

特別研究報告

題目

ISBSG リポジトリを用いた見積もり研究に対する追試

指導教員

楠本 真二 教授

報告者

山田 悠斗

平成 28 年 2 月 16 日

大阪大学 基礎工学部 情報科学科

ISBSG リポジトリを用いた見積もり研究に対する追試

山田 悠斗

内容梗概

ソフトウェア開発プロジェクトにおいては初期段階における規模の見積もりが重要である。そのため、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 リポジトリデータが多く用いられる。また、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。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。

本研究では、ISBSG リポジトリ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リポジトリ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 リポジトリ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、3 種類のリポジトリ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。さらに、異なる知見が得られた論文に対して追加調査を実施し、得られた知見に対する考察を行った。

主な用語

ソフトウェア開発見積もり
ファンクションポイント法
ISBSG リポジトリ
実証的ソフトウェア工学
追試 (replication study)

目次

1	まえがき	1
2	準備	2
2.1	見積もり分野における研究	2
2.1.1	工数	2
2.1.2	線形重回帰分析	2
2.1.3	ファンクションポイント法	3
2.1.4	ISBSG リポジトリ	4
2.2	実証的ソフトウェア工学	4
2.2.1	妥当性についての議論	4
2.2.2	追試 (replication study)	5
3	追試の準備	6
3.1	論文の選択	6
3.2	生産性に関する論文	6
3.2.1	内容	6
3.2.2	実験手順	7
3.2.3	得られた知見	9
3.3	カテゴリ変数に関する論文	10
3.3.1	内容	10
3.3.2	実験手順	12
3.3.3	得られた知見	13
3.4	ファンクションポイントに関する論文	13
3.4.1	内容	13
3.4.2	実験手順	13
3.4.3	得られた知見	14
3.5	ニューラルネットワークに関する論文	15
3.5.1	内容	15
3.5.2	実験手順	15
3.5.3	得られた知見	16
3.6	使用するリポジトリ	16
3.7	使用する統計ツール	17

4	追試の結果	18
4.1	生産性に関する論文	18
4.1.1	結果	18
4.1.2	考察	18
4.2	カテゴリ変数に関する論文	20
4.2.1	結果	20
4.2.2	考察	20
4.3	ファンクションポイントに関する論文	20
4.3.1	結果	20
4.3.2	考察	21
4.4	ニューラルネットワークに関する論文	21
4.4.1	結果	21
4.4.2	考察	22
5	ファンクションポイントに関する論文の追加調査	24
5.1	調査内容	24
5.2	調査結果	24
5.3	考察	25
6	あとがき	26
	謝辞	27
	付録	28
	参考文献	31

表目次

1	生産性に関する論文での ISBSG リポジトリの選別基準	9
2	生産性に関する論文の実験結果	10
3	カテゴリ変数に関する論文の実験結果	13
4	ファンクションポイントに関する論文での ISBSG リポジトリの選別基準 . . .	14
5	ファンクションポイントに関する論文の実験結果	14
6	ニューラルネットワークに関する論文での ISBSG リポジトリの選別基準 . . .	16
7	ニューラルネットワークに関する論文の実験結果	16
8	生産性に関する論文の追試結果	18
9	カテゴリ変数に関する論文の追試結果	20
10	ファンクションポイントに関する論文の追試結果	21
11	ニューラルネットワークに関する論文の追試結果	21
12	ISBSG リポジトリを用いた実験の再現	22
13	ファンクションポイントに関する論文の追加調査の結果	25

目 次

1	生産性を考慮に加えた見積もり手法	7
2	Two レベルモデルにおける実験の手順	8
3	推測誤り率の増加による MBRE の推移	11
4	推測誤り率の増加による MdBRE の推移	11
5	追試における推測誤り率の増加による MBRE の推移	19
6	追試における推測誤り率の増加による MdBRE の推移	19

1 まえがき

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規模の見積もり [1] がある．ソフトウェア開発においては初期段階でのソフトウェア開発規模の正確な見積もり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規模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 リポジトリデータ [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 リポジトリが利用されていた．

また，実証的ソフトウェア工学 [4, 5] 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 [6, 7]．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 リポジトリ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リポジトリ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 リポジトリが利用されていた論文の中から 4 本の論文 [8, 9, 10, 11] を選択し，ISBSG リポジトリと異なる 3 種類のリポジトリ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．また，異なる知見が得られた論文に関する追加調査を実施することによって，既存研究と結果が異なった原因に対する考察を行った．

以降，2 章では研究の背景となる関連研究について述べる．3 章では追試を行うための準備事項について述べる．4 章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 章では追試を行った結果を踏まえた追加調査の結果と考察について述べる．最後に 6 章で本報告のまとめを述べる．

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 見積もり分野における研究

ソフトウェア開発プロジェクトにおいては、初期段階におけるプロジェクト全体の規模の見積もりが重要であるとされており [2]，見積もりの誤りが開発費の超過や納期遅れといったプロジェクト失敗の大きな要因となる場合がある。このような失敗を未然に防ぐためにソフトウェア開発規模の見積もりに関する研究が盛んに行われている。

2.1.1 工数

ソフトウェア開発プロジェクトの規模を表す指標の 1 つに工数がある。工数とは以下の式 (1) から算出される延べ作業時間を表す数値である。ソフトウェア開発の初期段階で予めプロジェクト全体の工数を知ることによって、適切なスケジュールを立てたり、適切に人員を配置することが可能となる。ソフトウェア開発規模の見積もりではこの工数を用いる場合が多く、様々な工数予測手法がこれまでに考案されている [12]。

$$\text{工数} = \text{開発期間} \times \text{開発チームの人員数} \quad (1)$$

2.1.2 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 [13]。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 (2) 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。

$$\hat{Y} = a_1X_1 + a_2X_2 + \dots + a_nX_n + b \quad (2)$$

精度の推定には以下の式 (3)~(7) によって求められる 5 つの評価指標 [14] 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。

また、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。

$$AR = | \text{実測値} - \text{予測値} | \quad (3)$$

$$MRE = \frac{AR}{\text{実測値}} \quad (4)$$

$$MER = \frac{AR}{\text{予測値}} \quad (5)$$

$$BRE = \begin{cases} MRE & (\text{予測値} - \text{実測値} \geq 0) \\ MER & (\text{予測値} - \text{実測値} < 0) \end{cases} \quad (6)$$

$$\text{Pred}(25) = \frac{\text{評価指標の値が 0.25 以下であるデータ数}}{\text{全データ数}} \quad (7)$$

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。

2.1.3 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 [15] 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 (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 [16]。

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ファンクションポイントと言い、アプリケーションファンクションポイント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ファンクションポイントと言う。

2.1.4 ISBSG リポジトリ

ISBSG リポジトリ [3] とは、ISBSG (The International Software Benchmarking Standards Group) 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、最新のリポジトリには 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 リポジトリ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 リポジトリが使用されていた。

2.2 実証的ソフトウェア工学

実証的ソフトウェア工学 [4, 5] 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。

2.2.1 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 [17] 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 [18] 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。

2.2.2 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 [6, 7]。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以下のような事柄を調査することができる。

- 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

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 リポジトリ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年間で以下の示す主要国際会議，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 リポジトリが用いられている22本の論文を追試の対象とした．対象となる論文のタイトルと著者名，出典名，出版年を付録に示す。

- 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)

今回はその中から4本の論文を選択した．選択した論文を，生産性に関する論文 [8] (付録における [1])，カテゴリ変数に関する論文 [9] (付録における [3])，ファンクションポイント (FP) に関する論文 [10] (付録における [8])，ニューラルネットワーク (NN) に関する論文 [11] (付録における [19]) と名称付けて説明する。

3.2 生産性に関する論文

3.2.1 内容

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新しい工数見積もりモデルの作成方法を提案している．ここでの生産性は以下の式 (8) によって定義される。

$$\text{生産性} = \frac{\text{FP}}{\text{工数}} \quad (8)$$

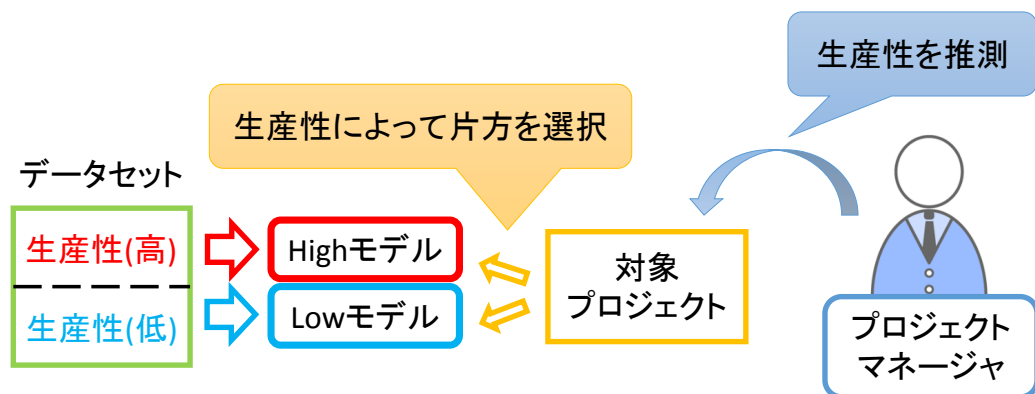


図 1: 生産性を考慮に加えた見積もり手法

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。実験では、プロジェクトマネージャが生産性の推測を誤る可能性も考慮に加えた方法が取られている。提案手法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。

3.2.2 実験手順

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。

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たモデル

Two レベルモデル

生産性を High(高), Low(低) の2段階に分類し、それらを考慮に加えて見積もりを行うモデル

Three レベルモデル

生産性を High(高), Medium(中), Low(低) の3段階に分類し、それらを考慮に加えて見積もりを行うモデル

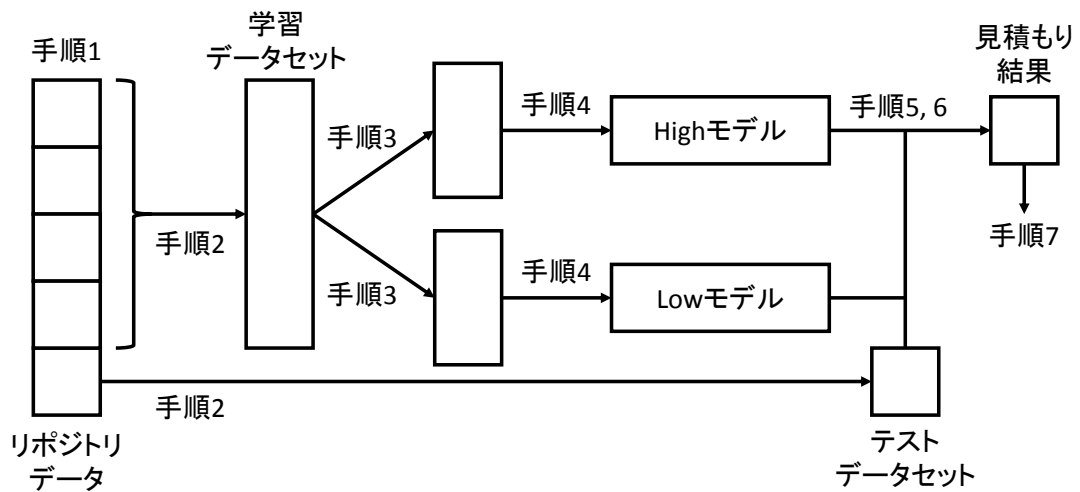


図 2: Two レベルモデルにおける実験の手順

この中で Two レベルモデルを実施する際の手順を以下に示す。また、この手順を図で表したものを図 2 に示す。

手順 1 リポジットリ内のデータを 5 つのサブセットに等分する

手順 2 5 つのサブセットのうち 1 つをテストデータセットとし、他の 4 つのサブセットを合わせたものを学習データセットとする

手順 3 学習データセットを生産性の高いデータから順に並べ直し、上から 50 %を High グループ、それ以外を Low グループに分類する

手順 4 High グループと Low グループからそれぞれ工数見積もりモデルを作成し、High モデル、Low モデルとする

手順 5 手順 3 で学習データセットを生産性によって分類した際の閾値に基づいて、テストデータセットを High グループと Low グループに分類する

手順 6 テストデータセット内の High グループには High モデルを、Low グループには Low モデルを適用して工数見積もりを行う

手順 7 テストデータセット全体の見積もり精度を評価する

手順 8 テストデータセット内のデータのうち、閾値に近い生産性を持つデータについてグループの分類を変更する (変更したデータ数の割合を推測誤り率と呼ぶ)

手順 9 手順 8 で推測誤り率を増加させながら、手順 5~8 を繰り返す

手順 10 手順 2 でテストデータセットとなるサブセットを変更し、全てのサブセットがテストデータセットとなるまで手順 2~9 を繰り返す

手順 11 試行回数を増やすため、リポジトリ内のデータをランダムに並び替えて手順 1~10 を計 4 回繰り返す

No レベルモデルでは手順 3, 手順 5, 手順 8 を省いて生産性による分類を行わずに見積もりを行う。Three レベルモデルでは手順 3, 手順 5 でのデータセットの分類を High グループ, Medium グループ, Low グループの 3 段階に変更して見積もりを行う。

手順 7 では各プロジェクトについて MRE, MER, BRE を計算した後、テストデータセット全体でのそれぞれの平均値 (MMRE, MMER, MBRE) と中央値 (MdMRE, MdMER, MdBRE) を計算する。

3.2.3 得られた知見

リリース 9 の ISBSG リポジトリに蓄積されている 3026 データのうち、表 1 の選別基準に従って抽出された 593 データを使用する。

既存研究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、手順 8 の推測誤り率が 0 %の時点での数値である。

表 1: 生産性に関する論文での ISBSG リポジトリの選別基準

FP Standard (規模計測手法)	IFPUG4
Normalised Work Effort (工数)	記録された工数と等しい
Data Quality Rating (データの信頼性)	A もしくは B
Resource Level (工数の計測対象部分)	1 (開発自体に必要とした工数)
Functional Size (アプリケーションファンクションポイント)	欠損していない
Language Type (開発言語)	欠損していない
Development Type (開発種別)	欠損していない
Development Platform (開発プラットフォーム)	欠損していない

さらに、手順8で推測誤り率を増加させた際の、各モデルにおける見積もり精度の推移を図3、図4に示す。図3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4は横軸に推測誤り率が、縦軸にMdBREの数値が示されている。図中のNo subsetがNoレベルモデル、Two subsetがTwoレベルモデル、Three subsetがThreeレベルモデルの見積もり精度である。

表2より、推測誤り率が0%の時はいずれの評価指標でもThreeレベルモデル、Twoレベルモデル、Noレベルモデルの順に見積もりの精度が高い。また図3、図4を見ると、推測誤り率が増加するほどTwoレベルモデルと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38%以下の時点では、MBREとMdBREのどちらにおいてもTwoレベルモデルとThreeレベルモデルの方が従来の手法を用いた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法の方が工数見積もりにおける見積もり精度が高いと言える。

3.3 カテゴリ変数に関する論文

3.3.1 内容

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。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。

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(9)で定義される。

表 2: 生産性に関する論文の実験結果

レベル	MMRE	MdMRE	MMER	MdMER	MBRE	MdBRE
No	1.04	0.51	0.84	0.57	1.45	0.86
Two	0.67	0.39	0.57	0.42	0.89	0.54
Three	0.45	0.32	0.42	0.30	0.58	0.38

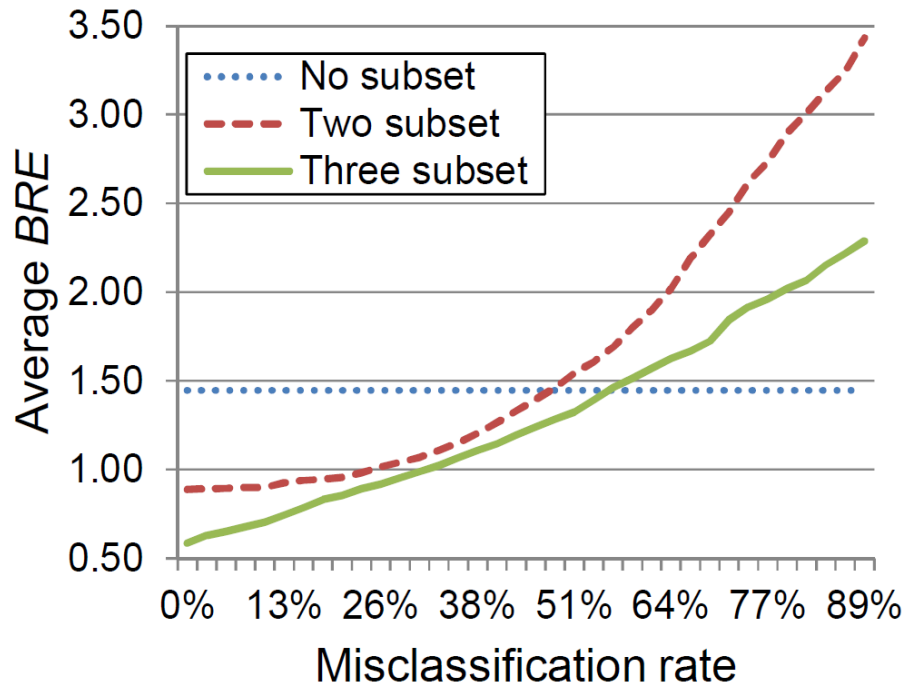


図 3: 推測誤り率の増加による MBRE の推移

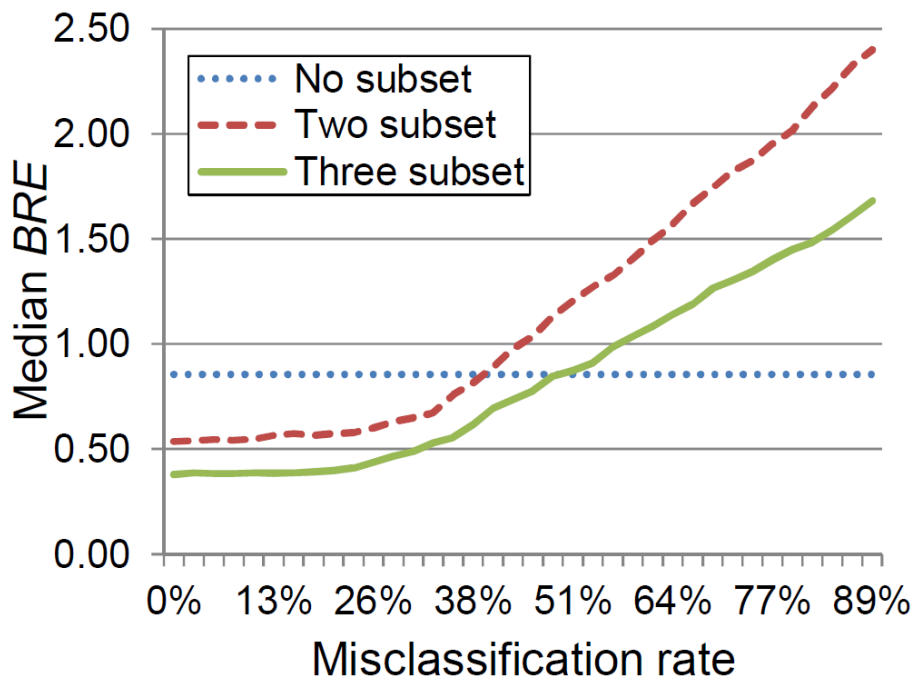


図 4: 推測誤り率の増加による MdBRE の推移

$$d_{ij}(c) = \begin{cases} 1 & (\text{カテゴリ } c \text{ に属する}) \\ 0 & (\text{カテゴリ } c \text{ に属さない}) \end{cases} \quad (9)$$

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、リポジトリ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新たな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。

これらのモデルを使用して見積もりを行い、その精度を比較する。

3.3.2 実験手順

FP と 3 種類のカテゴリ変数を使用して、上記の 4 種類の工数見積もりモデルを作成する。これらのモデルの精度を、FP のみを説明変数とする単回帰モデルと比較することで評価する。評価の指標には MBRE と MMRE を用いる。

また、各モデルの精度を推定する手順としては n 分割交差検証法 [19] を用いる。手順を以下に示す。

手順 1 使用するデータを n 個のサブセットに等分する

手順 2 n 個のサブセットのうち 1 つをテストデータセットとし、残りのサブセットを合わせたものを学習データセットとする

手順 3 学習データセットから見積もりモデルを作成する

手順 4 手順 3 で作成したモデルを用いてテストデータセットの見積もりを行い、精度を評価する

手順 5 手順 2 でテストデータセットとなるサブセットを変更し、全てのサブセットがテストデータセットとなるまで手順 2~4 を繰り返す

今回は $n = 10$ とした 10 分割交差検証法を用いて実験を行う。

3.3.3 得られた知見

リリース 9 の ISBSG リポジトリに蓄積されている 3026 データのうち、まず表 1 の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 (Language Type, Development Type, Development Platform) 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 558 データを使用する。

既存研究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ことがわかる。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。

3.4 ファンクションポイントに関する論文

3.4.1 内容

ファンクションポイント (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。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。

3.4.2 実験手順

まず、FP と 5 種類の基本機能要素の規模について相関関係を調べる。相関関係の調査には以下に示す 2 種類の相関係数 [20] を用いる。いずれも、相関係数の値が高いほど 2 変数の相関関係が強いと評価される。

ケンドールの順位相関係数

2 変数の順位値が各データの中でどれだけ一致しているかを表す指標

表 3: カテゴリ変数に関する論文の実験結果

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	166.1%	147.0%	151.6%	147.2%	148.9%
MMRE	112.5%	102.4%	106.6%	103.2%	103.9%

スピアマンの順位相関係数

2変数の順位値についてピアソンの積率相関係数を計算したもの

各相関係数の値が最も高い要素が、FPと最も相関の強い要素である。この要素の規模のみを説明変数とし、FPを目的変数とした単回帰モデルを作成し、それを簡易FP推定モデルとする。

3.4.3 得られた知見

リリース11のISBSGリポジトリに蓄積されている5052データのうち、表4の選別基準に従って抽出された600以上のデータを使用する。

既存研究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。

表5から、FPと最も相関の強い要素はEIとなることがわかる。EIを説明変数とすることによって、式(10)のような簡易FP推定モデルが得られる。

$$FP = 79 + 1.9 \times (\text{EIの規模}) \quad (10)$$

表4: ファンクションポイントに関する論文でのISBSGリポジトリの選別基準

Data Quality Rating (データの信頼性)	AもしくはB
UFP rating (FPに関するデータの信頼性)	AもしくはB
Input count (外部入力)	欠損していない
Output count (外部出力)	欠損していない
File count (内部論理ファイル)	欠損していない

表5: ファンクションポイントに関する論文の実験結果

要素	ケンドール	スピアマン
EI	0.658	0.839
EO	0.597	0.776
EQ	0.528	0.692
ILF	0.619	0.804
EIF	0.264	0.363

3.5 ニューラルネットワークに関する論文

3.5.1 内容

ニューラルネットワーク (NN) [21] 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する。NN とは脳機能に見られる特性を計算機上のシミュレーションによって表現するネットワークモデルであり、結合荷重を持つニューロン同士の結合により構成される。今回は以下の 2 種類の NN [22] を使用する。

MLP (Multi Layer Perceptron)

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 (Radial Basis Function Neural Network)

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

3.5.2 実験手順

以下の手順に基づいて実験を行う。なお，手順 2，手順 4 における LOOCV (Leave-One-Out Cross-Validation) とは n 分割交差検証法の n を使用するデータ数とした手法である。例えばデータセットに蓄積されたデータ数が 20 である場合，データセットを 20 分割してそのうち 1 つをテストデータとし，残りの 19 データを学習データとする。また，手順 1 と手順 4 でデータをモデルに与える際に対数を取るのは，データを正規化して利用しやすい形に変形するためである。

手順 1 MLP モデルと RBFNN モデルへの入力としてソフトウェアの規模を指す調整済みファンクションポイントの対数と開発チームの最大人員数の対数を，出力として開発期間の対数を与える

手順 2 各 NN の中間層にあるニューロンの数を 2~40 まで変化させ，LOOCV を用いて開発期間の見積もりを行う

手順 3 見積もり精度が最も良くなったニューロン数を持つ NN モデルを採用する

手順 4 調整済みファンクションポイントの対数と開発チームの最大人員数の対数を説明変数とし，開発期間の対数を目的変数とする重回帰モデルを作成し，LOOCV を用いて開発期間の見積もりを行う

手順 5 各 NN モデルと重回帰モデルの見積もり精度を，AR の平均値である MAR と中央値である MdAR，また Pred(25) を用いて比較する

3.5.3 得られた知見

リリース 11 の ISBSG リポジトリに蓄積されている 5052 データのうち、表 6 の選別基準に従って抽出された 49 データを使用する。

既存研究における実験の結果を表 7 に示す。ニューロン数は NN の中間層に置かれたニューロンの数を表す。

表 7 から、2 種類の NN モデルの見積もり精度が共に重回帰モデルよりも高いことがわかる。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 使用するリポジトリ

各論文の追試には、以下の 3 種類のリポジトリを使用する。

SEC リポジトリ

独立行政法人情報処理推進機構 [23] が日本に存在する組織や企業から実開発のデータ

表 6: ニューラルネットワークに関する論文での ISBSG リポジトリの選別基準

Data Quality Rating (データの信頼性)	A もしくは B
UFP rating (FP に関するデータの信頼性)	A もしくは B
FP Standard (規模計測手法)	IFPUG4+ もしくは NESMA
Max Team Size (開発期間)	2ヶ月 以上
Project Elapsed Time (開発チームの最大人員数)	2人 以上
Development Platform (開発プラットフォーム)	Mainframe
Language Type (開発言語)	3GL (第 3 世代言語)
Development Type (開発種別)	New (新規開発)

表 7: ニューラルネットワークに関する論文の実験結果

モデル	ニューロン数	MAR	MdAR	Pred(25)
重回帰	-	0.29	0.25	0.51
MLP	25	0.18	0.15	0.80
RBFNN	23	0.21	0.19	0.61

を収集し、整理したデータセット。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 (2014-2015 版)。

某社リポジトリ

金融保険業で用いるソフトウェアの開発を行う某社から提供を受けたデータセット。98 プロジェクトのデータが144 項目に分けて蓄積されている。

Chinese Data

tera-PROMISE リポジトリ [24] 中にある中国系ソフトウェア開発企業のプロジェクトに関するデータセット。499 プロジェクトのデータが18 項目に分けて蓄積されている。

ただし、生産性に関する論文、カテゴリ変数に関する論文、NN に関する論文については、某社リポジトリと Chinese Data は実験に必要な項目が欠如している。よって、それら3 本の論文では SEC リポジトリのみを、FP に関する論文では3 種類全てのリポジトリを用いて追試を行う。

3.7 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 [25] 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 生産性に関する論文

4.1.1 結果

SECリポジトリに蓄積されているデータのうち、既存研究と同様の選別を行うことによって得られた189データを使用する。追試における実験の結果を表8に示す。TwoレベルモデルとThreeレベルモデルの数値は、既存研究における手順8の推測誤り率が0%の時点での数値である。

さらに、手順8で推測誤り率を増加させた際の見積もり精度の推移を図5、図6に示す。図5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6は横軸に推測誤り率が、縦軸にMdBREの数値が示されている。

4.1.2 考察

表8を見ると、推測誤り率が0%の時はいずれの評価指標でもThreeレベルモデル、Twoレベルモデル、Noレベルモデルの順に見積もりの精度が高い。また図5、図6を見ると、推測誤り率が増加するほどTwoレベルモデルと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レベルモデルとThreeレベルモデルの方が従来の手法を用いた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。よって、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。

表 8: 生産性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

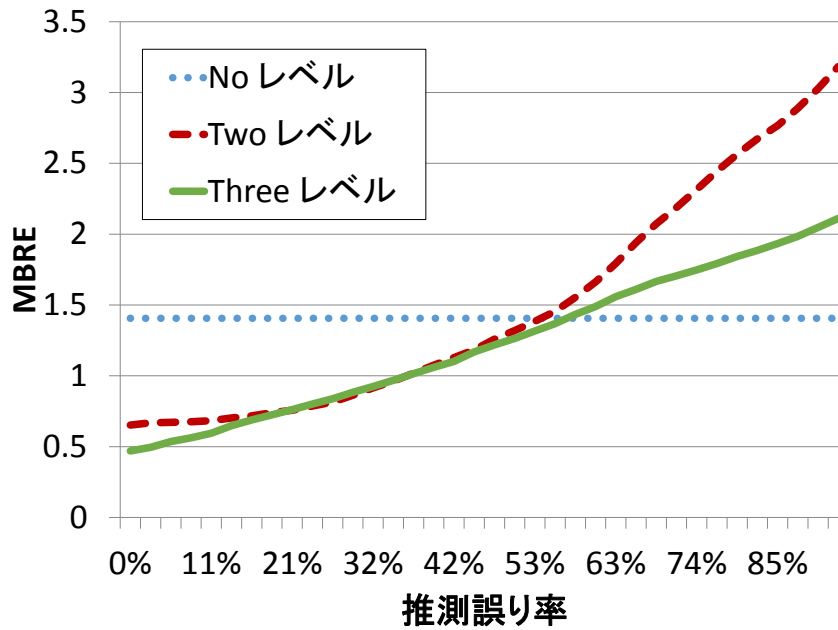


図 5: 追試における推測誤り率の増加による MBRE の推移

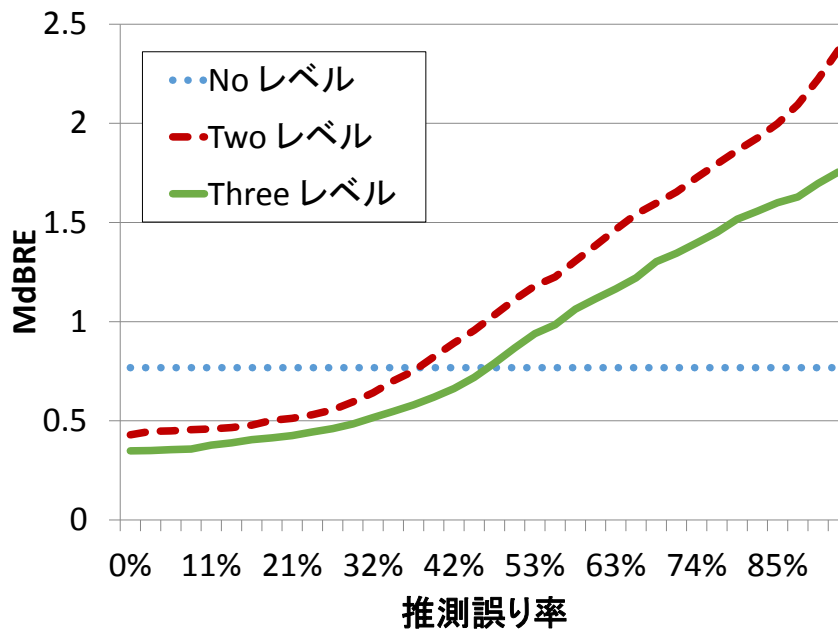


図 6: 追試における推測誤り率の増加による MdBRE の推移

4.2 カテゴリ変数に関する論文

4.2.1 結果

SEC リポジトリに蓄積されているデータのうち、既存研究と同様の選別を行うことによって得られた 88 データを使用する。追試における実験の結果を表 9 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。

4.2.2 考察

表 9 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、追試を行うことによって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。

4.3 ファンクションポイントに関する論文

4.3.1 結果

既存研究と同様の選別を行うことによって、SEC リポジトリに蓄積されているデータから得られた 122 データを、某社リポジトリに蓄積されているデータから得られた 55 データを、Chinese Data に蓄積されているデータから得られた 372 データを使用する。追試における実験の結果を表 10 に示す。表の縦軸はそれぞれ以下のことを表している。

リポジトリ (データ数)

使用するリポジトリの種類と、使用するデータの数

ケンドール

ケンドールの順位相関係数を用いた際に FP と最も相関の強くなった基本機能要素

表 9: カテゴリ変数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%

スパイマン

スパイマンの順位相関係数を用いた際に FP と最も相関の強くなった基本機能要素

4.3.2 考察

既存研究で ISBSG リポジトリを用いた場合、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 10 を見ると、FP と相関の強い要素は SEC リポジトリでは EI または EQ、某社リポジトリでは EO、Chinese Data では EI となっている。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、リポジトリ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、使用するリポジトリ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 ニューラルネットワークに関する論文

4.4.1 結果

SEC リポジトリに蓄積されているデータのうち、既存研究と同様の選別を行うことによって得られた 36 データを使用する。追試における実験の結果を表 11 に示す。

表 10: ファンクションポイントに関する論文の追試結果

リポジトリ (データ数)	ケンドール	スパイマン
SEC (122)	EI	EQ
某社 (55)	EO	EO
Chinese Data (372)	EI	EI

表 11: ニューラルネットワーク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

4.4.2 考察

SEC リポジトリで追試を行う前に、研究で用いられたものと同じリリースの ISBSG リポジトリを用いて実験の再現を試みた。その結果が表 12 である。

表 12 を見ると、2 種類の NN モデルの MdAR についてはいずれも重回帰モデルより値が低くなっており、Pred(25) についてはいずれも重回帰モデルより値が高くなっている。これは既存研究と同様の結果である。しかし、MAR については MLP モデルの値は重回帰モデルと同じであり、RBFNN モデルの値は重回帰モデルより高くなっている。これは既存研究と異なる結果である。平均値である MAR の値が高いことと、中央値である MdAR の値が低く Pred(25) の値が高いことから、こちらの準備した実験環境で NN モデルを用いると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ると考えられる。よって、追試では MdAR と Pred(25) の値からモデルの精度を比較する。

以上を踏まえて表 11 を見ると、2 種類の NN モデルの Pred(25) についてはいずれも値が重回帰モデルより高くなっている。しかし、MdAR については MLP モデルの値は重回帰モデルより高く、RBFNN モデルの値は重回帰モデルより低くなっている。よって、開発期間の見積もり精度が重回帰モデルより高くなる NN モデルは RBFNN モデルのみであると言える。これは既存研究と異なる結果となっている。

次に、SEC リポジトリを実験対象とした場合に 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 ISBSG リポジトリにおける NN モデルの入出力値の関係を調査している。その際、入力値である調整済みファンクションポイントの対数と開発チームの最大人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 (11) である。AFP (Adjusted Function Points) は調整済みファンクションポイントを、MTS (Max Team Size) は開発チームの最大人員数を、Duration は開発期間を指す。このモデルの決定係数は 0.56 となった。

$$\ln(\text{Duration}) = 0.15 + 0.438 \times \ln(\text{AFP}) - 0.187 \times \ln(\text{MTS}) \quad (11)$$

表 12: ISBSG リポジトリを用いた実験の再現

モデル	ニューロン数	MAR	MdAR	Pred(25)
重回帰	-	0.34	0.30	0.39
MLP	21	0.34	0.21	0.57
RBFNN	26	0.40	0.26	0.47

SEC リポジトリでも同様のモデルを作成した。そのモデルが以下の式 (12) である。このモデルの決定係数は 0.398 となった。

$$\ln(\text{Duration}) = -0.669 + 0.435 \times \ln(\text{AFP}) - 0.205 \times \ln(\text{MTS}) \quad (12)$$

ISBSG リポジトリから得られたモデルの方が、SEC リポジトリ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 SEC リポジトリより ISBSG リポジトリ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 [26]。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 [27]。

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 ファンクションポイントに関する論文の追加調査

本章ではファンクションポイント (FP) に関する論文についての追加調査を行った結果及び考察について説明する。

5.1 調査内容

FP に関する論文の実験を行った結果である表 2 と表 10 を見ると、ISBSG リポジトリ、SEC リポジトリ、Chinese Data では FP と相関の強い要素として EI が導かれているのに対し、某社リポジトリでは相関の強い要素として EO のみが導かれている。なぜ某社リポジトリを使用した結果が他の 3 つのリポジトリを使用した結果と大きく異なったのか、その原因を調査する。

調査するにあたり、各リポジトリが持つ以下の特徴に着目した。

ISBSG リポジトリ、SEC リポジトリ、Chinese Data が持つ特徴

様々な業種の企業からプロジェクトデータが収集されている

某社リポジトリが持つ特徴

1 つの企業からプロジェクトデータを収集しており、特定の業種のデータのみ存在する

これらの特徴と、某社リポジトリのみ相関の強い要素が異なっていたことを踏まえると、FP と相関の強い要素は業種ごとに偏りが生じると考えられる。

そこで、ISBSG リポジトリ、SEC リポジトリ、Chinese Data から使用するデータをそれぞれ業種別に分類し、業種ごとに FP と各基本機能要素の相関関係を調査する。しかし、SEC リポジトリについては業種別に分類すると全ての業種においてデータ数が 30 未満となり、統計的なデータとしては数が不十分となった。Chinese Data については業種を把握できる項目が存在せず、業種別に分類することができなかった。よって、ISBSG リポジトリのみを調査の対象とする。

5.2 調査結果

ISBSG リポジトリから使用するデータを業種別に分類してデータ数の多い上位 5 つを抽出した結果、「Insurance」、「Communications」、「Government」、「Banking」、「Manufacturing」が抽出された。ISBSG リポジトリから抽出された 5 つの業種について相関関係を調査した結果を表 13 に示す。なお、金融保険会社である某社との比較のために、Insurance と Banking を統合したデータについても調査を行っている。

5.3 考察

表 13 を見ると、業種ごとに導かれる要素が異なっていることがわかる。また、表 10 の某社リポジトリと表 13 の Insurance + Banking を比べると、共に相関の強い要素は EO となっている。

このことから、FP と相関の強い要素には業種ごとに決まった特徴があり、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業種ごとに適した要素を選択する必要があると考えられる。

次に、各業種と導かれた基本機能要素の関連性を考える。ISBSG リポジトリにおける 5 つの業種のうち、FP と最も相関の強い要素と他の要素について相関係数に有意差が存在する業種は Insurance と Communications となった。この 2 つの業種について、業種が持つ特徴から各要素が導かれた理由を考える。

表 13 から、Insurance において FP と相関の強い要素は EO であることがわかる。これは、様々な保険サービスが新規導入されていくことで、開発したソフトウェアに新たな出力が付け足されていくためと考えられる。一方、表 13 から Communications において FP と相関の強い要素は ILF であることがわかる。これは、膨大な通信情報や通信履歴が様々な種類別に分けられてシステム内で保持されるためと考えられる。

表 13: ファンクションポイントに関する論文の追加調査の結果

業種 (データ数)	ケンドール	スピアマン
Insurance (131)	EO	EO
Communications (113)	ILF	ILF
Government (90)	EI	EI
Banking (47)	EI	EI
Manufacturing (39)	EI	ILF
Insurance + Banking (178)	EO	EO

6 あとがき

本研究では、ISBSG リポジトリ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、3種類のリポジトリ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、追加調査も交えてその原因に対する考察を行った。

本研究の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、今回のような追試を続けていくことで、妥当性に関する議論を継続的に行い、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より詳細な追加調査も今後の課題として考えられる。例えば、ファンクションポイントに関する論文の追加調査によって業種ごとの差を検証したが、各業種で主に用いられているソフトウェアの種類に対する新たな調査が挙げられる。これによって各業種と導かれた要素との関連性について、より詳細な考察が可能となる。

謝辞

本研究の全過程を通し，理解ある親身なご指導を賜り，的確なご助言を頂きました 楠本 真二 教授に心より感謝申し上げます。

本研究を行うにあたり，的確なご指導およびご助言を頂きました 肥後 芳樹 准教授に心より感謝申し上げます。

本研究を行うにあたり，丁寧かつ適切なご指導を頂きました 松本 真佑 助教に深く感謝申し上げます。

本研究を進めるにあたって，様々な点でご助力を頂きました 近畿大学工学部情報学科の 角田 雅照 講師に心より感謝申し上げます。

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。

本研究を進めるにあたって，様々な点でご協力頂きました 鈴鹿 久佳 様に深謝致します。

本研究を進めるにあたり，適切なご助言および多大なるご助力を頂き，さらに日常でも声をかけて頂きました 大阪大学大学院情報科学研究科コンピュータサイエンス専攻博士前期課程2年の 江川 翔太 氏に心より感謝申し上げます。

また本研究を進めるにあたり，多大なるご助言及びご協力を頂きました 大阪大学大学院情報科学研究科コンピュータサイエンス専攻博士前期課程1年の 古田 雄基 氏に心より感謝申し上げます。

本研究を進めるにあたり，様々な形で励まし，ご助言を頂きました その他の楠本研究室の皆様のご協力に心より感謝致します。

最後に，本研究に至るまでに，講義，演習，実験等でお世話になりました大阪大学基礎工学部情報科学の諸先生方に，この場を借りて心から御礼申し上げます。

付録

- [1] Tsunoda, Masafumi, et al. "Incorporating expert judgment into regression models of software effort estimation." *Software Engineering Conference (APSEC), 2012 19th Asia-Pacific*. Vol. 1. IEEE, 2012.
- [2] Azzeh, Mohammad, Daniel Neagu, and Peter Cowling. "Improving analogy software effort estimation using fuzzy feature subset selection algorithm." *Proceedings of the 4th international workshop on Predictor models in software engineering*. ACM, 2008.
- [3] Tsunoda, Masateru, Sousuke Amasaki, and Akito Monden. "Handling categorical variables in effort estimation." *Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement*. ACM, 2012.
- [4] Fernández-Diego, Marta, and José-María Torralba-Martínez. "Discretization methods for NBC in effort estimation: An empirical comparison based on ISBSG projects." *Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement*. ACM, 2012.
- [5] González-Ladrón-de-Guevara, Fernando, and Marta Fernández-Diego. "ISBSG variables most frequently used for software effort estimation: A mapping review." *Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*. ACM, 2014.
- [6] Minku, Leandro L., and Xin Yao. "How to make best use of cross-company data in software effort estimation?." *Proceedings of the 36th International Conference on Software Engineering*. ACM, 2014.
- [7] Wen, Jianfeng, et al. "Systematic literature review of machine learning based software development effort estimation models." *Information and Software Technology* 54.1 (2012): 41-59.
- [8] Lavazza, Luigi, Sandro Morasca, and Gabriela Robiolo. "Towards a simplified definition of Function Points." *Information and Software Technology* 55.10 (2013): 1796-1809.
- [9] Seo, Yeong-Seok, Doo-Hwan Bae, and Ross Jeffery. "AREION: Software effort estimation based on multiple regressions with adaptive recursive data partitioning." *Information and Software technology* 55.10 (2013): 1710-1725.

- [10] Minku, Leandro L., and Xin Yao. "Ensembles and locality: Insight on improving software effort estimation." *Information and Software Technology* 55.8 (2013): 1512-1528.
- [11] Lokan, Chris, and Emilia Mendes. "Investigating the use of duration-based moving windows to improve software effort prediction: A replicated study." *Information and Software Technology* 56.9 (2014): 1063-1075.
- [12] Idri, Ali, Fatima azzahra Amazal, and Alain Abran. "Analogy-based software development effort estimation: A systematic mapping and review." *Information and Software Technology* 58 (2015): 206-230.
- [13] Zhang, Wen, Ye Yang, and Qing Wang. "Using Bayesian regression and EM algorithm with missing handling for software effort prediction." *Information and Software Technology* 58 (2015): 58-70.
- [14] Huang, Jianglin, Yan-Fu Li, and Min Xie. "An empirical analysis of data preprocessing for machine learning-based software cost estimation." *Information and Software Technology* 67 (2015): 108-127.
- [15] Mittas, Nikolaos, and Lefteris Angelis. "Visual comparison of software cost estimation models by regression error characteristic analysis." *Journal of Systems and Software* 83.4 (2010): 621-637.
- [16] Azzeh, Mohammad, Daniel Neagu, and Peter I. Cowling. "Analogy-based software effort estimation using Fuzzy numbers." *Journal of Systems and Software* 84.2 (2011): 270-284.
- [17] Nassif, Ali Bou, Danny Ho, and Luiz Fernando Capretz. "Towards an early software estimation using log-linear regression and a multilayer perceptron model." *Journal of Systems and Software* 86.1 (2013): 144-160.
- [18] Mittas, Nikolaos, et al. "Integrating non-parametric models with linear components for producing software cost estimations." *Journal of Systems and Software* 99 (2015): 120-134.
- [19] López-Martín, Cuauhtémoc, and Alain Abran. "Neural networks for predicting the duration of new software projects." *Journal of Systems and Software* 101 (2015): 127-135.

- [20] Minku, Leandro L., and Xin Yao. "Software effort estimation as a multiobjective learning problem." *ACM Transactions on Software Engineering and Methodology (TOSEM)* 22.4 (2013): 35.
- [21] Dejaeger, Karel, et al. "Data mining techniques for software effort estimation: a comparative study." *Software Engineering, IEEE Transactions on* 38.2 (2012): 375-397.
- [22] Mittas, Nikolaos, and Lefteris Angelis. "Ranking and clustering software cost estimation models through a multiple comparisons algorithm." *Software Engineering, IEEE Transactions on* 39.4 (2013): 537-551.

参考文献

- [1] Jane Radatz, Anne Geraci, and Freny Katki. IEEE standard glossary of software engineering terminology. *IEEE Std*, Vol. 610.12-1990, , 1990.
- [2] Barry Boehm, Chris Abts, and Sunita Chulani. Software development cost estimation approaches - A survey. *Annals of software engineering*, Vol. 10, No. 1-4, pp. 177–205, 2000.
- [3] International Software Benchmarking Standards Group (ISBSG). <http://www.isbsg.org>.
- [4] 井上克郎, 松本健一, 鶴保征城, 鳥居宏次. 実証的ソフトウェア工学環境への取り組み. 2004.
- [5] Barbara Kitchenham, Hiyam Al-Khilidar, Muhammed Ali Babar, Mike Berry, Karl Cox, Jacky Keung, Felicia Kurniawati, Mark Staples, He Zhang, and Liming Zhu. Evaluating guidelines for reporting empirical software engineering studies. *Empirical Software Engineering*, Vol. 13, No. 1, pp. 97–121, 2008.
- [6] Fabio QB Da Silva, Marcos Suassuna, A César C França, Alicia M Grubb, Tatiana B Gouveia, Cleviton VF Monteiro, and Igor Ebrahim dos Santos. Replication of empirical studies in software engineering research: a systematic mapping study. *Empirical Software Engineering*, Vol. 19, No. 3, pp. 501–557, 2014.
- [7] Forrest J Shull, Jeffrey C Carver, Sira Vegas, and Natalia Juristo. The role of replications in empirical software engineering. *Empirical Software Engineering*, Vol. 13, No. 2, pp. 211–218, 2008.
- [8] Masafumi Tsunoda, Akito Monden, Jacky Keung, and Kaname Matsumoto. Incorporating expert judgment into regression models of software effort estimation. In *Software Engineering Conference (APSEC), 2012 19th Asia-Pacific*, Vol. 1, pp. 374–379. IEEE, 2012.
- [9] Masateru Tsunoda, Sousuke Amasaki, and Akito Monden. Handling categorical variables in effort estimation. In *Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement*, pp. 99–102. ACM, 2012.

- [10] Luigi Lavazza, Sandro Morasca, and Gabriela Robiolo. Towards a simplified definition of Function Points. *Information and Software Technology*, Vol. 55, No. 10, pp. 1796–1809, 2013.
- [11] Cuauhtémoc López-Martín and Alain Abran. Neural networks for predicting the duration of new software projects. *Journal of Systems and Software*, Vol. 101, pp. 127–135, 2015.
- [12] Gavin R Finnie, Gerhard E Wittig, and Jean-Marc Desharnais. A comparison of software effort estimation techniques: using function points with neural networks, case-based reasoning and regression models. *Journal of Systems and Software*, Vol. 39, No. 3, pp. 281–289, 1997.
- [13] Barry W Boehm. *Software engineering economics*, Vol. 197. Prentice-hall Englewood Cliffs (NJ), 1981.
- [14] Tron Foss, Erik Stensrud, Barbara Kitchenham, and Ingunn Myrtveit. A simulation study of the model evaluation criterion MMRE. *IEEE Transactions on Software Engineering*, Vol. 29, No. 11, pp. 985–995, 2003.
- [15] IFPUG: Function Point Counting Practices Manual, Release 4.3. *International Function Point Users Group*, 2010.
- [16] 柏本隆志, 楠本真二, 井上克郎, 鈴木文音, 湯浦克彦, 津田道夫. 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.
- [17] Robert K Yin. *Case study research: Design and methods*. Sage publications, 2013.
- [18] 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, 2 2016. 修士学位論文.
- [19] Andrea De Lucia, Eugenio Pompella, and Silvio Stefanucci. Effort estimation for corrective software maintenance. In *Proceedings of the 14th international conference on Software engineering and knowledge engineering*, pp. 409–416. ACM, 2002.
- [20] Christophe Croux and Catherine Dehon. Influence functions of the Spearman and Kendall correlation measures. *Statistical methods & applications*, Vol. 19, No. 4, pp. 497–515, 2010.

- [21] Richard P Lippmann. An introduction to computing with neural nets. *ASSP Magazine, IEEE*, Vol. 4, No. 2, pp. 4–22, 1987.
- [22] Jung-Wook Park, Ronald G Harley, and Ganesh Kumar Venayagamoorthy. Adaptive-critic-based optimal neurocontrol for synchronous generators in a power system using MLP/RBF neural networks. *Industry Applications, IEEE Transactions on*, Vol. 39, No. 5, pp. 1529–1540, 2003.
- [23] IPA 独立行政法人 情報処理推進機構. <https://www.ipa.go.jp>.
- [24] Krishna R. Pryor D. Menzies, T. The Promise Repository of Empirical Software Engineering Data. <http://openscience.us/repo>, 2015.
- [25] R: The R Project for Statistical Computing. <https://www.r-project.org>.
- [26] 一言正之, 桜庭雅明, 小野寺勝. 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, No. 20, pp. 67–72, 2012.
- [27] Masayuki MURAKAMI. *Practicality of modeling systems using the IDS method: Performance investigation and hardware implementation*. PhD thesis, The University of Electro-Communications, 2008.