ISBN 978-981-11-3671-9
Proceedings of 2017 the 7th International Workshop on Computer Science and Engineering (WCSE 2017)
Beijing, 25-27 June, 2017, pp. 448-453
doi: 10.18178/wcse.2017.06.077

Predictive Analysis of Cloud Incidents

Yaman Roumani [1], Joseph K. Nwankpa [2] and Yazan F. Roumani [3]

[1] Eastern Michigan University, Ypsilanti, MI, USA
[2] The University of Texas Rio Grande Valley, McAllen, TX, USA
[3] Oakland University, Rochester, MI, USA

Abstract. With the widespread use of cloud computing, the number of cloud incidents involving outages, vulnerabilities, data loss, auto fails and hacks are constantly increasing. Although several prediction models have been proposed to forecast cloud incidents, such models do not consider trend, level, and seasonality components of cloud incidents. Using time series analysis, we create a predictive model for cloud incidents. Results show that the level of the series to be the best estimator of the prediction model and that time series model can be useful for prediction.

Keywords: cloud incidents, prediction, time series, ARIMA

1. Introduction

Cloud computing has been widely endorsed by IT companies as the next paradigm shift in technology. Cloud computing allows users to develop, deploy, and run scalable applications that work rapidly, reduce costs, and are available without the concerns about the properties and the locations of the underlying infrastructure. Any incident including: outages, vulnerabilities, data loss, auto fails and hacks could result in the loss of billions of dollars for businesses and users. Several examples illustrate this point. In February 2009, Gmail had an outage that lasted two and a half hours [1]. Similarly, there are many reports about incidents in popular cloud services including Amazon [2], Adobe [3], Microsoft [4], and Dropbox [5]. Security has also been a significant issue facing cloud service providers and users as they consider shifting their data and information to the cloud [6]. So, the key problem for cloud provider is minimizing cloud incidents in order to provide a reliable and secure service for their customers. In this study, we create a prediction model using time series analysis. Time series analysis takes into account that data points tend to have an internal structure such as autocorrelation, periodical, and seasonal components. It describes complex relationships among past data points and extends these relationships into the future. In this paper, we use one method of times series analysis namely, exponential smoothing, to predict the number of cloud incidents. Our work has several contributions. First, we identify an alternative method for cloud incidents prediction using time series analysis. Second, we consider trend, level and seasonality components of cloud incidents which provide information about regularity in the series that can help in prediction. To our knowledge, this is the first study to examine cloud incidents using this approach.

The paper proceeds as follows: Section 2 discusses background literature on cloud incidents and existing prediction techniques. Section 3 offers a review of time series modeling and exponential smoothing technique. This is followed by Section 4 which talks about the methodology, data collection, analysis, and examination of the fit and predictive capabilities of the proposed model. Finally, Sections 5 and 6 offer a discussion of implications, and limitations.

2. Related Work

Prior prediction studies on cloud incidents have used different techniques to build a prediction model including: machine learning, hidden Markov models, and Bayesian estimation. Using supervised and

unsupervised statistical learning methods, [7] proposed an anomaly detection approach for scale- out storage systems that predicts cloud anomalies caused by memory and network faults. The authors demonstrated that their approach can efficiently identify anomalies. Tan and Wang [8] proposed an adaptive runtime anomaly prediction system which can predict anomalies. The system employs classification to capture a special alert state in addition to the normal and anomaly states. However, the authors did not report the predictive efficiency of their model. Hagen et al. [9] analyzed Amazon's incident report of a recent cloud outage and proposed an object-oriented verification algorithm to detect the logical violation of safety constraints by IT changes. The proposed method has been shown to detect several configuration changes in static and dynamic routing environments that cause a network overload. Mills et al. [10] implemented design-time method to predict system failures prior to system deployment. The method uses a genetic algorithm to search system simulations for parameter combinations that result in system failures. The method was applied to an existing cloud simulator and concluded that their method can reveal insights about optimal parameter settings and finding failure scenarios. Guan et al. [11] presented an unsupervised failure detection and prediction technique using Bayesian classifiers and decision trees. The method characterizes normal execution states of a system and detects anomalous behaviors. The anomalies would be verified, labeled and used to predict future failure occurrences.

3. Time Series

Time series modeling involves the process of creating a model for a variable measured over a time period. Times series model does not explain or measure the causal factors underlying the behavior of the observed variable, it explores patterns in past movements in order to forecast future behavior [12]. Current time-series models include linear, non-linear and a combination of linear and non-linear models [13]. Linear models include exponential smoothing and Autoregressive Integrated Moving Average (ARIMA), while non-linear models include neural network and fuzzy systems. This study we will focus on linear models. More specifically, our analysis will focus on exponential smoothing.

3.1. Exponential smoothing

Exponential smoothing is a special case of ARIMA models, thus, it is important to explain ARIMA models first. ARIMA models, also known as the Box- Jenkins model, are a class of linear models for univariate time series [14]. Building an ARIMA model consists of three stages. The first stage involves model identification and it includes specifying the structure and order of the model. The ARIMA model structure is represented by (p, d, q) where p is the number of autoregressive (AR) terms, d is the number of non-seasonal differences, and q is the number of moving average (MA) terms. An AR term specifies whether the data values are autocorrelated, or affected by preceding values. For instance, p =0 implies that there is no autocorrelation, whereas p = 1 means the current value is affected by the previous data point. Nonseasonal differences (d) refer to the type of adjustment that is needed to achieve a stationary mean. When d = 0, it implies that the mean is stationary, d = 1 means that there is a linear trend, and d = 2 implies that there is a quadratic trend. MA term refers to the number of lagged forecast errors. For example, q = 0, means that there are no random shocks in the data. The general equation of an ARIMA(p, d, q) is:

$y'_t = c + (\phi_1 y'_{t-1} + \ldots + \phi_p y'_{t-p}) + (\theta_1 e_{t-1} + \ldots + \theta_q e_{t-q}) + e_t$ where: y'_t = the differenced series; c = a constant, ϕ, θ = coefficients; p = order of the AR term; q = order of the MA term; e_t = the estimated residual at time t

In order to determine the terms of p, d and q and identify the ARIMA model, plots of autocorrelation (ACF) and partial autocorrelation (PACF) functions are used. But, prior to the first stage, it is important to make sure the data is stationary, meaning that the mean of the data shows no trend over time. If data was found to be non-stationary, it can be transformed into stationary data by taking the successive differences between data points. The second stage of building an ARIMA model estimates the parameters of the model using maximum likelihood or non-linear least squares estimation methods. Estimation of parameters requires complicated iteration procedure [15]. Diagnostic checking and forecasting are done in the third stage. Diagnostic checking ensures that the residuals of the model are random and the estimated parameters are statistically significant. The fitting process is guided by the principle of parsimony where the best model is the simplest possible model. In order to identify any misspecification, it is important to plot the mean and

variance of residuals over time and perform a Ljung-Box test [16]. Unlike ARIMA, exponential smoothing models isolate seasonality from irregular variation [17]. Exponential smoothing assigns exponentially decreasing weights as data points get older [18]. Therefore, recent data points are assigned more weight than older data points. For our cloud incidents data, we will discuss Holt-Winters' additive which was the only model that provided the best fit to our data.

3.2. Holt-Winters' additive model

Holt-Winters' additive model takes into consideration data with a linear trend and a seasonal effect that is not dependent on the level of the series [19]. Holt-Winters' additive model produces three smoothed values: α (level), γ (trend of the forecast), and δ (seasonal adjustment to the forecast). For this study, level refers to the relative magnitude of the number of cloud incidents which may be constant or change with time. Trend refers to gradual upward or downward, long-term movement of the number of cloud incidents. Seasonality refers to short-term, regular variations in the number of cloud incidents at regular intervals. Methodology

3.3. Data collection

Cloud incidents data was collected from Cloutage.org, an Open Security Foundation (OSF) project geared towards providing cloud security knowledge and resources. Cloutage keeps track of all cloud incidents that have occurred since 1998. Each cloud incident consists of the following attributes: incident id, incident type, incident date, summary, affected organization, affected services, and duration. Additionally, each incident is classified into five categories: dataloss, autofail, hack, vulnerability, and outage. According to Cloutage, dataloss incidents include unforeseen loss of data as a result of issues such as poor backup and recovery. Autofail refers to an incident involving failure in update mechanisms such as virus definition and software updates. A hack incident includes breaches in cloud service providers or online services. A vulnerability incident tracks site specific vulnerabilities in cloud service providers and online services. Finally, an outage incident involves any unexpected availability or impact to cloud providers. This includes unavailable services and inaccessible features. For this study, the collected data included 2404 cloud incidents that occurred between January 2010 and February 2014. All cloud incidents were aggregated over a monthly period. Table 1 shows descriptive statistics of the dataset.

Table 1: Descriptive statistics – cloud Incidents dataset

Collection period	Jan 2010 – Feb 2014
Dataloss incidents	3
Autofail incidents	8
Hack incidents	28
Vulnerability incidents	40
Outage incidents	2325
Total number of	2404
Average monthly	48.08

3.4. Fitting time series model

To find the best fitting time series model, we used SPSS v.20 statistical software. The dependent variable included the number of cloud incidents (aggregated by month), whereas the independent variable included time (measured by month and year). All necessary assumptions were evaluated. Stationarity was assessed using the autocorrelation function and the augmented Dickey- Fuller test. Model parameter appropriateness and seasonality were assessed with the autocorrelation, partial autocorrelation and inverse autocorrelation functions. In order to identify any systematic patterns or outliers, plots of the residuals versus time were inspected. Once a model was fit, residual diagnosis was performed using Ljung–Box test to determine model adequacy. The Ljung–Box test is used to evaluate the null hypothesis that the residuals are white noise, meaning random spread of residuals in time [20]. If the null hypothesis is not rejected, it indicates an adequate fitted model. Holt-Winters' additive was found to be the best model for cloud incidents. Using Ljung-Box statistics, the accuracy of the fitted model was checked. The reported p-value of 0.253 implies that the model is correctly specified. Also, the reported R^2 value indicates that the model explains 81.6% of

the observed variation in the cloud incidents dataset. Table 2 shows the smoothing parameters α (level), γ (trend) and δ (seasonal) for the prediction model. The Holt-Winter's additive model includes three estimates α (level), γ (trend) and δ (seasonal). The level estimate is statistically significant ($p<0.05$) while the estimate for trend and seasonal components are not statistically significant ($p>0.05$).

Table 2. Holt-Winters' additive model parameters

Parameter	Estimate	SE	t	Sig.
α (Level)	0.701	0.147	4.779	0.000
γ (Trend)	0.000	0.036	0.014	0.989
δ (Seasonal)	0.001	0.277	0.004	0.997

Figure 1 shows the graphs of the original cloud incidents and the fitted values obtained from the prediction model.

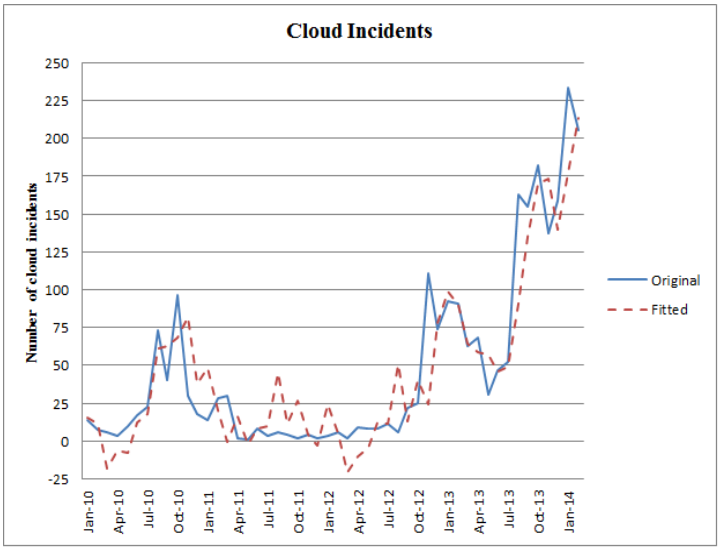

Fig 1. Cloud incidents: original vs. fitted values

3.5. Forecasting cloud incidents and predictive capability

Subsequently, the prediction model is used to generate the forecasting values for 10 months of 2014 as shown in Table 3. Based on the table, we can see that forecasted number of cloud incidents for 2014 is highest in October (232) and the lowest in April (176) and June (176).

Table 3. Forecasting number of cloud incidents for the year 2014

Month	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
# of Incidents	181	176	168	176	178	218	211	232	227	220

One of the important criterions for evaluating forecasting validity is prediction accuracy. In order to estimate model performance and reliability, an error analysis based on Mean Absolute Percentage Error (MAPE) is applied. Using MAPE, the predictive error of Holt- Winter's additive model was 43.66%.

4. Discussion

This study suggests that time series model provides a good fit for cloud incidents and can be used to predict the number of cloud incidents. With respect to forecasting, the prediction model had a 43.66% prediction error. It is our belief that the percentage error can be reduced by creating separate time series model for different cloud vendors and including additional factors in the prediction model such as: cloud infrastructure (storage, network, virtualization, servers…etc.), cloud service type (email, VoIP, banking…etc.), popularity, and market share. The proposed model in this study confirms that it is possible to use time series modeling to predict the number of cloud incidents, but its relevance to cloud vendor still needs further research. Based on an existing study that used text-mining techniques from the same data

source, we believe specific vendors such as Google and Microsoft can benefit from time series modeling to predict cloud incidents [21].

Our study showed that the level of the series was the only significant parameter while seasonality and trend were insignificant components of the prediction model. This means that cloud incidents cannot be predicted based on their monthly seasonal patterns or trends. Future research can examine if weekly, quarterly, or annual patterns may produce significant trends and seasonal components. Moreover, our findings suggest an increase in the number of incidents in 2014. More specifically, the number of cloud incidents during 2014 is expected to be higher than the previous years. It is important to note that since time series models assign more weight to more recent values, better forecasts can be obtained if more recent data are available for the prediction model.

5. Conclusion

The objective of this study was to predict the number of cloud incidents using time series models and to examine whether cloud incidents have trends, levels, and seasonality components. Using cloud incidents dataset, we built a time series model. Based on the fit statistics and the coefficients of the proposed model, the level of the time series was the only significant parameter for cloud incidents prediction. Our results also presented forecasting values for cloud incidents during 2014.

6. References

[1] Current Gmail outage. (2009). Google Blog.

[2] Whittaker, Z. (2013). Amazon Web Services suffers outage, takes down Vine, Instagram, others with it. ZDNet.

[3] Adobe's Cloud Outage Angers Users. (2014, May 16). Wall Street Journal.

[4] Microsoft Azure Outage: Questions Remain. (2014). Information Week.

[5] Dropbox hit by outage on Friday, denies that it was hacked. (2014, January 10). PCWorld.

[6] Wei, L., Zhu, H., Cao, Z., Dong, X., Jia, W., Chen, Y., & Vasilakos, A. V. (2014). Security and privacy for storage and computation in cloud computing. Information Sciences, 258, 371-386.

[7] Silvestre, G., Sauvanaud, C., Ka âniche, M., & Kanoun,K. (2014, October). An anomaly detection approach for scale-out storage systems. In IEEE, p. 294-301.

[8] Tan, Y., Gu, X., & Wang, H. (2010, July). Adaptive system anomaly prediction for large-scale hosting infrastructures. In ACM SIGACT- SIGOPS, p.173-182.

[9] Hagen, S., Seibold, M., & Kemper, A. (2012, April). Efficient verification of IT change operations or: How we could have prevented Amazon's cloud outage. In IEEE, p.368-376.

[10] Mills, K., Dabrowski, C., Filliben, J., & Ressler, S. (2013, October). Combining genetic algorithms and simulation to search for failure scenarios in system models. In SIMUL 2013, p. 81-88.

[11] Guan, Q., Zhang, Z., & Fu, S. (2012). Ensemble of bayesian predictors and decision trees for proactive failure management in cloud computing systems. Journal of Communications, 7(1), 52-61.

[12] Granger, C. W. J., & Newbold, P. (2014). Forecasting economic time series. Academic Press.

[13] Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159-175.

[14] Box, G. E., & Jenkins, G. M. (1970). Time Series Analysis: Forecasting and Control, Holden- Day, 1970.

[15] Chatfield, C. (2013). The analysis of time series: an introduction. CRC press.

[16] Anderson, O. D. (1976). Time series analysis and forecasting: the Box-Jenkins approach (p. 182). London and Boston: Butterworths.

[17] McKenzie, E. (1984). General exponential smoothing and the equivalent ARMA process. Journal of Forecasting, 3(3), 333-344.

[18] Shumway, R. H., & Stoffer, D. S. (2010). Time series analysis and its applications: with R examples. Springer Science & Business Media.

[19] IBM. (2017). Time Series Modeler. IBM Knowledge Center.

[20] Ljung, G. M., & Box, G. E. (1978). On a measure of lack of fit in time series models. Biometrika, 65(2), 297- 303.

[21] Hsia-Ching Chang; Chen-Ya Wang, (2015). Cloud Incident Data Analytics: Change-Point Analysis and Text Visualization," System Sciences (HICSS), 2015 48th Hawaii International Conference, pp.5320-5330.