

Using Time Series Models for Defect Prediction in Software Release Planning

James Tunnell and John Anvik
Computer Science Department
Central Washington University
Ellensburg, WA 98926, USA
[tunnellj, janvik]@cwu.edu

Abstract—A time series model is presented that uses historical project information to predict the number of future defects, given the number of future features and improvements completed. This would allow hypothetical release plans to be compared by assessing their predicted impact on testing and defect-fixing time. We selected the VARX time series model as a reasonable approach. The accuracy of the model appeared low for a single dataset, but the error was found to be normally distributed.

Keywords—*software defect prediction; quality assurance; release planning; time series model;*

I. INTRODUCTION

There are two primary concerns in software release planning: improving functionality and maintaining high quality. Both objectives are constrained by limits on development time and budget, so the scope of the planned work must be limited to accommodate fixing inevitable defects (bugs) that will arise. In this way, a high quality software product can be produced while also improving its functionality.

A significant consideration in the release planning process is the amount of time allocated for testing and bug-fixing. If this factor is not considered, the project risks a slip in the schedule or in the quality of the product. As the time and effort required for testing and bug-fixing will likely be a function of the defects introduced during development, it is desirable to be able to predict the number of expected defects.

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 [2], a key problem in Search-Based Software Engineering (SBSE) [9, 13].

Most approaches to defect prediction focus on either code analysis [1, 5, 6, 8, 11] or historical defect information [7, 10, 13]. However, for the defect prediction model to be useful in comparing release plans, the model should also depend on the planned features and improvements planned for the next release, as well as the defects from past releases.

This paper presents an approach to defect prediction that can be applied for a proposed release. A multivariate time series model is used that incorporates information about proposed features, improvements, and historical defect data.

The paper proceeds as follows. First, Section II presents further motivation for the use of a time series model for predicting defects. Next, we present an overview of concepts in time series modeling in Section III. Section IV presents our modeling methodology and Section V presents the application of the approach, which is applied to a software project dataset. Related work is presented in Section VI, and the paper concludes in Section VII.

II. 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, and may not provide any quantitative difference between release plans that are similar (but not the same). 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. Such a model would assume some explanatory relationship.

Since predictive models rarely have perfect accuracy, confidence levels are an important part of any prediction to allow release planners to assess the risk of relying on the defect prediction. Planners can choose a more narrow prediction window, in exchange for a larger risk that the prediction is inaccurate. Conversely, a wider prediction window means that the potential cost range is also wider with a lower risk of inaccuracy.

III. TIME SERIES MODELING

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

A. Time Series

A time series is a collection of observations that occur in order, with an underlying process that is stochastic. 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.

B. 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). 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.

The VAR model can be further extended by considering one or more variables to be exogenous, making a VARX model. Exogenous variables are used to explain the other non-exogenous, variables, but the model does not attempt to explain the exogenous variables themselves. 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.

C. Stationarity and Trends

A strictly stationary process has a probability distribution that is time-invariant. This means statistics such as mean and variance do not change. The AR, VAR, and VARX models discussed so far require time series data that is stationary, where the probability distribution of the underlying stochastic process is time-invariant. Testing can identify a time series as being stationary, trend stationary, or non-stationary.

A time series can be established as non-stationary by testing for the presence of a unit root in the underlying AR model. The unit root test used is the Augmented Dickey Fuller (ADF) test. On the other hand, a stationarity test establishes a time series as trend stationary by testing for the presence of a deterministic trend function (either a constant or a line). The stationarity test used is the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test.

IV. MODELING METHODOLOGY

The typical methodology used for building time series models involves specification, estimation, and diagnostics checking [4, p. 478]. Once specified and estimated, the diagnostic checking step ensures that only valid models are considered for selection. The final step of modeling is selection, where the models are compared by some model selection criterion [4, p. 581]. This section presents our approach to specifying, estimating, diagnostics checking, and model selection for defect prediction.

A. Model Specification & Estimation

The specification of a $VARX(p)$ model is accomplished by choosing an order p , which is the number of autoregressive

terms to include in the model. Then 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 is to avoid having too many parameters relative to the number of observations. To this end, we establish a ratio K of the number of observations to the number of parameters. By choosing a minimum value for this ratio, K_{min} , and using the formula for the number of parameters in a $VARX(p)$ model, the following equation can be used to obtain a maximum model order p_{max} :

$$p_{max} = \left\lfloor \frac{n}{mK_{min}} \right\rfloor$$

where there are m time series variables and n samples. The establishes an upper bound on model order, so model specification will include the generation of models having order 1, 2, ..., p_{max} . These models, with their estimated parameters, will be candidates for final model selection after undergoing diagnostic checking.

B. Diagnostics Checking

Diagnostic checking is performed to verify that a model can be accepted. This step includes testing for stability and for model inadequacy. A stability test checks that the roots of the AR process characteristic equation lie outside the unit circle [4, p. 56]. To test inadequacy, the Ljung-Box is used to compare the model residuals to white noise.

C. Model Selection

Model selection criteria are used to compare models by their fit, to minimize residual error, and to penalize the model to some degree based on the number of parameters. Of the commonly used selection criteria, the standard Akaike Information Criterion (AIC) was used because “[t]he penalty for introducing unnecessary parameters is more severe for BIC and AICC than for AIC” [3]. 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.

V. APPLICATION OF METHODOLOGY

To validate our approach of using a time series model to predict defects, we used historical data taken from a software project’s issue tracking system. Issue tracking systems are used by projects for tracking development tasks, features, enhancements, and bugs, both past and present.

We chose the *MongoDB*¹ Core Server project as the data set. This project was chosen as it has been active since May 2009 and uses *JIRA*² for issue tracking, which made it easy to collect data. Issues for versions 0.9.3 through 3.0.0-rc6 were exported from the project’s *JIRA* web interface into XML

¹ MongoDB is an open source, document-oriented database system.

² JIRA is an issue tracking and project management system made by Atlassian.

format. The fields collected from each issue report were: type, priority, creation date, and resolution date.

Only issues marked as *fixed*, *complete*, or *done* were for modeling. In the data collected, 18 (0.26%) issues did not meet this criterion and were excluded. Also, *JIRA* supports issues having sub-tasks. Those sub-tasks whose parent issue was not in the dataset were considered orphans and discarded. There were 20 (0.28%) orphaned sub-tasks in the dataset. The final dataset contained 7042 issues.

A. 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.

1) 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.

2) Establishing Stationarity

To establish stationarity, the ADF unit root and KPSS stationarity tests were applied. In both tests, it was assumed that the deterministic component was constant (without slope). These test results did not agree, so the time series data was differenced and the tests were rerun. The test results then agreed, establishing the stationarity of the differenced data.

3) Time Windowing

It can be assumed that the software development process underlying a given project changes over time. Rather than developing a model that also changes over time, the data was kept for modeling only if it occurred within a time window. This was done to limit the effect of process change on the model. A time window of 78 weeks (approximately 18 months) was selected to balance between more observations (to capture consistent long-term behaviors), and fewer observations (to limit exposure to behavioral changes).

Applying this time window, the data was divided into three 78-week windows. As the data was differenced, the first sample was skipped in each data period. These windowed periods are denoted W_{2-79} , W_{80-157} , and $W_{158-235}$.

B. VARX Modeling

1) Use of the VARX Model

The VARX model was chosen to model the time series because there are multiple time series to be considered jointly. The $Y_{\Delta imp}$ and $Y_{\Delta new}$ time series were both considered exogenous, so that hypothetical future values could be considered when comparing release plans. And by selecting $K_{min} = 4$, a maximum model order of $p_{max} = 6$ is obtained, so only model orders 1 through 6 were estimated for later diagnostic checking.

C. Model Diagnostic Checking

Candidate models were tested for stability and inadequacy. A 5% significance level was used in the Ljung-Box test. All model orders were found stable for all windowed periods.

Several model orders were found to be inadequate, specifically orders 1-2 for period W_{2-79} , and order 5 for period $W_{158-235}$.

D. Model Selection

Models that were not rejected for instability or inadequacy were then compared and the best for each windowed period was selected by AIC selection criterion. The best model orders found were 4, 1, and 1, for windowed periods W_{2-79} , W_{80-157} , and $W_{158-235}$, respectively. The fit for each of these models was demonstrated by plotting one-step predictions along with actual values, as shown for each model in Fig. 1. The fit for each appears to track well with many of the significant changes in the time series.

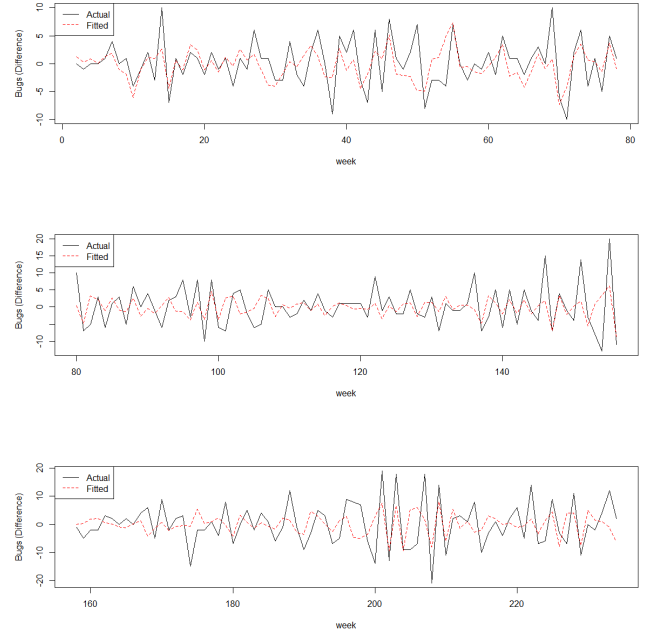


Figure 1. One-step predictions vs actual values, for each model selected by AIC score.

E. Forecasting

Selected models were used to forecast the number of defects in the next sample after the end of the window. The input for making these predictions was the number of improvements and features that were expected to be resolved.

Table I shows the resulting single-step, out-of-sample defect prediction data for the first time window, W_{2-79} , including the upper and lower bounds of the confidence intervals. The actual number of improvements, features, and bugs in the prediction sample period was 4, 0, and 18, respectively. Notice that the actual number of bugs, 18, is outside of the 90% confidence interval, which spans from 6.4 to 13.79 (see the outlined row in Table I). On the other hand, the actual number of future defects in the next window, W_{80-157} , was 17. This was inside the 90% confidence interval, which spans from 13.38 to 18.00.

TABLE I. FORECASTING AT THE END OF THE FIRST TIME WINDOW, $W_{2.79}$. FUTURE OUTPUT VALUES ARE PREDICTED FOR A NUMBER OF HYPOTHETICAL INPUT VALUES.

Improvements	Features	90% lo	75% lo	mean	75% hi	90% hi
2	0	5.61	6.72	9.31	11.89	13.00
2	1	5.54	6.66	9.24	11.82	12.93
2	2	5.48	6.59	9.17	11.75	12.86
2	3	5.41	6.52	9.10	11.69	12.80
4	0	6.40	7.51	10.09	12.68	13.79
4	1	6.33	7.44	10.03	12.61	13.72
4	2	6.27	7.38	9.96	12.54	13.65
4	3	6.20	7.31	9.89	12.48	13.59

To gauge how well prediction will work in general, a sliding 78-week window was applied, starting at the first sample period, and shifting by one sample period after modeling. Only actual numbers were used in this forecasting. The resulting distribution of errors between the mean forecasted bugs and the actual number of bugs is shown as a histogram in Fig. 2. Note that the histogram appears to be normally distributed. The actual number of bugs was inside the 90% confidence interval for 23.87% of the sliding window ranges.

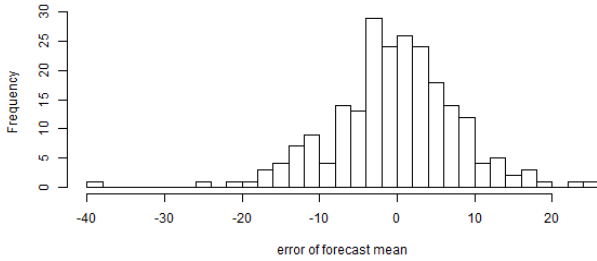


Figure 2. Histogram of forecast mean errors obtained using a 78-week sliding window.

VI. RELATED WORK

Prior defect prediction techniques generally fall into two categories: those based on code analysis and those based on statistical analysis.

Code analysis techniques typically involve a detailed analysis of code, using metrics such as lines of code (LOC) [1] or decision points [5]. Henry and Kafura [8] defined metrics from design document information for use in defect prediction.

Statistical analysis techniques create mathematical models based on historical defect occurrence information, such as regression analysis and extrapolation [10]. Graves et al. [7] developed a weighted time-damping model using a statistical analysis of change management data. And Singh et al. [12] applied the Box-Jenkins method to time series datasets from the Eclipse and Mozilla projects to predict defect counts using an ARIMA model, though their model is non-explanatory and is only in terms of past defects. We included past features and improvements as model inputs, so defects can be predicted using values from hypothetical release plans.

VII. CONCLUSIONS AND FUTURE WORK

The VARX modeling methodology was successfully applied to the time series data collected from the *MongoDB* project. A model was created for each of three time windows and then used to make defect predictions for a range of hypothetical values for the number of improvements and features. Also, a picture of the prediction performance was obtained by applying the approach with a sliding window. This resulted in a normally distributed error between the mean forecasted and actual number of bugs. A low proportion (23.87%) of the sliding window ranges included the actual number of bugs using a 90% confidence interval. These results indicate that the VARX model had a low prediction accuracy for the actual number of defects in the *MongoDB* dataset.

Having applied the VARX time series model to one project dataset, a next step is to apply the methodology to other software project data sets, such as *Eclipse* or *Firefox*, to better determine the applicability of the modeling approach.

REFERENCES

- [1] F. Akiyama. An example of software system debugging. In IFIP Congress (1), volume 71, pages 353–359, 1971.
- [2] A. J. Bagnall, V. J. Rayward-Smith, and I. M. Whitley. The next release problem. *Information and software technology*, 43(14):883–890, 2001.
- [3] S. Bisgaard and M. Kulahci. *Time series analysis and forecasting by example*. John Wiley & Sons, 2011.
- [4] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis*. John Wiley, 2008.
- [5] J. E. Gaffney. Estimating the number of faults in code. *Software Engineering, IEEE Transactions on*, SE-10(4):459–464, July 1984.
- [6] E. Giger, M. Pinzger, and H. C. Gall. Comparing fine-grained source code changes and code churn for bug prediction. In *Proceedings of the 8th Working Conference on Mining Software Repositories*, pages 83–92. ACM, 2011.
- [7] T. L. Graves, A. F. Karr, J. S. Marron, and H. Siy. Predicting fault incidence using software change history. *Software Engineering, IEEE Transactions on*, 26(7):653–661, 2000.
- [8] S. Henry and D. Kafura. The evaluation of software systems' structure using quantitative software metrics. *Software: Practice and Experience*, 14(6):561–573, 1984.
- [9] H. Jiang, J. Zhang, J. Xuan, Z. Ren, and Y. Hu. A hybrid ACO algorithm for the next release problem. In *Software Engineering and Data Mining (SEDM), 2010 2nd International Conference on*, pages 166–171. IEEE, 2010.
- [10] P. L. Li, M. Shaw, J. Herbsleb, B. Ray, and P. Santhanam. Empirical evaluation of defect projection models for widely-deployed production software systems. *SIGSOFT Softw. Eng. Notes*, 29(6):263–272, Oct. 2004.
- [11] N. Nagappan and T. Ball. Use of relative code churn measures to predict system defect density. In *Software Engineering, 2005. ICSE 2005. Proceedings. 27th International Conference on*, pages 284–292. IEEE, 2005.
- [12] L. L. Singh, A. M. Abbas, F. Ahmad, and S. Ramaswamy. Predicting software bugs using arima model. In *Proceedings of the 48th Annual Southeast Regional Conference*, page 27. ACM, 2010.
- [13] J. Xuan, H. Jiang, Z. Ren, and Z. Luo. Solving the large scale next release problem with a backbone-based multilevel algorithm. *Software Engineering, IEEE Transactions on*, 38(5):1195–1212, 2012.