工数予測における類似性に基づく欠損値補完法の実験的評価

田村 晃一　柿元 健　戸田 航史　角田 雅照　門田 暁人
松本 健一　大杉 直樹

ソフトウェア開発において，開発中もしくは将来のプロジェクトの計画立案や管理を目的として，重回帰モデルを利用した開発工数の予測が行われている．一般に，モデル構築に使用するプロジェクトデータには未記録の値 (欠損値) が存在するため，モデル構築を行う前に，欠損値を何らかの値で補完する (欠損値補完法)，または，欠損値を含むメトリクスやプロジェクトを削除することで欠損値のないデータセットを作成すること (無欠損データ作成法) が必要となる．ただし，いずれの手法がプロジェクトデータに適しているかは従来明らかにされていない．本論文では，複数の企業で収集された 706 件 (欠損率 47%) のプロジェクトデータに対し，4 つの欠損値補完法 (平均値挿入法，ペアワイズ除去法，k-nn 法，CF 応用法) 及び，無欠損データ作成法を適用し，重回帰モデルの構築を行った．各手法の予測精度を評価するために欠損のない 143 件のプロジェクトの工数予測を行った結果，類似性に基づく欠損値補完手法 (k-nn 法，CF 応用法) を用いる場合に高い精度のモデルが構築されることがわかった．

Multivariate regression models have been commonly used to estimate the software development effort to assist project planning and/or management. Since project data sets for model construction often contain missing values, we need to build a complete data set that has no missing values either by using imputation methods or by removing projects and metrics having missing values (removing method). However, while there are several ways to build the complete data set, it is unclear which method is the most suitable for the project data set. In this paper, using project data of 706 cases (47% missing value rate) collected from several companies, we applied four imputation methods (mean imputation, pair-wise deletion, k-nn method and applied CF method) and the removing method to build regression models. Then, using project data of 143 cases (having no missing values), we evaluated the estimation performance of models after applying each imputation and removing method. The result showed that the similarity-based imputation methods (k-nn method and applied CF method) showed the best performance.

1　はじめに

ソフトウェア開発プロジェクトの初期段階において開発工数を予測すること，及び各工程の終了段階で開発工数の再予測を行うことは，プロジェクト完遂に必要な資源の確保や，スケジュール管理を行う上で重要である．そのために，過去のソフトウェア開発プロジェクトの実績データ (以下，プロジェクトデータ) を予測の根拠に用いる定量的予測手法が数多く提案され，用いられてきた[1][17][18][21]．なかでも，プ

Empirical Evaluation of Similarity-Based Missing Data Imputation for Effort Estimation.

Koichi Tamura, Koji Toda, Masateru Tsunoda, Akito Monden, Ken-ichi Matsumoto, 奈良先端科学技術大学院大学情報科学研究科, Graduate School of Information Science, Nara Institute of Science and Technology.

Takeshi Kakimoto, 奈良先端科学技術大学院大学情報科学研究科, Graduate School of Information Science, Nara Institute of Science and Technology. 現在, 大阪大学大学院情報科学研究科, Graduate School of Information Science and Technology, Osaka University.

Naoki Ohsugi, 株式会社 NTT データ, NTT Data Corporation.

コンピュータソフトウェア, Vol.26, No.3 (2009), pp.44–55.
[研究論文] 2007 年 2 月 29 日受付.

本論文は，第 14 回ソフトウェア工学の基礎ワークショップ (FOSE 2007) の発表論文をもとに発展させたものである．

ロジェクトのメトリクス (開発規模，欠陥数など) を説明変数として用い，目的変数である開発工数との関係を一次式で表現する重回帰モデルを用いた予測は，ツールが普及しており適用が容易なことから，広く実施されてきた [2] [22].

重回帰モデルの構築にあたっては，欠損値を含まないプロジェクトデータが必要となるが，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 [10] [13] [21]．欠損値が生じる原因としては，収集メトリクスが異なる複数の組織のデータをマージしたことや，時間的制約や不注意による記録漏れなどが挙げられる．

そのため，重回帰モデルの構築にあたっては，何らかの方法で欠損値の含まないデータセットを用意することが必須となる．その 1 つの方法は，欠損値を含むメトリクスやプロジェクトを除去し，欠損値のないデータセットを作成すること (無欠損データ作成法) である [16] [19]．ただし，データセットのサイズが小さくなり，予測の根拠となる情報量を減らしてしまうため，モデルを構築したとしても十分な予測精度が得られない可能性がある．もう 1 つの方法は，欠損値を何らかの値で補完すること (欠損値補完法) により，データセットのサイズを保つことである [2] [9] [13] [14] [16] [18]．この場合の問題は，補完する値によってはデータセットにノイズが混じることとなり，必ずしも妥当なモデルが得られないことである．他に，重回帰モデルに特化した欠損値除去法として，ペアワイズ除去法が提案されている．ただし，いずれの手法が工数予測モデルに適しているのかは，必ずしも明らかでない．

本論文では，International Software Benchmarking Standards Group (ISBSG) [6] が収集した，複数の企業で収集されたプロジェクトデータから欠損を含む 706 件のプロジェクト (データ欠損率 47%) を選び，4 つの欠損値補完法 (平均値挿入法，ペアワイズ除去法，k-nn 法，CF 応用法) 及び，無欠損データ作成法を適用し，重回帰モデルの構築をそれぞれ行う．そして，ISBSG データセットから選んだ欠損のない別の 143 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，欠損値補完法および無欠損データ作成法の性能を実験的に比較する．ここで，CF 応用法とは，協調フィルタリングに基づく予測手法 [21] を応用した手法であり，欠損値補完への適用は本論文が初めてである．CF 応用法と k-nn 法はいずれもプロジェクト間の類似性に基づく欠損値補完法といえるが，類似度計算を求めるアルゴリズムが異なっている．各々のアルゴリズムの詳細は 3 章で述べる．

従来，欠損値補完・除去法の比較は，全く行われていないわけではないが，いずれも不十分である．Jonsson ら [9] や Strike ら [19] は，欠損の無いデータセットを人為的に (ランダムに) 欠損させ，欠損値補完法の比較を行っている．しかし，現実のデータ欠損はランダムというよりはむしろバースト的であるため，現実のデータセットを用いても同等の効果が得られるかは明らかでない．バースト的な欠損となる理由はいくつか存在する．企業ごとに計測対象のメトリクスが異なる場合，それらのデータを結合した際にバースト的な欠損が発生する．同一の企業であっても，時期によって計測対象のメトリクスが異なっている場合があり，バースト的な欠損が発生する原因となる．

また，Cartwright ら [3] や Myrtveit ら [16] は，欠損を含む現実のデータに対する欠損値補完・除去を行ってから重回帰モデルを構築し，データに対するモデルの適合度を評価している．しかし，適合度のみの評価では，実際の予測を行うにあたって有用であるかは不明である．特に，モデル構築用のデータセットに対して予測モデルが過剰に適合 (オーバーフィッティング [5]) した場合，モデル構築用データセット以外のデータセットに対して予測を行った際に高い予測精度が得られない可能性がある．そのため，適合度の評価に加えて，予測性能の評価を行うことが求められる．構築したモデルの予測性能を評価するためには，欠損のないデータセットが別途必要であり，この点についての評価は行われていない．

さらに，いずれの従来研究においても，欠損値補完法と無欠損データ作成法との比較は行われていない．工数予測モデルを構築する者にとって，多少ノイズが入ってでも情報量を減らさない (欠損値補完法) のが良いのか，情報量を減らしてでもノイズを避ける (無欠損データ作成法) のが良いのかを知ることは重要で

ある．

以降，2 章では，本論文の実験で用いた工数予測手法であるステップワイズ重回帰分析について述べ，3 章では，従来の欠損値補完・除去法，及び本論文で新たに欠損値補完法として適用する CF 応用法について述べる．4 章では，各手法の精度を評価するための実験について説明し，5 章で評価実験の結果と考察について述べる．6 章で関連研究について述べ，最後に 7 章で本論文の結論について述べる．

2 ステップワイズ重回帰分析による工数予測

重回帰分析は多変量解析の一手法であり，ソフトウェア開発に要する工数を予測するために広く用いられている．本論文では，工数予測手法として重回帰分析の一手法であるステップワイズ重回帰分析を用いた．

重回帰分析では，予測対象の変数 $\hat{Y}$(目的変数) と目的変数に影響を与える複数の変数 N(説明変数) との関係を表した一次式 (重回帰モデル) を作成する．ここで，変数の数を k，偏回帰係数を a，定数項を C とすると重回帰分析のモデル式は式 (1) で定義される．

$$\hat{Y} = a_1 N_1 + a_2 N_2 + \cdots + a_k N_k + C \qquad (1)$$

式 (1) の a と C は，残差平方和が最小となるように決定される．作成された回帰式に，予測対象のケース (プロジェクト) の説明変数の値を与えることで，目的変数の値を予測することが可能となる．

重回帰分析では，予測精度を向上させるために，多数の説明変数候補の中から，予測精度の向上に寄与すると考えられる変数を選択して回帰式を作成する方法がとられる．ステップワイズ重回帰分析は，ステップワイズ変数選択法により採用する変数を決定し，重回帰分析を行う手法である．ステップワイズ変数選択は次の手順で行われる [20].

手順 1． 変数を全く含まないモデルを初期モデルとして作成する．

手順 2． 作成されたモデルに対して，各説明変数の係数が 0 でないかの検定を行い，指定した有意水準 (本論文の評価実験では，偏 F 値の有意水準を $p_{in} = 0.05$，$p_{out} = 0.1$ とした) で棄却されない場合に変数を採択する．ただし，多重共線性を回避するために，採択する変数の分散拡大要因 (VIF) が一定値以上の場合，またはその変数を採択することによって，他の変数の VIF が一定値以上となる場合，その変数は採択しない．本論文の評価実験では，VIF は 10 とした．

手順 3． 検定により適切な変数が選択されたと判断されるまで手順 2. を繰り返す．

3 欠損値補完・除去法

欠損値を含むデータセットに対してステップワイズ重回帰分析を適用する前処理として，データセットに含まれる欠損値を何らかの値で補完する，もしくは欠損値を含むプロジェクト及びメトリクスを除外することが行われる．本論文では，欠損値補完法として従来よく使われている平均値挿入法，ペアワイズ除去法，k-nn 法に加えて，今回新たに欠損値補完に適用する協調フィルタリングを応用した補完法 (CF 応用法) を用いて欠損値補完を行う．また，過去のプロジェクトのデータから欠損値を多く含むメトリクスやプロジェクトを除去し，無欠損のデータセットを作成する手法 (無欠損データ作成法) も適用する．各欠損値補完法，及び無欠損データ作成法の詳細を以下に示す．

3.1 平均値挿入法

欠損値に対して，当該メトリクスの平均値を挿入することで，欠損値を補完する [19].

3.2 ペアワイズ除去法

重回帰分析に特化した手法で，重回帰分析の過程においてメトリクス間の共分散を求める際に，メトリクスのいずれかが欠損しているプロジェクトを除外して共分散を求める手法である [13].

3.3 類似性に基づく欠損値補完法 (k-nn 法)

類似性に基づく欠損値補完法 (k-nn 法) は欠損値に対して，類似したプロジェクトのメトリクス値を用いて欠損値を補完する [3]．k-nn 法は 3 つの手順 (正規化，類似度計算，補完値計算) から構成される．各手順の詳細とアルゴリズムについては以下で述べる．

手順 1．メトリクス値の正規化 各メトリクスは値域に大きなばらつきがあるため，値域をそろえるための正規化を行い，値域を [0,1] にする．ここで，p_i は i 番目のプロジェクト，m_j は j 番目のメトリクスと定義すると，プロジェクト p_i のメトリクス m_j の値 v_{ij} を正規化した値 v'_{ij} は式 (2) で定義される．

$$v'_{ij} = \frac{v_{ij} - \min(P_j)}{\max(P_j) - \min(P_j)} \qquad (2)$$

ここで，P_j はメトリクス m_j が計測されているプロジェクトの集合，$\max(P_j)$ と $\min(P_j)$ はそれぞれ $\{v_{x,j}|p_x \in P_j\}$ の最大値，最小値を表す．

手順 2．プロジェクト間の類似度計算 メトリクス値を補完するプロジェクトと類似したプロジェクトを見つけるため，プロジェクト間のユークリッド距離を計算し，それを類似度とする．メトリクス値を補完するプロジェクト p_a と他の各プロジェクト p_i とのユークリッド距離による類似度 $E(p_a, p_i)$ は式 (3) で定義される．

$$E(p_a, p_i) = \sqrt{\sum_{j \in M_a \cap M_i} \left(v'_{aj} - v'_{ij}\right)^2} \qquad (3)$$

ここで，M_a と M_i はそれぞれプロジェクト p_a と p_i に関して記録されている (欠損していない) メトリクスの集合を表している．

手順 3．類似度に基づく補完値の算出 補完値の算出には，補完対象のメトリクスが欠損していない，プロジェクト p_a と類似度の高い上位 k 個のプロジェクト (*k-nearestProjects*) を用いる．プロジェクト p_a のメトリクス m_b の値 v_{ab} を補完対象とすると，*k-nearestProjects* のメトリクス m_b の平均値を補完値とする．k は実験的に別途求める必要がある．

3.4 類似性に基づく欠損値補完法 (CF 応用法)

本論文では，新たに協調フィルタリングを応用した工数予測手法[21] を欠損値補完に適用する．CF 応用法も，メトリクス値が類似したプロジェクトから過去のプロジェクトの欠損値を補完する．

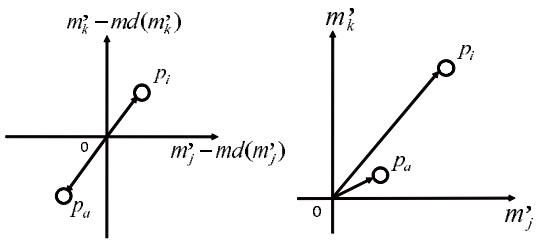

図 1 プロジェクト間の類似度計算

CF 応用法は k-nn 法と同様に，3 つの手順 (正規化，類似度計算，補完値計算) から構成される．各手順の詳細とアルゴリズムについては以下で述べる．

手順 1．メトリクス値の正規化 前節で述べた，手順 1．メトリクス値の正規化と同様の方法で正規化を行う．

手順 2．プロジェクト間の類似度計算 メトリクス値を補完するプロジェクトと類似した他のプロジェクトを見つけるため，プロジェクト間の類似度を算出する．メトリクス値を補完するプロジェクト p_a と他の各プロジェクト p_i との類似度 $sim(p_a, p_i)$ は式 (4) で定義される．

$$sim(p_a, p_i) = \frac{\sum_{j \in M_a \cap M_i} \left(v'_{aj} - md(m'_j)\right)\left(v'_{ij} - md(m'_j)\right)}{\sqrt{\sum_{j \in M_a \cap M_i} \left(v'_{aj} - md(m'_j)\right)^2} \sqrt{\sum_{j \in M_a \cap M_i} \left(v'_{ij} - md(m'_j)\right)^2}} \qquad (4)$$

ここで，m'_j はメトリクス m_j を正規化したものであり，$md(m'_j)$ は j 番目のメトリクスの正規化した値の中央値を表す．

v'_{ij} から $md(m'_j)$ を減算することで，中央値よりも大きなメトリクス値は正の値をとり，小さい値は負の値をとるようになる．類似度の計算例を図 1 に示す．図 1 の左側の図は正規化したメトリクス値から中央値を減算して類似度を計算した場合，右側の図は正規化したメトリクス値をそのまま使って類似度を計算した場合を示している．この計算により，類似度 $sim(p_a, p_i)$ の値域が [-1,1] をとるようになり，大きくメトリクス値が離れたプロジェクト間の類似度が小さくなる．

手順 3. 類似度に基づく補完値の算出 補完対象となる欠損値について，その補完値の算出に類似したプロジェクトの対応するメトリクスの実測値を用いる．手順 2. の類似度計算では，(ベクトルのなす角を用い，ベクトルの大きさを用いないため) 規模が異なるが傾向が似ているプロジェクト同士は類似度が高いとみなしている．そこで，補完値の算出において，類似度 $sim(p_a, p_i)$ を重みとして，プロジェクト p_a と類似したプロジェクトのメトリクス値 v_{ib} に，プロジェクトの規模を補正する $amp(p_a, p_i)$ を乗じた値で加重平均を行う．プロジェクト p_a のメトリクス m_b の補完値 $\hat{v}_{ab}$ は式 (5) で定義される．

$$\hat{v}_{ab} = \frac{\sum_{i \in k-nearestProjects} (v_{ib} \times amp(p_a, p_i) \times sim(p_a, p_i))}{\sum_{i \in k-nearestProjects} sim(p_a, p_i)} \quad (5)$$

ここで，*k-nearestProjects* は，メトリクス m_b が欠損しておらず，かつ，プロジェクト p_a と類似度の高い上位 k 個のプロジェクトの集合を表す．k の値は実験的に別途求める必要がある．

また，$amp(p_a, p_i)$ は，プロジェクト p_a のファンクションポイント (FP) を f_a と定義すると，式 (6) で定義される．

$$amp(p_a, p_i) = \frac{f_a}{f_i} \quad (6)$$

$amp(p_a, p_i)$ は，プロジェクト p_a の規模が p_i の規模のおよそ何倍になっているかを，正規化された FP により求めている．この $amp(p_a, p_i)$ により，多様な規模のプロジェクトを含むデータセットを用いた場合にも補完が可能となる．

3.5 無欠損データ作成法

無欠損データ作成法では，欠損値を含むデータセットに対して重回帰分析を行うために，欠損値を含むプロジェクトやメトリクスを除去する [16] [19]．従来研究では欠損を 1 つでも含むプロジェクトを全て除去する手法 (リストワイズ除去法) がよく用いられている．しかし，欠損値が多くのプロジェクトにおいて存在している場合，リストワイズ除去法を適用するとモデル構築に使用できるプロジェクトがほとんど残らない場合がある．そこで，無欠損データ作成法では，予め高い割合で欠損値を含んでいるメトリクスを除去してから，欠損を 1 つでも含むプロジェクトを除去することで欠損を含まないデータセットを作成する．ただし，多くのプロジェクトを残そうとすればメトリクスの数が少なくなり，逆に，多くのメトリクスを残そうとすればプロジェクトの数が少なくなるため，それぞれの除去のバランスには注意が必要である．

4 評価実験

4.1 実験概要

本論文では，各欠損値補完・除去法の有効性をソフトウェア開発企業で収集された実績データを用いて実験的に評価する．実験では，前章で述べた各欠損値補完・除去法によって欠損値を補完もしくは除去したデータセットを用いてプロジェクトの総開発工数を予測する．それぞれの欠損値補完・除去法を適用してから予測モデルを構築した際の予測精度を比較することにより評価を行う．なお，本論文ではもともと欠損の含まれるデータセットを用いるために，欠損値の真の値は不明である．そのため，補完値そのものの正確さの評価 (真の値との比較) をすることはできない．そこで，欠損値補完・除去を行ったデータセットを用いて工数予測モデルを構築し，その工数予測モデルの予測精度を (欠損のないデータセットを用いて) 評価することで，間接的に欠損値補完の精度を評価する．

4.2 実験用データセット

実験で利用したデータセットは，ISBSG (International Software Benchmarking Standards Group) が収集した，20ヶ国のソフトウェア開発企業の実績データ [6] から抽出した．ISBSG データセットは，一般公開されている実績データであり，多くの欠損値を含んでいる (欠損率 58%)．さらに，ISBSG データセットは企業横断データセットとして工数予測手法の評価に広く使用されている [12]．これらの特徴により，本論

文の目的である欠損値補完の効果を検証するのに適したデータセットであると考えられる．

本実験では，設計終了時を予測時期と想定し，総開発工数の予測を行った．ISBSG データセットから，FP 計測手法が IFPUG，開発形態が新規開発，データの質 (Data quality Rating) が A もしくは B(A〜D の全 4 段階中上位 2 段階)，及び総開発工数 (目的変数) が欠損していない 849 件 (欠損率 39%) のプロジェクトを実験に用いた．ISBSG は複数の企業から収集したデータセットであるため，データの質に大きなばらつきが存在している．そこで，データの質が欠損値補完及び工数予測の精度に大きな影響を与えることを防ぐため，文献[8][15] にならってデータセット抽出を行った．

実験に使用したメトリクスの詳細を表 1 に示す．データセットに含まれるメトリクスのうち総開発工数を目的変数とし，4 個のメトリクスを説明変数として用いている．本論文で欠損値補完に適用する平均値挿入法，k-nn 法，CF 応用法は数値変数の補完が前提となっており，数値変数の補完に焦点を当てた評価を行うために，開発プラットフォーム，言語タイプ，ビジネス領域タイプなどのカテゴリ変数については採用しなかった．開発期間は実測の値が記録されているため，説明変数として選択するか否かは，その因果関係から議論が分かれる．しかし，(1) 開発期間の計画値が ISBSG データセットに含まれていないこと，(2) 一般に，(失敗プロジェクトを除いて) 開発期間の計画値と実測値に極端に大きなずれはないと考えられること，(3) 本実験の目的は欠損値処理手法の評価であり，そのためには，生産性や工数に影響を与え，欠損値の比較的少ない開発期間を含めた方が望ましいことから，開発期間を計画値と見なし，説明変数に含めた．また，説明変数として用いるメトリクスを増やしすぎると，予測用のデータセットとして使用できるプロジェクトの数が大きく減少してしまう．以上の理由により，本実験では表 1 の 4 つのメトリクスを説明変数として選択した．

このデータセットを，欠損値を含むプロジェクトのみのデータセット (プロジェクト数 706 件 (欠損率 47%)) と欠損値を含まないプロジェクトのみのデータセット (プロジェクト数 143 件) の 2 つのデータセットに分割した．前者を予測モデルを構築するデータセット (以降フィットデータと呼ぶ) とし，後者を構築されたモデルを用いて実際に予測を行うデータセット (以降テストデータと呼ぶ) とした．

表 1 実験に用いたデータセットに含まれるメトリクス

変数の種類	名称	欠損率
説明変数	FP	0%
	開発期間 (単位:月)	8.8%
	システム化計画工数	77.0%
	要件定義・設計工数	71.2%
目的変数	総開発工数 (単位:人時)	0%

無欠損データ作成法を適用するにあたっては，欠損率の高いメトリクスを削除した．本実験では，システム化計画工数を削除，要件定義・設計工数を削除，及びシステム化計画工数と要件定義・設計工数を削除する 3 つの方法を用いた．各欠損値補完・除去法を適用した結果，モデル構築に使用できるプロジェクト数は表 2 のようになった．表 2 中の P はシステム化計画工数，S は要件定義・設計工数を表している．また，k-nn 法における *k-nearestProjects*，及び CF 応用法における式 (5) の *k-nearestProjects* は，それぞれ *k-nearestProjects* の値を 1〜20 まで変化させて欠損値補完を行ったデータセットに対して重回帰モデルを構築し，その重回帰モデルの残差平方平均が最小となった 3，及び 8 を用いた．

4.3 評価基準

予測精度の評価基準として，絶対誤差 (MAE)，実測値に対する誤差 (MRE)，予測値に対する誤差 (MER)，及び Pred(25) の 4 種類の評価基準を用いた．M 件のプロジェクトがあり，実測値と予測値をそれぞれ X_i，$\hat{X}_i (i = 1 \sim M)$ と定義すると，それぞれの評価基準は次の式 (7)〜(10) により計算される．

絶対誤差 (MAE)

$$MAE = \left| \hat{X}_i - X_i \right| \qquad (7)$$

実測値に対する誤差 (MRE)

表 2　欠損値補完・除去法適用後の無欠損プロジェクト数

欠損値補完・除去法	無欠損プロジェクト数
平均値挿入法	706 件
ペアワイズ除去法	706 件
類似性に基づく補完法 (k-nn 法)	706 件
類似性に基づく補完法 (CF 応用法)	706 件
無欠損データ作成法 (P 削除)	72 件
無欠損データ作成法 (S 削除)	28 件
無欠損データ作成法 (P，S 削除)	631 件

表中の P はシステム化計画工数，S は要件定義・設計工数を表す

$$MRE = \frac{\left|\hat{X}_i - X_i\right|}{X_i} \qquad (8)$$

予測値に対する誤差 (MER)

$$MER = \frac{\left|\hat{X}_i - X_i\right|}{\hat{X}_i} \qquad (9)$$

Pred(25)

$$Pred(25) = \frac{\sum_{i=1}^{M} isAccurate(R_i)}{M} \qquad (10)$$

$$isAccurate(R) = \begin{cases} 1 & R \leq 0.25 \\ 0 & R > 0.25 \end{cases}$$

これらの基準の内，MRE は実測値を分母としているため，過大予測したモデルの誤差を求めるには有用な評価基準である一方で，過小予測したモデルの誤差が小さくなってしまうという問題点がある．これに対し，MER は予測値を分母としているため，MRE とは逆に，過大予測したモデルの誤差が小さくなってしまうという問題点がある．そこで，MRE と MER 両方の評価基準を用いることで，どの欠損値補完・除去を行ったモデルが過大予測や過小予測でもない最も予測精度の高いモデルなのか，正しく判断できる [4] [11]. Pred(25) は予測値の MRE が 25%以下となったプロジェクトの割合を示す．Pred(25) は値が大きいほど予測精度が高いことを表し，その他の評価基準は値が小さいほど予測精度が高いことを表す．

4.4　実験手順

評価実験は次の手順で行った．

1. フィットデータに対し，平均値挿入法，ペアワイズ除去法，k-nn 法，CF 応用法，及び無欠損データ作成法によって欠損値補完・除去を行う．
2. それぞれの手法で欠損値を補完あるいは削除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)．

5　結果と考察

5.1　実験結果

各欠損値補完・除去法によって欠損値の補完または除去を行い，ステップワイズ重回帰分析で予測を行った．この時の自由度調整済み決定係数 (調整済み R^2) と，表 1 に示した説明変数のうちステップワイズ変数選択法により選択された変数を表 3 に示す．調整済み R^2 はフィットデータに対するモデルの適合度を表す尺度であり，0〜1 の範囲の値を取る．調整済み R^2 が 1 に近いほど適合度が高いことを示す．表 2 に示したとおり各モデルで構築に用いるプロジェクト件数が異なるため，厳密な適合度の比較は行えないが，参考として記述している．また，予測した時の誤差の中央値を表 4 に示す．T は開発期間，P はシステム化計画工数，S は要件定義・設計工数を表している．

表 3 ステップワイズ重回帰分析適用時の調整済み R^2 と選択された変数

	調整済み R^2	選択された変数 (標準化係数)
平均値挿入法	0.562	FP(0.568)，S(0.231)，T(0.180)，P(0.102)
ペアワイズ除去法	0.957	S(0.742)，FP(0.455)
類似性に基づく補完法 (k-nn 法)	0.923	S(0.526)，P(0.391)，FP(0.117)，T(0.042)
類似性に基づく補完法 (CF 応用法)	0.925	S(0.606)，P(0.412)
無欠損データ作成法 (P 削除)	0.822	S(0.737)，FP(0.317)
無欠損データ作成法 (S 削除)	0.964	P(0.653)，T(0.379)
無欠損データ作成法 (P，S 削除)	0.548	FP(0.626)，T(0.202)

表中の T は開発期間，P はシステム化計画工数，S は要件定義・設計工数を表す

表 4 ステップワイズ重回帰分析に各欠損値補完・除去を適用した際の予測精度

	MAE 中央値	MRE 中央値	MER 中央値	Pred(25)
平均値挿入法	2648	0.818	1.112	20%
ペアワイズ除去法	1036	0.461	0.609	28%
類似性に基づく補完法 (k-nn 法)	760	0.304	0.268	43%
類似性に基づく補完法 (CF 応用法)	829	0.274	0.295	46%
無欠損データ作成法 (P 削除)	1050	0.458	0.416	28%
無欠損データ作成法 (S 削除)	1463	0.479	0.526	27%
無欠損データ作成法 (P，S 削除)	1875	0.555	0.563	18%

表中の P はシステム化計画工数，S は要件定義・設計工数を表す

MAE，MRE，MER における各手法の箱ひげ図を図 2〜図 4 に示す．グラフの上方は省略している．グラフの縦軸は予測誤差を示し，箱の下端は第 1 四分位，上端は第 3 四分位，箱中の線分は中央値，ひげの下端は第 1 四分位から IQR(Inter-Quartile Range, 箱の幅) の 1.5 倍以内に含まれる最小誤差のプロジェクト，ひげの上端は第 3 四分位から IQR の 1.5 倍以内に含まれる最大誤差のプロジェクト，○は外れ値の可能性があるプロジェクトを表す．

さらに，各欠損補完・除去法間の評価基準の差が統計的に有意かどうかを検定した．本論文の実験では同じフィットデータに対して各手法を適用しているが，MAE，MRE，MER の値は正規分布していないため，ノンパラメトリックな検定である Wilcoxon の符号付順位和検定を行った [20]．また，Pred(25) に対しては，二群の比率の差の検定を行った．それぞれの検定において有意水準は 0.05 とした．それぞれの検定結果は以降の節で述べる．

5.2 類似性に基づく補完法とその他の補完法との比較

まず，類似性に基づく補完法 (k-nn 法，CF 応用法) と平均値挿入法，ペアワイズ除去法の比較を行う．表 4 より全ての評価基準において類似度に基づいた欠損値補完法 (k-nn 法，CF 応用法) の方が平均値挿入法，ペアワイズ除去法よりも高い精度が得られたことがわかる．また，図 2〜図 4 の箱ひげ図より，類似性に基づく補完法の方が平均値挿入法，ペアワイズ除去法よりも IQR が小さい，すなわち誤差のばらつきが小さくなっていることがわかる．MAE において平均値挿入法は他手法と比べて最も IQR が小さく，誤差のばらつきが小さくなっているが，箱の位置が上部に位

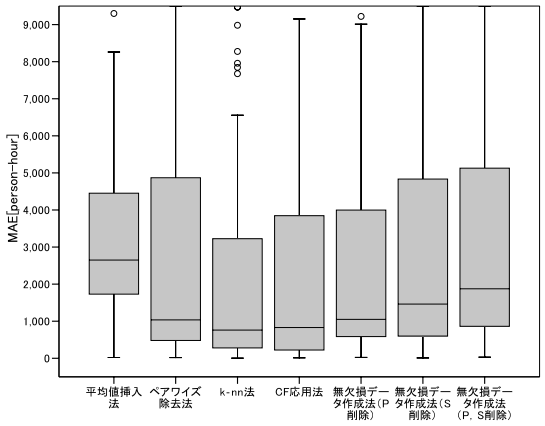

図 2　MAE の箱ひげ図

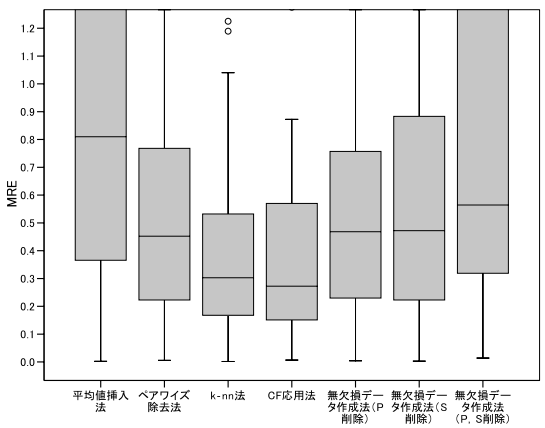

図 3　MRE の箱ひげ図

置しており予測精度は悪い．類似性に基づく補完法，その他の補完法の評価基準の差について検定を行った結果，全ての評価基準において有意差がみられた．

以上の結果より，類似性に基づく補完法を用いることで平均値挿入法，ペアワイズ除去法を用いるよりも高い精度が得られる工数予測モデルを構築できるといえる．

5.3　類似性に基づく補完法と無欠損データ作成法との比較

次に，類似性に基づく補完法と無欠損データ作成法の比較を行う．表 4 より全ての評価基準において類

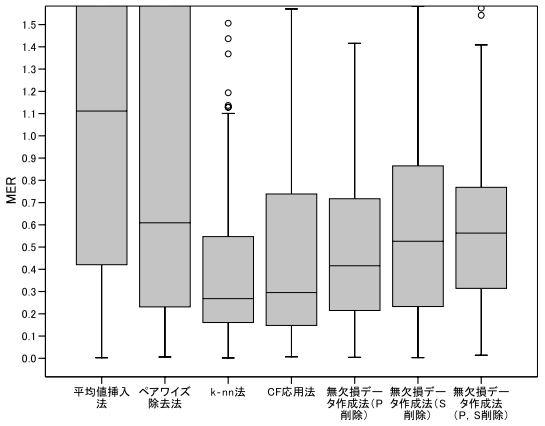

図 4　MER の箱ひげ図

似度に基づいた補完法 (k-nn 法，CF 応用法) の方が 3 つの無欠損データ作成法よりも高い精度が得られたことがわかる．無欠損データ作成法の中では，MRE 中央値と Pred(25) に関しては P 削除と S 削除間に明確な差が認められないが，MAE 中央値，MER 中央値に関しては P 削除の方が小さな誤差となっていることがわかる．

また，図 2〜図 4 の箱ひげ図においても，無欠損データ作成法の中では P 削除の IQR が概ね最も小さくなっている．以上より，システム化計画工数 (P) を削除した方法 (P 削除) が最も高い精度が得られたといえる． P 削除では 72 件のプロジェクトが残ったのに対し，S 削除では 28 件となっており，予測モデル構築に使用できるプロジェクトが減少することにより，予測値の信頼区間が広くなってしまったことが精度を低下させた一因であると考えられる．また，P, S 削除では 631 件のプロジェクトが残ったが，今度は説明変数が少なくなりすぎた (FP と開発期間の 2 つのみとなった) ことが精度が低くなった原因であると考えられる．

図 2〜図 4 の箱ひげ図より，類似性に基づく補完法 (k-nn 法) は全ての評価基準において 3 つの無欠損データ作成法よりも IQR が小さいことがわかる．また，類似性に基づく補完法 (CF 応用法) は，MRE において 3 つの無欠損データ作成法よりも IQR が小さく，MAE 及び MER において無欠損データ作成法 (P

削除) よりも IQR が大きいものの箱の位置は下部にあり，より高い精度であることがわかる．類似性に基づく補完法と無欠損データ作成法の評価基準の差について検定を行った結果，MER の CF 応用法と無欠損データ作成法 (P 削除) 間を除いて有意差がみられた．

以上の結果より，無欠損データ作成法における，プロジェクト数および説明変数を減らしすぎることの弊害が確認できた．ただし，P 削除，S 削除，P, S 削除のいずれにおいても，平均値挿入法よりは高い精度が得られたことから，欠損値補完が常に優れているとはいえない．あくまでも適切に欠損値を補完できる場合に限り，欠損値を含むプロジェクトデータを捨てずに利用することが望ましく，その有力な方法が類似性に基づく欠損値補完法であることがわかった．

この結果は，プロジェクトデータを収集する者にとって，欠損値を含むデータであっても収集する価値があることを示している．

5.4 類似性に基づく補完法間の比較

類似性に基づく補完法である CF 応用法と k-nn 法を比較すると，表 4 より予測精度に大きな差はないが，MAE 中央値及び MER 中央値においては k-nn 法の方が高い精度が得られており，MRE 中央値及び Pred(25) においては CF 応用法の方が高い精度が得られたことがわかる．図 2〜図 4 の箱ひげ図より，全評価基準において k-nn 法の方が IQR が小さい，即ち誤差のばらつきが小さくなっている．CF 応用法と k-nn 法の評価基準の差について検定を行った結果，MAE と MER において有意差がみられた．

以上の結果より，k-nn 法と CF 応用法は大きな差はないものの，どちらかといえば k-nn 法の方が高い精度が得られるといえる．また，MER において精度が低いことから，CF 応用法は k-nn 法と比べて過小予測する傾向があるといえる．一方，k-nn 法は CF 応用法と比べて過大予測する傾向があるといえる．欠損値の補完を行って工数予測を行う者は，k-nn 法と CF 応用法の両方を実施し，予測値に大きな差がないことを確認することが望ましいと考えられる．

さらに，予測に用いるメトリクスが異なるため厳密な評価は行えないが，ISBSG データセットを用いた従来研究と予測精度を比較する．本論文において精度の高かった CF 応用法の MRE 中央値が 0.274, Pred(25) が 46%であり，文献 [15] において精度の高かったステップワイズ重回帰分析による工数予測では MRE 中央値が 0.617，Pred(25) が 21%，文献 [8] において精度の高かった Robust Regression による工数予測では MRE 中央値が 0.683，Pred(25) が 0%であり，文献 [7] において精度の高かった Ordinary least-squares regression による工数予測は MRE 中央値が 0.38，Pred(25) が 21%であった．以上より，類似性に基づく欠損値補完法 (CF 応用法) を用いることで，従来研究と比較しても高い予測精度が得られたといえる．

5.5 ステップワイズ重回帰分析により選択された説明変数の比較

表 3 のように，各手法適用後にステップワイズ重回帰分析を行った時の選択された変数がばらついている理由は，現段階では不明である．しかし，高い精度で工数を予測できた手法は変数の選ばれ方も良いと仮定すると，要件定義・設計工数は開発工数の予測に有効なメトリクスであることがわかる．また，FP は無欠損のメトリクスであるにも関わらず，必ずしも選ばれないことがわかる．しかし，平均値挿入法及び無欠損データ作成法 (P，S 削除) では FP が最も高い標準化係数となっている．特に，平均値挿入法においては，他の欠損値補完手法と比較して FP 以外のメトリクスの欠損値に不適切な値が補完されたため，元々無欠損であった FP が工数予測において最も有用なメトリクスとして選択されたと考えられる．さらに，開発期間も必ずしも選ばれないことがわかる．以上より，後工程で計測されたメトリクスほど工数予測を行うための変数としては有効である可能性がある．

6 関連研究

Jonsson ら [9] は，欠損値を正確に補完することに焦点を当てており，欠損のないデータセットに対して人為的に欠損値を設けてから補完を行い，真の値との比較を行っている．この結果から，ランダムに値を欠損させた場合には k-nn 法の補完精度が高いことが示

されている．

Strike ら [19] は，欠損のないデータセットに対して人為的に欠損値を設けてから補完・除去を行い，工数予測を行っている．この結果から，k-nn 法により欠損値補完を行ったデータセットを用いることで，高い精度で工数予測が行えることが示されている．しかし，現実のデータセットの欠損値の発生は，ランダムというよりはむしろバースト的であり，欠損値の分布に大きな偏りがある．

また，Cartwright ら [3]，Myrtveit ら [16] は，欠損値を含むデータセットに対して欠損値補完・除去を行い，重回帰モデルを作成している．しかし，モデル構築に使用したデータセットに対して当てはまりの良いモデルを作成することに焦点があてられており，実際に作成したモデルがどの程度の精度で予測を行えるのか評価を行っていない．

本論文では，意図的に欠損させたのではなく実際に欠損を多く含んだデータセットに対して欠損値補完・除去を行ったことと，モデル構築後に (欠損値のないデータセットを用いて) 工数予測を行い，精度の評価を行った点が従来と異なる．さらに，本論文では，欠損値補完法と無欠損データ作成法との比較を行った点が従来と異なる．

7 おわりに

本論文では，欠損値を含む過去のソフトウェア開発プロジェクトデータセットを用いて開発工数予測モデルを構築する際の，欠損値補完手法の効果を実験的に明らかにした．得られた主な結果は次の通りである．

- 類似性に基づく欠損値補完手法 (CF 応用法，k-nn 法) によって欠損値を補完することで，比較対象とした欠損値補完・除去法である平均値挿入法，ペアワイズ除去法，無欠損データ作成法よりも高い精度の工数予測モデルが構築できた．
- 無欠損データ作成法よりも，類似性に基づく欠損値補完手法が優れていたことから，プロジェクトデータを収集する者にとって，欠損値を含むデータであっても収集する価値があることが示された．
- CF 応用法はやや過小予測する傾向があり，k-nn 法はやや過大予測する傾向があることから，欠損値の補完を行って工数予測を行う者は，k-nn 法と CF 応用法の両方を実施し，予測値に大きな差がないことを確認することが望ましいと考えられる．

今後は，他のプロジェクトデータセットが入手できたならば，同様の評価実験を行う予定である．また，欠損値をより適切な値で補完できるようにさらなる手法の改善，及び欠損値を前提とした工数予測手法 (Optimized Set Reduction 法など) との精度比較を行う予定である．

謝辞

本研究の一部は，文部科学省「e-Society 基盤ソフトウェアの総合開発」及び「次世代 IT 基盤構築のための研究開発」の委託に基づいて行われた．

参 考 文 献

[1] Boehm, B.W.: *Software engineering economics*, Prentice Hall, New Jersey, 1981.

[2] Briand, L., Langley, T. and Wieczorrek, I.: A replicated assessment and comparison of common software cost modeling techniques, in *Proc. 22nd IEEE International Conf. on Softw. Eng.*, Limerick, 2000, pp. 377–386.

[3] Cartwright, M., Shepperd, M.J. and Song, Q.: Dealing with Missing Software Project Data, in *Proc. 9th IEEE International Softw. Metrics Symposium* (*Metrics'03*), Sydney, Australia, 2003, pp. 154–165.

[4] Foss, T., Stensrud, E., Kitchenham, B. and Myrtveit, I.: A Simulation Study of the Model Evaluation Criterion MMRE, *IEEE Trans. Softw. Eng.*, Vol. 29, No. 11(2003), pp. 985–995.

[5] Hawkins, D.M.: The Problem of Overfitting, *Journal of Chemical Information and Modeling*, Vol. 44, No. 1(2004), pp. 1–12.

[6] ISBSG Estimating, Benchmarking and Research Suite Release 9: International Software Benchmarking Standards Group, 2004, http://www.isbsg.org/

[7] Jeffery, R., Ruhe, M. and Wieczorek, I.: A comparative study of two software development cost modeling techniques using multi-organizational and company-specific data, *Information and Softw. Technology*, Vol. 42(2000), pp. 1009–1016.

[8] Jeffery, R., Ruhe, M. and Wieczorek, I.: Using Public Domain Metrics to Estimate Software Development Effort, in *Proc. 7th IEEE International Softw. Metrics Symposium* (*Metrics'01*), London, England, 2001, pp. 16–27.

[9] Jonsson, P. and Wohlin, C.: An evaluation of k-nearest neighbour imputation using likert data, in *Proc 10th IEEE International Softw. Metrics Symposium* (*Metrics'04*), Chicago, Illinois, 2004, pp. 108–118.

[10] 柿元健, 角田雅照, 大杉直樹, 門田暁人, 松本健一: 協調フィルタリングに基づく工数見積り手法のデータの欠損に対するロバスト性の評価, 電子情報通信学会論文誌, Vol. J89-D, No. 12(2006), pp. 2602–2611.

[11] Kitchenham, B., Pickard, L., MacDonell, S.G. and Shepperd, M.J.: What accuracy statistics really measure, IEE Proc. Softw. Vol. 148, No. 3(2001), pp. 81–85.

[12] Kitchenham, B., Mendes, E. and Travassos, G.: Cross versus Within-Company Cost Estimation Studies: A Systematic Review, *IEEE Trans. Softw. Eng.*, Vol. 33, No. 5(2007), pp. 316–329.

[13] Kromrey, J. and Hines, C.: Nonrandomly missing data in multiple regression: An empirical comparison of common missing-data treatments, *Educational and Psychological Measurement*, Vol. 54, No. 3(1994), pp. 573–593.

[14] Little, R.J.A. and Rubin, D.B.: *Statistical Analysis with Missing Data*, 2nd ed., John Wiley and Sons, New York, 2002.

[15] Mendes, E., Lokan, C., Harrison, R. and Triggs, C.: A Replicated Comparison of Cross-company and Within-company Effort Estimation Models using the ISBSG Database, in *Proc. 11th IEEE International Softw. Metrics Symposium* (*Metrics'05*), Como, Italy, 2005, p. 36.

[16] Myrtveit, I., Stensrud, E. and Olsson, U. H.: Analyzing data sets with missing data: An empirical evaluation of imputation methods and likelihood-based methods, *IEEE Trans. Softw. Eng.*, Vol. 27, No. 11(2001), pp. 999–1013.

[17] Shepperd, M. and Schofield, C.: Estimating software project effort using analogies, *IEEE Trans. Softw. Eng.*, Vol. 23, No. 12(1997), pp. 736–743.

[18] Srinivasan, K. and Fisher, D.: Machine learning approaches to estimating software development effort, *IEEE Trans. Softw. Eng.*, Vol. 21, No. 2(1995), pp. 126–137.

[19] Strike, K., El Eman, K. and Madhavji, N.: Software cost estimation with incomplete data, *IEEE Trans. Softw. Eng.*, Vol. 27, No. 10(2001), pp. 890–908.

[20] 田中豊, 垂水共之 (編): Windows 版 統計解析ハンドブック ノンパラメトリック法, 共立出版, 東京, 1999.

[21] 角田雅照, 大杉直樹, 門田暁人, 松本健一, 佐藤慎一: 協調フィルタリングを用いたソフトウェア開発工数予測方法, 情報処理学会論文誌, Vol. 46, No. 5(2005), pp. 1156–1164.

[22] Walston, C. and Felix, C.: A Method of Programming Measurement and Estimation, *IBM Systems Journal*, Vol. 16, No. 1(1977), pp. 54–73.