The Use of a CUSUM Residual Chart to Monitor Respiratory Syndromic Data

Huifen Chen
Department of Industrial and Systems Engineering,
Chung-Yuan University, Chung Li, TAIWAN; Email: huifen@cycu.edu.tw
Chaosian Huang
Department of Manufacturing, Lextar Electronics Corp., Hsinchu, Taiwan

December 7, 2012

Abstract

We construct a respiratory syndromic surveillance mechanism for the respiratory syndrome. The data used for illustration are the daily counts of respiratory syndrome sampled from the National Health Insurance Research Database in Taiwan. The population size is 160,000. We first fit a regression model with an ARIMA (autoregressive integrated moving average) error term to the data and then construct CUSUM (cumulative sum) residual charts to detect the aberration in visit frequencies of respiratory syndrome. The day-of-the-week, seasonal, and holiday effects are considered in the regression model. Our results show that the CUSUM residual chart is useful in detecting abnormal increases of respiratory symptoms.

Keywords: ARIMA; CUSUM chart; regression analysis; respiratory syndrome; syndromic surveillance; time series

1 Introduction

An epidemic, or outbreak, means that the occurrence of a disease is at an unexpectedly high frequency (Baxter et al., 2000). Recent epidemics, e.g., SARS in 2003, avian influenza in 2003-2005 and H1N1 in 2009, have caused deaths of many people in the world. Early detection of outbreaks is important for timely public health response to reduce morbidity and mortality. By early detecting the aberration of diseases, sanitarians can study or research into the causes of diseases as soon as possible and prevent the cost of the society and medical treatments. Traditional disease-reporting surveillance mechanisms might not detect outbreaks in their early stages because laboratory tests usually take long time to confirm diagnoses.

Syndromic surveillance was developed and used to detect the aberration of diseases early (Henning, 2004). The syndromic surveillance mechanism is to collect the baseline data of prodromal phase symptoms and detect the aberration of diseases from the expected baseline by placing the variability of data from the expected baseline. Such surveillance methods include the SPC (statistical process control) methods, scan statistics, and forecasting methods (Tsui et al. 2008). See Section 2 for literature review.

In this work, we study the implementation of CUSUM (CUmulative SUM) residual chart for detecting the outbreak of the respiratory syndrome in Taiwan. Since the daily visits of the respiratory syndrome are time series data with seasonal effect, we use a regression model with an ARIMA (AutoRegressive Integrated Moving Average) error term to model the daily counts from ambulatory care clinic data. The CUSUM of residuals are then plotted in the CUSUM chart for detecting unusual increase in daily visits. The test data are the 2005-2008 ambulatory care clinic data from the National Health Insurance Research Database (NHIRD) in Taiwan.

This paper is organized as follows. In Section 2, we review related literature. In Section 3, we summarize the data, propose a regression model whose error term follows an ARIMA model, and construct the CUSUM chart using the residuals. The regression model is fitted to the daily counts data of respiratory symptoms in years 2005 and 2006. In Section 4, we assess the performance of the CUSUM residual chart by applying it to monitor the daily counts data in 2007 and 2008. The conclusion is given in Section 5.

2 Literature review

We discuss here the syndromic surveillance methods including the forecast-based, scan statistics, and SPC-based methods. Detailed reviews can be found in Tsui et al. (2008, 2011) and Unkel et al. (2012).

The forecast-based methods are useful to model non-stationary baseline data. To detect aberration, an upper threshold value is determined using the fitted forecast model. When the actual value of the response variable exceeds the threshold, an outbreak alarm is sent. Two popular forecasting methods are time-series and regression models. Goldenberg et al. (2002) used the AR (AutoRegressive) model to forecast the over-the-counter medication sales of the anthrax and built the upper prediction interval to detect the outbreak. Reis and Mandl (2003) developed generalized models for expected emergence-department visit rates by fitting historical data with trimmed-mean seasonal models and then fitting the residuals with ARIMA models. Lai (2005) used three time series models (AR, a combination of growth curve fitting and ARMA error, and ARIMA) to detect the outbreak of the SARS in China.

The scan statistics have been widely used in retrospective detection of temporal clustering of diseases (Glaz et al. 2001). For example, Heffernan et al. (2004) applied the scan statistic method to monitor respiratory, fever diarrhea and vomiting syndromes by the chief complaint data of the emergency department. This method scans a window of time; if an observed cluster of diseases is significantly unusual for the underlying probability model, a signal is sent. Scan statistics are also used for prospective detection of unusual clusters, where the time-window length can vary over a range of values (Kulldorff 2001, Naus and Wallenstein 2006).

Recently the control charts have been applied in health-care and public-health surveillance (Woodall 2006). The SPC methods were first applied in the industrial statistical control (Montgomery 2005). Since the Shewhart chart is insensitive at detecting small shifts, CUSUM and exponentially weighted moving average (EWMA) charts are more commonly used in public-health surveillance than the Shewhart chart. Hutwagner et al. (1997) developed a computer algorithm based on the CUSUM scheme to detect salmonella outbreaks using the laboratory-based data. Morton et al. (2001) applied Shewhart, CUSUM and EWMA charts to detect and monitor the hospital-acquired infections. Their results showed that when used together, Shewhart and EWMA work well for monitoring bacteremia and multiresistant organism rates and that CUSUM and Shewhart charts are suitable for monitoring surgical infection.

Modifications of CUSUM charts were proposed for the incidence rate with a changing population size. Some modifications are based on a Poisson model (e.g., Mei et al. 2011, Jiang et al. 2012) and some are based on a Bernoulli model (e.g., Sego et al. 2008). If a Poisson model is used, the counts of incidents at regular time intervals are needed. The Bernoulli CUSUM chart can detect increase in incidence rate earlier than the Poisson CUSUM chart because the Bernoulli CUSUM chart monitors sequential Bernoulli data without waiting for aggregated counts (Shu et al. 2011).

Some literature modeled the baseline data with a forecast model before applying an SPC scheme because the baseline data may not be independent and identically distributed and the mean of the data may be a function of time. Rogerson and Yamada (2004) applied a Poisson CUSUM residual chart to detect the lower respiratory tract infections for 287 census tracts simultaneously, where the baseline data were fitted by logistic regression models. Miller et al. (2004) used the regression model with autoregressive error to fit the influenzalike illness data in an ambulatory care network, where the regression terms include weekend, holiday and seasonal adjustments (sine and cosine functions). They then used the standardized CUSUM residual chart for detecting the outbreak. Fricker et al. (2008) applied the adaptive regression model with day-of-the-week effects using an 8-week sliding baseline and then used the CUSUM chart of the adaptive regression residuals. They showed that this approach performed better than the Early Aberration Reporting System (EARS) for baseline data with day-of-the-week effects.

Literature comparing the three types of methods exists. Cowling et al. (2006) compared time series, regression, and CUSUM models using influenza data from Hong Kong and the US. They found that the time series model was the best in the Hong Kong setting, while both the time series and CUSUM models worked equally well on the US data. Woodall et al. (2008) showed that the CUSUM chart approach is superior to the scan statistics. Han et al. (2010) compared CUSUM, EWMA and scan statistics for surveillance data following Poisson distributions. Their results showed that CUSUM and EWMA charts outperformed the scan statistic method.

3 Methods

Here we discuss the implementation of CUSUM residual charts for respiratory syndromic surveillance. Before implementing the CUSUM residual chart, a regression model with an ARIMA error term is fitted to the daily counts of respiratory syndrome. The test data are the 2005 to 2008 daily counts of respiratory syndrome from the National Health Insurance Research Database (NHIRD) in Taiwan. Section 3.1 introduces the data source, Section 3.2 summarizes the data, Section 3.3 shows the regression model for the respiratory daily counts, and Section 3.4 shows the monitoring scheme, CUSUM residual chart.

3.1 *Data source*

The data used in this study are the 2005 to 2008 daily counts (i.e. the number of daily visits) of respiratory syndrome for 160,000 people sampled from the National Health Insurance Research Database (NHIRD) by the Bureau of National Health Insurance, Taiwan. Patients' diagnoses in NHIRD were encoded using the ICD-9-CM (International Classification of Diseases, 9th Revision, Clinical Modification Reference) code. In this study, the ICD-9 codes of the respiratory syndrome are adopted from the syndromic classification criteria of the Centers for Disease Control and Prevention (CDC) in the United States (CDC 2003) as listed in Appendix A.

3.2 *Data summary*

Here we summarize the daily-counts data of respiratory syndrome from 2005 to 2008 with population size 160,000. Figure 1, the run chart of the daily counts from 2005 to 2008, shows that the daily counts are time-series data with seasonal variation. The epidemic peak period for the respiratory syndrome is from November to April next year.

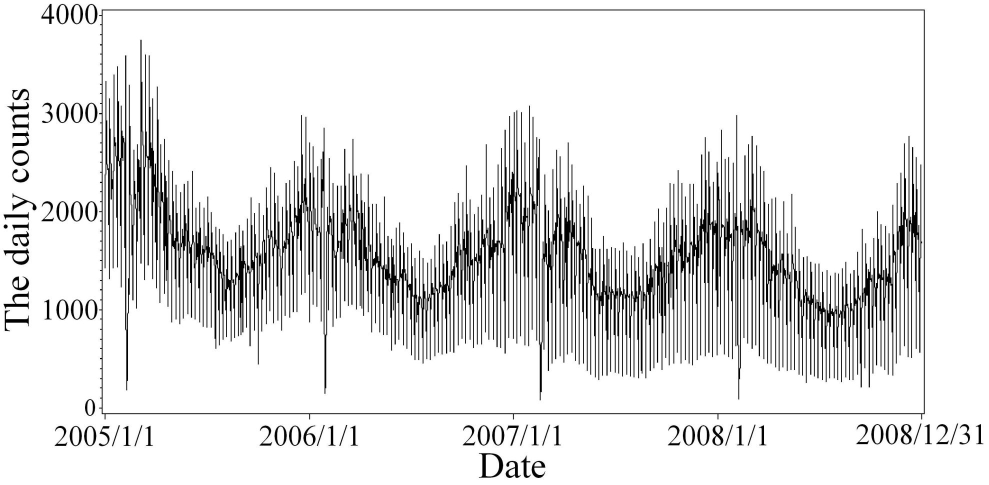


Figure 1: The daily counts of the respiratory syndrome from 2005 to 2008

Besides the month-of-the-year effect discussed above, the daily counts also have the day-of-the-week and holiday effects. Figure 2 provides a closer look of Figure 1 with time zone from January 14 to February 16 in 2006 in Subfigure 2(a) and April 1 to 17 in 2006 in Subfigure 2(b). Subfigure 2(a) shows the day-of-the-week effect: for a given week, the daily count is highest on Monday (unless it is a holiday) and lowest on Sunday because most clinics are closed on Sunday. The holiday effect is shown in both Subfigures 2(a) and 2(b). Subfigure 2(a) shows that the daily counts are yearly lowest during the Chinese New Year holidays (January 29 to 31, the first three days in the Chinese New Year) because the clinic service of the outpatient department usually is not available during this period. Subfigure 2(b) shows that the daily count also drops in the national holidays to about the same level as that on Sunday.

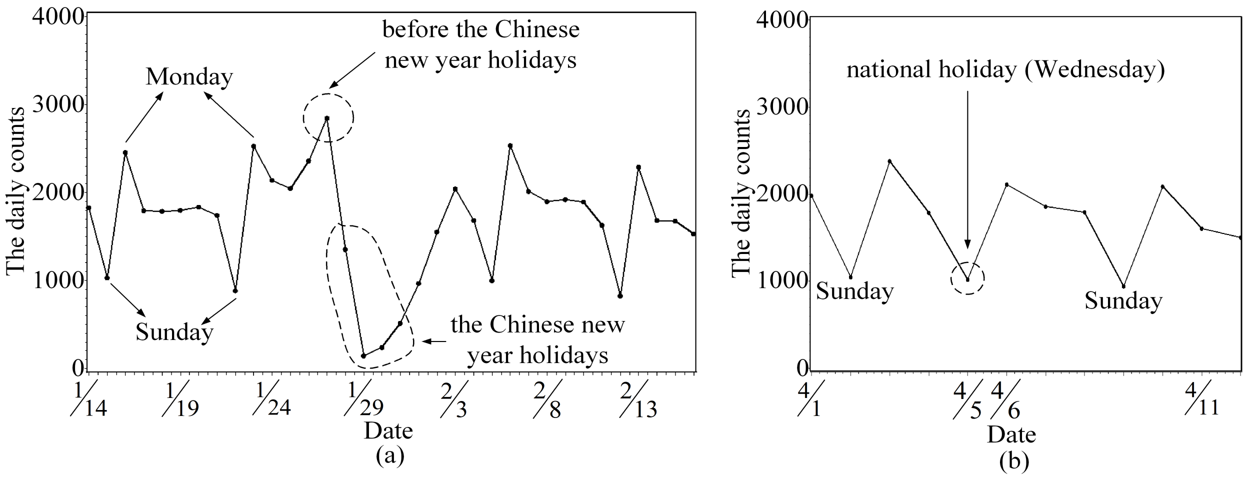


Figure 2: Subfigure (a) shows the daily counts of the respiratory syndrome from January 14 to February 16 in 2006; Subfigure (b) shows the daily counts from April 1 to April 17 in 2006 with April 5 (Wednesday) being a national holiday

3.3 *Regression models with an ARIMA error term*

Since the daily counts are time series data with seasonal variation, we use the regression model with an ARIMA error term to fit the daily counts data for the respiratory syndrome. For normality, we use the Box-Cox transformation to transform the daily counts data. The general form of the regression model is

$$W_t = \mu_t + \epsilon_t, \qquad t = 1, 2, \ldots, \tag{1}$$

where $W_t = Y_t^\lambda$ is the transformed response variable, λ is a constant, Y_t denotes the daily count for the respiratory syndrome in day t, μ_t is the mean response depending on a set of predictor variables (e.g., the day of the week) and ϵ_t is the error term following an ARIMA(p, d, q) process with non-negative integers p, d, and q referring to the orders of autoregressive, integrated, and moving-average parts in the ARIMA model (Box et al. 1994).

The predictor variables are set based on the day-of-the-week, month-of-the-year, holiday, typhoon, and trend effects. For the day-of-the-week effect, we set dummy variables D_1 to D_6 to stand for Monday to Sunday, excluding the reference day Wednesday. For the month-of-the-year effect, dummy variables M_1 to M_{11} stand for January to December, excluding the reference month June. For the holiday effect, we set dummy variables C_{B2}, C_{B1}, C, C_4, C_5, and C_6 to stand for the day that is two days before, just before, within the first three days of, the fourth day of, the fifth day of, and the sixth day of the Chinese New Year, respectively. We also set dummy variables D, D_{ten}, and H to denote the Dragon-Boat Festival, the National Day, and the other national holidays, and set H_A to denote the day after one of these holidays (except the Chinese-New-Year holidays). The dummy variables T and T_A stand for a typhoon day and the day after, respectively. The sine and cosine functions are used in the model to show the seasonal effect. Finally, we include the trend variable t and the interaction terms (e.g., the interaction effect between typhoon and Monday is denoted by $T{*}D_1$). In summary, the mean response is modeled as

$$\begin{aligned}\mu_t \;=\; & \beta_0+\sum_{i=1}^{6}\beta_iD_i+\sum_{j=1}^{11}\beta_{6+j}M_j+\beta_{18}C_{B2}+\beta_{19}C_{B1}+\beta_{20}C+\beta_{21}C_4+\beta_{22}C_5+\beta_{23}C_6\\ & +\beta_{24}D+\beta_{25}D_{ten}+\beta_{26}H+\beta_{27}H_A+\beta_{28}T+\beta_{29}T_A+\beta_{30}\sin(\frac{2\pi t}{365.25})\\ & +\beta_{31}\cos(\frac{2\pi t}{365.25})+\beta_{32}C_{B2}{*}D_1+\beta_{33}C_{B2}{*}D_7+\beta_{34}C_{B1}{*}D_1+\beta_{35}C_{B1}{*}D_7\\ & +\beta_{36}C{*}D_1+\beta_{37}C{*}D_7+\beta_{38}C_4{*}D_1+\beta_{39}C_4{*}D_7+\beta_{40}C_5{*}D_1+\beta_{41}C_5{*}D_7\\ & +\beta_{42}C_6{*}D_1+\beta_{43}C_6{*}D_7+\beta_{44}D{*}D_1+\beta_{45}D{*}D_7+\beta_{46}D_{ten}{*}D_1+\beta_{47}D_{ten}{*}D_7\\ & +\beta_{48}H{*}D_1+\beta_{49}H{*}D_7+\beta_{50}H_A{*}D_1+\beta_{51}H_A{*}D_7+\beta_{52}T{*}D_1+\beta_{53}T{*}D_7\\ & +\beta_{54}T_A{*}D_1+\beta_{55}T_A{*}D_7+\beta_{56}t. \qquad (2)\end{aligned}$$

Once the fitted regression model with a fitted ARIMA error term is obtained, the residuals can be used to construct CUSUM charts for monitoring abnormal increases in the respiratory-syndrome frequency.

3.4 *CUSUM residual charts*

The CUSUM chart is a useful tool to monitor the occurrence of epidemics. The residuals calculated from the fitted regression model with an ARIMA error term can be used to construct an upper one-sided standardized CUSUM chart (Montgomery 2005) for detecting abnormal increases in daily counts of respiratory syndrome.

The CUSUM value at time t, called C_t^+, is defined as

$$C_t^+=\max(0,R_t/\sigma_R-k+C_{t-1}^+),\qquad t=1,2,\ldots, \qquad (3)$$

where C_0^+ is the starting value of the CUSUM statistic, $R_t=W_t-\hat{W}_t$ is the residual at time t, $\hat{W}_t$ is the prediction of W_t, σ_R is the standard deviation of the residual R_t, and the constant

k is half of the shift amount in mean. The value of σ_R can be estimated by the square root of the mean square error (MSE) of the fitted regression model. In our application, we set $C_0^+ = 0$ and $k = 0.5$ (for detecting the situation where the mean of the residual increases by one standard deviation σ_R).

Like Miller et al. (2004), we set the upper control limit h of the upper one-sided standardized CUSUM chart so that the in-control average run length (denoted as ARL_0) is 50. To compute the value of h, we use the Siegmund's approximation (Siegmund 1985 and Montgomery 2005, p. 396) for the average run length (ARL):

$$\mathrm{ARL}_\delta \approx \frac{e^{-2\Delta b} + 2\Delta b - 1}{2\Delta^2},$$

where ARL_δ is the ARL when the process mean shifts by δ standard deviations, $b = h + 1.166$, $\Delta = \delta - k$. By letting $\mathrm{ARL}_0 = 50$ (with $\delta = 0$), the upper control limit $h = 2.225$.

4 Results

To illustrate the monitoring method discussed in Section 3, we use the 2005 to 2006 daily counts data to fit a regression model in Section 4.1. The fitted regression model is then applied to the 2007 and 2008 data to construct upper one-sided standardized CUSUM charts in Section 4.2 for purposes of illustration and validation.

4.1 *The fitted regression model*

We use the regression model with an ARIMA error term, shown in Section 3.3, to fit the 2005 to 2006 daily counts data of respiratory syndrome. The estimated value of λ in the Box-Cox transformation is $\lambda = 0.96$. The level of significance is set to 5%.

Table 1 lists the estimates of regression coefficients and the associated p-values. Only significant predictor variables are listed. For better prediction, all interaction terms that have data are included in the fitted model even if their p-values are higher than 5%. (Some interaction terms have no data so that they are not included in the model and some have only a few data resulting in large p-values.) The fitted ARIMA(p, d, q) model for the error term is

$$\begin{aligned}\tilde{\epsilon}_t &= \frac{(1-\hat{\omega}_1 B)(1-\hat{\omega}_2 B^7)(1-\hat{\omega}_3 B^9 - \hat{\omega}_4 B^{11} - \hat{\omega}_5 B^{12} - \hat{\omega}_6 B^{17} - \hat{\omega}_7 B^{22})a_t}{(1-\hat{b}_1 B)(1-\hat{b}_2 B^7)} \\ &= \hat{\phi}_1 \tilde{\epsilon}_{t-1} + \hat{\phi}_7 \tilde{\epsilon}_{t-7} + \hat{\phi}_8 \tilde{\epsilon}_{t-8} + (1 - \sum_{k=1}^{30} \hat{\theta}_k B^k) a_t, \end{aligned} \tag{4}$$

where B is the backward shift operator and $\hat{\phi}$'s and $\hat{\theta}$'s are

$$\begin{aligned}&\hat{\phi}_1 = \hat{b}_1,\, \hat{\phi}_7 = \hat{b}_2,\, \hat{\phi}_8 = -\hat{b}_1\hat{b}_2,\, \hat{\theta}_1 = \hat{\omega}_1,\, \hat{\theta}_7 = \hat{\omega}_2,\, \hat{\theta}_8 = -\hat{\omega}_1\hat{\omega}_2,\, \hat{\theta}_9 = \hat{\omega}_3,\, \hat{\theta}_{10} = -\hat{\omega}_1\hat{\omega}_3, \\ &\hat{\theta}_{11} = \hat{\omega}_4,\, \hat{\theta}_{12} = \hat{\omega}_5 - \hat{\omega}_1\hat{\omega}_4,\, \hat{\theta}_{13} = -\hat{\omega}_1\hat{\omega}_5,\, \hat{\theta}_{16} = -\hat{\omega}_2\hat{\omega}_3,\, \hat{\theta}_{17} = \hat{\omega}_6 + \hat{\omega}_1\hat{\omega}_2\hat{\omega}_3, \\ &\hat{\theta}_{18} = -(\hat{\omega}_1\hat{\omega}_6 + \hat{\omega}_2\hat{\omega}_4),\ \hat{\theta}_{19} = -\hat{\omega}_2\hat{\omega}_5 + \hat{\omega}_1\hat{\omega}_2\hat{\omega}_4,\ \hat{\theta}_{20} = \hat{\omega}_1\hat{\omega}_2\hat{\omega}_5,\ \hat{\theta}_{22} = \hat{\omega}_7, \\ &\hat{\theta}_{23} = -\hat{\omega}_1\hat{\omega}_7,\ \hat{\theta}_{24} = -\hat{\omega}_2\hat{\omega}_6,\ \hat{\theta}_{25} = \hat{\omega}_1\hat{\omega}_2\hat{\omega}_6,\ \hat{\theta}_{29} = -\hat{\omega}_2\hat{\omega}_7,\ \hat{\theta}_{30} = \hat{\omega}_1\hat{\omega}_2\hat{\omega}_7, \\ &\text{and the rest of } \hat{\theta}_k\text{'s are zero.}\end{aligned} \tag{5}$$

The fitted ARIMA model has an AR order $p = 8$, integrated order $d = 0$, and MA order $q = 30$. The estimates $\hat{b}$'s and $\hat{\omega}$'s are listed in Table 1. The fitted distribution of the white noises $\{a_t\}$ is normal with mean 0 and variance being the MSE = 6961.2. The residual analysis (not reported here) indicates that the assumption of identically and independently normally distributed white noises $\{a_t\}$ is appropriate.

Table 1: The coefficient estimates of the fitted regression model and their associated p-values, where the standard-error estimates (S.E.) for the coefficient estimates are shown in parentheses

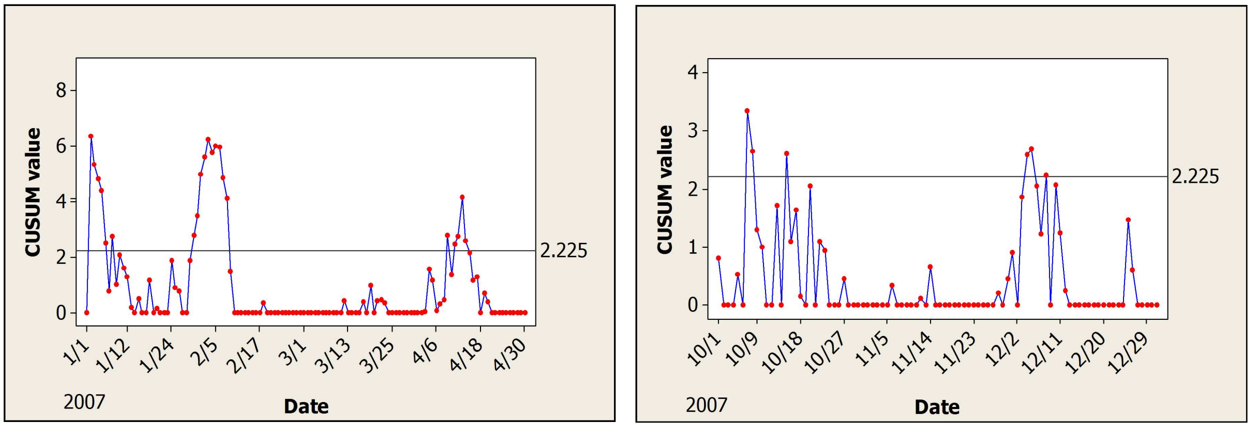
Parameter	Estimate (S.E.)	p-value	Parameter	Estimate (S.E.)	p-value
b_1	0.709 (0.048)	< 0.0001	$\beta_{24}(D)$	-210.4 (47.6)	< 0.0001
b_2	0.943 (0.017)	< 0.0001	$\beta_{25}(D_{ten})$	-293.0 (66.4)	< 0.0001
ω_1	0.222 (0.067)	0.0009	$\beta_{26}(H)$	-547.3 (30.6)	< 0.0001
ω_2	0.573 (0.042)	< 0.0001	$\beta_{27}(H_A)$	275.3 (31.0)	< 0.0001
ω_3	-0.125 (0.039)	0.0013	$\beta_{28}(T)$	-252.4 (45.9)	< 0.0001
ω_4	-0.138 (0.038)	0.0003	$\beta_{29}(T_A)$	176.5 (40.1)	< 0.0001
ω_5	-0.087 (0.039)	0.0246	$\beta_{30}(\sin)$	175.6 (54.1)	0.0012
ω_6	-0.150 (0.039)	0.0001	$\beta_{31}(\cos)$	301.1 (51.9)	< 0.0001
ω_7	0.136 (0.039)	0.0005	$\beta_{33}(C_{B2}{*}D_7)$	-19.7 (103.1)	0.8486
β_0(Intercept)	1317 (75.6)	< 0.0001	$\beta_{34}(C_{B1}{*}D_1)$	-143.6 (105.2)	0.1723
$\beta_1(D_1)$	476.3 (45.7)	< 0.0001	$\beta_{36}(C{*}D_1)$	-545.5 (75.4)	< 0.0001
$\beta_6(D_6)$	-674.3 (46.3)	< 0.0001	$\beta_{37}(C{*}D_7)$	375.6 (81.2)	< 0.0001
$\beta_7(M_1)$	-112.1 (43.4)	0.0098	$\beta_{41}(C_5{*}D_7)$	31.1 (76.3)	0.6837
$\beta_8(M_2)$	-218.3 (38.2)	< 0.0001	$\beta_{43}(C_6{*}D_1)$	-208.1 (102.4)	0.0421
$\beta_{18}(C_{B2})$	256.4 (71.7)	0.0003	$\beta_{46}(D_{ten}{*}D_1)$	-236.7 (92.6)	0.0106
$\beta_{19}(C_{B1})$	540.2 (73.6)	< 0.0001	$\beta_{49}(H{*}D_7)$	510.2 (60.8)	< 0.0001
$\beta_{20}(C)$	-1217.9 (49.9)	< 0.0001	$\beta_{50}(H_A{*}D_1)$	-205.6 (61.1)	0.0008
$\beta_{21}(C_4)$	-605.1 (54.6)	< 0.0001	$\beta_{51}(H_A{*}D_7)$	-169.1 (57.9)	0.0035
$\beta_{23}(C_6)$	288.8 (67.2)	< 0.0001	$\beta_{52}(T{*}D_1)$	-729.4 (81.4)	< 0.0001

In Table 1, Column 1 lists the nine time-series and ten regression coefficients (with the corresponding predictors indicated in parentheses), Column 2 lists the estimates of the coefficients in Column 1 and their standard-error estimates in parentheses, Column 3 lists the associated p-values, and Columns 4 to 6 are the same as Columns 1 to 3 but for the rest of 19 regression coefficients.

Table 1 shows that the coefficient estimate for D_1 (Monday) is positive (476.3) and the coefficient estimate for D_6 (Sunday) is negative (-674.3), meaning that the mean daily count is expected to increase by 476.3 on Monday and decrease by 674.3 on Sunday. The coefficient estimates for C_{B2} and C_{B1} (indicators for the days that are two days and one day before the Chinese New Year, respectively) are positive, meaning that the mean daily counts increase when the Chinese New Year holidays are approaching. The coefficient estimate for C (indicator for the first three days in the Chinese New Year) is -1217.9 resulting the lowest mean daily counts in a year, because only medical service of emergency departments is available during these three holidays. Comparing coefficient estimates for C_4 (indicators for the fourth day in the Chinese New Year) and C_6 (indicator for the sixth day in the Chinese New Year), we see that the daily count increases gradually after the first three holidays in the Chinese New Year. Comparing the national holiday variables, the coefficient estimates for D (the Dragon-Boat Festival) and D_{ten} (indicator for the National Day) are higher than the coefficient estimate of H (indicator for the other national holidays). The coefficient estimate for T shows that the reduction effect of typhoons in the daily count is about the same as that for the predictors D and D_{ten}. The coefficients of H_A (indicator for the day after a holiday) and T_A (indicator for the day after a typhoon) show that the mean daily count after a holiday or day off due to typhoon goes up but is lower than the mean daily count on Monday. The coefficient value of T_A is smaller than that of H_A because typhoons often occur in summer, an epidemic off-peak period. The coefficients of the paired interaction terms show that if a holiday, typhoon day, or the day before/after them is on Monday or Sunday, the mean daily count is different from that in other weekdays.

4.2 *CUSUM residual charts*

We apply the upper one-sided standardized CUSUM chart (Section 3.4) on the 2007 and 2008 respiratory data based on the fitted regression model (Section 4.1) with the 2005 and 2006 data. Recall that the CUSUM value is $C_t^+ = \max(0, R_t/\sigma_R - k + C_{t-1}^+)$, $t = 1, 2, \ldots$, where the residuals R_t is the difference between W_t and its fitted value $\hat{W}_t$ and σ_R is estimated by $\sqrt{\text{MSE}}$



(a) $^1\!/_1 \sim {}^4\!/_{30}$

(b) $^{10}\!/_1 \sim {}^{12}\!/_{31}$

Figure 3: The CUSUM charts for (a) January 1 to April 30 and (b) October 1 to December 31 in 2007

$= \sqrt{6961.2} = 83.4.$

Figure 3 contains the CUSUM charts for peak periods (January 1 to April 30 in Subfigure 3(a) and October 1 to December 31 in Subfigure 3(b)) in 2007. A few alarms are shown in Subfigure 3(a): January 2 to 6, January 30 to February 8, and April 9. In Figure 3(b), the alarms occur in October 7, October 15, and December 4. Figures 4 and 5 show the CUSUM charts for peak periods in 2008 (January 1 to April 30 for Figure 4 and September 1 to December 31 for Figure 5). In Figure 4, the alarms occur during January 2 to 9, and February 1 to 10, while in Figure 5, the alarms occur during September 29 to 30. Since the CUSUM charts are constructed based on real data, it is difficult to identify whether these alarms are false. All the alarms seem to be reasonable except the one occurring during September 29 to 30 in 2008. Because of typhoon, September 29 (Monday) 2008 was not a work day. However, many clinics were still open since the weather was better than expected. Consequently the number of visits is higher than that computed from the fitted model. Such a false alarm is caused by an unpredicted event and can be identified easily.

5 Conclusions

This work discusses the implementation of CUSUM residual charts for monitoring daily counts of respiratory syndrome. The population size is 160,000. Before using the CUSUM chart, we fit a regression model with an ARIMA error term to the daily counts data. The numerical results indicate that the CUSUM residual chart seems to work well in showing abnormal increases in

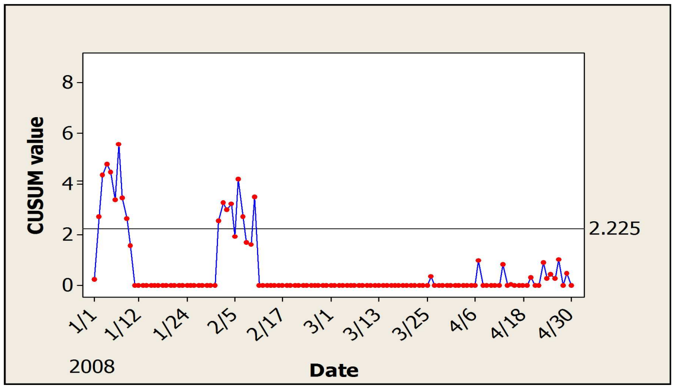


Figure 4: The CUSUM chart from January 1 to April 30 in 2008

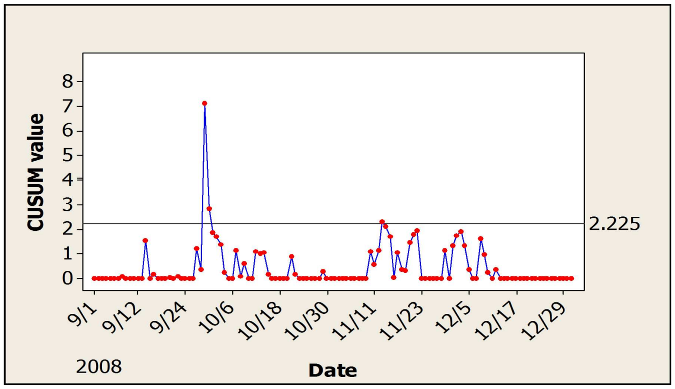


Figure 5: The CUSUM chart from September 1 to December 31 in 2008

daily counts of respiratory syndrome.

We conclude this section by discussing three issues:

1. Some research uses the weekly counts data, rather than daily counts, to eliminate the day-of-the-week effect. Our numerical results, however, show that the model for weekly counts is not much simpler. Since the weekly data are not as effective to identify outbreaks as daily data, this work chooses to use the daily data.

2. Our fitted regression model is based on historical data of the past two years. The time window can be longer so that more data can be used for model fitting. The shortage though is that the coefficient estimates would have larger variance and hence the prediction interval would be wider. Furthermore, the behavior of daily counts may not be the same each year, using historical data that are long ago may hurt the prediction accuracy for the future observations.

3. In this work, some interaction terms can not be included in the regression model because of

lack of data. To overcome this situation, one way is to include more historical data for model fitting. The payoff is inducing more variation in parameter estimates as discussed in the previous issue. Another way is to modify the regression model based on expert experiences so that the interaction terms with no data can be included in the model.

Acknowledgments

This study is based in part on data from the National Health Insurance Research Database provided by the Bureau of National Health Insurance, Department of Health and managed by National Health Research Institutes in Taiwan. The interpretation and conclusions contained herein do not represent those of Bureau of National Health Insurance, Department of Health or National Health Research Institutes in Taiwan. This research was supported by Taiwan National Science Council via Grant NSC100-2221-E-033-045.

Appendix A: Respiratory-syndrome ICD-9-CM code

In this study, we adopt the respiratory syndrome definitions from the syndromic classification criteria of the Centers for Disease Control and Prevention (CDC) in the United States (CDC 2003). The ICD-9 codes of the respiratory syndrome are listed in Table 2.

Table 2: The list of respiratory ICD-9-CM code

020.3	020.4	020.5	021.2	022.1	460	462	463	464.00	464.01	464.10	464.11
464.20	464.21	464.30	464.31	464.4	464.50	464.51	465.0	465.8	465.9	466.0	466.11
466.19	478.9	480.8	480.9	482.9	483.8	484.5	484.8	485	486	490	511.0
511.1	511.8	513.0	513.1	518.4	518.84	519.2	519.3	769	786.00	786.06	786.1
786.2	786.3	786.52	799.1	075	381.00	381.01	381.03	381.04	381.4	381.50	381.51
382	382.0	382.00	382.01	382.02	382.4	382.9	461.0	461.1	461.2	461.3	461.8
461.9	493.00	493.01	493.02	493.10	493.11	493.12	493.90	493.91	493.92	511.9	514
518.0	518.81	518.82	782.5	784.1	786.05	786.07	786.09	786.50	786.51	786.59	786.7
786.9	003.22	031.0	031.8	031.9	032.0	032.1	032.2	032.3	032.89	032.9	033.0
033.1	033.8	033.9	034.0	052.1	055.1	055.2	073.0	073.7	073.8	073.9	079.0
079.1	079.2	079.3	079.6	079.81	098.6	114.5	114.9	115.00	115.05	115.09	115.10
115.15	115.90	115.95	115.99	116.0	116.1	117.1	117.3	117.5	130.4	136.3	480.0
480.1	480.2	481	482.0	482.1	482.2	482.30	482.31	482.32	482.39	482.40	482.41
482.49	482.81	482.82	482.83	482.84	482.89	483.0	483.1	484.1	484.3	484.6	484.7
487.0	487.1	487.8									

References

Baxter, R., Rubin, R., Steinberg, C., Carroll, C., Shapiro, J., and Yang, A. (2000) Assessing core capacity for infectious diseases surveillance. Falls Church (VA): The Lewin Group, Inc.

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1994) *Time Series Analysis: Forecasting and Control, revised edition.* Holden-Day, San Francisco.

Centers for Disease Control and Prevention, USA. (2003) Syndrome definitions for diseases associated with critical bioterrorism-associated Agents. Available from: http://emergency.cdc.gov/surveillance/syndromedef/. (November, 2012)

Cowling, B. J., Wong, I. O. L., Ho, L.-M., Riley, S., and Leung, G. M. (2006) Methods for monitoring influenza surveillance data. *International Journal of Epidemiology* **35**, 1314–1321.

Fricker, R. D. Jr., Hegler, B. L., and Dunfee, D. A. (2008) Comparing syndromic surveillance detection methods: EARS versus a CUSUM-based methodology. *Statistics in Medicine* **27**, 3407–3429.

Glaz, J., Naus, J., and Wallenstein, W. (2001) *Scan Statistics.* Springer-Verlag Inc., New York.

Goldenberg, A., Shmueli, G., Caruana, R. A., and Fienberg, S. E. (2002) Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 5237–5240.

Han, S. W., Tsui, K. L., Ariyajunya, B., and Kim, S. B. (2010) A comparison of CUSUM, EWMA, and temporal scan statistics for detection of increases in Poisson rates. *Quality and Reliability Engineering International* **26(3)**, 279–289.

Heffernan, R., Mostashari, F., Das D., Karpati A., Kulldorff M., and Weiss, D. (2004) Syndromic surveillance in public health practice, New York City. *Emerging Infectious Diseases* **10**, 858–864.

Henning, K. J. (2004) What is syndromic surveillance? *Morbidity and Mortality Weekly Report* **53** (Supplement), 5–11.

Hutwagner, L. C., Maloney, E. K., Bean, N. H., Slutsker, L., and Martin, S. M. (1997) Using laboratory-based surveillance data for prevention: An algorithm for detecting Salmonella outbreaks. *Emerging Infectious Diseases* **3**, 395–400.

Jiang, W., Shu, L. J., Zhao, H. H., and Tsui, K. L. (2013) CUSUM procedures for health care surveillance. *Quality and Reliability Engineering International* **29**(6), 883–897.

Kulldorff, M. (2001) Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society A* **164**(1), 61–72.

Lai, D. (2005) Monitoring the SARS epidemic in China: A time series analysis. *Journal of Data Science* **3**, 279–293.

Mei, Y. J., Han, S. W., and Tsui, K. L. (2011) Early detection of a change in Poisson rate after accounting for population size effects. *Statistica Sinica* **21**, 597–624.

Miller, B., Kassenborg, H., Dunsmuir, W., Griffith, J., Hadidi, M., Nordin, J. D., and Danila, R. (2004) Syndromic surveillance for influenzalike illness in an ambulatory care network. *Emerging Infectious Diseases* **10**, 1806–1811.

Montgomery, D. C. (2005) *Introduction to Statistical Quality Control, 5^{th} edition.* John Wiley & Sons, Inc., New York.

Morton, A. P., Whitby, M., McLaws, M. L., Dobson, A., McElwain, S., Looke, D., Stackelroth, J., and Sartor, A. (2001) The application of statistical process control charts to the detection and monitoring of hospital-acquired infections. *Journal of Quality in Clinical Practice* **21**, 112–117.

Naus, J., and Wallenstein, S. (2006) Temporal surveillance using scan statistics. *Statistics in Medicine* **25(2)**, 311–324.

Reis, B. and Mandl, K. D. (2003) Time series modeling for syndromic surveillance. *BMC Medical Informatics and Decision Making* **3**:2.

Rogerson, P. A. and Yamada, I. (2004) Approaches to syndromic surveillance when data consist of small regional counts. *Morbidity and Mortality Weekly Report* **53** (Supplement), 79–85.

Sego, L. H., Woodall, W. H., Reynolds, Jr, M. R. (2008) A comparison of surveillance methods for small incidence rates. *Statistics in Medicine* **27(8)**, 1225–1247.

Shu, L.J., Jiang, W., and Tsui, K. L. (2011) A comparison of weighted CUSUM procedures that account for monotone changes in population size. *Statistics in Medicine* **30**, 725–741.

Siegmund, D. (1985) *Sequential Analysis: Tests and Confidence Intervals.* Springer-Verlag, New York.

Tsui, K. L., Chiu, W., Gierlich, P., Goldsman, D., Liu, X., and Maschek, T. (2008) A review of healthcare, public health, and syndromic surveillance. *Quality Engineering* **20(4)**, 435–450.

Tsui, K. L., Wong, S. Y., Jiang, W., and Lin, C. J. (2011) Recent research and developments in temporal and spatiotemporal surveillance for public health. *IEEE Transactions on Reliability* **60(1)**, 49–58.

Unkel, S., Farrington, C. P., Garthwaite, P. H., Robertson, C., and Andrews, N. (2012). Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society A* **175**(1), 49–82.

Woodall, W. H. (2006) The use of control charts in health-care and public-health surveillance. *Journal of Quality Technology* **38**, 89–104.

Woodall, W. H., Marshall, J. B., Joner, M. D., Jr, Fraker S. E., and Abdel-Salam, A.-S. G. (2008) On the use and evaluation of prospective scan methods for health-related surveillance. *Journal of the Royal Statistical Society A* **171**(1), 223–237.

BIOGRAPHIES

Huifen Chen is professor of Industrial and Systems Engineering Department at Chung-Yuan University, Taiwan. She completed her Ph.D. in Industrial Engineering at Purdue University in 1994 and master in statistics at Purdue University in 1990. Her research interests include statistical process control, public health, random-vector generation, and stochastic root finding. Her email address is huifen@cycu.edu.tw.

Chaosian Huang received a master degree in Industrial and Systems Engineering Department at Chung-Yuan University, Taiwan in 2009. He is currently a senior engineer in Department of Manufacturing, Lextar Electronics Corp., Hsinchu, Taiwan.