組込みソフトウェア開発における設計関連メトリクスに基づく下流試験欠陥数の予測



角田　雅照[†1, †2]　　門田　暁人[†1, †3]　　松本　健一[†1]

本研究では，組込みソフトウェア開発の設計工程から得られるメトリクスを用いて，テスト後半（結合テスト，総合テスト）における欠陥数の予測を試みた．モデルの説明変数として，基本・詳細設計におけるレビュー工数，検出欠陥数といった品質保証に関するメトリクスに加えて，試験仕様書に関するメトリクスを用いた．さらに，各仕様書のドキュメント量に関するメトリクスも採用した．分析の結果，*BRE* 平均値がおおむね 35% 程度となり，比較的高い予測精度となった．また，基本設計終了後に予測を行っても，比較的予測精度が高くなることがわかった．

Predicting Faults After Unit Testing Using Design Phase Metrics in Embedded Software Development

Masateru Tsunoda[†1, †2], Akito Monden[†1, †3], Kenichi Matsumoto[†1]

This research tried to predict the number of faults after unit testing using design phase (i.e., integration testing and system testing) metrics in embedded software development. As explanatory variables of the model, we used review effort in basic and detail design phase, metrics related to quality assurance such as number of found faults. In addition, we used metrics related to test specification documents and metrics about amount of specification documents. Experimental results shows that average *BRE* (balanced relative error) of the models were almost 35%, and we can say that prediction accuracy of them were relatively high. Also, the results show that when the number of faults is predicted after basic design phase, the prediction accuracy is relative high.

1. はじめに

近年，組込みソフトウェアの開発は非常に増加しており，その重要性が増している．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，GDP では 13% 以上の比率となっている [Tamaru 2008].

本研究では，近年大きな課題となっている組込みソフトウェアの品質確保 [Mihara2013] に着目し，開発の設計工程から得られるメトリクスを用いて，テストの後半（結合テ

【脚注】
† 1　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
† 2　近畿大学理工学部情報学科
Department of Informatics, Kindai University
† 3　岡山大学大学院自然科学研究科
Graduate School of Natural Science and Technology, Okayama University

表１　分析対象のメトリクス一覧

	メトリック	プロセス	試験	詳細	詳細	関連 RQ
v_1	基本設計仕様書分量	N	N	N	基本設計時の仕様書の分量を示す.	1, 2, 3
v_2	基本設計仕様書レビュー工数	Y	N	N	基本設計時の仕様書をレビューした際の工数を示す.	1, 2
v_3	基本設計仕様書レビュー欠陥数	Y	N	N	基本設計時の仕様書をレビューした際に発見された欠陥数を示す.	1, 2
v_4	基本設計試験仕様書分量	N	Y	N	基本設計に関してシステムテストにて試験するための仕様書の分量を示す.	1, 2, 3, 4
v_5	基本設計試験仕様書レビュー工数	Y	Y	N	基本設計のテスト仕様書をレビューした際の工数を示す.	1, 2, 4
v_6	詳細設計仕様書分量	N	N	Y	詳細設計時の仕様書の分量を示す.	1, 3
v_7	詳細設計仕様書レビュー工数	Y	N	Y	詳細設計時の仕様書をレビューした際の工数を示す.	1
v_8	詳細設計仕様書レビュー欠陥数	Y	N	Y	詳細設計時の仕様書をレビューした際に発見された欠陥数を示す.	1
v_9	詳細設計試験仕様書分量	N	Y	Y	詳細設計に関して結合テストにて試験するための仕様書の分量を示す.	1, 3, 4
v_{10}	詳細設計試験仕様書レビュー工数	Y	Y	Y	詳細設計のテスト仕様書をレビューした際の工数を示す.	1, 4
v_{11}	欠陥数	-	-	-	結合テストとシステムテスト時に発見された欠陥数を示す.	-

スト，総合テスト）における欠陥数の予測を試みた結果について報告する．開発初期にテストの後半に発見される欠陥を予測することで，出荷後品質を確保するためのテスト計画立案や，手戻り工数の確保などに役立つと期待される．また，欠陥数の増大に寄与するメトリクスを明らかにすることで，欠陥を減らすためのプロセス改善にも役立つと期待される．

本研究では，ある組込みソフトウェア開発企業で収集されたデータに基づき，テスト後半の欠陥数がどの程度定量的モデルに基づいて予測可能であるかを分析する．また，テスト後半の欠陥数を予測するために優先的に収集すべきメトリクスが存在するのかどうかを明らかにする．モデルの説明変数に用いるメトリクスの候補として，基本・詳細設計におけるレビュー工数，検出欠陥数といった品質保証に関するメトリクスに加えて，試験仕様書に関するメトリクスを用いる．さらには，各仕様書のドキュメント量に関するメトリクスも採用し，ドキュメントの分量が欠陥予測に有効であるかを分析する．

これらのメトリクスは多くの企業において計測されており，テスト時に発見される欠陥数の予測に取り組んでいる研究も存在する [Takata1994][Tsunoda2009][Komuro2011]．ただし，我々の知る限り，プロジェクトの早期にテスト後半の欠陥数を予測することを目的として，設計ドキュメントに関するメトリクスのみを用いて予測モデルの構築を試みた研究は存在しない．

以降，２章では分析対象のデータについて説明し，３章では分析方法について説明する．４章では分析結果について述べ，５章では関連研究について説明する．最後に６章でまとめを述べる．

2.　分析に用いたデータ

分析に用いたデータは，ある組込みソフトウェア開発企業の２つの部署において収集されたデータである．２つの部署で開発される組込みソフトウェアが対象とするものは，有線・無線通信，画像処理，公共交通関連などであり，開発プロセスはウォータフォールとなっている．２つの部署はそれぞれ独立しており，別々の顧客に対してソフトウェアを開発している．大部分のプロジェクトでは，顧客から与えられた要求に基づき，請負開発を行っており，基本設計，詳細設計，コーディング，単体テスト，結合テスト，システムテストを行っている．なお，システムテストではハードウェアのテストは含んでいない．開発規模については部署 A では新規開発部分の中央値が約 10,000 ステップ，流用部分が約 21,000 ステップであり，部署 B ではともに約 20,000 ステップであった．

データは 2009 年から 2012 年に収集されたものであり，部署 A には 53 件，部署 B には 54 件のデータが含まれている．データセットに含まれるメトリクスを表 1 に示す．メトリクスは，基本設計時のドキュメントに関するものと，詳細設計時のドキュメントに関するものが含まれる．前者は列「詳細」が N，後者は Y となっているメトリクスが該当する．ドキュメントには，設計書と試験仕様書の両方が存在する．前者に関するメトリクスは列「試験」が N，後者は Y となっている．さらに，それぞれに対してレビューが実施されており，それらに関するプロセスメトリクス，すなわちレビュー工数とレビュー時の欠陥数が記録されている．これらのメトリクスでは列「プロセス」を Y とし，それ以外を N としている．なお，試験仕様書のレビュー時の欠陥数は記録されていなかった．

予測対象であるテスト後半の欠陥数は，結合テストでの欠陥数とシステムテストでの欠陥数を合算したものである．データセットにはソースコードの行数，テストケース数，コードレビュー工数なども記録されていたが，これらのメトリクスは分析に使用しなかった．本研究ではできるだけ早期に，すなわち遅くとも詳細設計後にテスト後半の欠陥を予測することを前提としており，それ以降に収集されるメトリクスは予測モデルの説明変数として利用しないため，これらを分析から除外している．

分析では，表 1 のメトリクスに欠損を含むケースを除外して分析した．これは一般的にリストワイズ除去法 [Little 2002] と呼ばれる方法であり，欠損，すなわち変数に値が記録されていないケースを分析から除外する．欠損値処理の方法には，リストワイズ除去法の他に，欠損値に各変数の平均値を挿入する平均値挿入法 [Little2002] などがあるが，Strike ら [Strike2001] は，ソフトウェア開発データを用いて予測モデルを構築する場合，リストワイズ除去法による欠損値処理が妥当であることを示している．そこで本研究では Strike らの実験結果に基づき，リストワイズ除去法を適用した．その結果，部署 A のデータ件数は 24 件，部署 B のデータ件数は 19 件となった．

3. 分析方法

3.1. リサーチクエスチョン

本研究の目的は，組込みソフトウェアの品質管理を支援するために，設計ドキュメントに関するメトリクスを用いてテスト後半の欠陥数の予測を行うことである．そのために，以下の 4 つのリサーチクエスチョンを設定した．

- RQ1: 設計ドキュメントに基づくメトリクスのみを用いて，テスト後半の欠陥数を予測できるのか？
- RQ2: 基本設計時に収集されるメトリクスのみで欠陥数を予測できるのか？
- RQ3: プロダクトメトリクスのみを用いて欠陥数を予測できるのか？
- RQ4: 試験仕様書に関するメトリクスを用いることにより，欠陥数の予測精度が高まるのか？

表 1 の列「関連 RQ」に，それぞれのリサーチクエスチョンの分析において各メトリクスを用いるかどうかを示している．我々の知る限り，組込みソフトウェアにおいて，ドキュメントに関するメトリクスを用いてテスト後半の欠陥数を予測できるのかどうかは明らかでない．そこで RQ1 を設定した．基本設計時に収集されるメトリクスのみを用いて予測することができれば，より早期にテスト計画を立案することができる．そこで RQ2 を設定した．一般に，プロセスメトリクスはプロダクトメトリクスよりも収集するほうがコストが掛かるため，収集が容易ではない．実務者がプロセスメトリクスを収集するかどうか判断材料の一つとできるように，RQ3 を設定した．設計書に対応した試験仕様書は必ずしも作成されるわけではなく，また，設計時に作成されるとも限らない．もしこれらを作成していた場合，これらのメトリクスを活用したほうが良いのかどうかを実務者が判断できるように RQ4 を設定した．

実験では，各リサーチクエスチョンに答えるため，以下の 3 つを組み合わせてモデルを作成し，それぞれのモデルの予測精度を比較した．

- 詳細設計時のメトリクスの有無
- プロセスメトリクスの有無
- 試験仕様書に関するメトリクスの有無

3.2. 重回帰分析

欠陥数予測モデルは，重回帰分析を用いて構築した．重回帰分析はソフトウェアプロジェクトにおけるメトリクスを予測するモデルを構築する際に広く用いられる方法である．重回帰分析では，以下のようなモデルが構築される．

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \quad (1)$$

ここで y は予測対象である目的変数を表し，本研究では欠陥数に該当する．$x_1, x_2, \ldots, x_k$ は k 個の説明変数であり，本研究では設計ドキュメントに関するメトリクスとなる．β_0 は回帰定数，$\beta_1, \beta_2, \ldots, \beta_k$ は偏回帰係数，ε は誤差項である．偏回帰係数同士を比較可能なように標準化したものを標準化偏回帰係数と呼ぶ．偏回帰係数の絶対値が大きいほど，目的変数に対する該当の説明変数の影響が大きいことを示す．

対数変換：重回帰分析を用いてソフトウェアプロジェクトのメトリクスを予測する場合，モデルの構築前に説明変数，目的変数に対し，対数変換が適用される場合がある．ソフトウェアプロジェクトのデータセットには多数の小規模プロジェクトと少数の大規模プロジェクトが存在する場合が多く，そのためメトリクスの分布も偏っている場合が多い．このような場合，対数変換を適用することより，モデルの予測精度が高まることが多い．そのため，本研究でも対数変換を適用する．

多重共線性：構築されたモデルに多重共線性が発生していないかどうかを確認するために，VIF(Variance inflation factor) と条件指標を用いる．多重共線性とは，説明変数間 に強い線形の関係が存在していることであり，多重共線性 が発生していると，偏回帰係数の分散が大きくなり，偏回帰係数の推定が不安定になる [Onodera2004]．一般に，各変数の VIF が 10 を超える場合やモデルの条件指標が 30 を超える場合に，多重共線性が発生しているとされる [Tabachnick1996] [Tanaka1995].

外れ値：予測モデルを構築する際，外れ値を除外するこ

表 2　構築された予測モデル一覧（部署 A）

プロセス	試験	詳細	Adj. R^2	p 値	条件指標	Cook の距離	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}
Y	Y	N	0.62	0.00	16.0	0.66	3			1	2	-	-	-	-	-
Y	Y	Y	0.62	0.00	16.0	0.66	3			1	2					
N	Y	N	0.56	0.00	12.4	0.47	2	-	-	1	-	-	-	-	-	-
N	Y	Y	0.56	0.00	12.4	0.47	2	-	-	1	-		-	-		-
N	N	N	0.42	0.00	8.0	0.49	1	-	-	-	-	-	-	-	-	-
N	N	Y	0.42	0.00	8.0	0.49	1	-	-	-	-		-	-	-	-
Y	N	N	0.42	0.00	8.0	0.49	1			-	-	-	-	-	-	-
Y	N	Y	0.42	0.00	8.0	0.49	1			-	-				-	-

表 3　構築された予測モデル一覧（部署 B）

プロセス	試験	詳細	Adj. R^2	p 値	条件指標	Cook の距離	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}
Y	Y	Y	0.75	0.00	15.4	1.16						1			3	2
N	Y	Y	0.68	0.00	11.7	1.84		-	-	2	-	1	-	-		-
N	N	Y	0.65	0.00	9.1	1.66		-	-	-	-	1	-	-	-	-
Y	N	Y	0.65	0.00	9.1	1.66				-	-	1			-	-
N	Y	N	0.53	0.00	13.1	0.45	1	-	-	2	-	-	-	-	-	-
Y	Y	N	0.53	0.00	13.1	0.45	1			2		-	-	-	-	-
N	N	N	0.50	0.00	10.4	0.31	1	-	-	-	-	-	-	-	-	-
Y	N	N	0.50	0.00	10.4	0.31	1			-	-	-	-	-	-	-

とが行われる．外れ値とはメトリクスに何らかの特異な値を含むケースである．特異な値とは，他のプロジェクトと比較して大きく異なる値のことであり，発生する理由は単なる記録ミスの場合もあれば，仕様変更などの特殊な事情により起こるなど様々である．重回帰分析において外れ値を特定して除外する方法として，Cook の距離がある．Cook の距離とは，あるケースをモデルの推定の計算から除外した場合に，すべてのケースの残差がどの程度変化するかを示す距離である．Cook の距離が大きい時は，そのケースを除外して回帰統計量を計算した場合に係数が大きく変化したことを示す [Onodera2004]．一般に，Cook の距離が 1 より大きいケースはモデルの分析結果に大きな影響を与えているとみなされ，モデル構築から除外される．

3.3. 変数選択

重回帰分析により予測モデルを構築する際，変数選択を行った．変数選択は説明変数の候補から有用な変数を選択してモデルを構築する方法である．本研究では，変数増減法により変数選択を行った．具体的には以下の手順により行う [Tanaka1995]．

1. どの変数も入っていない状態から開始する．
2. もし，すべての変数が含まれていれば，取り込むべき変数はないという情報を持ってステップ 3 に進む．すべての変数が含まれていなければ，残りの変数を順番に 1 つずつ採用してみて，偏回帰係数の検定のための F 値を計算し，その値が最大となる変数を選ぶ．その F 値に対応する確率が指定された p_{in} より小さく，かつその変数を採用することになって多重共線性が生じない（VIF が 10 を超えない）ならば，その変数を取りこんで次のステップに進む．p_{in} より大きければ，取りこむべき変数はないという情報を持ってステップ 3 に進む．
3. モデルに含まれている変数について，偏回帰係数の検定のための F 値を計算し，F 値が最小となる変数を選ぶ．その F 値に対応する確率が指定された p_{out} より小さい時，取りこむべき変数がないという情報があれば終了する．そうでなければ，どの変数も落さずステップ 4 に進む．F 値に対応する確率が p_{out} より大きい時，その変数を落とし（取りこむべき変数がないという情報があれば，それをキャンセルして），再びステップ 3 に戻る．
4. すべての変数が取りこまれていれば終了する．そうでなければステップ 2 に戻る．

p_{in}，p_{out} は 0.05 から 0.5 の範囲で，重要な変数を落さないことに重点をおくなら大きい値，余分な変数を取りこまないことに重点をおくなら小さい値を指定する [Tanaka 1995]. 本研究では重要な変数を落さないことに重点をおき，p_{in} に 0.2，p_{out} に 0.4 を指定する．

3.4. 予測モデルの評価

自由度調整済み決定係数：構築されたモデルがどの程度適切に構築されたのかを確かめるために，本研究では自由度調整済み決定係数を用いる．決定係数はデータに対するモデルの当てはまりの良さを評価する指標であり，1 に近いほどモデルのデータに対する適合度がよいことを示す [Misono2007]．ただし，決定係数は説明変数の数が増加するに従って高くなる傾向がある [Misono2007]．この問題を

表 4　各予測モデルの精度評価（部署 A）

プロセス	試験	詳細	*AE* 平均値	*AE* 中央値	*BRE* 平均値	*BRE* 中央値	R1	R2	R3	R4	平均順位
N	N	N	9.4	6.1	34.7%	28.2%	1	3	1	1	1.5
Y	N	N	9.4	6.1	34.7%	28.2%	1	3	1	1	1.5
N	N	Y	9.4	6.1	34.7%	28.2%	1	3	1	1	1.5
Y	N	Y	9.4	6.1	34.7%	28.2%	1	3	1	1	1.5
N	Y	N	11.9	4.9	37.6%	35.9%	5	1	7	7	5.0
N	Y	Y	11.9	4.9	37.6%	35.9%	5	1	7	7	5.0
Y	Y	N	12.5	6.7	36.8%	33.3%	7	7	5	5	6.0
Y	Y	Y	12.5	6.7	36.8%	33.3%	7	7	5	5	6.0

表 5　各予測モデルの精度評価（部署 B）

プロセス	試験	詳細	*AE* 平均値	*AE* 中央値	*BRE* 平均値	*BRE* 中央値	R1	R2	R3	R4	平均順位
N	Y	Y	42.2	35.6	35.9%	30.0%	1	4	1	1	1.8
N	Y	N	46.3	34.4	36.1%	32.5%	3	2	2	2	2.3
Y	Y	N	46.3	34.4	36.1%	32.5%	3	2	2	2	2.3
Y	Y	Y	43.2	23.6	36.3%	41.0%	2	1	6	8	4.3
N	N	Y	46.4	36.7	36.8%	32.7%	5	5	7	4	5.3
Y	N	Y	46.4	36.7	36.8%	32.7%	5	5	7	4	5.3
N	N	N	49.6	39.3	36.3%	36.9%	7	7	4	6	6.0
Y	N	N	49.6	39.3	36.3%	36.9%	7	7	4	6	6.0

回避するために，データの数と説明変数の数によって決定係数を補正した，自由度調整済み決定係数を用いる．

リーブワンアウト法：構築されたモデルの予測精度を評価するために，リーブワンアウト法 [Kocaguneli2013] を適用する．リーブワンアウト法とは，データセットからあるケースを 1 件取り出して予測対象のテストデータとし，残りのケースをモデル構築に用いるラーニングデータとする．これを，すべてのケースに対して繰り返す方法である．テストデータでは欠陥数を未知とみなし，予測後に実際の欠陥数と比較して予測がどの程度正確であるかを評価する．

予測精度評価指標：構築されたモデルによる欠陥数の予測精度を評価するため，*AE*(Absolute Error)，*MRE*(Magnitude of Relative Error)[Conte1986]，*BRE* (Balanced Relative Error) [Miyazaki1994] の 3 つの指標の平均値と中央値を用いた．欠陥数の実測値を x，予測値を $\hat{x}$ とするとき，それぞれの指標は以下の式により求められる．

$$AE = |x - \hat{x}| \tag{2}$$

$$MRE = \frac{|x - \hat{x}|}{x} \tag{3}$$

$$BRE = \begin{cases} \dfrac{(\hat{x} - x)}{x}, \hat{x} - x \geq 0 \\ \dfrac{(x - \hat{x})}{\hat{x}}, \hat{x} - x < 0 \end{cases} \tag{4}$$

それぞれの指標の値が小さいほど，予測値と実測値との差が小さい，すなわち予測精度が高いことを示す．直感的には *MRE* と *BRE* は実測値との相対誤差であるといえる．ただし，*MRE* は過大予測に対し，アンバランスな評価になるという問題がある．過少予測の場合，*MRE* は最大でも 1 にしかならない．例えば実測値が 10，予測値が 0 の場合，*MRE* は 1 となる．そこで本研究では *MRE* に加え，過大予測と過少予測をバランスよく評価する指標 [Mølokken-Østvold2005] である *BRE* を予測精度の評価指標に用いる．

4.　分析結果

部署 A と部署 B で目的変数である欠陥数の値域が大きく異なっていたため，データの層別を行い，それぞれの部署向けに予測モデルを構築した．データを層別して個別のモデルを構築することは，精度の高い予測モデルを構築するために一般に行われる方法の一つである．

構築したモデルの概要を 4.1 節で述べ，4.2 節において *BRE* などの予測精度を中心にモデルを評価するとともに，リサーチクエスチョンに対する答えを述べる．

4.1. モデルの概要

部署 A 向けに構築した予測モデルの概要を表 2 に，部署 B 向けに構築した予測モデルの概要を表 3 に示す．プロセス，試験，詳細の各列は，それぞれプロセスメトリクス，試験仕様書に関するメトリクス，詳細設計時のメトリクスを説明変数の候補として含めたかを示す．含めた場合には Y とし，含めない場合は N としている．その結果，各変数が採用されたかどうかを v_1 から v_{10} の列に示す．v_1 から v_{10} に対応する変数は表 1 に示している．列中の数値は，構築されたモデルにおいて，標準偏回帰係数の絶対値の大きさの順位を示している．数値が小さいほど偏回帰係数が大きい，すなわち目的変数の値に対する影響が大きかったことを示

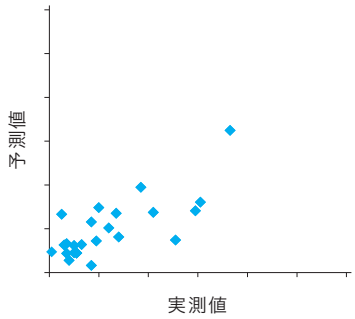

図１　最も精度の高いモデルによる予測（部署 A）

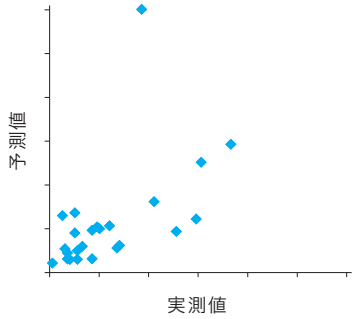

図２　最も精度の高いモデルによる予測（部署 B）

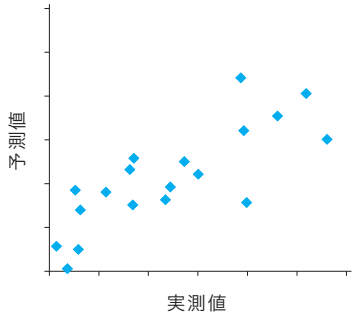

図３　最も精度の低いモデルによる予測（部署 A）

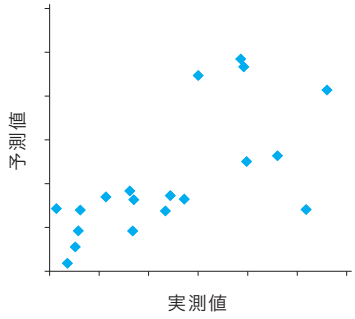

図４　最も精度の低いモデルによる予測（部署 B）

す．空白は採用されなかったことを示し，「-」は説明変数の候補に含まれていなかったことを示す．Adj. R^2 は自由度調整済み決定係数を示し，この値が大きい順に各行，すなわちモデルを並べ替えている．

モデルが適切に構築されたことを示すために，p 値，条件指標，Cook の距離を示している．p 値は構築されたモデルが統計的に有意であったかどうかを示し，全て有意水準 5% で有意であった．表に示すように，すべてのモデルの条件指標は 30 を下回っていた．また，結果は省略するが，各変数の VIF は 10 を下回っていたことから，多重共線性は発生していないと考えられる．Cook の距離は，各モデルにおける Cook の距離の最大値を示す．部署 A では全てのモデルが，部署 B では半数のモデルにおいて 1 を下回っていたことから，除外すべき（明らかな）外れ値プロジェクトは含まれていないと考えられる．

自由度調整済み決定係数によってモデルを並べ替えた場合，部署 A では説明変数に関して以下の結果が観察された．それぞれの項目は RQ2 から RQ4 に関連しているが，各リサーチクエスチョンに対する答えは次節の分析結果に基づいて述べる．

- 全てのモデルにおいて，詳細設計時のメトリクスは説明変数として採用されなかった．
- 自由度調整済みの上位 2 つのモデルにおいて，プロセスメトリクス v_5 が採用されたが，その他のモデルでは採用されなかった．
- 上位 4 つのモデルにおいて，試験仕様書に関するメトリクス v_4 が採用され，かつ偏回帰係数が最も大きくなっていた．

同様に，部署 B では以下のような結果となった．

- 上位 4 つのモデルにおいて，詳細設計時のメトリクス v_6 が採用され，かつ偏回帰係数が最も大きくなっていた．
- 最上位のモデルにおいて，プロセスメトリクス v_{10} が採用されたが，その他のモデルでは採用されなかった．
- 最上位のモデルにおいて，試験仕様書に関するメトリクス v_9，v_{10} が採用された．上位から 2 番目のモデルにおいて同様のメトリクス v_4 が採用された．

4.2. モデルの予測精度

部署 A 向けに構築したモデルの予測精度を表 4 に，部署 B 向けに構築したモデルの予測精度を表 5 に示す．プロセス，試験，詳細の各列の意味は表 2，表 3 と同様である．列 R1 から R4 は，*AE* 平均値から *BRE* 中央値の 4 つの指標それぞれに関して，全モデルの中で何番目に精度が高かったのかを順位付けしたものである．平均順位は，その 4 つの順位を平均したものであり，それぞれの表は平均順位の昇順，すなわち予測精度が高かった順にモデルを並び替えている．

RQ1 に関する分析：最も精度の高かったモデルにおける予測精度に着目する．部署 A，部署 B において，最も精度の高いモデルでは *BRE* 平均値が 35% 程度，*BRE* 中央値が 30% 程度となっており，誤差が小さかった．部署 B の場合，*AE* の平均値と中央値が部署 A よりも大きくなっているが，これは欠陥数の値域が異なるためであり，相対的な誤差（*BRE*）を考慮すると，予測精度が低いわけではない．図 1，図 2 は，予測精度を視覚化するために，横軸を実測値，縦軸を予測値として散布図としたものである．図には部署

A，B において最も精度が高かった（平均順位が最も小さい）モデルを示している（守秘義務上，各軸に値を示していないが，各グラフの縦軸と横軸の値域は同一としている）．図からも良好な予測結果といえる．これらの結果から，RQ1 に対する答えは Yes となる．なお，部署 A，B において最も精度が低かった場合でも，比較的予測精度は良好であった（図 3，図 4）．

RQ2 に関する分析：部署 A では，最も精度の高いモデルでは基本設計時に収集されるメトリクスのみが採用されていた．部署 B の場合，予測精度の上位 2 番目と 3 番目のモデルにおいて，基本設計時のメトリクスのみが採用されており，1 番目のモデルと比べて予測精度に大きな差はなかった．外れ値の影響を避けるため，Wilcoxon の符号付き順位検定を用い，予測精度に差があるかを検定したが，有意水準 5% で有意差がなかった．よって，RQ2 に対する答えは Yes となる．

RQ3 に関する分析：部署 A，部署 B とも，最も精度の高いモデルではプロダクトメトリクスのみが採用されていた．このことから，RQ3 に対する答えは Yes となる．ただし表 2，表 3 において，プロセスメトリクスを含んでいるモデルのほうが説明力（自由度調整済み決定係数）が大きかったことから，これらのメトリクスを収集している企業の場合，説明変数として用いることを検討してもよいと考えられる．

RQ4 に関する分析：部署 A では，最も精度の高いモデルでは試験仕様書に関するメトリクスが採用されていなかった．部署 B の場合，予測精度の上位 4 つのモデルでは，全て試験仕様書に関するメトリクスを用いていた．ただし，予測精度に大きな違いはなく，Wilcoxon の符号付き順位検定においても，有意水準 5% で有意差がなかった．このことから，RQ4 に対する答えは No となる．なお，表 2，表 3 において，説明力が上位のモデルでは試験仕様書に関するメトリクスを含んでおり，また，有意差がなかったとはいえ，予測精度がわずかながら高かったことから，これらのメトリクスが既に収集されている場合，活用を検討してもよいと考えられる．

4.3. 考察

目的変数に単体テストの欠陥数を含めて予測した場合のモデルの精度を表 6 に示す．表には，説明変数の候補として全ての変数を用いた場合と，基本設計仕様書分量のみを用いた場合の精度を部署別に示している．単体テストの欠陥数を含めない場合（表 4，表 5）と比較して，部署 A では予測精度が大きく低下し，部署 B では予測精度の若干の低下が見られた．Wilcoxon の符号付き順位検定を用い，単体テストの欠陥数を含めない場合と予測精度に差があるかを検定すると，部署 A の場合，*AE*，*BRE* とも有意に悪化していた．なお，検定では各表で最も精度の高いモデル同士を比較している．

表 6　単体テスト欠陥数を含めた場合の予測精度

部署	説明変数候補	*AE* 平均値	*AE* 中央値	*BRE* 平均値	*BRE* 中央値
A	v_1	31.5	16.2	159.4%	74.4%
A	v_1 - v_9	46.4	13.6	121.8%	56.5%
B	v_1	61.9	33.8	43.2%	26.0%
B	v_1 - v_9	67.1	36.6	40.7%	36.5%

表 7　単体テストの欠陥数と結合テスト以降での欠陥数との相関係数

部署	相関係数	p 値
A	0.38	0.07
B	0.71	0.00

単体テストの欠陥数と，結合テスト以降での欠陥数について，部署ごとにスピアマンの順位相関係数を求めた結果を表 7 に示す．部署 B では相関が高かったのに対し，部署 A では相関が低く，かつ有意水準 5% で有意となっていなかった．部署 A と B で単体テストの欠陥数を含めた場合と含めない場合で予測精度に違いが生じた原因は，単体テストの欠陥数と，結合テスト以降での欠陥数の関連の強さの違いが原因であると考えられる．

この分析結果より，設計関連メトリクスを用いてプロジェクトの早期に欠陥数を予測する場合，単体テストの欠陥数を含めると予測精度が低下することがあるといえる．

5.　関連研究

これまで，プロセスに関するメトリクスを用いてテスト時の欠陥数の予測を試みた研究はいくつか存在する．小室ら [Komuro2011] は，レビュー実績などを用いてテスト工程で発見される欠陥数を予測するモデルを提案している．ただし，説明変数に設計ドキュメントに関するプロダクトメトリクスを用いておらず，また，テスト工程全体の欠陥数を予測対象としている点が異なる．文献 [Tsunoda2009] では，システムテスト欠陥密度を予測するために，コードレビュー指摘密度などを説明変数に用いている．ただし，ソフトウェアのコード行数などを説明変数として用いているため，示されているモデルを設計完了時に適用することはできない．

設計ドキュメントに関するプロダクトメトリクスとプロセスメトリクスの両方を用いて品質の予測を試みた研究もわずかながら存在する [Katayama2009][Takata1994]．ただし，試験仕様書に関するメトリクスについては，我々の知る限りこれまで用いられておらず，従って予測精度に寄与するのかも明らかではなかった．また，それらの研究では目的変数に単体テストの欠陥数を合算した場合と合算しない場合の予測精度についても比較していない．

片山ら [Katayama2009] は，組込みソフトウェア開発に

おいて，低品質モジュールを予測するために，基本・詳細設計での，設計ドキュメントに関するプロセス・プロダクトメトリクスを説明変数としてモデルを構築している．ただし，片山らはテストの順番を決めるために，単体テストからシステムテストまでに発見される欠陥の多いモジュールを予測に基づきランキングしており，予測モデルの使用目的が本論文と異なる．このため，欠陥数の予測精度が明らかでない．また，各メトリクスの予測に対する効果も評価していない．

高田ら [Takata1994] は単体テストからシステムテストにおいて発見される欠陥数を予測するために，片山らと同様のメトリクスを説明変数として用いている．ただし，コードレビューでの欠陥数も説明変数に含めているため，モデルを設計完了時に適用することはできない．

当然ではあるが，従来研究では本研究のリサーチクエスチョンに答えることはできない．テスト後半の欠陥数に着目し，各種のメトリクスを組み合わせて欠陥数の予測精度を評価し，それぞれのメトリクスの予測精度に対する寄与を確かめたことが本研究の主要な貢献である．

6. おわりに

本研究では，組込みソフトウェアの開発における定量的な品質管理を支援することを目的とし，設計ドキュメントに関するメトリクスを説明変数として，テスト後半の欠陥数を予測することを行った．テスト後半の欠陥数とは，結合テストとシステムテストにおいて発見される欠陥の合計である．実験では，組込みソフトウェアを開発しているある企業において収集されたデータを用いた．予測モデルの説明変数として用いたものは，仕様書の分量などのプロダクトメトリクス，レビュー工数やレビュー欠陥数などのプロセスメトリクス，試験仕様書の分量など，試験仕様書に関するメトリクスである．これらのメトリクスは基本設計時と詳細設計時に収集される．データは２つの部署から収集されており，それぞれの部署において予測モデルを作成し，精度を評価した．その結果より，以下のことがいえる．

- どちらの部署においても，リーブワンアウトにより予測した場合，*BRE* 平均値がおおむね 35% 程度となり，比較的高い精度となった．よって，表１のようなメトリクスを収集している企業では，欠陥数の予測を試みる価値があると考える．この場合，基本設計終了後に予測を試みても，ある程度精度が高いことが期待される．
- 実験結果からは，どのメトリクスを優先的に用いるべきかは一概にはいえなかった．ただし，採用された説明変数が異なっていても（表２，表３），予測精度に大きな違いはなかった．このことから，表１のようなメトリクスを一部でも収集している場合，欠陥数の予測を試みるとよいと考えられる．
- 部署によって，どのメトリクスを説明変数の候補とするとよいのかが大きく異なっていた．このことから，最初から各部署で共通の予測モデルを構築することを目指すのではなく，部署ごとに別個に構築することも検討すべきであるといえる．
- 単体テストの欠陥数を目的変数に合算すると予測精度が低下することがあった．このため，設計関連メトリクスのみを用いて予測モデルを構築する際には，単体テストの欠陥数を目的変数に合算しないほうがよいといえる．

欠陥数の値域が異なる２つの部署どちらにおいても類似の結果が得られたことから，結果の信頼性は比較的高いと考えられる．今後の課題は，設計ドキュメントのメトリクスとソフトウェアの品質に関連があるかどうかを明らかにすることである．

謝辞

本研究の一部は，文部科学省科学研究補助費（基盤 C：課題番号 25330090）による助成を受けた．

【参考文献】

[Conte1986] S. Conte, H. Dunsmore, and V. Shen, Software Engineering, Metrics and Models, Benjamin Cummings, 1986.

[Katayama2009] 片山真一 , 大蔵君治 , 伏田享平 , 川口真司 , 名倉正剛 , 門田暁人 , 飯田元 , “ ソフトウェアタ グを用いた設計文書メトリクスからの低品質モジュールの予測 ,” 電子情報通信学会技術研究報告 , Vol.109, No.343, pp.67-72, 2009.

[Kocaguneli2013] E. Kocaguneli, and T. Menzies, “Software effort models should be assessed via leave-one-out validation,” Journal of Systems and Software, Vol.86, No.7, pp. 1879-1890, 2013.

[Komuro2011] 小室睦 , 薦田憲久 , “ ピアレビューデータに基づく品質予測モデル ,” 電子情報通信学会論文誌 D, Vol.J94-D, No.2, pp.439-449, 2011.

[Little2002] R. Little, and D. Rubin, Statistical Analysis with Missing Data, 2nd ed., John Wiley & Sons, 2002.

[Mihara2013] 三原幸博 , “ 組込みソフトウェア開発における品質向上の勧め [バグ管理手法編],” http://www.ipa.go.jp/files/000030363.pdf

[Misono2007] 御園謙吉 , 良永康平 (編) , よくわかる統計学 II 経済統計編 , ミネルヴァ書房 , 2007.

[Miyazaki1994] Y. Miyazaki, M. Terakado, K. Ozaki, and H. Nozaki, “Robust Regression for Developing Software Estimation Models,” Journal of Systems and Software, Vol.27, No.1, pp.3-16, 1994.

[Mølokken-Østvold2005] K. Mølokken-Østvold, and M. Jørgensen, “A Comparison of Software Project Overruns-Flexible versus Sequential Development Models,” IEEE Transactions on Software Engineering, Vol.31, No.9, pp.754-766, 2005.

[Onodera2004] 小野寺孝義 , 山本嘉一郎 (編) , SPSS 事典：BASE 編 , ナカニシヤ出版 , 2004.

[Strike2001] K. Strike, K. El Eman, and N. Madhavji, “Software Cost Estimation with Incomplete Data,” IEEE Transactions on Software Engineering, Vol.27, No.10, pp.890-908, 2001.

[Takata1994] 高田義広 , 松本健一 , 鳥居宏次 , “ ニューラルネットを用いたソフトウェア信頼性予測モデル ,” 電子情報通信学会論文誌 D-I, Vol. J77-D-I , No.6 , pp.454-461, 1994.

[Tamaru2008] 田丸喜一郎 , “10 年で変わった「組込み」、変わらない「組込み」,” ZIPC WATCHERS, pp.6-7, 2008.

[Tabachnick1996] B. Tabachnick, and L. Fidell, Using Multivariate Statistics, 3rd Edition, Harper Collins College Publishers, 1996.

[Tanaka1995] 田中豊 , 垂水共之 (編) , Windows 版 統計解析ハンドブック 多変量解析 , 共立出版 , 1995.

[Tsunoda2009] 角田雅照 , 玉田春昭 , 森崎修司 , 松村知 子 , 黒崎章 , 松本健一 , “ コードレビュー指摘密度を用 いたソフトウェア欠陥密度予測 ,” 情報処理学会論文誌 , Vol.50, No.3, pp.1144-1155, 2009.