ISBSGデータを用いた見積もり研究に対するIPA/SECデータを用いた外的妥当性の評価


山田 悠斗※1


江川 翔太※1,※2


松本 真佑※1


角田 雅照※3


楠本 真二※1

ソフトウェア開発プロジェクトの開発工数・期間の見積もり技術に関する研究が盛んに行われている．その際，評価のためのデータとして豪州の団体が世界各国より収集したプロジェクトデータであるISBSGデータが用いられることが多い．一方，一般に既存研究で得られた知見に対する追試（replication study）が重要であるとされている．追試を行うことによって特定の条件や環境において得られた知見に対する再現性や，異なる実験条件から得られる知見の差異を調査することができる．また，追試を行い様々な条件から得られた知見を統合することによって，新規開発プロジェクトにおいてそれと類似した条件での知見を再利用することが可能となる．本稿では，ISBSGデータが用いられている見積もりの既存研究を対象にして別データを用いた追試を実施する．既存研究とは異なるデータセットを用いて実験を行うことで，得られる知見に差異が生じるかを調査する．過去5年間でISBSGデータが利用されていた論文22本の中から4本の論文を選択し，IPAが日本の各種ベンダ企業より収集したプロジェクトデータであるIPA/SECデータを用いて追試を行った．その結果，2本の論文では既存研究と類似した知見が，残りの2本の論文では既存研究と異なる知見が得られた．

Evaluation of External Validity Using IPA/SEC Dataset for Effort Estimation Studies Using ISBSG Dataset

Yuto Yamada※1, Shota Egawa※1,※2, Shinsuke Matsumoto※1, Masateru Tsunoda※3, Shinji Kusumoto※1

Many research studies of effort and duration estimation of the software development project have been conducted. In the studies, ISBSG dataset which includes actual software development data collected by the Australian organization from all over the world has been frequently used to evaluate the results. In the field of empirical software engineering, replication studies for the knowledge and experience finding from past studies are very important. By conducting the replications, we can evaluate the repeatability of the results of past studies got from the specific context and the difference between the results and the replication ones. This paper conducted replication studies for past ones of software effort estimation where ISBSG dataset were used in the evaluation. We selected four existing studies from twenty-two papers published within last five years. In the replications, we used IPA/SEC dataset that Information-technology Promotion Agency (IPA), Japan, have collected from Japanese software development companies. As the results, one replication produced the similar results to previous two and other two did different ones. Then, we discussed the reasons of the different results.

※1 大阪大学 大学院情報科学研究科
※2 ソニーグローバルマニュファクチャリング&オペレーションズ株式会社 設計部門 ソフトウェア設計部 テスト技術2課
※3 近畿大学 理工学部情報学科

1　はじめに

ソフトウェア工学はソフトウェアの開発，運用，保守に関して体系的，定量的にその応用を考察する分野であり，この分野で扱われている技術の中にソフトウェア開発の見積もり［Radatz1990］がある．ソフトウェア開発においては初期の段階から全体のコストや工数・開発期間などを正確に見積もることが重要であるとされており［Boehm2000］，見積もりの誤りがプロジェクト失敗へとつながる場合がある．この問題を解決するためにソフトウェア開発の見積もりに関する研究が盛んに行われている．見積もりに関する研究の評価には豪州の団体が世界各国より収集したプロジェクトデータであるISBSGデータ［ISBSG］が用いられることが多い．過去5年間における主要国際会議の論文誌を調査した結果では，見積もりに関する研究が行われている83本の論文のうち22本の論文でISBSGデータが利用されていた．

また，実証的ソフトウェア工学［Kitchenham2008］の分野では，既存研究で得られた知見に対する追試（replication study）が重要であるとされている［Silva2014, Shull2008］．研究に対する追試とは，ある研究に関して，実験の条件や環境を部分的に変更して実験を再現することである．追試を行うことによって特定の条件において得られた知見が別の条件においても再現できるかどうか，異なる条件では別の知見が得られるかどうかなどを調査することができる．これらを調査することで，研究成果に対する妥当性の評価を行うことができる．また，様々な条件から得られた知見を統合することで，新規開発プロジェクトにおいてそれと類似した条件での知見を再利用することが可能となる．

そこで，本研究ではISBSGデータが用いられており，開発工数・期間の見積もりに関する既存研究を対象にした追試を実施する．追試を行うにあたって明らかにするResearch Questionを3つ設定した．（RQ1）は，過去5年間における見積もり研究のうちISBSGデータが用いられた研究はどの程度存在するか，である．（RQ2）は，追試が可能で，現場でもすぐに適用できる研究はどの程度存在するか，である．（RQ3）は，ISBSGデータを用いた場合とIPA/SECデータを用いた場合から得られる知見に差は生じるか，である．異なる知見が得られた場合は結果が異なった原因の考察や追加の調査につなげることができ，同様の知見が得られた場合は既存研究の外的妥当性がより高められたと判断することができる．

以降，2.では研究の背景となる関連研究について述べる．3.では追試を行うための準備事項について述べる．4.では選択した論文に対して追試を行った結果と考察について述べる．5.では追試全体に対する考察について述べる．6.では主な結果と今後の課題をまとめる．

2　準備

本節では研究の背景となる諸用語や関連研究について簡単に述べる．

2.1　線形重回帰分析

ソフトウェア開発規模の見積もりには，多変量回帰分析の一手法である線形回帰分析が多く用いられている［Boehm1981］．線形回帰分析では予測対象となる目的変数と予測に必要とする説明変数との関係を一次式で表したモデルが作成される．一般的なモデルは式（1）の形で表される．目的変数である$\hat{Y}$には予測の対象となる工数などが，説明変数である$\hat{X}$にはソフトウェアの規模や要因といった予測対象を導くために必要となる要素が当てはめられる．a_iとbは予測対象の実測値と予測値の残差が最小となるように決められる．

$$\hat{Y} = a_1X_1 + a_2X_2 + \cdots + a_nX_n + b \quad (1)$$

精度の推定には以下の式（2）～（6）によって求められる5つの評価指標［Foss2003］であるAR（Absolute Residuals），MRE（Magnitude of Relative Error），MER（Magnitude of Error Relative to the estimate），BRE（Balanced Relative Error），Pred（25）が多く用いられる．MREは実測値から見た予測値の相対誤差を，MERは予測値から見た実測値の相対誤差を表す．BREは過大見積もりや過小見積もりに対しバランスの良い評価を行うことができる．AR，MRE，MER，BREは値が小さいほど，Pred（25）は値が大きいほど見積もり精度が良いと評価される．また，各指標の平均値は指標名の先頭にM，中央値にはMdを付与することとする．例えば，MREの平均値はMMRE，中央値をMdMREと示す．

$$AR = |\text{実測値} - \text{予測値}| \quad (2)$$

$$MRE = \frac{AR}{\text{実測値}} \quad (3)$$

$$MER = \frac{AR}{\text{予測値}} \quad (4)$$

$$BRE = \begin{cases} MRE & (\text{予測値} - \text{実測値} \geq 0) \\ MER & (\text{予測値} - \text{実測値} < 0) \end{cases} \quad (5)$$

$$Pred(25) = \frac{\text{評価指標の値が 0.25 以下であるデータ数}}{\text{全データ数}} \quad (6)$$

回帰モデルの予測精度を表す指標として，ほかに決定係数がある．これは重相関係数の2乗に等しく，説明変数が目的変数をどの程度説明できるかを表す．この値が大きいほど説明変数と目的変数の相関が強く，得られたモデルの予測能力が高いことを意味する．

2.2　ファンクションポイント法

ソフトウェアの規模を見積もる手法の一つにファンクションポイント法［IFPUG］がある．この手法では，まずソフトウェアの持つ機能から5種類の基本機能要素を抽出し，それぞれの処理内容の複雑度からファンクションポイント（以降，FP）と呼ばれる点数を付ける．このFPから工数などの推定が行われる．5種類の基本機能要素とは以下に示す要素のことを言う［Kashimoto2000］．

内部論理ファイル (ILF)：計測対象のアプリケーション内で更新される，論理的な関連を持ったデータの集合

外部インターフェースファイル (EIF)：計測対象のアプリケーションによって参照されるデータの集合（更新されない）

外部入力 (EI)：計測境界外からのデータ入力によってILFの更新を行う処理

外部出力 (EO)：計測境界外へのデータ出力を含む処理のうち，出力に派生データを含むもの

外部参照 (EQ)：計測境界外へのデータ出力を含む処理のうち，出力に派生データを含まないものであり，処理がILFを更新しないもの

アプリケーション全体でのFPの合計をアプリケーションFPと言い，アプリケーションFPにシステムの特性を考慮に入れて調整を加えた値を調整済みアプリケーションFPと言う．FP値は見積もり研究における規模の尺度としてよく利用されている．

2.3 ISBSGデータ

ISBSGデータは，豪州の団体であるISBSG（The International Software Benchmarking Standards Group）[ISBSG] が世界24カ国に存在する組織や企業から実開発のデータを収集・整理したデータセットである．データの項目には開発工数や開発言語，開発プラットフォームなどが含まれる．また，FPに対する品質ランクも収録されており，信頼性の高いデータを選別することができる．リリースごとにデータ数は異なるが，最新のデータセットには5,000以上のプロジェクトデータが118項目に分けて蓄積されている．

2.4 IPA/SECデータ

IPA/SECデータは，独立行政法人情報処理推進機構 [IPA] が日本に存在する組織や企業から実開発のデータを収集し，整理したデータセットである．データの項目には開発工数や開発言語など，ISBSGデータと共通した項目が多く含まれる．また，データの品質に関するランクも収録されており，信頼性の高いデータを選別することができる．2014-2015版では，3,541プロジェクトのデータが194項目に分けて蓄積されている．

2.5 実証的ソフトウェア工学

実証的ソフトウェア工学 [Kitchenham2008] とは，ソフトウェア開発現場での作業や実績に対する計測，定量化とその評価，そしてフィードバックによる改善という実証的手法を行う研究分野である．ソフトウェア開発の課題である生産性の向上や品質の確保に対する有用なアプローチとして注目されている．

(1) 妥当性についての議論

実証的ソフトウェア工学では，実在するソフトウェア開発データを用いたケーススタディを通じて提案手法の評価が行われることが多いが，このとき妥当性に関する議論が行われなければならない．妥当性には以下の分類 [Robert2013] が存在する．

内的妥当性
研究成果が研究の際に操作した要因から影響を受けている程度を指す

外的妥当性
ある研究から得られた成果を，違った母集団，環境，条件へ一般化し得る程度を指す

構成概念妥当性
結果を得るために行った操作が適切である程度を指す

信頼性
他者が同様の手順で研究を行った場合，研究結果が再現可能となる程度を指す

実証的ソフトウェア工学における妥当性に関する研究として，文献 [Egawa2016] の研究がある．この研究では，見積もり研究において研究成果の外的妥当性がどの程度意識されているかを調査するため，過去の研究論文を対象とした網羅的なレビューが行われている．調査の結果，対象となる89本の論文のうち，研究成果の外的妥当性についての議論を行っていない研究論文が26本存在しており，結論部分においてのみ言及している論文が31本存在することが示されている．このことから，見積もりの研究に携わる研究者は，研究成果の外的妥当性に関してより注意を払うべきであるということが主張されている．

(2) 追試 (replication study)

実証的ソフトウェア工学の分野では，既存研究で得られた知見に対する追試（replication study）が重要であるとされている [Silva2014, Shull2008]．研究に対する追試とは，ある研究に関して，実験の条件や環境を部分的に変更して実験を再現することである．追試を行うことによって，特定の条件において得られた知見が別の条件においても再現できるか，異なる条件や環境では別の知見が得られるかなどを調査することができる．追試によって異なる知見が得られた場合は結果が異なった原因の考察や追加調査につなげることができる．同様の知見が得られた場合は既存研究の研究成果の外的妥当性がより高められたと判断することができる．また，追試を行い，様々な条件から得られた知見を統合することで，新規開発プロジェクトにおいてそれと類似した条件での知見を再利用することが可能となる．

3 追試の準備

本節では追試を行うために必要な準備事項について述べる．

3.1 Research Questionの設定

追試を行うにあたって明らかにしたい問をResearch Question (RQ) として設定する．本研究では，以下の3つを設定した．

- **RQ1**：過去5年間における見積もり研究のうちISBSGデータが用いられた研究はどの程度存在するか
- **RQ2**：追試が可能で，現場でもすぐに適用できる研究はどの程度存在するか

●**RQ3**：ISBSGデータを用いた場合とIPA/SECデータを用いた場合から得られる知見に差は生じるか

RQ1は，ISBSGデータには複数企業のデータが含まれているため，ほかのデータセットに比べて外的妥当性が高いと思われるが，それがどの程度使用されているのかを明らかにするために設定した．RQ2に関しては，工数見積もりは現場での利用可能性が高くないといけないが，実際には適用が難しい研究成果も多いと思われる．適用可能な研究がどの程度の割合で存在するかを明らかにするために設定した．RQ3に関しては，IPA/SECデータは日本国内の複数の企業から収集したデータセットであるため，このデータセットで成立する知見は日本国内の企業に適用してもある程度成立すると考えられる．ISBSGデータで得られた知見が，IPA/SECデータ，すなわち日本の企業でもどの程度同様に得られるのかを調査するために設定した．

3.2 論文の選択

見積もりに関する研究の評価にはISBSGデータが用いられることが多い．過去5年間における以下の主要国際会議の論文誌を調査した結果では，見積もりに関する研究が行われている83本の論文のうち22本の論文でISBSGデータが使用されていた．これらの論文を追試の対象とする．

- ACM Transactions On Software Engineering and Methodology
- ASIA-PACIFIC Software Engineering Conference
- Empirical Software Engineering and Measurement
- nternational Conference on Software Engineering
- International Conference on Software Maintenance
- IEEE Transactions on Software Engineering
- Information and Software Technology
- Journal of Systems and Software

22本の論文の中から，追試が可能であり，かつ追試を行うのにふさわしい論文を選択する必要がある．選択のために提示した条件と，各条件を満たしている研究を順に抽出した際の論文数の推移を表1に示す．表の通り，最終的に4本の論文［Tsunoda2012-1，Lavazza2013，López-Martín2015，Tsunoda2012-2］が追試の対象として選択され，それぞれに関して追試を行った．4本の論文をそれぞれ，生産性に基づくモデル作成に関する論文［Tsunoda2012-1］，FPの簡易推定に関する論文［Lavazza2013］，ニューラルネットワーク（以降，NN）を利用した見積もりに関する論文［López-Martín2015］，カテゴリ変数を用いたモデルに関する論文［Tsunoda2012-2］と名称付けて説明する．

ただし，本論文では紙面の都合上，追試の詳細な結果に関しては，生産性に基づくモデル作成に関する論文，FPの簡易推定に関する論文，NNを利用した見積もりに関する論文についてのみ報告する．

表1　追試の対象となる論文の条件と抽出された論文数

論文の条件	論文数
評価の段階でプロジェクトデータが用いられている	20
入手可能なデータセットが用いられている	19
使用されているデータセットが5種類未満である	16
再現実験のために必要な情報が正確に記載されている	11
ISBSGデータの入手可能な情報が用いられている	9
現場での適用容易性が高い	4

3.3 生産性に基づくモデル作成に関する論文の概要

回帰分析に基づく工数見積もりモデルに関する研究が行われており，対象プロジェクトの生産性も考慮に加えた工数見積もりモデルの作成方法を提案している［Tsunoda2012-1］．なお，ここでの生産性はFPを工数で割った値として定義される．

通常，工数見積もりモデルを作成する際にはデータセットに蓄積されたデータすべてを対象として回帰分析を行い，1つのモデルを作成する．提案手法ではデータセットを生産性の高さによって複数に分類し，それぞれからモデルを作成する．そしてテストデータが持つ生産性の値の高さに基づき，複数のモデルのうち1つを選択して工数の見積もりを行う．現場で提案手法を実施する際，現行プロジェクトの生産性の推測はプロジェクトマネージャが行う．プロジェクトマネージャが生産性の推測を誤る確率を「推測誤り率」とする．

実験では，以下に示す3種類のモデルを作成し，見積もりの精度を比較する．Twoレベルモデルを用いて対象プロジェクトの見積もりを行う場面を図1に示す．また，実験を行う際の条件を表2に示す．

Noレベルモデル

生産性を考慮せずに回帰分析を行う従来の見積もり手法を用いるモデル

Twoレベルモデル

生産性の値の高さによってデータセットをHigh（高），Low（低）の2段階に分類して提案手法を用いるモデル

Threeレベルモデル

生産性の値の高さによってデータセットHigh（高），Medium（中），Low（低）の3段階に分類して提案手法を用いるモデル

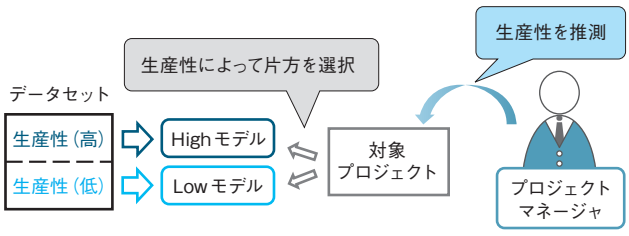

図1　Twoレベルモデルを用いた見積もり手法

表2　生産性に基づくモデル作成に関する論文における実験の条件

入力	FP，開発言語，開発種別，プラットフォーム
出力	工数
使用するモデル	Noレベルモデル，Twoレベルモデル，Threeレベルモデル
評価指標	MRE，MER，BRE

リリース9のISBSGデータ内の593データを対象として実験が行われた．その結果，推測誤り率が38%以下の場合，つまり現場のプロジェクトマネージャが生産性の推測を誤る可能性が低いと判断できる場合には，従来の手法よりも提案手法のほうが工数見積もりにおける見積もり精度が向上するという知見が得られている．

3.4 FPの簡易推定に関する論文の概要

FPの計測方法に関する研究が行われており，従来よりも簡易化されたFP推定モデルを提案している［Lavazza2013］．FPを計算する際に抽出される5種類の基本機能要素のうち，FPと最も相関の強い要素の規模のみを説明変数とした簡易FP推定モデルを作成する．相関関係の調査にはケンドールの順位相関係数とスピアマンの順位相関係数を用いる［Croux2010］．実験を行う際の条件を表3に示す．

表3　FPの簡易推定に関する論文における実験の条件

入 力	FP, EI, EO, EQ, ILF, EIF
出 力	FPと各基本機能要素の相関値
評価指標	ケンドールの順位相関係数，スピアマンの順位相関係数

リリース11のISBSGデータが持つ600以上のデータを対象として実験が行われた結果，EIにおけるケンドールの順位相関係数が0.658，スピアマンの順位相関係数が0.839となり，FPと最も相関が強くなった．

3.5 NNを利用した見積もりに関する論文の概要

NN［Richard1987］を利用したソフトウェア開発期間の見積もりモデルの精度を調査している［López-Martín2015］．NNとは，人間の脳が問題を解く際の振る舞いを計算機上のシミュレーションによって表現したネットワークモデルである．今回は以下の2種類のNNであるMLP（Multilayer Perceptron）とRBFNN（Radial Basis Function Neural Network）を使用する［Park2003］．これらのモデルの精度を，重回帰モデルを使用した場合と比較して調査する．実験を行う際の条件を表4に示す．なお，表4におけるPred（25）は，ARに対する値である．

MLP

内部のニューロンが入力層，中間層，出力層に分かれており，ループせず単一方向にのみ信号が伝播するネットワーク

RBFNN

MLPの中間層で放射基底関数を用いて出力を計算するネットワーク

表4　NNを利用した見積もりに関する論文における実験の条件

入 力	FP, 社内要員数
出 力	開発期間
使用するモデル	重回帰モデル, MLP, RBFNN
評価指標	AR, Pred（25）

リリース11のISBSGデータ内の49データを対象として実験が行われた結果，2種類のNNモデルはいずれも重回帰モデルと比べて，MAR,MdARに関して精度が6%以上高くなった．このことから，NNを利用することによって開発期間の見積もり精度をより高めることができると言える．

3.6 カテゴリ変数を用いたモデルに関する論文の概要

工数見積もりにおけるカテゴリ変数の扱い方に関する研究が行われている［Tsunoda2012-2］．カテゴリ変数とは性別，職業など一般に数や量では測れない変数を指す．回帰モデルの説明変数としてカテゴリ変数を使用する際には，対応方法の異なる様々なモデルが用いられる．以下の4種類のモデルを対象とし，これらの工数見積もりモデルを，カテゴリ変数を使用しないモデルと比較することで評価する．

ダミー変数化を用いたモデル

カテゴリ変数から0.1の2値を取る複数のダミー変数を生成し，それらを説明変数として回帰モデルを作成する．

層別を用いたモデル

変数の値の組み合わせにより，データセット内のデータをサブセットに分割する．そしてそれぞれのサブセットから回帰モデルを作成する．

交互作用モデル

ある説明変数の値によって，ほかの説明変数の効果が変化することを交互作用と言う．今回はダミー変数化を用いたモデルに，各ダミー変数とFPの積により作成した説明変数を加えたモデルを作成する．

階層線形モデル

グループごとにまとまりがあるデータを分析する際に用いられる．層別によって分割したサブセット間の関係性を考慮に加えて，サブセットごとに切片と傾きが変化するモデルを作成する．

リリース9のISBSGデータ内の558データを対象として実験を行った結果，カテゴリ変数を使用する4種類のモデルはいずれもカテゴリ変数を使用しない場合と比べて精度が約10%向上した．また，4種類のモデル間での精度の差は5%未満となり，見積もりにおいて同程度の精度が得られた．

3.7 使用する統計ツール

追試の中で回帰分析などの統計処理を行う際は，統計分析のフリーソフトであるR［Rproject］を使用する．

4 追試の結果

本節では，対象となる4本の論文に対してIPA/SECデータで追試を行った結果及び考察について説明する．

4.1 生産性に基づくモデル作成に関する論文

（1）データの選別基準

IPA/SECデータに蓄積されているデータのうち，表5に示す既存研究と同様の選別基準に従って抽出された189データを使用する．「本データの信頼性」とは，当該プロジェクトデータ

の合理性や整合性に関する信頼度を4段階（A～D）で評価した値であり，最も信頼度が高い場合はAと評価される．

表5　生産性に基づくモデル作成に関する論文でのIPA/SECデータの選別基準

項 目	選別基準
FP実績値の計測手法	IFPUG
本データの信頼性	AもしくはB
FP実績値（調整前）	欠損していない
主開発言語グループ	欠損していない
開発プロジェクトの種別	欠損していない
開発対象プラットフォームのグループ	欠損していない

(2) 結果と考察

各モデルの説明変数にはFPの対数，主開発言語グループ，開発プロジェクトの種別，開発対象プラットフォームのグループを用いる．追試における実験の結果を表6に示す．Twoレベルモデルと Threeレベルモデルの数値は，推測誤り率が0%の時点での数値である．更に，推測誤り率を増加させた際の見積もり精度の推移を図2，図3に示す．図2は横軸に推測誤り率が，縦軸にMBREの数値が示されている．図3は横軸に推測誤り率が，縦軸にMdBREの数値が示されている．

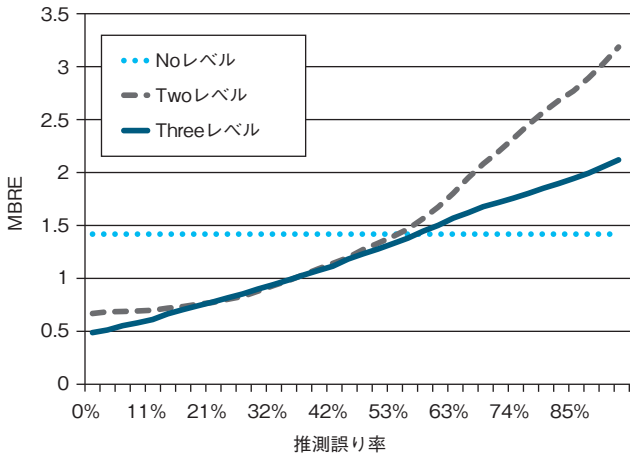

図2　追試における推測誤り率の増加によるMBREの推移

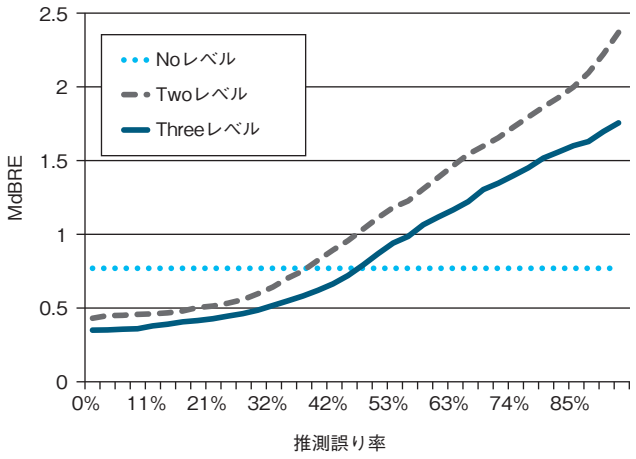

図3　追試における推測誤り率の増加によるMdBREの推移

表6　生産性に基づくモデル作成に関する論文の追試結果

レベル	MMRE	MdMRE	MMER	MdMER	MBRE	MdBRE
No	1.00	0.53	0.85	0.54	1.41	0.77
Two	0.50	0.33	0.46	0.35	0.65	0.43
Three	0.38	0.28	0.35	0.29	0.47	0.35

表6を見ると，推測誤り率が0%のときはいずれの評価指標でもNoレベルモデルよりTwo, Threeレベルモデルのほうが見積もりの精度が高い．図2，図3を見ると，推測誤り率が増加するほどTwo, Threeレベルモデルの精度は低下している．しかし，推測誤り率が37%以下の時点では，MBREとMdBREのどちらにおいてもTwo, Threeレベルモデルのほうがレベルモデルより精度が高い．

以上のことから，推測誤り率が低い状態，つまり現場のプロジェクトマネージャが生産性の推測を誤る可能性が低いと判断できる場合には，従来の手法よりも提案手法のほうが工数見積もりにおける見積もり精度が高くなると考えられる．これは既存研究から得られた知見と同様の知見である．

海外からデータを収集したISBSGデータと日本からデータを収集したIPA/SECデータで同様の知見が得られたことから，追試を行うことにより既存研究の研究成果の外的妥当性が高められたと言える．

4.2　FPの簡易推定に関する論文

(1) データの選別基準

IPA/SECデータに蓄積されているデータのうち，表7に示す既存研究と同様の選別基準に従って抽出された，122データを使用する．

表7　FPの簡易推定に関する論文でのIPA/SECデータの選別基準

項 目	選別基準
本データの信頼性	AもしくはB
EI	欠損していない
EO	欠損していない
ILF	欠損していない

(2) 結果と考察

追試における実験の結果を表8に示す．ケンドールとスピアマンの順位相関係数を用いた際の，FPと各基本機能要素の相関係数の値が示されている．

表8　FPの簡易推定に関する論文の追試結果

要 素	ケンドール	スピアマン
EI	0.663	0.836
EO	0.607	0.791
EQ	0.656	0.850
ILF	0.655	0.837
EIF	0.384	0.525

既存研究でISBSGデータを用いた場合，FPと最も相関の強い要素としてEIが得られている．しかし表8より，IPA/SECデータを用いた場合におけるFPと最も相関の強い要素は，ケンドールの順位相関係数の場合はEI，スピアマンの順位相関係数の場合

はEQとなった．このことから，FPと相関の強い基本機能要素は必ずしもEIではなく，データセットにより異なることが分かる．

よって，基本機能要素を用いて簡易FP推定モデルを作成する際は，使用するデータセットごとに適した要素を選択する必要があると考えられる．

4.3 NNを利用した見積もりに関する論文

(1) データの選別基準

IPA/SECデータに蓄積されているデータのうち，既存研究と同様の選別基準に従って抽出された，表9に示す36データを使用する．

表9 NNを利用した見積もりに関する論文でのIPA/SECデータの選別基準

項 目	選別基準
本データの信頼性	AもしくはB
実績月数（開発期間）	2カ月以上
社内ピーク要員数	2人以上
開発対象プラットフォームのグループ	Windows系
主開発言語	第3世代言語
開発プロジェクトの種別	新規開発

(2) 結果と考察

各モデルの説明変数またはNNの入力値には調整済みアプリケーションFPの対数と社内ピーク要員数の対数を用いる．追試における実験の結果を表10に示す．

なお，リリース11のISBSGデータを用いて実験の再現を行った結果，こちらの準備した実験環境でNNモデルを用いると，MdARとPred（25）に関しては重回帰モデルより精度が向上したが，MARに関しては精度がやや低下した．これは，NNモデルを用いた際に，ごく一部のプロジェクトデータに関して本来の値と大きく外れた異常な予測値が発生しやすくなったためである．元となる論文では，NNモデルの動作がブラックボックス化されていることや，最適なノード数などを見つけることが困難であるということが問題として述べられている．このことから，NNモデルの環境や得られる結果を完全に一致させることは困難であると考えられる．よって今回は，ごく一部の異常値を無視すれば元の実験と類似した傾向の結果が得られたことから，元の論文と近い実験環境を構築できたと判断し，この実験環境で追試を行う．異常値による影響を減らすために，MARよりも異常値に関して頑健であるとされているMdARとPred（25）の値からモデルの精度を比較する．

表10を見ると，Pred（25）とMdARのどちらにおいても重回帰モデルより精度が向上しているNNモデルはRBFNNモデルのみである．これは既存研究と異なる結果である．

表10 NNを利用した見積もりに関する論文の追試結果

モデル	ニューロン数	MAR	MdAR	Pred（25）
重回帰	−	0.33	0.27	0.44
MLP	5	0.44	0.30	0.50
RBFNN	14	0.37	0.20	0.53

次に，追試ではMLPモデルの精度が重回帰モデルより低下した原因を考える．既存研究ではNNモデルの入出力値の関係性を調査している．その際，入力値である調整済みアプリケーションFPの対数と社内ピーク要員数の対数を説明変数とし，出力値である開発期間の対数を目的変数とする重回帰モデルを，データセットに蓄積されたすべてのデータに対して回帰分析を行い作成している．そのモデルが以下の式（7）である．AFP（Adjusted Function Points）は調整済みアプリケーションFPを，MTS（Max Team Size）は社内ピーク要員数を，Durationは開発期間を指す．このモデルの決定係数は0.560となった．

$$\ln(\mathrm{Duration}) = 0.150 + 0.438 \times \ln(\mathrm{AFP}) - 0.187 \times \ln(\mathrm{MTS}) \quad (7)$$

追試においても同様のモデルを作成した．そのモデルが以下の式（8）であり，決定係数は0.398となった．

$$\ln(\mathrm{Duration}) = -0.669 + 0.435 \times \ln(\mathrm{AFP}) - 0.205 \times \ln(\mathrm{MTS}) \quad (8)$$

既存研究でISBSGデータから得られたモデルのほうが追試でIPA/SECデータから得られたモデルより決定係数が高いことから，入出力の相関関係はIPA/SECデータよりISBSGデータのほうが強いことが分かる．また，NNモデルは入出力値の相関関係が強いほど予測精度が高くなるとされている［Hitokoto2012］．更に，RBFNNモデルはMLPモデルより安定した学習が可能であるとされている［Murakami2008］．以上のことから，既存研究よりも入出力値の相関関係が弱くなった影響をMLPモデルのみが受け，MLPモデルの見積もり精度が低下したと考えられる．

4.4 カテゴリ変数を用いたモデルに関する論文

紙面の都合上，詳細な結果は省略する．生産性に基づくモデル作成に関する論文と同様に，既存研究から得られた知見と類似した知見が得られた．

5 考察

ある手法を用いた際のコンテキストが，どの程度自身の環境に対応しているかを特定することは容易ではない，という課題が存在する．実際に2本の論文では，既存研究とはコンテキストが異なるデータで構成されたIPA/SECデータを用いた場合は異なる結果が得られた．ただし，FPの簡易推定に関する論文では，ISBSGデータを業種で層別した場合は業種ごとに異なる結果が得られたことから［Yamada2016］，コンテキストを特定するには層別を行い，データのフィルタリングを行うことが有効であると考えられる．

また，上記の問題に対し，コンテキストに強く依存しない手法を用いることも一つの解決策であると考えられる．類似した知見が得られた2本の論文では比較的コンテキストの依存が強くないため，この手法に関してはコンテキストを厳密に合致させなくても見積もり改善の効果が期待できる．

6　おわりに

本稿では，ISBSGデータが評価に用いられている見積もりの既存研究に対する追試を実施した．設定したRQへの回答として，RQ1については，ISBSGデータが用いられた研究は22本存在するという結果が得られた．RQ2については，4本の論文［Tsunoda2012-1，Lavazza2013，López-Martín2015，Tsunoda2012-2］が選択された．RQ3については，2本の論文では既存研究と異なる知見が得られたことから，追加調査も交えてその原因に対する考察を行った．ほかの2本の論文では既存研究と類似した知見が得られたことから，既存研究の研究成果の外的妥当性が高められたという結果が得られた．以上のことから，IPA/SECデータで同様の知見が得られた研究については開発現場でも同様の成果が得られることが期待でき，かつ，適用も容易なので，工数見積もりの精度を高めるために適用すべきであると言える．

今後の課題としては，まず見積もり分野におけるほかの既存研究に対する追試が考えられる．実証的ソフトウェア工学では妥当性への脅威に関する議論が重要となるが，その必要性に対する認識はいまだに不十分である．よって，今回のような妥当性に関する議論を継続的に行うことで，その必要性に対する認識をより広めていくべきである．また，今回の追試から得られた知見に対する追加調査も今後の課題として考えられる．例えば，FPの簡易推定に関する論文について，各業種と導かれた要素との関連性に対する調査が挙げられる．更に，外的妥当性をより高めるために，今回用いたデータセットとは異なる種類のデータセットで追試を行うことも必要であると考えられる．加えて，実環境に対する実験結果の対応度合いをいかにして特定するかという課題に対する，より詳細な調査が考えられる．様々な実環境との適応度合いや期待できる効果などを調査するには，実企業の協力を得た更なる追試が必要となる．

謝辞

本研究を行うにあたり，データを提供していただくと共に多大なご助言をいただきました独立行政法人情報処理推進機構技術本部ソフトウェア高信頼化センターの関係各位に深く感謝申し上げます．

本研究は一部，日本学術振興会科学研究費補助金基盤研究(S)(課題番号：25220003)，及び基盤研究(C)(課題番号：16K00113)の支援を受けている．

【参考文献】

［Radatz1990］J. Radatz, A. Geraci, and F. Katki. IEEE standard glossary of software engineering terminology. IEEE Std, 610.12-1990.

［Boehm2000］B. Boehm, C. Abts, and S. Chulani. Software development cost estimation approaches - A survey. Annals of software engineering, 10, 1-4, 177–205, 2000.

［ISBSG］International Software Benchmarking Standards Group (ISBSG). http://www.isbsg.org.

［Kitchenham2008］B. Kitchenham, et al. Evaluating guidelines for reporting empirical software engineering studies. ESE, 13, 1, 97–121, 2008.

［Silva2014］F. QB Da Silva, et al. Replication of empirical studies in software engineering research: a systematic mapping study. ESE, 19, 3, 501–557, 2014.

［Shull2008］F. J Shull, J. C Carver, S. Vegas, and N. Juristo. The role of replications in empirical software engineering. ESE, 13, 2, 211–218, 2008.

［Tsunoda2012-1］M. Tsunoda, A. Monden, J. Keung, and K. Matsumoto. Incorporating expert judgment into regression models of software effort estimation. In APSEC, 1, 374–379. IEEE, 2012.

［Lavazza2013］L. Lavazza, S. Morasca, and G. Robiolo. Towards a simplified definition of Function Points. IST, 55, 10, 1796–1809, 2013.

［López-Martín2015］C. López Martín and A. Abran. Neural networks for predicting the duration of new software projects. JSS, 101, 127–135, 2015.

［Tsunoda2012-2］M. Tsunoda, S. Amasaki, and A. Monden. Handling categorical variables in effort estimation. In ESEM, 99-102. ACM, 2012

［Boehm1981］Barry W Boehm. Software engineering economics. 197. Prentice-hall Englewood Cliffs (NJ), 1981.

［Foss2003］T. Foss, E. Stensrud, B. Kitchenham, and I. Myrtveit. A simulation study of the model evaluation criterion MMRE. TSE, 29, 11, 985–995, 2003.

［IFPUG］IFPUG: Function Point Counting Practices Manual, Release 4.3. IFPUG, 2010.

［Kashimoto2000］柏本，楠本，井上，鈴木，湯浦，津田．イベントトレース図に基づく要求仕様書からのファンクションポイント計測手法．2000.

［IPA］IPA 独立行政法人情報処理推進機構．https://www.ipa.go.jp.

［Robert2013］Robert K Yin. Case study research: Design and methods. Sage publications, 2013.

［Egawa2016］江川翔太．見積もり研究における外的妥当性の調査を目的とした系統的レビューと追試．修士学位論文，大阪大学，2016.

［Croux2010］C. Croux and C. Dehon. Influence functions of the Spearman and Kendall correlation measures. Statistical methods & applications, 19, 4, 497–515, 2010.

［Richard1987］Richard P Lippmann. An introduction to computing with neural nets. ASSP Magazine, IEEE, 4, 2, 4–22, 1987.

［Park2003］J. Park, R. G Harley, and G. Kumar Venayagamoorthy. Adaptive-critic-based optimal neurocontrol for synchronous generators in a power system using MLP/RBF neural networks. TIA, 39, 5, 1529–1540, 2003.

［Rproject］R: The R Project for Statistical Computing.

［Hitokoto2012］一言，桜庭，小野．ニューラルネットワークを用いた洪水予測システムの開発．こうえいフォーラム：日本工営技術情報，20, 67–72, 2012.

［Murakami2008］Masayuki Murakami. Practicality of modeling systems using the IDS method: Performance investigation and hardware implementation. PhD thesis, The University of Electro-Communications, 2008.

［Yamada2016］山田悠斗．ISBSGリポジトリを用いた見積もり研究に対する追試．特別研究報告，大阪大学 基礎工学部，2016.