An Empirical Investigation of Alert Notifications: A Temporal Analysis Approach

Babajide Osatuyi
New Jersey Institute of Technology
bjo4@njit.edu

Michael Chumer
New Jersey Institute of Technology
chumer@njit.edu

ABSTRACT

As the deployment of situational awareness mechanisms such as geothermal sensors, use of social network sites, and information and communication technologies (e.g., cell phones) become increasingly widespread to emergency responders, the problem of alert analysis has become very important. Broadcast of large amounts of alerts sent back to command centers for processing may impair the ability of analysts to connect dots that may otherwise adequately enable them to make informed decisions in a timely fashion. This paper investigates trends and patterns embedded in alert notifications generated over a given period of time in order to uncover correlations that may exist in the data. Data for this study are obtained from the National Center for Crisis and Continuity Coordination (NC4). We employ classical time series analysis to understand, explain and predict trends and patterns in the data. This work presents results obtained thus far in the quest for the effect of passage of time on alert patterns. Implications of this work in practice and research are discussed.

Keywords

Alert notification, time series analysis, intelligence gathering.

INTRODUCTION

The need for adequate intelligence (or information) analysis has become highly important to organizations such as emergency response organizations (EROs), the Department of Homeland Security (DHS) and Natural Disaster Management (Janeja et al., 2005) due to the wide range of sources from which information is gathered during response to a catastrophic event. While multiple sources of information is required to gain and/or validate shared situational awareness, the amount of information available can very quickly become overwhelming and administrators tend to miss out on crucial information that may enable them make informed decisions in a timely manner (Malizia et al., 2008). As an example, the recent failed terror attack on the American jetliner might have been averted if the responsible agencies proactively connected crucial dots in the information gathered over time. Alert notifications are one of the sources of intelligence gathering techniques employed by organizations such as DHS and EROs. In fact, alert notification is recognized as one of the most important role in Emergency Response Information Systems (van de Walle and Turoff, 2007). Therefore the management of incoming alerts is imperative to these organizations.

Prior studies (e.g., Janeja et al., 2005, Fawcett and Provost, 1999, Fawcett and Provost, 1997, Grossman, 2004) have presented approaches to assess incoming alerts for relevance and integrity (i.e., true versus false positives). These studies emphasized the evaluation of algorithms used to assess the validity of alerts received in order to discern if they were true or false. Rigorous relevance and integrity analyses suggested from prior research have been applied to the study data from the National Center for Crisis and Continuity Coordination (NC4), hence considered highly classified and reliable for further analysis such as that employed in this study.

In a related research, Watier et al (1991) proposed an autoregressive integrated moving average (ARIMA) type model-based warning system where the alert threshold value is a function of the upper side of the prediction interval. Farrington et al. (1996) developed a regression algorithm to assist in detecting outbreaks of infectious diseases reported to the Communicable Surveillance Center. Prediction intervals for the model base-line rate were used to construct a threshold for the number of potential cases. This study combines regression analysis and ARIMA methodologies (i.e., classical time series approach) in order to explain most of the variability in the alert notifications data. Given the sequential nature of the data, we specifically seek to understand time dependence of incidents in order to make informative predictions. The ability to predict future incidents based on past occurrences is an important contribution of this work.

Reviewing Statement: This paper represents work in progress, an issue for discussion, a case study, best practice or other matters of interest and has been reviewed for clarity, relevance and significance.

The rest of the paper will be organized as follows: The methodology employed is presented in the next section followed by the presentation of the results till date. Implications of the results in research and practice are discussed as well as future directions.

METHODOLOGY

Classical time series is used for the analysis in this paper as it combines regression analysis and ARIMA model that takes into account the serial dependency (autocorrelation) of observations in an uncontrolled setting, allowing for prediction of future incidents without an attempt to measure independent relationships that influence it (Linden, 2003, SAS Institute Inc., 1996). In general, the ARIMA model introduced by Box and Jenkins (1976) includes autoregressive (p) and moving average parameters (q) as well as differencing (d) in the formulation of the model. Box-Jenkins models are summarized as ARIMA (p, q, d). For example, given a time series process y_t, a first order autoregressive process is denoted as ARIMA (1,1,1) or simply AR(1) means that it contains one autoregressive parameter and one moving average parameter after it was differenced once to attain stationarity. AR(1) can be expressed mathematically as:

$$y_t = theta + phi * y_{t-1} + error_t \quad (1)$$

where *theta* is a constant, *phi* is the coefficient of the one-step time series process (y_{t-1}), *phi* is bounded within unity, $error_t$ is called the white noise assumed to be normal with zero standard deviation and mean of unity.

The identification and estimation of ARIMA models introduced by Box and Jenkins (1976) includes three steps: model identification, model parameter estimation and diagnostic checking for the identified model appropriateness for modeling and forecasting. Following the setup steps for a given time series, a calibrated model is developed with the basic statistical properties of the time series into its parameters.

Study Data

Study data was mined from NC4's open database for a period of two years. An excerpt from the data used for the study is shown in Table 1.

category	type	count	severity	qtr	year
Hazmat	Chemical Spill	2	Minor	1	2007
Transportation	Tractor Trailer Accident	2	Moderate	1	2007
Fire	1 Alarm Fire	5	Minor	1	2007
Fire	6+ Alarm Fire	1	Severe	1	2008
Hazmat	Fuel Spill	1	Minor	2	2007

Table 1. Incidents received from 2007-2008.

The first column of the table is the *category* of an incident followed by the type of the incident. *Category* is classified as hazmat, transportation, fire, advisory, security, aviation, health, structural, infrastructure, geophysical, or other. *Type* is the kind of a category incident. For instance, an earthquake is a *type* of geophysical incident. The frequency of a type of incident is recorded as *count*. *Severity* column is used to capture the severity level of the incident reported. Quarter (*qtr*) and *year* of the incident are recorded as well. The fourth record, for instance, in Table1 is an alert that was sent out about a category 6+ fire in the first quarter of year 2008 and rated as highly severe. Time of occurrence is also captured but not used in this study due to the unequal lapse in incident time, which is inappropriate for time series analysis. Time intervals for the analysis has been divided into quartiles to correspond to the four seasons and to determine if the alerts as data suggest seasonal patterns of hurricanes, tornadoes, wildfires, and other natural or manmade events that rise in severity to an emergency requiring a level of response. These patterns, during certain seasons, can be predicted and may be intuitive but additional temporal patterns related to emergency situations could be identified through a time series analysis.

RESULTS

Figure 1 shows alert notifications by quarter, received between 2007 and 2008. The pattern shown in Figure 1 suggests that the overall number of incidents received in the first quarter of 2007 and 2008 are fairly equal. A steady increase in number of incidents over the quarters is noticed in 2008.

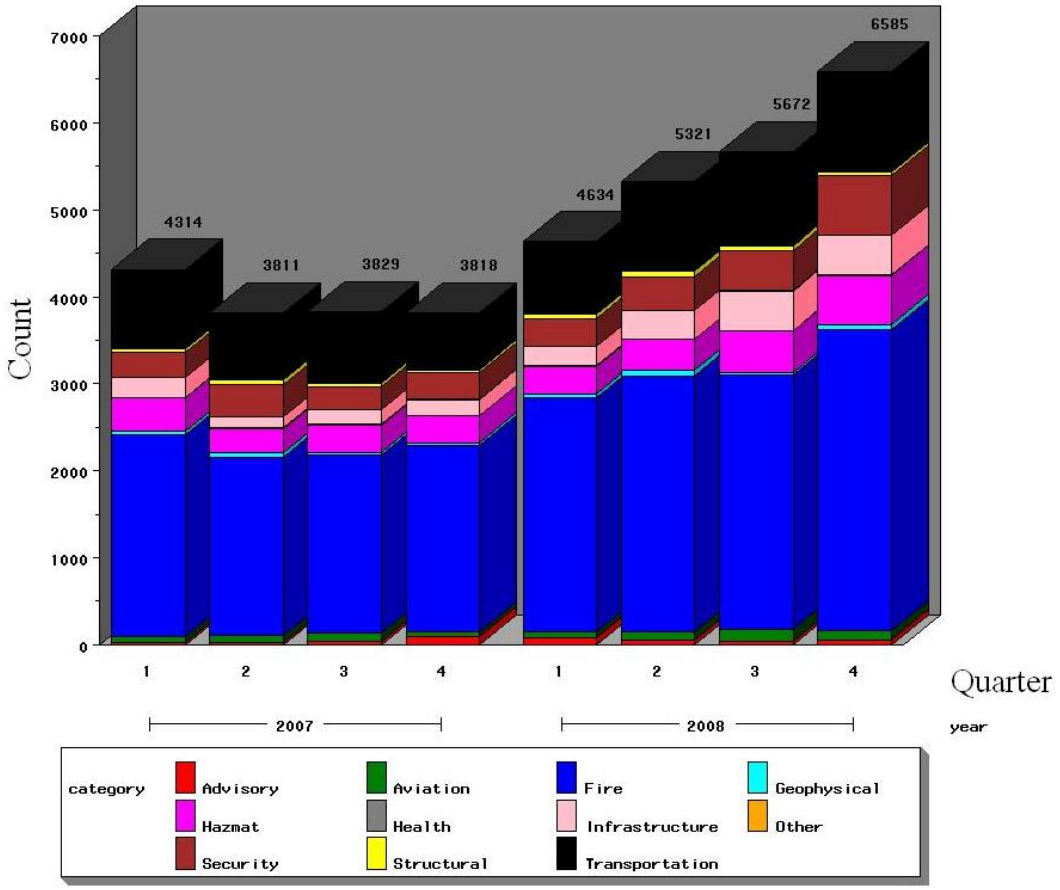

Figure 1. Quarterly plot of alert notifications

A plot of sequential data without reference to the quarter (see Figure 2) suggests the presence of a periodic *shock* to the system.

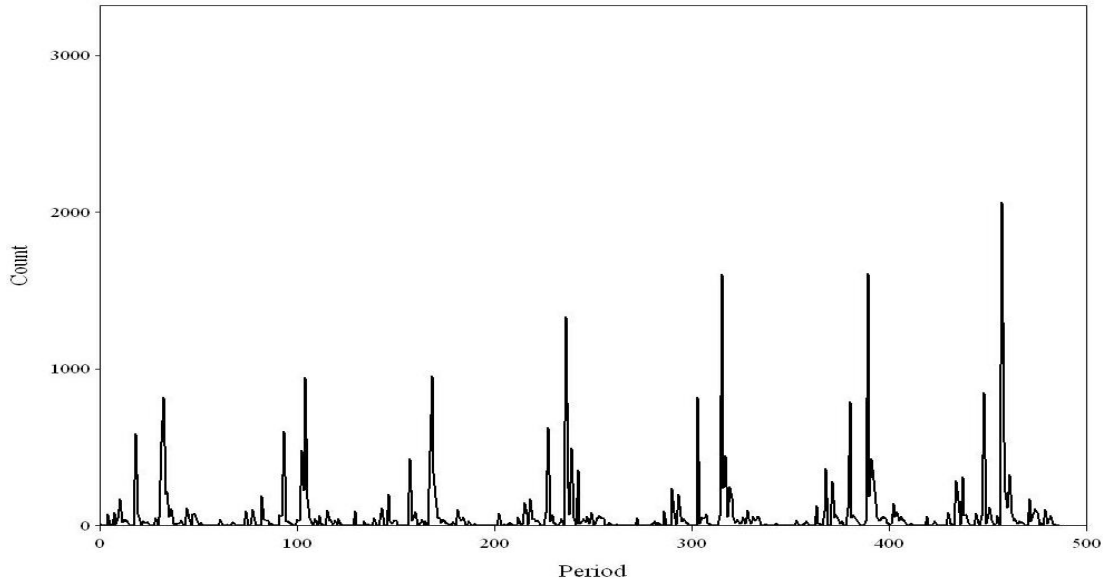

Figure 2. Incident alerts over time

Minor and moderate severity levels were found to explain 11.64% (*p<000.1*) of the variability in the number of incidents reported between 2007 and 2008. No evidence for linear correlation was found between year and quarter with the number of incidents reported. Residuals from the linear regression tests above are then analyzed using ARIMA modeling process.

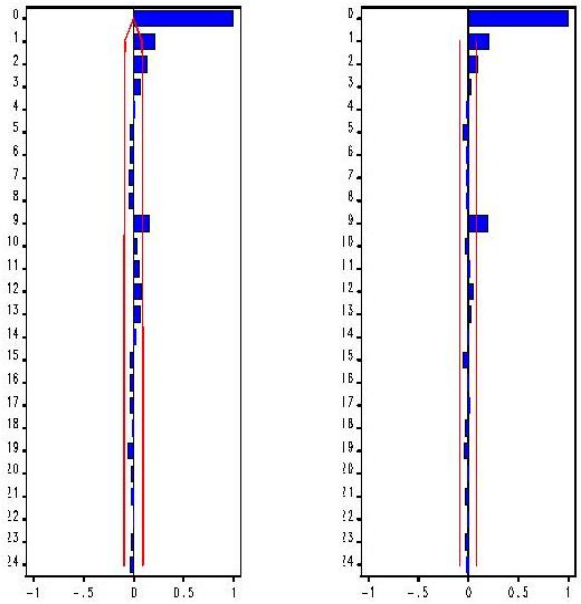

Figure 3. Autocorrelation (left) and partial autocorrelation (right) plots.

The autocorrelation plots from Figure 3 above suggests that an AR(1) model might be appropriate for explaining the variability in the data. An estimation method known as, Extended Sample Autocorrelation Function (ESACF), method is specified in the ARIMA procedure to estimate the model that best represents the data. Result from the estimation test shows that an ARMA (1,1) model is best fitted to the residuals from the linear regression test. This means there is an autoregressive (AR with coefficient = 0.537, *p=0.004*) and a moving average (MA with coefficient = 0.345, *p=0.0418*) component in the model.

Later diagnostics tests about the autocorrelation existing in the residuals, confirms that the model for incidents follow an ARMA(1,1) time series model at lag 1. The overall fitted model is presented as follows:

$$y_t^* = theta + 78.952(minor) + 53.995(moderate) + 0.345(a_t) + 0.537(y_{t-1}) + error \quad (2)$$

From the fitted model in (2) above, we can infer that the likelihood of future incident occurrence has one period ahead relationship with the current incident at a constant rate (0.537). The constant rate of incidence suggests that regardless of the category, the number of incident increases over time. It will therefore behoove security administrators to analyze incoming alerts on a shorter time frame (e.g., weekly) in order to detect shocks to the systems that can then be further investigated before it escalates to a much serious disaster.

DISCUSSION

Results from this exploratory study are discussed in terms of their implications to design as well as practice. A linear correlation of minor and moderate severity levels on the total number of incidents suggests that a slice into minor and moderate-severity level incidents should be conducted to understand potential causation effects so that it can be curtailed before it surfaces. More specifically, future studies will use the mechanism employed by Stroup et al (1989) where seasonal effects are introduced by basing the incident calculation on counts from comparable periods in past years to conduct extensive auxiliary analysis on slices of the data.

The data also suggest that there is no correlation between a high level of severity of the incidents and the number of incidents. This finding can be explained from statistical and pragmatic viewpoints. Statistically, the number of highly severe incidents is fewer than those of minor and moderate incidents; hence, they contribute less to the variability in the data. From a pragmatic point of view however, highly severe incidents such as a plane crash, bridge collapse, and wild fire, do not occur as often as moderate and minor incidents; hence their statistical insignificance.

Most importantly, the model confirms serial dependency of incidents, which is consistent with the notion that certain incidents (e.g., conscious, malicious attackers) build on previous attempts (Lee et al., 2006, Ning et al., 2002). Studies have shown that an attacker is likely to retool and re-attack using a different strategy if caught on a first attempt. Related studies in intrusion detection systems research area show that this is a familiar pattern from attackers (Ning et al., 2002) and that organization must keep abreast of these activities in order to prevent, control, or mitigate their effects. Our model in (2) provides empirical evidence that there is at least a 50 % chance that a previous incident (e.g., security breach) will reoccur in subsequent attacks. Hence, the

development of tools for operational emergency preparedness and management should be structured around the assumption that incidents are likely to reoccur once detected.

CONCLUSION

This study sought to investigate patterns embedded in alert notifications received over a period of two years. Several approaches to analyzing time stamp data were presented. The analyses suggest that measuring the passage of time provide a useful way of (1) screening data to characterize incidents, (2) identifying patterns of incidents that might be of interest to intelligence analysts, and (3) trends and relationships between incidents that may suggest potential future occurrence. These elements provide a set of converging measures helpful in understanding, interpreting and explaining the dimensions that are embedded in alert notifications.

An understanding of the existence and characteristics of patterns in alerts are important for organizations to be more agile and effective as they respond to incidents (e.g., conscious and unconscious attacks) that may induce staggering disasters if not detected early. This study provides evidence for the applicability of temporal analysis for investigating phenomena in the crisis management domain.

A direction of future work is to collect more region specific data to see if causality could be explained by other factors other than characteristics of the alerts. Case studies can then be conducted to surface variables and factors that might provide richer understanding of causalities in the trends and patterns observed in the alert notifications.

REFERENCES

1. Box, G. E. P. and Jenkins, G. M. (1976) Time series analysis: Forecasting and control, Holden-Day, San Francisco: CA.
2. Farrington, C. P., Andrews, N. J., Beale, A. D. and Catchpole, M. A. (1996) A statistical algorithm for the early detection of outbreaks of infectious disease, Journal of the Royal Statistical Society. Series A (Statistics in Society), 159, 3, 547-563.
3. Fawcett, T. and Provost, F. (1997) Adaptive fraud detection, Data Mining and Knowledge Discovery, 1, 3, 291-316.
4. Fawcett, T. and Provost, F. (1999) Activity monitoring: Noticing interesting changes in behavior, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,
5. Grossman, R. L. (2004) Alert management systems: A quick introduction, Springer US.
6. Janeja, V. P., Atluri, V., Gomaa, A., Adam, N., Bornhoevd, C. and Lin, T. (2005) Dm-ams: Employing data mining techniques for alert management, Proceedings of the 2005 national conference on Digital government research, Atlanta, Georgia
7. Lee, S., Chung, B., Kim, H., Lee, Y., Park, C. and Yoon, H. (2006) Real-time analysis of intrusion detection alerts via correlation, Computers & Security, 25, 3, 169-183.
8. Linden, A. (2003) Using time analysis for evaluating disease management program effectiveness. In: Academy Health Meeting, Vol. 20, pp., Nashville, Tennessee.
9. Malizia, A., Astorga-Paliza, F., Onorati, T., Diaz, P. and Aedo, I. (2008) Emergency alerts for all: An ontology based approach to improve accessibility in emergency alerting systems, 5th International ISCRAM Conference, Washington, DC, USA
10. Ning, P., Cui, Y. and Reeves, D. (2002) Analyzing intensive intrusion alerts via correlation. pp. 74--94. North Carolina State University at Raleigh.
11. SAS Institute Inc. (1996) Forecasting examples for business and economics using the sas system, SAS Institute Inc., Cary, NC.
12. Stroup, D. F., Williamson, G. D., Herndon, J. L. and Karon, J. M. (1989) Detection of aberrations in the occurrence of notifiable diseases surveillance data, Statistics in Medicine, 8, 3, 323-329.
13. van de Walle, B. and Turoff, M. (2007) Introduction, Communication of the ACM, 50, 3, 29-31.
14. Watier, L., Richardson, S. and Hubert, B. (1991) A time series construction of an alert threshold with application to s. Bovismorbificans in france, Statistics in Medicine, 10, 10, 1493-1509.