A System for Simulation: Introducing Outbreaks into Time Series Data

Colin R. Goodall[1], Ph.D., Sylvia Halasz[1], Ph.D., John R. Allegra[2], M.D., Ph.D., Dennis Cochrane[2], M.D.

[1] *AT&T Labs – Research.* [2] *Emergency Medical Associates of New Jersey Research Foundation.*

Objective

Several authors have described ways to introduce artificial outbreaks into time series for the purpose of developing, testing, and evaluating the effectiveness and timeliness of anomaly detection algorithms, and more generally, early event detection systems. While the statistical anomaly detection methods take into account baseline characteristics of the time series, these simulated outbreaks are introduced on an ad hoc basis and do not take into account those baseline characteristics. Our objective was to develop statistical-based procedures to introduce artificial anomalies into time series, which thus would have wide applicability for evaluation of anomaly detection algorithms against widely different data streams.

Methods

Extending earlier approaches in the literature, we identify several key features of a system for introducing artificial outbreaks into time series. These are the shape of each outbreak, the size of each outbreak, and the spacing of the outbreaks. For each combination of features, we introduce multiple outbreaks into a given time series multiple times, with different offsets for the first artificial outbreak.

SHAPE OF OUTBREAKS

The *shape* of the artificial outbreak should follow one of several templates, or epicurves; these epicurves are based on epidemiological and infectious disease models for the progression of an outbreak through a population [Sartwell]. Figure 1 shows, in red, the shapes of epicurves with durations 14, 6, 4, 4, 1, and 4 days respectively from a set with 9 different durations.

SIZE OF OUTBREAKS

The *size* of the artificial outbreak, measured by its maximum count in any time interval (amplitude), should be calibrated automatically to the data sequences. Motivated by principles of robust and resistant data analysis, we calculated the size Y of the outbreak from the time series using a rule derived from the formula for the boxplot,

$$Y = \text{Round}[\ \alpha \times \text{IQR} + (\ Q3 - \text{Med}\ )/2 \times E\ ]$$

where E = epicurve normalized to have maximum height 1, Med is the median of the time series, and Q3 and IQR its upper (third) quartile and interquartile range respectively. α is a parameter controlled by the user to give artificial outbreaks of varying sizes.

SPACING OF OUTBREAKS

Multiple artificial outbreaks introduced into the same time series at random separated points can increase the efficiency of using time series data several fold compared to when just a single artificial outbreak is introduced into a time series. The *spacing* of the artificial outbreaks should not distort measurements in the testing and evaluation of anomaly detection algorithms derived from these simulations.

We provide a protocol to measure the effect of the interaction of the presence of an earlier outbreak on the detection of a later outbreak introduced into the data sequence. Additional outbreaks can be introduced when the interaction is measurably negligible. For example, we can measure the effect of changing spacing on sensitivity and specificity.

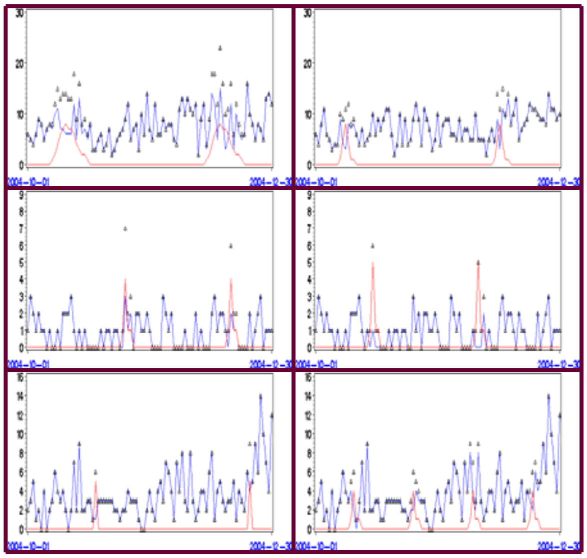

Figure 1: **Comparative Time Series of Emergency Department Visits with Artificial Outbreaks**

Results

Data sequences differ according to the typical count (0-10, 10-100, 100-1000, etc), their variability, their skewness, autocorrelation or spikiness, the proportion of zero values, and other measures of their distribution. A selection of 4 different time series from two hospital ED's and 3 syndromes is shown in Figure 1 (blue lines). The artificial outbreaks are shown alone (red lines) and added to the time series (black triangles). The size parameter is α = 1.5 and 2.0 (middle panels) and otherwise 1.75. These are all challenging choices. We added artificial outbreaks at spacing 57 days (top panels), 39 days (middle panels), and 57 and 22 days (lower panels).

Next we ran HWR (see companion poster), C3, and Cusum 7 anomaly detection algorithms on time series with artificial outbreaks with spacing 22, 39, 57, and 89 days. The variability in sensitivity with spacing is shown in Figure 2, and the variability of specificity with spacing is shown in Figure 3.

With the settings used here the moving average-based HWR has higher average sensitivity and specificity than C3 or Cusum 7, but HWR loses some of its specificity when the outbreaks were placed very close. This loss is strongest for the 1 day outbreak and 22 day spacing.

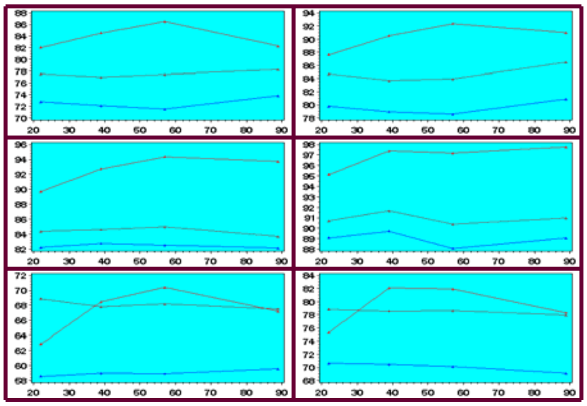

Figure 2: **Sensitivity as a function of spacing for HWR, C3, and Cusum 7; 4 (upper), 6 (middle) & 1 (lower) day; smaller (left) & larger (right) outbreaks**

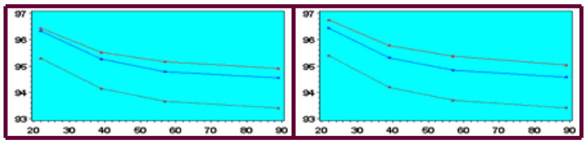

Figure 3: **Specificity as a function of spacing for HWR, C3, and Cusum 7; 4 day; smaller (left) & larger (right) outbreaks**

Conclusions

- We have demonstrated that, as part of a simulation system for syndromic surveillance, we can introduce artificial outbreaks into data sequences robustly and efficiently. We provide explicit formulae. This allows effective comparisons of algorithms and processes across multiple types of data sequences.
- We have developed a statistical framework for introducing artificial outbreaks, where the size is statistically related to the underlying time series.
- We are showing a method which is applicable uniformly, and also to data that is not normally distributed.
- We have shown that the specificity and sensitivity depend on the spacing of the outbreaks introduced and for the algorithms studied they stay close to constant as the spacing increases beyond a certain value (about 57 days).

Advances in Disease Surveillance 2007;2:199