Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–14

DOI: 10.1177/0962280215592427
smm.sagepub.com

SAGE

CUSUM control charts to monitor series of Negative Binomial count data

Airlane Pereira Alencar,[1] Linda Lee Ho[2] and Orlando Yesid Esparza Albarracin[1]

Abstract

To detect outbreaks of diseases in public health, several control charts have been proposed in the literature. In this context, the usual generalized linear model may be fitted for counts under a Negative Binomial distribution with a logarithm link function and the population size included as offset to model hospitalization rates. Different statistics are used to build CUSUM control charts to monitor daily hospitalizations and their performances are compared in simulation studies. The main contribution of the current paper is to consider different statistics based on transformations and the deviance residual to build control charts to monitor counts with seasonality effects and evaluate all the assumptions of the monitored statistics. The monitoring of daily number of hospital admissions due to respiratory diseases for people aged over 65 years in the city São Paulo-Brazil is considered as an illustration of the current proposal.

Keywords

Surveillance, seasonal effect, deviance residual, likelihood ratio

1 Introduction

The concept of quality control is largely applied in production processes, but it can be also applied in services. Monitoring a production process can avoid large damages if problems are detected as soon as they occur. In the public health, this concept is also employed in surveillance of diseases. For example, monitoring the mortality or morbidity of a disease to decide whether it reached an epidemic level (or not) is crucial to take urgent decisions and to plan health services.

Statistical process control has been used to monitor production processes, and recently it also has been applied in surveillance problems in public health.[1] The last approaches based on these methods have been proposed to detect outbreaks of infectious diseases which can be viewed as shifts in the industrial process production.[2] However, in surveillance problems, the monitored variable presents

[1]Department of Statistics, University of São Paulo, São Paulo, Brazil
[2]Department of Production Engineering, University of São Paulo, São Paulo, Brazil

Corresponding author:
Linda Lee Ho, Department of Production Engineering, University of São Paulo, Av Professor Almeida Prado, Travessa 2, 128, São Paulo 05508070, Brazil.
Email: lindalee@usp.br

some specific features. It is a time series of count data, like the daily number of hospitalizations, which is affected by the population size. Additionally, it presents seasonality pattern since more admissions are expected from several diseases during the winter season. Due to the theses characteristics, usually the hospitalization rates are analyzed. Many methods of surveillance in public health have been suggested in the literature to detect small shifts.[1,2]

Cumulative Sum (CUSUM) control charts are frequently used in health surveillance as alternative to the Shewhart control chart. They are more efficient than the Shewhart chart for detection of small shifts[3,4] because they take into account the cumulative information of the sequence of observations. Rogerson and Yamada (RY),[5] Woodall,[1] and Höhle and Paul[6] proposed CUSUM charts for count data.

Recently, Höhle and Paul[6] compare many methods of infectious disease surveillance and emphasize the importance of including the seasonality effect, the population in risk, and available explanatory variables. Basically, they recommend employing generalized linear models (GLMs) to monitor these count series. Control charts for regression models under a Gaussian distribution are found in engineering[7] and in health surveillance.[5,8] In the literature, usually the Poisson distribution is adopted to model count series,[9] but recently, a Negative Binomial distribution has been used.[6] In general, most papers[8,9] build their control charts proposing transformed variables to normalize the count data and thereafter the control limits are calculated based on the Gaussian distribution. The assumption of normality, although important, usually is not checked in most analyses.

The main goal of the current paper is to evaluate CUSUM control charts to monitor count series with seasonal effects. A GLM is fitted under a Negative Binomial distribution with a logarithm link function, including the population size as offset and a set of covariates. The monitoring of daily number of hospital admissions due to respiratory diseases for people aged over 65 years in the city São Paulo-Brazil is considered as an illustration of the current proposal.

The main contribution is to propose a CUSUM chart based on other transformed variables and on the deviance residual (DR) for the Negative Binomial distribution, as this later is supposed to follow a Gaussian distribution, and evaluate all these CUSUM control charts. The CUSUM chart proposed by RY,[5] extended for the Negative Binomial distribution, and the CUSUM chart proposed by Höhle and Paul[6] based on the likelihood ratio (LR) statistic for the Negative Binomial distribution are also evaluated. The performance is compared by a simulation study. Control limits are adjusted to meet the desired level of average run length (ARL_0), to make fair the comparison among the statistics, and to reach an optimum value of ARL_1 for different out-of-control shifts.

This paper is organized as follows. In Section 2, the usual control charts and new proposals to monitor count data are presented. Section 3 presents the results of the fitted GLM for the hospitalization data from 2006 to 2010 and the detection of epidemic periods in 2011. All the presented methods are empirically evaluated by simulations in Section 4. Final comments and conclusions are presented in Section 5.

2 Control charts to monitor count data series

In general, control charts are built to identify shifts in the expected values, and these shifts may be related to an increase or decrease of these expectations. In surveillance studies, only larger counts are of concern since they may be associated to epidemics. Therefore, in the present study, all control charts are built only with upper control limits to detect increases in the expectations. In order to build control charts to monitor count data series, let us consider some assumptions:

- Only a single observation X_t is considered at each time t and X_t assumes non-negative integer values in $\{0, 1, \ldots, \}$;

- The expected value may vary over time as a function of explanatory variables;
- X_t follows a Poisson or Negative Binomial distribution.

Since our main goal is to monitor a count time series and its mean value is supposed to change over time as observed in practice, even for an in-control process, from now on the expected value of X_t is denoted as $\mu_{0,t}$ and $\mu_{1,t}$ when the process is respectively in-control and out-of-control and as μ_t when it is not specified the state of the process.

The probability density functions of X_t following a Poisson distribution and a Negative Binomial distribution, written as members of the exponential family, are respectively

$$f(X_t|\mu_t) = exp\{X_t \ln \mu_t - \mu_t - \ln(\Gamma(X_t + 1))\} \tag{1}$$

with $E(X_t) = Var(X_t) = \mu_t$, and

$$f(X_t|\mu_t, \phi) = exp\left\{X_t \ln\left(\frac{\mu_t}{\mu_t + \phi}\right) + \phi \ln\left(\frac{\phi}{\mu_t + \phi}\right) + \ln\left(\frac{\Gamma(\phi + X_t)}{\Gamma(X_t + 1)\Gamma(\phi)}\right)\right\} \tag{2}$$

with $E(X_t) = \mu_t$ and $Var(X_t) = \mu_t + \mu_t^2/\phi$. The gamma function in (1) and (2) is an extension of the factorial function and for a positive integer n, $\Gamma(n) = (n-1)!$. The parametrization in (2) is known as the NB-2 in the literature and obtained as a Poisson model with a gamma random effect or as the distribution of the number of failures until the rth success, implying a non-constant coefficient of variation $(1 + \mu/\phi)$.[10]

To monitor non-normally distributed data, a first approach often used in statistical process control consists of applying some transformation on the original count variable in order to get approximately a normal distribution. Then, control charts are built for the transformed variable with control limits determined under a normal distribution.

Traditionally, the transformations for a sequence $X_1, X_2, \ldots$ of independent Poisson random variables (with mean μ_t) based on the asymptotic normality and on the square root transformation as presented in Rossi, Lampugnani and Marchi (RS)[8] are respectively

$$Z^*_{0,t} = \frac{X_t - n_t\mu_t}{\sqrt{n_t\mu_t}}, \quad t = 1, 2, \ldots,$$

and

$$Z^{**}_{0,t} = 2(\sqrt{X_t} - \sqrt{n_t\mu_t}), \quad t = 1, 2, \ldots,$$

where $n_1, n_2, \ldots$ are the corresponding sample sizes with $E(X_t) = n_t\mu_t$.

Based on these two transformations, as they are asymptotically normal standardized distributed, RS[8] proposed the statistic $Z_{1,t}$ which also asymptotically follows a normal standardized distribution to build a control chart

$$Z_{1,t} = 0.5Z^*_{0,t} + 0.5Z^{**}_{0,t} = \frac{X_t - 3n_t\mu_t + 2\sqrt{X_t n_t\mu_t}}{2\sqrt{n_t\mu_t}} \tag{3}$$

Similarly, transformations for a sequence $X_1, X_2, \ldots$ of independent Negative Binomial random variables to stabilize the variance can be found in the literature in Laubscher[11] or Johnson, Kemp and Kotz (JK)[12] as

$$Z_{2,t} = \sqrt{\phi - a}\left(\sinh^{-1}\sqrt{\frac{X_t + b}{\phi - 2b}} - \sinh^{-1}\sqrt{\frac{\mu_t + b}{\phi - 2b}}\right) \tag{4}$$

with $a = b = 0$ or $a = 0.5$; $b = 0.375$.

Recently, Guan (GN)[13] suggested the transformed variable

$$Z_{3,t} = \sqrt{\phi - 0.5}\left(\sqrt{\frac{X_t + 0.385}{\phi - 0.75}} - \sqrt{\frac{\mu_t + 0.385}{\phi - 0.75}}\right) \tag{5}$$

Finally, the transformation suggested by Jorgensen (JG)[14] for counts following a Negative Binomial consists of a standardization of X_t, expressed as

$$Z_{4,t} = \frac{X_t - \mu_t}{\sqrt{\phi\pi_t/(1-\pi_t)^2}} \tag{6}$$

with $\pi_t = \mu_t/(\mu_t + \phi)$. Although the transformations (4) to (6) are not new, they have not been used to build CUSUM control charts with their performance evaluated.

Other approach consists in building control charts with the standardized residuals obtained after fitting a GLM for count data. For non-normal distributed data, the DR is a candidate as it presents good properties as variance stability and also it is supposed to follow a standardized normal distribution.[15,16] In this paper, the DRs are proposed as an alternative statistic to build a control chart for count data. In case of X_t following a Negative Binomial distribution, with mean μ_t, the DR is defined[10] as

$$Z_{5,t} = sign(X_t - \mu_t)\sqrt{(d_t^2)} \tag{7}$$

where

$$d_t^2 = \begin{cases} 2\phi\ln(1 + \mu/\phi), & if\ X_t = 0 \\ 2X_t\ln\left(\frac{X_t}{\mu_t}\right) - 2\phi(1 + X_t/\phi)\ln\left(\frac{1+X_t/\phi}{1+\mu_t/\phi}\right) & if\ X_t > 0 \end{cases}$$

and ϕ is a constant dispersion parameter. According to McCulloch and Searle,[15] $Z_{5,t}$ follows approximately a standardized normal distribution.

For all the presented statistics, in general, the parameters are replaced by their corresponding estimates obtained in a training dataset. In this paper, they are used to build CUSUM control charts as follows

$$C_{i,t} = max(0, C_{i,t-1} + Z_{i,t} - k_t), i = 1, \ldots, 5 \tag{8}$$

Whenever $C_{i,t} > h$, it is decided that the process is out-of-control, meaning that the monitored parameter has shifted and a search for special causes starts. The values of the control parameters k_t and h are firstly determined under a normal distribution to get in-control average run length equal to 500 ($ARL_0 = 500$) as proposed in previous contributions about public health surveillance.[9]

In general, the value of k_t is constant. It is recommended to perform simulation studies in order to estimate the empirical ARL_0 and change the control limit h to achieve the desired ARL_0 as presented in the next section, even for the statistics that are supposed to have a Gaussian distribution.

RY[5] proposed a CUSUM control chart for Poisson data with parameters k_t and h_t that change over time to detect shifts in time series with seasonal effects. They argue that equivocal results could be obtained if the CUSUM control chart is implemented with fixed parameters as the in-control average rate is not constant between distinct periods. Expressions for the series of k_t for several distributions of the exponential family which include Poisson and Negative Binomial distributions can be found in Hawkins and Olwell.[17] The monitored statistic is based on the ratio of densities with out-of control and in-control parameters. For density probability functions belonging to the exponential family, this statistic corresponds to the minimal sufficient statistic, and the values of k_t depend on the parameters $\mu_{0,t}$, $\mu_{1,t}$, and ϕ. Hawkins and Olwell[17] obtained the expression of k_t for the Negative Binomial distribution using a parametrization different from the presented in (2).

Following RY,[5,18] the CUSUM chart is expressed as

$$C_{6,t} = max[(0, C_{6,t-1} + c_t(X_t - k_t))] \tag{9}$$

and is proposed to monitor the count series with the value of k_t determined under a Negative Binomial distribution as

$$k_t = \frac{-\phi \ln\{(\phi + \mu_{0,t})/(\phi + \mu_{1,t})\}}{\ln\{\mu_{1,t}(\phi + \mu_{0,t})/\mu_{0,t}(\phi + \mu_{1,t})\}} \tag{10}$$

where X_t is supposed to follow a Negative Binomial distribution (as in (2)) with in-control and out-of-control means, respectively $\mu_{0,t}$ and $\mu_{1,t}$.

This control chart signals whenever $C_{6,t} > h_t$ and $c_t = h/h_t$ is the ratio between h and h_t, where h_t is the threshold associated with the desired ARL_0 and the constant h depends on the average of k_t, for $t = 1, \ldots, n$. An easier algorithm adopted in this study is to consider $c_t = 1$. In this sense, a signal is given whenever $C_{6,t} > h$ and the constant h is searched to meet the desired ARL_0

To complete the list of monitored statistics, the approach proposed by Höhle and Paul[6] based on the LR for Negative Binomial distribution is presented. The idea of a sequence of hypothesis tests based on the LR statistics to detect a change point is proposed in Lorden[19] and recent procedures for generalized likelihood tests are in Xu et al.[20]

In order to detect a change point using a sequence of n observations, $x_1, \ldots, x_n$, Höhle and Paul[6] defines the instant of the shift in the expected value as

$$N = min\left\{ n \geq 1 : \max_{1 \leq \tau \leq n} \left[lnL(\tau) = \sum_{t=\tau}^{n} \ln\left\{ \frac{f(x_t|\mu_{1,t}, \phi, w_t)}{f(x_t|\mu_{0,t}, \phi, w_t)} \right\} \right] \geq h \right\} \tag{11}$$

where w_t are covariates, ϕ is the dispersion parameter, and $f(x_t|\mu_t, \phi, w_t)$ is the probability density function of the Negative Binomial distribution defined in (2) including the covariates w_t to model μ_t.

The LR in (11) corresponds to the statistic test for a null hypothesis that all observations come from the same in-control distribution against the alternative where the observations $\tau, \ldots, n$ are from the out-of-control process with mean $\mu_{1,t}$.

Based on a GLM with a Negative Binomial distribution and the values of $\mu_{0,t}$ and $\mu_{1,t}$ when the process is in-control and out-of-control, respectively, the CUSUM control chart can be recursively written as

$$C_{7,0} = 0, \quad C_{7,t} = max\left(0, C_{7,t-1} + ln\left\{\frac{f_{\theta_0}(x_t)}{f_{\theta_1}(x_t)} - k\right\}\right), \quad t \geq 1 \tag{12}$$

A signal is given when $C_{7,t} > h$ and the value of h must be searched using simulations to achieve ARL_0 as the previous methods. In Höhle and Paul,[6] the value of k is equal to 0, since it is based on the LR statistic.

Höhle and Paul[6] also present a generalization of LR to build a CUSUM chart. It consists on replacing the parameters when the process is out-of-control by its estimates at each iteration. Due to its complexity, they consider estimating the change of the process mean from $\mu_{0,t}$ to $\mu_{1,t}$ for a Poisson process and incorporate this estimation in the CUSUM proposed to a Negative Binomial distribution. Values of ARL_0 can be obtained from Monte Carlo Markov Chain algorithm.[21] Estimate of the run length for the stationary distribution is implemented in the Surveillance Library of software R.[22] As pointed in Höhle and Paul,[6] the results of their generalization are similar to the usual LR. In this sense, only the usual LR statistic for fixed parameters in (11) is included in the current study for comparative purpose in this study.

It is worth noting that the methods proposed by RY[5] and Höhle and Paul[6] are based on the LR and both depend on fixing a value of $\mu_{1,t} = \Delta\mu_{0,t}$. The monitored statistics depends on the chosen value of Δ.

To find the optimal values of h and k, i.e. find the smaller ARL_1 fixing the desired value of $ARL_0 = 500$, and to evaluate the performance of CUSUM charts, the values of ARLs are empirically calculated based on simulations. More details of the algorithm to find these values are detailed in Fricker.[23] All data analysis and simulations are implemented in the R software.

3 GLM for count time series

In this section, GLMs are fitted to analyze the count time series of the daily hospital admissions due to respiratory diseases for people aged over 65 years in the city of São Paulo-Brazil. Daily data from January 2006 to December 2010 are used to fit the model, which include explanatory variables, and predicted values for 2011 are calculated.[2,23] Only a Negative Binomial distribution is considered since the dispersion parameter is significantly larger than one. Additionally, the population size is included as an offset, and the logarithm function is the chosen link function, so the hospitalization rate per 100,000 inhabitants over time is modeled. The daily admission time series is obtained from Hospital Information System at Health Secretary of São Paulo (PRO-AIM) and depicted in Figure 1.

As shown in Figure 1, the daily hospitalization series presents a seasonal behavior, and to control part of this seasonality, sine and cosine functions are included in the model.[24] Additionally, days of week are included as categorical variables to explain the variability of daily hospitalizations. Let X_t denote the daily number of hospitalizations at day t with $X_t \sim NegBin(\mu_{0,t}, \phi)$ and the expectation $\mu_{0,t}$ written as

$$ln\left(\frac{\mu_{0,t}}{pop_t}100000\right) = \beta_0 + \beta_1 cos(2\pi t/365) + \beta_2 sin(2\pi t/365) + \beta_3 Sat_t + \beta_4 Sun_t + \beta_5 Mon_t \tag{13}$$

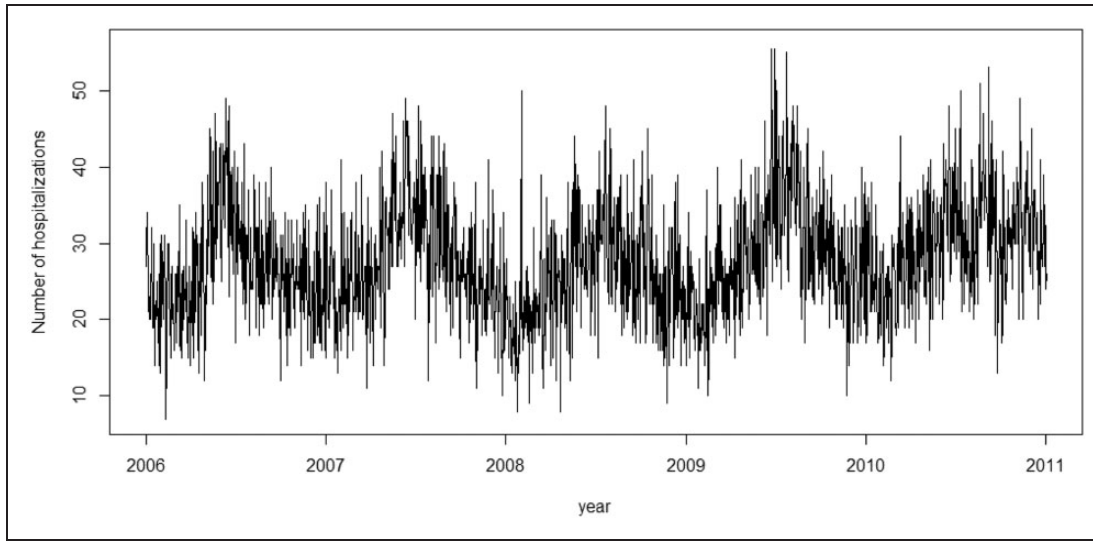

Figure 1. Daily admissions due to respiratory diseases for people aged over 65 years in São Paulo from 2006 to 2011.

Table 1. Estimates, standard errors, and p values of the parameters of model (13).

Coefficient	Estimate	SE	p value
Intercept (β_0)	1.227	0.007	<0.001
Cosine (β_1)	-0.173	0.007	<0.001
Sine (β_2)	-0.045	0.007	<0.001
Saturday (β_3)	-0.158	0.016	<0.001
Sunday (β_4)	-0.207	0.016	<0.001
Monday (β_5)	0.040	0.015	0.008
Dispersion (ϕ)	69.99	8.22	

SE: standard error.

where pop_t is the population in risk on the tth day; Sat_t, Sun_t, and Mon_t are dummy variables equal to 1 respectively for Saturday, Sunday, and Monday, and zero otherwise. The inclusion of the offset term $g_t = (pop_t/100000)$ allows to model the average daily rate of admissions. The estimates of the coefficients of model (13) are presented in Table 1. Note that the estimated dispersion parameter larger than one is an indicative of overdispersion.

The fitness of the proposed model is evaluated by a residual analysis. The residual deviance is 1847.7 for 1825 degrees of freedom, which is an indicative of good fit. The usual residuals plots indicated that there is no departure of the assumptions of the model, excepted the independence of the residuals, since the autocorrelations of the DRs are around 0.2 for the first and second lags. Figure 2 presents the quantile-quantile plot of the DRs, and it seems they are normally distributed. Also, the normality assumption is accepted using the Shapiro Wilk test ($p = 0.1972$), what implies that the assumption of a Negative Binomial distribution for the counts is appropriate. Additionally, there is no evidence of outliers among the observations.

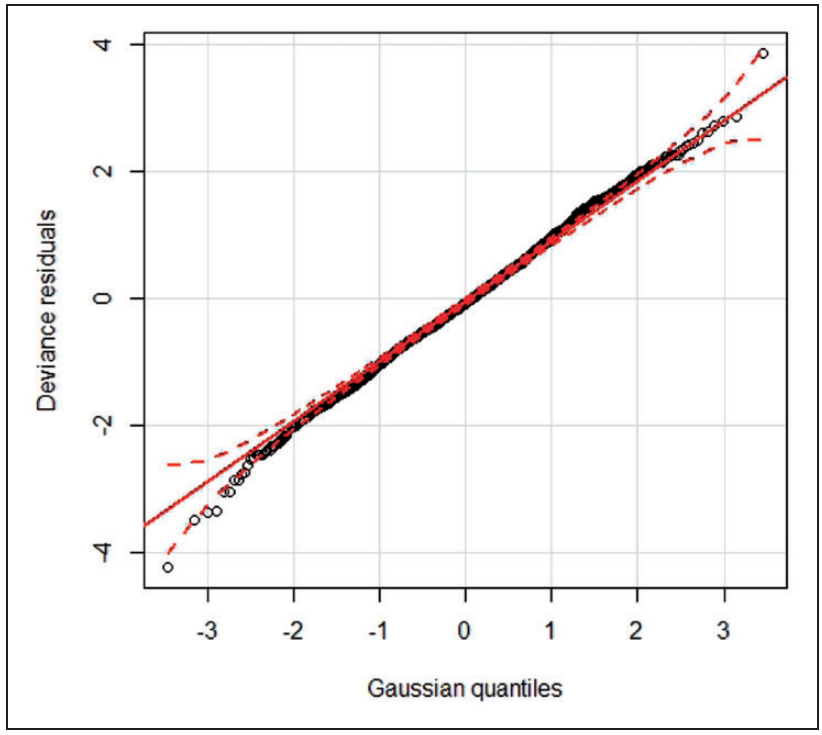

Figure 2. Quantile-quantile plot of deviance residuals.

The transformation proposed by RS[8] seems to follow a Gaussian distribution ($p = 0.0550$), and the normality is rejected for the transformation proposed by JG[14] ($p < 0.0001$).

Based on this fitted model for the daily number of hospitalizations from 2006 to 2010 and assuming that this period is non-epidemic, the time-varying in-control average $\mu_{0,t}$ is estimated from (13) using the estimates in Table 1. Figure 3 shows the fitted series from 2006 to 2010 (red line) and the in-control average μ_0 for the daily number of hospitalizations from 2011 (blue line).

The CUSUM charts were built to detect increases of 50% ($\delta = 1.5$) in 2011, using the values of h and k presented in the next section, and are depicted in Figure 4. All control charts indicate epidemics in January and in the beginning of February. Only Jorgensen's method identified 11 days as epidemics, including the days 85 and 91 (respectively 26 March and 1 April). In January 2011, several days (days = 3, 7, 10, 13, 17, 25, 26, 28, and 31) presented a hospitalization rate more than 50 % higher than the average of the rates in previous years. This is consistent with the results of the control charts in Figure 4.

4 Performance of control charts

The predicted in-control averages $\mu_{0,t}$ for the daily number of hospitalizations are obtained based on the estimates of Table 1. Then, future hospitalization data are simulated 10,000 times according to a Negative Binomial distribution with mean $\delta\mu_{0,t}$ and variance $\delta\mu_{0,t} + (\delta\mu_{0,t})^2/\phi$, where ϕ is the dispersion parameter.

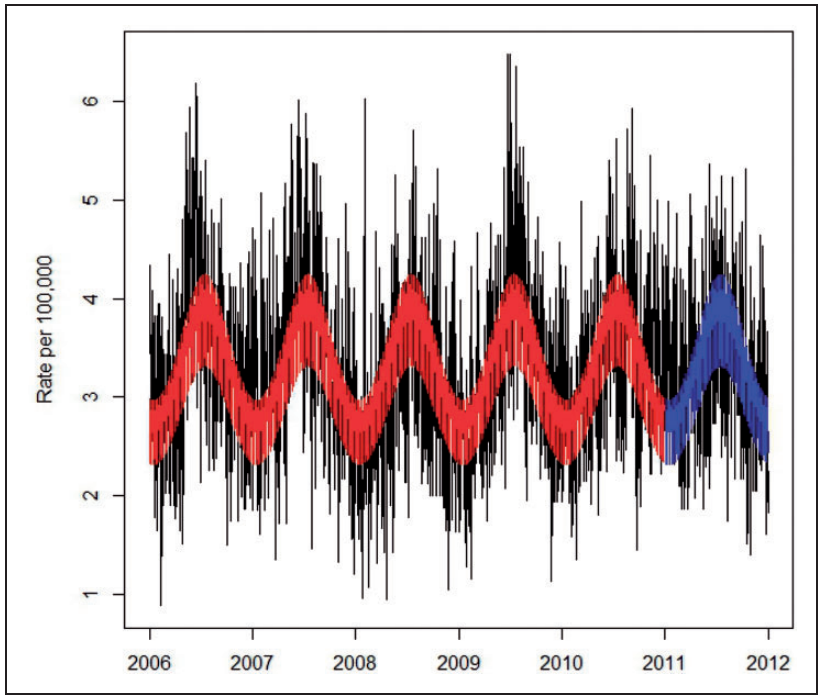

Figure 3. Daily admissions in São Paulo from 2006 to 2011 and adjusted values.

With in-control simulated samples, control limits are computed and the overall in-control ARL_0 is calculated. Then, out-of-control samples are simulated assuming $\mu_{1,t} = \delta\mu_{0,t}$ with $\delta = \{1; 1.25; 1.5; 1.75; 2\}$. The performance of the proposed control charts is evaluated by comparing the values of out-of-control ARL_1.

Table 2 presents the values of ARL_1 for the CUSUM chart built with the statistics $C_{1,t}$(JK) (using the transformation presented in JK[12]). Different thresholds h, depending on k, are obtained to meet the desired value of $ARL_0 \approx 500$ (the median of length run (MLR_0) is also included). Then, values of ARL_1 are determined for different sizes of shifts δ. In general, ARL_1 decreases as k increases until a tipping point. After this point, ARL_1 starts to increase. This behavior is observed for other statistics than $C_{1,t}$ so tables like Table 2 for other statistics are not shown here. For the CUSUM chart with the statistics $C_{1,t}$, in a case of smaller shifts (as $\delta = 1.25$), the better result is observed for lower values of k ($k = 0.4$) while for larger shifts, better performance is associated with higher values of k values ($k = 1.0$).

For comparative purposes, the "best" pair (k, h) of the CUSUM chart (for each shift size), which provides the lowest value of ARL_1 for each statistic are included in Table 3, maintaining the $ARL_0 \approx 500$. These parameters do not reach the exact optimum argument but certainly they are very close to get the optimum ARL_1. The LR and RY CUSUMs were calculated to detect increases of 100% ($\Delta = 2$) in the mean rate, even simulating data for different values of the true increase, $\delta = 1.25, 1.5, 1.75$ and 2.0.

Analyzing these results, all statistics present tiny differences in ARL_1 for large shifts, but differences of one or two days are observed until detecting an increase of 25% in the mean

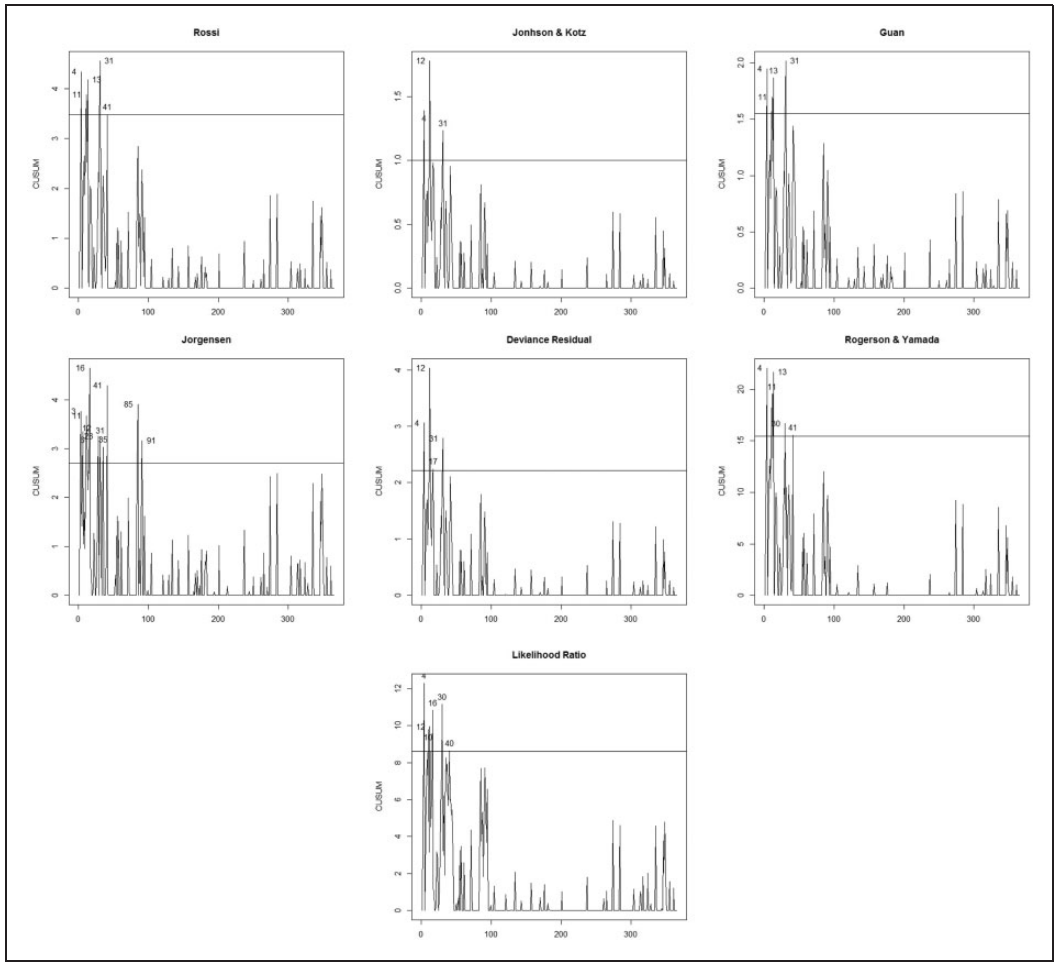

Figure 4. CUSUM charts for hospitalizations in 2011 for $ARL_0 = 500$ and $\delta = 1.5$.

hospitalization rate. Lowest ARL_1s are achieved using the LR and the RY statistics. On average, it may take a week to detect an increase of 25% and three and two days to detect a larger increase.

In CUSUM charts, the pair (k,h) plays an important role. So many contributors[18,25–28] have dedicated to their determination as function of desired values of ARL_1 and ARL_0. RY[18] suggest to calculate k and h depending on the desired ARL values as

$$k \approx \sqrt{\frac{\ln(-\ln(\nu) - \ln(\nu))}{2ARL_1} - \frac{1}{2ARL_0}} \tag{14}$$

$$h \approx \left(\frac{2k^2 ARL_0 + 2}{2k^2 ARL_0 + 1}\right) \frac{\ln(1 + 2k^2 ARL_0)}{2k} - 1.166 \tag{15}$$

with $\nu = \frac{ARL_1}{ARL_0} \exp\left(1 - \frac{ARL_1}{ARL_0}\right)$

Table 2. Average run length and standard errors (SE) for CUSUM chart – JK.

		$\delta = 1.0$			$\delta = 1.25$		$\delta = 1.5$		$\delta = 1.75$		$\delta = 2.0$	
k	h	ARL_0	SE	MLR_0	ARL_1	SE	ARL_1	SE	ARL_1	SE	ARL_1	SE
0.0	13.47	500.71	4.13	373.00	20.40	0.05	10.65	0.02	7.32	0.01	5.74	0.01
0.1	7.14	500.77	4.36	359.00	13.01	0.04	6.39	0.02	4.43	0.01	3.54	0.01
0.2	4.94	499.79	4.35	355.50	10.94	0.05	5.06	0.01	3.50	0.01	2.80	0.01
0.3	3.80	500.21	4.35	362.00	10.87	0.05	4.60	0.01	3.13	0.01	2.48	0.01
0.4	3.07	500.25	4.28	371.50	**10.66**	0.06	4.11	0.01	2.74	0.01	2.15	0.01
0.5	2.57	499.85	4.29	365.00	11.56	0.08	3.94	0.02	2.55	0.01	1.95	0.01
0.6	2.20	500.00	4.28	366.50	12.92	0.10	**3.87**	0.02	2.44	0.01	1.84	0.01
0.7	1.92	499.73	4.26	376.00	14.82	0.12	3.90	0.02	2.35	0.01	1.74	0.01
0.8	1.69	500.81	4.31	359.00	17.36	0.15	4.00	0.02	2.29	0.01	1.65	0.01
0.9	1.51	500.85	4.33	364.00	20.17	0.17	4.16	0.02	2.31	0.01	1.65	0.01
1.0	1.35	500.17	4.37	352.00	23.11	0.20	4.38	0.03	**2.29**	0.01	**1.61**	0.01
1.1	1.21	499.36	4.35	357.00	26.33	0.23	4.67	0.03	2.33	0.01	1.62	0.01

ARL: average run length; MLR: median of length run; JK: Johnson, Kemp and Kotz.
Values in bold mean are the lowest of ARL1.

Table 3. The "best" design parameters (k and h) which provide the best performance in terms of ARL_1.

		$\delta = 1.25$			$\delta = 1.5$			$\delta = 1.75$			$\delta = 2.0$		
		k	h	ARL_1	k	h	ARL_1	k	h	ARL_1	k	h	ARL_1
C_{0t}	RS	0.6	5.28	9.30	1.1	3.20	3.47	1.5	2.37	2.02	1.6	2.22	1.46
C_{1t}	JK	0.4	3.07	10.66	0.6	2.20	3.87	1.0	1.35	2.29	1.0	1.35	1.61
C_{2t}	GN	0.3	2.42	9.43	0.5	1.56	3.50	0.8	0.96	2.05	1.0	0.70	1.43
C_{3t}	JG	0.5	4.74	8.76	0.9	2.97	3.25	1.2	2.30	1.89	1.2	2.30	1.39
C_{4t}	DR	0.5	4.10	8.84	0.9	2.46	3.27	1.2	1.84	1.90	1.2	1.84	1.39
C_{5t}	RY	−8.5	26.4	**7.18**	−5.1	15.64	**3.07**	−3.7	13.3	**1.88**	−3.7	13.3	**1.39**
C_{6t}	LR	−3.8	11.90	**7.35**	−3.1	8.92	**3.02**	−1.5	5.56	**1.80**	−1.5	5.56	**1.34**
(k,h)		0.36	5.59	7.00	0.59	3.80	3.00	0.79	2.93	1.80	0.94	2.45	1.30
in (14) and (15)		0.34	5.92	8.00	0.54	4.09	3.50	0.76	3.02	1.90	0.90	2.56	1.40
		0.31	6.22	9.00	0.50	4.36	4.00	0.74	3.10	2.00	0.87	2.66	1.50

ARL: average run length; JK: Johnson, Kemp and Kotz; RS: Rossi, Lampugnani and Marchi; GN: Guan; JG: Jorgensen; DR: deviance residual; RY: Rogerson and Yamada; LR: likelihood ratio.
Values in bold mean are the lowest of ARL1.

Fixing $ARL_0 = 500$, $\delta = 1.25$, and $ARL_1 = 7; 8; 9$, the values of k and h are simply calculated using (14) and (15). These values of ARL_1 are chosen to be close to the optimum ARLs for the considered statistics. For other shift sizes, new values of ARL_1 are chosen, and the values of k and h are also presented in the last rows of Table 3.

Note that the pairs of (k,h) differ considerably from the optimum values shown in the top block of Table 3 for most of CUSUM charts (excepting similar values of k for JK and GN and h for RS). It is worth to mention that these three exception cases are charts built based on the transformations on the original count data to get approximately normal distribution. Values of k (with desired ARL_0

and ARL_1) given by (14) may be used to build the CUSUM charts for transformed variables; however, it was found by simulation that the threshold h in (15) do not produce a real $ARL_0 = 500$.

Considering the values of h and k for $\delta = 1.25$, the normality hypothesis is tested and the p values of the Shapiro Wilk tests are calculated for each simulated sample for 2011. This procedure is repeated for each method based on transformed variables. Only the DR seems to be Gaussian, presenting a 5% quantile of the p values equals to 0.0516. In general, for all other statistics, the assumption of normality is rejected. For example, Rossi's transformation presented a median p value of 0.009, indicating that more than one half of the simulated samples (under the Negative Binomial distribution) after the transformation are far from normality. Similar results are observed for other values of δ.

5 Conclusion and perspectives

In this paper, different statistics (as transformations of the count data or residuals or LR) are used to build CUSUM control charts to monitor count data series. Their performances are compared in a simulation study, where it is assumed that the count data follows a Negative Binomial distribution with mean value varying over time. The values of the time-varying mean used in the simulations are based on the estimated mean of a real hospitalization rate.

Among the evaluated statistics, the CUSUM chart based on the LR (proposed by Höhle[22]) presents the best performance based on the comparison of out-of-control ARL_1 values. The method proposed by RY[5] for the Negative Binomial distribution presents also good performance, detecting a 25% increase in hospitalizations in one week, like the likelihood method. It is worth to note that the extended method of RY,[5] considering time-varying control limits h_t, may be more difficult to be implemented since it requires too many calculations to obtain h_t to achieve the target in-control $ARL_0 = 500$. However, maintaining constant values of h and k, this method is simple and also yields low ARL_1s.

Despite the best performance of the LR method proposed by Höhle,[22] some public health authorities may find easier to understand control charts based on transformed variables, since they consist in cumulative sums of standardized variables, than calculating likelihoods. However, the idea behind the LR is only the ratio between the probabilities of observing time series assuming that the process is out-of-control and in-control, what also has an appealing interpretation.

The choice of control limits for charts based on the transformed variables usually assumes the normality of these variables even when the normality is rejected. This choice may lead to equivocal decision as they provide values of in-control ARL_0s lower than the desired levels of $ARL_0 = 500$.[18,25]

Similarly, the calculation of k and h proposed by RY[18] may not produce optimum values of ARL_1. For this reason, the values of h and k must be exhaustively searched by simulation to get optimized results as discussed for example in Fricker[23] and RY.[18] Based on simulated results, the initial values of k for the CUSUM charts based on the transformed variables may be calculated as proposed in RY.[18]

In general, other papers do not assume that k may be a negative value to reach the optimum ARL, but better results may be achieved using this search mainly for CUSUM charts based on the LR and the method proposed by RY.[18]

Also, it is important to remind that better performances may be achieved for the LR CUSUM proposed by Höhle and Paul[6] based in the generalized LR, but this implies more computational efforts to be obtained for the Negative Binomial model, but it is implemented for a Poisson distribution in the library Surveillance available in the R software.[22]

Although the transformed variables are not normally distributed, except for the DR, the simulation study indicates that the performance of the CUSUMs is similar for all considered statistics.

The monitoring of daily number of hospital admissions due to respiratory diseases for people over 65 years old in São Paulo city illustrates the current proposal. For these data, a GLM is fitted for data from January 2006 to December 2010, taking into account the population in risk, the seasonality and week days. Then, with the estimated parameters, predicted values for 2011 are calculated and all charts are drawn. All statistics signal higher expected hospitalization rates in the beginning of 2011. These days really present higher hospitalization rates (an increase higher than 50%), indicating that all methods are able to detect deviations as concluded based on the simulation study.

As in production processes, one possible extension of the control charts discussed here is the inclusion of other limits, as a warning limit, beyond the current control limit. A signal of out-of-control is given whenever $C_{i,t} > h$ or a sequence of last j values of $C_{i,t}$ lies in the interval $]w; h]$, that is, $(w < C_{i,t-j+1} \leq h) \bigcup \ldots \bigcup (w < C_{i,t-1} \leq h) \bigcup (w < C_{i,t} \leq h)$. The limits ($w$ and h) and the length of the sequence j may be obtained by simulation. Usually in production processes, the inclusion of the supplementary rules improves the performance of control charts, and the same may be implemented in surveillance monitoring.

Funding

This work was supported by FAPESP (Grant number 13/00506-1); CNPq (Grant number 301618/2010-0) and CAPES.

References

1. Woodall W. The use of control charts in health-care and public health surveillance. *J Qual Technol* 2006; **38**: 88–103.
2. Unkel S, Farrington C, Garthwaite P, et al. Statistical methods for the prospective detection of infectious disease outbreaks: a review. *J R Stat Soc Ser A Stat Soc* 2012; **175**: 40–82.
3. Lucas J. Counted data CUSUM's. *Technometrics* 1985; **27**: 129–144.
4. Hawkins D. A CUSUM for a scale parameter. *J Qual Technol* 1981; **13**: 228–233.
5. Rogerson PA and Yamada I. Approaches to syndromic surveillance when data consist of small regional counts. *Morb Mortal Wkly Rep* 2004; **53**: 79–85.
6. Höhle M and Paul M. Count data regression chart for the monitoring of surveillance time series. *Comput Stat Data Anal* 2008; **52**: 4357–4368.
7. Skinner KR, Montgomery DC and Runger GC. Process monitoring for multiple count data using generalized linear model-based control charts. *Int J Prod Res* 2003; **41**: 1167–1180.
8. Rossi G, Lampugnani L and Marchi M. An approximate CUSUM procedure for surveillance of health events. *Stat Med* 1999; **18**: 2111–2122.
9. Rossi G, Del Sarto S and Marchi M. A new risk-adjusted Bernoulli cumulative sum chart for monitoring binary health data. *Stat Methods Med Res*. Epub ahead of print. 22 April 2014. DOI: 10.1177/0962280214530883.
10. Hardin J and Hilbe J. *Generalized linear models and extensions*, 2nd ed. College Stations, TX: Stata Press, 2007.
11. Laubscher N. On stabilizing the binomial and negative binomial variances. *J Am Stat Assoc* 1961; **56**: 143–150.
12. Johnson N, Kemp A and Kotz S. *Univariate discrete distribution*, 3rd ed. Hoboken, NJ: John Wiley & Sons, 2005.
13. Guan Y. Variance stabilizing transformations of Poisson, binomial and negative binomial distributions. *Stat Probab Lett* 2009; **79**: 1621–1629.
14. Jorgensen B. *The theory of dispersion models*. London: Chapman and Hall, 1996.
15. McCulloch CE and Searle SR. *Linear and generalized linear mixed models*, 3rd ed. New York: Wiley, 2001.
16. McCullagh P and Nelder JA. *Generalized linear models*, 2nd ed. London: Chapman & Hall/CRC, 1989.
17. Hawkins D and Olwell D. *Cumulative sum charts and charting for quality improvement*. New York: Springer, 1998.
18. Rogerson P and Yamada I. *Statistical detection and surveillance of geographic clusters*. Boca Raton, FL: Chapman & Hall/CRC, 2009.
19. Lorden B. Likelihood ratio tests for sequential k-decision problems. *Ann Math Stat* 1972; **43**: 1412–1427.
20. Xu L, Wang S and Reynolds M. A generalized likelihood ratio control chart for monitoring the process mean subject to linear drifts. *Qual Reliab Eng Int* 2013; **29**: 1099–1638.
21. Brook D and Evans D. An approach to the probability distribution of CUSUM run length. *Biometrika* 1972; **59**: 539–548.
22. Höhle M. Surveillance: an R package for the surveillance of infectious diseases. *Comp Stat* 2007; **22**: 571–582.
23. Fricker R. *Introduction to statistical methods for biosurveillance – with emphasis on syndromic surveillance*. New York: Cambridge University Press, 2013.

24. Serfling RE. Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public Health Rep* 1963; **78**: 494.
25. Siegmund D. *Sequential analysis test and confidence intervals*, 3rd ed. New York: Springer-Verlag, 1985.
26. Woodall WH and Adams BM. The statistical design of CUSUM charts. *Qual Eng* 1993; **4**: 559–570.
27. Alwan L. Designing an effective exponential CUSUM chart without the use of nomographs. *Commun Stat Theory Methods* 2000; **29**: 2879–2893.
28. Rogerson P. Formulas for the design of CUSUM quality control charts. *Commun Stat Theory Methods* 2006; **35**: 373–383.