Revista Colombiana de Estadística
July 2017, Volume 40, Issue 2, pp. 243 to 262
DOI: http://dx.doi.org/10.15446/rce.v40n2.59925

Monitoring Aggregated Poisson Data for Processes with Time-Varying Sample Sizes

Monitoreo de datos Poisson agregados para procesos con tamaños de muestra que varían en el tiempo

Víctor Hugo Morales[1,a], José Alberto Vargas[2,b]

[1]Departamento de Matemáticas y Estadística, Facultad de Ciencias, Universidad de Córdoba, Montería, Colombia

[2]Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia

Abstract

This article deals with the effect of data aggregation, when Poisson processes with varying sample sizes, are monitored. These aggregation procedures are necessary or convenient in many applications, and can simplify monitoring processes. In health surveillance applications it is a common practice to aggregate the observations during a certain time period and monitor the processes at the end of it. Also, in this type of applications it is very frequent that the sample size vary over time, which makes that instead of monitor the mean of the processes, as would be in the case of Poisson observations with constant sample size, the occurrence rate of an adverse event is monitored.

Two control charts for monitoring the count Poisson data with time-varying sample sizes are proposed by Shen, Zou, Tsung & Jiang (2013) and Dong, Hedayat & Sinha (2008). We use the average run length (ARL) to study the performance of these control charts when different levels of aggregation, two scenarios of generating of sample size and different out-of-control states are considered. Simulation studies show the effect of data aggregation in some situations, as well as those in which their use may be appropriate without significantly compromising the prompt detection of out-of-control signals. We also show the effect of data aggregation with an example of application in health surveillance.

Key words: Data aggregation, EWMAG and EWMAe charts, Health surveillance, Levels of aggregation, Time-varying sample sizes.

[a]PhD(c). E-mail: vmorales@correo.unicordoba.edu.co
[b]PhD. E-mail: javargasn@unal.edu.co

Resumen

Este artículo trata sobre el efecto de la agregación de datos cuando se monitorean procesos Poisson con tamaño de muestra variable. Estos procedimientos de agregación resultan necesarios o convenientes en muchas aplicaciones y pueden simplificar los procesos de monitoreo. En aplicaciones de vigilancia de la salud, es una práctica común agregar las observaciones durante un cierto período y monitorear el proceso al final de éste. También, en este tipo de aplicaciones es muy frecuente que el tamaño de muestra varíe sobre el tiempo, lo cual hace que en lugar de monitorear la media del proceso, como sería en el caso de observaciones Poisson con tamaño de muestra constante, se monitorio la tasa de ocurrencias de un evento adverso.

Dos cartas de control para monitorear el conteo de datos Poisson con tamaños de muestra que varían en el tiempo han sido propuestas por Shen et al. (2013) and Dong et al. (2008). Usamos la longitud de corrida promedio (ARL) para estudiar el desempeño de estas cartas de control cuando se consideran diferentes niveles de agregación, dos escenarios de generación de tamaños de muestra, y diferentes estados fuera de control. Estudios de simulación muestran el efecto de la agregación de datos en algunas situaciones, así como otras en las que su uso puede ser apropiado sin comprometer significativamente la pronta detección de situaciones fuera de control. También mostramos el efecto de la agregación mediante un ejemplo de aplicación en vigilancia de la salud.

***Palabras clave*:** agregación de datos, cartas EWMAG y EWMAe, vigilancia de la salud, niveles de agregación, tamaños de muestras variables.

1. Introduction

There are situations for which aggregating count events is a recurrent practice. Dubrawski & Zhang (2010) indicate that in areas such as public health surveillance, the available data for analysis is presented in diverse and increasing volumes, which can compromise the reliability of models used in the analysis and the importance of statistical conclusions. The authors consider that the suitable use of data aggregation can be a solution to this problem.
As Schuh, Woodall & Camelio (2013) pointed out, this practice has become common and necessary in many cases. This is especially true in areas related to health surveillance. In such situations, the data are frequently aggregated, reported and monitored after a certain period of time, as for example, the number of surgical errors in one month or the number of cases of dengue (a tropical disease) reported monthly in a city. In the last case, it is common to find that during some weeks of the year, because of climatic factors, dengue events are equal to zero. An alternative to handle the large number of zeros obtained during an equal number of weeks, is often, aggregating the number of events obtained over periods of two or more weeks.

As Shen et al. (2013) pointed out, when the sample size is a constant, detecting a change in the rate could be achieved simply by detecting a change in the Poisson mean. However, in many cases, especially those related to health surveillance, the

sample size is not constant. In this cases, as in other related to health surveillance, the purpose is to detect the increase in the rate of occurrence of an adverse event. For that purpose, and according to Dong et al. (2008), in these cases one must know not only the number of events recorded, but also the corresponding sample size.

For many processes of interest in health surveillance, the counts of events recorded in regular time intervals are strongly related to the product exposure, i.e., in these processes, the counts of events recorded in regular time intervals are related to the population at risk, which frequently changes over time, so that these are not identically distributed. This feature makes many monitoring methods that assume constant sample sizes, such as Frisén & De Maré (1991) and Gan (1990), unadequate in this case. According to Dong et al. (2008), surveillance methods for these processes have not yet been studied thoroughly.

Jiang, Shu & Tsui (2011), pointed out that a simple monitoring scheme for Poisson observations when sample size varies over time, is the u chart. As is known, this chart is part of Shewhart schemes, which only uses the most recent information to determine the state of the process, and are not sensitive to small changes. To overcome these shortcomings of the u chart, CUSUM and EWMA schemes have been developed for the same purpose. Several proposals have been made within the CUSUM schemes, related to the monitoring of Poisson counts with variable sample sizes. Rossi, Lampugnani & Marchi (1999) propose a CUSUM scheme based on the normal approximation of a Poisson process in order to overcome the drawback of the standard CUSUM model, when the size, population structure at risk and reference rate may not be constant during the surveillance. Jiang et al. (2011) propose a weighted CUSUM chart (WCUSUM) with general weight functions applied to the likelihood-ratio statistic, in order to detect changes efficiently in the incidence rate when sample size varies. Shu, Jiang & Tsui (2011) compare the WCUSUM chart with the conventional CUSUM procedure in the presence of monotonous changes in population size. The simulation results show that the WCUSUM method may be more efficient than the conventional CUSUM methods in detecting increases in the incidence rate, especially for small shifts.

Within the group of EWMA schemes, Dong et al. (2008) studied the monitoring of Poisson data with sample sizes varying over time using pre-specified control limits, and explored three different ways to find the control limits of their EWMA schemes. Ryan & Woodall (2010) studied the ability in detecting increases in the Poisson rate of various cumulative sum (CUSUM) and exponentially weighted moving average (EWMA) control charts, recommended to monitor a process with Poisson count data when the sample size varies. These authors extend the work of Dong et al. (2008) and propose a modification of this (EWMA-M), adding a lower reflecting barrier. Zhou, Zou, Wang & Jiang (2012) proposed a new EWMA method based on weighted likelihood estimation and testing. However, as Shen et al. (2013) pointed out, several works have been built on the assumption that the sample size is assumed to follow a pre-specified random or deterministic model, which is known a priori when establishing appropriate control limits before the control charts initiate. In practice, our knowledge about the time-varying sample sizes is rarely available. To overcome this drawback, Shen et al. (2013) propose

the use of probability control limits in an EWMA control chart for monitoring Poisson count data with time-varying sample sizes in Phase II. This chart uses dynamic control limits that are calculated at each monitoring point. The dynamic probability control limits (DPCLs) have been used in various applications because of their advantages over standard control limits. Huan, Shu, Woodall & Tsui (2016) indicate that DPCLs are more general and able to accommodate more complex situations than the constant control limits. Shang & Woodall (2015) use DPCLs in risk-adjusted Bernoulli cumulative sum (CUSUM) chart, for monitoring the surgical performance in specific patient sequences.

If we aggregate data when the sample size varies over time, and the count of events or non-conformities follows an (conditional) independent Poisson distribution, given the corresponding sample size, the result is a process with time-varying sample sizes and with (conditional) Poisson distribution. Given the importance of determining the effect of aggregation of data in process monitoring, and considering the large number of applications that require or use time-varying sample sizes, we study the effect of aggregating data in situations that involve Poisson count data and whose sample sizes are not constant over time. If the sample sizes are constant over time, these surveillance schemes also work properly. For this purpose we use the schemes proposed by Shen et al. (2013) and Dong et al. (2008).

EWMAG and EWMAe charts, are two EWMA type control charts, that can be used when there are processes in which the sample size is not constant over time. However, these are constructed in different ways. The EWMAG chart does not have a certain structure to find the control limits for observations that are obtained over time. This chart uses dynamic control limits, which are determined online and depend only on the current and past sample size observations. Simulation processes are necessary to calculate these control limits. The EWMAe chart uses a predetermined structure to find the control limits. Also, as pointed out Shen et al. (2013), this chart was built on the assumption that the sample size follows a prespecified random or deterministic model, which is known a priori when establishing appropriate control limits before the control chart initiates.

The remainder of this article is organized as follows. In section 2, data aggregation is introduced, and it is done a review of some investigations on this topic, emphasizing those dealing with Poisson data. In section 3, a description of the EWMAG chart, proposed by Shen et al. (2013), is done. Also, a adaptation of this chart to case of aggregated data, is done. In section 4, the EWMAe chart, proposed by Dong et al. (2008), is described and adapted to situations in which aggregated data is considered. In section 5, the implemented simulation processes are described and their results are discussed. In section 6, the effect of data aggregation using an application with real data is shown.

2. Aggregated Poisson Data

In many applications, it is frequent to use methodologies and practices that facilitate the analysis of certain information, allowing to correct or at least atten-

uate the difficulties that can generate some characteristics in the data, such as the zero-excess presence. One of these procedures, is the data aggregation.

The effect of this practice was studied by Reynolds & Stoumbos (2000). They considered two CUSUM chart types for monitoring changes in the proportion of defective items, p. The first chart was based on binomial variables that result from counting the total number of effective items in a sample of size n. The second chart was based on Bernoulli observations which correspond to the individual items checked in the samples. As result of this procedures, it was concluded that there is little difference between the binomial CUSUM chart and the Bernoulli CUSUM chart, in terms of the expected time required to detect small and moderate shifts in p, but the Bernoulli CUSUM chart is better for detecting large shifts in this parameter.

Reynolds & Stoumbos (2004*b*) investigated if it is convenient or not to group observations. In this approach, they investigated whether it is better to use $n = 1$ or $n > 1$ as sample size. As a result they found that, combinations of CUSUM charts for the mean and the variance, in general, produce best statistical performance in detecting small and big shifts, sustained or transitories in μ or in σ, when the sample consists of one observation. Besides, if the Shewhart chart is used, $n = 1$ is better when it is required to detect small sustained shifts. Reynolds & Stoumbos (2004*a*) investigate whether using $n = 1$ is better than $n > 1$ from the perspective of statistical performance in monitoring the mean and the variance process. In order to get this, the performance of Shewhart, EWMA and CUSUM charts are compared. The result shows that it is not reasonable to use the Shewhart control chart when there are individual observations, and the EWMA and CUSUM charts have a better statistical performance for a wide range of sample sizes and out-of-control situations like drift processes. In the same way, significant differences in the statistical performance of the EWMA and CUSUM charts, were not found. With these charts, using $n = 1$ produces a better statistical performance than $n > 1$.

One of the biggest areas of application of data aggregation is related to all those situations in which a Poisson type variable is generated, such as in the public health surveillance. In this area, the use of data aggregation is a common practice. For example, Burkom, Elbert, Feldman & Lin (2004) explore the data aggregating by space, time and categories of data, and discuss the impact of this technique on the efficiency of alert algorithms relevant to public health surveillance, among others. The authors concluded that a judicious strategy of data aggregation has an important function within the improvement of biomonitoring systems.

Gan (1994) compares the performance of two CUSUM charts. In one chart he considered the time between Poisson events, which has exponential distribution, and in the other chart he considers aggregate counts, which have Poisson distribution. He considers aggregation time periods of length 1 and 10 times units. Schuh et al. (2013) extend Gan (1994) investigation by exploring to a greater extent the relative performance of the exponential CUSUM and Poisson CUSUM control charts, considering aggregation time periods of length 1, 7, 14, and 30, taking into

account that weekly, biweekly, and monthly, are more commonly used aggregation periods in public health and safety.

3. EWMAG Chart

As indicated earlier, Shen et al. (2013) proposed the EWMAG chart, which uses probability control limits in the EWMA control chart for monitoring Poisson count data with time-varying sample sizes in Phase II.

The proposed EWMA chart is called EWMAG chart because its in-control run length distribution, is theoretically identical to the geometric distribution, i.e., the false alarm rate does not depend on the time of the monitoring, nor does the sample size being monitored.

Let X_t be the count of an adverse event during the fixed time period $(t-1, t]$ (count of events at time t). Suppose X_t independently follows the Poisson distribution with the mean θn_t conditional on n_t, where θ and n_t denote the occurrence rate of the event and sample size at time t respectively. The objective is to detect an abrupt change in the occurrence rate from θ_0 to another unknown value $\theta_1 > \theta_0$. The EWMAG chart uses

$$Z_i = (1-\lambda)Z_{i-1} + \lambda\frac{X_i}{n_i} \tag{1}$$

as the charting statistic, where $Z_0 = \theta_0$, and $\lambda \in (0, 1]$ is a smoothing parameter which determines the weights of past observations.

The control limit of the EWMAG chart must satisfy the following equations

$$\begin{gathered} P(Z_1 > h_1(\alpha) \mid n_1) = \alpha \\ P(Z_t > h_t(\alpha) \mid Z_i < h_i(\alpha), 1 \leq i < t, n_t) = \alpha \quad for \quad t > 1 \end{gathered} \tag{2}$$

where $h_t(\alpha)$ is the control limit at time t and α is the prespecified false alarm rate. At time t, the probability control limit is determined right after we observe the value of n_t. Consequently, the EWMAG chart does not need the assumption of future sample sizes and does not suffer from wrong model assumptions. This property makes the proposed EWMAG chart significantly different from previous control charts.

Shen et al. (2013) consider that, because of the intricacy of the conditional probability (2) it is impossible to solve $h_t(\alpha)$ analytically. Thus, computational procedures are necessary. The procedure in order to find the probability control limits is summarized in the following algorithm:

1. At time t, and under the in-control condition, X_t should follow the Poisson distribution with mean $\theta_0 n_t$, where n_t is exactly known. If there is no out-of-control signal at time $t-1$ $(t = 1, 2, ...)$, $\widehat{X}_{t,i}$ $(i = 1, \ldots, M)$ are generated from the distribution Poisson $(\theta_0 n_t)$. Accordingly, M values of the pseudo charting statistic Z_t are obtained through

$$\widehat{Z}_{t,i} = (1-\lambda)\widehat{Z}_{t-1,j} + \lambda\frac{\widehat{X}_{t,i}}{n_t} \tag{3}$$

where $i = 1,\ldots,M$, $j \in \{1,\ldots,M'\}$, with $M' = \lfloor M(1-\alpha)\rfloor$, and $\widehat{Z}_{t-1,j}$ is uniformly selected from $\widehat{\mathbf{Z}}'_{[t-1]M'}$. Here, $\lfloor M(1-\alpha)\rfloor$ denotes the largest integer less than or equal to $M(1-\alpha)$, and $\widehat{\mathbf{Z}}'_{[t-1]M'}$ contains the ranked values $\widehat{Z}_{t-1,(1)},\ldots,\widehat{Z}_{t-1,(M')}$ which are less than or equal to $h_{t-1}(\alpha)$. When $t = 1$, $\widehat{Z}_{t-1,j} = \theta_0$, for all j.

2. Sort the values $\widehat{Z}_{t,1}$, $\widehat{Z}_{t,2},\ldots,\widehat{Z}_{t,M}$ in ascending order, and the α upper empirical quantile of these M values, is used for approximating the control limit $h_t(\alpha)$.

3. Compare the value of $\widehat{Z}_t$, which is calculated based on observed X_t and n_t, with $h_t(\alpha)$, to decide whether to issue an out-of-control signal or to continue toward the next time point.

4. If it is decided to continue, the values $\widehat{Z}_{t,(M'+1)},\ldots,\widehat{Z}_{t,(M)}$ are removed. Then go back to step 1.

3.1. EWMAG chart for Aggregated Poisson Data

In this section, the proposal of Shen et al. (2013) is adapted to the scenario in which data are aggregated. We suppose that $X_1, X_2, \ldots, X_i, \ldots, X_m,\ldots$ are the counts of events during periods of time of equal length $(t_1-1, t_1]$, $(t_2-1, t_2],\ldots,(t_i-1, t_i],\ldots,(t_m-1, t_m],\ldots$ For simplicity, and in accordance to Shen et al. (2013), we will call these the counts in the times $t_1, t_2, \ldots, t_i, \ldots, t_m,\ldots$ respectively. We suppose that $X_i \sim P(\theta_0 n_{t_i} \mid n_{t_i})$, where θ_0 is the occurrence rate of the event, and n_{t_i} is the sample size at time t_i. If we aggregate the counts from time t_i until time t_m, it is obtain $Y_{t_{im}} = \sum_{k=i}^{m} X_k$, that have distribution $P(\theta_0 \sum_{k=i}^{m} n_{t_k} \mid \sum_{k=i}^{m} n_{t_k})$. Thus, according to Shen et al. (2013), it is possible to implement the EWMAG chart for the variable Y. We will assume that the periods of aggregation have the same length, for example, a week, a month, a year, and so on. Let $Y_{t_{1r}}$ the variable Y observed from t_1 until t_r (union of the time interval from $(t_1-1, t_1]$ until $(t_r-1, t_r]$). Under the in-control condition, $Y_{t_{1r}}$ should follow the Poisson distribution with mean $\theta_0 \sum_{k=1}^{r} n_{t_k}$ conditional on $\sum_{k=1}^{r} n_{t_k}$, where $\sum_{k=1}^{r} n_{t_k}$ is exactly known. Therefore we can obtain the control limits during the first time period of aggregation, by randomly generating $\widehat{Y}_{t_{1r},i}$, where $i = 1,\ldots,M$, from the distribution $P(\theta_0 \sum_{k=1}^{r} n_{t_k})$ and correspondingly calculating M values of pseudo $Z_{t_{1r}}$ from (1) with $Z_0 = \theta_0$, say $\hat{Z}_{t_{1r},1},\ldots\widehat{Z}_{t_{1r},M}$. Let us store in a vector $\widehat{\mathbf{Z}}_{1rM}$ the values $\widehat{Z}_{t_{1r},1},\ldots\widehat{Z}_{t_{1r},M}$, sorted in ascending order. Thus control limit $h_{t_{1r}}(\alpha)$ can be approximated as the $M' = \lfloor M(1-\alpha)\rfloor$ largest values in $\widehat{\mathbf{Z}}_{1rM}$. After, $h_{t_{1r}}(\alpha)$ is compared with $Z_{t_{1r}}$, which is calculated based on the observed $Y_{t_{1r}}$ and $\sum_{k=1}^{r} n_{t_k}$. An out-of-control signal is issued if $Z_{t_{1r}} > h_{t_{1r}}(\alpha)$. Otherwise, it is

possible to move forward to the next time period, from t_{r+1} until t_s (union of the time interval from $(t_{r+1}-1, t_{r+1}]$ until $(t_s-1, t_s]$), $s > r+1$.

According to (2) in order to determine the control limit $h_{t_{r+1,s}}(\alpha)$ corresponding to time period indicated above, we should ensure that the value of pseudo $Z_{t_{1r}}$ is less than or equal to $h_{t_{1r}}(\alpha)$. Hence only the ranked values $\widehat{Z}_{t_{1r},(1)}, ...\widehat{Z}_{t_{1r},(M')}$ should be kept to determine $h_{t_{r+1,s}}(\alpha)$. We store the M' ranked pseudo $Z_{t_{1r}}$ into a vector $\widehat{\mathbf{Z}}'_{1rM'}$. Let $Y_{t_{r+1,s}}$ be the variable Y observed for the time period of aggregation t_{r+1} y t_s. Given $\sum_{k=r+1}^{s} n_{t_k}$, a vector $\widehat{\mathbf{Z}}_{[r+1]sM}$ with dimension M can be obtained throughout

$$\widehat{Z}_{t_{r+1,s},i} = (1-\lambda)\widehat{Z}_{t_{1r},j} + \lambda \frac{\widehat{Y}_{t_{r+1,s},i}}{\sum_{k=r+1}^{s} n_{t_k}} \tag{4}$$

where $i = 1, \ldots, M$, $\widehat{Z}_{t_{1r},j}$ is uniformly selected from $\widehat{\mathbf{Z}}'_{1rM'}$ with $j \in \{1, \ldots, M'\}$, and $\widehat{Y}_{t_{r+1,s},i}$ are randomly generated from $P(\theta_0 \sum_{k=r+1}^{s} n_{t_k})$. After sorting the M elements of $\widehat{\mathbf{Z}}_{[r+1]sM}$ in ascending order, the control limit $h_{t_{r+1,s}}(\alpha)$ is obtained by setting it at the $(1-\alpha)$-quantile of the M elements. An out-of-control signal is issued if $\widehat{Z}_{t_{r+1,s}} > h_{t_{r+1,s}}(\alpha)$. Otherwise, it is possible to move forward to the next time period, $[t_{s+1}, t_u]$, $s > r+1$. Repeat the above procedures by simulating M samples of $P(\theta_0 \sum_{k=s+1}^{u} n_{t_k})$,...etc.

4. EWMAe Chart

According to Dong et al. (2008), let the discrete time stochastic process under surveillance be denoted by $X = \{X^*(t), t = 1, 2, ...\}$, where $X^*(t), t \geq 1$ are assumed to be conditionally independent given a random changepoint τ. They also assume that $X^*(t)$ is distributed as Poisson $(N_t\theta_0 I\{t < \tau\} + N_t\theta_1 I\{t \geq \tau\})$, where N_t is a constant representing the number of product exposures at time interval t and $I\{t < \tau\}$ and $I\{t \geq \tau\}$ are the indicator functions. The authors study three types of EWMA methods based on an exponentially weighted moving average, Z_s, of all accumulated observations. The alarm statistic can be equivalently represented by the recursive formula

$$Z_s = (1-\lambda)Z_{s-1} + \lambda \frac{X^*(s)}{N_s}, \quad Z_0 = \theta_0 \tag{5}$$

or

$$Z_s = (1-\lambda)^s \theta_0 + \lambda \sum_{t=1}^{s} (1-\lambda)^{s-t} \frac{X^*(t)}{N_t}$$

where the weight parameter is $\lambda \in (0, 1]$.

When the process is in-control, the mean of Z_s, $E(Z_s)$, is equal to θ_0 and the variance of Z_s is

$$\sigma_s^{(\infty)^2} = \lambda^2 \sum_{t=1}^{s} (1-\lambda)^{2s-2t} \frac{\theta_0}{N_t} \tag{6}$$

Let $N_0 = \min\{N_t\}$. Then

$$\sigma_s^{(\infty)^2} \leq \frac{\theta_0}{N_0}\frac{\lambda}{2-\lambda}\{1-(1-\lambda)^{2s}\} = \sigma_s^{*(\infty)^2} \tag{7}$$

and

$$\lim_{s\to\infty}\sigma_s^{*(\infty)^2} = \frac{\theta_0}{N_0}\frac{\lambda}{2-\lambda} = \sigma^{*(\infty)^2} \tag{8}$$

Hence, $\sigma^{*(\infty)^2}$ is the asymptotic variance of Z_s.

The superscript (∞) represents $\tau = \infty$, which corresponds to the in-control state of the process. Here, τ represents the occurence time of change.

Dong et al. (2008) define three EWMA methods. The first one, called EWMAe-type, has the time of an alarm as

$$t_A = \min\{s;\, Z_s > \theta_0 + L\sigma_s^{(\infty)}, s \geq 1\} \tag{9}$$

The second, designated EWMAa1-type, has the time of an alarm as

$$t_A = \min\{s;\, Z_s > \theta_0 + L\sigma_s^{*(\infty)}, s \geq 1\} \tag{10}$$

The third, designated EWMAa2-type, has the time of an alarm as

$$t_A = \min\{s;\, Z_s > \theta_0 + L\sigma^{*(\infty)}, s \geq 1\} \tag{11}$$

If $\lambda = 1$, then just the last observation is used in the alarm statistic and the EWMAa1-type and EWMAa2-type methods coincide.

As Ryan & Woodall (2010) pointed out, the equations (7) and (8) leads to a problem when we are in phase II, since they require the knowledge of the minimum sample size, N_0, of the entire set of samples. In practice, this is unlikely since the samples are taken in real time.

We use (6) when establishing the control limits, since with this expression it is not necessary to know N_0.

4.1. EWMAe chart for Aggregated Poisson Data

Analogous to that described in subsection 3.1, in this section we adapted the proposal of Dong et al. (2008), which is described in section 4, to the case in which aggregated information is used.

If in times $t_1, t_2, \ldots, t_i, \ldots, t_{n,\ldots}$ are observed Poisson data $X^*(t_1), X^*(t_2), \ldots, X^*(t_i), \ldots, X^*(t_n), \ldots$, with means $\theta_0 n_{t_i}$ conditional on n_{t_i} respectively, then in the time period $[t_i, t_j]$, $i, j \in \mathbb{Z}^+$ the observation $Y_{ij} = \sum_{k=i}^{j} X^*(k)$ will be obtained, which have Poisson distribution with mean $\theta_0 \sum_{k=i}^{j} N_{t_k}$ conditional on $\sum_{k=i}^{j} N_{t_k}$. Thus, according to Dong et al. (2008), it is possible to implement the EWMAe chart for the variable Y.

Analogous to the first type of EWMA method described by Dong et al. (2008), the monitoring statistic in this case is

$$Z_{s_a} = (1-\lambda)Z_{s_a-1} + \lambda\frac{Y(s_a)}{\sum_{k=i}^{j} N_{t_k}}, \quad Z_0 = \theta_0 \tag{12}$$

or

$$Z_{s_a} = (1-\lambda)^{s_a}\theta_0 + \lambda\sum_{t=1}^{s_a}(1-\lambda)^{s_a-t}\frac{Y(t)}{\sum_{k=i}^{j} N_{t_k}} \tag{13}$$

where s_a indicate the times in which the process is monitored after a certain period of aggregation. When the process is in-control, the mean of Z_{s_a} is

$$\begin{aligned} E(Z_{s_a}) &= E\left\{(1-\lambda)^{s_a}\theta_0 + \lambda\sum_{t=1}^{s_a}(1-\lambda)^{s_a-t}\frac{Y(t)}{\sum_{k=i}^{j} N_{t_k}}\right\} \\ &= \theta_0\left\{(1-\lambda)^{s_a} + \lambda\sum_{t=1}^{s_a}(1-\lambda)^{s_a-t}\right\} \\ &= \theta_0 \end{aligned}$$

and the variance of Z_{s_a}, $\sigma_{s_a}^{(\infty)^2}$, is

$$\begin{aligned} \sigma_{s_a}^{(\infty)^2} &= Var(Z_{s_a}) \\ &= Var\left\{(1-\lambda)^{s_a}\theta_0 + \lambda\sum_{t=1}^{s_a}(1-\lambda)^{s_a-t}\frac{Y(t)}{\sum_{k=i}^{j} N_{t_k}}\right\} \\ &= \lambda^2\left\{\sum_{t=1}^{s_a}((1-\lambda)^{2(s_a-t)})\frac{\theta_0}{(\sum_{k=i}^{j} N_{t_k})}\right\} \end{aligned} \tag{14}$$

The variance $\sigma_{s_a}^{(\infty)^2}$ can be equivalently represented by the recursive formula

$$Var(Z_{s_a}) = \sigma_{s_a}^{(\infty)^2} = \lambda^2\frac{\theta_0}{\sum_{k=i}^{j} N_{t_k}} + (1-\lambda)^2 Var(Z_{s_a-1})$$

The stopping rule of EWMAe chart for aggregated Poisson data, t_{A_a}, is

$$t_{A_a} = \min\{s_a;\, Z_{s_a} > \theta_0 + L\sigma_{s_a}^{(\infty)}, s_a \geq 1\}$$

When aggregated observations are used, the control limits use (14), which is analogous to (6).

5. Simulation and Results

In this section we study the behavior of ARL_1 when we consider different levels of aggregation, combined with different changes in θ_0. The study provides

information about the effect of aggregation on monitoring processes, when the EWMAG and EWMAe charts are used. Simulation processes are used for this purpose. Two scenarios of sample size generation, different out-of-control signals caused by increasing the in-control occurrences rate $\theta_0 = 1$, and different levels of aggregation are considered.

The aggregation level represents the number of registered points whose results are added to be monitored later. The level of aggregation is associated with the length of the interval at which the observations are taken, and are added to the end of this. A larger aggregation interval implies a higher level of aggregation. For example, if from a process, daily data is taken, an aggregation period of one week implies a level of aggregation equal to 7.

As in Shen et al. (2013), we consider $\lambda = 0.1$, and we assume that the out-of-control states occur when $\tau = 1, 5, 10, 20$ and 50. The scenarios of sample size generation, called I and II respectively and considered in Shen et al. (2013), are

$$n_t = \frac{13.8065}{8 \times (0.5 + exp(-(t - 11.8532)/26.4037))}$$

and

$$n_t \sim U(1, 4)$$

In each chart, the ARLs are obtained from 30000 replicates. Furthermore, for the EWMAG chart, we use $M = 30000$ simulated Poisson observations in each case. Also a simulation study has been done to determine the effect of data aggregation under the presence of outliers. In this last case, in which we used the scenario I only, a fixed percentage of all monitored data is contaminated using data from a different distribution from the one where the data are generated in-control, and the number of alarms generated by each chart is determined. The data coming from in-control processes were contamined with outliers from Poisson distributions with rates of adverse events equal to $1.025, 1.100, 1.250, 1.500, 2.000$ and 2.500. In each of these cases, 1000 data were generated, from which, 5% was contaminated. That is, in each of these cases, 1000 data were monitored, of which 50 were outliers. Then, the total number of detections (nd), and the number of correctly detected outliers (cd) by the charts, were compared taking into account aggregated and non aggregated data.

For the EWMAe chart, values L were adjusted for each level of aggregation and for each scenario related to the sample size, such that, the in-control ARL, with aggregated and non aggregated information, is approximately equal to 370 in all cases. Then, different changes were introduced in the rate of adverse events θ, from an in-control value θ_0, toward an out-of-control value $\theta_1 > \theta_0$. For the EWMAG chart, the same scenarios of simulation of sample sizes were used, and a false alarm probability (α) equal to 0.0027 was used in absence and presence of aggregation. This value of α ensures an in-control average run length equal to 370.

Table 1 considers sample sizes generated from scenario I, and $\tau = 1$. Here it can be seen how the out-of-control ARL of the EWMAG and EWMAe charts decreases, when the aggregation level increases. This behavior indicates the way in

which the charts increase their sensitivity, by increasing the aggregation level used in each time point. However, it should be noted that the number of individual sampling points, used when there are aggregated data, is greater as indicated by the ARL. For example, when we have data without aggregate in each monitoring time point, a shift toward 1.1 in rate of adverse events, θ, is detected by the EWMAe and EWMAG charts, in average, after 100 time points (99.2 and 97.6) approximately. These charts, however, detect this same change in θ, in average, after about 62 monitoring points (62.4 and 62.7), when the aggregation level is equal to two, which is equivalent to, approximately, 124 individual sampling points. When the level aggregation is equal to three, the charts detect this change, on average, after about 46 monitoring points (46.4 and 45.9), which is equivalent to about 138 individual sampling points. This shows that the sensitivity of charts increases when the aggregation level increases, but also increases the number of samples to be used.

Table 1: Out-of-control ARL comparison of EWMAG and EWMAe charts for different levels of aggregation. Scenario I and $\tau = 1$.

	EWMAG						EWMAe					
	Aggregation levels						Aggregation levels					
θ	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
1.000	368.8	369.1	368.3	368.7	368.5	368.4	371.1	369.2	371.4	370.2	369.1	369.8
1.025	243.8	204.3	182.2	164.9	150.2	139.1	245.6	207.1	182.0	165.0	151.2	141.6
1.100	99.2	62.4	46.4	37.8	31.7	27.2	97.6	62.7	45.9	37.1	31.0	26.8
1.250	39.4	22.1	14.7	12.4	10.3	8.7	36.7	21.2	15.8	11.7	9.5	8.1
1.500	18.2	10.1	7.1	5.5	4.5	3.8	16.0	8.9	6.0	4.7	3.7	3.1
2.000	7.5	4.0	2.8	2.1	1.6	1.3	6.1	3.2	2.0	1.5	1.0	0.9
2.500	4.4	2.2	1.5	1.1	0.7	0.6	3.2	1.6	0.9	0.6	0.4	0.3

From Table 1, also can be seen that, for example, a change of 25% in θ, with respect to its value in control, is detected by the EWMAG chart after monitoring approximately 39 disaggregate points, while that for data with a level of aggregation equal to 2, is detected after monitoring approximately 22 points, which are equivalent to 44 individual sampling points. A similar result is obtained for the EWMAe chart. Depending on the real situation, these differences in the total number of samples used for aggregated and non aggregated data, could be of little significance. For this same change in θ and great aggregation levels, in general, there are significant differences between aggregated and non aggregated data.

Table 2 considers sample sizes generated from scenario II, and $\tau = 1$. Here, similar to what is shown in Table 1, the out-of-control ARL of the EWMAG and EWMAe charts, decreases when the aggregation level increases in each time point monitored, as well as it also increases the total of samples used. In this case, the out-of-control ARL is generally smaller than in scenary I for large changes in θ, and bigger than in scenary I for small changes in θ.

According to Shabbak & Midi (2012), in statistical quality control, a process changes into an out-of-control situation when outliers appear in two different ways, namely, outliers that are randomly distributed within a data set and outliers that sequentially occur after a specific observation during a specific period of time in the data set.

TABLE 2: Out-of-control ARL comparison of EWMAG and EWMAe charts for different levels of aggregation. Scenario II and $\tau = 1$.

	EWMAG						EWMAe					
	Aggregation levels						Aggregation levels					
θ	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
1.000	370.1	369.0	369.3	371.1	370.2	369.2	371.3	370.6	368.7	369.2	368.8	370.2
1.025	259.7	218.6	201.9	179.8	168.6	157.1	261.1	223.8	198.6	181.5	166.0	157.6
1.100	105.4	68.7	51.0	42.2	35.3	30.0	106.8	68.2	49.9	39.3	32.9	28.5
1.250	32.7	18.0	12.5	9.9	8.0	6.9	31.3	17.4	11.2	8.6	7.0	5.9
1.500	11.1	5.8	4.0	3.0	2.4	1.9	9.8	4.8	3.2	2.3	1.8	1.5
2.000	3.7	1.7	1.1	0.7	0.5	0.4	2.8	1.2	0.7	0.4	0.3	0.2
2.500	1.8	0.7	0.4	0.2	0.1	0.1	1.2	0.4	0.2	0.05	0.04	0.02

The second case was studied in Tables 1 and 2. Table 3 shows the results of a simulation study for the first case. Here, a process in-control with occurrences rate $\theta_0 = 1$ is contaminated with outliers generated from a process with occurrences rate greater than θ_0. More specifically, five Poisson processes in control are considered, and each of them is contaminated with outliers generated from one of the five out-of-control processes, depending on the considered changes in θ, that is, $1.025, 1.100, 1.500, 2.000$ or 2.500. In each case, 1000 observations are monitored, of which 5% correspond to outliers. Outliers are generated at specific points, so that they can be fully identified . The total number of detections (nd) is determined, as well as the number of correctly detected outliers (cd) by the charts. All of the above was done considering aggregated and non aggregated data. For aggregated data, three aggregation levels were considered: 2, 4 and 5. In order to have more stable results, all the above was replicated 10,000 times. As an example, of 1000 data monitored individually, with a contamination of 5%, and with an increase of 50% in the rate of adverse events, the EWMAG chart with disaggregated data had a total of 16 detections approximately (15.6), and 2 correctly detected outliers approximately (1.9). However, with the same conditions as the ones above, but now considering an aggregation level equal to 2, the values corresponding to nd and cd, are 9.5 and 1.7 respectively. A similar behavior occurs in the EWMAe chart. Table 3 shows that in general the aggregation level affects the total number of detections, but has very little effect on the number of correctly detected outliers.

TABLE 3: Comparison of the total number of detections and correctly detected outliers, when we consider aggregated and non aggregated data with 5% contamined. Scenario I.

	EWMAG								EWMAe							
	Aggregation levels								Aggregation levels							
	Level 1		Level 2		Level 4		Level 5		Level 1		Level 2		Level 4		Level 5	
θ_1	nd	cd	nd	cd	nd	cd	nd	cd	nd	cd	nd	cd	nd	cd	nd	cd
1.025	8.1	0.4	4.1	0.4	2.1	0.4	1.8	0.4	7.8	0.4	4.0	0.4	2.0	0.4	1.8	0.4
1.100	9.0	0.6	4.8	0.6	2.4	0.5	2.0	0.5	8.6	0.6	4.6	0.5	2.4	0.5	2.0	0.5
1.500	15.6	1.9	9.5	1.7	5.8	1.7	4.9	1.8	14.9	1.8	9.3	1.7	5.7	1.7	4.8	1.8
2.000	30.6	5.7	20.4	5.2	15.8	5.7	15.2	6.2	29.5	5.6	19.9	5.1	15.5	5.6	14.8	6.1
2.500	60.0	12.5	43.4	11.6	33.6	12.8	32.2	13.9	58.2	12.3	42.6	11.5	33.1	12.7	31.7	13.7

Frequently, the performance of a control chart in Phase II, is measured in terms of its ARL. According to Ryan & Woodall (2010), the measured ARL values can either be zero state or steady-state ARL values. Zero-state ARL values are based on sustained shifts in the parameter that occur under the initial startup conditions of the control chart, while steady-state ARL values are based on delayed shifts in the parameter.

Table 4 shows the out-of-control ARL of EWMAG and EWMAe charts when there is an increase of 25% in occurrence rate of the event. Two levels of aggregation and five changepoints are considered. In both charts it can be noted that when the changepoint time increases, so does the ARL_1, both in the case of aggregated data as in disaggregated data. The effect that the aggregation procedures have on the values of the ARL_1, varies very little for the different changepoints. This can be observed by comparing, through a quotient, the ARL_1 of disaggregated case, with its corresponding of the aggregate case, for each of the values of τ, and in each chart. For example, by doing the quotient between the values corresponding to level 1 of the EWMAe chart, with the corresponding values of level 2, the values $1.7, 1.8, 1.8, 1.8, 1.9$ are obtained, which indicates that the proportion between the ARL_1 of two cases, does not change significantly when the value of τ increases. Something similar occur when it is compared level 1 with level 3. A similar behavior, have the ARL_1 of the EWMAG chart.

Table 4: Out-of-control ARL comparison of EWMAG and EWMAe charts with different τ, and two aggregation levels. Scenario I.

	EWMAG			EWMAe		
	Aggregation levels			Aggregation levels		
τ	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
1	39.4	22.1	14.7	36.7	21.2	15.8
5	42.7	23.7	16.9	42.6	23.9	17.0
10	46.1	25.5	18.1	47.3	25.7	18.8
20	52.8	29.0	20.5	53.5	29.5	21.2
50	77.5	41.1	28.3	78.1	41.5	29.2

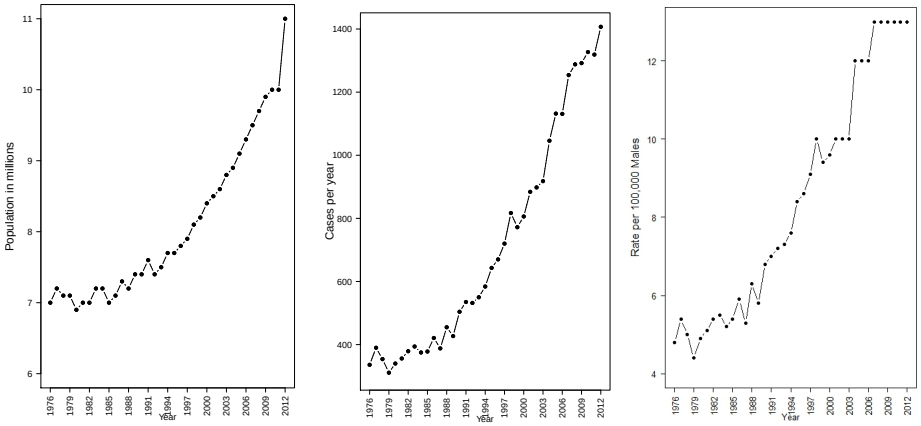
6. Example

In this section, the effect of aggregation of data is displayed in a real case. The information analyzed is filed at the Department of Health of New York, from 1976 to 2012. It corresponds to data related to liver cancer in men in the state of New York in the indicated period. From the available information, one can calculate the sample size in each year, which, together with the number of registered cases, allows to do the calculus of the incidence of this disease in the state and periods indicated. These data is available at www.health.ny.gov[1]

Figure 1(a) shows the population of New York between 1976 and 2012, estimated from available information at the website above mentioned, 1(b) shows the behavior of the number of cancer cases throughout the period, and 1(c) shows the

[1] http://www.health.ny.gov/statistics/cancer/registry/table2/tb2prostatenys.htm

incidence rate of liver cancer per 100,000 males in New York during the period of study. The increasing trend in the number of cases and incidence rate, occurs in almost all the period of study. The interest here is to monitor the incidence rate of patients with cancer in this period.



(a) Population New York (b) Cases of liver cancer (c) Incidence rate of liver cancer

FIGURE 1: Male liver cancer incidence and related population.

According to the pattern described by the incidence rate and taking into account the significant change between the time point 12 (year 1987) and 13 (year 1988), the period from 1976 to 1987 is chosen as the period reference for estimation of the incidence rate in control of liver cancer, i.e., this period is considered Phase I of process. This allows to obtain $\theta_0 = 0.73$. With this estimated value for θ_0, begins the monitoring in phase II, from the year 1988. According to Shen et al. (2013), a calibration sample of this size may not be large enough to precisely determine the true value of θ_0, but it suffices to illustrate the effect of data aggregation in a real-world setting.

From Figure 2 it can be seen that when the EWMAG chart is used with non-aggregated data, it sends an out-of-control signal at point 10 (year 1997). When aggregations of length 2, 4 and 6 are used, the out-of-control signals are given at points 5 (year 1997), 3 (year 1999), and 2 (year 1999) respectively. When the EWMAe chart is used, as shown in Figure 3, similar results are found. In this application it can be appreciated that an aggregation level equals to 2, does no affect on the sensibility of control charts to detect a change in θ. Levels aggregation of length 4 and 6 only affect the detection time at two periods. In Figure 2 and 3, the dotted line without markers represents the control limit of each chart, and the dotted line with markers represents the charting statistics.

From the data of liver cancer, it can be determined that in 1997, year in which the charts, with disaggregated data, issue a signal, the rate of occurrence is equal to 1.15, i.e., there is an increase in θ of about 56% with respect to θ_0. With this large change in θ, using data without aggregating, or using an aggregation level

equal to 2, is unimportant. Similar reasoning can be made when aggregation levels of 4 and 6 are used.

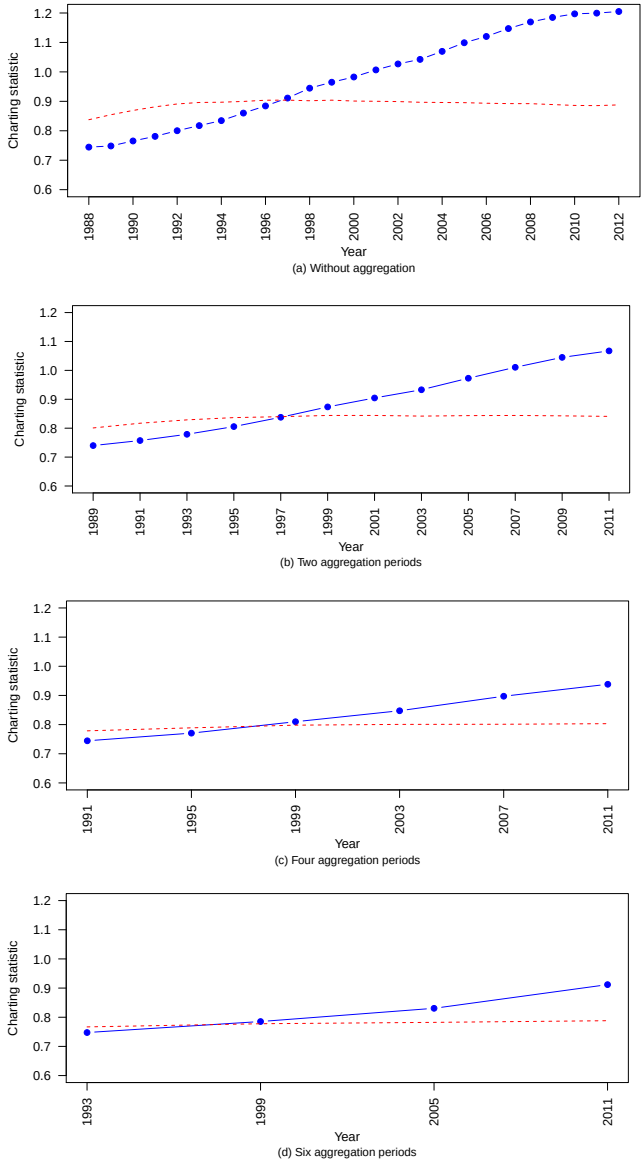


FIGURE 2: Monitoring of cancer liver data with the EWMAG Chart for different levels of aggregation.

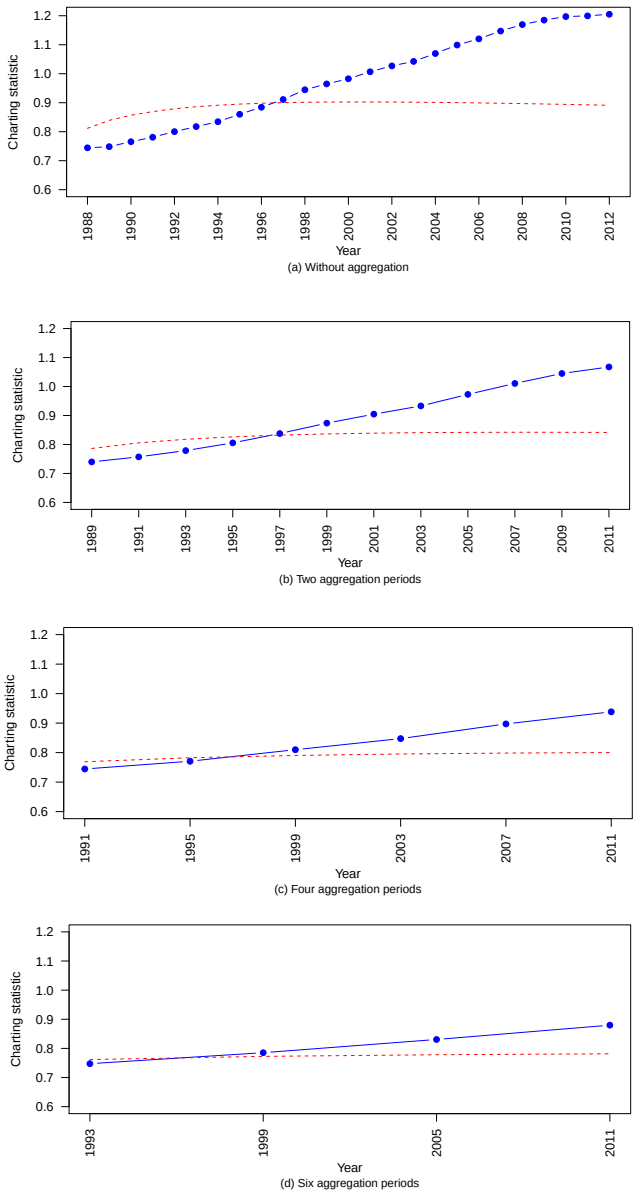


FIGURE 3: Monitoring of cancer liver data with the EWMAe Chart for different levels of aggregation.

This example shows that there are situations in which aggregation data processes can be implemented, affecting very little the sensitivity of the monitoring process. However, when we have little knowledge of a particular situation, it is advisable to use low levels of aggregation.

7. Conclusions

Data aggregation is a practice used in many cases. Although it is an important issue, there are few studies to evaluate its impact on the monitoring of processes. In areas such as health monitoring, the late detection of changes in the rate of adverse events is of vital importance. Consequently, it is necessary to further study the effects of this practice on the monitoring processes.

Two aspects should be taken into account at the moment of considering data aggregation: the aggregation level and the maximum magnitude of change in the parameter of interest that can be tolerated. In this article, we have identified some cases in which data aggregation does not involve important adverse effects, as well as those in which only low levels of this, are recommended.

Simulation studies allow us to conclude that in monitoring of Poisson count data, with a sample size not constant over time, data aggregation has some adverse effects, especially if it is desired to swiftly detect small changes in the rate of adverse events, being the effect more acentuated, as the level of aggregation increases. However, this practice which is appropriate and needed in many cases, properly implemented, can provide good results. In many cases, low levels of aggregation do not affect monitoring processes, or do very little.

We found out that data aggregation does not significantly affect the early detection of out-of-control states when changes in the parameter of interest of about 25% are tolerated, even for large levels of aggregation. For changes of 50% or more, data aggregation practically has no adverse effects on the efficiency of monitoring procedures, regardless of the level of aggregation. More detailed simulation studies could reveal more precisely the magnitudes of changes in the rate of adverse events, for which it could be considered workable data aggregation, without significantly affecting the efficiency of monitoring processes. The little effect of data aggregation in the detection of outliers should be noted. It can also be concluded that the effect that the aggregation procedures have on the values of the ARL_1, varies very little for different changepoints.

The EWMAG and EWMAe charts, designed to monitor the rate of adverse events in Poisson count data, when sample sizes are not constant over time, show a similar behavior in the presence of aggregated data, observing a better performance of the EWMAG chart, only when small changes in the rate of adverse events are detected, and there are low levels of aggregation. In other situations, the EWMAe chart usually shows lower values of out-of-control ARL. Even though in studies related to health surveillance, is essential to detect as soon as possible small changes in the rate of adverse events, there may be some situations in which moderate changes can be tolerated. In these cases, the data aggregation could be considered feasible without causing major problems, even for various aggregation levels.

Though there are sophisticated procedures for the proper handling of data with zeros excess in the context of control charts, data aggregation can be a simple and convenient alternative in this and many other cases.

Acknowledgments

The authors thank the referees for their valuable comments and suggestions that have resulted in significant improvements in the paper.

[Received: September 2016 — Accepted: March 2017]

References

Burkom, H. S., Elbert, Y., Feldman, A. & Lin, J. (2004), 'Role of data aggregation in biosurveillance detection strategies with applications from essence', *Morbidity and mortality weekly report* **53**, 67–73.

Dong, Y., Hedayat, A. & Sinha, B. (2008), 'Surveillance strategies for detecting changepoint in incidence rate based on exponentially weighted moving average methods', *Journal of the American Statistical Association* **103**(482), 843–853.

Dubrawski, A. & Zhang, X. (2010), 'The role of data aggregation in public health and food safety surveillance', *Biosurveillance: Methods and Case Studies* pp. 161–179.

Frisén, M. & De Maré, J. (1991), 'Optimal surveillance', *Biometrika* **78**(2), 271–280.

Gan, F. F. (1990), 'Monitoring poisson observations using modified exponentially weighted moving average control charts', *Communications in Statistics-Simulation and Computation* **19**(1), 103–124.

Gan, F. F. (1994), 'Design of optimal exponential CUSUM control charts', *Journal of Quality Technology* **26**(2), 109–124.

Huan, W., Shu, L., Woodall, W. H. & Tsui, K. L. (2016), 'CUSUM procedures with probability control limits for monitoring processes with variable sample sizes', *IIE Transactions* **48**(8), 759–771.

Jiang, W., Shu, L. & Tsui, K. L. (2011), 'Weighted CUSUM control charts for monitoring poisson processes with varying sample sizes', *Journal of Quality Technology* **43**(4), 346–362.

Reynolds, M. R. & Stoumbos, Z. G. (2000), 'A general approach to modeling CUSUM charts for a proportion', *IIE Transactions* **32**(6), 515–535.

Reynolds, M. R. & Stoumbos, Z. G. (2004*a*), 'Control charts and the efficient allocation of sampling resources', *Technometrics* **46**(2), 200–214.

Reynolds, M. R. & Stoumbos, Z. G. (2004*b*), 'Should observations be grouped for effective process monitoring?', *Journal of Quality Technology* **36**(4), 343–366.

Rossi, G., Lampugnani, L. & Marchi, M. (1999), 'An approximate CUSUM procedure for surveillance of health events', *Statistics in Medicine* **18**(16), 2111–2122.

Ryan, A. G. & Woodall, W. H. (2010), 'Control charts for poisson count data with varying sample sizes', *Journal of Quality Technology* **42**(3), 260–275.

Schuh, A., Woodall, W. H. & Camelio, J. A. (2013), 'The effect of aggregating data when monitoring a poisson process', *Quality control and applied statistics* **45**(3), 260–272.

Shabbak, A. & Midi, H. (2012), 'An improvement of the hotelling statistic in monitoring multivariate quality characteristics', *Mathematical Problems in Engineering* .

Shang, X. & Woodall, W. H. (2015), 'Dynamic probability control limits for risk-adjusted bernoulli CUSUM charts', *Statistics in Medicine* **34**(25), 3336–3348.

Shen, X., Zou, C., Tsung, F. & Jiang, W. (2013), 'Monitoring poisson count data with probability control limits when sample sizes are time varying', *Naval Research Logistics (NRL)* **60**(8), 625–636.

Shu, L., Jiang, W. & Tsui, K. L. (2011), 'A comparison of weighted CUSUM procedures that account for monotone changes in population size', *Statistics in medicine* **30**(7), 725–741.

Zhou, Q., Zou, C., Wang, Z. & Jiang, W. (2012), 'Likelihood-based EWMA charts for monitoring poisson count data with time-varying sample sizes', *Journal of the American Statistical Association* **107**(499), 1049–1062.