Vol.2017-MPS-115 No.3
2017/9/25

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]

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．そこで本稿では，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

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元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，実験設備の更新を伴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．そこで本稿では，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の拡充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．

[1] 電気通信大学／高輝度光科学研究センター
[2] 奈良先端科学技術大学院大学
[3] 電気通信大学
[a)] shinbe@spring8.or.jp

[*1] http://www.spring8.or.jp/

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．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，専用のポータルサイト[*3]でユーザー登録を行い，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，概ね 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．

[*2] http://www.spring8.or.jp/ja/about_us/whats_sp8/facilities/bl/
[*3] https://user.spring8.or.jp/

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，ポータル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」[1] 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け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．そこで本稿では，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・手法のカテゴリー，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，例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事前に促したり，中長期的な整備投資の判断資料としても用いることが可能となる．

SPring-8 では，各課題の責任者を「実験責任者」と定義し，実験責任者及び共同で実験を行う研究者の総称を「ユーザー」と呼ぶため，本稿でもこれらの呼称を使用する．

3. 先行研究

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，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について発表から 2 年後の引用論文数を予測した Lokker ら [2] 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ひとつである 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 [3] の研究，Web Of Knowledge の書誌情報及び著者情報を元に 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 [4] の研究がある．

これらの先行研究では，被引用論文の予測モデルに関する提案は行われてきたものの，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．そこで本稿では，SPring-8 で利用して創出された認定成果である登録論文数を機械学習モデルによって予測し，実測値との差異を検証の上，予測精度について議論する．

4. 予測モデルの構築

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，ポータル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」「ユーザー属性情報」にグループ分けした．各グループを，本稿では“データセット”と呼ぶ．

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 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上，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ー都合による実験未実施といった事態も発生しうるため，ユーザー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・希望シフト数並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．

4.2.2 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，グループ分けした上で行われる．認定成果の公開（研究成果の論文化）に必要な平均期間や一課題あたりの平均成果登録数は研究分野・手法毎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・手法及び希望審査分野は，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ーを選択する必要がある[*6]．本稿では，このうち大分類のみを特徴量に使用した．

4.2.3 データセット C：ユーザー属性情報

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 に来所したユーザーののべ人数，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．

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[*7]及び統合開発環境の RStudio[*8]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，本稿では集団学習アルゴリズムの 1 つであるラ

[*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
[*6] 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
[*7] https://www.r-project.org
[*8] https://www.rstudio.com

ンダムフォレスト [5] 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．

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，データセットの組み合わせを変えながら，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

5.1 評価実験の概要

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 724 件を用いた[*9]．

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 及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及び重回帰分析モデルを構築し，10 交差検証法[*10]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[*11]による変数選択を行っている．

次に，データセット A・B・C 及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，年間を通じて都度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．従って，本稿における 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 期のデータが全て確定した日時は，2016 年 9 月末であった．

[*9] SPring-8 の供用開始は 1997B 期であるが，ポータルサイトは 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

[*10] 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

[*11] 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

表 1 相関係数の比較

	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

	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	8.31	7.469	7.646	6.857

表 3 データセットの組み合わせによる相関係数と RMSE の比較（2013A 期の成果登録数の予測）

	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

5.2 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，相関係数及び 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 に近い値であるほど予測値と実測値との乖離が小さいことを示す．

$$RMSE = \sqrt{\frac{1}{n}\sum_{k=0}^{n}(y_i - \hat{y_i})}$$

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も，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．相関係数及び 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．

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（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，特徴量を重要度（IncMSE）の高い順に並び替え，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，本稿では「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．

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毎に取得し，実測値の高いビームラインから並び替えたグラフを **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．

6. 考察

6.1 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全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，

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

順位	特徴量名	種別
1	実施期	共通
2	実施課題件数	A
3	共用ビームライン来所のべ数	C
4	実施課題実験責任者分類［大学等教育機関］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	来所初利用数	C
10	実施課題実験責任者分類［国公立研究機関等］のべ数	C
11	実施課題研究手法［X 線回折］件数	B
12	共用ビームライン初利用数	C
13	実施課題研究分野［産業利用］件数	B
14	実施課題実験責任者分類［海外］のべ数	C
15	実施課題研究分野［生命科学］件数	B
16	来所のべ数	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	専用ビームライン初利用数	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	実施課題実験責任者分類［産業界］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	共用ビームライン来所のべ数	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

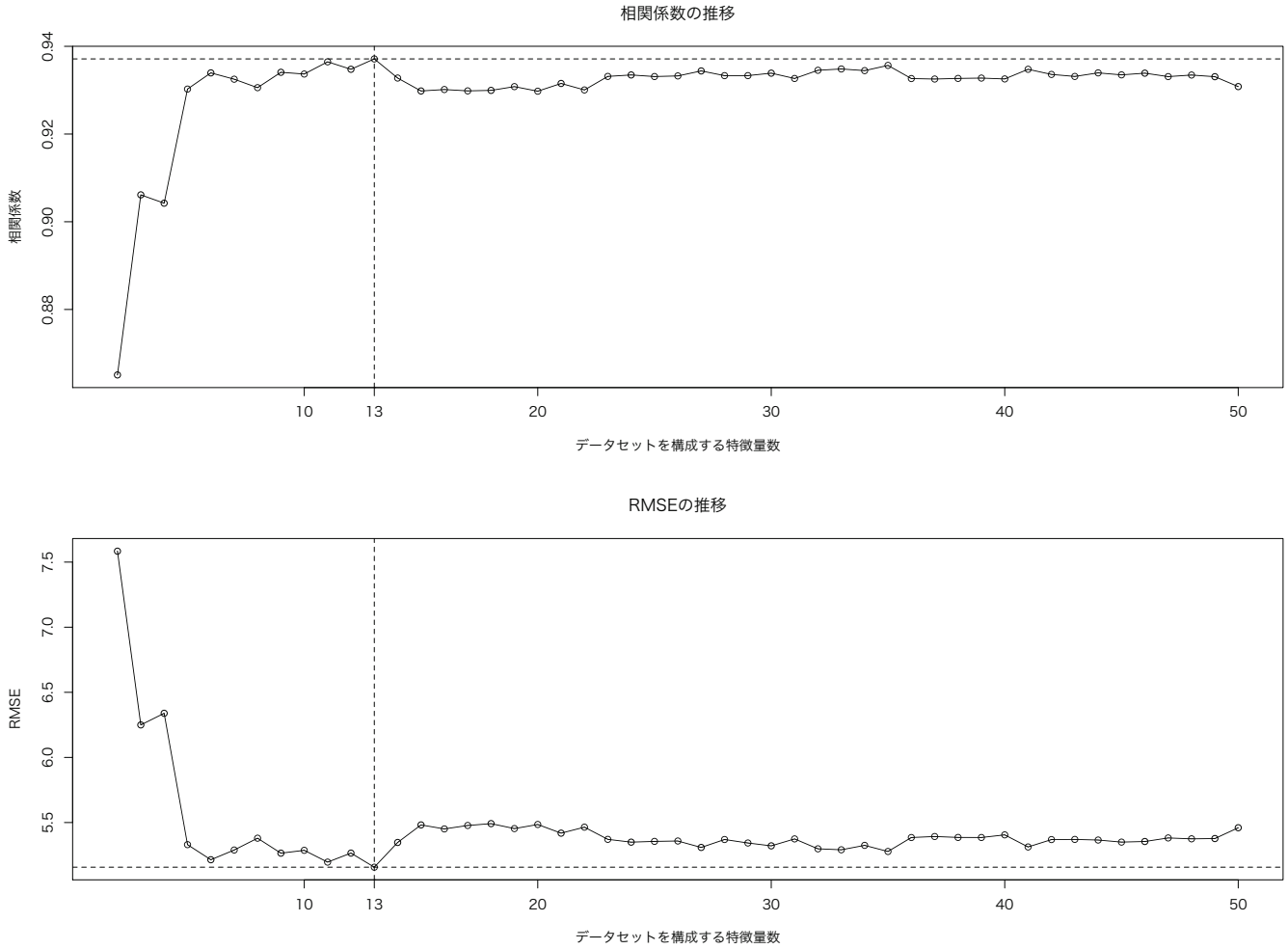

図 1 重要度（IncMSE）の高い順に特徴量を 1 つずつ加えて作成した計 50 個のデータセットのランダムフォレスト予測モデルに基づく，2013A 期の予測値・成果登録数の適合度の推移（上図：相関係数，下図：RMSE）

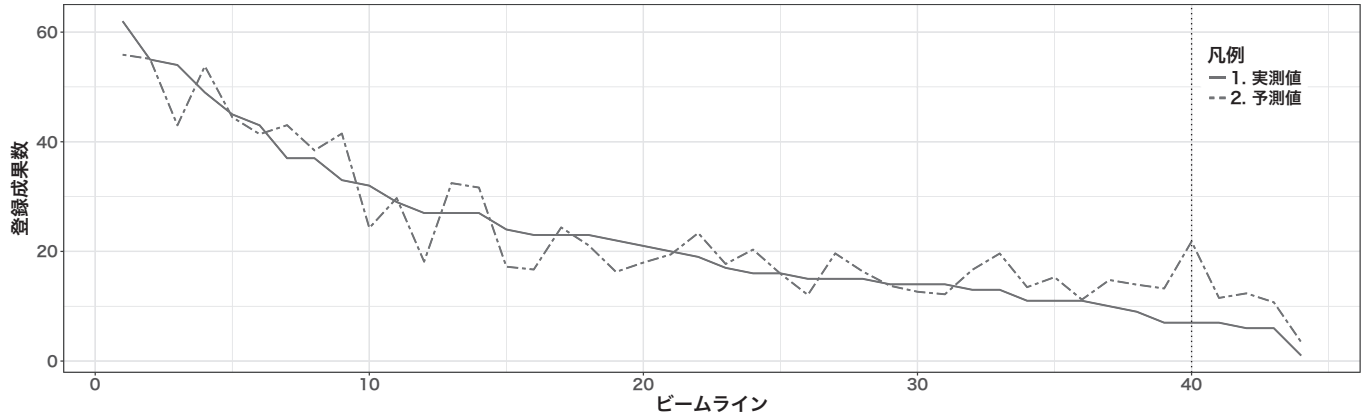

図 2 2013A 期の成果登録数の予測値と実測値
※ Y 軸方向の点線は，予測値と実測値が最も乖離したビームラインを示す

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」毎に一意に決まり，同一ビームライン内や実施期毎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・手法毎に一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 つのカテゴリー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初めて利用したユーザーのユニーク数である「来所初利用数」に加え，「共用ビームライン来所のべ数」「共用ビームライン初来所者数」といった共用ビームラインに限定したユーザー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，各期のユーザー層に大きな変化がない一方，共用ビームラインは新規ユーザーの流入が継続的に発生しているため，共用ビームラインの成果登録数の傾向予測にこれらのパラメータが寄与したためと考えられる．また，実験責任者の所属分類として，「大学等教育機関」「国公立等研究機関等」ののべ人数が特徴量として用いられているが，これは「産業界」のユーザー層よりも当該ユーザー層の方が，論文等による成果公表を行う意識が相対的に高いため，結果として成果登録数の予測パラメータとしての重要度が高くなったと推測される．

6.2 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 17.24 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 24.24 に対し，実測値は 7 であった．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通りに測定できなかったといった事実が判明した．従って，当該ビームラインにおける予測値と実測値との乖離は，本稿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．

7. 結論

本稿では，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」「ユーザー属性情報」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み合わせを検証した結果，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作成の上，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，本稿では「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及び成果登録数の減少」といった，本稿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．ビームライン毎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] 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⟩ (2017.06.23).
[2] 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, 336, pp. 655-657 (2008).
[3] 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, 85(1), pp. 257-270 (2010).
[4] 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, 4(2), pp.139-141 (2014).
[5] Breiman, L.: *Random forests*, Machine learning, 45(1), pp.5-32 (2001).

正誤表

下記の箇所に誤りがございました．お詫びして訂正いたします．

訂正箇所	誤	正
2 ページ 3 章	論文ににについて	論文について
3 ページ 5.2 節の式	$RMSE = \sqrt{\frac{1}{n}\sum_{k=0}^{n}(y_i - \hat{y}_i)}$	$RMSE = \sqrt{\frac{1}{n}\sum_{k=0}^{n}(y_i - \hat{y}_i)^2}$
4 ページ表 4 の第 3 位	共用ビームライン来所のべ数	実施課題実験責任者分類［大学等教育機関］のべ数
4 ページ表 4 の第 4 位	実施課題実験責任者分類［大学等教育機関］のべ数	来所のべ数
4 ページ表 4 の第 9 位	来所初利用数	実施課題実験責任者分類［国公立研究機関等］のべ数
4 ページ表 4 の第 10 位	実施課題実験責任者分類［国公立研究機関等］のべ数	共用ビームライン来所のべ数
4 ページ表 4 の第 12 位	共用ビームライン初利用数	実施課題実験責任者分類［産業界］のべ数
4 ページ表 4 の第 14 位	実施課題実験責任者分類［海外］のべ数	来所初利用数
4 ページ表 4 の第 16 位	来所のべ数	共用ビームライン初利用数
4 ページ表 4 の第 19 位	専用ビームライン初利用数	実施課題実験責任者分類［海外］のべ数
4 ページ表 4 の第 29 位	実施課題実験責任者分類［産業界］のべ数	専用ビームライン来所のべ数
4 ページ表 4 の第 47 位	共用ビームライン来所のべ数	専用ビームライン初利用数
6 ページ 6.1 節	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初めて利用したユーザーのユニーク数である「来所初利用数」に加え，「共用ビームライン来所のべ数」「共用ビームライン初来所者数」といった共用ビームラインに限定したユーザー数の集計値が含まれていた．これは，専用ビームライン	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ーののべ数である「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ー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，各期のユーザー層に大きな変化がない一方，共用ビー

	の場合，各期のユーザー層に大きな変化がない一方，共用ビームラインは新規ユーザーの流入が継続的に発生しているため，共用ビームラインの成果登録数の傾向予測にこれらのパラメータが寄与したためと考えられる．また，実験責任者の所属分類として，「大学等教育機関」「国公立等研究機関等」ののべ人数が特徴量として用いられているが，これは「産業界」のユーザー層よりも当該ユーザー層の方が，論文等による成果公表を行う意識が相対的に高いため，結果として成果登録数の予測パラメータとしての重要度が高 くなったと推測される．	ムラインはユーザーの流入・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」「国公立等研究機関等」「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ー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．