様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　　元 年　　６ 月 １０ 日現在

機関番号：３４４１９
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ００１１３

研究課題名（和文）ユーザを考慮したソフトウェア開発支援手法評価フレームワーク

研究課題名（英文）Evaluation framework considering users for software development support methods

研究代表者
角田　雅照（TSUNODA, Masateru）

近畿大学・理工学部・講師

研究者番号：６０４５７１４０

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：本研究では，定量的開発支援手法の使いやすさを向上させるために，ユーザを考慮した手法評価フレームワークを確立することをゴールとした．従来の評価では，主に以下の3つの点について十分に考慮されていない．(1) ユーザの予測精度に対する要求水準の評価方法，(2) 予測精度における誤差分散，(3) 予測モデル利用時における欠損値．これらを考慮することにより，定量的開発支援手法を適切に評価することができる．そこで，研究期間内にこれらの3点の解決に取り組んだ．また，定量的開発支援手法の普及に有用な方法について検討を行った．具体的には，ゲーミフィケーションや身体姿勢の効果について評価を行った．

研究成果の学術的意義や社会的意義
定量的開発支援手法に関して，各手法がユーザを十分に考慮して評価されていないために，実際にはユーザにとって問題となる点が改善されず，ユーザにとって使いにくい手法となっている可能性がある．これまで，定量的開発支援手法は予測精度，すなわち，予測値と実測値の差にのみ着目し，評価されることが主であった．ユーザを考慮して手法を評価するために必要な要素を明らかにし，それらを統合して定量的開発支援手法の評価することにより，よりユーザにとって使いやすい手法が明確となるとともに，各手法が評価され，改善されていく．これにより，ユーザにとって使いやすい手法が増加し，定量的開発支援手法の普及が促進すると期待される．

研究成果の概要（英文）：The goal of the study is to establish evaluation framework considering users for software development support methods based on measurements. Existing methods did not take into accounts the followings: (1) how to evaluate requirements of users for prediction accuracy, (2) influence of variance to evaluation of prediction accuracy, (3) influence of missing values when utilizing prediction models. If we consider the above points, we can evaluate development support methods properly. Therefore, we coped with the above points. In addition, we analyzed the effect of methods which are useful to spread development support methods. The methods are gamification and posing.

研究分野：ソフトウェア工学

キーワード：一対比較法　ゲーミフィケーション　欠損値

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

近年，ソフトウェアの市場規模は非常に大きくなっており，その重要性が増している．例えば，テレビなどの機器に搭載されるソフトウェアである，組込みソフトウェアの開発費は 2004 年度には 2 兆円であったが，2008 年度には 3 兆 5 千億円に増加した．さらに，同年には組み込みソフトウェアがインストールされた機器の生産額は製造業の生産額の 50%以上であり，GDP では 13%以上の比率となっている．

ソフトウェア開発プロジェクトは大規模化，多様化，短納期化が進んでおり，プロジェクトを QCD の面で成功に導く，すなわち，品質（Quality）の低下，納期（Delivery）の遅れ，コスト（Cost）の超過を避けることは必ずしも容易ではない．例えば，組込みソフトウェアでは品質確保が近年大きな課題のひとつとなっている．このような状況において，プロジェクトの成功を支援するために，これまで様々なツールや手法が数多く提案されてきた．我々はソフトウェア開発を支援するために，バグモジュール予測や開発コスト予測などの定量的開発支援手法に取り組んできた．

ただし，ソフトウェア開発を支援するツールや手法が開発現場において広く利用されているとはいえないという研究が，近年複数発表されている．その理由として，ツールや手法の適用にコストが掛かる，ツールや手法の存在をユーザが知らない，ツールや手法がユーザにとって使いにくいことなどが指摘されている．例えばソースコードの問題点を発見する，静的検査ツールの場合，問題点であるとツールが指摘している部分が，実際には問題点でないことが多い（false-positive が多い）ことが使いにくい理由として挙げられている．

これらの従来研究より，定量的開発支援手法に関しても，各手法がユーザを十分に考慮して評価されていないために，実際にはユーザにとって問題となる点が改善されず，ユーザにとって使いにくい手法となっている可能性があるといえる．すなわち，手法の評価が不十分であることが，手法普及の障害のひとつになっているとの着想に至った．これまで，定量的開発支援手法は予測精度, すなわち, 予測値と実測値の差にのみ着目し, 評価されることが主であった．ただし, 例えば平均的には 2 つの予測精度が同じでも， 数回に 1 回大きく予測が外れる手法のほうが，ユーザにとって使いにくい可能性がある．我々の調査においても，ユーザは手法の安定性を重視することが示唆されている．

ユーザを考慮して手法を評価するために必要な要素を明らかにし，それらを統合して定量的開発支援手法の評価フレームワークを確立することにより，よりユーザにとって使いやすい手法が明確となるとともに，各手法がフレームワークに従って評価され，改善されていく．これにより，ユーザにとって使いやすい手法が増加し，定量的開発支援手法の普及が促進すると期待される．

２．研究の目的

本研究では，定量的開発支援手法の使いやすさを向上させるために，ユーザを考慮した手法評価フレームワークを確立することである．従来の評価では，主に以下の 3 つの点について十分に考慮されていない. そこで, 研究期間内にこれらの 3 点をどう考慮すべきか明らかにした.

- ユーザの予測精度に対する要求水準
- 予測精度における誤差分散
- 予測モデル利用時における外れ値

これらを考慮することにより，定量的開発支援手法を適切に評価することができる．また，定量的開発支援手法の普及に有用な方法について検討を行った．外れ値については研究開始直前にある程度検討できたため，欠損値についても検討を行った．

３．研究の方法

ユーザを考慮した評価方法を確立するため，ユーザに対するアンケート結果を中心に分析を進める．

(1) ユーザの予測精度に対する要求水準の明確化

定量的開発支援手法の予測精度に対して，ユーザが求める性質とその水準を明らかにする．そのために，ユーザが求める性質の候補とその水準を変化させた質問票を複数用意し，開発者に回答してもらい，そのデータを分析する．

(2) 誤差分散を考慮した，予測精度評価指標の確立

定量的開発支援手法の予測精度評価のために，誤差平均と誤差分散の両方を考慮した，新たな評価指標を作成するとともに，その妥当性をユーザによるアンケートに基づき確かめる．

(3) 現在データにおける外れ値の影響を考慮したモデル評価

予測モデルに入力する現在データに，外れ値を実験的に追加して予測を行い，現在データの外れ値に対してロバストな予測方法を明らかにする．

４．研究成果

(1) イロレーティングに基づく一対比較法の改善

表 1　各順位付け方法に対する評価

順位付け方法	レーティング	比較回数	勝	負
平均順位	2304	40	25	15
10 段階	2126	40	16	24
イロレーティング	2190	40	21	19
平均評価	2155	40	18	22

表 2　各見積もり方法の誤差平均と分散

見積方法	*BRE* 平均値	*BRE* 分散
A	25%	6%
B	25%	16%
C	32%	7%
D	50%	7%
E	50%	16%

この研究成果では授業間比較を対象としているが，提案方法はユーザによる予測モデル評価の比較にも容易に適用可能である．授業間比較の妥当性を高めるために，本研究では従来の研究と同様に，一対比較法に着目する．一対比較法では，2 つの評価対象のうち，どちらが良いかにより評価を行う．ただし，評価対象の組み合わせが膨大な数になるため，回答者の負担が課題となる．

そこで本研究では，回答者の負担を減らしつつ，授業間比較の妥当性を高めるために，全組み合わせの一部の一対比較結果から，イロレーティングに基づいて授業の順位付けを行う方法を提案する．

イロレーティングはチェスなどの一対一の対戦ゲームの勝敗を基に，プレイヤーの実力を順位付けする方法である．イロレーティングでは，対戦が総当たりではなく偏りがあることを前提としており，実力に大きな差がある場合の勝利と，そうでない場合について同等に扱わない．例えば，ランキング上位のプレイヤーに勝利した場合，その勝利を重視し，逆に下位のプレイヤーに対する勝利は重視しない．勝敗結果に基づいてレーティング（スコア）が算出され，レーティング順に順位付けされる（値が大きいほど優れていることを示す）．

本研究では，以下の 3 種類の方法により授業の順位付けを行い，妥当性を評価する．

- イロレーティングに基づく順位付け: 授業の一対比較を n 段階評価の評価回数と同程度の回数行った後，評価に基づきイロレーティングを適用し，各授業を順位付けする．
- 平均順位に基づく順位付け: イロレーティングによる授業の順位と，n 段階評価（n は自然数）による授業の順位の平均値を算出し，それを順位とする．例えば授業 A において，イロレーティングによる順位が 2 位，5 段階評価による順位が 4 位とすると，(2 + 4) / 2 = 3 位を順位とする．
- 平均評価に基づく順位付け: ある授業に対するレーティングの値と，n 段階評価の値の平均値を算出し，それに基づき順位付けする．平均値の計算時には，それぞれの値域を [0, 1] にする．ここでは値域変換時に広く用いられる方法（最大値と最小値に基づく）を用いた．

以下の手順によりデータを収集した．

(1)　被験者は一対比較法により授業を評価する．
(2)　被験者は 10 段階評価により授業を評価する．
(3)　手順 1，2 の結果に基づき，レーティングを算出する．
(4)　手順 3 の結果に基づいて順位表を作成する．
(5)　被験者は手順 4 で作成された順位表について，一対比較法により評価する．
(6)　手順 5 の結果に基づき，順位表のレーティングを算出する．

被験者による，4 種類の順位付け方法の納得度の評価を表 1 に示す．表において，「勝」は一対比較において他方より優れていると評価された回数，「負」は逆に劣っていると評価された回数である．

平均順位に基づく順位付けの評価が最も高く，その他の順位付け方法よりもレーティングが 100 以上離れていた．その他の方法では，レーティングに大きな差がなったが，10 段階評価に基づく順位付けの評価が最も低く，イロレーティングに基づく順位付けが，平均順位に基づく順位付けの次に評価が高かった．

これらの結果より，10 段階評価とイロレーティングの平均順位を用いて授業を順位付けすることにより，最も納得度（妥当性）が高くなるといえる．この場合，従来の 10 段階評価と比較すると，回答者負担は 2 倍（実験では最大で 108 回）となるが，単純な一対比較法（1431 回．3.2 節参照）と比べると負担は小さい．イロレーティングの適用時は全組合せの一対比較法と比べて回答数が少なく，一部の授業に対する評価に偏りが生じる可能性があるが，10 段階評価の結果により補正ができ，妥当性が高まった可能性がある．

(2) 工数見積における誤差分散がユーザ評価に与える影響

本研究では，実験において被験者に，ある見積もり手法による見積もり工数と実際の工数を複数提示し，その手法が好ましいかどうかを評価してもらった．見積もり手法として，表 2 に示す 5 種類を用意した．これらは重回帰分析などの実際の見積もり方法に基づくものではなく，見積もり誤差の平均値と分散が一定となるようにランダムで値を作成したものである．例えば見積もり手法 A により見積もり結果は，*BRE* 平均値を 25%，分散が 6%となるように，工数の

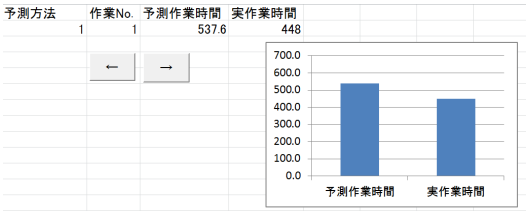

図 1　見積もり結果の提示

表 3　イロレーティングによる評価

見積方法	レーティング	比較回数	選択回数	非選択回数
A	2381	33	27	6
B	2298	34	24	10
C	2153	30	13	17
D	2111	24	7	17
E	2057	39	9	30

実測値と見積もり値をランダムで生成した．なお，見積もり手法 C は，実験ツールに不具合があったため，被験者によりわずかに *BRE* 平均と分散が異なっている．

被験者は情報科学を専攻する大学の 4 年生 16 人である．各手法では 20 個の見積もり結果を図 1 の形式で提示した．ユーザは GUI のボタンを操作することにより，それぞれの見積もり結果を閲覧できる．また，見積もり値と実測値は棒グラフにより視覚化して提示した．

どの手法が好ましいかを比較する方法として，5 つの見積もり方法全てを比較し，それぞれの優劣を評価することが最も一般的である．ただし表 2 に示すように，全ての手法間に非常に大きな見積もり誤差があるわけではいため，全てを比較して評価することは容易ではない．そこで一対比較法を用いた．

一対比較法の結果を単純に順位付けに用いるとすると，優劣の結果を集計することになる．ただしこの場合，見積もり精度に大きな差がある場合と，見積もり精度にあまり差がない場合の優劣について，どちらも同様に評価するため，必ずしも正確に評価できないという問題がある．この問題を解決するために，イロレーティングを用いる．

イロレーティングによる評価を表 3 に示す．被験者が不足しているため，見積もり方法 D については比較回数が少なくなっている．見積もり方法 A と B については，誤差平均は同じだが分散の小さい A のほうが，被験者からの評価が高かった．このことから，工数見積もり手法の評価において，分散は無視できないと考えられる．

見積もり方法 B と C については，誤差平均が小さく誤差分散が大きい B のほうが，誤差分散の小さい C よりも被験者からの評価が高かった．　　B と C の誤差平均の差は大きくないが，被験者はその違いを重視するとともに，誤差分散の大きさについてはあまり重視しなかったといえる．このことから，誤差分散の小ささよりも，誤差平均の小ささのほうが重視される可能性が高いといえる．

見積もり方法 D と E については，誤差平均は同じだが分散の大きい E のほうが，被験者からの評価が高かった．なお，AE の平均値と最大値，実測値の平均値，*BRE* の最大値は D と E でほとんど差がない，もしくは E のほうが大きかった，評価結果が予想と逆となった理由として，E の評価回数が少ないことが影響している事が考えられる．その他の理由として，誤差平均の大きさと比較して分散の大きさが小さい場合，分散の違いに被験者が気づきにくかった可能性もある．

まとめると，以下の傾向があるといえる．

- 誤差平均が 2 つの見積もり方法で同様に小さい場合，分散の小さい方法の評価が高かった．
- 誤差平均が小さく分散が大きい見積もり方法のほうが，誤差平均が大きく分散の小さい方法よりも評価が高かった．
- 誤差平均が 2 つの見積もり方法で同様で，誤差分散が誤差平均と比較して小さい場合，誤差分散の差は考慮されない可能性がある．

(3) 重回帰分析を用いた工数予測における欠損値補完手法の性能比較

重回帰モデルの構築にあたっては，欠損値を含まないデータセットが必要となるが，一般に，多数の部署・組織から収集されたプロジェクトデータには欠損値が含まれる．このため，何らかの手法で欠損を補完し，欠損の無いデータを作成することが必須となる．

ひとつの方法として，欠損値を何らかの値で補完(欠損値補完法)の実施により，欠損値が無いままにデータセットのサイズを保つことある．ただし，欠損値補完法によって適切な値を補完できない場合，それはデータセットにとってのノイズとなり，妥当なモデルが得られなくなる可能性がある．

本研究では，1364 件のプロジェクト（欠損率 34%）から，欠損を含む 1204 件のプロジェクトを選び，4 つの欠損値補完法（k-nn 法，CF 応用法，Miss Forest 法，多重代入法）を適用し，重回帰モデルの構築を行う．そして欠損を含まない 160 件のプロジェクトを用いて各モデルの予測精度を評価することで，欠損値補完法の性能を実験的に比較する．

評価実験は次の手順で行った．

1. フィットデータに対し，k-nn 法，CF 応用法，多重代入法，Miss Forest 法によって欠損値補完を行う．
2. それぞれの手法で欠損値を補完したフィットデータに対して，ステップワイズ重回帰分析

表 4　欠損値補完法適用時の予測誤差中央値

手法	MRE	MER
k-nn 法	0.364	0.375
CF 応用法	0.405	0.656
Miss Forest 法	0.499	0.540
多重代入法	0.326	0.354

を行い，総開発工数を目的変数としたモデルを構築する．

3. 構築したモデルを用いてテストデータの総開発工数を予測し，各評価基準の値を算出する（テストデータの総開発工数は未知数とみなす）．

ただし多重代入法，および Miss Forest 法はランダム性を含むため，上記の作業を 10 回繰り返した．

各欠損値補完法によってデータセット中の欠損値の補完を行い，ステップワイズ重回帰分析で予測を行った．予測誤差 MRE，MER を表 4 に示す． MRE，MER ともに多重代入法で欠損値補完を行った場合に，最も高い精度で予測できることがわかった．ただし MER において多重代入法の IQR が k-nn に比べてやや大きいことから，予測結果のばらつきを抑えたい場合には k-nn の採用も考えられる．

５．主な発表論文等

〔雑誌論文〕（計 5 件）

1. 村上優佳紗，角田雅照，“ ソフトウェア開発における開発者のリスク認識の分析 ”，コンピュータソフトウェア，vol.35，no.4，pp.37-43，2018 査読有．
2. 加納豊之，角田雅照，“ イロレーティングに基づく授業間比較の試み ”，電子情報通信学会論文誌 D，vol.J101-D，no.6，pp.989-993，2018 査読有．
3. 山田悠斗，江川翔太，まつ本真佑，角田雅照，楠本真二，“ 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 ”, SEC journal, vol.13, no.3, pp.10-17, 2017 査読有．
4. 戸田航史, 角田雅照,“ 重回帰分析を用いた工数予測における欠損値補完手法の性能比較 ”, コンピュータソフトウェア，vol.34，no.4，pp.150-155，2017 査読有．
5. 角田雅照，天嵜 聡介，“ ソフトウェア開発プロジェクトの生産性分析に対する傾向スコアの適用 ”，情報処理学会論文誌，vol.58，no.4，pp.855-860，2017 査読有．

〔学会発表〕（計 13 件）

1. U. Yukizawa, M. Tsunoda, and A.Tahir, “Please Help! A Preliminary Study on the Effect of Social Proof and Legitimization of Paltry Contributions in Donations to OSS,” Proc. of International Conference on Software Analysis, Evolution and Reengineering (SANER), pp.609-613, 2019.
2. M. Tsunoda, and H. Yumoto, “Applying Gamification and Posing to Software Development,” Proc. of Asia-Pacific Software Engineering Conference (APSEC), 2018.
3. M. Tsunoda, T. Hayashi, S. Sasaki, K. Yoshigami, H. Uwano, and K. Matsumoto, “How Do Gamification Rules and Personal Preferences Affect Coding?” Proc. of International Workshop on Empirical Software Engineering in Practice (IWESEP), pp.13-18, 2018.
4. K. Nakasai, M. Tsunoda, H. Hata, and K. Matsumoto, “Identifying Spoofing Accounts on Twitter Based on Relationships of Accounts,” Proc. of International Conference on Big Data, Cloud Computing, and Data Science Engineering (BCD), pp.85-90, 2018.
5. M. Tsunoda, K. Matsumoto, S. Ohiwa, and T. Oshino, “Analyzing Software Maintenance Cost Based on Work Efficiency and Unit Cost,” Proc. of International Conference on Big Data, Cloud Computing, and Data Science Engineering (BCD), pp.102-108, 2018.
6. S. Yamashita, M. Tsunoda, and T. Yokogawa, “Visual Programming Language for model-checkers Based on Google Blockly,” Proc. of International Conference on Product-Focused Software Process Improvement (Profes), pp.597-601, 2017.
7. M. Tsunoda, and S. Amasaki, “On Software Productivity Analysis with Propensity Score Matching,” Proc. of International Symposium on Empirical Software Engineering and Measurement (ESEM), pp.436-441, 2017.
8. Y. Murakami, M. Tsunoda, and H. Uwano, “WAP: Does Reviewer Age Affect Code Review Performance?”' Proc. of International Symposium on Software Reliability Engineering (ISSRE), pp.164-169, 2017.
9. Y. Murakami, and M. Tsunoda, “Is Cutting-Edge Software Engineering Attractive for Developers in SMEs?” Proc. of International Conference on Big Data, Cloud Computing, and Data Science (BCD), pp.332-338, 2017.

10. T. Kakimoto, M. Tsunoda, and A. Monden, “Should Duration and Team Size Be Used for Effort Estimation?” Proc. of International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2017.
11. K. Yoshigami, M. Tsunoda, Y. Yamada, and S. Kusumoto, “Should Function Point Elements be Used to Build Prediction Models?” Proc. of International Workshop on Empirical Software Engineering in Practice (IWESEP), pp.41-46, 2017.
12. Y. Murakami, M Tsunoda, and K. Toda, “An Empirical Evaluation of the Tobit Model on Software Defect Prediction,” Proc. of Applied Computing and Information Technology (ACIT), pp.196-201, 2016.
13. **長濱優樹，角田雅照,“ 工数見積における誤差分散がユーザ評価に与える影響 ”，情報処理学会関西支部 支部大会 講演論文集，**D-03，2016.

〔図書〕（計 1 **件）**

1. T. Kakimoto, M. Tsunoda, and A. Monden, ``Should Duration and Team Size Be Used for Effort Estimation?'' In Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (Studies in Computational Intelligence, Vol. 721), pp.91-105, 2017.

〔産業財産権〕

○出願状況（計 0 **件）**

○取得状況（計 0 **件）**

〔その他〕

ホームページ等

http://www.info.kindai.ac.jp/~tsunoda/publication.php