Chapter

Change Detection by Monitoring Residuals from Time Series Models

Tom Burr and Kim Kaufeld

Abstract

Change detection in time series can be approached by fitting a model to the no-change, ordinary background data and then monitoring time series of residuals, where a residual is defined as residual = data – fit. In many applications, models that fit time series data lead to residuals that exhibit no patterns unless the signal of interest is present. Therefore, an effective signal or change detection approach is to first fit a time series model to the background data without any signal and then monitor the time series of residuals for evidence of the signal. This chapter briefly reviews a few time series modeling options and then focuses on statistical tests for monitoring residuals, including Page's cumulative sum (cusum, a type of scan statistic), the ordinary cumulative sum (cumsum), the matched filter (a version of the Neyman-Pearson test statistic), and pattern tests, such as those used in quality control. Simulation and analytical approximation methods are recommended for studying test behavior, as illustrated in three examples.

Keywords: time series models, residuals, scan statistics, cusum, matched filter

1. Introduction

This chapter's focus is on change detection by monitoring residuals arising from fitted models to time series data. The residual at time t is defined as $r_t = x_t - \hat{x}_t$, where $\hat{x}_t$ is the predicted (estimated) value of the data x_t. Large residuals or patterns in residuals could indicate that some type of change has occurred compared to usual behavior during the analysis period that was used to fit the model. For example, the number of positive test results for a disease could show a sharp rise or decline compared to the recent past, perhaps indicating that a signal of interest is present, such as a more infectious strain emerging. As another example, assembly line productions monitor product quality, such as the diameter of a machined part, which can drift due to measurement effects and/or machining effects. Diameter drifting can lead to detectable residual patterns, where the residual is the measured diameter—target diameter. Prior to diameter drifting, the time series should vary randomly around a mean value that is close to the target mean diameter. Therefore, time series fitting of the "in control" process data simply requires estimation of the mean and standard deviation of the measured diameter of each part. Other time series fitting options are less simple.

There are many types of time series, such as series of the unit or system failure times in reliability data, series of new disease cases or deaths, series of measured product quality, such as geometric dimensions in machined parts or salt content in bags of chips. This chapter does not consider predicting the next failure time, the time to the next spike in disease counts, or the time for the machined part mean dimension to shift. Instead, the chapter focuses on monitoring for possible changes in the mean-time to failure, changes in the distribution of disease counts, or machined part dimensions. There are many applications in which a training period for model fitting is assumed to define normal behavior, and then the testing period monitors for various types of changes from the normal behavior, such as a shift to a different mean value.

Figure 1 is a time series of $n = 50$ values that have mean 0 and standard deviation 1 (independently and identically distributed normal random values, denoted iid N(0,1) in the figure caption) that exhibit no change in (a), a mean shift of 2 units on period 25 in (b) a mean shift of 2 units at time indices 26–30 in (c), and a mean shift at time indices 26–50 in (d). The human brain/eye is reasonably effective at spotting such changes but is vulnerable to being fooled by spurious patterns. Statistical methods, some very simple and some less simple have been developed to detect changes of interest, as this chapter explains.

Let $x_1, x_2, \ldots, x_n$ denote a time series, which is a sequence of values at times $1, 2, \ldots, n$. **Figure 2** plots an example simulated time series with $n = 50$.

An effective time series model leads to residuals that are approximately independently and identically distributed (iid) [1, 2]. The residual at time t is defined

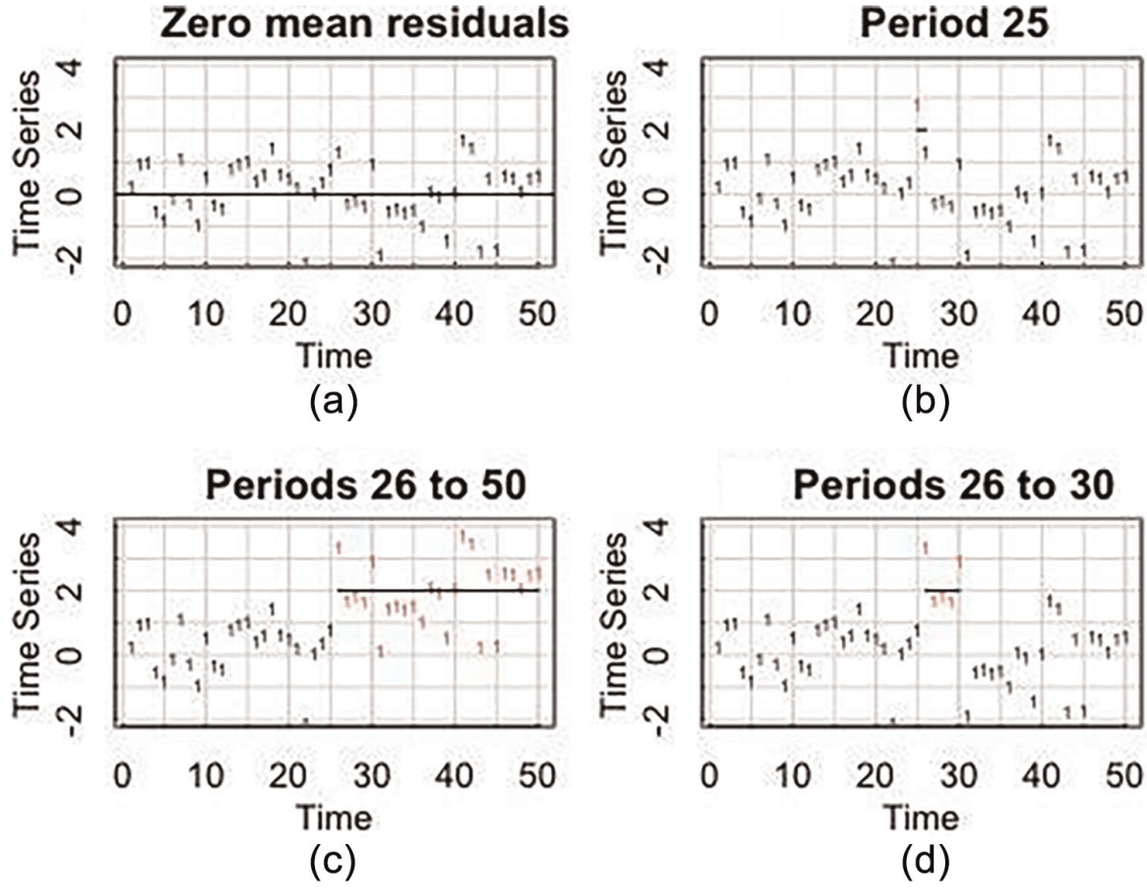

Figure 1.
Time series, iid N(0,1).

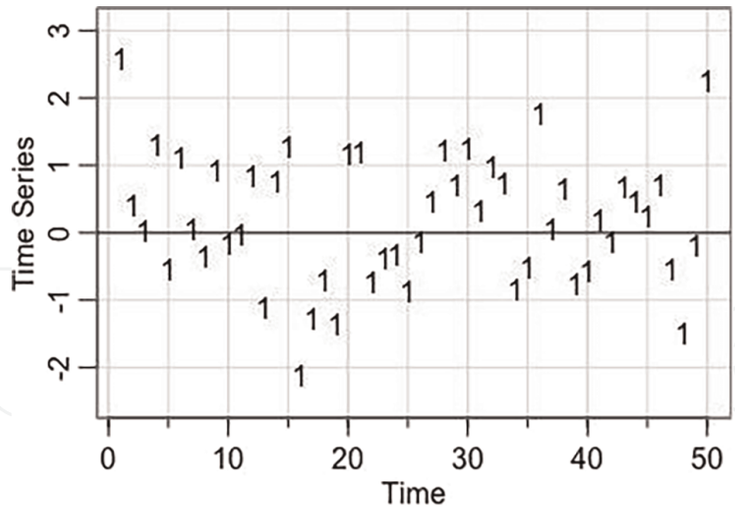

Figure 2.
Time series, MA(1) generated from iid N(0,1).

as $r_t = x_t - \hat{x}_t$, where $\hat{x}_t$ is the predicted (estimated) value of x_t. For example, $f()$ could be linear in the x's, resulting in the well-known auto-regressive (AR) model, $f() = \rho_1 x_{t-1} + \rho_2 x_{t-2} + \ldots + \rho_{l_A} x_{t-l_A}$ or $f()$ might be linear in both the x's and an underlying iid noise sequence $e_1, e_2, \ldots, e_n$, resulting in the auto-regressive moving average (ARMA) model $f() = \rho_1 x_{t-1} + \rho_2 x_{t-2} + \ldots + \rho_{l_A} x_{t-l_A} + \theta_1 x_{t-1} + \theta_2 x_{t-2} + \ldots + \theta_{l_M} x_{t-l_M}$ where l_A is the AR lag, and l_M is the moving average (MA) lag [1, 2]. The lag-one MA, AR, and ARMA models are given in (Eqs. (1)–(3)).

$$x_t = \theta_1 e_{t-1} + e_t \tag{1}$$

$$x_t = \rho_1 x_{t-1} + e_t \tag{2}$$

$$x_t = \rho_1 x_{t-1} + \theta_1 e_{t-1} + e_t \tag{3}$$

Conditions on the magnitudes of θ and ρ ensure stationarity (constant mean and variance over time). Often, time series can be transformed to stationarity by taking first differences, as is commonly done in stock market price series [1, 2]. Of course, $f()$ might not be linear in prior x values or e values and in general could be an arbitrarily complicated function, $f(x_{t-1}, \ldots, x_{t-l_A}, e_{t-1}, \ldots, e_{t-l_M})$.

Figure 2 is simulated data from a lag one MA model, $x_t = \theta_1 e_{t-1} + e_t$. **Figure 3** is the cumulative sum (cumsum, $C_t = \sum_{i=1}^{t} x_i$) and Page's cusum S_t with parameter k (see [3] and examples 2 and 3) defined as.

$$S_t = \max(0, S_{t-1} + x_i - k \tag{4}$$

for the data plotted in **Figure 2** [3]. Because of the reset-to-0 feature of S_t if the sum goes negative and because of the parameter k, Page's S_t does not have the large drift behavior that the cumulative sum does. **Figure 4** is the estimated underlying residual sequence and the actual underlying simulated residual sequence. **Figure 5** is the same as **Figure 4** but plots the difference between the estimated and true residuals. All plots and analyses are performed in R [4]. Statistical tests for changes in the background due to signal are applied to estimated residuals, so alarm thresholds and signal detection probabilities should be estimated using estimated residuals obtained via simulation.

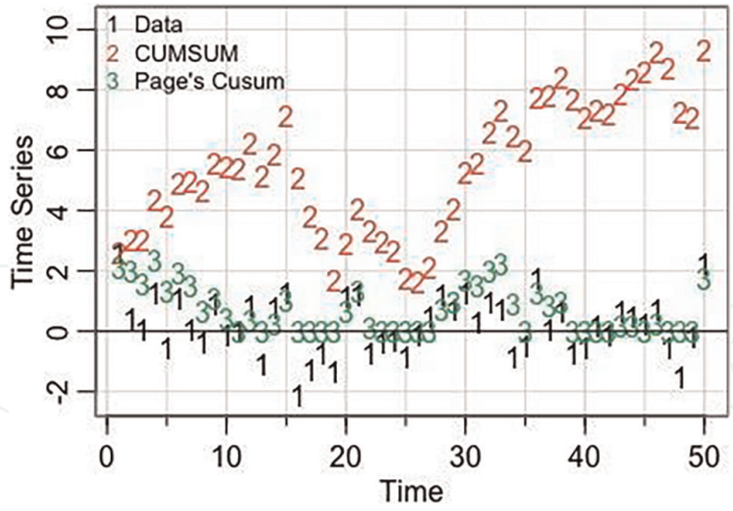

Figure 3.
IID N(0,1) time series, cumulative sum, and Page's Cusum.

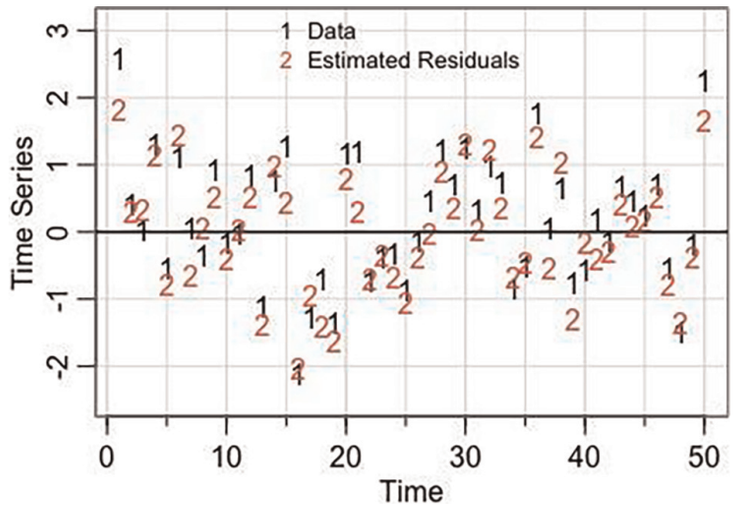

Figure 4.
Estimated residuals for MA(1) from N(0,1) data.

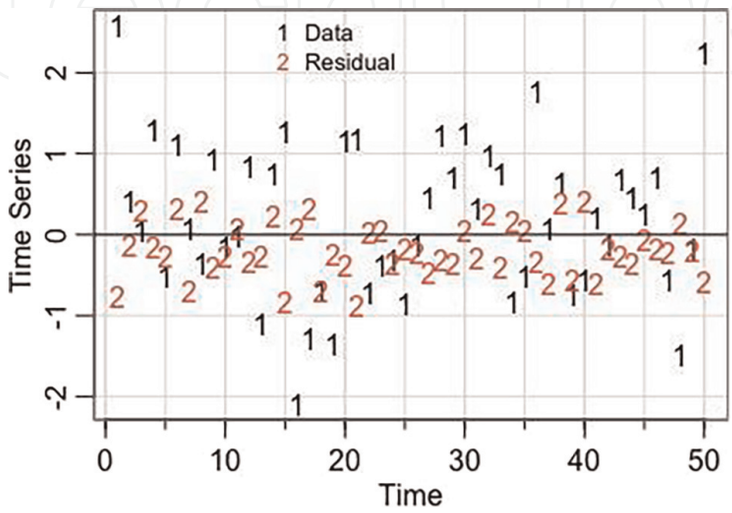

Figure 5.
*Same as **Figure 4**, but plotting the difference, residual-estimated residual.*

This chapter describes three change-detection examples. Example 1 monitors for patterns of large residuals, such as a consecutive string of three residuals exceeding a threshold. Example 2 monitors for excessive numbers of tweets in any of the 65 Florida counties. Example 3 monitors for nuclear material loss. Portions of Examples 1 and 3 have been published. Example 2 is entirely new.

2. Example 1: monitoring for patterns of large residuals

The Stein-Chen (SC) method approximates the probability density function (pdf) that assigns probabilities to the number of times that a pattern such as $I_t, I_{t+1}, I_{t+2} = \{1\ 0\ 1\}$ occurs, starting at position t in a binary time series of length n. In example 1, the original time series that is converted to binary is assumed to consist of a sequence of independent iid residuals that result from fitting any type of time series model. Recently the SC method was shown to provide an accurate Poisson-based approximation and corresponding total variation distance bounds in a time series context [5]. The binary values.

$I_1, I_2, I_3, \ldots, I_n$ are assumed to be independent and identically distributed with constant probability $p = P(I_i = 1)$. The probability p is the probability that the original time series X exceeds a threshold, and the I notation denotes an indicator or binary variable. As an aside, the SC method can also be applied if p is not constant over time, but the independence assumption is difficult to avoid [5–10]. Any type of time series model [1, 2] can be fit, and then the resulting residuals become the original series that is thresholded to convert to binary; therefore, the application is quite general. **Figure 6** is the estimated residuals from **Figure 4** but thresholded at 1.4 (values of 1.4 or larger are set to 1; values less than 1.4 are set to 0) to convert to binary.

Note that if $\{1\ 0\ 1\}$ is known to not occur, for example, starting at position $t = 1$, then this information impacts the probability that $\{1\ 0\ 1\}$ occurs starting at position $t = 2$ or $t = 3$, because the trials to obtain $\{1\ 0\ 1\}$are overlapping and thus not independent, so the Poisson distribution assumptions are not met. Nevertheless, Ref. [5] showed that Poisson-based approximation (that is strictly correct only for independent trials) can be remarkably accurate, and the SC method provides a bound on the total variation distance between the true and approximate pdf.

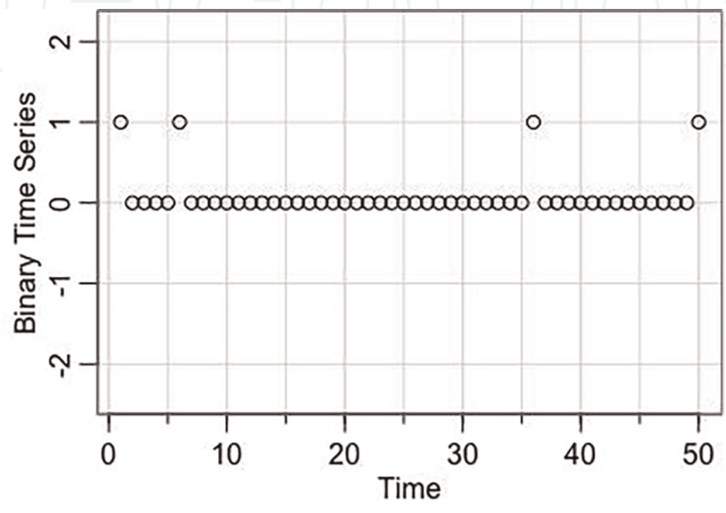

Figure 6.
Binary version of the estimated residuals in ***Figure 3****.*

Consider scanning for {1 *x* 1} with x = 0 or 1, with $p = P(I_i = 1)$ being quite small, such as 0.10 or less in a residual series of length n = 10. Then the probability of the pattern {1 *x* 1} is $p_p = p^2$, and there are $n - 2 = 8$ possible starting locations for the pattern in N = 10 trials. Because there are only 2^{10} = 1024 possible patterns of 0s and 1s, all 1024 patterns could be listed, and the probabilities assigned to each set of 10 binary values that include {1 × 1} at least once could be summed to provide an exact calculation. For larger values of n, this exact calculation is unwieldy, so an approximate method is desired, provided the approximation is highly accurate with provable error bounds.

Start at index i = 1 and check whether {1 × 1} occurs in positions {1 2 3}, then start at index i = 2 and check whether {1 × 1} occurs starting at index 2 in positions {2 3 4}, then start at index 3, etc. Note, for example, that if {1 × 1} occurs starting at position i = 1, then the probability that {1 × 1} also occurs starting at index 3 is p. Clearly, there is a small neighborhood of dependence around each starting index, as just illustrated. This neighborhood of dependence violates the assumptions for a Poisson distribution (as a limit distribution for a sequence of N Bernoulli trials, each with a small probability of success), but Ref. [6] shows that provided the dependence neighborhood is modest, the Poisson distribution can still provide an excellent approximation to the pdf defined on the number of times {1 × 1} occurs in a series of length N.

2.1 Stein-Chen method

The Poisson pdf with mean parameter $\lambda = (N-2)p_p$ provides an approximation Y to the true pdf W for the number of times {1 × 1} occurs in a series of length n [5, 6]. The value $(n-2)$ is used instead of n because the length 3 pattern could only be found starting at index 1, 2, … , $n-2$.

The quality of the Poisson(λ) approximation can be measured by computing b_1 and b_2, where $b_1 = \sum_{i=1}^{n} \sum_{j \in N_i} p_i p_j$, with $N_i = \{i-2, i-1, i, i+1, i+2\}$ being the dependent neighborhood N_i of index i, and $b_2 = \sum_{i=1}^{n} \sum_{j,j' \in N_i} E\{I_j I_{j'}\}$, where $j \neq j'$. The term b_3 is equal to 0 in Theorem 2 of Ref. [6] by the construction of N_i in this example. Then, the total variation distance (TVD) satisfies.

$$d_{TVD}(Y, W) \leq 4(b_1 + b_2) = 4\left(n-2, 9p_p^2 + 3p_p p\right) \tag{5}$$

The TVD is a general distance measure between two pdfs. The TVD is defined here as the maximum absolute difference between the probability assigned by Y and the probability assigned by W to any specified subset of possible integer values. In the current scanning context, the most important subset of possible values to consider is the single value {0}, which would imply that the pattern {1 × 1} never occurred (occurred 0 times) in the $n-2$ overlapping trials. Then, the SC method, in this context, uses the Poisson approximation to assign a value to P {0} and this method ensures that the Poisson approximation to P{0} is quite accurate, as shown by the numerical example below.

According to the Poisson approximation, $P(\{1\,x\,1\} \text{ never occurs}) = e^{-\lambda}$. For example, using $n = 1000$ and $p = 0.01$, $\lambda = 998p_p = 0.0998$, then $e^{-\lambda}$= 0.905 is the approximate probability that the pattern never occurs, with an SC-based bound of $4(n-2)\left(9p_p^2 + 3p_p p\right) = 0.0123$. Therefore, the maximum difference between the

true probability defined by the Y random variable and the approximate probability assigned to any subset of the possible number of occurrences of $\{1 \times 1\}$ defined by the approximating W (Poisson random variable) is 0.01236. So, for example, if the probability that $(\{1 \times 1\}$ never occurs$) = e^{-\lambda} = 0.905$, then the true probability of 0 occurrences of the pattern is between 0.89 and 0.92. Section 2.3 uses simulation to confirm the quality of the SC approximation in Eq. (5) in this context.

Simulation can be used to closely approximate the true probabilities, but only for moderate values of n. For example, in 10^6 repeated sets of n = 1000 Bernoulli trials with $p = 0.01$, then $\lambda = 998p_p = 0.0998$, and $e^{-\lambda}$ = 0.905, the simulation-based P(0 occurrences of $\{1\,x\,1\}$) = 0.907. The Poisson-based approximation gives 0.905 with a SC-based TVD bound from Eq. (4) of 0.0123.

2.2 Summary

The SC method was described to approximate the pdf for the number of occurrences of an example pattern in an independent binary time series. In scanning for whether a pattern, such as $\{1 \times 1\}$, occurs starting at index i, there are overlapping tries to achieve the pattern, resulting in many non-independent trials consisting of the values in three successive indices. As the time series length increases and the probability $p = P(I_i = 1)$ decreases, the SC method shows that the Poisson approximation is excellent, with a small total variation distance bound.

The SC bound does not seem to be commonly used; however, related references are available [5–10]. For example, Ref. [7] applies the SC method to calculate coincidence probabilities. References [8, 9] apply the SC method in different time series contexts than considered here. To simplify the calculation of bivariate Poisson moments, Ref. [10] applies the SC identity $Xf(X) = \mu E(f(X+1))$, where X is a Poisson(μ) random variable, E denotes expected value, and $f()$ is any bounded function defined on the nonnegative integers. The SC identity was used to develop the SC approximation method used in Ref. [5] and the example above. Reference [5] showed that the SC bound defends the use of the Poisson approximation in real applications (as opposed to unwieldy combinatorial calculations for long time series), and provides a small bound on the approximation error. Simulation and/or analytical approximation are needed to estimate $p = P(I_i = 1)$ to apply the Poisson approximation and associated SC bound.

3. Example 2: monitoring for excessive numbers of tweets in 65 Florida counties

3.1 Introduction

Example 2 analyzes two types of daily tweet counts. The first type of counts is available from March 8, 2010, to December 31, 2015, in each of 65 Florida counties. In the available data from 65 counties (instead of the full 67), Lafayette is merged with Madison, and Liberty is merged with Gadsden; see Appendix 1. Such merging changes the spatial resolution available to detect spatial–temporal outbreaks for the four relevant counties. Reporting counts at the county (or merged county) level also changes the available spatial resolution, compared, for example, to geo-tagged counts. The second type of counts is time-tagged (to the nearest second) and geo-tagged (latitude and longitude) tweets, not aggregated to county-level counts.

One approach to monitor any numerically-valued time series has two main steps—(1) use training data to fit a model or models to the daily counts by county; and (2) monitor the corresponding residuals during training and testing to detect departures from the fitted model(s). For (1), effects such as day-of-week or time-of-year effects with multiple seasonal trends could be present, so the model building should be comprehensive, including assessment of simple exponentially weighted moving average (EWMA) models as well as models to fit trends and/or seasonal effects. For (2), the anomalies of interest could arise in neighboring counties (such as a reaction to a severe weather event such as a hurricane), and so could cluster in time and/or space.

3.2 Exploratory data analysis

Figure 7 is daily tweet counts from March 8 to December 31, 2010 (236 days) for (a) Brevard, (b) Broward, (c) Duval, and (d) Flagler counties (4 of 65 counties). This report addresses what types of models might fit these data, and options for monitoring residuals from the fitted model(s).

Question 1: Is the distribution of daily counts stable or stationary? Stationary means that, for example, the mean and variance of the counts is constant over time [1, 2]. Informally, it appears that spiking occurs in Broward county and that a mean shift occurs in Duval county. If the counts are not stationary, sometimes, for example, the first differences in counts are at least approximately stationary, perhaps with occasional spikes (**Figure 8**) [1, 2].

Question 2: Is the background ("training") data such as the 2010 counts adequately fit by a Poisson distribution [11–13]? **Figure 9** plots the variance/mean ratio (which

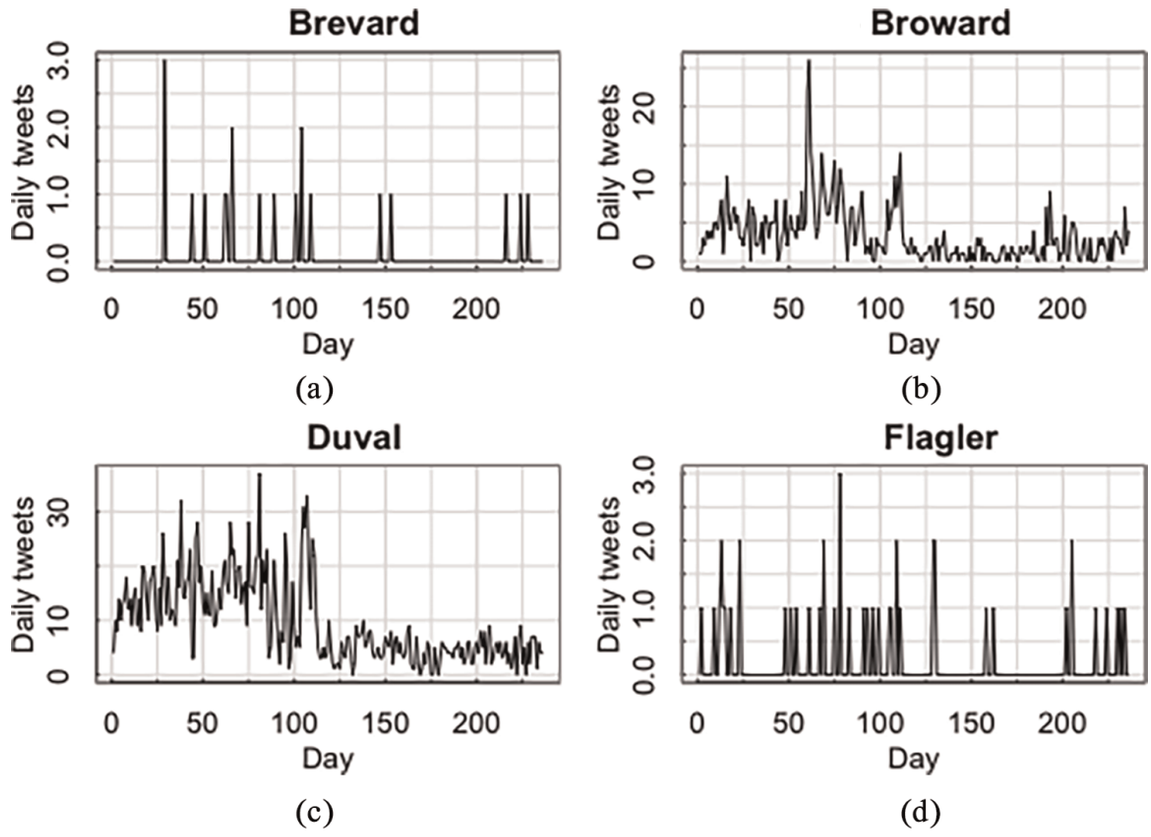

Figure 7.
The daily tweet counts for four of 65 Florida counties from march 8 to December 31, 2010.

should be approximately 1 for Poisson data) for each of the 65 counties. Hillsborough, Miami-Dade, and Orange counties appear to be outliers. In view of **Figure 9**, question 2 is academic here, because it is not expected that a simple constant-mean Poisson model will be adequate [11–13]. **Figure 9** suggests non-Poisson behavior.

Figure 10 plots the daily number of tweets for Broward county and simulated counts assuming a constant mean (equal to the mean of the Broward county counts) Poisson model. **Figure 10** also suggests non-Poisson behavior.

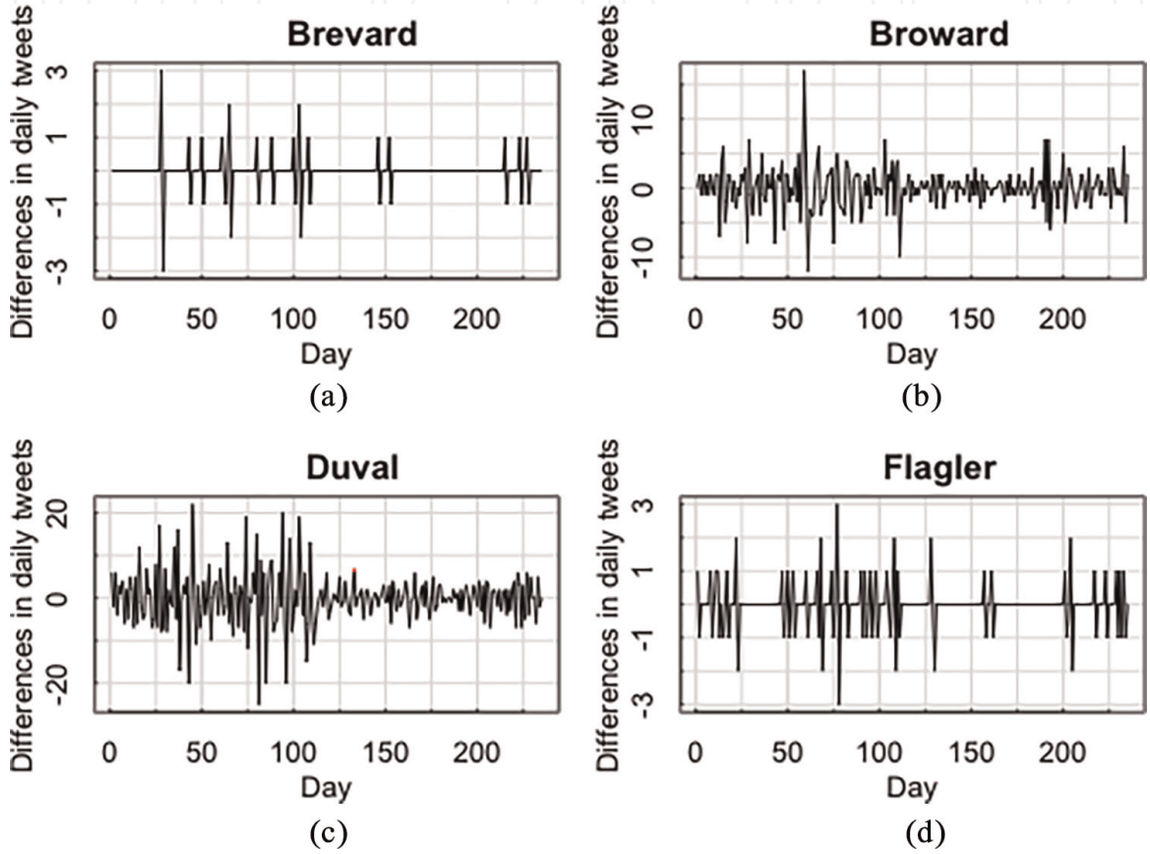

Figure 8.
The first differences in daily tweet count for four of 65 Florida counties from March 8 to December 31, 2010.

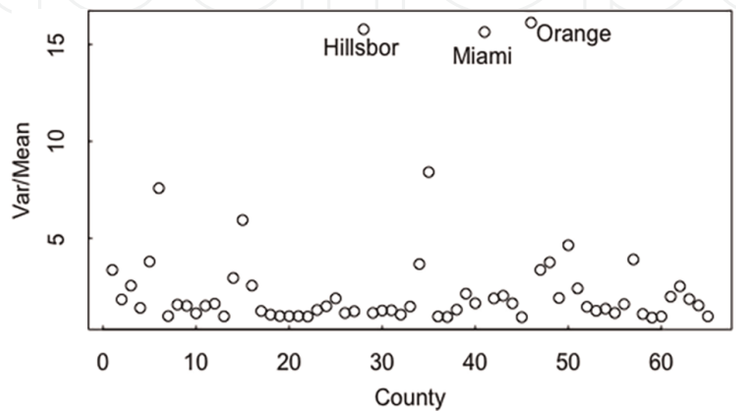

Figure 9.
The variance/mean for each of 65 Florida counties in 2010. The three largest variance/mean ratios are Hillsborough, Miami-Dade, and Orange counties.

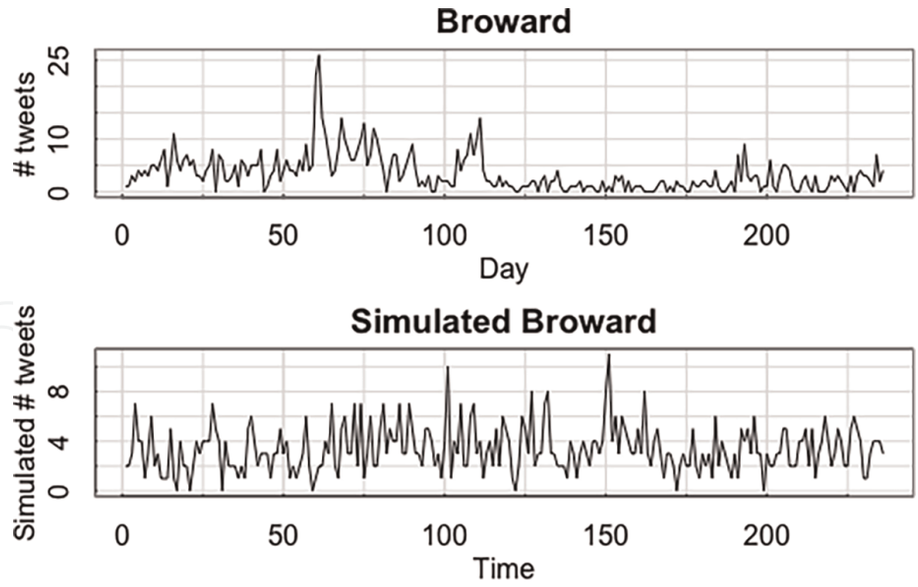

Figure 10.
The daily number of tweets in Broward county from March 8 to December 31, 2010, real (top), and simulated Poisson (bottom).

In **Figure 10**, the largest simulated count is 11 (the mean count is 3.4) but 7 of the 236 real counts exceed 11, and 2 of those exceed 20. More formally, the 99th percentile of the variance/mean ratio when sampling 236 observations from Poisson (3.4) is 1.23, but $\frac{s^2}{\bar{x}} = 3.4$ (**Figure 8**), where $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ is the sample mean and $s^2 = \frac{\sum_{i=1}^{236} (x_i - \bar{x})^2}{235}$ is the sample variance. All analyses are performed using R [4]. Together, **Figures 9** and **10** strongly suggest that a constant mean Poisson model is not adequate. Possibly, residuals around a fitted model could display approximate Poisson behavior, or perhaps another model such as the negative binomial for which variance/mean > 1 would be more appropriate [11].

Figure 11 is the total counts overall 65 Florida counties for each year in 2010–2015. **Figure 12** is the autocorrelation function $ACF_i = \rho_i = \frac{\sum_{j=1}^{n} (x_i - \bar{x})\left((x_{j-i} - \bar{x}\right)}{\sqrt{\sum_{j=1}^{n} (x_i - \bar{x})^2}}$ (with a minor adjustment to avoid negative indices [1]) using indices 1–117, prior to the large drop at index 118 for the total Florida counts in 2015. The ACF is useful for selecting possible models, such as those included in the class of ARMA models [1]. Modern machine-learning (ML) methods can also be evaluated, typically using AR modeling [1, 2, 12, 13]. An example of linear AR model is the lag one model $x_t = \mu + \rho x_{t-1} + e_t$, where μ is the long-term mean and the condition $|a| < 1$ ensures that the $\{x_t\}$ series is stationary. The noise term is denoted e_t. An example of a linear MA model is in Eq. (1); fitting an MA model requires estimating the noise sequence e_t [1, 2]. **Figure 13** is (a) simulated MA(1) data and (b) estimated residuals versus true residuals as was also shown in **Figures 4** and **5**.

3.3 Model selection and fitting

Exploratory data analysis (EDA) indicates no significant correlations in the residuals from the ARIMA(0,1,1) fit, so a simple MA(1) model could be competitive. The

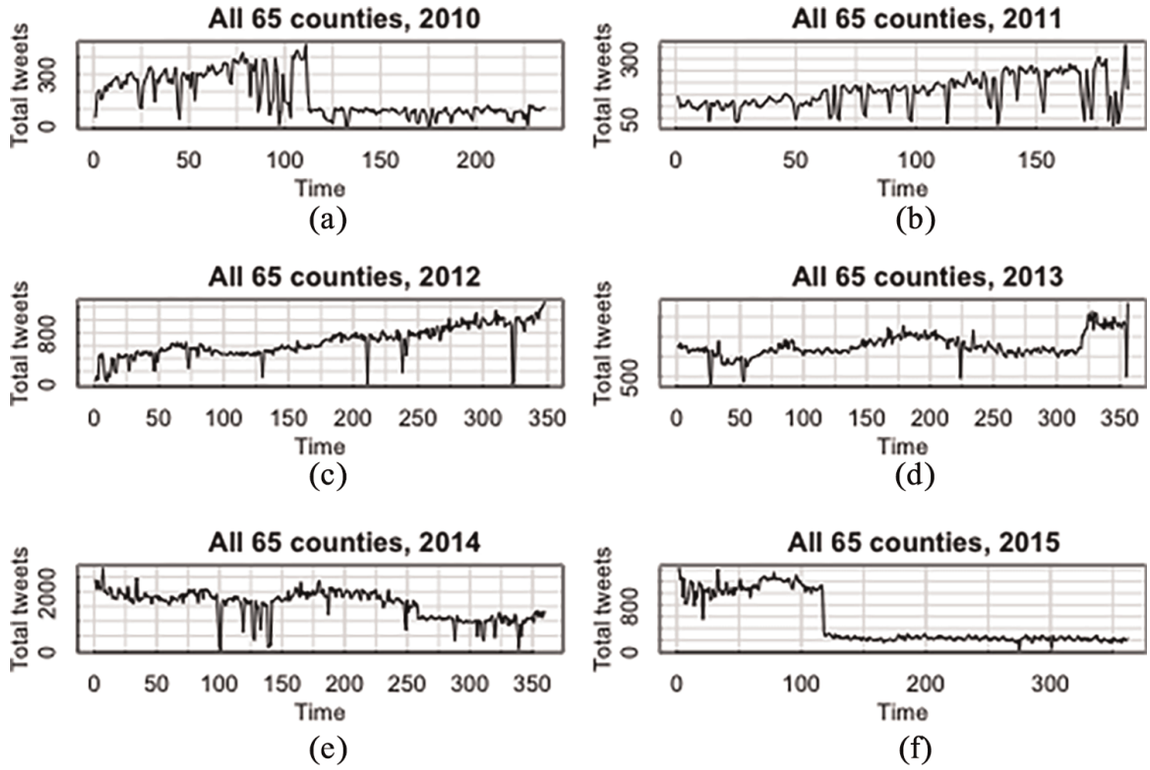

Figure 11.
The total counts overall 65 Florida counties for each year in 2010–2020,105.

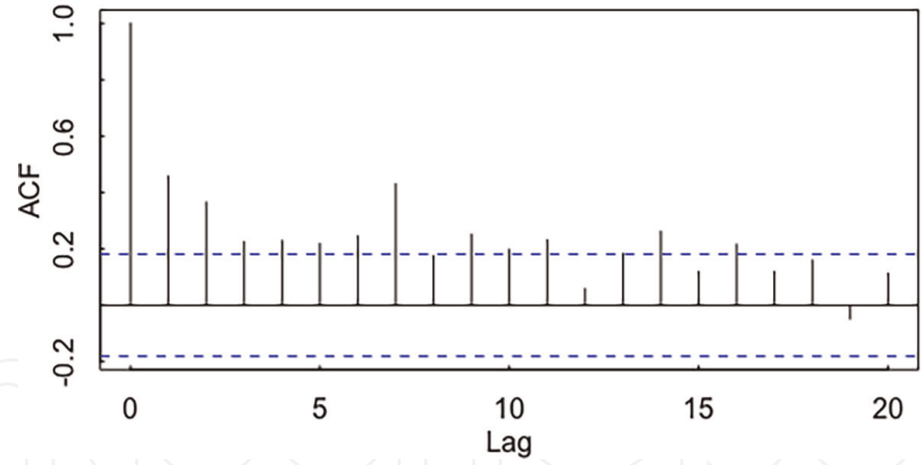

Figure 12.
The ACF for total Florida counts in 2015 over indices 1–117 before the large drop.

exponentially weighted moving average (EWMA) fit leads to the same fits as an MA (1) fit to the lag-one differences, which can be easily seen [1] as follows:

$$\hat{x}_t = (1-\lambda)\hat{x}_{t-1} + \lambda x_{t-1} = x_{t-1} + (1-\lambda)(x_{t-1} - \hat{x}_{t-1}) \tag{6}$$

$$x_t = (1-\lambda)e_{t-1} + e_t \tag{7}$$

$$\hat{x}_t = x_{t-1} - (1-\lambda)e_{t-1} \tag{8}$$

The forecast $\hat{x}_t$ in Eq. (6) from EWMA is the same as the forecast $\hat{x}_t$ in Eq. (7) from ARMA(0,1,1) which is an MA(1) after differencing [1, 2]. It should be pointed out that Poisson EWMA control charts as described in Ref. [11] are designed to monitor

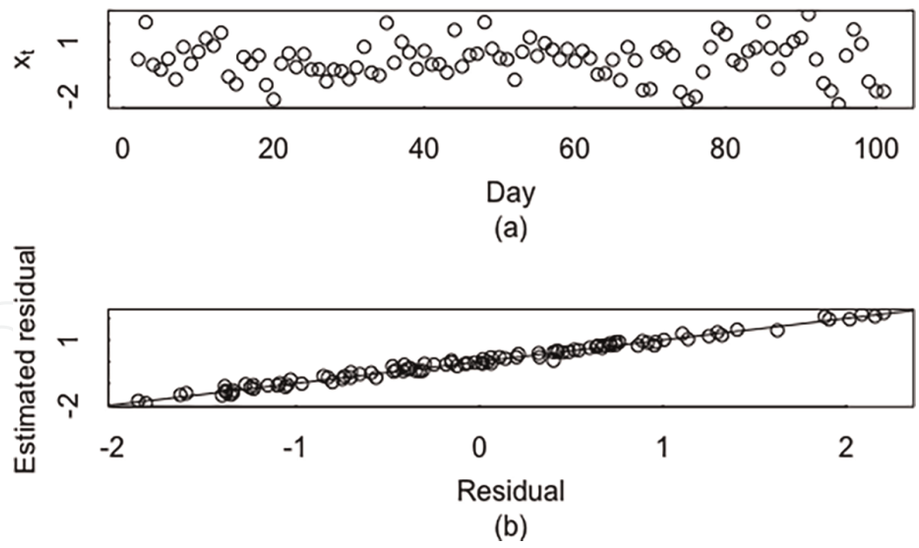

Figure 13.
(a) Simulated MA(1) data and (b) estimated residuals versus true residuals.

whether the Poisson mean μ_t is constant over time. In that case, $\hat{\mu}_t = (1-\lambda)\hat{\mu}_{t-1} + \lambda x_{t-1}$ is monitored rather than monitoring residuals from a local fit to the time-varying mean as in the twitter count monitoring.

The EWMA is among the simplest and most effective methods to forecast a time series. However, anomalies that persist for more than one day will impact the EWMA forecast in a manner that leads to reduced detection probability (DP) to detect the anomaly. For example, suppose an anomaly persists for 5 days. The EWMA forecast will be quite effective for day 1 of the anomaly, but the day 2 EWMA forecast will tend to increase due to the elevated expected count on day 1; this increased forecast leads to a reduced residual, thus reducing the DP.

Smooth fits using wavelets [13], EWMA [1, 2], and iterative bias reduction (IBR) [14–16] have been compared on this data. Edge effects are present in the wavelet and EWMA smoothers at the beginning of the data. The RMSE for wavelets, EWMA, and IBR are 4.05, 3.33, and 3.58, respectively, so EWMA is the simplest and has the smallest RMSE in this example. In data regions that have peaks, EWMA also performs acceptably well. The smoothing parameter λ in Eq. (6), $\hat{x}_t = (1-\lambda)\hat{x}_{t-1} + \lambda x_{t-1}$ can be chosen by using a grid search in Ref. [0,1] to minimize the RMSE in the training data. It is not unusual for EWMA to provide a competitive RMSE compared to other methods. As an example, **Figure 14** is the daily counts and the EWMA fit for Bay county counts.

Figure 15 is the residuals from the EWMA fit in **Figure 14**.

The standard deviation of the residuals in **Figure 15** is 3.43 while the standard deviation of the counts is 4.43. **Figure 16** is the ACF of the residuals in **Figure 15**.

3.4 Monitoring residuals

One effective option to monitor residuals from a fitted model is Page's statistic applied in this case to residuals in a single county, or to sums of residuals from neighborhoods of counties [17–20]. Reference [20] describes a scan statistic that spans temporally and spatially. Reference [21] illustrates that if the temporal and/or spatial spanning window is selected to provide maximum evidence of an anomaly, then the false alarm probability can be undesirably large unless the variable spanning is taken into account.

To monitor for positive mean shifts due to anomalously-large count(s), Page's statistic applied to the residuals $e_t = x_t - \hat{x}_t$ is defined in Eq. (4) ($S_t = \max(0, S_{t-1} + e_t - k)$). The parameter k is chosen to have good DP for a specific mean shift, but monitoring whether $S_t \geq h$ for threshold h can be effective for a range

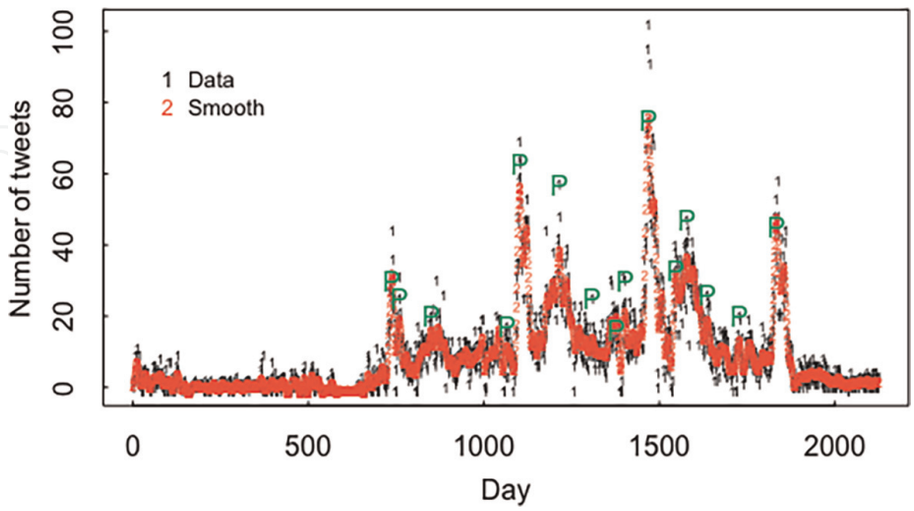

Figure 14.
Daily counts and a smooth fit for bay county. The green "P's" are local maxima in the smooth fit.

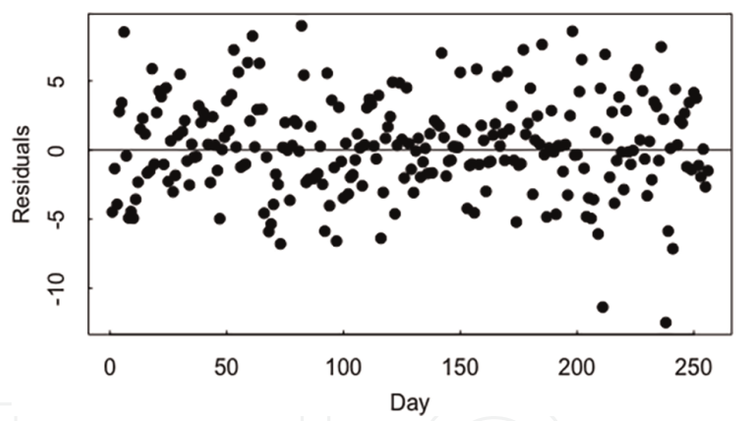

Figure 15.
Residuals $e = x - \hat{x}$ for the EWMA fit in ***Figure 14****.*

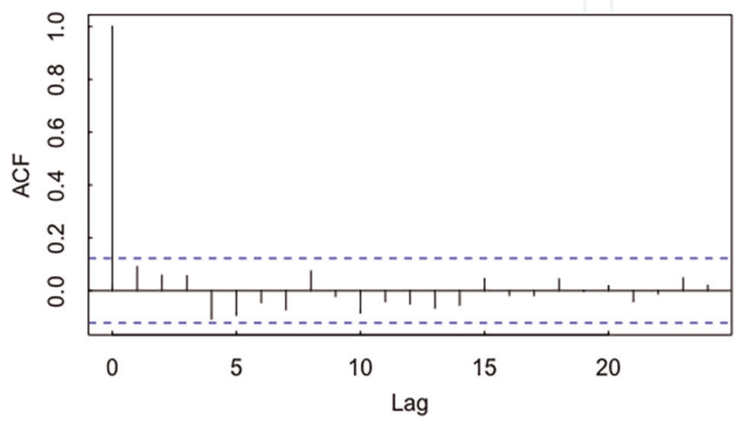

Figure 16.
ACF of the residuals in ***Figure 15****.*

of values of k. For the Poisson distribution, by using the ratio of the likelihood under the background mean μ_B to the shifted mean μ_S, the optimal value of k (leading to the largest DP) is given by $k = \frac{\mu_S - \mu_B}{\ln\left(\frac{\mu_S}{\mu_B}\right)}$. A statistical test based on the maximum of Page's statistic over an analysis window is equivalent to a statistical test based on a scan statistic defined as $\max_{t,j} \sum_{i=0}^{j} \{e_{t+i} - k\}$ over the analysis window.

Because the mean clearly changes over time in the background/training counts, the approach taken here is to fit a model and monitor the residuals e_t that are scaled to have variance 1, so $k = 0.5$ is chosen on the assumption that the residuals are symmetric around 0, and approximately normally distributed, so $k = 0.5$ is optimal (leads to the largest DP) for a mean shift of one standard deviation. However, $k = 0.5$ can lead to large DPs for other mean shift magnitudes. **Figure 17** plots Page's statistic applied to residuals from an EWMA fit for days 1–256 in 2012 in the first four counties (alphabetically: Alachua, Baker, Bay, and Bradford). **Figure 18** is similar to **Figure 17**, but Page's statistic is applied to the net residual in each county plus the residuals in all the nearest-neighbor of each respective county.

To select a threshold (see **Figure 19**) for monitoring all 65 residuals (one from each county) and all 65 residuals (one from each county including the neighboring counties), the 0.95 or 0.99 quantiles of the distribution of the maximum of the 65 or 130 Page's statistic values can be estimated (corresponding to a FAP or either 0.05 or 0.01 per analysis period (256 days in this example).

Figure 20 plots the DP for one-county-at-a-time monitoring and one-county-plus-nearest-neighbors monitoring. The 65 Florida counties are mapped in Appendix 1, defining the neighborhood scheme. For example, the five neighbors of Brevard county are Indian River, Orange, Osceola, Seminole, and Volusia. The nearest-neighbor county DPs are slightly smaller than the one-county-at-a-time DPs because the injected anomalies only impacted one county, and the alarm thresholds are slightly

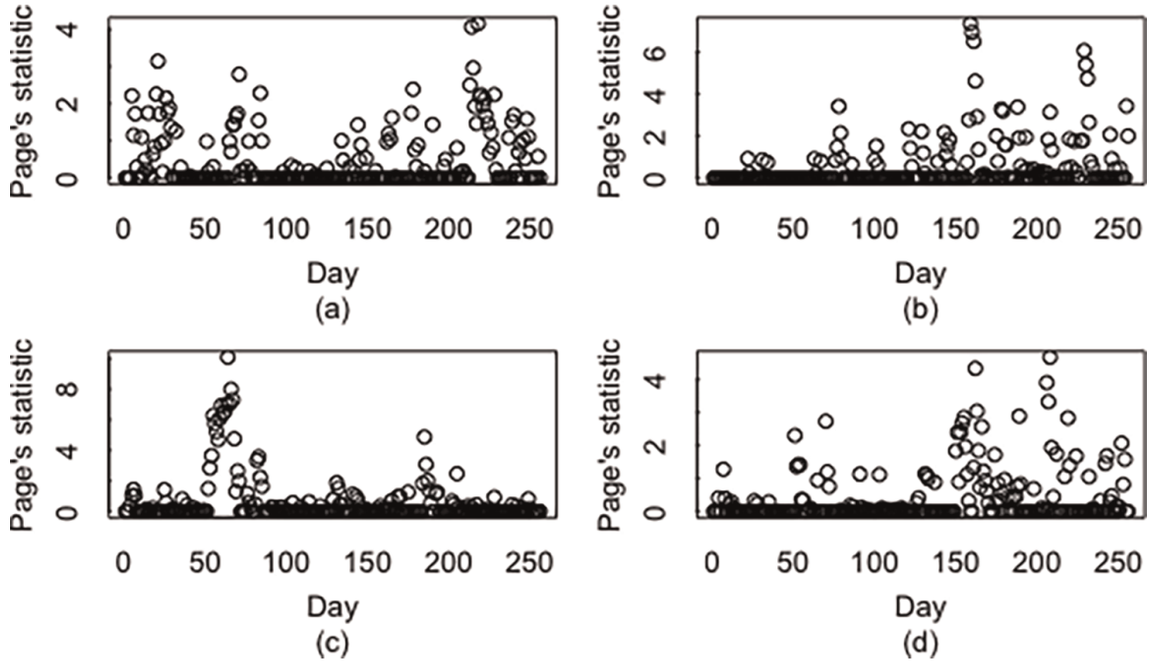

Figure 17.
Page's statistic versus day for Alachua, baker, bay, and Bradford.

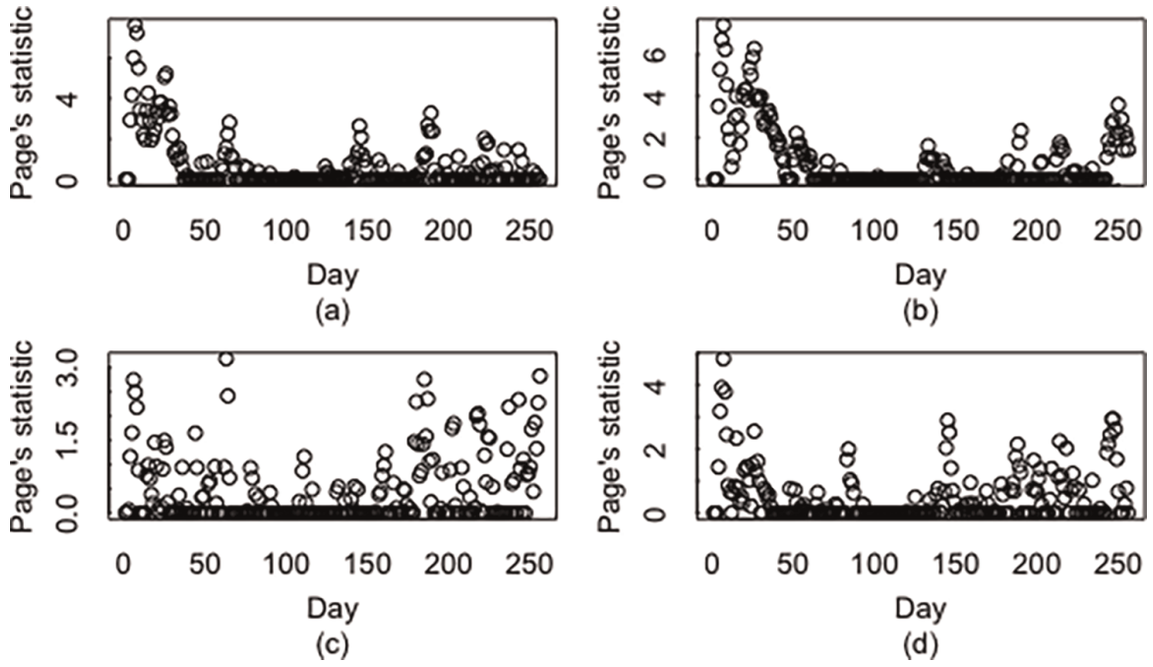

Figure 18.
Page's statistic versus day for Alachua, baker, bay, and Bradford and their respective nearest neighbors.

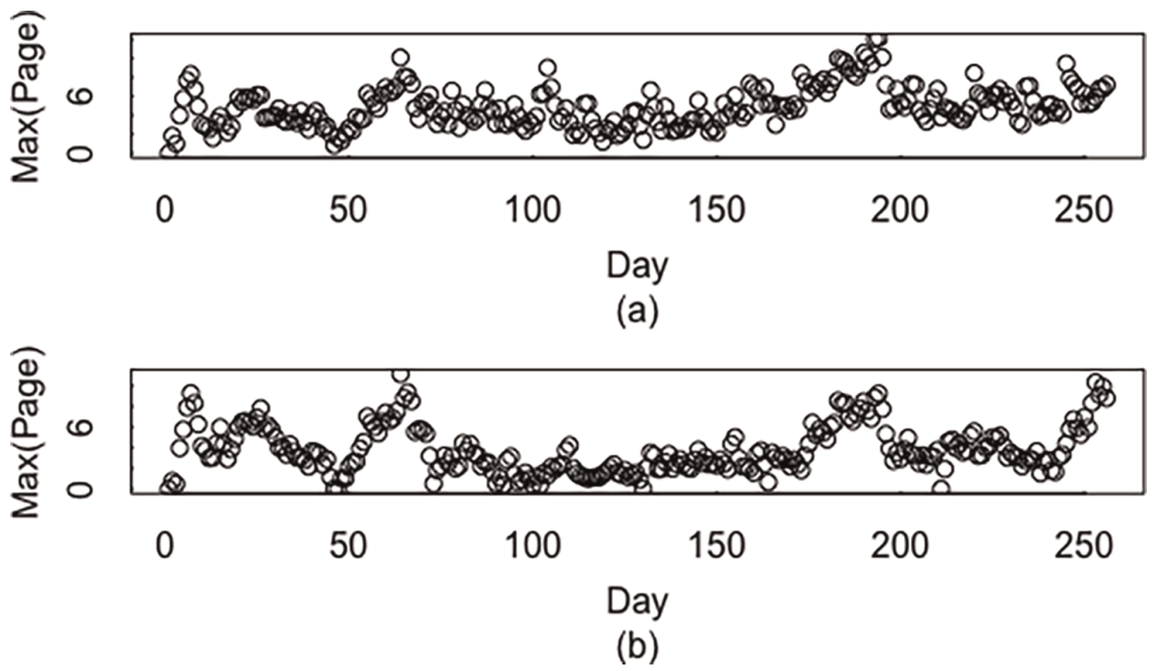

Figure 19.
The maximum value of Page's statistic overall 65 counties (a) and overall 65 counties with residuals from the respective county's neighbors (b).

larger when monitoring both individual counties and individual plus nearest-neighbor counties.

If the injected anomaly impacts all five neighbors of Brevard county then the DPs are much larger than those in **Figure 19**. For example, the DPS are 0.36, 0.78, and 0.99 for a mean shift of 1, 2, or 3 standard deviations in each of the five neighbors.

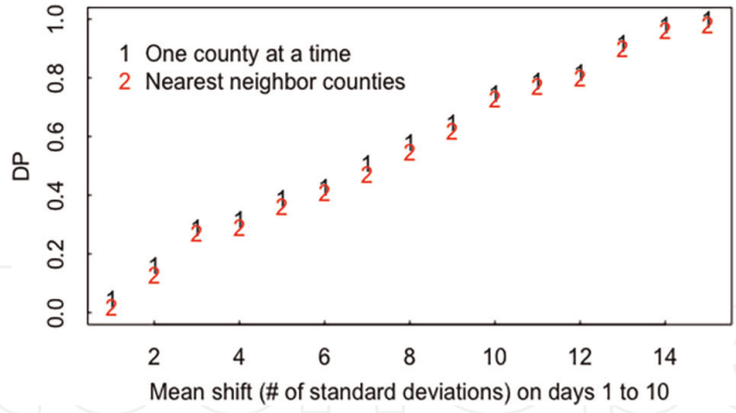

Figure 20.
The DP versus the mean shift occurring on days 1–10 in Brevard county. The nearest-neighbor monitoring includes both one-county-at-a-time and nearest neighbor monitoring.

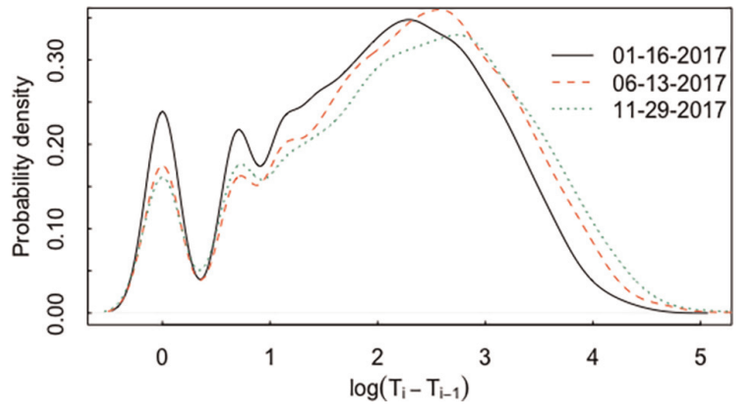

Figure 21.
The probability density of log($T_i - T_{i-1}$) for each of the 3 days.

Another Florida data set records each tweet to the nearest second by geo-location in latitude and longitude for each of 3 days during 2017 (January 16, June 13, and November 29). **Figure 21** plots the log of the lag-one time differences,

$\log(T_i - T_{i-1})$ for each of the 3 days (with the T values measured in seconds, but the time unit is ignored in applying the log function). If the inter-arrival times followed an exponential distribution with mean $1/\gamma$, then **Figure 20** would exhibit a single peak, and the number of counts in any given time interval of duration t would be distributed as Poisson($\mu = \gamma t$). Clearly, the counts do not exhibit exponential inter-arrival times, and the lag-one time differences between the tweets show considerable similarity across the 3 days. The 3 days have a total of 8102, 6376, and 5839 counts, respectively.

The custom R function generate.data() can be extended to include seasonality such as a single dominant peak per year if appropriate; however, typically more than one local peak is present per year. For example, find peaks applied to the IBR-based smooth finds an average of 4.9, 5.0,3.9, 4.8, 4.1, and 4.7 peaks (across all 65 counties) in years 2010, 2011, … , 2015, respectively. Simulated data from generate.data() can be used to assess candidate fitting options and to estimate DPs when synthetic anomaly effects are added. However, DP estimates tend to be too optimistic if a data generator such as generate.data() does not include enough realistic effects [18, 19].

Model fitting includes both model selection and estimating parameters in the selected models [13, 18, 19]. Recall that ARMA models are linear models but more generally, for example, AR models such as $x_t = f\left(x_{t-1}, x_{t-2}, \ldots, x_{t-p}\right) + e_t$ can be linear in the previous x values or not. Multivariate adaptive regression with splines (MARS) is a flexible nonlinear fitting option that was evaluated using the first 300 days to train and the next 300 days to gest for Brevard county counts. Although the RMSE from MARS fits was 2.48 in training (EWMA has an RMSE of 2.78 in training), the RMSE increased to 3.82 in testing (EWMA has an RMSE of 3.67 in testing), so EWMA remains competitive. Bayesian additive regression trees (BART) are another flexible option to fit AR models. The RMSE for BART (using gbart in R) was 5.07, which is larger than the standard deviation of the daily counts in the testing data (4.04).

Together with the time record of each tweet (to the nearest second), the latitude and longitude provide an option to monitor for spatial and/or temporal clustering. If there is no space–time clustering, then the event of being close in time is independent of the event of being close in time. It is, therefore, possible to check for independence by comparing the number of tweets that are close in space and time to the expected number assuming independence. Arbitrarily defining close in time to be the 0.1 quantiles of all the pairs of time gaps between tweets (and defining close in space to be the 0.1 quantiles of all the spatial distances between pairs of tweets), the 2-by-2 tallies for January 16 is in **Table 1**.

The χ^2test for independence strongly rejects the independence of space and time. Alternatively, the latitude and longitude values can be randomly reordered, breaking any true possible connection between space and time. The resulting test for independence is then not expected to be rejected; however, as an aside, while the random resorting of latitude and longitude reduced the 0.026–0.014, it turns out that 0.014 is large enough to 0.01 in this example, that some type of discretization phenomenon leads to this unexplained behavior. This same discretization phenomenon occurs for other arbitrary definitions of close, such as the 0.05 or 0.2 quantiles instead of the 0.1 quantiles.

Time-tagged Twitter counts with geo-locations are also available for Minnesota, Ohio, and Texas from January 01, 2016 to December 31, 2018 (1093 days with 3 missing days). Reference [19] provides plots of daily Twitter counts for Minnesota, Ohio, and Texas. Decreasing trends are obvious in all three states, and a t-test comparing the first 500 counts to the last 500 counts strongly rejects stationarity.

Reference [19] provides historgrams of the daily counts for Minnesota, Ohio, and Texas, respectively, along with simulated daily counts from a Poisson distribution having that respective state's mean count rate. The real data are much more dispersed than a corresponding Poisson distribution.

	Close in time (0.1 quantile)	**Not close in time**
Close in space (0.1 quantile)	856,752	23,761,044
Not close in space	2,893,098	5,306,257

Table 1.
*For the 8102 geo-located tweets on January 16, 2017, there are 8102*8101/2 = 32,817,151 comparisons (the sum of the four entries in Table) of time and space. The expected number in the "close in time and space" cell is 0.1 × 0.1 × 32,817,151 = 328171.51, while the observed counts are 856,752 and 856,752/32817151 = 0.026, which is statistically significantly larger than 0.01.*

3.5 Summary

Example 2 and Ref. [19] focused on monitoring residuals from simple EWMA fits. Other possible fitting options, such as MARS or BART for autoregressive modeling are described in Refs. [13, 18, 19]. If there were a more consistent seasonal peak, then model fitting could appropriately include seasonal peak fitting as in Ref. [22] for influenza forecasting. Another data source is Google's flu query data (counts of google searches that seek information about influenza symptoms) as a real-time option to monitor for flu outbreaks [23]. It would be valuable to investigate why the google flu data monitoring option has not been effective. Also, in monitoring for spatial–temporal clustering, it was assumed that "close in space" and "close in time" were defined arbitrarily at the 0.1 quantiles of their respective distributions. If instead several possible quantile values were examined for statistical significance, and the quantile choice leading to the highest evidence of clustering is used, then a simulation-based method [21] could adjust for such maximal selection of statistical evidence of clustering.

4. Example 3: monitoring for nuclear material loss

In nuclear material accounting (NMA), the material balance (MB) is defined as $MB = I_{begin} + T_{in} - T_{out} - I_{end}$, where T_{in} is transferred in; T_{out} is transferred out; I_{begin} is beginning inventory; and I_{end} is ending inventory [24–34]. All terms involve measured material, so the MB values should vary around 0 if there is no NM loss. The measurement error standard deviation of the MB is denoted σ_{MB}. Typically, many measurements are combined to estimate the terms T_{in}, I_{begin}, T_{out}, and I_{end} in the MB; therefore, the central limit effect and years of experience suggests that MBs in most facilities will be approximately normally distributed with a mean equal to the true NM loss μ and standard deviation σ_{MB}, which is expressed as $X \sim N(\mu,\sigma_{MB})$, where X denotes the MB. If the MB at a given time ("balance period") exceeds $k\ \sigma_{MB}$ with k in the 2–3 range, then the NMA system "alarms."

A sequence of n MBs is often assumed to have approximately a multivariate normal distribution $X = X_1, \ldots, X_n \sim N(\mu, \Sigma)$, where the n-by-n covariance matrix $\Sigma = \begin{pmatrix} \sigma_1^2 \sigma_{12}^2 \ldots \sigma_{1n}^2 \\ \sigma_{21}^2 \sigma_2^2 \ldots \sigma_{2n}^2 \\ \ldots \\ \sigma_{n1}^2 \sigma_{n2}^2 \ldots \sigma_n^2 \end{pmatrix}$. Estimating Σ is often one of the most tedious steps in frequent NMA (near real-time accounting, NRTA). A simplified example of estimating a component of Σ using a model of a generic electrochemical facility with one input stream, one output stream, and two inventory items is as follows. First, each individual measurement method is modeled with a measurement error model. A typical model for multiplicative errors is $M_{ij} = T_i(1 + S_i + R_{ij})$ with $S_i \sim N(0, \delta_S^2)$ and $R_{ij} \sim N(0, \delta_R^2)$, where the jth measurement results (often $j = 1$) M_{ij} of item i, T_i is the true value of item i, R_{ij} is a random error of item i, S_i is a short-term systematic error for item i. Then, the error variance for the two inventory items is $\sigma_I^2 = (T_1 + T_2)^2 \delta_S^2 + T_1^2 \delta_R^2 + T_2^2 \delta_R^2$, just as one example [26–29, 34] of error modeling and variance propagation used to estimate a component of Σ.

In the early 1980s, some believed that a plant reporting an MB every 30 days would have a larger detection probability (DP) than that same plant reporting an MB every year (typically a facility is inventoried and cleanout out approximately once per year). However, Ref. [27] then showed that for optimal (from the diverter's viewpoint, meaning that the DP is minimized) protracted diversion with the per-period loss being proportional to the row sums of the covariance matrix, Σ of the MB series, annual MB testing has larger DP than monthly mass balance testing. Reference [27] dampened hopes that frequent NMA, referred to as NRTA, would allow challenging diversion DP goals to be met. However, Ref. [27] conceded that NRTA has shorter detection times and higher DPs against abrupt diversion. Reference [27] showed that the best statistical test, the Neyman-Pearson (NP)-based matched filter for the worse case loss with $\mu_i^* = \frac{K\sum_{j=1}^{n}\Sigma_{i,j}}{\sum_{i=1}^{n}\sum_{j=1}^{b}\Sigma_{i,j}}$, is based on the cumsum, $C_t = \sum_{i=1}^{t} x_i$.

There are several reasons to apply statistical tests to a transformed sequence defined as $Y_i = \{X_i - E(X_i|X_{i-1}, X_{i-2}, \ldots, X_1)\}/\tilde{\sigma}_i$ where E denotes the expectation and the standard deviation $\tilde{\sigma}_i$ of $\{X_i - E(X_i|X_{i-1}, X_{i-2}, \ldots, X_1$ is $\tilde{\sigma}_i = \sqrt{\sigma_{ii}^2 - f\Sigma^{-1}f^T}$ where $f = \Sigma_{i,1(i-1)}$, the 1 to (i-1) entries in the ith row of Σ [2, 7]. From properties of the MVN, each component Y_i vector can be computed by calculating the conditional mean $E(X_i|X_{i-1}, X_{i-2}, \ldots, X_1) = f\Sigma_{i-1}^{-1}X_{i-1}$ where Σ_{i-1}^{-1} is the inverse of the (i-1)-by-(i-1) matrix Σ_{i-1} that corresponds to balance period 1 through period i-1. The Cholesky factorization $\Sigma = LU$ leads to a more computationally efficient recursive approach that avoids matrix inversion [28]. The transformed sequence $Y = L^{-1}$ X has $\Sigma_Y = I$, so Y is a residual time series.

This is a logistic advantage and a DP advantage to transform the $X_1, X_2, \ldots, X_n$ time series to the series of residuals $Y_1, Y_2, \ldots, Y_n$ known in NMA as the SITMUF (standardized, independently transformed material unaccounted for sequence, here MUF is another term for the MB). The logistic advantage is that the SITMUF time series is iid N(0,1) if the loss $\mu = 0$, so alarm thresholds depend only on the sequence length n and the desired FAP. The DP advantage is a DP increase for many loss vectors arises because the variance of the SITMUF sequence decreases over time, so particularly if a diversion occurs late in the analysis period, the DP is larger for the Y sequence than for the X sequence. Note that one cannot claim higher DP for the Y sequence than for the X sequence in general, because the true loss scenario is never known, and the DP can be larger for X than for Y for some loss scenarios. Modern NRTA systems use a suite of several statistical tests, usually applied to the Y series.

Example DPs are plotted in **Figures 22–24** (three different loss vectors for n = 16) for these five tests: (1) SITMUF test, (2) Page's test applied to the SITMUF series, (3) CUMSUM test, (4) a combination of (1–3), and (5) the NP-based matched filter is also useful to compute the largest possible DP for a specified loss. The alarm threshold h is chosen so that the FAP per analysis period (usually one year) is 0.05 or whatever FAP is specified. In **Figures 22–24**, the covariance matrix Σ from Ref. [34] has 1 on the diagonal, −0.48 on the lag-one off diagonals, and 0.01 on all higher-lag off diagonals. The loss 1 vector (**Figure 21**) is 0 on periods 1 to 5, constant on periods 6–10, then 0 on periods 11–16 (the nonzero entries summing to the quantity plotted on the horizontal axis). The loss 2 vector (**Figure 22**) is all 0 except for an abrupt loss on period 6. The loss 3 vector (**Figure 23**) is constant for all 16 periods.

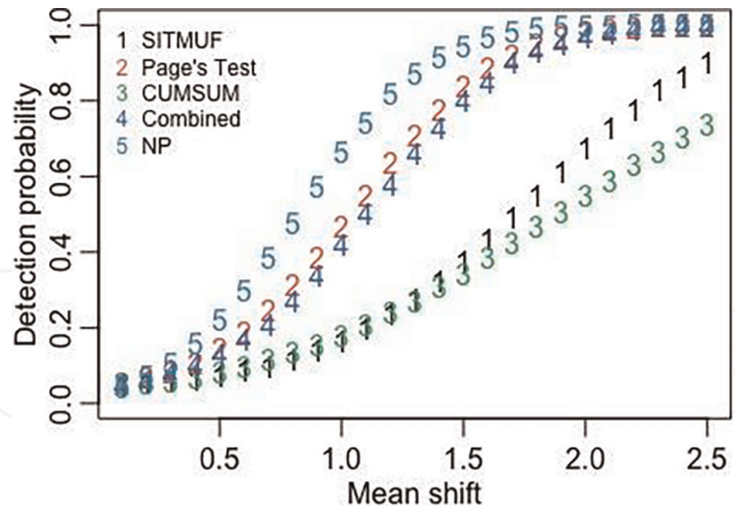

Figure 22.
DP versus the total mean shift (loss) for n = *16, for loss vector 1 (constant loss on periods 6 to 10).*

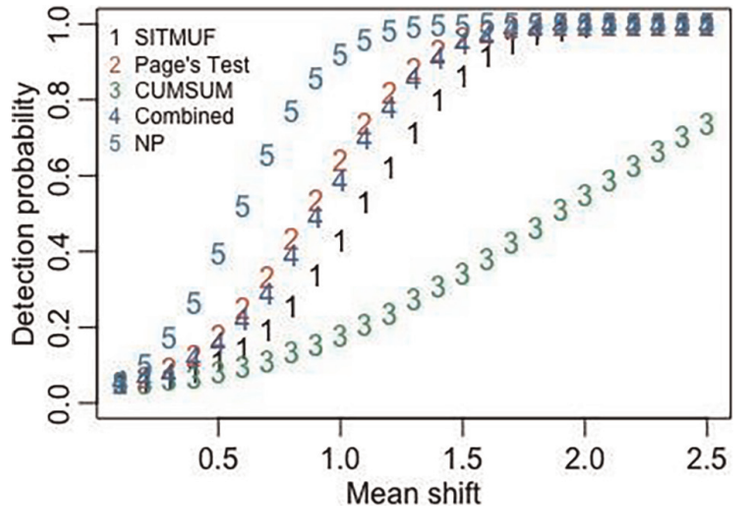

Figure 23.
DP versus the total mean shift (loss) for n = *16, for loss vector 2 (abrupt loss on period 6).*

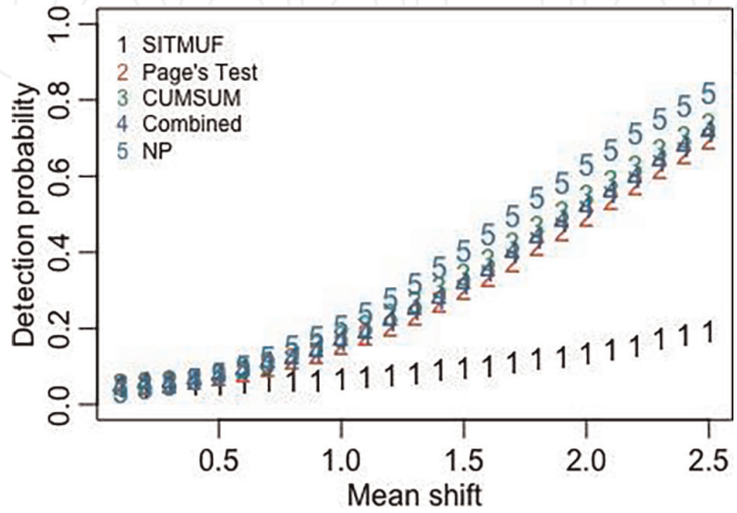

Figure 24.
DP versus the total mean shift (loss) for n = *16, for loss vector 3 (constant loss).*

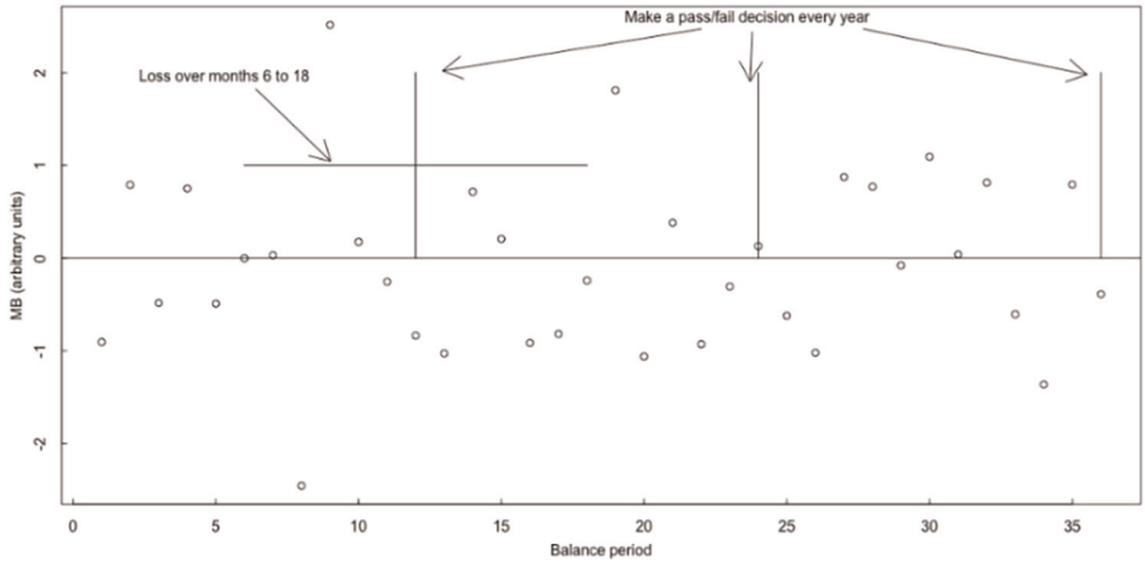

Figure 25.
MB sequences over 36 months using fixed-period (annual) decision periods.

This example is concluded with two remarks.

Remark 1. Reference [24] showed that assuming the model $X_1, \ldots, X_n \sim N(\mu, \Sigma)$ leads to larger DPs than fitting an ARMA model on training data for which it would have to be assumed that the NM loss $\mu = 0$ [31], provided Σ is well estimated. If Σ is not well estimated, a Bayesian updating scheme to improve the estimate of Σ could be used on training data for which the NM loss $\mu = 0$ [35].

Remark 2. **Figure 25** illustrates fixed-period testing and data-driven testing [29]. Some versions of NRTA use the most recent 1-year length sequences, so for $n = 12$ balance periods per year, the first evaluation period is months 1–12, the second evaluation period is months 2–13, etc. This scheme allows for a statistical decision to be made at every annual physical inventory, such as at months 12, 24, and 36. An alternate scheme consisting of a hybrid of period-driven and data-driven testing is described in Ref. [29], where it is pointed out that one should not simply truncate sequential statistical tests at the time of the annual physical inventory because the adversary could remove a portion of an SQ during year 1 and the remaining portion during year 2. The scan statistic has the highest DP (0.95 in this case, verified by simulation) if one knows that a loss will occur over a 12-month period with an unknown start period (such as month 7). The scan statistic computes a moving sum of months 1–12, 2–13, 3–14, etc.

5. Summary

This chapter has described three change-detection examples. Example 1 monitored for patterns of large residuals, such as a consecutive string of three residuals exceeding a threshold. Example 2 monitored for excessive numbers of tweets in any of the 65 Florida counties. Example 3 monitored for nuclear material loss. Portions of examples 1 and 3 have been published as cited in Refs. [5, 34]. Example 2 is entirely new. Page's statistic is generally recommended because of its reasonably large DP for a range of change patterns, such as any of those in **Figure 1**.

Appendix 1: The Florida counties in example 2

> ctylist
[1] “alachua” “baker” “bay” “bradford”
[5] “brevard” “broward” “calhoun” “charlotte”
[9] “citrus” “clay” “collier” “columbia”
[13] “desoto” “dixie” “duval” “escambia”
[17] “flagler” “franklin” “gadsden” “gilchrist”
[21] “glades” “gulf” “hamilton” “hardee”
[25] “hendry” “hernando” “highlands” “hillsborough”
[29] “holmes” “indian river” “jackson” “jefferson”
[33] “lake” “lee” “leon” “levy”
[37] “madison” “manatee” “marion” “martin”
[41] “miami-dade” “modroe” “nassau” “okaloosa”
[45] “okeechobee” “orange” “osceola” “palm beach”
[49] “pasco” “pinellas” “polk” “putnam”
[53] “st. johns” “st. lucie” “santa rosa” “sarasota”
[57] “seminole” “sumter” “sumannee” “taylor”
[61] “union” “volusia” “wakulla” “walton”
[65] “washington”.

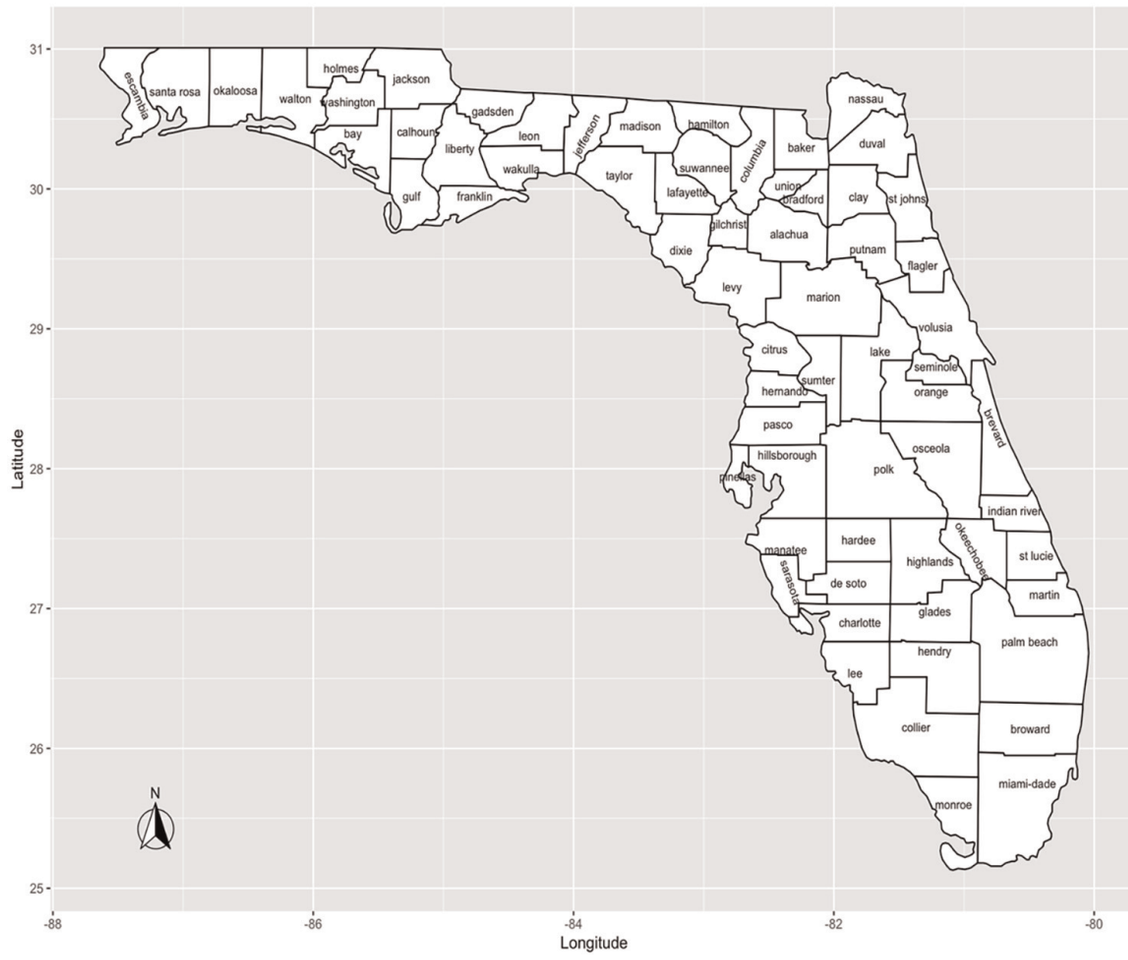

Figure A.1.
The 65 Florida countries. In the available data, liberty county is merged into Gadsden and Lafayatte county is merged into Madison to reduce the 67 Florida counties to 65. Therefore, there are 65 identified regions (counties) for which spatial and/or temporal residuals patterns can be monitored for change.

Figure A.1 is the 67 Florida counties, and this data merged Liberty into Gadsden and merged Lafayatte into Madison to reduce the 67 Florida counties to 65.

```
> ctynbrs[5,] # Brevard county: neighbors of Brevard county
[1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
[38] 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0.
ctylist[ctynbrs[5,]
> ctylist[as.logical(ctynbrs[5,])]
[1] “Indian river” “orange” “osceola” “seminole” “volusia”
```


Author details

Tom Burr* and Kim Kaufeld
Los Alamos National Laboratory, Los Alamos USA

*Address all correspondence to: tburr@lanl.gov

References

[1] Chatfield C. The Analysis of Time Series: An Introduction. 6th ed. London, United Kingdom: Chapman and Hall; 2004

[2] Shumway R, Stoffer D. Time Series Analysis and its Applications with R Examples. 4th ed. Pittsburgh: Springer; 2016

[3] Lucas J. Counted data Cusums. Technometrics. 1985;**27**(2):129-144

[4] R Core Team. R. A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. Available from: https://www.R-project.org/: R Foundation for Statistical Computing; 2017

[5] Burr T, Henderson B. Scanning for clusters of large values in time series: Application of the Stein-Chen method. Applied Mathematics. 2021;**12**:1031-1037

[6] Arratia R, Goldstein L, Gordon L. Poisson approximation and the Chen-Stein methods. Statistical Science. 1990; **5**(4):403-434

[7] Sahatsathatsana C. Applications of the Stein-Chen method for the problem of coincidences. International Journal of Pure and Applied Mathematics. 2017; **116**(1):49-59

[8] Kim S. A use of the Stein-Chen method in time series analysis. Journal of Applied Probability. 2000;**37**(4):1129-1136

[9] Aleksandrov B, Weis C, Jentsch C. Goodness-of-fit tests for Poisson count time series based of the Stein-Chen identity. Statistica Neerlandica. 2021;**76**: 35-64

[10] Weis C, Aleksandrov B. Computing bivariate Poisson moments using stein-Chen identities. The American Statistician. 2022;**76**(1):10-15

[11] Borror C, Champ E, Rigdon S. Poisson EWMA control charts. Journal of Quality Technology. 2018;**30**(4):352-361. DOI: 10.1080/00224065.1998.11979871

[12] Venables W, Ripley B. Modern Applied Statistics with S-Plus. New York: Springer; 1999

[13] Hastie T, Tibshiranii R, Friedman J. Elements of Statistical Learning. New York: Springer; 2001

[14] Burr T, Hengartner N, Matzner-Løber E, Myers S, Rouviere L. Smoothing low resolution gamma spectra. IEEE Transactions on Nuclear Science. 2010;**57**:2831-2840

[15] Cornillon P, Hengartner N, Jegou N, Matzner-Løber E. Iterative bias reduction: A comparative study. Statistics and Computing. 2013;**23**(6):777-791

[16] Hengartner N, Matzner-Lober E, Rouviere L, Burr T. Multiplicative Bias Corrected Nonparametric Smoothers with Application to Nuclear Energy Spectrum Estimation, Nonparametric Statistics. 3rd ISNPS ed. Avignon, France: Springer; 2016 arXiv Preprint arXiv:0908.0128

[17] Mathes R, Lall R, Levin-Rector A, Sell J, Paladini M, Konty K, et al. Evaluating and implementing temporal, spatial, and spatio-temporal methods for outbreak detection in a local syndromic surveillance system. PLoS ONE. 2017;**12** (9):e0184419. DOI: 10.1371/journal.pone.0184419

[18] Burr T, Graves T, Klaman R, Michalek S, Picard R, Hengartner N.

Accounting for seasonal patterns in syndromic surveillance data for outbreak detection, BioMedCentral. Medical Informatics and Decision Making. 2006; **6**:40

[19] Burr T, Kaufeld K. Statistical Evaluation of Daily Tweet Counts from Florida, Minnesota, Ohio, and Texas. New Mexico, United States: Los Alamos National Laboratory Report; 2021

[20] Kulldorff M. Prospective time periodic geographical disease surveillance using a scan statistic. Journal of the Royal Statistical Society. 2001; **A164**:61-72

[21] Burr T. Maximally selected measures of evidence of disease clustering. Statistics in Medicine. 2001;**20**: 1443-1460

[22] Osthus D, Moran K. Multiscale influenza forecasting, nature communications 12. Art. 2021;**2991**

[23] Pervais F, Pervaiz M, Rehman N, Saif U. FluBreaks early epidemic detection for google flu trends. Journal of Medical Internet Research. 2012;**14**(5): e125. DOI: 10/2196jmir.2002

[24] Burr T, Hamada MS. Smoothing and time series modeling of nuclear material accounting data for protracted diversion detection. Nuclear Science and Engineering. 2014;**177**:307-320

[25] Burr T, Hamada MS. Statistical Challenges in Integrated Nuclear Safeguards, Nuclear Science in the Series Energy Science and Technology. Vol. 4 (12). Vienna, Austria: IAEA; 2014

[26] Burr T, Hamada MS. Revisiting Statistical Aspects of Nuclear Material Accounting Science and Technology of Nuclear Installations. London, United Kingdom: Hindawi Publishing Corporation; 2013. pp. 1-15. DOI: 10.1155/2013/961360

[27] Avenhaus R, Jaech J. On subdividing material balances in time and/or space. Journal of Nuclear Materials Management. 1981;**10**:24-34

[28] Picard R. Sequential analysis of material balances. Journal of Nuclear Materials Management. 1987;**15**(2):38-42

[29] Burr T, Hamada MS, Ticknor L, Sprinkle J. Hybrid statistical testing for nuclear material accounting data and/or process monitoring data in nuclear safeguards. Energies. 2015;**8**:501-528

[30] Prasad S, Booth T, Hu M, Deligonul S. The detection of nuclear materials losses. Decision Sciences. 2007; **26**(2):265-281

[31] Speed T, Culpin D. The role of statistics in nuclear materials accounting: Issues and problems. Journal of the Royal Statistical Society A. 1986;**149**(4): 281-313

[32] Downing D, Pike D, Morrison G. Analysis of MUF data using ARlMA models. Journal of Nuclear Material Management. 1978;**7**(4):80-86

[33] Bonner E, Burr T, Krieger T, Martin K, Norman C. Comprehensive Uncertainty Quantification in Nuclear Safeguards, Science and Technology of Nuclear Installations. London, United Kingdom: Hindawi Publishing Corporation; 2017. pp. 1-16. DOI: 10.1155/2017/2679243

[34] Burr T, Hamada MS. Bayesian updating of material balances covariance matrices using training data. International Journal of Prognostics and Health Monitoring. 2014;**5**(1):006-013