測パラメータにこれらの集計値が影響したものと推測される．

6.3　予測値と実測値の乖離に関する考察

2013A期の成果登録数の予測値と実測値は，最大で14.82の差異が生じた．図2のY軸方向に点線を引いた部分（成果登録数第40位のビームライン）が該当箇所にあたり，予測値21.82に対し，実測値は7であった[*15]．当該ビームラインの現場の担当者に，2013A期の成果登録数の落ち込みについて確認したところ，当該期は機器の不調により採択課題数が通常よりも少なくなってしまったこと，また実施された課題についても当初の予定どおりに測定できなかったといった事実が判明した．したがって，当該ビームラインにおける予測値と実測値との乖離は，本論文のモデルに含まれていない要因による影響が大きかったものと考えられる．

6.4　将来予測の起点と予測先の期間に関する考察

関ら[11]は，X年を起点としてY年経過後の実測値を目的変数する予測モデルにおいて，X＋Z年後からY年経過時点の予測値を検証データとして用いる場合，Z≦Yの区間では未来のデータから生成されたモデルをもとに値を推定することになり，予測可能性に関する議論が行えないと論じている（**図 3**）．

本論文の目的は，認定成果の公表年限である「期終了後3年経過時点」の成果登録数の予測することであるが，関らの指摘に基づくと，2013A期の期終了「直後」に目的変数を推定するには，学習データとして2010A期までの確定値[*16]からモデルを構築する必要があることになる（**図 4**）．

そこで，2010A期までのデータを学習に用いたモデル（以下，制限モデル）を別途構築し，2013A期の予測精度

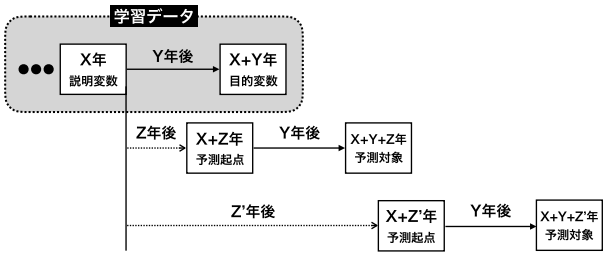

図 3　学習データに使用する期間と予測対象期との概念図
※Z≦Yの場合，X＋Z年時点では確定していない未来の値を予測モデルの学習データに含む．

Fig. 3　Conceptual diagram of periods used for learning data and prediction target period.

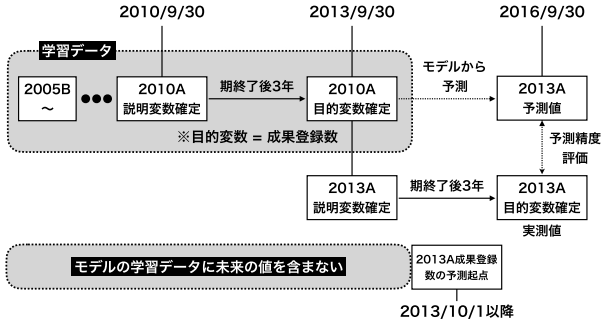

図 4　本論文における学習データと予測対象期との関係

Fig. 4　Relationship between learning data and prediction target period in this paper.

の追加検証を行った．制限モデルにおけるビームラインごとの成果登録数の予測値と実測値の差異を示したグラフを**図 5**に示す．なお，予測値は，すべて1.5倍掛けて補正を行っている（理由は後述）．

制限モデルの相関係数は0.847，RMSEは8.811となり，チューニングモデル（図2）と比べ，予測精度が低下することが確認された．これは，学習データの減少に加え，2011B期から，期終了後3年以内に成果登録を行わなかった実験責任者に対し，新たな課題申請書の受付けを行わない制度が開始されたことにより，実施課題総数に対する成果登録数すなわち成果登録率が劇的に改善された影響があげられる[*17]．2005B期以降の期別の成果登録率の推移を**図 6**に図示する．

つまり，2010A期までの学習データには，成果登録率の改善にともなう成果登録数の増加傾向が織り込まれていないため，予測精度の低下につながったものと考えられる．また，制限モデルの予測結果を俯瞰すると，ほぼすべてのビームラインにおいて，予測値が実測値よりも下振れしていたことから，成果登録の義務化前の2011A期と義務化後の2011B期では，成果登録率が57.5%から89.3%まで上昇したことに着目し，登録率の上昇倍率（1.553 ≒ 1.5倍）を予測値に掛けることで実測値との乖離を補正した．これに

[*15] 成果登録数は必ず整数値をとる．
[*16] 2010A期の目的変数である成果登録数が確定したのは期終了後3年経過した2013年9月末であり，2013A期の説明変数も同じ日に確定した．したがって，翌日以降に2013A期の終了後3年経過時点の成果登録数を予測するのであれば，学習データに未来の値は含まないことになる．
[*17] 表5における，予測対象期：2011Bの相関係数が表中の最低値を示したものも，同様の原因によるものと考えられる．

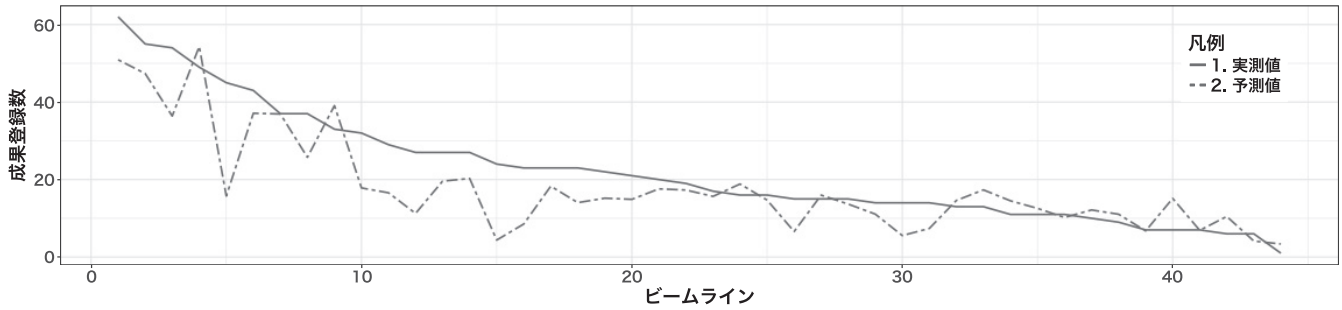

図 5　学習データを 2010A 期までに制限した場合の成果登録数の予測値と実測値
※予測値は，すべて 1.5 倍掛けて補正を行っている．

Fig. 5　Predicted and actual value of registered publications when learning data is limited to 2010A period.

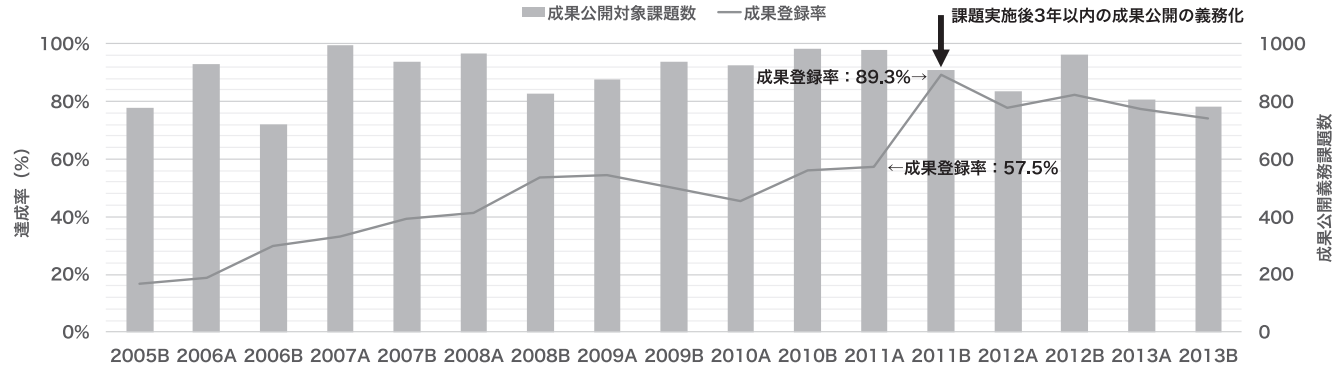

図 6　2005B～2013B 期の成果登録総数と成果登録率の推移

Fig. 6　Trends in the total number of registered publications and the publication registration rate for 2005B–2013B periods.

より，RMSE は 13.477 から 8.811 まで改善されている．

チューニングモデルと（図 2）と制限モデル（図 5）の 2013A 期の予測精度を比較した場合，前者の最大誤差が 14.82（予測値：21.82，実測値：7）に対し，後者は 29.26（予測値：15.74，実測値：45）であった．したがって，制限モデルはビームライン全体の大まかな成果登録数の予測には利用可能ではあるものの，当該期の全ビームラインにおける平均成果登録数 22.27 以上の最大誤差が生じていることから，成果登録数の将来予測を精緻に行うには，2011B 期以降のパラメータを学習データに組み込むことが必要であるといえる．だが，「未来の値」を学習データに含まずに，予測対象期の終了直後に 3 年経過時点の成果登録数を予測するには，少なくとも 2011B 期の目的変数が確定した 2015 年 3 月末時点の学習データをもとにモデルを構築し，検証用データとして 2014B 期以降の実績値と比較する必要がある．そのため，モデルの妥当性については 2014B 期の終了後 3 年経過時点の 2018 年 3 月末以降にしか議論できないことになることから，2012B 期までの確定データをもとにモデルを構築し，2013A 期の予測値と実測値との適合度について検証した本論文では，「期終了後 2 年半経過時点で，半年後の成果登録数を予測」した場合のモデルであるととらえ，学習データの確定までの日時と検証用データの時間差を最小にすることを優先した．

なお，2011B 期以降の成果登録率は大きく変動していないことから，今後，運用制度等に大幅な見直しがなく，同様の傾向が当面続くと仮定すれば，実際の運用上は，期終了後直後に 3 年経過時点の成果登録数を予測するモデルとして「チューニングモデル」は利用できるものと考える．

7.　結論

本論文では，大型放射光施設 SPring-8 で実施された成果非専有課題に対する期終了後 3 年経過時点の成果登録数をビームライン単位で予測するモデルを構築した．予測モデルのアルゴリズムにはランダムフォレストを使用し，学習データについては「課題情報」「研究分野・手法情報」「ユーザ属性情報」に関する特徴量を用いた．各特徴量群に対し予測精度が高くなる組合せを検証した結果，複数のデータセットを結合した学習モデルの方が単体のデータセットよりも良好な値を示した．

さらに，ランダムフォレストの計算過程で算出される特徴量の重要度（IncMSE）の高いものから特徴量を 1 つずつ足し合わせたデータセットを用意し，それぞれの学習データに対してモデルを構築のうえ，予測精度の評価を行った．その結果，重要度上位 13 位までの特徴量を足し合わせた学

習モデルの相関係数が最も高く，RMSE は最小となった．当該モデルを，本論文では「チューニングモデル」と位置付けている．

また，チューニングモデルに対して，予測値と実測値との乖離が大きいビームラインの状況を確認したところ，予測対象期においては「機器不調による実施課題数および成果登録数の減少」といった，本論文の特徴量には含まれていない要素が影響していたことが判明した．ビームラインごとの運転時間や機器の稼働状況といった，予測精度のさらなる向上に寄与しうる特徴量の組み込みと評価については今後の課題である．

成果登録数は，研究分野・手法によって差はあるものの，実施課題数の母数が多いほど増加する傾向にある．実施課題数は，ビームラインの特性や研究分野，競争倍率，実験に供出できる時間等の複合的な要素によって決まるため，数の大小によってビームラインのアクティビティを単純に評価することはできず，また成果登録数についても同様である．研究領域の盛衰を映し出すビームラインの成果創出効果を総合的に評価するには，実施課題に対する成果登録数すなわち成果登録率や，登録論文自体のインパクト，被引用論文数といった複数の指標が必要となる．ビームラインの将来計画に寄与する指標として，次は成果登録率の予測を行い，成果登録数との関係について分析を進めたい．

参考文献

[1] 神辺圭一, 松本　亘：共同利用施設利用者を支援する Web 申請システムの開発と運用, デジタルプラクティス, Vol.3, No.2, pp.155–163 (2012).

[2] 高輝度光科学研究センター：成果公開の促進に関する選定委員会からの提言, 入手先 ⟨https://user.spring8.or.jp/ui/wp-content/uploads/recommendation_20101027.pdf⟩ (参照 2017-06-23).

[3] 江端新吾，伊藤裕子：大学の先端研究機器共用施設の研究活動への効果の把握—北大オープンファシリティを事例として，文部科学省 科学技術・学術政策研究所 DISCUSSION PAPER，No.113 (2015).

[4] 米谷　悠，池内健太，桑原輝隆：大学の論文生産に関するインプット・アウトプット分析：Web of Science と科学技術 研究調査を使った試み，文部科学省 科学技術政策研究所 DISCUSSION PAPER，No.89 (2013).

[5] Jordan, M.I. and Mitchell T.M.: Machine learning: Trends, perspectives, and prospects, *Science*, Vol.349, Issue 6245, pp.255–260 (2015).

[6] Li B., Yang G., Wan R., Dai X. and Zhang Y.: Comparison of random forests and other statistical methods for the prediction of lake water level: a case study of the Poyang Lake in China, *Hydrology Research*, Vol.47, Issue S1, pp.69–83 (2016).

[7] 河村一輝，諏訪博彦，小川祐樹，荒川　豊，安本慶一，太田敏澄：飲食店向け不動産営業を支援する申込み顧客推薦モデルの提案，人工知能学会論文誌，Vol.32, No.1, pp.WII-O_1–10 (2017).

[8] Lokker, C., McKibbon, K.A., McKinlay, R.J., Wilczynski, N.L. and Haynes, R.B.: Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study, *Bmj*, Vol.336, Issue 7645, pp.655–657 (2008).

[9] Lawrence D.F. and Aliferis, C.F.: Using content-based and bibleo-metric features for machine learning models to predict citation counts in the biomedical literature, *Scientometrics*, Vol.85, Issue 1, pp.257–270 (2010).

[10] Matsui, T., Kanamori K. and Ohwada H.: Predicting Future Citation Count Using Bibliographic and Author Information of Articles, *International Journal of Machine Learning and Computing*, Vol.4, No.2, pp.139–141 (2014).

[11] 関　喜史，松尾　豊：論文の引用情報を用いた論文被引用数予測，第 25 回人工知能学会全国大会論文集，Vol.25, pp.1–4 (2011).

[12] Breiman, L.: Random forests, *Machine learning*, Vol.45, Issue 1, pp.5–32 (2001).

神辺 圭一　（正会員）

2001 年九州大学理学部生物学科卒業．2003 年同大学大学院人間環境学府発達・社会システム専攻（教育学コース）修士課程修了．2004 年（公財）高輝度光科学研究センター入社．以来，SPring-8・SACLA 利用者支援システムの開発・運用・高度化ならびにデータ分析業務に従事．現在，電気通信大学大学院情報システム学研究科社会知能情報学専攻博士後期課程（社会人枠）在学中．日本教育工学会，CIEC 各会員．

諏訪 博彦　（正会員）

1998 年群馬大学社会情報学部卒業．2006 年電気通信大学大学院情報システム学研究科博士後期課程修了．博士（学術）．2014 年 10 月より奈良先端科学技術大学院大学助教．社会情報システムに関する研究に従事．

篠田 孝祐　（正会員）

2004 年北陸先端科学技術大学院大学知識科学研究科博士後期課程修了．博士（知識科学）．現在，電気通信大学大学院情報理工学研究科助教．マルチエージェントシステム，社会シミュレーション，複雑ネットワーク分析に興味を持つ．人工知能学会会員．

栗原 聡　（正会員）

慶應義塾大学大学院理工学研究科修了．NTT 基礎研究所，大阪大学大学院情報科学研究科/産業科学研究所を経て，2012 年より電気通信大学大学院情報理工学研究科教授．同大学人工知能先端研究センターセンター長．博士（工学）．人工知能，複雑ネットワーク科学，ユビキタスコンピューティング等の研究に従事．『人工知能とは』（近代科学社）．翻訳『スモールワールド』（東京電機大学出版）等．人工知能学会，電子情報通信学会，日本ソフトウェア科学会，ACM 各会員．