CHAPTER I

# Introduction

Two primary concerns in software release planning are: improving functionality and maintaining quality. Both objectives are constrained by limits on development time and cost. In order to respect these constraints and still pursue both objectives, the scope of planned work must be limited so that time is available to properly deal with the inevitable defects (bugs) that will arise. In this way, a software release can better ensure quality while also improving functionality.

A critical step in this planning process is to factor in a suitable amount of time for testing and bug-fixing. Otherwise, there is a risk of slip in the development schedule and/or software quality. As the time and effort required for testing and bug-fixing will likely be a function of the number of defects introduced during development, it is desirable to be able to predict how many bugs can be expected as development proceeds.

A potential application for defect prediction is to compare different release plans according to their estimated bug fallout and subsequent impact on testing and bug-fixing times. This would assist release planners in ensuring that the total development time does not exceed the project’s time budget for a release. The comparison of different release plans is integral to release plan optimization, which is the focus of The Next Release Problem (discussed in detail in the MOTIVATION chapter).

Many approaches to defect prediction focus on either code analysis or historical defect information. To make the defect prediction model useful for comparing release plans, the model must be dependent in some way on the basic elements of the release plan: planned new features and improvements. The historical defect models discussed in the LITERATURE REVIEW chapter are limited in this respect, as they depend only on the past defects.

An approach to defect prediction is presented using a multivariate time series model. This model can be applied for a proposed release, because predictions can be made using only information about proposed features and improvements.

The paper is organized as follows. First, related work is presented in Section \ref{sec:related\_work}. Then, Further motivation for the use of a time series model for predicting defects is presented in Section \ref{sec:motivation}. Next, an overview of time series modeling concepts is provided in Section \ref{sec:time\_series\_modeling}. Methodologies for data collection and time series modeling are detailed in Sections \ref{sec:data\_methodology} and \ref{sec:modeling\_methodology}, respectively. The results of applying these methodologies are given in Section \ref{sec:results}. Last, the paper concludes and poses future work in Section \ref{sec:conclusions}.

CHAPTER II

# LITERATURE REVIEW

Software defect (bug) prediction typically involves a detailed analysis of code or proposed design changes. Some of these analytical methods are mentioned next. Then several statistical approaches to prediction are discussed.

Akiyama \cite{1971\_akiyama} predicted defect counts based on lines of code (LOC), number of decisions, and the number of subroutine calls. Gafney \cite{1984\_gaffney\_estimating} likewise predicted defect count based on LOC. Rather than code itself, Henry and Kafura \cite{1984\_henry\_evaluation} define metrics that are based on information taken from design documents, to be used in defect prediction. Nagappan and Ball \cite{2005\_nagappan\_codechurn} use relative code churn (lines modified) as a metric for predicting the density of defects. Giger, Pinzger, and Gall \cite{2011\_giger\_finegrained} compare the use of code churn to a more fined-grained approach, capturing \say{the exact code changes and their semantics down to statement level.}

## Statistical Approaches to Defect Prediction

Rather than requiring a detailed code analysis to predict defects, the approach proposed in this paper is to develop a mathematical model based on historical data on defect occurrences. Specifically, the proposed approach is to develop a defect prediction model using previous software features, improvements, and defects.

A related approach, used by Li, Shaw, Herbsleb, Ray, and Santhanam\cite{2004\_li\_emperical\_eval}, is to study only the defect occurrences themselves, and attempt to develop a mathematical model for defect projection. In their work, functions were fitted to a time series of defect occurrences, then the function parameters themselves were extrapolated for each new release. They found that the Weibull model fit best in 73\% of the tested software releases. They attempted to extrapolate model parameters using naive methods, moving averages, and exponential smoothing, but found these techniques to be \say{...inadequate in extrapolating model parameters of the Weibull model for defect-occurrence projection}. The reason given for this ineffectiveness is the changing nature of the software development system. For example, development practices, staffing levels, and usage patterns may all change between releases.

In another related approach, Graves, Karr, Marron, and Siy\cite{2000\_graves\_predicting} developed several models that predict the future distribution of software faults in a given code module. The basis of their predictive models is a statistical analysis of change management data, which describes only the changes made to code files. The best model they found was a weighted time damping model, where every change in the module files contributed to fault prediction, with time-damping to account for age of changes. They achieved \say{slightly less successful performance} by basing a generalized linear model on just the modules age and the number of past changes. They also found factors that did not improve model performance: module length, number of developers making changes in the module, and how often a module is changed simultaneously with another module.

In the final approach discussed here, by Singh, Abbas, Ahmad, and Ramaswamy\cite{2010\_singh\_predicting}, the Box-Jenkins method is applied to datasets from the \textit{Eclipse} and \textit{Mozilla} software projects, which are represented as time series data, and defect count is predicted using an ARIMA model. Their modeling effort is focused at the component-level, and they conclude that \say{current bug count of a component is linearly related to its previous bug count}.

CHAPTER III

# MOTIVATION

Release planners typically rely on both their experience and project conventions to generate a release plan by selecting planned features and improvements such that the estimated time to test for and fix defects will not cause a schedule slip.

However, if the defect estimation technique is only loosely based on past experience, as with a rule-of-thumb, then it may prove too coarse for comparing multiple release plans. Specifically, such a technique may not provide any quantitative difference between release plans that are similar (but not the same). For example, suppose two different release plans are being considered. Both include two features, but one has five improvements and the other has seven. A rule-of-thumb approach may provide the same estimate for each. Even for dissimilar release plans, such an approach still has the disadvantage of lacking confidence intervals to quantify prediction uncertainty.

An alternative approach is to develop a model that will take into account the differences in composition of features and improvements between the release plans. In this case, one would expect that the predicted number of defects would vary across the release plans and that prediction uncertainty can be quantified by confidence intervals. Such a model would assume some explanatory relationship, such as shown in Figure \ref{fig:use\_explanatory\_model}.

Since predictive models will have some inaccuracy, but confidence levels can be used to quantify the uncertainty of future prediction based on past accuracy. This would allow release planners to assess the risk of relying on the defect prediction. A higher confidence level results in less risk because it encompasses a larger window for the prediction. Conversely, a lower confidence level results in more risk and a more narrow prediction window.

## Application to the Next Release Problem

Release plan optimization is exactly the goal of The Next Release Problem (NRP), but there is a gap between the abstract domain of the NRP and the detailed, messy data found in software projects. By applying an explanatory predictive model there is a path toward bridging this gap, opening up the potential for using NRP optimization techniques in real-world release planning. In this section, first the NRP is described, then the gap between it and practical planning is discussed, and finally it is shown how the explanatory model suggested earlier would be applied to help bridge this gap.

### Defining the NRP

The Next Release Problem (NRP) was defined by Bagnall, Rayward-Smith, and Whittley \cite{2001\_bagnall\_nrp}, and was shown to be NP-Hard. Being abstract in its treatment of feature cost, a broad range of optimization techniques can be applied to the NRP, such as integer programming, hill climbing, simulated annealing, genetic algorithms, etc. The NRP is the subject of academic research in the area of Search-Based Software Engineering \cite{2010\_jiang\_hybrid,2012\_xuan\_solving,2007\_zhang\_multi\_obj\_nrp}.

The NRP describes the situation where software project planners, who have multiple customers to satisfy, would like to maximize the revenue produced from completing the project. This is all described mathematically as follows.

A software project has a set of all possible requirements (new features and enhancements) that might be included in the next software release. A customer is satisfied by completing a subset . The importance of a customer is given by the weight, .

Requirements may have acyclic dependencies, or prerequisites, that must be completed first. A subset that includes all prerequisite requirements, recursively, is indicated by , and should be taken to mean

For example, if , and is a prerequisite for , then .

A requirement $r \in R$ has a cost $cost(r) \in \mathbb{Z}^+$, associated with its implementation, not considering the cost of any prerequisite requirements. Then, the cost for some subset $R' \subseteq R$ will be

\begin{equation}

cost(R') = \sum \{cost(r) | r \in \hat{R}\_i \}

\end{equation}

Once customer $i$ is satisfied, their weight $w\_i$ contributes to the total revenue from the project, as in

\begin{equation}

\sum\_{i \in S} w\_i

\end{equation}

So, the NRP is posed as follows: for a group of $n$ customers, select the subset $S \subseteq \{ 1,2,...,n \}$ that maximizes total revenue, while keeping the total cost within some budget constraint $B$. This is given by

\begin{equation}

\begin{split}

maximize~~& \sum\_{i \in S} w\_i \\

subject~to~~~& cost(\bigcup\_{i \in S} \hat{R}\_i) \le B

\end{split}

\end{equation}

### The Gap Between Abstraction and Reality

As was discussed in the previous section, a planner would need several things to be able to implement a NRP-like optimization:

\begin{enumerate}

\item{A set of requirements that could potentially be implemented.}

\item{A set of customers that are satisfied by some subset of the requirements, and have an associated weight.}

\item{A cost function, to quantify the cost of each requirement.}

\item{A cost budget that should not be exceeded.}

\end{enumerate}

Having all these in hand, a planner could proceed to optimize the subset of requirements planned for the next release. One difficulty with this that can be highlighted is in the definition of a cost function. It might be suggested that the estimated time to implement a requirement alone might be used to determine cost, but there is a practical detail that prevents this: in order to maintain quality software, the total cost of any requirement should take into consideration both the cost of implementation \emph{and} the cost of fixing associated defects. Otherwise, a release plan would appear to be within budget, when there is a risk that the budget will be exceeded when defect costs are also considered.

### Bridging the Gap

We use the explanatory model to address the need to consider defect cost. Such a model, given some subset of proposed requirements, can be used to predict defects and to find additional cost which should be considered. This use of the predictive model is illustrated in Figure \ref{fig:apply\_model\_in\_nrp}.

\begin{figure}[htbp]

\begin{center}

\tikz[nodes={text height=1em, text depth=.2em, draw=black!20, thick, fill=white, font=\large}, rounded corners, semithick]

\graph[tree layout]{

"Requirements Subset" -> "Predictive Model" -> "Predicted Defects" -> "Defect Cost Function" -> "$\Sigma$" -> "Total Cost";

"Requirements Subset" -> "Requirements Cost Function" -> "$\Sigma$";

};

\caption[Overall cost of requirements]{Defect prediction model used in determining overall cost of some requirements subset.}

\label{fig:apply\_model\_in\_nrp}

\end{center}

\end{figure}

Since predictive models cannot be perfectly accurate, instead we would expect that any forecasting would include confidence levels. Taking into account the confidence of a prediction allows planners to account for risk in the use of the defect prediction. If more risk is acceptable, then planners will get a narrower prediction window, and in exchange take more of a chance that the prediction is inaccurate. A wider prediction window means, though, that when the defect prediction is used to determine requirements cost, that potential cost range will also be wider.

CHAPTER IV

# BACKGROUND

In this section, time series and autoregressive models are introduced. Then, further concepts related to modeling, exogeneity and stationarity, are discussed.

## Time Series

A time series is a collection of observations that occur in order. The process underlying a time series is assumed to be stochastic, so the model must correspondingly be probabilistic. Critically, the sequence of observations cannot be re-arranged, as each observation is typically dependent on one or more previous observation. This dependence is termed autocorrelation and accounting for it is one of the primary functions of a time series model.

## Autoregressive Models

A basic autoregressive (AR) model is formed as a linear combination of previous values, plus a white noise term that accounts for random variations (the stochastic portion). An $AR(p)$ model for predicting a value $X$ at time $t$ can be written

\begin{equation}

X\_t=c+\sum\_{i=1}^{p}{\phi\_t X\_{t-1}+\epsilon\_t}

\end{equation}

where $\phi\_1, \phi\_2, ..., \phi\_p$ are the $p$ parameters, $c$ is a constant, and $\epsilon\_t$ is the white noise term.

When the AR model is extended to the multivariate case (i.e. allowing for multiple time series), a Vector AR (VAR) model is formed. This model will support not only a time series for defect count, but also time series for the two release plan variables: improvements and new features

## Endogeneity and Exogeneity

Under the VAR model, the behavior of each time series is explained by both its own past values and the past values of the other time series. This makes the variables endogenous.

The alternative is that a time series should not be explained by itself, and is only used to explain other time series. This type of explanatory variable is called exogenous, and could be considered an input.

By also considering exogenous variables, a VAR model would become a VARX model. This model meets the requirements of the explanatory model described in the Motivation section, since it would allow release plan variables to be kept exogenous and used only to explain defect count.

## Trends and Stationarity

AR, VAR, and VARX models do not account for non-stationary data. If a time series is not stationary, differencing may produce a stationary series. Trending time series are challenging to analyze, because the summary statistics of mean, variance, and autocovariance vary over time, and are therefore not interpretable \cite{franses1998time}. Two trend types are discussed here: deterministic and stochastic.

A deterministic trend will move upward or downward, meaning that the time series mean is non-constant. However, the time series will be constant according to a deterministic function and the time series movements will generally follow the deterministic function, with non-permanent fluctuations above or below. Such a time series is said to be stationary around a deterministic trend.

In contrast, a stochastic trend shows permanent effects whenever random variations occur, and the series will not necessarily fluctuate only close to the area of a deterministic function. The application of differencing can be used to remove a stochastic trend.

Stationarity can be strict or weak (of some order). Strict stationarity occurs when statistical properties are invariant with respect to shifts of the time origin\cite{moon2000mathematical}. Alternatively, a weak stationarity (of second order) can be established, and from this strict stationarity can be established by then assuming normality\cite{box\_jenkins\_reinsel\_2008}.

For a multivariate time series, stationarity holds if all the component univariate time series are stationary\cite{yang2005stationarity}, so the goal of stationarity testing will be to establish second-order stationarity for each univariate time series component, and then show that the assumption of normality is reasonable. This will establish the stationarity of the multivariate time series as a whole. Next, tests are discussed for assessing if a deterministic or stochastic trend is present.

## Unit Root and Stationarity Testing

A time series that contains a stochastic trend is non-stationary. A pure auto-regressive (AR) model of such a time series contains a unit root \cite{franses1998time}. Testing for the presence of a unit root can therefore be used to test for non-stationarity. A unit-root test poses as the null hypothesis that an AR model has a unit root. Then, a test statistic is measured. If the p-value is below some significance, the null hypothesis can be rejected, and it is established that the time series does not have a stochastic trend. The Augmented Dickey Fuller (ADF) test is often used for unit root testing.

On the other hand, a stationarity test uses the null hypothesis that a time series is stationary around a deterministic trend. If the test statistic shows that this hypothesis can be rejected, at some significance level, then a stochastic trend should be considered, by the unit root test. The Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test can be applied for testing stationarity.

CHAPTER V

# METHODOLOGY

In this chapter, we consider methodologies for both obtaining time series data (a data methodology) and for obtaining a model using that data (a modeling methodology).

## Data Methodology

In this section, the data source and data collection method are detailed. Then, the method of preparing data for the modeling phase is presented.

### Data Sources

The empirical datasets used to establish predictive models came from two software projects’ historical data, and were taken from their issue tracking systems[[1]](#footnote-1).

The first project was the *MongoDB[[2]](#footnote-2)* Core Server software project, which has been actively developed since 2009. Issues from versions 0.9.3 through 3.0.0-rc6 of the project were used, leading to a dataset containing 7,042 issues.

TODO: Netbeans data source

### Data Collection & Cleansing

MongoDB uses JIRA[[3]](#footnote-3) for issue tracking. Issue data was exported from the project's JIRA web interface[[4]](#footnote-4) as XML data. Then, issue data is extracted from the JIRA XML, and the following fields are kept from each issue: type, priority, creation date, resolution date.

Not all of the data was preserved for modeling. The modification or removal of data is discussed next.

#### Unfixed Issues

The proposed model structure assumes that bug creation can be explained by software changes. Therefore, issues that do not result in any change should not be included in the dataset. For this reason, only issues with resolution *fixed*, *complete*, or *done* will be kept. Other possible issue resolutions are: *unresolved*, *won't fix*, *duplicate*, etc. In the data used, 18 (0.26%) of the issues were unfixed.

#### Sub-tasks

Issues that are sub-tasks are first converted to be the same type as the parent issue. Those sub-tasks whose parent issue is not in the dataset are considered orphans and discarded. There were 20 (0.28%) orphaned sub-tasks encountered in the dataset, so this decision is not expected to have much impact on the outcome.

### Data Preparation

After creation, the dataset was operated on to prepare it for time series modeling. The data was sampled, made stationary, and windowed. These three steps are discussed next.

#### Sampling

First, the data was sampled at regular periods to measure the following: number of improvements resolved, number of features resolved, and number of bugs created. A 7-day sampling period was used. As an example, this sampling process is illustrated in Figure \ref{fig:sampling\_issue\_data}, with results shown in Table \ref{tab:sampling\_issue\_data}.

\begin{figure}[htbp]

\begin{center}

\begin{tikzpicture}

\node (bb) at (-2.3,2) [draw] {|Bug|};

\node (n) at (-1.6,1) [draw] {|New Feature|};

\node (ii) at (-1.9,0) [draw] {|Improvement|};

\node (bbbb) at (-1,-1) [draw] {|Bug|};

\node (i) at (0,-2) [draw] {|Improvement|};

\node (bbbbb) at (1.1,-3) [draw] {|Bug|};

\draw[dashed] (-3,3) -- (-3,-4);

\draw[dashed] (0,3) -- (0,-4);

\draw[dashed] (3,3) -- (3,-4);

\draw (-4.5,3.5) node {Period 1};

\draw (-1.5,3.5) node {Period 2};

\draw (1.5,3.5) node {Period 3};

\draw (4.5,3.5) node {...};

\end{tikzpicture}

\caption[Sampling issue data]{Sampling issue data by dividing time into equally-spaced periods.}

\label{fig:sampling\_issue\_data}

\end{center}

\end{figure}

\begin{table}[htbp]

\caption[Sampling issue data]{Results of sampling example issues shown in figure \ref{fig:sampling\_issue\_data}.}

\centering

\begin{tabular}{ c | c | c | c }

\hline

Period & Improvements & New Features & Bugs \\

~& Resolved & Resolved & Created \\

\hline

1 & 0 & 0 & 1 \\

2 & 1 & 1 & 1 \\

3 & 1 & 0 & 1 \\

\hline

\end{tabular}

\label{tab:sampling\_issue\_data}

\end{table}

#### Establishing Stationarity

To establish stationarity, we first need to see if we can rule out the presence of a stochastic trend by applying the augmented Dickey-Fuller (ADF) test. If we can indeed rule out a stochastic trend, we should be able to confirm stationarity by applying the KPSS test. Or, if a stochastic trend cannot be ruled out, then KPSS test should be applied to check that trend stationarity is also rejected. If the data is found to have a stochastic trend, it should be differenced and then retested to confirm (trend) stationarity. In both tests, it will be assumed that the deterministic component is constant, with an intercept but no trend.

The *urca[[5]](#footnote-5)* library provides ur.df and ur.kpss functions for performing these tests.

#### Time Windowing

It is assumed that the software development process underlying a given project might change over time. Rather than developing a model that also changes over time, the data will be kept for modeling only if it occurs within a time window. This will limit the amount of process change the model is exposed to. Taking this approach means that the entire modeling methodology will be executed for each time-windowed part of the data.

## Modeling Methodology

The typical methodology used for building time series models involves specification, estimation, and diagnostics checking \cite[p. 478]{box\_jenkins\_reinsel\_2008}. Once specified and estimated, the diagnostic checking step ensures that only valid models are considered for selection. The final step of modeling would be selection, where models are compared by some model selection criterion \cite[pg. 581]{box\_jenkins\_reinsel\_2008}. This section presents the approach used to specify, estimate, check, and select a VARX model to be used for defect prediction.

### Model Specification & Estimation

Specification of a model is accomplished by choosing an order , which is the number of autoregressive terms to include in the model. Once an order is specified, the model parameters can be estimated by a procedure such as least squares regression.

The model order will directly affect the number of parameters included in the model. One goal of specification will be to avoid having too many parameters relative to the number of observations. The following derivation will lead to a simple rule for limiting the model order in this respect. First, let be the number of time samples in a time series. When there are time series, each sample contains observations, so there are total observations for all time series. Next, for a model of the time series variables, there are unknown parameters to be estimated. Let the ratio of observations to parameters be denoted by

To keep at or above some minimum ratio , we form the inequality

In terms of this becomes

Then, for a fixed value of , an upper bound on the model order would be

With this upper bound, model specification will include the generation of models having order 1, 2,..., . These models, with their estimated parameters, will be candidates for final model selection after undergoing diagnostic checking.

To estimate the parameters of a VARX model, the *dse[[6]](#footnote-6)* library provides the estVARXar function.

### Diagnostics Checking

Diagnostic checking is performed to verify that a model can be accepted. This step includes testing for stability and for model inadequacy.

For model with an AR portion to be stable, the roots of the process characteristic equation must lie outside the unit circle \cite[p. 56]{box\_jenkins\_reinsel\_2008}. Equivalently, the inverse of the roots must lie inside the unit circle.

The *dse* library provides the stability function for performing this test.

#### Portmanteau Test

For an adequate ARMA model, it can be shown that “As the series length increases, the [model residuals] become close to the white noise...” \cite[p. 338]{box\_jenkins\_reinsel\_2008}. For this reason, there are model inadequacy tests formed around a study of the residuals.

One of these tests, the Ljung-Box test, forms a statistic from the autocorrelation of the residuals (up to some lag). In this test, the null hypothesis is that residuals are independent, so their autocorrelation is not high enough to be distinguished from a white noise series. To support this hypothesis, the test p-value should be above some level of significance, say 5%.

The *stats[[7]](#footnote-7)* library provides the Box.test function for performing the Ljung-Box test.

### Model Selection

Model selection criteria are used to compare models according to their fit, by penalizing for residual error and the number of parameters. There are a number of different selection criteria, including Akaike Information Criterion (AIC), AIC with correction (AICc), and Bayesian Information Criterion (BIC). Bisgaard and Kulahci noted that “[t]he penalty for introducing unnecessary parameters is more severe for BIC and AICC than for AIC” \cite{bisgaard2011time}. A less severe penalty for the number of parameters would be preferred in this case, since we are already limiting the number of parameters in the model specification step, and because additional parameters may in fact be necessary to account for time series autocorrelations with higher lags. Therefore, AIC was chosen as the selection criterion.

The *dse* library provides the bestTSestModel function for performing model selection.

CHAPTER VI

# RESULTS

CHAPTER VII

# CONCLUSIONS

CHAPTER VIII

# FUTURE WORK

1. An issue tracking system, in addition to tracking bugs, past and present, can be used to track features, enhancements, or any other type of software process issue. [↑](#footnote-ref-1)
2. MongoDB is a scalable document-oriented database system (http://www.mongodb.org/). [↑](#footnote-ref-2)
3. JIRA is an issue tracking and project management system made by Atlassian, who provides free JIRA subscriptions for qualified open source projects. [↑](#footnote-ref-3)
4. See https://jira.mongodb.org/browse/SERVER. [↑](#footnote-ref-4)
5. The *urca* (http://cran.r-project.org/web/packages/urca) library provides tests for time series data, and is freely available as a package for the *R* computing environment. [↑](#footnote-ref-5)
6. The *dse* (http://cran.r-project.org/web/packages/dse) library provides tools for time series models, and is freely available as a package for the *R* computing environment. [↑](#footnote-ref-6)
7. The *stats* (http://stat.ethz.ch/R-manual/R-patched/library/stats/html/00Index.html) library provides core statistics functions, and is freely available as a package for the *R* computing environment. [↑](#footnote-ref-7)