ISBSG データを用いた見積もり研究に対する IPA/SEC データを用いた追試

山田悠斗
大阪大学　大学院情報科学研究科
y-yuto@ist.osaka-u.ac.jp

江川翔太
大阪大学　大学院情報科学研究科
s-egawa@ist.osaka-u.ac.jp

角田雅照
近畿大学　理工学部情報学科
tsunoda@info.kindai.ac.jp

楠本真二
大阪大学　大学院情報科学研究科
kusumoto@ist.osaka-u.ac.jp

要旨

ソフトウェア開発プロジェクトの見積もり技術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 *ISBSG* データが用いられることが多い．一方，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．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/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．

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 [1] 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を正確に見積もることが重要であるとされており [2]，見積もりの誤りがプロジェクト失敗へと繋がる場合がある．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 ISBSG データ [3] 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 [4] 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 [5, 6]．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，異なる条件では別の知見が得られるか等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いる見積もりの既存研究を対象にした追試を実施する．既存研究とは異なるデータセットで実験を行うことで，得られる知見に差異が生じるかを調査する．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繋げることができる．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．

過去 5 年間で ISBSG データが利用されていた論文の中から 4 本の論文 [7, 8, 9, 10] を選択し，IPA/SEC データを用いて追試を行った．その結果，4 本中 2 本の論文では既存研究と類似した知見が得られたが，他の 2 本の論文では既存研究と一部異なる知見が得られた．

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. では主な結果と今後の課題をまとめる．

2. 準備

本章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 [11]．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等が当てはめられ，説明変数である X_i 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．予測対象の実測値と予測値の残差が最小となるように a_i と b が決められる．

$$\hat{Y} = a_1X_1 + a_2X_2 + \cdots + a_nX_n + b \tag{1}$$

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 [12] 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，例えば MRE の平均値を MMRE などと表し，中央値を MdMRE などと表す．

$$\text{AR} = |\ \text{実測値} - \text{予測値}\ | \tag{2}$$

$$\text{MRE} = \frac{\text{AR}}{\text{実測値}} \tag{3}$$

$$\text{MER} = \frac{\text{AR}}{\text{予測値}} \tag{4}$$

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

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

回帰モデルの予測精度を表す指標として，他に決定係数がある．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 1 つにファンクションポイント法 [13] 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 から工数等の推定が行われる．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 [14]．

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

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

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，出力データ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，出力データ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

特に，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．

2.3. ISBSG データ

ISBSG データとは，ISBSG (The International Software Benchmarking Standards Group) [3] が世界 24 ヶ国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．開発工数やソフトウェアの規模，開発言語等のデータが収録されている．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 5000 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．

見積もりに関する研究の評価にはこの ISBSG データ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 データが使用されていた．

2.4. IPA/SEC データ

IPA/SEC データとは，独立行政法人情報処理推進機構 [15] 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．2014-2015 版では，3541 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 [4] 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．

2.5.1 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．妥当性については以下の分類 [16] 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 [17] 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．

2.5.2 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 [5, 6]．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等の事柄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の調査に繋げることができる．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

本章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. 論文の選択

ISBSG データが見積もり研究において一般的に用いられていることから，過去 5 年間での以下に示す主要国際会議，論文誌において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，ISBSG データが用いられている 22 本の論文を追試の対象とした．

- ACM-TOSEM (ACM Transactions On Software Engineering and Methodology)
- APSEC (ASIA-PACIFIC Software Engineering Conference)
- ESEM (Empirical Software Engineering and Measurement)
- ICSE (International Conference on Software Engineering)
- ICSM (International Conference on Software Maintenance)
- IEEE-TSE (IEEE Transactions on Software Engineering)
- IST (Information and Software Technology)
- JSS (Journal of Systems and Software)

追試を行うには，提案手法の詳細や実験手順，データの選別基準といった実験を再現するために必要となる情報が正確に記載されていなければならない．今回はそれらの情報が記載されている 4 本の文献 [7]，[8]，[9]，[10] を追試の対象として選択した．選択した論文をそれぞれ，生産性に基づくモデル作成に関する論文 [7]，カテゴリ変数の扱い方に関する論文 [8]，FP の簡易推定に関する論文 [9]，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 [10] と名称付けて説明する．

3.2. 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 [7]．ここでの生産性は以下の式 (7) によって定義される．

$$\text{生産性} = \frac{\text{FP}}{\text{工数}} \tag{7}$$

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全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．

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を 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

リリース 9 の ISBSG データ内の 593 データを対象として実験を行った．その結果，推測誤り率が 38 %以下

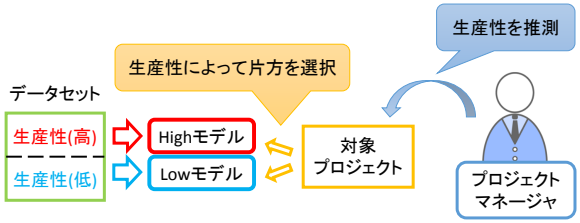

図 1. Two レベルモデルを用いた見積もり手法

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方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た．

3.3. カテゴリ変数の扱い方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 [8]．カテゴリ変数とは性別，職業など一般に数や量で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．今回は以下の 4 種類のモデルを対象とする．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．プロジェクト p_i がカテゴリ変数 m_j においてカテゴリ値 c に属するかどうかを表すダミー変数 $d_{ij}(c)$ は，以下の式 (8) で定義される．

$$d_{ij}(c) = \begin{cases} 1 & (\text{カテゴリ } c \text{ に属する}) \\ 0 & (\text{カテゴリ } c \text{ に属さない}) \end{cases} \quad (8)$$

層別を用いたモデル

各カテゴリ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，他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．

3.4. 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 [9]．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 [18]．

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を行った結果，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 [19] 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 [10]．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 [20]．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．

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

リリース 11 の ISBSG データ内の 49 データを対象として実験を行った結果，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いえる．

3.6. 使用する統計ツール

追試の中で回帰分析等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 [21] を使用する．

4. 追試の結果

本章では，対象となる 4 本の論文に対して IPA/SEC データを用いて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

4.1.1　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 1 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．

4.1.2　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 2 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．さらに，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．

表 2 を見ると，推測誤り率が 0 %の時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

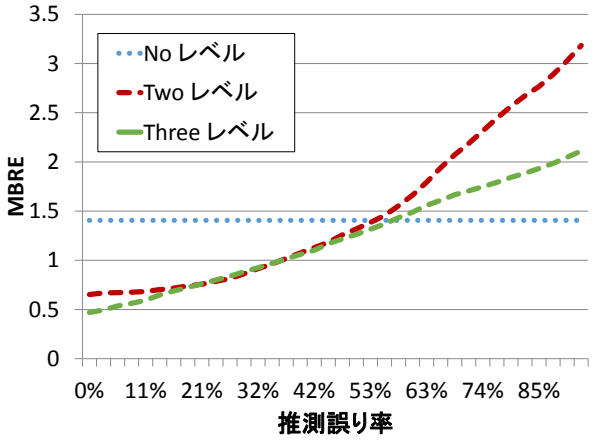

図 2. 追試における推測誤り率の増加による MBRE の推移

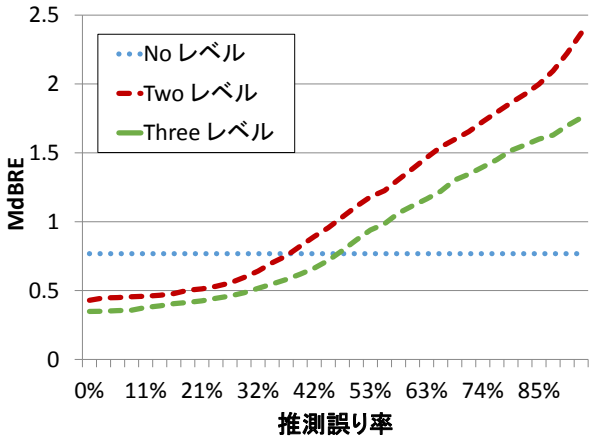

図 3. 追試における推測誤り率の増加による MdBRE の推移

の方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 レベルモデルの方が No 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

表 1. 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

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

表 2. 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

法の方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いえる．

4.2. カテゴリ変数の扱い方に関する論文

4.2.1 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，既存研究と同様に，まず表 1 に示す選別基準に従ってデータを抽出する．そして 3 つのカテゴリ変数 (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) の組み合わせからデータを分類し，10 データ未満のグループを省いた 88 データを使用する．

4.2.2 結果と考察

各モデルの説明変数には FP の対数と，カテゴリ変数である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 3 に示す．カテゴリ変数を使用しないモデルと，カテゴリ変数への対応方法が異なる 4 種類のモデルの見積もり精度が示されている．

表 3 から，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向上している．また，4 種類のモデルの精度に特徴的な差は無く，見積もりにおいて同程度の精度が得られる．

この結果を既存研究から得られた結果と比較すると，類似した傾向を示しており，追試によって既存研究と同様の知見が得られたと判断できる．このことから，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の研究成果の外的妥当性が高められたといえる．

4.3. FP の簡易推定に関する論文

4.3.1 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 4 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．

4.3.2 結果と考察

追試における実験の結果を表 5 に示す．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．

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は EI であるという結果が得られている．しかし表 5 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

表 3. カテゴリ変数の扱い方に関する論文の追試結果

評価指標	カテゴリ変数無し	ダミー変数	層別	交互作用	階層線形
MBRE	111.2%	103.3%	107.7%	107.7%	103.4%
MMRE	85.5%	75.8%	78.8%	78.8%	75.5%

表 4. FP の簡易推定に関する論文での IPA/SEC データの選別基準

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

場合は EI，スピアマンの順位相関係数の場合は EQ である．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わかる．

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．

4.4. NN を利用した見積もりに関する論文

4.4.1 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 6 に示す既存研究と同様の選別基準に従って抽出された 36 データを使用する．

4.4.2 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 7 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，一部のデータで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．MdAR と Pred(25) は MAR よりも外れ値に対して頑健であることから，追試では MdAR と Pred(25) の値からモデルの精度を比較する．

表 7 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．

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全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 (9) 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．

$$\begin{aligned}\ln(\text{Duration}) = 0.15 + 0.438 \times \ln(\text{AFP}) \\ -0.187 \times \ln(\text{MTS})\end{aligned} \tag{9}$$

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 (10) であり，決定係数は 0.398 となった．

$$\begin{aligned}\ln(\text{Duration}) = -0.669 + 0.435 \times \ln(\text{AFP}) \\ -0.205 \times \ln(\text{MTS})\end{aligned} \tag{10}$$

既存研究で ISBSG データから得られたモデルの方が追試で 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 データの方が強いことがわかる．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 [22]．さらに，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 [23]

表 5. 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

表 6. 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

本データの信頼性	A もしくは B
実績月数 (開発期間)	2ヶ月 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

．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．

5. 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．過去 5 年間の論文から 4 本の論文を対象とし，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論が得られた．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ことから，その原因に対する考察を行った．

今後の課題としては，まず見積もり分野における他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未だに不十分である．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．さらに，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．

6. 謝辞

本研究を行うにあたり，データを提供して頂くとともに多大なご助言を頂きました 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）の支援を受けている．

参考文献

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

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

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

表 7. 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

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

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

[6] 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.

[7] 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.

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

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

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

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

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

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

[14] 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.

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

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

[17] 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.

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

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

[20] 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.

[21] R: The R Project for Statistical Computing. https://www.r-project.org.

[22] 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.

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