An Adaptive Anomaly Detection Algorithm for Biological Outbreaks using Open-Source Media

Dongjin Kim*and James M. Wilson

Division of Integrated BioDefense
ISIS Center, Georgetown University
Washington, DC 20057

April 2, 2007

Abstract

We introduce an adaptive anomaly detection algorithm for early detection of socially disruptive biological outbreaks through monitoring of open-source media. We use indirect indicators covering various aspects of social disruptions to improve the timeliness of outbreak detection and use the exponentially weighted moving average and the exponentially weighted moving variance to build an adaptive baseline model. The algorithm is applied to detect the biological outbreaks of severe acute respiratory syndrome and dengue fever in India in 2003.

Keywords: anomaly detection; SARS; EWMA; EWMV

1 Introduction

As discussed in [1] biological outbreaks can result in substantial social disruptions such as avoidance of public places, hospital closings and hospital

*Corresponding author. Email: dkim@isis.georgetown.edu, Tel: 202-687-9087, Fax: 202-784-3479.

quarantines. These disruptions, in turn, are reported in open-source media even when the medical nature of that outbreak is uncertain. For example, our preliminary retrospective analysis of the biological outbreak of Severe Acute Respiratory Syndrome (SARS) in China [2] shows that a substantial collapse of the Chinese social and medical infrastructure was occurring in Chinese open-source media in November 2002, far earlier than March 2003 when World Health Organization (WHO) issued its global alert on SARS. Although WHO was aware of "a situation" in China earlier than March 2003, their approach to assessing an outbreak is focused on collecting direct diagnostic medical information, and they had no basis for estimating the nature or severity of the problem until such information was received. We suggest that monitoring of open-source media for indirect indicators covering various aspects of social disruptions can provide comparatively early detection of socially disruptive biological outbreaks.

With decreases in cost of computing power and storage space and with increases in the use of the internet around world, the use of open-source media has become an attractive means of biosurveillance (see [3, 4] for example). We monitor the daily occurrence rates of a set of pre-selected indicators through the search of keyword queries; for each indicator, in open-source media we count and record the number of articles that match the keyword query on a daily basis. Then, we compare the daily article counts to their normal behavior under the basic assumptions that as the situation in a region changes, the changing ground truths will be reflected in the open-source media produced by the region and as social disruptions become more significant, more reports in the open-source media are related to the social disruptions. An alert is issued when a significant deviation from normal behavior is detected.

In this article, we introduce an adaptive anomaly detection algorithm for early detection of anomalies that represent socially disruptive events including biological outbreaks which are of our interest. The purpose of the algorithm is to issue an "analyst alert" that is an automated recommendation for analysts to monitor a potential situation. Therefore, the main goal in this algorithm is how early in the evolution of an outbreak the analyst alert occurs with less false-alarm rate. We monitor a large number of equispaced time series simultaneously and build a baseline model to capture normal behavior of each individual indicator. We make the baseline models automatically adapt to keep themselves update as data are collected; so adaptive baseline models (see [5, 6, 7] for example). We detect anomalies when normal scores of observations at the current time significantly deviate from the baseline

models. Moreover, since each time series represent various aspects of social disruptions and behavior differently in time, we introduce a strategy to aggregate multiple time series to improve the timeliness of anomaly detection with less false-alarm rate.

The paper is arranged as follows: in the next section we describe an adaptive anomaly detection algorithm in detail. In Section 3, we discuss the results investigating the performance of the algorithm in terms of the timeliness and false alarms by tuning the parameters. Some concluding remarks are summarized at the end.

2 Algorithm Description

This section outlines in detail the adaptive anomaly detection algorithm for early detection of socially disruptive biological outbreaks through monitoring of open-source media.

The first step is to identify leading indirect indicators which can represent socially disruptive biological outbreaks in open-source media and to define keyword queries to search for instances of those indicators. We use our preliminary retrospective analysis of biological outbreaks such as SARS [2] and West Nile Virus [8] to identify leading indirect indicators for early detection of social disruptions especially due to biological outbreaks. We consider the following nine indirect indicators: 1. general population illness, 2. health worker illness, 3. avoidance of public places, 4. hospital inundated, 5. surveillance, 6. disinfection procedures, 7. hospital closings, 8. official investigations, and 9. official prosecutions. Figure 1 shows the keyword query to represent avoidance of public places in India.

Figure 1: Keyword query for avoidance of public places in India

(temple OR "farmers market" OR street OR mall OR square OR mosque OR festival OR "public places") AND (avoid OR avoided OR avoiding)

For each indicators, say $I_1, \cdots, I_9$, and a given time period, which in our case is a day, we record the number of articles per day that match corresponding keyword queries. We use our own search engine, the ISIS MiTAP system (see [9] for the detail) to examine selected open-source media for relevant articles and simply count and record the number of articles that match

the keyword queries on a daily basis. Due to variation in the number of articles, as shown in Figure 2, the daily frequencies of each indicators make little sense and hence we monitor the daily percentage of relevant articles, say $A_i(t)$, for each keyword query I_i, $i = 1, 2, \cdots, 9$. When the total number of articles at t is less than some small number (we use a threshold of 50), we conclude there is little information to get reliable article percentages and define $A_i(t) = 0$ for all $i = 1, 2, \cdots, 9$. We note that each $A_i(t)$ is a non-stationary discrete time series; its mean and variance change over time. To simplify the notation, index i is suppressed when we consider a time series individually for different indicators $I_1, \cdots, I_9$.

The article percentage $A(t)$ has inherent noise due to the error of the search engine, the delay of updating open-source media, and time discrepancy in multiple news sources. To reduce noise especially due to time discrepancy in open-source media we take a noise reduction process using last ω-day mean; let $X(t)$ be the time series after de-noising the daily article percentage $A(t)$, then

$$X(t) = \frac{1}{\omega} \sum_{i=1}^{\omega} A(t-i). \tag{1}$$

Figure 3 compares the de-noised time series based on $\omega = 2, 3, 5$,and 7 with raw time series for the indicator I_1 (general population illness). We note that such de-noised time series can keep the values high when there are sudden big changes or moderate but persistent changes. We will use last 7-day mean to treat common day-of-week effects in open-source media.

Next step is to build a baseline model that captures normal behavior of de-noised daily article percentage $X(t)$. We introduce an adaptive baseline distribution which can keep itself up-to-date as $X(t)$ is recorded everyday, without requiring access to past data. To initialize such an adaptive baseline distribution, we use daily percentages of article counts during a quiet period, say $[0, n]$, when no significant socially disruptive biological outbreaks occurred. And we construct the baseline distribution at $t = n$ under the assumption of Gaussian distribution which is fully characterized by the mean $E(n)$ and variance $V(n)$:

$$E(n) = \frac{1}{n} \sum_{t=1}^{n} X(t), \qquad V(n) = \frac{1}{n-1} \sum_{t=1}^{n} (X(t) - E(n))^2.$$

To verify the assumption of Gaussian distribution for the baseline model, we

show the normal probability plots of all the pre-selected indicators during the first two months, a quiet period, in Figure 4. Although the size of data is small, each indicator gives a good fit for Gaussian distribution.

Under the initialized baseline distribution with the mean $E(n)$ and variance $V(n)$, new incoming de-noised article percentage $X(n+1)$ is standardized by taking the normal score, $(X(n+1) - E(n))/\sqrt{V(n)}$ and used to update the baseline distribution by computing $E(n+1)$ and $V(n+1)$. We define a threshold, say Γ, for the boundary of the normal scores which can be allowed to be normal behavior. For example, $\Gamma = 2$, which represents 2 standard deviations away from the mean. We discuss in detail how to update $E(n+1)$ and $V(n+1)$ depending on the condition of $X(n+1)$.

If $|X(n+1) - E(n)|/\sqrt{V(n)} < \Gamma$, we use the Exponentially Weighted Moving Average (EWMA) and the Exponentially Weighted Moving Variance (EWMV) to update $E(n+1)$ and $V(n+1)$, respectively [10, 11]:

$$\begin{aligned} \text{If} \quad & |X(n+1) - E(n)|/\sqrt{V(n)} < \Gamma, \quad \text{then} \\ & E(n+1) = (1-\alpha)\, E(n) + \alpha\, X(n+1) \qquad (2) \\ & V(n+1) = (1-\beta)\, V(n) + \beta\, (X(n+1) - E(n+1))^2 \qquad (3) \end{aligned}$$

where α and β are weights ($0 < \alpha, \beta \leq 1$) that control the rate of exponential discounting of past data. The weighting for each day decreases exponentially, giving much more importance to recent data while still not discarding past data entirely. As shown in Figure 5, large values of α (β) measure short term average (variance), however, small values allow the EWMA (EWMV) to reach far back into the data and estimate longer term process average (variance). The EWMA responds to the changes in the mean; the EWMV changes in the variance. Thus, the Gaussian baseline distribution can keep being dynamic instead of static.

If $|X(n+1) - E(n)|/\sqrt{V(n)} \geq \Gamma$, then $X(n+1)$ is too extreme to be modeled normal behavior. Using this outlier may inflate the mean and variance. On the other hand, ignoring the outlier underestimates them. Thus outliers cannot be used as-is or ignored during updating as suggested in [5]. In our algorithm we replace the outlier $X(n+1)$ with $E(n) \pm \Gamma\sqrt{V(n)}$ and

update $E(n+1)$ and $V(n+1)$ using equations (2) and (3), respectively:

$$\begin{aligned}
\text{If} \quad & (X(n+1) - E(n))/\sqrt{V(n)} \geq \Gamma, \quad \text{then} \\
& E(n+1) = E(n) + \alpha\Gamma\sqrt{V(n)} \\
& V(n+1) = (1 - \beta + \beta(1-\alpha)^2\Gamma^2)\, V(n). \\
\text{If} \quad & (X(n+1) - E(n))/\sqrt{V(n)} \leq -\Gamma, \quad \text{then} \\
& E(n+1) = E(n) - \alpha\Gamma\sqrt{V(n)} \\
& V(n+1) = (1 - \beta + \beta(1-\alpha)^2\Gamma^2)\, V(n).
\end{aligned}$$

Especially when $(X(n+1) - E(n))/\sqrt{V(n)} \geq \Gamma$, we declare $X(n+1)$ as an anomaly.

Finally, we introduce how to aggregate multiple time series at each time point in order to improve the timeliness of anomaly detection. As mentioned in [3], different subsets of indicators show anomalies from one day to the next during an outbreak. The average normal score of all indicators can depreciate the timeliness of anomaly detection. Instead, we suggest the average normal score of no more than δ percent of leading indicators at each time. For example, $\delta = 50\%$, i.e. we use the average of four highest normal scores out of nine indicators at each time.

3 Test Results

We apply the algorithm introduced in the previous section to early detect socially disruptive biological outbreaks in India in 2003. In this period there were 19 times WHO biological outbreak alerts between April 17 and May 14 for SARS, and 2 times WHO alerts on October 30 and November 16 for dengue fever. Using the ISIS MiTAP system, we collect all on-line articles for the period January 1, 2003, to December 31, 2003, from the same English-language news sources in India as [3]: Indian Express, The Statesman India, Times of India, Hindu News, Telegraph India, and Tribune India. This test corpus comprises about 300,000 articles. Then we generate discrete time series by recording the daily occurrence rates of pre-selected nine indirect indicators introduced in the previous section through the search of keyword queries as shown in Figure 1. And we build an adaptive baseline model based on the first two months data because during the 59 days of January and February 2003, no major socially disruptive events occurred in India.

Our goal is to detect socially disruptive biological outbreaks early with less false-alarm rate. To accomplish our goal, we investigate the effect of changing the parameters of the algorithm on the timeliness of outbreak detection and false alarms. The tuning parameters of the algorithm are the de-noising factor ω for a noise reduction process using last ω-day mean, an anomaly threshold Γ, the EWMA weight α, the EWMV weight β, and a percentage parameter δ for aggregating multiple time series. In this numerical experiment, we fix two parameters out of five; we fix $\omega = 7$ in the noise reduction process to treat common day-of-week effects in news sources and $\delta = 50\%$, i.e. we use the average of four highest normal scores out of nine indicators in aggregating multiple time series. Then we will determine an optimal triple (Γ,α,β) that can improve the timeliness of outbreak detection and reduce false alarms.

Since the purpose of the algorithm is to issue an analyst alert as mentioned in Introduction, we compute the timeliness of detection as the difference between the first date when the algorithm generated an analyst alert and the reference date when the outbreak was actually detected by WHO. Remark that the timeliness defined in this paper is different from the general definition introduced in [12]: the first date is usually *after* the reference date, defined as the date the outbreak began. Furthermore, because the total number of non-outbreak days is not determined exactly, instead of a false-alarm rate, we measure false alarms by counting non-outbreak days on which the algorithm issues an analyst alert to mislead analysts.

We consider two anomaly thresholds: $\Gamma = 1.645$ and 2. The first threshold $\Gamma = 1.645$ is the upper 5% critical value from the standard normal distribution for the significance level of 0.05 and the other threshold $\Gamma = 2$ represents 2 standard deviations away from the mean. We note that we are focusing on the timeliness rather than less false alarms; if we increase the anomaly threshold as large as 3 standard deviations, which is commonly used in control chart methodology [10], we can reduce false alarms but lose a chance to issue an earlier analyst alert. Moreover, in order to improve the timeliness, we consider small values of the EWMA weight α and the EWMV weight β, which are useful for detecting small mean changes and small variance changes, respectively. Based on Figure 5, we study four pairs of two different values of $\alpha = 0.05$, 0.1 and $\beta = 0.001$, 0.01.

Table 1 presents the evaluation results for 8 different cases by combining two values of each component in the parameter triple (Γ,α,β). Here, Dengue 1 and 2 represent one of 2 times dengue fever, for which WHO alerts on

October 30 and November 16. These results are also plotted in Figure 6. From the result, we can easily see the tradeoff between the false alarms and timeliness; larger anomaly threshold ($\Gamma = 2$) cases reduce the false alarms from about 25 days to 8 days on average but decrease the timeliness from about 44 days to 34 days on average compared to the cases of $\Gamma = 1.645$.

Figure 7 shows four aggregated time series of two best cases for each thresholds. Case $(1.645, 0.05, 0.001)$, Figure 7 (A), shows consistent alarms during WHO alerts as well as earlier analyst alerts but more false alarms compared to others. If less false alarms are required, case $(2, 0.05, 0.001)$, Figure 7 (C), is a better choice since it can show more alarms reliably during WHO alerts than case $(2, 0.1, 0.01)$, Figure 7 (D). Thus, we suggest to keep $(\alpha, \beta) = (0.05, 0.001)$ and choose a different anomaly threshold either $\Gamma = 1.645$ or 2 depending on tradeoff between the false alarms and timeliness; $\Gamma = 1.645$ can guarantee the timeliness, however, $\Gamma = 2$ the less false alarms.

Table 1: Timeliness of Biological Outbreaks, SARS and 2 times Dengue Fever, in India in 2003

(Γ, α, β)	$(1.645, 0.05, 0.001)$	$(1.645, 0.05, 0.01)$	$(1.645, 0.1, 0.001)$	$(1.645, 0.1, 0.01)$
SARS	18	18	18	18
Dengue 1	14	15	14	15
Dengue 2	11	11	11	11
Timeliness	43	44	43	44
False Alarms	24	29	22	26
(Γ, α, β)	$(2, 0.05, 0.001)$	$(2, 0.05, 0.01)$	$(2, 0.1, 0.001)$	$(2, 0.1, 0.01)$
SARS	18	18	16	16
Dengue 1	11	13	13	13
Dengue 2	7	4	0	7
Timeliness	36	35	29	36
False Alarms	9	10	5	8

4 Concluding Remarks

We presented an adaptive anomaly detection algorithm through monitoring of open-source media and applied to early detect the biological outbreaks of SARS and dengue fever in India in 2003. We determined the optimal

parameters of the EWMA and EWMV weights as $\alpha = 0.05$ and $\beta = 0.001$, respectively. We also showed that the threshold $\Gamma = 2$ could reduce the false alarms but decrease the timeliness compared to smaller threshold $\Gamma = 1.645$. However, either case could improve the timeliness as many as 18 days for SARS and 10 days on average for dengue fever compared to the real WHO global alerts.

A baseline model in our approach is built under the Gaussian distribution. Although the Gaussian distribution is broadly applicable to describe the normal behavior of the article percentage in most indicators, it is not appropriate for all indicators. For example, as shown in Figure 8, the time series of the indirect indicator hospital quarantines can clearly reflect the abnormal behavior from both SARS and dengue fever outbreaks in India in 2003. However, as shown in Figure 9, its normal behavior is not normally distributed; most direct indicators behavior similarly. Therefore, we could not use such good indicators in the early detection of anomalies based on our algorithm. Thus, development of distribution-free detection algorithms can solve this problem and will be the subject of ongoing work.

5 Acknowledgements

The authors are thankful for the support of Dr. Paul E. Lehner, the Chief Engineer of the Information Technology Division of the MITRE Corporation. The authors are indebted to Seong Ki Mun, a director of the ISIS Center, Sergio Govoni, Guilan Huang, Peter Li, and Mark Polyak, for their encouragement and support of this research.

References

[1] J.M. Wilson, M.G. Polyak, J. Blake and J. Collmann; *A heuristic indication and warning staging model for analysis and interpretation of biological events* (2006)

[2] M.G. Polyak, J. Blake, J. Collmann and J.M. Wilson; *Emergence of Severe Acute Respiratory Syndrome (SARS) in the People's Republic of China, 2002-2003: a case study to define requirements for detection and assessment of international biological threats*

[3] P.E. Lehner and J.M. Wilson; *Automated detection of social disruption related to disease outbreaks: an empirical test*

[4] D. Agarwal, J. Feng and V. Torres; *Monitoring massive streams simultaneously: a holistic approach*, AT&T research report (2006)

[5] D. Lambert and C. Liu; *Adaptive thresholds: monitoring streams of network counts*, Journal of American Statistical Association Vol. 101, No. 473, (2006) 78-88

[6] V.A. Siris and F. Papagalou; *Application of anomaly detection algorithms for detecting SYN flooding attacks*, Computer Communications 29 (2006) 1433-1442

[7] N. Ye, C. Borror and Y. Zhang; *EWMA techniques for computer intrusion detection through anomalous changes in event intensity*, Qual. Reliab. Engng. Int. 18 (2002) 443-451

[8] J.M. Wilson, J. Blake, M. Turell and J. Davis-Cole; *The introduction of West Nile Virus to Washington, DC, 2002: a case study in using the Argus approach to characterize the evolution of bioevent*

[9] L. Damianos; *Description of upgraded MITAP* Journal of the American Medical Informatics Association (2006)

[10] D.C. Montgomery; *Introduction to statistical quality control*, John Wiley & Sons, New York (2001)

[11] J.F. MacGregor and T.J. Harris; *The exponentially weighted moving variance*, Journal of Quality Technology Vol. 25, No. 2, (1993) 106-118

[12] M.M. Wagner and G. Wallstrom; *Methods for algorithm evaluation*, in Wagner, et al., eds., Handbook of Biosurveillance, 2005

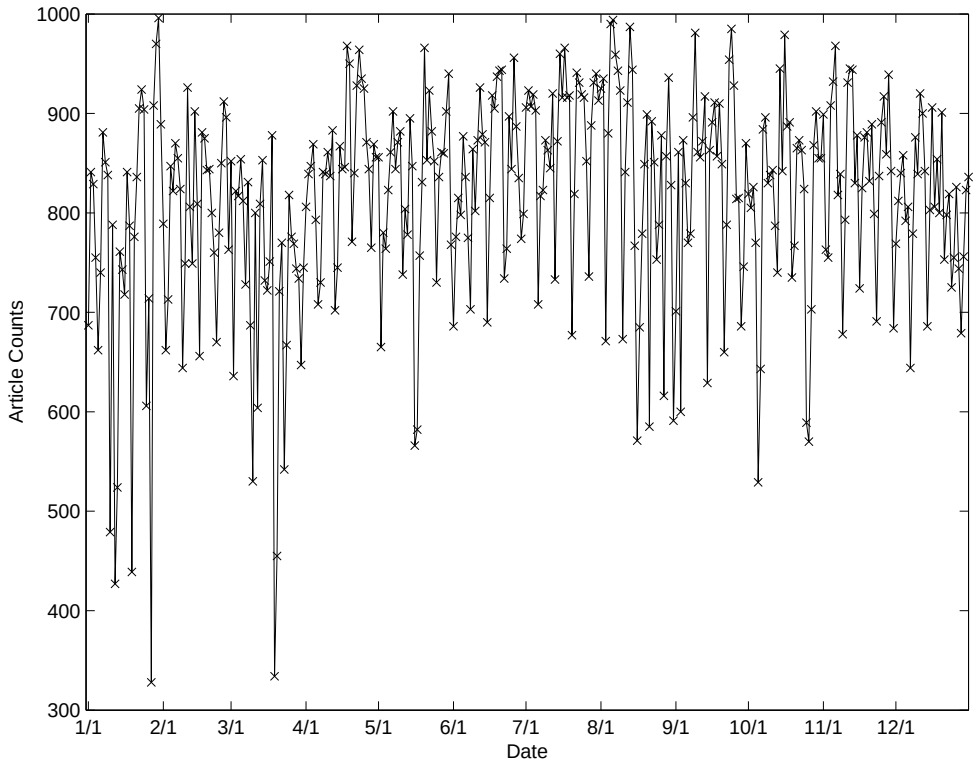

Figure 2: The variation in the total number of articles in Indian open-source media during 2003.

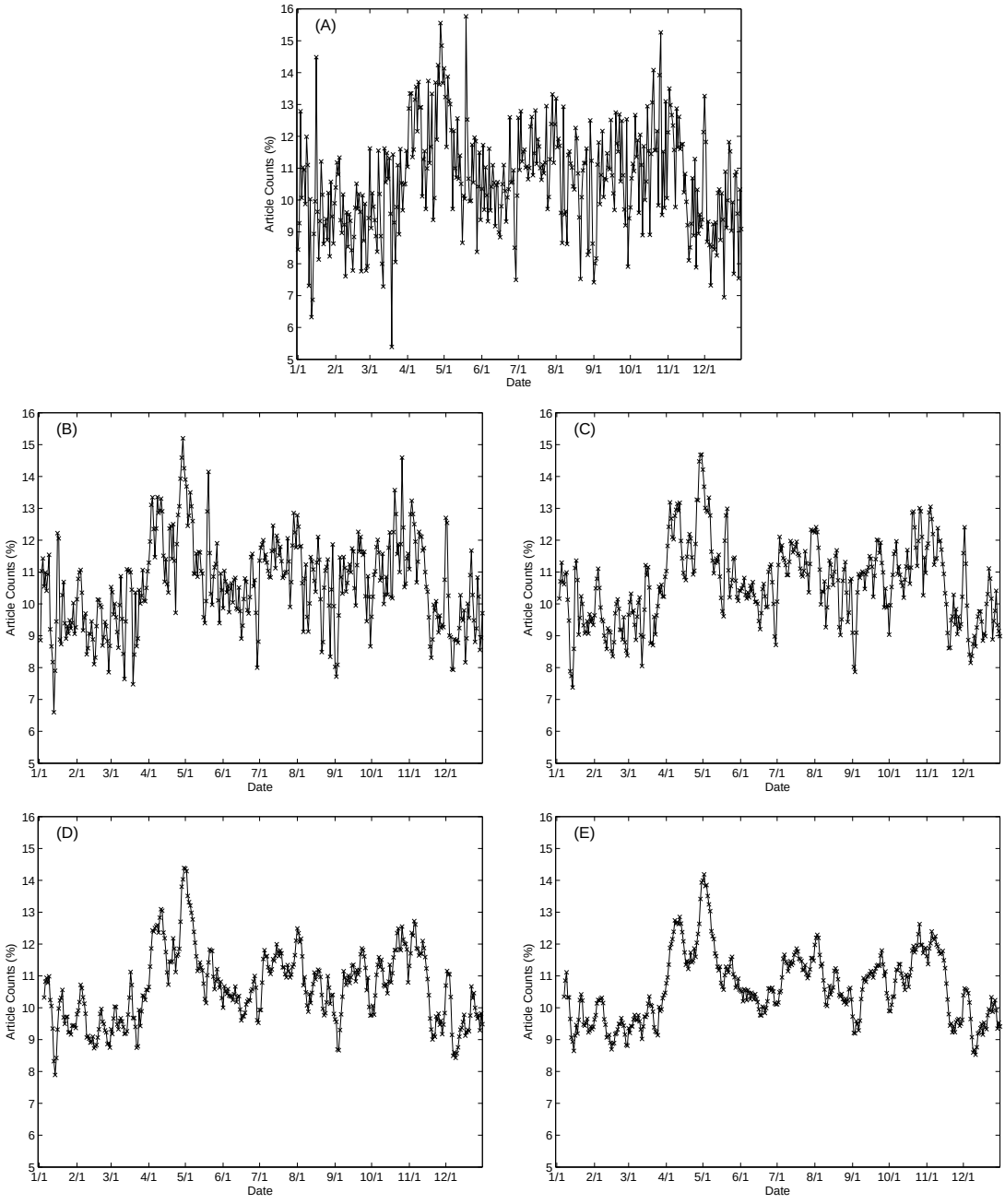

Figure 3: The daily percentage of relevant article counts to the indicator, general population illness.
(A) is raw time series and in turn the time series after the noise reduction using last ω-day mean with (B) $\omega = 2$ (C) $\omega = 3$ (D) $\omega = 5$ (E) $\omega = 7$.

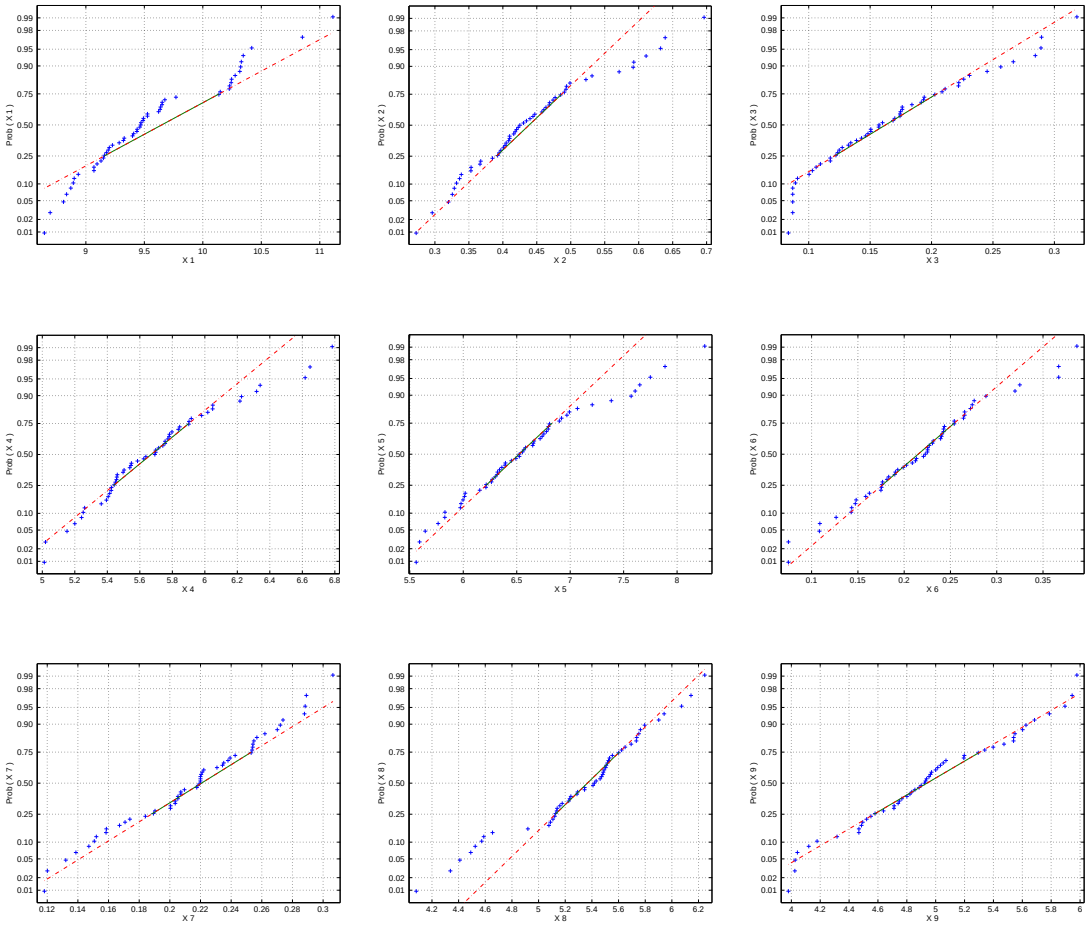

Figure 4: Normal probability plots of the de-noised daily article percentages X_i for nine indirect indicators I_i, $i = 1, 2, \cdots, 9$, during a quiet period, the first two months.

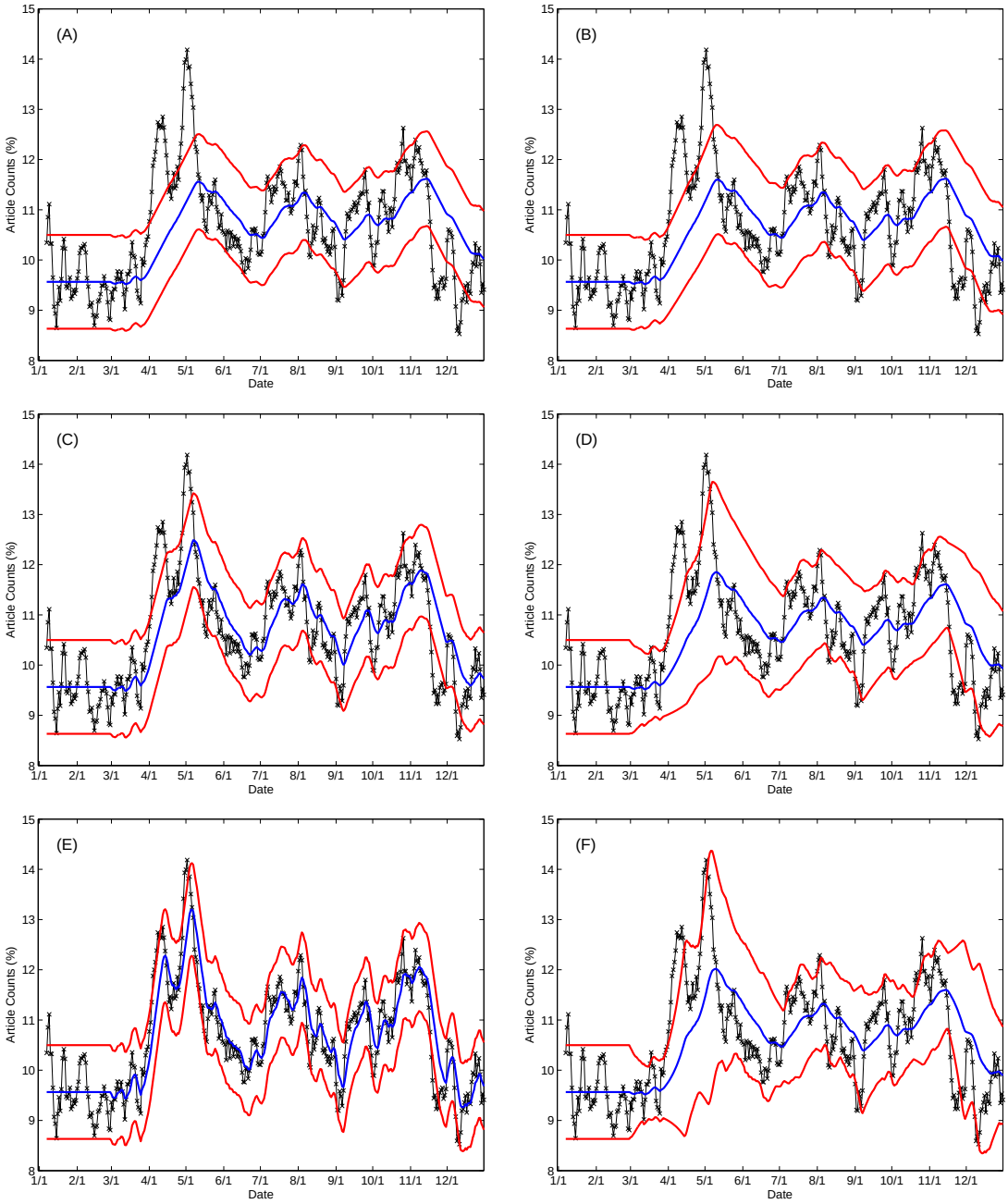

Figure 5: Effects of the EWMA weight α and the EWMV weight β. (A) $\alpha = 0.05$ (C) $\alpha = 0.1$ (E) $\alpha = 0.2$ when $\beta = 0.001$ in the left column and (B) $\beta = 0.01$ (D) $\beta = 0.05$ (F) $\beta = 0.1$ when $\alpha = 0.05$ in the right column for the indicator I_1. The blue center curve shows the evolution of the EWMA, and the upper and lower red curves show the boundaries to define outliers.

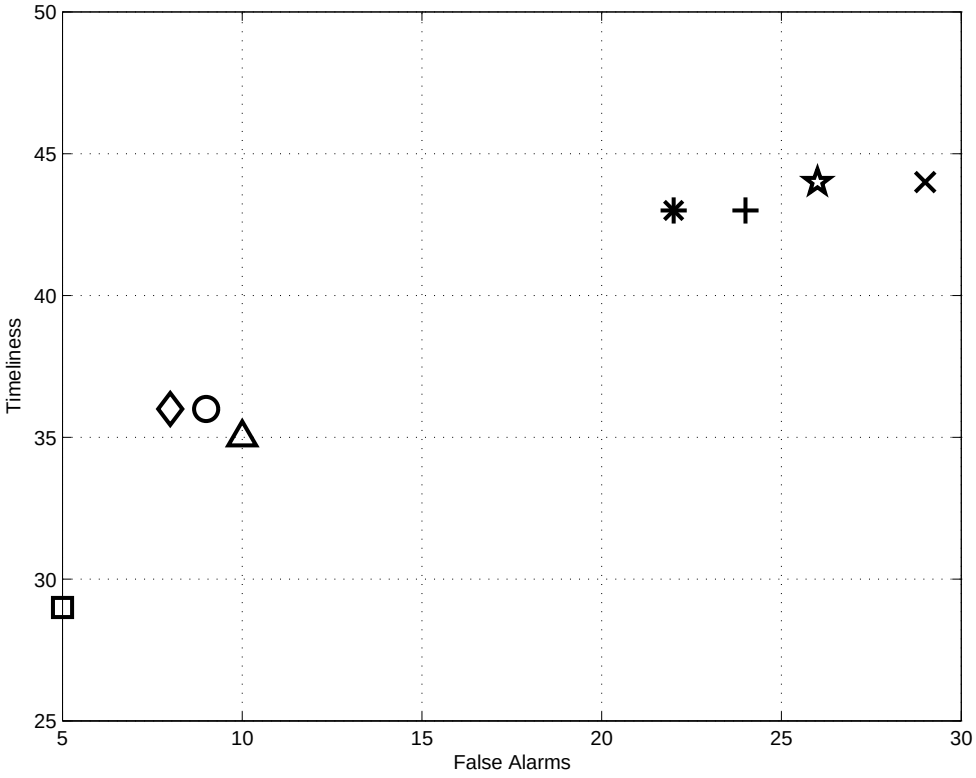

Figure 6: Timeliness versus false alarms for eight parameter triples (Γ,α,β)=(anomaly threshold, EWMA weight, EWMV weight).
$+ = (1.645, 0.05, 0.001)$, $\times = (1.645, 0.05, 0.01)$, $* = (1.645, 0.1, 0.001)$, $\star = (1.645, 0.1, 0.01)$, $\circ = (2, 0.05, 0.001)$, $\triangle = (2, 0.05, 0.01)$, $\square = (2, 0.1, 0.001)$, $\diamond = (2, 0.1, 0.01)$.

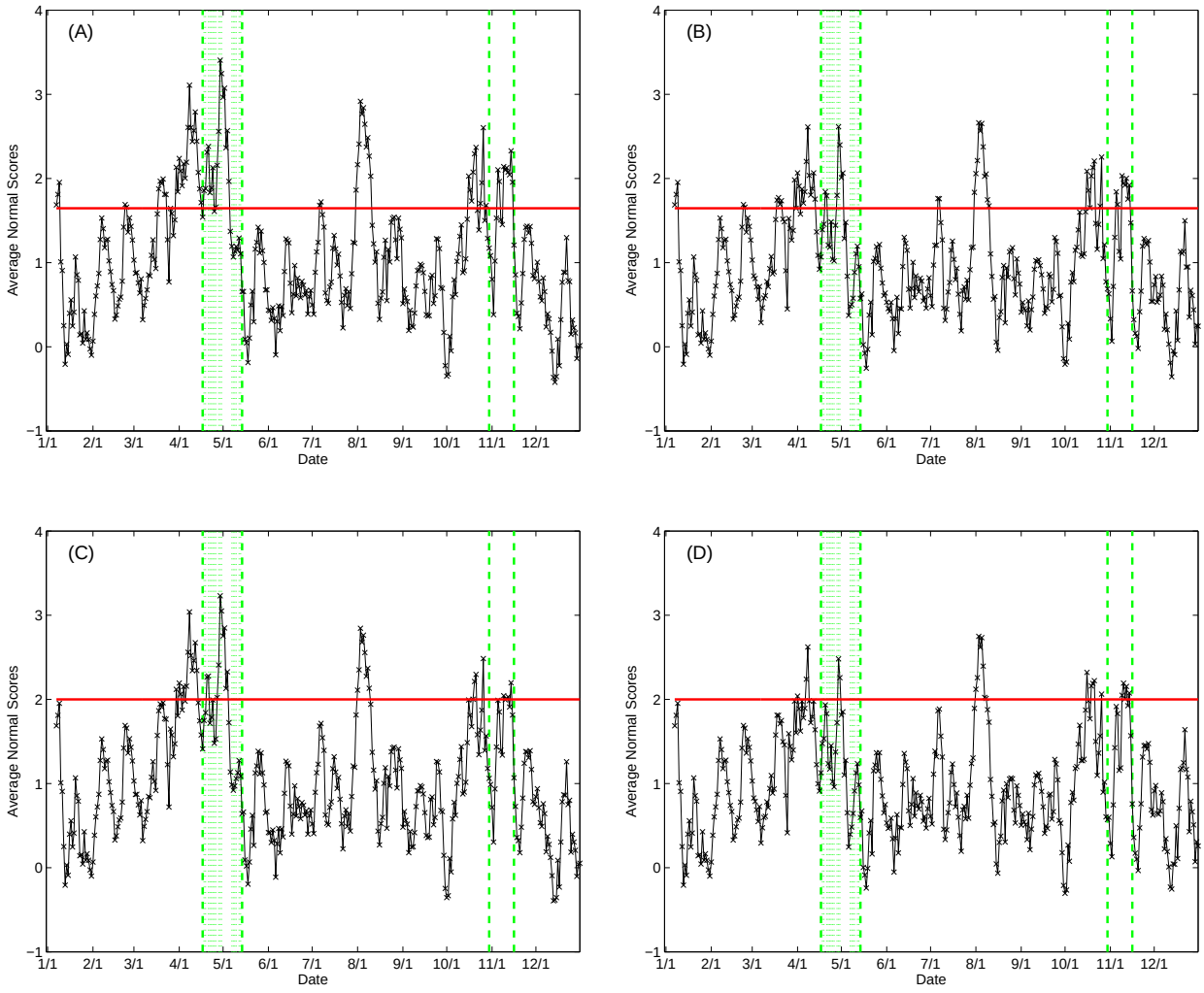

Figure 7: Time evolution of average normal scores for four parameter triples (Γ,α,β)=(anomaly threshold, EWMA weight, EWMV weight). (A)=(1.645,0.05,0.001), (B)=(1.645,0.1,0.001, (C)=(2,0.05,0.001) and (D)=(2,0.1,0.01). Each vertical red line represents the anomaly threshold. The horizontal green dotted lines show the real 19 times WHO biological outbreak alerts between April 17 and May 14 for SARS, and 2 times WHO alerts on October 30 and November 16 for dengue fever in India in 2003.

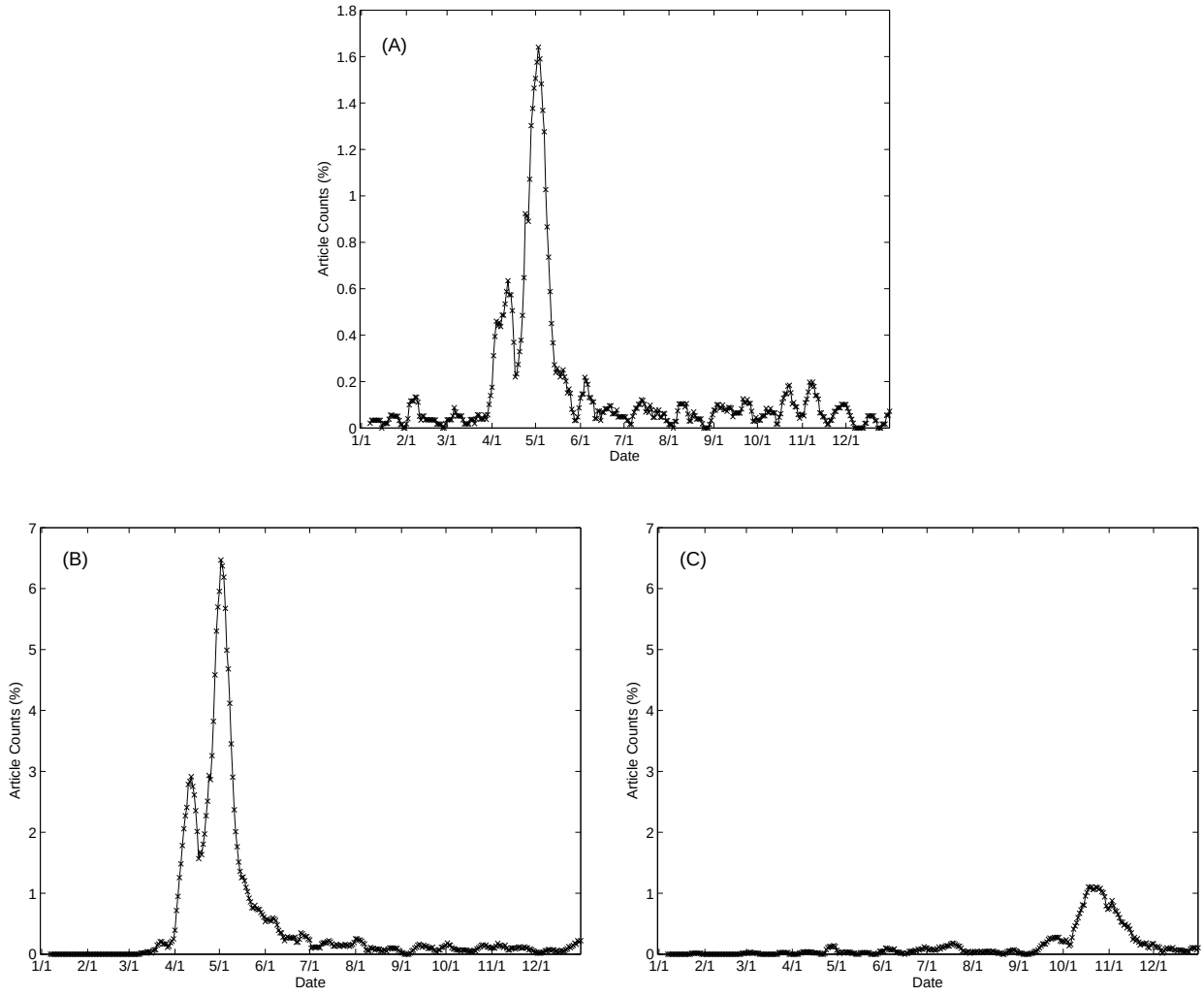

Figure 8: The de-noised daily article percentages for the indirect indicator (A) Hospital Quarantines and the direct indicators (B) SARS (C) Dengue Fever in India in 2003.

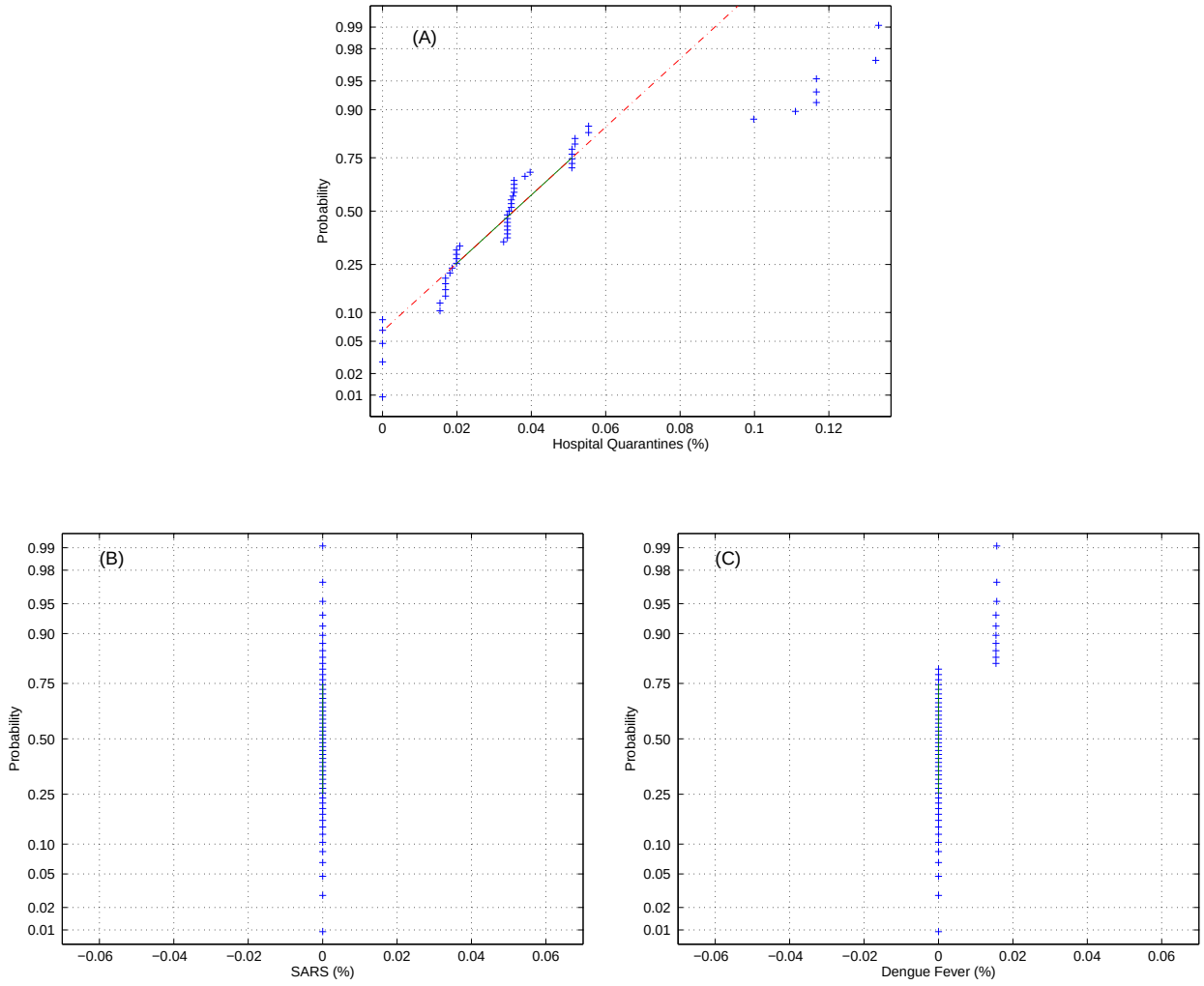

Figure 9: Normal probability plots of the de-noised daily article percentages for (A) Hospital Quarantines (B) SARS (C) Dengue Fever during a quiet period, the first two months.