

# A Study into Detection of Bio-Events in Multiple Streams of Surveillance Data

Josep Roure, Artur Dubrawski, and Jeff Schneider

The Auton Lab, Carnegie Mellon University, Pittsburgh, PA, USA

**Abstract.** This paper reviews the results of a study into combining evidence from multiple streams of surveillance data in order to improve timeliness and specificity of detection of bio-events. In the experiments we used three streams of real food- and agriculture-safety related data that is being routinely collected at slaughter houses across the nation, and which carry mutually complementary information about potential outbreaks of bio-events. The results indicate that: (1) Non-specific aggregation of p-values produced by event detectors set on individual streams of data can lead to superior detection power over that of the individual detectors, and (2) Design of multi-stream detectors tailored to the particular characteristics of the events of interest can further improve timeliness and specificity of detection. In a practical setup, we recommend combining a set of specific multi-stream detectors focused on individual types of predictable and definable scenarios of interest, with non-specific multi-stream detectors, to account for both anticipated and emerging types of bio-events.

## 1 Introduction

Maintaining the safety of agriculture and food supply is essential to the well-being of people and the economy. U.S. agriculture encompasses over \$1 trillion in economic activity, including more than \$50 billion in exports [4]. The U.S. food and agriculture systems are naturally vulnerable to disease, pest and contamination. That is due to several factors such as the relative ease of spreading communicable livestock and crop diseases, and simply the traditional methods of breeding and caring for livestock and growing crops.

In the experiments presented in this paper we use real food- and agriculture-safety data made available to us by the United States Department of Agriculture in the framework of an ongoing research project. One of the key objectives of that project is to provide USDA food-, animal- and plant-safety analysts with a surveillance tool capable of effectively monitoring multiple streams of heterogeneous data which is routinely collected by the department. Specific objec-

following parts of the paper), counts of positive and negative microbial tests of meat samples (set B), and counts of passed and failed sanitary inspections of slaughter houses (set C), conducted over a period of about 16 consecutive months in one of the Western U.S. states.

Multi-stream surveillance is attractive because it can lend improvements in sensitivity, specificity and timeliness of detection over more common univariate alternatives. Recently, researchers and practitioners in the field have turned their attention to exploiting the benefits of simultaneous tracking of multiple sources of complementary evidence [6,1,5]. The study presented in this paper focuses on evaluating the utility of approaches to multi-stream analysis in the context of a practical application.

## 2 Methodology

### 2.1 Temporal Scan

Each of the data streams we consider consists of two time series of counts. One represents the daily count of detects or positives (such as the animals discarded as being susceptible to a certain illness in the case of stream A), while the other represents the daily count of non-detects or negatives (such as healthy animals approved for slaughter in stream A). We apply statistical analysis to each of these streams in order to measure the extent of a possible departure of the counts observed on a given day from their expected normal levels. We then end up with a set of time series of daily p-values, computed independently for each of the individual streams of source data.

There are many ways of computing such p-values. We apply the method of temporal scan using the popular Chi-Square test and Fisher's exact tests of significance [7]. In temporal scan, a time window of interest,  $d$ -days wide, is moved along the time axis and at each discrete step (each day in our case), the numbers of detects and non-detects are aggregated across the  $d$  days inside the window and, separately, across all of the remaining days outside of the window. As our data exhibits a very strong day of the week effect, we modify the above procedure such that for the counts outside the window, we only consider the same days of the week as the ones inside the window of interest.

The resulting four counts form a 2-by-2 contingency table for the Chi-Square or Fisher's test procedure. We use Fisher's test if any of the component counts is less than 10 and if the sum of four counts is less than 100, otherwise we apply Chi-Square. The more the observed counts of detects and their proportion to non-detects inside the time window differ from the expectation based on the counts aggregated outside, the lower the p-value obtained for this stream of data on that day.

A simple uni-variate surveillance system would then monitor the individual streams of p-values and trigger an alert if one or more of them went below a pre-set threshold, say  $\alpha = 0.05$ . It would likely produce many false positives and yet still have only modest power because each of the streams is considered in isolation from the others.

## 2.2 Multi-stream Non-specific Detector

We next address the question of how, for a given time period, the evidence from each of the individual data streams can be combined into a single detector with more power than any of the individuals. In instances where we have a strong model of the relationships between the streams we might work directly on the raw data [2]. Often no such information is available and that is the case of the data we consider in our empirical results below. Therefore, we adopt a more general approach based on combining the p-values into a single detector.

At first glance, one might be tempted to apply a Bonferroni correction and signal an alarm if the smallest p-value passes the corrected test. This corresponds, however, to an aggregation method based on the *Min* function. The “correction” part of the method actually has no effect on AMOC curves (described in the empirical results section) since the threshold for signaling an alarm is varied to produce the curve.

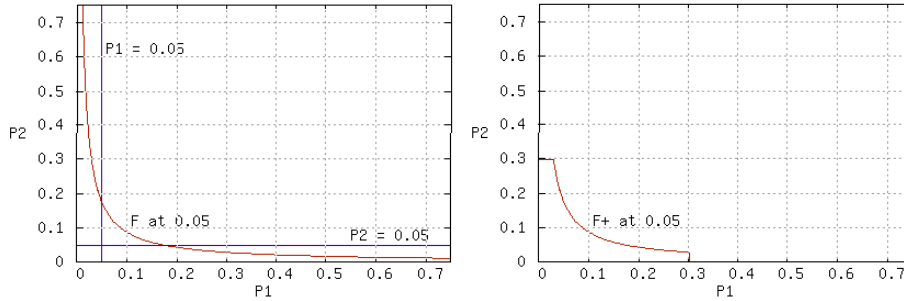
A more common method that makes better use of all data streams is Fisher’s method of combining p-values [3]. He observes that since p-values have a uniform distribution under the null hypothesis, the sum of logs of independent p-values have a  $\chi^2$  distribution with  $2n$  degrees of freedom:  $\sum_{i=0}^{n-1} \ln p_i \sim \chi_{2n}^2$ . Conceptually, this approach is easier to understand as computing an aggregate statistic which is the product of the p-values. It also turns out that there is an equivalent closed form solution for the combined p-value,

$$k \sum_{i=0}^{n-1} \frac{(-\ln k)^i}{i!}$$

where  $k = \prod_{i=0}^{n-1} p_i$ .

In order to illustrate the increased detection power of the combination methods, Figure 1 (right graph) shows the rejection regions for two independent p-values separately ( $P1 < 0.05$  and  $P2 < 0.05$ ) as well as for their Fisher’s combination ( $Fisher's < 0.05$ ). Note that for the sake of simplicity this example deals only with two data streams. The axes in the graph represent two p-values coming from two independent data streams. If we take into account only one of the p-values, for example  $P1$ , the null hypothesis rejection region is restricted to the left of the line  $P1 = 0.05$ . In this case, the null hypothesis is rejected only if  $P1 < 0.05$  no matter what is the value of  $P2$ . Whereas, when we use the *Min* aggregation the rejection region is extended and the null hypothesis is rejected when either  $P1$  or  $P2$  is below 0.05. Fisher’s method also adds to the rejection region cases where both p-values are out of but close to their individual rejection regions. In these cases, Fisher’s method is able to combine independent weak signals of departure from the null hypothesis and reject it even when neither  $P1$  nor  $P2$  individually would lead to reject it.

We refer to Fisher’s method as a non-specific detector because it is intended to detect any departure from the null distribution. It is useful even when we have no information about the type of outbreak we want to detect. We note that the generality sacrifices statistical power in the case where we have a model of the outbreak.



**Fig. 1.** Left graph: Individuals’ and Fisher’s aggregated rejection regions. Right graph: Fisher’s-based specific rejection region.

### 2.3 Fisher’s-Method-Based Specific Detector

Assume that we have been handed a “scenario” describing a particular kind of outbreak and we would like to improve our detection ability for it. In the empirical tests below, the scenario is one where a ramp up in positive observations occurs simultaneously in three data streams. Note though that the null hypothesis being tested with Fisher’s method is that none of the data streams departs from the null distribution.

As a demonstration of the increased power possible from specific detectors, we handcraft an extension to Fisher’s method. The modified detector uses the same combined p-value as the original, but then chooses not to signal an alarm if less than two of the uncombined p-values are below a given threshold. Those are cases where the evidence definitely does not match the scenario we are looking for, since only one stream departs from the null distribution.

The graph on the right in Figure 1 shows the rejection region for our example with only two data streams. The additional condition to signal an alarm removes from the Fisher’s *non-specific* rejection region areas that do not match our artificial “scenario”. The removed areas correspond to those cases where only one of the streams rejects the null hypothesis (low p-value) while the other one does not (large p-value). The effect of the specific detector is to reduce the number of false positives which in turn may allow us to adopt a higher threshold for signaling an alarm and consequently achieve a quicker detection.

## 3 Experiments and Evaluation

### 3.1 Artificial Scenario

In order to evaluate the performance of the plain Fisher’s method of p-value aggregation and our Fisher-based specific detector we injected artificial outbreaks into the actual data streams. We augmented the actual counts in the streams by using a multiplying factor that linearly ramps up over the period of the simulated outbreak. Such an outbreak is specified with three parameters: (1) the maximum

factor by which the actual counts are multiplied,  $\Delta$ , that represents the strength of the outbreak, (2) the total outbreak duration,  $Od$ , in days, and (3) the ramp duration,  $Rd$ . The multiplying factor grows linearly during the ramp duration and then it is kept constant at  $\Delta$  until the end of the outbreak. More precisely, the multiplying factor for the  $i$ 'th day of the outbreak is calculated as follows:

$$\begin{cases} 1 + (\Delta - 1)/Rd \cdot (i + 1) & \text{if } 0 \leq i < Rd \\ \Delta & \text{if } Rd \leq i \leq Od \\ 1 & \text{otherwise} \end{cases}$$

We injected outbreaks in all of the three data streams simultaneously, beginning at the same day, so that their combination would loosely match the pattern for which our specific detector is designed. The total duration  $Od$  was set to 14 days and the ramp duration  $Rd$  to 7 days. The  $\Delta$  parameter was set to 2.0 for stream A, 2.5 for stream B, and 2.3 for stream C. Those parameters were chosen so that the outbreaks would not be immediately detectable in the individual streams during the ramp-up periods.

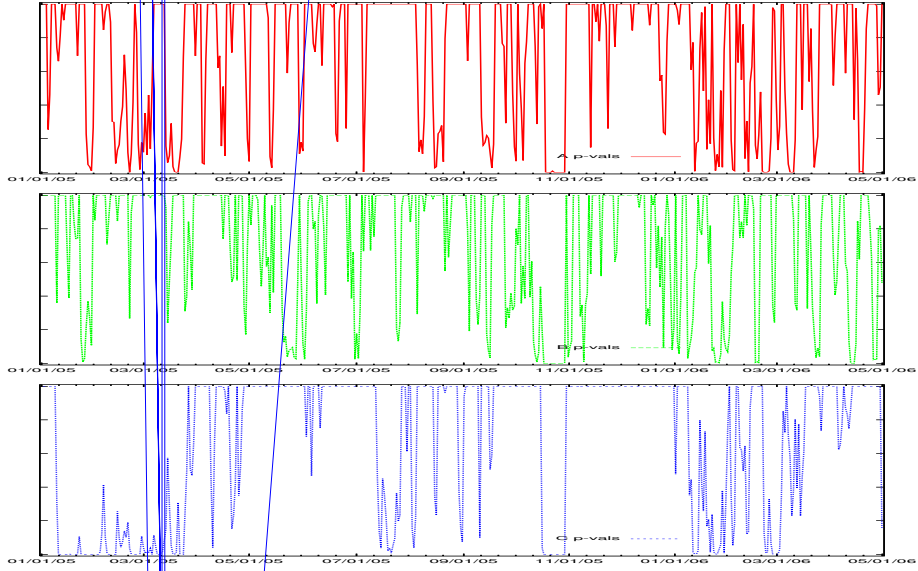
Figure 2 depicts the time series of p-values for the individual streams of considered data, computed using the 2-day-wide temporal scan window. Note that wider windows render more smoothing and lower the sensitivity of the univariate detectors (for brevity we do not discuss such effects in this paper). The synthetic outbreak can be seen to begin in the second half of October 2005.

It should be noted that it is best to experiment using labeled, known outbreaks identified in the historical data. Unfortunately, at the moment of writing this paper we had not had access to such information, and therefore resorting to realistic but artificial injections was a requirement.

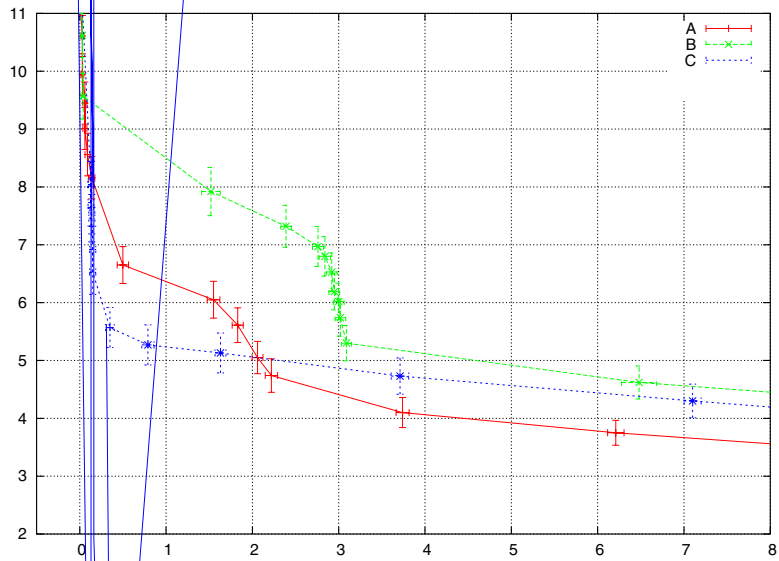
### 3.2 Min and Fisher's Multi-stream Methods

Figure 3 presents the Activity Monitoring Operating Characteristic (AMOC) curves for the univariate detectors set up for the individual streams A, B and C, and the corresponding curves for the two non-specific multi-stream detectors: *Min* and plain Fisher's (labeled F in the graph). The horizontal axis of the graph corresponds to the number of detects outside of the period of the injected synthetic outbreak, and the vertical axis denotes the time to detection in days from the first day of the outbreak ( $i = 0$  in the formula above). The more powerful the detector, the more its characteristic bends towards the lower left corner of the graph. The points and error bars shown are obtained as means and standard errors based on 100 independent injections of simultaneous outbreaks into the individual streams of data, with randomly selected start dates.

It is clear that the non-specific detector implementing Fisher's method of p-value aggregation has a superior detection power over the univariate detectors set on the individual streams of data, as well as over the Bonferroni-correction-motivated *Min* aggregate. This is due to Fisher's method ability to accumulate evidence from independent sources, or data streams, and to include into the rejection region cases where individual streams are close but above the threshold.

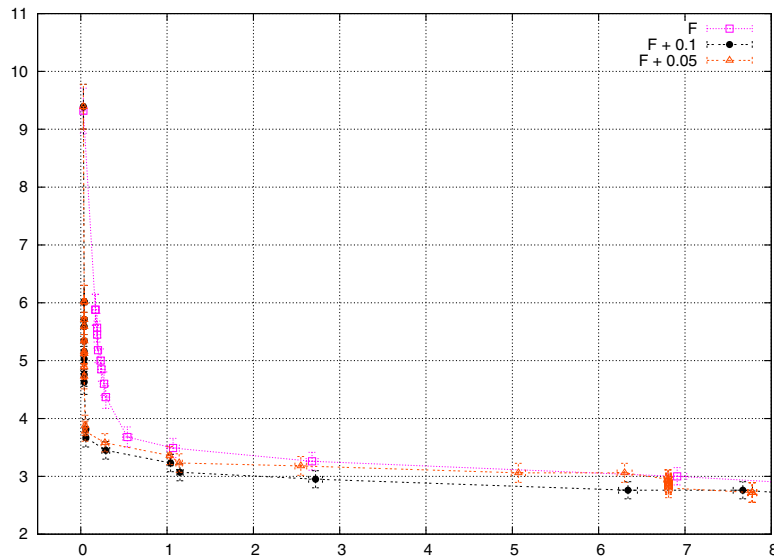


**Fig. 2.** P-values computed for the individual streams of data using temporal scan with the window width of 2 days. The synthetic outbreak starts on October 15, 2005.



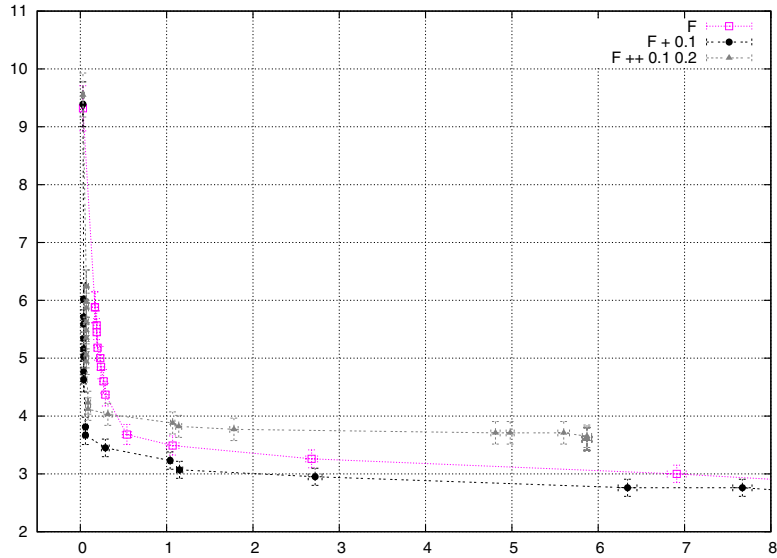
### 3.3 Fisher’s-Method-Based Specific Detector

Figure 4 shows a comparison of performance between the non-specific detector and its specific alternative (labeled F+ in the graph). Recall that our handcrafted specific detector uses the Fisher’s method but it signals an alarm only if at least two of the streams are below some given threshold. We plot curves for the specific detector with the threshold set to 0.05 and 0.1. We can see that both cases outperform the plain Fisher’s method. We observe though that setting the threshold to 1 would prevent the specific detector from filtering out any alerts and thus it would reduce to the plain Fisher’s method. On the other hand, setting the threshold set to 0 would lead to filtering out all the alerts, including true positives. In our experiments we found that a threshold of 0.1 works best with the considered data.

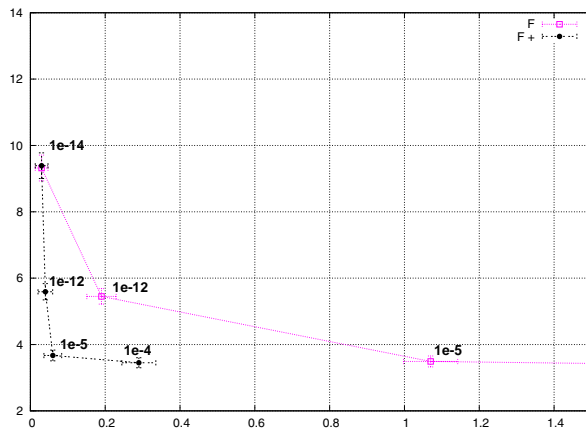


**Fig. 4.** AMOC curves for multi-stream detectors: the non-specific (Fisher’s) and the specific described in Section 2.3 with threshold set to 0.1 and 0.05

We have experimented with other Fisher’s-method-based hand-crafted specific multi-stream detectors. There are many potentially useful filters that one could think of. For example, since the considered outbreaks are associated with linear ramp-ups of positive counts during the first few days, one could require that Fisher’s-combined p-values for preceding two days are arranged in decreasing order with the current day’s p-value, i.e.  $p_{day-2} \geq p_{day-1} \geq p_{day0} < \alpha$ , in order to signal an alert on the given day. Another possibility could be to rise an alert only if both the current day and the previous day have critical p-values associated with them. That condition would bank on the fact that the outbreaks are known to last several days. Yet another condition could be to require that on



**Fig. 5.** AMOC curves for multi-stream detectors: the non-specific (F), specific (F+), and specific with a condition on the current and previous day (F++)



**Fig. 6.** An improvement of detection time can be achieved using a specific detector in place of its general-purpose counterpart

the current day at least two streams must produce p-values below some threshold and that on the previous day at least one of them was critical. Empirically, none of those ideas were able to outperform the plain Fisher’s method. Figure 5 shows the result for the last idea, labeled F++, where the threshold for the current day was set to 0.1 and for the previous day was set to 0.2. We can see that it outperforms Fisher’s method until the fourth day of detection latency, but then it also filters out the first days of the outbreak together with false positives

and thus it is not able to signal an alert faster than Fisher's method at lower detection latency settings.

Apparently, it is possible to take advantage of knowing the characteristic pattern of the particular type of a bio-event, and to construct detectors with higher specificity and timeliness of detection than their general, non-specific alternative.

Figure 6 illustrates how a specific detector can improve detection time over a non-specific detector. On the plain Fisher curve, using an alert threshold of  $1e-12$  produces the best result with only 0.2 false positives on average and an average delay in detection of 5.5 days. When the specific detector is added, most of the false positives are eliminated at the alert threshold of  $1e-12$ . Thus for the specific detector, the alert threshold can safely be increased up to  $1e-5$  or  $1e-4$  which brings the average time to detect down to 4 days.

## 4 Conclusions and Future Work

We demonstrated the usefulness of aggregating information from multiple data streams on real food- and agriculture-safety data from the USDA using a general aggregation method. We also showed how a handcrafted aggregation method tuned to specific outbreaks of interest can detect an outbreak faster. We recommend that a fielded system be comprised of both a non-specific detector and as many specific detectors as possible.

We want to stress that manual design of specific detectors might be a difficult task even when outbreak patterns are as simple as the one used in the presented results. In the future work, we will investigate how to learn specific detectors from data streams labeled with various outbreak types.

## Acknowledgments

This work was partially supported by the United States Department of Agriculture prime contract number 53-3A94-03-11 (Task 19) and by the Centers for Disease Control and Prevention award number 8-R01-HK000020-02.

## References

1. H.S. Burkom, S. Murphy, J. Coberly, and K. Hurt-Mullen. Public health monitoring tools for multiple data streams. *MMWR Morbidity and Mortality Weekly Report*, August 2005.
2. A. Dubrawski, K. Elenberg, A. Moore, and M. Sabhnani. Monitoring food safety by detecting patterns in consumer complaints. In *Proceedings of the National Conference on Artificial Intelligence AAAI/IAAI 2006*, 2006.
3. R. Fisher". *Statistical methods for research workers*. Oliver and Loyd, 1925.

4. Protecting against agroterrorism. GAO-05-214. Technical report, Government Accountability Office, March 2005.
5. D.B. Neill, A.W. Moore, and G.F. Cooper. A multivariate bayesian scan statistic. In *National Syndromic Surveillance Conference 2006*, 2006.
6. M.M. Wagner, A.W. Moore, and R.M. Aryel, editors. *Handbook of Biosurveillance*. Academic Press, 2006.
7. L. Wasserman. *All of Statistics*. Springer, 2004.