OSS の参加者は何を求め，何を得ているのか？ - 利用者と参加者の行動の視覚化 -

増田 礼子
フェリカネットワークス株式会社
Ayako.Masuda@FeliCaNetworks.co.jp

松尾谷 徹
有限会社 デバッグ工学研究所
matsuodani@biglobe.jp

要旨

オープンソース・ソフトウェア *(OSS : Open Source Software)* 作品は広く利用され，利用者視点からの分析は消費者行動の枠組みで調査・研究が行われている．一方，その作品を支えているエンジニア *(*広く参加者と呼ぶ*)* の行動についての調査・分析には課題がある．オープンサイエンスにおけるナレッジの有用性評価の一つの指標として参照量が考えられるが，ナレッジが持つ価値は多様であり，参照量だけでは評価できない．我々はエンジニアが *GitHub* からどのように学び，成長するのかについて，ナレッジの参照や影響を有用性の特性として捉え，指標化を試みている．我々の研究目的はエンジニアの育成や組織化における「意欲」を *OSS* 活動の「行動」から学び活用することである．ここでは，その入口として探索的な計測と分析について報告する．

課題は,「行動」を分析するために何を測りどのように分類するのか，である．本研究では，有用性の特性を捉えるため，時系列での型の違いに注目した．本研究では，時系列分析を用いて行動記録の視覚化を行い，他の分野では見られない特徴的な型を抽出できることを示す．この分析の前処理としての課題に，膨大な数の対象から，分析対象を合理的に選び出すための層別がある．この課題については順位を用いたクロス表により対処した．

1. はじめに

現代の ICT (Information and Communication Technology) 社会において，オープンソース・ソフトウェア (OSS : Open Source Software) 作品は，高度化し，かつ広く利用され，社会インフラとなっている．先端企業がビジネス戦略として先端技術を OSS に公開することもあり，利用技術中心に膨大な数のエンジニアが参加し活動を支えている．

エンジニアの働き方や意欲という観点から，彼らの活動の推進力を研究し活用できれば有用性は非常に高いが，調査や分析には課題がある．OSS として一括りにするには，あまりにも対象が大きすぎて多様であることと，行動を測り分析する方法が開発されていないことが要因であると考える．

オープンサイエンスにおけるナレッジの有用性評価の一つの指標として参照量が考えられるが，ナレッジが持つ価値は多様であり，参照量だけでは評価できない．我々は，インターネット・コミュニティ[1)] における，エンジニアの働き方に関して動的な観点から研究しており，GitHub [2] のデータを用いたデータサイエンスの探索的な分析 [3] を行っている．具体的には，エンジニアが GitHub のナレッジからどのように学び，成長するのかについて，ナレッジの参照や影響を有用性の特性として捉え，指標化を試みている．ナレッジの参照や影響を指標化するには，ナレッジの有用性を特性に分解し焦点を絞る必要がある．本研究では，この分解の手段として，GitHub のタイムスタンプによる記録データを時系列で視覚化する方法を試みた．

本研究では，アンケートやインタビューなどの「認知」アプローチではなく，log 情報等の記録を用いた「行動」アプローチでの分析を試みる．行動を特徴づける変数として時系列に着目し，作品の利用者と参加者の行動記録を分析した．分析の結果，得られた行動を特徴づける

「型」が観察された．ここでは，観察された型と，分析の方法について報告する．

この研究の前段階として，膨大な OSS 作品の中から，エンジニアの行動を探るための調査対象をどのように絞り込むのかという課題がある．本研究では，順位データを用いたクロス表を用いて対応した．

3 章では，分析に用いたデータの説明と，べき分布の視覚化，利用者行動と参加者行動の組合せを層別した結果を示す．4 章では，時系列での視覚化の方法とその結果を示し，5 章で時系列的な相関について分析し，特徴的な結果を示す．最後に，6 章で本研究のまとめとして総括する．

2. 行動に着目する理由と分析の手段

我々はエンジニアの働き方や意欲の研究対象として，OSS 作品を選び，その参加者と利用者について「行動」からの分析を試みている．本章では，本研究の背景や進め方について述べる．

2.1. 研究の背景

本研究で対象とするのは，OSS 作品の参加者と利用者の行動である．利用者の作品に対する評価や選択に関する研究は，消費者行動として盛んに行われ，ランキングなどが公開されている．一方，参加者の行動は就業活動や企業内での業務活動とは明らかに異なるが，その詳細は明らかになっていない．「OSS の参加者」の行動とは，作品を提供している何らかの組織に参加し，協働することである．何らかの組織とは，一種のインターネット・コミュニティであるが，OSS 作品を開発し公開するなど，何らかの活動目的を持っており，参加自由な交流の場としてのコミュニティとは異なる．

「プロのエンジニアでも難儀な作品に向かって活躍する彼らの原動力は何か」,「原動力はどこにあるのか」,「コミュニティとの関係はどのようなものなのか」など，これらの謎が解ければ，エンジニアの成長と働き方に寄与できると考えている．

2.2. どこに集まるのか

本研究で調査対象とした GitHub [2] の作品提供の基となる Repository[2] の新規登録数は，1 日で 数万件を超えている．既存の Repository に新規に参加する参加者の数も膨大だと推測される．

そこで，参加者の行動を調べる対象として，膨大な数の作品のどこに利用者や参加者が集まるのかを調べ，サンプルを層別して絞り込む必要がある．この課題の背景には，OSS における量的な分布の型がべき分布であるということがある [4]．べき分布を可視化するには，Zipf グラフ [5,6] や，分布の割合に着目したローレンツ曲線 [7–9] が有用である．これらの先行研究では，順位に基づく分析が行われていることから，本研究でも順位を使った層別を試みた．この層別については，3 章で述べる．

2.3. 時系列変化はあるのか

ソフトウェアの活動分析で時系列を扱う場合は少ない．バグ成長曲線や課題管理の見積もりと実績に用いる程度で，科学的な技法の応用には至っていない．

データサイエンスにおける時系列分析は，さまざまな分野で予測に利用されている実用技術である [10,11]．時系列分析では，定期周期でサンプリングされた情報 (時系列データ[3]) を用意して，変化の成分を「トレンド[4]」「周期変動[5]」「ノイズ[6]」に分解してモデル化を行う．

ここでの分析目的は，多種多様な作品や参加者の行動に対し，共通した尺度で比較することが出来る分析手法である．分析の精度は，定量的なモデル化まで求めず，探索的な分析で定性的な差を視覚化することを求めている．この実用化については，4 章で示す．

2.4. 時系列での相関はあるのか

OSS においては，利用者の評価が開発者をモチベートするのか，といった利用者と参加者の行動間の相関についても明らかになっていない．特に，時間軸で観測した行動間で相関がある場合があるのかについて分析する．

4 章で示すように，時系列の型は多様であるため，5 章では特徴的な型の組合せから利用者行動と参加者行動の影響の有無を概観する．明らかに影響していないケースが多数ある一方，連動しているものも観測された．これらの観測結果から，明らかになったことを 6 章でまとめ，報告する．

3. 分析対象とその分布

膨大な OSS 活動の場において，利用者や参加者はどこに集まるのか．本章では，膨大な OSS 作品の中から，参加者である開発者の行動分析のための調査対象をどのように絞り込み，収集するのかという課題に対し，事例を通して考える．3.1 節で分析データについて説明し，3.2 節で分布特性を考慮し，3.3 節で分析データを用いた層別を例示する．

3.1. 分析データ

本研究で用いるデータは，GitHub から取得したものであり，参加者や利用者の識別は，GitHub の Repository を単位に行っている．詳細な定義については GitHub の資料 [12] に従っているが，本研究で用いる主なものは次の 2 つである．

- **利用者の行動を代用する特性**：Repository に付与された Star[7] の数や付与した利用者と付与日などのデータ．
- **参加者の行動を代用する特性**：Repository の Issue[8] の数や起票者，起票日などのデータ．

分析には，次に示すサンプルデータ A と サンプルデータ B の 2 種類のデータを用いた．

- **サンプルデータ A (以降，サンプル A と記す)**：n = 1,287．2015 年 3 月 15 日から 1 週間の間に新規に登録された 117,757 件から 6 年経過後も保守が行われているものを抽出した．
- **サンプルデータ B (以降，サンプル B と記す)**：n = 1,395．先行研究 [4] で用いたデータを 2022 年 2 月現在の値でアップデートし，抽出した．データの概略は，2010 年 1 月 1 日から 2019 年 5 月 31 日までの期間内の 200 日をサンプリングして取得したものである．

データの収集は両サンプルとも GitHub API v3 [15] を用いて行った．

3.2. べき分布の扱い方

Star 数の分布や，Issue 数の分布はべき分布と呼ばれる分布を示し，正規分布を基礎とする一般的な統計手法では取り扱うことが困難である．具体的な例で示すと，サンプル A の作品が獲得した総 Star 数は，435,295 件であり，上位 12 件 (0.9 %) で総 Star 数の 50 % を獲得している．

この種の分布を扱う方法としては，順位と対数変換した獲得数や活動量 (Star 数や Issue 数) を用いる方法がある．図 1 は，サンプル A の Star 数の密度分布を示し，図 2 はサンプル A の Star 数，サンプル B の Star 数と Issue 数の対数変換 (y 軸) と順位 (x 軸) を示している．

対数変換し順位で並べると直線に近づく特性は，文書中の単語の出現数の研究において古くから知られており，Zipf の法則と呼ばれている [5,6,16]．もう一つの表現方法は，経済学で用いられているローレンツ曲線であり，母集団の大きさに差があっても比較することができる [7–9]．例としてサンプル A と B の Star 数のローレンツ曲線を図 3 と図 4 に示し，差を確認した．

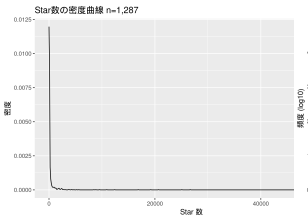

図 1: A：Star 数の分布

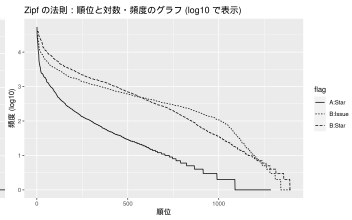

図 2: B：Zipf グラフ

図 3: A：Star 数

図 4: B：Star 数

分布の数値まで見たい場合には図 2 で示した Zipf グラフが，分布の割合を他と比較したい場合には図 3 と図 4 で示した正規化して視覚化されるローレンツ曲線が有用である．いずれも順位を基にデータを取り扱うので，本研究でも順位を使った層別を行う．

3.3. データの層別

4 章以降では，サンプル B を対象に，さらに詳細な分析として Repository 単位で時系列分析を行う．合理的に研究を進めるには，データを層別して探索的な分析を進める必要がある．ここでは，利用者行動 (Star) と参加者行動 (Issue) の 2 変数から層別を行う．

連続量である順位をカテゴリ化して，クロス表を作成する．カテゴリ化はデータを四分割して行った．Star 数を少ない順に {A,B,C,D}，同様に，Issue 数を {a,b,c,d} と表し，表 1 に示すクロス表を作成した．

表 1: Star 数と Issue 数のクロス表

		Issue 数			
Rank		a	b	c	d
Star 数	A	227	72	25	22
	B	92	116	91	51
	C	22	96	124	108
	D	8	63	110	168

このクロス表を基にした層別の結果およびコミュニティ特性の差については，4 章で視覚化して示す．

4. 時系列特性

本章では，活動の変化を時間経過から捉え，さらに層別を行う．時間経過による分析は，時系列分析と呼ばれ「時系列データ」を用いる．本研究では，時系列分析を行うに当たり，データ収集や前処理に課題がある．この課題について，4.1 節で述べる．

もう一つの課題は，膨大な数の作品の中から，調査にコストのかかる時系列分析対象をどのように選ぶのかである．4.2 節では，3 章の分析結果を利用し，精度の高い時系列分析が可能で，時系列のタイプ分けが可能なサンプル集団について述べ，4.3 節では，実際に層別した時系列の型について，参加者と利用者に分けて紹介する．

4.1. 時系列の視覚化の方法

時系列分析を用いた研究では，時系列の数学モデルを作成し，例えば季節商品の需要予測など定量的な取り扱いができる [10, 11]．モデルの要素は「トレンド」と呼ばれる時間的な定常状態と「周期変動」と「ノイズ」である．

ここでの分析は，探索的に時間特性を概観することが目的であるため，数学モデルには立ち入らず，そもそもトレンドのような定常状態が存在するのか，といった事前調査的な位置付けである．そのため，分析に用いるデータも，整形された時系列データではなく，利用者や参加者の活動記録のタイムスタンプを利用する．具体的な手法は，タイムスタンプのあるデータ列を時間順に並べ，平滑化帯域幅を用いて帯域内のサンプル数を数え，数えた数を視覚化して表示し，時系列を示す．原理的には，対象とするデータに活動記録の欠損がなければ，正確な時系列データと同じ値を得ることができる．ただし，利用者や参加者の活動頻度が低い場合には，時間周期を長くしないと欠損になる．具体的なツールは R 言語の ggplot2 ライブラリで提供されている「geom_freqpoly」[17] を用いて求めた．

4.2. 探索対象の選択

時系列データの観測周期は，分析精度に影響を与える．株価の変動や電力消費の分析は，対象とする変動が短時間で生じるため短い時間間隔で行う必要がある．本研究は，参加者や利用者の特性変化を明らかにすることが目的であり，月間程度で十分と推測されたが，月間の活動数が平均 10 件としても，7 年間で 800 件を超えるデータが必要となる．

実際のデータを調べると，3 章の結果では，数年間で 100 件以下のデータが多数を占めており，探索の対象条件を設定して選ぶ必要がある．探索的分析の目的は，対象のトレンドや周期変動を検出するのに必要な周期を明らかにすることである．そのため，大量の時間観測点を含むデータを選定し，トレンドや周期変動の有無と変動の速さ (時間微分) を観測した．具体的には，1,000 件以上のデータを対象としたサンプル B の Star と Issue の観測値から Repository をサンプリングして探索を行った．

4.3. 探索の例とまとめ

まず，利用者行動に対して時系列特性の観測を実施した．作品の利用者行動は，一般には消費者行動に近いと考えられる．図 5 に観測された代表的な例を示す．

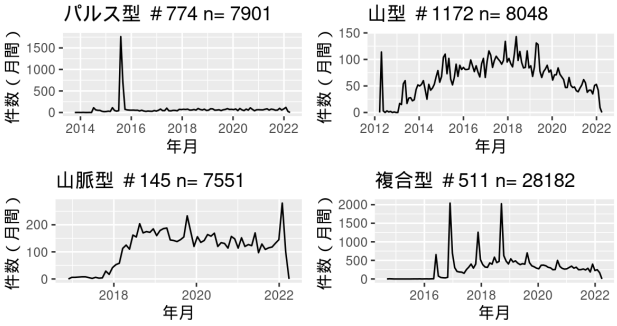

図 5: サンプル B の Star の時系列型例

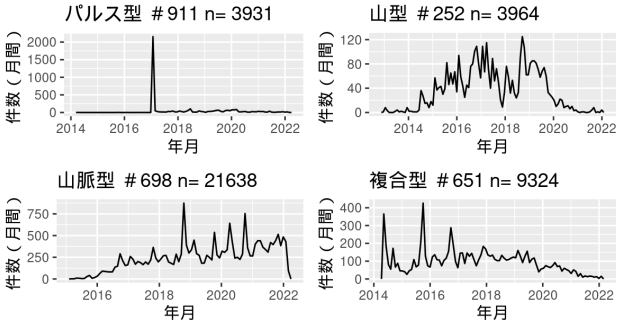

図 6: サンプル B の Issue の時系列型例

- **サンプル #774**：パルス型は，非常に短い期間 (2 〜 3 ヶ月) において，数年間で得られる評価のほとんどを獲得するパターンである．利用者の Star 評価がパルス以降にほとんど生じないが，利用していないことを表すものではない．新規に評価する利用者がいない状態を示している．本研究では，層別をテーマとしているため，その原因推測には立ち入らない．

- **サンプル #1172**：山型は，徐々に増加し，徐々に減少するもので，たとえば 10 年に渡るライフサイクル的な状態を示す．ここで選んだ事例は，最初にパルス的な傾向が見られる．また，途中のギザギザに周期性があるのか否かについては，目視判断なので不明である．

- **サンプル #145**：山脈型は，山型ほど明快な凹凸ではなく，なだらかな減少や増加を示す．ただし，時間軸の幅によっては，山型と変わらない．

- **サンプル #511**：複合型は，パルス型と山脈型や山型との組合せである．

作品の利用は，モノの購入や消費とは異なるので消費財の消費者行動とはかなり異なっている．「どのような分野と共通なのか」や「反応速度」など，詳細については次の課題である．ここで明らかになったことは，異なる型があり，OSS の共通特性ではないこと，パルス型のように非常に短期間に集中するものがあることである．

図 6 は，参加者側の時系列特性を示している．参加者側でも，利用者側と同様な型が観測されている．

- **サンプル #911**：パルス型は，利用者側では多く観測されたが参加者側では少数であった．短期間に月間で 1,000 件を超える案件が処理され，その後は月間数件で継続されるケースが観測された．サンプルデータ外になるが，東京都の Covid-19 サイト [18] も同様の振舞いであった．

- **サンプル #252**：この事例の山型は 6 年間ほどのライフサイクルで活動が活性化され，その後も安定した活動が続いている．山が 2 つのケースも観測された．

- **サンプル #698**：この事例の山脈型は 7 年間に渡り上昇が継続する山脈型の活動であり，上昇方向に振れることはあるが，大きく下がることは少ない．

- **サンプル #651**： この事例の複合型は初期に小さなパルスがいくつかあり，継続し，少し衰退している．

参加者側の型も多様であり，OSS の共通型ではない．パルス型を除けば，長期間に渡って活動が継続している．参加者のデータは活動記録であるため，新規に参加する活動と継続する活動に分離した分析も可能である．

層別で観測された型は本節で示した 4 種類に分類されるが，型分類の閾値などの具体的な値の分析・検証は，次の課題である．

5. 利用者行動との関係

本章では，参加者の行動，すなわち OSS 開発に参加し労力を提供する者の行動と，そのユーザに当たる利用者の行動との関係について調べる．OSS においても「作品の評価」と「作品開発」が連動するケースを層別できれば，そのメカニズムの研究に貢献できると考え，探索的な分析を行った．

OSS は巨大な場であり,「作品の評価」と「作品開発」が連動する / しないのどちらも存在する．本研究は，測定し層別することにより，インターネット空間を介した「働き方」に関する知識を得ることである．本章では，4 章で示した活動の型に着目して，利用者行動と参加者行動の関係が観測されるのかについて概観する．

分析は，次のような組合せでサンプルを選定して実施した．3 章の表 1 で示した利用者と参加者のクロス表から観測対象を選び，両者の関係を時系列で可視化して観察した．利用者側の層別は，活動が少ない順に {A,B,C,D}，参加者側を {a,b,c,d} と表している．これらの組合せを Aa，Db といった形で表現し，Repositry に対応するサンプル番号で識別した．

5.1. 量的相関と時系列相関

利用者の活動量，すなわち Star を付けることで代用される変量と，参加者の活動量に相関はあるのか．表 1 では，Aa から Dd への対角上で値が膨れていることから，明らかに相関が認められる．

たとえば，利用者が多いから参加者が増える，あるいは，その逆，といった因果関係については，今回の分析で用いたような量的な組合せデータだけでは明らかにすることはできない．

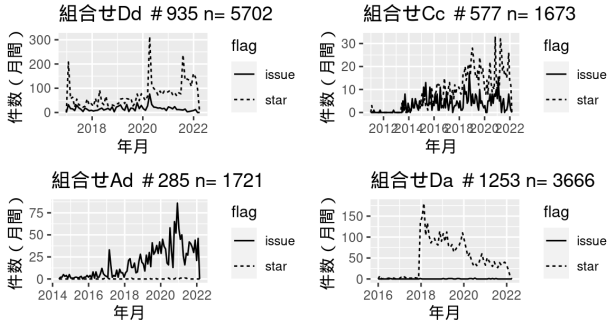

図 7: 時系列における相関の観察例

時系列的相関とは，利用者の増加に続いて参加者が増える，あるいはその逆などの時間遅れで反応があることを指す．反応がある可能性のあるサンプルの組合せは Dd や Cc など，参加者も利用者も多い場合が考えられる．一方，Ad や Da の組合せは，そもそも一方の活動が低調なのに，他方が活発であるデータの集団であり，因果関係は認められなかった．この 4 つの組合せで観測した時系列の例を図 7 に示し，その説明を次に述べる．

- **サンプル ＃935, ＃577**：組合せ Dd と Cc における例で，参加者活動と利用者活動は，時系列的にも連動し相関が認められる．
- **サンプル ＃285**：組合せ Ad は利用者の評価がほぼないが，参加者側は非常に活発な活動を示している．この事例では参加者の活動は上昇型のトレンドが 8 年に渡って継続している．
- **サンプル ＃1253**：組合せ Da は参加者活動が低いが利用者の評価が高い事例を示している．

利用者と参加者の行動に強い相関のある例，まったくない例などがあり，大きな層別では表 1 により，ある程度の絞り込みが可能である．相互の関係として，依存関係 (必要条件など) が強いものは見当たらなかった．

5.2. 利用者型の影響

4 章において，利用者と参加者の活動パターンを時系列の型で層別した．本節では，利用者側の型が参加者側の活動に影響を及ぼしているのかについて概観する．

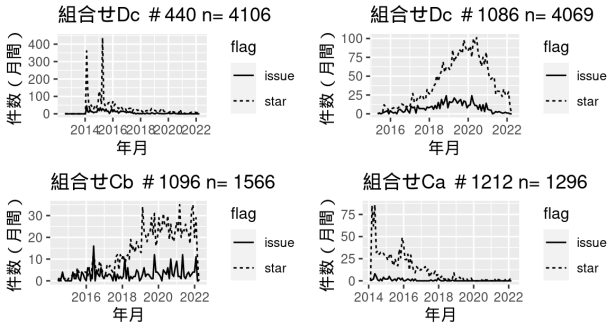

図 8: 利用者特性の型の影響例

図 8 に示した例を用いて，利用者行動の特徴的な型がある場合，参加者側に影響が生じているのか，あるい

は，参加者側の何かが利用者側に影響を及ぼしているのかについて概観する．

- **サンプル #440**：利用者側がパルス状に急増する例で，この事例では 2 回ほど生起している．一方で，参加者側に目立った変化は見られない．
- **サンプル #1086**：利用者側は山型であり，参加者側にも同様な山型が見られる．大きさの大小については多様である．
- **サンプル #1096**：利用者側が増加する山脈型を示している．参加者側も山脈型であるが利用者の増加とは強く連動していないように見える．
- **サンプル #1212**：利用者側が減少傾向を示し，参加者側も同様の変化を示している．

観測した範囲内で，利用者側の変化，特に急激な変化が参加者側に作用していると思えるような連動は見当たらなかった．このことから，商品が売れ始めると量産し，利益を追求するようなビジネスモデルとは異なることが分かる．

5.3. 参加者型の影響

ここでは，参加者側の行動が利用者にどのような影響を及ぼしているのかを概観するため，特徴的な参加者の型に対する利用者側の動きを調べた．参加者側の多様な型が多く見られたのは Cd の組合せであったため，次の 4 例もその中から選んでいる．

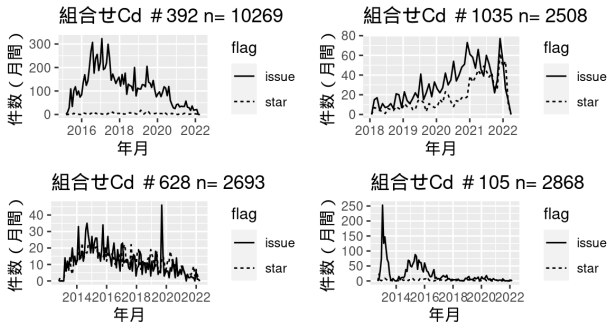

図 9: 参加者特性の型の影響例

図 9 は，図 8 とは逆に，参加者側に顕著な型が見られる場合について例示した．

- **サンプル #392, #105**：連動が見られない，あるいは大きな参加者の活動に比べ，利用者側の相対的な大きさが小さいケースとして層別される．
- **サンプル #1035, #628**：連動して変化しており，何らかの結びつきの研究サンプルとして層別できる．

参加者側の型と利用者側の特性についても，連動がないもの / あるものが存在し，定性的には層別として型が使えることが推測できる．

5.4. 時系列分析の要件のまとめ

OSS 作品は，その種類や数の巨大さだけでなく，動的な変化の多様性についても，商用作品とは全く異なるものであった．4.2 節で述べた通り，探索的分析の目的は，対象のトレンドや周期変動を検出するのに必要な周期を明らかにすることである．データが少ない Repository が数多く存在するため，データが少ない場合には四半期程度の間隔での対応が必要となる．一方，パルス型という短期変動が認められた．パルス型を観察するためには，短い間隔での対応が求められる．

本研究で示した時系列変化の観察により，時系列分析の要件として明らかになったことは，次の 2 点である．

- **短期変動**：パルス状の変化が，特に利用者側で観測された．分析の詳細化には時間幅を 1 ヶ月以下，できれば週単位で観測する必要がある．
- **短期変動以外の変動**：緩やかな変動であり四半期の間隔で対応できると思われる．

6. おわりに

本研究では，OSS の作品に関わる参加者と利用者の行動分析を試みた．膨大な数の OSS 作品から，その参加者や利用者の「行動」に関する情報を得るために，対象作品の層別と選択を行う課題があった．ここでは，前処理で選んだデータ群に対し，参加者と利用者の活動量の順位をカテゴリ化し，クロス表を作成して対処した．

中心となる課題は，行動を特徴づける説明変数として何を用い，どのように分析するかである．ここでは，時間経過による変化に着目し，分析の時間周期を時系列データに倣って，共通化する方法を用いた．この方法に

より，異なる種類の作品であっても，参加者や利用者の行動を比較したり，層別することが可能になった．

今回の分析では，定量的な行動のモデル化には至っていないが，視覚化による定性的な分析に耐える方法を開発することができた．この方法により，パルス型の急峻な行動の型が存在すること，利用者行動と参加者行動の関係は多様であり，さまざまな関係が存在することが明らかになった．

次の課題として，参加者の役割に着目し，その時系列変化の分析がある．これは，エンジニアの成長過程を行動記録から分析する課題である．我々の研究目的であるエンジニアの育成や組織化における「意欲」を OSS 活動の「行動」から学び活用できるよう，次の課題に取り組んでいく所存である．

脚注

1) インターネットのアプリケーションを通じて共通の関心分野，価値観や目的を持った利用者が集まって持続的に相互作用する場であり，提供されるネットワークサービスの総称 [1].
2) 一元的にデータの構成管理を行う単位であり，ファイルやディレクトリを格納すると共に変更履歴が記録される．
3) 周期的な時間間隔で観測したデータ.
4) 時系列の定常性が認められる成分.
5) 季節変動や景気変動などに相当.
6) ノイズはホワイトノイズを意味し，平均 0 の変化成分.
7) GitHub では，Repository に Star を付けるということは，リポジトリメンテナに対してその作業についての感謝を示すことでもあり，GitHub のリポジトリランキングの多くは，リポジトリに付けられた Star の数を考慮している [13].
8) Issue とは作業に関するアイデア，フィードバック，タスク，バグを追跡するために用いられる機能である [14]

参考文献

[1] インターネットコミュニティ，https://ja.wikipedia.org/wiki/インターネットコミュニティ (accessed:2022/03/18)

[2] GitHub, https://github.com/ (accessed:2022/03/18)

[3] Wickham, Hadley and Grolemund, Garrett, “R for data science: import, tidy, transform, visualize, and model data,” O’Reilly Media, Inc. 2016

[4] 増田礼子，松尾谷徹,「オープンソース・ソフトウェア開発コミュニティにおけるライフサイクルの視覚化」，ソフトウェア・シンポジウム 2020 論文集，pp.44-54，ソフトウェア技術者協会，2020

[5] Zipf, George Kingsley, “The Psycho-Biology of Language,” Boston-Cambridge Mass. Houghton Mifflin, 1935

[6] Zipf, George Kingsley, “Human Behavior & The Principle of Least Effort, An Introduction to Human Ecology,” Addison-Wesley Press Inc., 1949

[7] Gini, Corrado, “Measurement of inequality of incomes,” The economic journal, Vol.31, No.121, pp.124–126, JSTOR, 1921

[8] Gastwirth, Joseph L, “The estimation of the Lorenz curve and Gini index,” The review of economics and statistics, pp.306–316, JSTOR, 1972

[9] 中村和之，「経済指標の見方・使い方：所得格差を測る指標－ジニ係数とローレンツ曲線－」，http://www.pref.toyama.jp/sections/1015/ecm/back/2005apr/shihyo/ (accessed:2022/03/18)

[10] Rob J Hyndman and George Athanasopoulos, “Forecasting: Principles and Practice (3rd ed),” https://otexts.com/fpp3/, Monash University, Australia, 2022 (accessed:2022/03/18)

[11] 北川源四郎,「4-4 時系列データ解析」，http://www.mi.u-tokyo.ac.jp/consortium2/pdf/4-4_literacy_level_note.pdf，東京大学 数理・情報教育研究センター，2020 (accessed:2022/03/18)

[12] GitHub Docs，https://docs.github.com/ja，(accessed:2022/03/18)

[13] GitHub Docs：Star について，https://docs.github.com/ja/get-started/exploring-projects-on-github/saving-repositories-with-stars#about-stars (accessed:2022/03/18)

[14] GitHub Docs,：Issue について，
https://docs.github.com/ja/issues/
tracking-your-work-with-issues/about-issues
(accessed:2022/03/18)

[15] GitHub API v3, https://developer.github.com/
v3/ (accessed:2022/03/18)

[16] 金明哲, 「[連載] フリーソフトによるデータ解析・マイニング第 60 回：統計的テキスト解析 (5)〜統計法則と指標〜」, https://www1.doshisha.ac.jp/
~mjin/R/60/60.html
(accessed:2022/03/18)

[17] geom_freqpoly,
Histograms and frequency polygons,
https://ggplot2.tidyverse.org/reference/
geom_histogram.html (accessed:2022/03/18)

[18] 東京都 新型コロナウイルス感染症対策サイト,
https://github.com/tokyo-metropolitan-gov/
covid19 (accessed:2022/03/18)