A Tool for Comparing Outbreak Detection Algorithms

Yasin Şahin
Hacettepe University,
Computer Engineering Department
06800 Ankara, Turkey
yasin@cs.hacettepe.edu.tr

Abstract. Despite of the main objective of recent biosurveillance researches is bioterrorist attack threats, detection of natural outbreaks are also being tried to solve by governments all over the world. Such that, international foundations as WHO, OECD and EU publish public declaration about necessity of an international central surveillance system. Each data source and contagious disease carries its own patterns. Therefore, standardizing the process of outbreak detection cannot be applicable. Various methods have been analyzed and published on test results in biosurveillance researches. In general, these methods are the algorithms in literature of SPC and Machine Learning, although specific algorithms have been proposed like Early Aberration Reporting System (EARS) methods. Differences between published results show that, the characteristic of time series are tested with algorithm and the chosen parameters of this algorithm are also determine results. Our tool provides preprocessing of data; testing, analyzing and reporting on anomaly detection algorithms specialized at biosurveillance. These functionalities make it possible to use outputs for comparing algorithms and decision making.

Keywords: outbreak detection, data smoothing, anomaly detection, time series analyzing.

1 Introduction

Outbreak is the occurrence of monitoring disease symptoms on a particular area and time more than normally expected. The expected value is quite hard to be expressed by a formula because of the impact of changing geographical locations and climatic conditions on it. Even in the cases of outbreak definition is given as statistically significant increase, it is not possible to see a description or a formula of this increase. Outbreaks which caused great deaths in the history also brought economic burdens together. During the world war I when the world’s population was around 2 billion, Spanish Flu that predicted spread of 500 million people, caused the death of 20 to 50 million people all over the world. Besides health problems, just SARS has costed 10 billions of dollars only for Hong Kong government as a recent experiment[1].

1 http://en.wikipedia.org/wiki/1918_flu_pandemic
1.1 Surveillance Systems

Surveillance systems have a major mission to struggle against the outbreak. Nowadays, the surveillance systems are specialized as the early warning systems which are being used for outbreak detection of time series provided from diagnostic or prediagnostic data[2]. Not only surveillance systems should be limited in use of bio-terrorist attacks, but also should address the natural outbreaks in order to reduce the number of deaths and prevent the financial losses. An idealized surveillance system consists of three stages. First stage is data collection phase. Second stage is, detecting outbreak from dataset provided by first stage. Warning the authorities about the outbreak possibility is the last step of the system. Outbreak detection phase should have adequate performance and this expectation relies on data collection stage’s organization quality. The data collected asynchronously causes unreliable results.

1.2 Biosurveillance Data

Modern surveillance systems are expected to identify a possible outbreak as quick as possible. For this purpose, prediagnostic data such as increases in drug sales, number of absenteeism or tendencies of searches on the internet by regions could be inputs for the detection of outbreak[3]. Common patient pattern may consist of three steps in general; internet searching, treatment with the available drugs and visiting hospital lastly. Prediagnostic data might make more sense for this estimation.

The datasets used in the outbreak detection contain patterns which lead to different statistics over time[4]. The occurrence of a surging on the surveillance between the weekdays and weekend can be an obvious sample to this. Considering the outpatient services does not work on weekends, visitors to the hospital earlier in the week and so the number of cases may increase cumulatively. Goldenberg observed that drug sales increases at weekends[5]. Therefore, the time series even recorded for the same purpose may lead to significant differences on what conditions are covered. Holidays and holiday returns could be a similar effect as the day of the week[6]. The complexity of surveillance data became more prominent when considering the foresight of seasonal effects and the changes in the air temperature[3].

1.3 Data Preprocessing Methods

Various algorithms are proposed in anomaly detection in time series. Outbreaks are also special cases that arise in time series. Traditional methods provide highly effective process monitoring independent identically distributed (iid) time series. To make an effective performance on surveillance data with traditional methods, the data should be made iid by cleaning from the explainable patterns such as day-of the week effect and seasonal effect[4]. Evaluation metrics of preprocessing methods are Mean Square Error (MSE), Root Mean Square Error (RMSE), Median Absolute Percent Error (MdAPE) and Median Absolute Deviation (MdAD) [3][3] [4].
**Regression**

Regression analysis is a statistical technique for estimating the relationships between dependent variable and independent variable(s). Number of regression methods has been recommended and eq. 1 is the notation all of these methods.

\[
Y \approx f(X, \beta)
\]

where \(Y\) is dependent variable, \(X\) is independent variable(s) and \(\beta\) is unknown parameters. Regression models are common methods in time series forecasting. It is also known, regression is most preferred forecasting technique for daily health series[2]. Regression models are common methods in time series forecasting. It is also known that regression is most preferred forecasting technique for daily health series. In the context of biosurveillance, linear regression handles the observations components such as DOW effect, seasonal effect, holiday effect etc. by dummy variables. Linear regression computes additive relation among variables. Sometimes predictors used have a multiplicative rather than additive effect[7]. This nonlinearity can be derived by using \(\log(y_t)\) rather \(y_t\).

**Adaptive Regression**

Adaptive regression model proposed for surveillance datasets by Burkom et al. [4] for the first time. Fricker et al. [8] applied Cusum to the prediction errors of adaptive regression and noted well-performing of that composed algorithm. Adaptive regression model predicts the day’s observation count by sliding baseline consist of most recent observations. Although Burkom proposed 8 weeks sliding window, Fricker reported that the best window size could be varying from 15 to 90. See Dunfee and Hegler et al. [9] to study how to determine size of sliding window in detail.

\[
y(i) = \beta_0 + \beta_1 \times (i - t + n + 1) + \varepsilon
\]

where \(y(i)\) is observation count, \(\beta_0\) is intercept, \(\beta_1\) is slope and \(\varepsilon\) is prediction error.

**7 Day Differencing**

7 day differencing proposed by Muscotello et al. [10] is known as the simplest forecasting algorithm. It predicts observation of the same day of previous week as next forecast: \(y_t = x_{t-7}\). Applying any detection algorithm after extracting residual from prediction may be a simple and useful algorithm.

**Moving Average**

Moving average smoothing generates a new smoothed dataset by using subsets belongs to raw dataset. Simple moving average predicts mean of sliding window for next observation: \(y_t = \frac{1}{n} \sum_{i=1}^{n} x_{t-i}\). Each observation within window has the same weight on this prediction. Actually some outbreak detection methods such as EARS methods involve simple moving average into method’s algorithm.

Weighted moving average is not a familiar smoothing technique for biosurveillance researches. It provides weighting for more recent observations.
where \( \alpha \) is smoothing constant, \( \beta \) is linear trend smoothing constant, \( F_t \) is estimated level term at time \( t \) and \( T_t \) is estimated trend term at time \( t \). \( H_{t+m} \) is overall forecast value of model for time \( t+m \). \( m \) should be 1 if the model supposed to estimate next observation count. Although it is known as a simple and flexible smoothing technique, it performs well while extracting trends from datasets.

Level term initializes as \( L_1 = x_1 \) and \( T_1 \) can be computed by 3 separate ways: \( T_1 = x_2 - x_1 \) or \( T_1 = \frac{(x_p - x_1)}{p-1} \) or \( T_1 = 0 \) where \( p \) is number of observation within cycle.

**Holt**

Charles H. Holt et al. [11] proposed Holt exponential smoothing method for the first time. Holt method adds linear trend component to computation in additional to EWMA smoothing method.

\[
F_t = \alpha x_{t-1} + (1 - \alpha)(F_{t-1} + T_{t-1}) \\
T_t = \beta (F_t - F_{t-1}) + (1 - \beta)T_{t-1} \\
H_{t+m} = F_t + mT_{t+m}
\]

(5)

where \( \alpha \) is smoothing constant, \( \beta \) is linear trend smoothing constant, \( F_t \) is estimated level term at time \( t \) and \( T_t \) is estimated trend term at time \( t \). \( H_{t+m} \) is overall forecast value of model for time \( t+m \). \( m \) should be 1 if the model supposed to estimate next observation count. Although it is known as a simple and flexible smoothing technique, it performs well while extracting trends from datasets.

**Holt&Winters**

Peter Winters et al. [12] suggested handling seasonal components extended by Holt’s method.

\[
L_t = \alpha (x_t - S_{t-s}) + (1 - \alpha)(L_{t-1} + T_{t-1}) \\
T_t = \beta (L_t - L_{t-1}) + (1 - \beta)T_{t-1} \\
S_t = \gamma (x_t - L_t) + (1 - \gamma)S_{t-s} \\
HW_{t+m} = L_t + mT_t + S_{t-s+m}
\]

(6)

where \( S_t \) is seasonal component of the model. Holt&Winters smoothing method cannot make any estimation at first seasonal cycle because of model’s dependence on values provided by previous cycle. Holt&Winters needs also be initialized. Proposed initializing equations given as:

\[
L_s = \frac{1}{s}(x_1 + x_2 + \cdots + x_s) \\
T_s = [(x_{s+1} + x_{s+2} + \cdots + x_{s+s}) - (x_1 + x_2 + \cdots + x_s)]/m^2 \\
S_i = x_i - L_s \quad (1 \leq i \leq s)
\]

(7)

Level and trend smoothing and their initializations are computed as additive model given above, but computations of seasonal component of model changes as eq. (8).
Lastly, multiplicative model of Holt&Winters’ seasonal initialization can be calculated as given eq.(9)[13].

\[ S_i = \frac{x_i}{L_s} \quad (1 \leq i \leq s) \]  

**Decomposition**

Time series consists of four components in general. These are **Trend**, **Cyclical**, **Seasonal** and **Irregular** components. In many time series, occurrence of a pattern could be in two ways; additive and multiplicative. An additive model assumes that, the difference between the Monday and Saturday values is approximately the same each week for daily data. A component of a multiplicative model causes a pattern with equal ratio in time series under the same conditions. To give an example, if observation count of each Monday is 10 greater than Sunday of the same week, it means additive model exists, but if observation count each Monday is greater than Sunday of the same week 3/2 times, it means multiplicative model exists.

\[ x_t = T_t + C_t + S_t + I_t \]  
\[ x_t = T_t \times C_t \times S_t \times I_t \]  
\[ \log(x_t) = \log(T_t) + \log(C_t) + \log(S_t) + \log(I_t) \]

Decomposition is a statistical technique to remove explainable pattern(s) from time series for smoothing. Decomposition has separate means for different study areas. While a statistician use it for curve-fitting, an electronics engineer may use it prevent to signals from noisy [14]. Subtraction of an additive component or division of a multiplicative component is adequate to decompose a pattern.

The day of week effect and seasonal effect are also smoothed by decomposition. Applying linear regression to dataset as given eq. (13) can provide the predictions values of the effects.

\[ y_t = \beta_0 + \beta_{mon} I_{mon} + \cdots + \beta_{sun} I_{sun} + \beta_1 \sin \left( \frac{2\pi}{365.25} \right) i + \beta_2 \cos \left( \frac{2\pi}{365.25} \right) i \]

where \( \beta_0 \) is intercept, the \( i \)'s are dummy indicators, where \( I_{mon} = 1 \) on Monday, otherwise \( I_{mon} = 0 \) to disable it. \( \sin \left( \frac{2\pi}{365.25} \right) i \) and \( \cos \left( \frac{2\pi}{365.25} \right) i \) components are estimator of the seasonal effects, where \( i \) is the distance to last 1 October at time \( t \).

## 2 Outbreak Detection Methods

Control charts (Shewhart charts) were proposed by Walter Shewhart purpose of quality control at the beginnings of 1920s [15]. Statistical Process Control (SPC) has built
on Shewhart charts in times. Control charts check two sided thresholds called Upper Control Limit (UCL) and Lower Control Limit (LCL) to monitor the process. Falling outside of this acceptable interval between limits generates an alarm in case of possibility of an anomaly existence. Surveillance systems only interested in UCL threshold. SPC algorithms such as Shewhart, EWMA and Cusum have been widely studied for adapting to biosurveillance. Methods extended by SPC such as EARS methods or methods does not belong to SPC also proposed with several studies such as Negative Binomial Cusum (NBC) [16], Historical Limits Method (HLM) [17], Hidden Markov Model (HMM) [18] etc.

Shewhart
In the biosurveillance studies, the Shewhart Control Charts, which are basic for the EARS System models, are discussed frequently.

\[ S_t = \frac{x_t - \mu}{\sigma} \]  \hspace{1cm} (14)

where \( \mu \) and \( \sigma \) are estimated from observation counts of days that does not carry outbreak. It is checked that how much an observation moves away from the average with Shewhart method given eq. (14). Most important advantage of Shewhart for the biosurveillance systems is warning a quick response with a little historical data when the outbreak is sharp and suddenly raised. It is powerful when the shifts of process are not small.

Cumulative Sum
Counter to Shewhart, Cusum method is used to determine small shifts [19] encountered during process which is proposed by Page [20]. Calculating recent observations effect into score make Cusum method skillful algorithm in detecting small shifts. If the observations show seasonal waving, Cusum might be weak to make a right decision about process [21].

\[ CS_t^+ = \text{maks}[0, x_t - (\mu_0 + k)] + CS_{t-1} \]  \hspace{1cm} (15)

In the context of biosurveillance, it needs to be prevented falling below zero in order not to have negative meaningless values obtained by Cusum because of increasing is only be dealt with [22]. \( k \) reference value could be middle of standard deviation [9]. See Fricker et al. [8] for choosing \( k \) in details.

EWMA
EWMA, which is also a smoothing method, is successful in the determination of the small changes during process.

\[ E_t = \alpha x_t + (1 - \alpha)E_{t-1} \]  \hspace{1cm} (16)

where, \( \alpha \) is EWMA constant can be interval of \( 0 < \alpha \leq 1 \) equation.

\[ \sigma_{ewma}^2 = \frac{\alpha}{2 - \alpha} \sigma^2 \]

\[ E_t^+ = E_0 + k\sigma_{ewma} \]  \hspace{1cm} (17)
Mean value of observations called target value is selected for initialization of EWMA score: \( E_0 = \mu \). When the calculated EWMA score moves away from target value for \( k \) times, it exceeds the threshold and anomaly will have been detected. See Lucas et al. [28] which has study about determining \( k \) and threshold.

\[
E_t = \text{maks}[\mu, \alpha x_t + (1 - \alpha)E_{t-1}]
\]  

(18)

If EWMA algorithm is applied to processed data instead of raw data, eq. (18) could be changed as \( E_t = \text{maks}[0, \alpha x_t + (1 - \alpha)E_{t-1}] \).

**EARS Methods**

EARS methods adapted to detect sharp and sudden increments caused by bioterrorism without long historical baseline. EARS has three methods called C1, C2 and C3. Although EARS methods are claimed as Cusum-based, in fact they are based on Shewhart.

EARS methods may be the most discussed methods within surveillance context. For this reason, a researcher, who proposes a new algorithm, considers comparing the algorithm performance with EARS methods firstly.

\[
C_1(t) = \frac{x_t - \bar{X}_t}{s_t}
\]  

(19)

where \( \bar{X}_t \) and \( s_t \) are mean and standard deviation of sliding baseline with fixed size 7. Minimum standard deviation can be applied to statistic to prevent it from zero division as 0.2. Alarm is raised when the statistic of C1 exceeds threshold. Although CDC, proposer of EARS methods, tests C1 and C2 with 3 as threshold, Fricker et al. [8] proposed adjusted thresholds determined by fixed ATFS to obtain comparable performance. The only difference between C2 and C1 is two days lag preference of C2. C2 calculates both of standard deviation and mean of baseline excluding last 2 days: \( \bar{X}_t = \frac{1}{7} \sum_{i=t-3}^{t-9} x_i \) and \( s_t = \frac{1}{6} \sum_{i=t-3}^{t-9} [x_i - \bar{X}_t]^2 \). Thus, manipulation of an outbreak caused by step-wise increments within time series could be prevented. C3, another method of EARS, is calculated by an equation uses C2 statistics at time t with most recent two statistics as given below.

\[
C_3(t) = \sum_{i=t}^{t+2} \text{max}[0, C_2(i) - 1]
\]  

(20)

While threshold of C1 and C2 set as 3, C3 is being tested with 2 within EARS.

**Negative Binomial Cusum**

NBC is Cusum based method minimized false alert rates with right configuration. This ability can make it useful for designing a system purposes low false alert rate. Some of biosurveillance researches noted that NBC has better performance than EARS methods over outbreak detection [18]. NBC is described by two parameters; \( r \) and \( c \). Anomaly is determined by parameter \( c \).
\[ c_0 = \frac{\mu}{\mu - \sigma^2} \]
\[ r = \frac{\mu^2}{\mu - \sigma^2} \]

where mean and variance are being calculated based on sliding baseline. If there is an outbreak, in-control level \( c_0 \) will shift to out-of-control level \( c_1 \). Decision interval of NBC can be monitored with the equation given below [16].

\[
NBC_t = \text{max}(0, NBC_{t-1} + x_t - k) \\
k = r \cdot \ln \left[ \frac{c_0(1+c_1)}{c_1(1+c_0)} \right] / \ln \left[ \frac{1+c_0}{1+c_1} \right].
\]

Watkins et al. [18] proposed out of control level as value of a fixed interval of two baseline standard deviations greater than in control level \( c_0 \).

**Historical Limits Method**

HLM was designed to deal with seasonal effects [24].

\[ \frac{x_0}{\mu} > 1 + \frac{2\sigma_x}{\mu} \]  

(22)

where \( \mu \) and \( \sigma_x \) are the mean and standard deviation of the historical data. If the size of sliding baseline set as 7 days for 3 years historical data, 45 daily points totally joins equation. 15 observations for each year come from 7 preceding days, 7 subsequent days and current day. As seen in eq. (22), HLM needs a great historical data. This requirement makes it incomparable with EARS because of EARS’ purposes. A study claims that HLM has better performance than EARS methods has also tested statistic with weekly datasets [24].

### 3 System Design of Tool

The tool has developed with one of the most popular programming language Java. All abilities provided by the tool could have been implemented with Matlab or R except availability on Internet and user interfaces. Author’s experiment over Java and flexibility of Java made it one step ahead while language choosing phase. The tool was properly designed and implemented for requirements of Model View Controller (MVC) architecture. Thus, by the abstraction between layers, modifying business logic or user interface could be adapted quickly.

ZK ajax framework which is suitable for MVC preferred while developing. ZK has both of free community edition and enterprise edition but free version, Community Edition licensed with LGPL, was found adequate to the tool’s development process. A small database was also modeled within tool development. Another free product has been chosen as Database Management System. PostgreSQL is a free DBMS application provides any functionalities tool needs. Communication between application and database was handled by Hibernate, which has been incorporated to standard library of Java Persistence API.
Apache Common Math was used to linear regression operations. Preprocessing methods such as decompositions, Simple Linear Regression, Multiple Linear Regression and Adaptive Regression exploits apache common math library. Comma-separated values (CSV) file format was chosen in order to users could freely and easily edit their

---

2 [http://books.zkoss.org/wiki/ZK_Developer's_Reference/MVC](http://books.zkoss.org/wiki/ZK_Developer's_Reference/MVC)
files according to tool’s file structure. CSV file format based on plain text, so any text editor is able to edit it. File structure has to be as date, observation count, anomaly existence. OpenCsv library imported to project to have an ability to make csv files I/O operations on Java language. At the end of process, result charts generated by JFreeChart library.

The only requirement of web-based tool is having a user profile within the system.

4 Outbreak Algorithm Testing

In this section, process of an anomaly detection beginning by uploading a dataset will be told. At the end of the process, comparison of the algorithms will be done.

Each step of the process can be saved by the user with this tool. Rest of the process can be done with separate ways to get different results. First step of the process after uploading is determining a smoothing process. One or more preprocessing method can be applied to a dataset or skipping this step may be more suitable for the researcher.

![ODS process flow chart](image)

**Fig. 3.** ODS process flow chart

Report#1 is a chart for reporting processed dataset with its raw dataset. Report#2 shows daily scores of the algorithm with threshold. Report#3 provides XY line charts such as ROC, AMOC or whatever user wants to place within X,Y coordinates. AUC is calculated during XY line drawing to measure algorithm’s performance.
After smoothing operations, detection methods would be applied to source. User can test and see the comparing results of a method with various parameters or various methods with similar parameters. To see summary of performance statistic of results, user needs to filter to make a comparing list. At this step, user can display a daily score chart or provides XY line charts such as ROC, AMOC or whatever user wants to place within X,Y coordinates. AUC is calculated during XY line drawing to measure algorithm’s performance.

A sample analyzing process is comparison of Cusum and EWMA methods. Dataset we tested is available from CDC website. Both of two algorithms were tested with ATFS values as 20, 50, 100 and 200. The threshold adjustment to achieve a particular ATFS estimated empirically.

Algorithm #1: Cusum method was tested with dataset that extracted residual from adaptive regression proposed as Burkom et al. [4]. Cusum scored reset after detection to prevent cumulative increasing. Reference value \( k \) was calculated from 
\[
\frac{1}{2} \left( \frac{(n+1)(n+2)}{n(n-1)} \right) \sigma_e
\]
equation proposed by Fricker et al. [8].

Algorithm #2: EWMA method was applied to dataset that had been smoothed by DOW effect and seasonal effect decomposition. Sunday selected as reference day and both of seasonal components applied for multiplicative decomposition. EWMA was applied to processed data with smoothing factor as 0.3.

Parameters were not the best performers of the algorithms for applied dataset. The main idea of sampling is sampling the process within tool.

5 Results

Statistical evaluation metrics widely chosen for outbreak detection are sensitivity, specificity and timeliness. Starting and ending of the outbreaks are not same as anomalies which defined for other contexts. In general, definition of sensitivity is the proportion of conditional anomalies which signed as anomaly for it. But if the anomaly is an outbreak, then conditional anomaly is duration instead of observation. For example, consider an outbreak lasts 10 sequential days and it is detected only for 7th day of outbreak duration. Sensitivity would be measured 0.1 as usual way, but 1.0 for the context of biosurveillance. For this case, timeliness will be valuable metric to handle delay. Specificity is the proportion of conditional negatives which signed as negative for it. In other words, it is the probability that an outbreak test will correctly indicates that an individual does not have a particular condition.

Evaluation of the algorithms given previous section is based on sensitivity, specificity and timeliness metrics. The performance of each algorithm was also evaluated by comparing the area under the receiver operating characteristic (ROC) curve (AUC). According to AUC values, EWMA has better performance than Cusum. Areas under ROC curve of EWMA and Cusum are 0.021 and 0.009 respectively. Actually performance of Cusum could only approach to EWMA’s for applied data source with chosen parameters when ATFS was leveled to 20. Both algorithms have similar
timeliness and specificity but EWMA has better sensitivity than Cusum. It can be seen that, larger ATFS values make Cusum weak for detecting anomalies with adaptive regression.

Obtained performances of the algorithms were modeled without concern about finding best algorithms. Explaining skills of the tool we developed was purposed. These scores do not make better EWMA than Cusum or reverse. When we apply the algorithms to chosen dataset gives us these results. Defining parameters, determining comparison metrics or preprocessing methods are related to research’s expectations, so the performance evaluation weighting depend on that expectations.

Using the skills of the tool, a user can processes modeling and get reports showed on fig. 3.

Table 1. Evaluation Metrics of Cusum and EWMA for different threshold adjusted to achieve given ATFS

| ATF S (day) | Sensitivity (%) | Specificity (%) | Timeliness (day) |
|------------|----------------|-----------------|-----------------|
|            | Cusum          | EWMA            | Cusum           | EWMA           | Cusum | EWMA |
| 20         | 63,889         | 63,889          | 97,204          | 97,707         | 3,478 | 3,000 |
| 50         | 25,000         | 44,444          | 98,764          | 99,218         | 4,778 | 3,688 |
| 100        | 13,889         | 36,111          | 99,731          | 99,739         | 4,000 | 4,692 |
| 200        | 2,778          | 22,222          | 99,463          | 99,896         | 4,000 | 5,500 |

6 Conclusion

“The syndromic surveillance literature documents quite a number of efforts to develop and measure the performance of various individual detection algorithms. However, it contains very few comparisons between algorithms in order to assess the relative strengths and weaknesses of the algorithms. It is as if everyone is trying to develop a new hammer, but few are comparing among the hammers to determine which are to be preferred.”

Dunfee and Hegler et al. [9]

Anomaly detection in time series is a common research area of network security (intrusion detection), credit card security (fraud detection) and epidemiology (disease outbreak detection). SPC methods and Machine Learning algorithms have been imported to outbreak detection recent years. Outbreak detection differs from other anomaly detections because of surveillance data features.

Determining the outbreak detection algorithms may play a key role for the surveillance systems. To give a right decision about algorithm, questions listed below must be answered smoothly:

- What is the outbreak’s spreading type?
- Which type of outbreak spreading is more important to detect?
- What is the acceptable false alert rate?
What is the least true alert rate?
What is the longest acceptable out of control ARL (timeliness).

When we can give right answers to these questions, it is possible to determine the most convenient algorithm for data source. The tool we developed can give an idea about corresponding of expectations and results.

Adding new functionalities to search out patterns within time series and more algorithms from the literature of SPC and Machine Learning can be future works. Additional information such as provided from GIS can be added to process for detecting outbreak rapidly and robustly. It is also seen that, available surveillance datasets need to be increased to have field knowledge more than we do. An ideal surveillance system can be achieved by two ways. First one is defining the outbreak definition statistically for each disease and the second one is developing a system that is able to determine to the best algorithm related to expectations.

References

1. Siu, A., Wong, Y.C.R.: Economic Impact of SARS: The Case of Hong Kong. Asian Economic Papers 3(1), 62–83 (2004)
2. Lotze, T.H.: Anomaly Detection in Time Series: Theoretical and Practical Improvements for Disease Outbreak Detection, Ph.D. Thesis. University of Maryland (2009)
3. Shmueli, G., Burkom, H.S.: Statistical Challenges Facing Early Outbreak Detection in Biosurveillance. Technometrics (Special Issue on Anomaly Detection) (1), 39–51 (2010)
4. Burkom, H.S., Murphy, S.P., Shmueli, G.: Automated Time Series Forecasting for Biosurveillance. Statist. Med. 26(22), 4202–4218 (2007)
5. Goldenberg, A., Shmueli, G., Caru, R.A., Fienberg, S.E.: Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. Proceeding of the National Academy of Sciences 99(8), 5237–5240 (2002)
6. Fienberg, S.E., Shmueli, G.: Statistical issues and challenges associated with rapid detection of bio-terrorist attack. Statistics in Medicine 24(4), 513–529 (2005)
7. Kleinman, K., Lazarus, R., Platt, R.: A Generalized Linear Mixed Models Approach for Detecting Incident Clusters of Disease in Small Areas, with an Application to Biological Terrorism. Am. J. Epidemiol. 159(3), 217–224 (2004)
8. Fricker Jr., R.D., Hegler, B.L., Dunfee, D.A.: Comparing syndromic surveillance detection methods: EARS’ versus a CUSUM-based methodology. Statistics in Medicine (17), 3407–3429 (2008)
9. Dunfee, D.A., Hegler, B.L.: Biological Terrorism Preparedness: Evaluating the Performance of the Early Aberration Reporting System (EARS) Syndromic Surveillance Algorithms (2007)
10. Muscatello, D.: An adjusted cumulative sum for count data with day-of-week effects: application to influenza-like illness. In: Syn. Surv. Conf., Boston (2004)
11. Holt, C.C.: Forecasting seasonals and trends by exponentially weighted moving averages. International Journal of Forecasting 20(1), 5–10 (2004)
12. Winters, P.R.: Forecasting Sales by Exponentially Weighted Moving Averages. Management Science 6(3), 324–342 (1960)
13. Makridakis, S.G., Wheelwright, S.C., Hyndman, R.J.: Forecasting: Methods and Applications. John Wiley and Sons, New York (1998)
14. Brown, R.G.: Smoothing, forecasting and prediction of discrete time series. Dover Publications, Mineola (2004)

15. Shewhart, W.A.: Quality Control Charts. Bell System Technical Journal 5(4), 593–603 (1926)

16. Hawkins, D.M., Olwell, D.H.: Cumulative Sum Charts and Charting for Quality Improvement. Springer, New York (1998)

17. Stroup, D.F., Williamson, G.D., Her, J.L.: Detection of aberrations in the occurrence of notifiable diseases surveillance data. Statistics in Medicine 8(3), 323–329 (1989)

18. Watkins, R.E., Eagleson, S., Veenendaal, B., Wright, G., Plan, A.J.: Applying cusum-based methods for the detection of outbreaks of Ross River virus disease in Western Australia. BMC Medical Informatics and Decision Making 8(37) (2008)

19. Gan, F.F.: An optimal design of CUSUM control charts for binomial counts. Journal of Applied Statistics 20(4), 445–460 (1993)

20. Page, E.S.: Continuous Inspection Schemes. Biometrika Trust 41(1/2), 100–115 (1954)

21. Kartal, M.: İstatistiksel Kalite Kontrolu, Erzurum: Şafak Yaynevi (1999)

22. Wagner, M.M., Moore, A.W., Aryel, R.M.: Handbook of biosurveillance. Academic Press Inc., Amsterdam (2006)

23. Lucas, J.M., Sanucci, M.S.: Exponentially Weighted Moving Average Control Schemes: Properties and Enhancement. Technometrics 32(1) (1990)

24. Pelecanos, A.M., Ryan, P.A., Gatt, M.L.: Outbreak detection algorithms for seasonal disease data: a case study using Ross River virus disease. BMC Medical Informatics and Decision Making 10(74) (2010)