

# Predicting Coding Effort in Projects Containing XML

S. Karus

University of Zurich, Switzerland  
University of Tartu, Estonia  
siim.karus@ut.ee

M. Dumas

University of Tartu  
Estonia  
marlon.dumas@ut.ee

**Abstract**—This paper studies the problem of predicting the coding effort for a subsequent year of development by analysing metrics extracted from project repositories, with an emphasis on projects containing XML code. The study considers thirteen open source projects and applies machine learning algorithms to generate models to predict one-year coding effort, measured in terms of lines of code added, modified and deleted. Both organisational and code metrics associated to revisions are taken into account. The results show that coding effort is highly determined by the expertise of developers while source code metrics have little effect on improving the accuracy of estimations of coding effort. The study also shows that models trained on one project are unreliable at estimating effort in other projects.

**Keywords**-XML; XSLT; metrics; coding effort; estimation

## I. INTRODUCTION

Estimating the cost of software projects is a long-standing research problem in the software engineering field [1]. It is both highly challenging from a research perspective and highly relevant from a practical perspective. Indeed, misestimates of software project costs can result in high losses due to overdue and over-budget projects [2, 3].

It has been shown that expert judgement is highly inconsistent and thus not a reliable means of software cost estimation [4]. Accordingly, a number of algorithmic or statistical models for software cost estimation have emerged. In general, these models focus on estimating development and maintenance effort since human resources are the source of a significant portion of software development costs.

We can broadly classify development effort estimation models into *a priori models* and *evolutionary models*. The former category includes models like COCOMO [3], which aim at predicting effort based on organisational features that are usually known during the analysis and design phases, such as expected number of function points, size, experience and cohesion of the team, software processes and tools employed, etc. A key characteristic of these models is that they do not take into account the code base and history of the project at the time the prediction is made. In contrast, evolutionary models aim at predicting future development and maintenance effort of a project, by using metrics extracted from the project's code base, version control system, bug tracking system and other collaboration systems [5, 6, 7, 8, 9]. In other words, these methods take into account the intermediary snapshots and evolution of the software project. The study reported in this paper falls in this latter category, and focuses specifically on the use of metrics extracted from version control systems.

Another way of classifying effort estimation models is based on the measure that they aim at predicting. Models in the style of COCOMO aim at predicting man-hours. Of course, this is highly dependent on the past experience of developers in previous projects and many other organisational parameters that are exogenous to the project in question. Due to these exogenous factors, the predictive power of these models can vary across projects and careful calibration is required [3]. In contrast, other models tend to predict a measure that can be directly derived from the software product itself: namely *code churn* [10]: the sum of lines of code (LOC) added, removed or modified during a certain timeframe. While code churn does not directly measure total development or maintenance costs, it is nonetheless an indicator of coding effort during a given timeframe of a project. Additionally, code churn is an indirect indicator of the cost of design, testing, deployment and project management. The more concise are the modifications and additions, the less code is there to test, and deploy. Also, a project with a reduced number of small modifications is indicative of less major re-design decisions, compared to a project with large and frequent rewrites. Thus, estimating code churn can give an insight into total development effort – be it during the initial development phase or during the maintenance phase of a project. Furthermore, code churn has been shown to be a predictor of defects in software projects [11]. The present study deals with the prediction of three code churn metrics separately: number of added LOC, modified LOC and deleted LOC over a one-year period, which corresponds to the length of a typical long-term software development planning period and is in line with the choice of timeframe made in related studies on long-term code churn [12, 13].

Most *evolutionary* coding effort estimation models assume that coding effort can be estimated based on software design metrics (e.g. coupling and cohesion metrics [8]), or code metrics (e.g. code complexity metrics [5]). Relatively little attention has been given to estimating coding effort based on organisational metrics extracted from version control systems, such as developers' activity. In this respect, a recent study has shown high correlation between organisational metrics and code churn metrics in the development of the Windows operating system [6].

This and similar studies suggest that there is a wealth of information available in version control systems – beyond the code itself – that can be used to predict code churn.

In evolutionary estimation models, information extracted from the current snapshot and evolution of the project is used to predict future coding effort. Many techniques can be used for this purpose. In this paper, we rely on standard machine

learning techniques. These techniques require training prior to their use. The training is done on a subset of snapshots, and the trained models are used to predict one-year code churn on other snapshots. A question that arises in this setting is whether models trained on snapshots from a project can be used to make predictions on other projects.

In light of the above, this paper addresses the following general research questions:

- RQ1. Can coding effort for the subsequent year, measured in terms of added LOC, deleted LOC, and modified LOC, be estimated based on only organisational or code metrics extracted from version control systems?
- RQ2. Can a model trained on one project be used to make accurate predictions on other projects (i.e. are the main factors determining the future coding effort the same across multiple projects)?
- RQ3. Are the relationships between coding effort and organisational metrics fundamentally different from one project to another, or can a single unified model be used to make estimations of added, modified and removed LOC across multiple projects?

The bulk of previous research addressing the question of coding effort estimation in general and evolutionary effort estimation in particular, deal with code written in procedural and object-oriented programming languages. In the meantime, the rapid uptake of the Web has resulted in more and more software systems containing code written using the eXtensible Markup Language (XML) code and associated languages such as the Extensible Stylesheet Language Transformations (XSLT). XML-based languages are used to encode, among other things, build and deployment information (e.g. Ant), configuration information, application data and data schemas (XML Schema), document transformations (XSLT), images (SVG) and other software artefacts pertaining to the presentation layer of Web applications [14]. Statistics extracted from Ohloh.net<sup>1</sup> show that about 20% of open-source projects make use of XML, while less than 10% make use of Java (and the same applies to C). Yet, there are only a few studies that attempt to predict coding effort from data associated with XML files in project repositories.

Accordingly, this paper addresses the above research questions in the context of software projects that make use of XML. The underpinning hypothesis is that XML files in such software projects contain valuable information that can be used to build accurate code churn prediction models. In particular, the paper considers the use of machine learning algorithms to build code churn prediction models based on code and organisational metrics extracted from XML files, combined with organisational features extracted from the project's version control system

This paper starts with an overview of related work (Section II), followed by a description of the method (Section III),

---

<sup>1</sup> Language usage values were taken on 15th August 2010 from the open-source projects database <http://ohloh.net>, which tracks more than 437,000 open-source projects Worldwide.

dataset (Section IV), metrics (Section V), and algorithms (Section VI) employed for creating churn estimation models. The experimental results are discussed in Section VII. Finally, threats to validity are reviewed in Section VIII while conclusions and possible future directions are outlined in Section IX.

## II. RELATED WORK

Software development effort is commonly measured in man-hours, project cost, or code churn. The first two can only be used when studying commercial software projects. In contrast, open-source public (especially community-driven) projects do not capture data about development time or project cost. The only data available for analysis of open-source projects has to come from version control systems, mailing lists, and bug tracking systems. These sources give information about developer past activity, which can be incorporated into development effort estimation models to balance the lack of other information in case of open-source software [7]. The work reported here is based on open-source projects and accordingly, it aims at predicting coding effort in terms of product characteristics (churn metrics) rather than man-hours.

The data used for estimation also differs between studies [15]. Commonly used input data for estimation includes: analysis documents, source code/design metrics, expert judgement, and organisational metrics. Most algorithmic approaches base their estimations on either organisational metrics or code metrics but not on both. This paper studies the question of whether combining these two approaches would result in better estimations.

Zhou, et al. [8] used linear regression to investigate relations between object-oriented design metrics and maintainability in open source Java projects. This study is representative of a body of studies aimed at testing hypotheses such as “low coupling between classes in an object-oriented software leads to better maintainability measured in terms of amount of changes during the maintenance phase” [12]. Similar studies consider the use of procedural code complexity metrics to predict maintainability [5]. These previous studies focus on code metrics, while our study combines organisational and code metrics, and focuses on projects containing XML code. In a separate study of 8 open source software projects, we have established that code churn can be estimated with high accuracy using models built on organisational metrics [16]. In the present study, we aim at predicting individually each of the three components of code churn (LOC added, deleted and modified) in an attempt to better understand the drivers of coding effort. Also, the present study is more focused on the role of XML for coding effort prediction.

Nagappan, et al. [6] used organisational metrics to estimate code defects in the development of the Windows operating system. This study addresses research questions similar to RQ1. However, the organisational metrics considered by Nagappan, et al. are only available in commercial software projects. Our study is restricted to metrics that can be extracted from a project's source code management system, and therefore the study can be applied both in commercial and open-source projects. That is, we extend the idea of using organisational metrics for evolutionary effort estimation to a broader range of software projects.

TABLE I. DETAILS ON PROJECTS USED IN THE STUDY (REVS – # OF FILE REVISIONS, FILES – # OF FILES, DEVS – # OF DEVELOPERS).

Project	Common files	Revs	Files	Devs	Years
<b>Commons</b> http://www.wso2.org/	no ext.&java (31%), <b>XML (11%)</b>	1711	1517	37	3
<b>Dia</b> http://www.gnome.org/projects/dia/	shape (28%), xpm (14%), c&png (13%)	3571	632	139	11
<b>Docbook</b> http://docbook.sourceforge.net/	<b>XML (34%)</b> , gen (25%), no ext. (14%)	7136	3028	28	8
<b>Docbook2X</b> http://docbook2x.sourceforge.net/	<b>XSL (36%)</b> , <b>XML (17%)</b> , no ext. (11%)	1081	159	2	8
<b>Esb</b> http://www.wso2.org/	<b>XML (27%)</b> , java (18%), <b>XSL (14%)</b>	1070	546	10	3
<b>eXist</b> http://exist.sourceforge.net/	java (60%), <b>XML (7%)</b> , jar (6%)	6300	3329	31	7
<b>feedparser-read-only</b> http://www.feedparser.org/	<b>XML (78%)</b> , HTML (19%), no ext. (1%)	227	1287	4	3
<b>Gnome-doc-utils</b> http://live.gnome.org/GnomeDocUtils	<b>XML (23%)</b> , po (19%), <b>XSL&amp; no ext. (16%)</b>	968	121	115	5
<b>Groovy</b> http://groovy.codehaus.org/	java (40%), groovy (29%), jar (7%)	6920	2549	53	7
<b>Tei</b> http://tei.sourceforge.net/	odd (41%), <b>XML (34%)</b> , <b>XSL (12%)</b>	4318	1633	12	6
<b>Valgrind</b> http://valgrind.org/	c (25%), exp (19%), vgtest (14%)	4090	1000	19	2
<b>Wsas</b> http://www.wso2.org/	java (31%), <b>XML (18%)</b> , no ext. (17%)	1180	823	22	3
<b>Wsf</b> http://www.wso2.org/	c (15%), h (14%), no ext. (13%)	1405	2285	21	3

In a similar vein, Pendharkar, et al [9] studied the relation between team size and software development cost (including initial development and maintenance). They uncovered a significant correlation between “active” team size and coding effort measured in terms of added, modified and deleted lines of code. Our study follows a similar line but considers a broader set of metrics.

### III. METHOD

The aim of the study is to determine if it is possible to build statistical models to predict the components of future code churn of a project (specifically added LOC, deleted LOC, and modified LOC) based on metrics extracted from version control systems. Given this aim, we have a choice between hypothesising that certain relations exist between a set of input metrics and the above three code churn metrics, or to uncover such relations using exploratory analysis. In the first approach we would start with a set of hypothesis, and we would use statistical conformance testing to validate these hypotheses on the chosen dataset. However, as mentioned above, we are not aware of previous studies on possible relations between code/organisational metrics and code churn components in the context of projects containing XML code. Hence, there is little basis for formulating a priori hypotheses about such relations.

Accordingly, we adopt a bottom-up approach based on data mining and exploratory data analysis.

The adopted data mining approach comprises the following steps:

1. Data pre-processing: choice of prediction targets and proposition of input features (attributes that might influence the value we need to predict), data gathering, normalisation, and cleansing.
2. Learning: choice of data mining algorithms and application of these algorithms.
3. Results validation: evaluation of models fit using standard statistical techniques.

This data mining approach allows us to identify interference between input features and the predicted variable and in doing so they uncover the existence of a predictive model. However, the data mining approach itself does not allow us to explain the cause of the interference. In order to compensate for this shortcoming, exploratory data analysis was used in order to gain an understanding of the models created by the data mining algorithms.

Compared to a conformance testing approach, data mining and exploratory data analysis offer the benefit of not requiring

an a priori specific model to test. The aim of exploratory data analysis is to propose models that can then be conformance-tested. Moreover, data mining and exploratory data analysis can uncover non-intuitive relations. In fact, the results of this study show, among other things, that there are no generally-applicable straightforward models to estimate the components of code churn – that is, linear models based on one or a very small number of interactions between input features. The models with better predictive power uncovered in the study involve a non-trivial number of interactions between input features.

#### IV. DATASET

Eight Open Source Software Project repositories with XSLT and XML code were used to train the models:

- WSO2 commons
- WSO2 Wsas
- WSO2 Esb
- Docbook
- Docbook2X
- Exist
- Gnome-doc-utils
- Tei

Additionally, in order to achieve higher level of data separation, five projects were used only for testing:

- Feedparser-read-only
- WSO2 wsf
- Dia
- Groovy
- Valgrind

These projects were chosen so that they would represent different types of software systems. For example, “WSO2” is an enterprise service bus type of a project, “docbook”, “docbook2X” and “gnome-doc-utils” are used for documentation formatting, “dia” for graph drawing and “eXist”, “tei”, “groovy” and “valgrind” are software development tools. The projects also varied in the use of languages, team size, and project maturity. More than 118,000 file revisions (of more than 24,000 files) were used for data preparation, analysis, model creation, and testing. It is noteworthy that project “gnome-doc-utils” did not have any lines of code modified or removed and did not therefore produce models for predicting modified or removed LOC. An overview of the project repositories used is given in Table I. The number of revisions (Revs) is the number of revisions to project (commits made to the project). Thus, if multiple files are changed in one commit, the number of files in a project can be higher than the number of revisions, which is the case with “feedparser-read-only”.

A common term for long-term code churn prediction is a year [12, 13], which is used in this study as well. Added, modified and removed lines of code for each file revision was calculated based on GNU diff output. Yearly modified, added and removed LOC are considered to be the sums of corresponding file revision operation measurements during the year following the date of commit of the revision (“cumulative yearly added LOC”, “cumulative yearly deleted LOC”, and “cumulative yearly modified LOC”). For the dataset, the cumulative yearly added LOC ranges from 0 to 1,850,000,

modified LOC ranged from 0 to 680,000, and removed LOC ranged from 0 to 820,000 LOC.

As XML is also used for storing project information (e.g. ant build scripts), we did have a look into the types of XML files used in the projects. It turned out that out of all XML files 13% were project definitions and ant build files while the rest were mostly project specific files (that is, they used namespaces defined in the project itself)<sup>2</sup>. All projects had XML files that were neither build nor project definition files. For more general study of languages used in open source projects, we refer to a separate study [14], which shows that there are no major XML based file formats as majority of XML used in the projects is project specific. Thus, it makes sense to treat all XML file similar to each other just like we treat C/C++ and Java files similarly.

The projects studied had life spans from 2 to 11 years and some of these are still active. The data was collected during spring-summer of 2009. For each project, all revisions over the entire project’s lifetime were extracted. We could have partitioned the dataset into one-year periods (e.g. calendar years). However, this would have reduced the amount of one-year periods available for training and validation of the algorithms. Instead, we decided to compute for each project revision, the code churn over the one-year period starting from the date of the commit of the revision. In this way, there were as many predictions as the number of commits. Of course, it was not possible to compute one-year-forward code churn metrics for revisions with commit dates within one year of the date of extraction of the dataset, so these revisions were not included in the set of revisions for training and validation of the learning algorithms (though they were used to calculate one-year-forward code churn of previous revisions).

#### V. FEATURE SELECTION

In order to train models for estimation, features to base the estimations on, need to be selected. In our study, we selected two sets of features independently: organisational metrics extracted from version control systems, and code metrics extracted from XML and XSLT files in a given snapshot of the project.

##### A. Organisational features

Organisational features describe the project team and the team’s familiarity with some of the most popular technologies and languages. In addition, project size (in number of files) and project age/maturity (in revisions) were taken into account. Current and historical number of file extensions was used as an indicator of technologies used in the project. The committers’ previous activities were also included in the feature set as it can be assumed that the developer making the commit is still active whilst others might no longer be active participants.

For each file revision in the version control system, the following metrics were calculated:

---

<sup>2</sup> We identified project files by checking for ant, maven, netbeans, or eclipse namespaces in XML files or the presence of common project definition file structure (e.g. root element named ‘project’).

- Number of developers, who have made commits to version control system before and including the time of the commit of the revision at hand
- Number of developers, who have made commits to XSL, XML, Java, HTML, C, or Graphics files in version control system before and including the time of the commit of the revision at hand (6 metrics)
- Number of previous commits made by the committing developer
- Number of previous commits made by the committing developer to XSL files
- Number of previous commits made to project
- Current number of files in project
- Current number of different file extensions used in the project
- Number of all file extensions used in the project to date
- Cumulative yearly added LOC, modified LOC, and deleted LOC (3 prediction targets, used only for validation and training)

The number of developers by area of expertise was calculated by counting committers who made commits to files with the corresponding file extension. For example, a Java developer is defined as one who has made at least one commit to a .java file, while a C developer is one who has made at least one commit to a .c or .h file, a Graphics developer is one with commit to .gif, .png or .jpg file, and an XML/XSL/HTML developer is one with commit to .xml, .xsl (or .xslt) or .html (or .htm) file, respectively. We only considered five areas of expertise (Java, C, Graphics, XML, XSL and HTML) because these were the only file types that appeared in abundant amounts in the studied data set.

### B. XML Code features

Maintainability estimation studies on Object-Oriented and structural languages commonly use lines of code, number of classes, inheritance metrics and different program flow based complexity metrics. XML and XSLT, however, have no classes nor inheritance. Additionally, program flow graphs of XSL transformations are not as meaningful as program flow graphs on procedural languages due to the non-procedural nature and path matching satisfiability complexities of XSLT [17]. Thus, new metrics need to be considered for XML and XSLT.

On the other hand, because of its structure, it is straightforward to define “count metrics” on XML (e.g. number of elements and number of attributes). Furthermore, elements have different types (i.e. names) and thus we can potentially define one count metric per element. However, the set of possible XML element or attribute names is infinite, making the definition of one count metric per possible type of XML element impractical. Thus, we chose to count only occurrences of each type of XSLT 1.0 element, as well as the total number of elements and attributes in an XML file not belonging to the XSLT namespace. We did not count occurrences of elements that can be present only once per transformation (e.g. <transformation>, <stylesheet>, <output>). XML metrics were extracted from all XML based files. That is, XML metrics were counted also from XSL files, config files, shape files (in project “dia”), xsd files and any other files, which could be parsed as XML.

XSLT uses XPath expressions for matching, selecting and testing input data. These expressions can only be present in the XSLT attributes “match”, “select”, “test” and in attributes outside the XSLT namespace enclosed between curly braces (these are called “inline expressions”). Inline XPath expressions can only be used for selecting data. As these expressions denote decision points in XSLT, counting them gives us information on the flow of transformations.

It is important to differentiate between two types of XPath expressions: simple expressions identifying specific elements or attributes by their name and namespace versus complex expressions identifying wider ranges of nodes in input documents. Complex expressions can be written using wildcards or using function calls. Simple expressions are those that do not contain neither a wildcard nor a function call. By counting separately simple and complex expressions we can get an indication of the complexity of the transformation.

In total, 61 primary metrics (also called *features*) were collected for each project snapshot, that is, for each commit found in each project repository. The 61 metrics include:

1. count of XML nodes,
2. count of different types of XML nodes (4 metrics),
3. count of each type of XSLT element (28 metrics),
4. count of XSL output literals,
5. count of elements in XSL target namespace,
6. count of direct children of the root element,
7. count of XSL global params and variables (2 metrics),
8. count of inline expressions,
9. count of XSL attributes that contain test expressions (“select”, “match”, and “test”) for each expression type (“simple”, “complex due to wildcard”, and “complex due to function call”) and in total (12 metrics),
10. average number of child elements of XSLT “message” elements,
11. sum of XML attribute and element type nodes,
12. count of attributes and elements inside XSLT “param”, “variable”, and “message” elements (3 metrics),
13. count of XSLT output attributes and elements (sum of XSLT “element” elements, XSLT “attribute” elements, and XML elements and attributes in target namespace,
14. count of “complex” expressions by attribute (“select”, “match”, “test”) and total number of complex expressions in the file (4 metrics).

## VI. TRAINING ALGORITHMS

Two different algorithms were used for training models for coding effort estimation:

- Neural Networks (NN) – Back-Propagated Delta rule network with three layers [18]
- Decision Trees (DT) with regression equations on the leaves [19]

The neural networks algorithm is capable of identifying various complex relations between input features and prediction target. Due to this ability, NN is commonly used to build models for estimation of features that have complex relations with other features. As such, NN can be used to build highly accurate estimation models in cases where relations between features are not known well or involve many interactions

TABLE II. MEAN AND MEDIAN CORRELATIONS AND NORMALISED MAE AND RMSD FOR MODELS TRAINED ONLY ON ORGANISATIONAL METRICS.

	Operation	Pearson		Kendall		NMAE		NRMSD	
		Mean	Median	Mean	Median	Mean	Median	Mean	Median
DT	Added	0.9101	0.9216	0.7308	0.7064	0.1509	0.1078	0.1050	0.0964
	Removed	0.9301	0.9443	0.7759	0.7978	0.1556	0.1073	0.1051	0.1026
	Modified	0.9287	0.9420	0.7877	0.7785	0.1471	0.0939	0.1008	0.1007
NN	Added	0.8169	0.9147	0.6695	0.7140	0.0336	0.0146	0.1526	0.1192
	Removed	0.9106	0.9502	0.7253	0.8048	0.0329	0.0172	0.1336	0.1233
	Modified	0.9147	0.9502	0.7297	0.7424	0.0354	0.0199	0.1332	0.1237

between features. On the other hand, the performance of NN algorithm and the models built using the algorithm is highly dependent on the amount of training data available.

Decision Trees allow one to build simpler and more explainable models. This also makes models built by Decision Trees algorithm easier to infer and interpret. The performance of the algorithm and models built by it are also less dependent on the amount of training data used. However, Decision Trees algorithm cannot handle complex relations well. Thus, Decision Trees (with regression equations) and Neural Networks are somewhat complementary. This is confirmed by a study by MacDonell & Gray [20] who empirically tested various learning techniques and found linear regression (with removal of outliers) and Neural Networks the best learning methods for estimations on software projects.

Microsoft SQL Server 2008 Analysis Services was used to train and test the models. The dataset was split into training (70%) and test set (30%) using random sampling and the split between projects outlined in section IV. Most parameters of the algorithms were left to default values. The only differences were made for Decision Trees algorithm where “Complexity penalty” and “Minimum support” were set to 0.99 and 0.2 accordingly. The parameter values were chosen after experimenting with different values and evaluating the performance of the resulting models.

To aid the Neural Networks algorithm, added, removed and modified LOC values were normalised for training and testing of Neural Networks trained models. Cross-validation (3- and 4-fold) on subsets of up to 7000 entities was used to verify that the algorithms behaved consistently with low chance of overfitting (identification of random patterns). Cross-validation on full dataset or with more folds (e.g. 10-fold) failed due to database limitations.

## VII. RESULTS

In order to evaluate the predictive power of the models, we computed the following statistical validation measures after applying each model to the testing data set:

- Pearson Correlation coefficient – indicates linear correlation between actual and predicted values (higher value is better).
- Kendall Correlation coefficient – indicates rank based correlation between actual and predicted values (higher value is better).
- Mean Absolute Error (MAE) – Mean difference between actual and predicted values (lower value is better).

- Normalised Mean Absolute Error (NMAE) – Normalised MAE (MAE divided by the mean of actual values). One can think of NMAE as the average error ratio. A model estimating constant 0 would have NMAE 1.0.
- Root Mean Square Deviation (RMSD) – measures differences between predicted values and actual values. Compared to MAE, RMSD is more sensitive to high errors (lower value is better).
- Normalised Root Mean Square Deviation (NRMSD) – min-max normalised RMSD.

In the following discussion only normalised values are mentioned as absolute values are not comparable between projects and algorithms used.

It should be noted that the predicted variable (churn) belongs to infinite countable set. Thus, measures of accuracy such as precision, recall, F-score or information gain are not applicable. On the other hand, we could have included Bias and Mean Absolute Percent among the validation measures, but these two measures are influenced by the distribution of the predicted variable.

### A. RQ1: Project-Specific Models

The first research question is whether we can construct prediction models based on organisational metrics and code metrics extracted from version control systems. To this end, we constructed models for each project using organisational metrics alone, code metrics alone and a combination of both.

The means and medians of validation measures for models trained on organisational metrics alone are given in Table II. This table shows that the generated models have a reasonable ability to estimate coding effort. Most models have high predictive power, however, the mean values are penalised by the low predictive power of some models (as shown by significantly better median values in most cases). Specifically, models trained on “docbook2X” project had rather low predictive power both when using Decision Trees and Neural Networks (e.g. NMAE of around 0.28 using Decision Trees and around 0.08 using Neural Networks). This may be explained by the fact that “docbook2X” had only two developers.

For models trained on both organisational and code metrics, the validation measures show a slight but noticeable difference in predictive power on average – improving the performance in some cases, but deteriorating it in other cases. For models trained using Decision Trees on all metrics, NMAE values decreased by 0.0549 on average and NRMSD values decreased by 0.0327 on average compared to models trained only on organisational metrics. This is a relative improvement of around

TABLE III. MEAN AND MEDIAN CORRELATIONS AND NORMALISED MAE AND RMSD FOR MODELS TRAINED ON ALL METRICS.

	Operation	Pearson		Kendall		NMAE		NRMSD	
		Mean	Benefit	Mean	Benefit	Mean	Benefit	Mean	Benefit
DT	Added	0.9540	+5%	0.8225	+13%	0.1007	+33%	0.0727	+31%
	Removed	0.9665	+4%	0.8166	+5%	0.0898	+42%	0.0632	+40%
	Modified	0.9499	+2%	0.8206	+4%	0.0984	+33%	0.0769	+24%
NN	Added	0.8330	+2%	0.6546	-2%	0.0313	+7%	0.1321	+13%
	Removed	0.9267	+2%	0.6876	-5%	0.0250	+24%	0.1150	+14%
	Modified	0.9286	+2%	0.6930	-5%	0.0280	+21%	0.1321	+1%

30%. The other performance metrics experienced a modest improvement. However, estimations of models trained on the “eXist” project clearly improved when code metrics were excluded from the training of the Decision Trees models.

On the other hand, models trained using Neural Networks had consistently better NMAE and NMRSD values when code metrics were included. This can be explained by the fact that, in the majority of the cases, Neural Networks models had lower performance when trained on only organisational metrics compared to corresponding models trained using Decision Trees algorithm. It is also noteworthy that with low amounts of training data, the inclusion of code metrics improved the accuracy of Neural Networks models more than it did when high amounts (more than 2500 training cases) of training data was employed.

Some models trained on all metrics (organizational plus code metrics) displayed an interesting characteristic of identifying organisational metrics either as having low-influence dependencies or sometimes ignoring them completely. Still, the differences in terms of validation metrics compared to models built on only organizational metrics were small. For example, models trained using Decision Trees on project “esb” had Pearson correlation improve from 0.96 to 0.99, Kendall correlation improve from 0.89 to 0.94, NMAE improve from 0.06 to 0.02 and NMRSD improve from 0.08 to 0.05 when code metrics were included during training (reducing error at 90% confidence by 2-7%). Similarly models trained on project “tei” improved from Pearson correlations from 0.92 to 0.96, Kendall correlations from 0.75 to 0.84, NMAE from 0.13 to 0.08, NRMSD from 0.12 to 0.09 (reducing error at 90% confidence by 2-6%). This could be caused by correlations between organisational and code metrics.

*Overall, the results show that organisational features, particularly number of developers, are important drivers of coding effort. Thus, it is possible to train accurate models for coding effort estimation based only on organisational features. Nonetheless, the inclusion of XML code metrics noticeably improves the predictive power of the models in most cases. As shown in Table III, the relative improvement in NMAE for models trained on all metrics, compared to models trained on organisational metrics only, ranged from 7% to 42%.*

#### B. RQ2: Applying Models across Projects

In order to address RQ2, we tested whether models trained on only a given project can give good estimations on other projects.

The test results show that models trained on single project are not generalisable. Estimations on all the projects did not

have any predictive power – the correlations with actual values were close to zero or even negative<sup>3</sup>. Only the models trained on project “eXist” using Decision Trees algorithm and models trained on project “docbook” with Neural Networks algorithm showed consistently low positive correlations (Pearson ranged from 0.1073 to 0.3364 and Kendall from 0.1990 to 0.3365).

It is interesting to note that model trained on project “tei” made estimations with strong negative correlations on project “esb” – Pearson less than -0.70, Kendall less than -0.15 (-0.69 in case of “Added LOC”). This strong distinction implies that these projects have distinctly different drivers for coding effort. Figure 1 shows the actual decision trees produced by training on “esb” and “tei” projects. It can be seen that despite the dependencies for modified and removed LOC are similar, the way these dependencies affect modified and removed LOC is very different. For one, the number of developers on project has negative effect on modified and removed LOC in project “esb”, but positive effect in project “tei”. In fact, while studying the models created by Decision Trees algorithm, we were not able to find any feature that had the same kind of effect (either always positive or always negative) in all projects. Thus, the features studied are not suitable for generalised models on their own, implying the need to include additional features or non-linear and non-sigmoid transformations of features or analysis of interactions between features in order to build models that could explain drivers of coding effort for all projects in general.

Whilst models trained on all features performed slightly better, the errors in models trained using Decision Trees algorithm were still too high to make the models useful on projects other than the one used for training.

The fact that having more developers could reduce the coding effort (number of modified, added or removed lines of code) intrigued us as it is highly counter-intuitive. We made an additional test to see whether high developer turnaround could have caused this phenomenon by counting the number of active developers (developers who have made commits in the past and will make future commits during the period studied) in addition to all developers. The test showed that even though the number of active developers was considered an important influencer for some models, the number of total developers had always similar influence (the multipliers were differed less than 5%). Thus, high developer turnaround was not confirmed as the cause of the phenomenon.

<sup>3</sup> Even though negative correlations nearing -1 are strong, they are useless in this context as negative correlations were not consistent nor expected.



Figure 1. Models trained using Decision Trees algorithm on projects "esb" and "te" on only organisational metrics.

Another possible cause of the phenomenon is more diverse code ownership. That is, developers are reluctant to modify other developer's source code or the diversity of developers makes the source code more difficult to comprehend. This hypothesis was not tested as part of this study and would still need to be put on test.

Finally, we trained models using Neural Networks on a given project and tested them on all other projects. Models trained using neural networks predicted code churn on other projects better than models trained using Decision Trees, occasionally having Pearson correlation higher than 0.9, Kendall correlation higher than 0.8, NMAE lower than 0.05 and NRMSD lower than 0.2 for estimations on other projects. These occasions were exceptions and in the vast majority of cases the models did not have significant predictive power on projects they were not trained on.

Thus, we conclude that models trained on the repository of one project are generally weak when used to predict code churn on other projects. Models trained using Decision Trees algorithm are not usable for estimations on projects other than the project the model was trained on.

### C. RQ3: Existence of a Unified Model

To test the viability of using a single model, as opposed to training models on a project-specific basis, we trained and tested models using the combined data from all the projects using a 70:30 random split. Models trained using Decision Trees had a reasonable Pearson correlation (higher than 0.69) and NRMSD (lower than 0.14). However, NMAE values of higher than 0.4372 showed that the model was not reliable for

general purpose. Models trained using Neural Networks algorithm were significantly better having Pearson correlations from 0.7131 to 0.7431; NMAE values from 0.0802 to 0.1222; and NRMSD values from 0.0893 to 0.1278. However, Kendall correlations were weak: less than 0.4674. Thus, *these models can be used with some success for estimations on multiple projects*. Per project analysis confirms that models trained using Neural Networks algorithm are good at making estimations on projects, which had the most influence during training due to their higher number of training cases.

As sub-projects belonging to one larger project are likely to have the same influencers, it is of some interest to see, whether models trained on complex project "WSO2" subprojects "common", "esb" and "wsas" are usable on each other. It turns out they are not usable for predictions on each other as NMAE value for estimations on other projects is around 1 or even higher, meaning that the range of mean error for these predictions is close to or higher than the range of actual values.

We also established that similar to models trained on only one project, *models trained using Decision Trees algorithm on multiple projects are not generalisable*. However, we also noted that *models trained using Neural Networks algorithm on multiple projects do have limited predictive power*.

## VIII. THREATS TO VALIDITY

### A. External threats

One external threat to our studies is their generalisability. This threat only affects RQ1, which in its nature is internally focused. It is possible that there are projects for which coding

effort estimation models cannot be successfully trained using the algorithms discussed in this study. However, all of the cases studied (some on more than 100 revisions) did yield statistically significant and useful models. As it has been suggested that 4-10 samples are usually enough to find any counter-examples [21], the likeliness of a project not having a coding effort estimation model based on organisational features is low. The diversity of the projects studied and the models achieved support the generalisability of the conclusions. There might still be some common nominators among the projects, which are the cause of the presence of these relatively simple relations.

For one, all the projects studied were open source non-commercial projects, which may differ substantially from closed source commercial projects, which might display other characteristics and be less influenced by the organisational features of the project. Another common threat among the projects studied was their rather high amount of XML and XSL code, which would definitely influence the models, especially as many of the features studied were XSL specific. Nevertheless, similar metrics can be constructed for other technologies. The main conclusion that organisational metrics are good input features for models estimating coding effort in XML rich open source non-commercial projects is not influenced by these threats, but generalisability to larger set of projects would need further study even though the implied presence of similar models in other open source non-commercial projects is strong.

The substantial differences between the models trained on projects ‘esb’ and ‘tei’ (discussed above), do raise the question of generalisability of case studies conducted on individual software projects. This is especially an issue for estimation-related case studies where generalisability is often overlooked as a secondary issue. This study demonstrated a lack of cross-project generalisability of models based on software metrics alone as well as models based on code metrics plus organisational metrics.

### B. Internal threats

The study of models generated is subject to the threat of not having considered all possible influencers and being missing some input features or feature transformations or interactions that could be the key to creating a single unified model that could be applicable on all projects with very good accuracy. Nevertheless, the study of actual models shows that even if there are some features independent of the features studied, linear models would be impossible even with the inclusion of them as regressions with opposite correlations with features were identified for different feature value ranges. As such, one would need to apply transformations on the features or take into account the interactions between different features in attempt to create unified models that could be successfully used for multiple projects. This would, however require the use of more complex training algorithms or additional data pre-processing.

One might also consider the option that code features and organisational features are correlating or derivable from each other. Even if this were the case, it would not discredit using models trained on all features as reference models as the assumption of the independence of code and organisational metrics is never assumed for that model.

When doing cross-project predictions (in the context of RQ2), one could think about the possibility that different projects might have values in different scales, which might affect the accuracy of applying a model trained in one model to another project. In some cases, this type of problem can be addressed by normalizing the data using Z-scores. However, this option is applicable if the data is normally distributed, which is not the case in the data sets in this study. We could not use min-max normalization because of the presence of extreme values in the data set (i.e. some commits with very large number of LOC). Accordingly, we adopted the best-effort approach of using absolute values.

Finally, one might consider the possibility that during the early phases of a project, there may be some irregularities in the values of the metrics that might affect the predictions obtained by the models. In order to check the likelihood of this threat, we conducted a test where we split the data set according to revision number. Specifically, we constructed a data set containing the first 120 revisions of all projects. We then repeated the whole procedure for constructing and testing the models using this “early-revisions data set” and using the complementing data set (revisions above 120). The performance of the models was comparable across these two data sets.

### C. Construct validity

The algorithms used in the study are not capable of identifying complex regressions – Decision Trees algorithm only identifies linear regressions and Neural Networks algorithm only combines sigmoid functions. Thus, it is possible there are relations not identified that could yield models with greater accuracy and generalisability. This does not influence the conclusions that the algorithms studied do not generate generalisable models, which has been shown by presenting counter examples (projects with contradicting models).

Pearson’s correlation coefficient is generally used under the assumption of normal distribution of values, which does not necessarily hold on the datasets in the study. However, only the general form and interpretation of Pearson’s correlation coefficient for testing correlation between two value sets is used in the study. We also make use of Kendall correlation, which is not affected by the distribution of values – only their order is relevant for the computation and interpretation.

Models with low error can have low Kendall correlation in case of value distributions with strong peaks. That is why MAE and RMSD are used along with Kendall and Pearson correlations to validate the models’ predictive power. High Kendall correlation with high MAE and RMSD is a sign that there exists a function of the predicted value, which gives very accurate estimations of actual values (i.e. there is a missing step in the model).

## IX. CONCLUSIONS AND FUTURE WORK

We have shown that, in the context of projects with XML code, Decision Trees can be successfully used to train models to predict coding effort for the subsequent year of a project based on organisational metrics extracted from version control systems (RQ1). The inclusion of code metrics does not always improve the performance of models. The results also show that different projects have different drivers of coding effort as

models trained for specific projects perform poorly when used to predict coding effort on other projects. Thus, a model trained on one specific project cannot be used for estimations on other projects (RQ2). Even models trained on data from all projects performed poorly when making effort predictions on some of the projects, showing that a single uniform model cannot be created based on the organisational metrics used in this study (RQ3).

A more in-depth study of project-specific models aimed at identifying predictive patterns and project characteristics that affect the predictive power of these models, is a possible avenue for future work. Such a study would lead to a deeper understanding of the drivers of coding effort and possibly to the identification of early signs of code maintenance and extensibility issues. It would be of special interest to reverse the models in order to give models for optimising software project team composition and/or choice of platform and technologies used or the planned software feature-set.

Another avenue for future work is the construction of additional features to be given as input to the models. For example Smith, et al. made use of genetic algorithms to construct and select features [22]. Their approach could be used to successfully construct complex features that might lead to better models by solving and explaining non-linear relations. This could also lead to the discovery of some of the general rules and practices that affect coding effort.

It might be possible to extract additional metrics from the project snapshots. Also, additional features could be extracted from other sources, such as bug tracking systems. Those additional features might prove to have significant influence on coding effort. Identifying metrics with greater influence can result in significantly improved models.

## X. ACKNOWLEDGEMENTS

This research was started during a visit of the first author to the software evolution and architecture lab at University of Zurich. We thank Harald Gall and the members of his group for their valuable advice. The work is also funded by ERDF via the Estonian Centre of Excellence in Computer Science. The analysis of the models was carried out in the High Performance Computing Center of University of Tartu.

## XI. REFERENCES

- [1] B. Boehm, C. Abts and S. Chulani, "Software development cost estimation approaches — A survey," *Annals of Software Engineering*, vol. 10, no. 1, pp. 177-205, 2000.
- [2] T. C. Jones, *Estimating software costs*, Hightstown, NJ: McGraw-Hill, Inc., 1998.
- [3] B. W. Boehm, Clark, Horowitz, Brown, Reifer, Chulani, R. Madachy and B. Steece, *Software Cost Estimation with Cocomo II*, Upper Saddle River, NJ: Prentice Hall PTR, 2000.
- [4] S. Grimstad and M. Jørgensen, "Inconsistency of expert judgment-based estimates of software development effort," *Journal of Systems and Software*, vol. 80, no. 11, pp. 1770 - 1777, 2007.
- [5] J. F. Ramil and M. M. Lehman, "Metrics of Software Evolution as Effort Predictors - A Case Study," in *ICSM '00: Proceedings of the International Conference on Software Maintenance (ICSM'00)*, Washington, DC, USA, 2000.
- [6] N. Nagappan, B. Murphy and V. Basili, "The influence of organizational structure on software quality: an empirical case study," in *ICSE '08: Proceedings of the 30th international conference on Software engineering*, Leipzig, Germany, 2008.
- [7] J. J. Amor, G. Robles and J. M. Gonzalez-Barahona, "Effort estimation by characterizing developer activity," in *EDSER '06: Proceedings of the 2006 international workshop on Economics driven software engineering research*, Shanghai, China, 2006.
- [8] Y. Zhou and B. Xu, "Predicting the maintainability of open source software using design metrics," *Wuhan University Journal of Natural Sciences*, vol. 13, no. 1, pp. 14-20, February 2008.
- [9] P. C. Pendharkar and J. A. Rodger, "The relationship between software development team size and software development cost," *Communications of the ACM*, vol. 52, no. 1, pp. 141-144, January 2009.
- [10] J. Munson and S. Elbaum, "Code churn: a measure for estimating the impact of code change," in *Proceedings. International Conference on Software Maintenance, 1998.*, 1998.
- [11] N. Nagappan and T. Ball, "Use of relative code churn measures to predict system defect density," in *ICSE '05: Proceedings of the 27th international conference on Software engineering*, St. Louis, MO, USA, 2005.
- [12] C. v. Kotten and A. Gray, "An application of Bayesian network for predicting object-oriented software maintainability," *Information and Software Technology*, vol. 1, no. 48, pp. 59-67, January 2006.
- [13] M. M. T. Thwin and T.-S. Quah, "Application of neural networks for software quality prediction using object-oriented metrics," *Journal of Systems and Software*, vol. 76, no. 2, pp. 147-156, May 2005.
- [14] S. Karus and H. Gall, "A study of language usage evolution in open source software," in *Proceedings of the 8th International Working Conference*, Honolulu, HI, USA, 2011.
- [15] M. Jørgensen and M. Shepperd, "A Systematic Review of Software Development Cost Estimation Studies," *Software Engineering, IEEE Transactions on*, vol. 33, no. 1, pp. 33 -53, January 2007.
- [16] S. Karus and M. Dumas, "Code Churn Estimation Using Organisational and Code Metrics: An Experimental Comparison," *Information and Software Technology*, vol. 54, no. 2, pp. 203-211, February 2012.
- [17] M. Benedikt, W. Fan and F. Geerts, "XPath satisfiability in the presence of DTDs," *Journal of the ACM*, vol. 55, no. 2, pp. 1-79, May 2008.
- [18] Microsoft, [Online]. Available: <http://msdn.microsoft.com/en-us/library/cc645901.aspx>. [Accessed 9 April 2011].
- [19] Microsoft, 9 April 2011. [Online]. Available: <http://msdn.microsoft.com/en-us/library/cc645868.aspx>.
- [20] S. G. MacDonell and A. R. Gray, "Alternatives to Regression Models for Estimating Software Projects," in *Proceedings of the IFPUG Fall Conference*, Dallas, 1996.
- [21] K. M. Eisenhardt, "Building Theories from Case Study Research," *The Academy of Management Review*, vol. 14, no. 4, pp. 532-550, October 1989.
- [22] M. G. Smith and L. Bull, "Feature Construction and Selection Using Genetic Programming and a Genetic Algorithm," in *Lecture Notes in Computer Science: Genetic Programming*, vol. 2610, Springer Berlin / Heidelberg, 2003, pp. 93-100.