Applying Outbreak Detection Algorithms to Prognostics

Artur Dubrawski[1], Michael Baysek[1], Maj. Shannon Mikus[2], Charles McDaniel[3], Bradley Mowry[4], Laurel Moyer[4], John Ostlund[1], Norman Sondheimer[5], Timothy Stewart[6]

(1) Auton Lab, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, awd@cs.cmu.edu .
(2) AF/A4ID, 1030 Air Force, Pentagon, Washington DC 20330.
(3) C&M Consulting, 24441 Woodbridge Way, Schertz, TX 78154.
(4) 84th Combat Sustainment Wing, Hill Air Force Base, UT 48056.
(5) University of Massachusetts Amherst, Amherst, MA 01003.
(6) Lockheed Martin Aeronautics Company, Hill Air Force Base, UT 48056.

Abstract

Fleet maintenance and supply management systems are challenged to increase the availability and reliability of equipment. Prognostics can help. This paper examines the utility of selected statistical data mining algorithms, originally developed for bio-surveillance applications, in achieving fleet prognostics. Preliminary experimental evaluation suggests that it is possible, useful and practical to apply such algorithms to rapidly detect emerging patterns of systematic failures of equipment or support processes, and to continuously monitor relevant data for indications of specific types of failures. The key remaining technical challenge is to tame down a potentially large number of plausible pattern detections without compromising high detectability rates. The key practical consequences to maintenance and supply managers include the ability to be notified about emergence of a possible problem substantially earlier than before, the ability to routinely screen incoming data for indications of problems of all conceivable types even if their number is very large, and the ability to pragmatically prioritize investigative efforts according to the statistical significance of the detections.

Introduction

The challenges of maintaining fleets of equipment are in certain ways similar to the issues arising in the management of public health. In the public health domain, specialists attempt to identify outbreaks of infectious diseases in their early stages in order to facilitate effective mitigation of their consequences to a human population. For example, when a new pathogen enters a local community, it typically leads to unusual patterns of demand on the health care system. Public health officials watch for these indications to alert them to take precautions to avoid catastrophic failure – an uncontrolled spread of the disease. Similarly, in the case of fleet prognostics, fleet managers attempt to identify an outbreak of identical equipment failures early in their cycle in order to mitigate their consequences..

The existence of relevant data makes the automated detection of emerging outbreaks conceivable. In case of public health, the informative data streams include emergency room admission records, pharmacy sales, school/work absenteeism counts, etc. Based on such data, appropriate outbreak detection algorithms can be developed and deployed to assist public health officials. Methods stemming from statistical process control, outlier detection, time series analysis and forecasting are being used to tackle the detection problem. Statistical machine learning nicely complements those approaches by allowing learning detectors automatically from historical data. It has been demonstrated to enable warnings of the advent of epidemics by observing the experience of the healthcare system. Our team has been actively participating in these developments and it has fielded a variety of systems that are used in government agencies world-wide.

Our current research is looking at the opportunity to apply bio-surveillance methodology to fleet prognostics. This paper discusses the initial results of experimental application of the outbreak detection algorithms to aircraft fleet management. The study uses as a test bed avionics components of the United States Air Force F-16 fighter fleet and it analyzes already existing databases recording aircraft logistics and maintenance activity. The initial stage results have been very encouraging. The prototype algorithm has discovered a statistically significant pattern of maintenance activity weeks before a set of parts began to fail fleet-wide. Such patterns can easily go unnoticed without an automated surveillance system. Our next steps include comprehensive evaluation of a range of potentially relevant techniques of event detection, drill-down and statistical explanation of findings. The long term objective is to adapt the technology to account for the differences between biological organisms (the original domain of development of the underlying methodology) and man-made devices (the target domain).

This paper explains the need for prognostics in fleet management, reviews related efforts including our previous work on bio-surveillance, details the current avionics experiment and it ends with the conclusions and plans for future work we have drawn from the experiment.

The Need for Prognostics in Supply Chain Management

In planning for the maintenance and supply of a fleet of equipment, fleet managers make assumptions about the scope and tempo of future operations, the quality and lifespan of the replacement parts, and the effectiveness of maintenance activities, among other factors. When any of those assumptions are violated, for example, through a change in operations tempo, delivery of an order of out-of-spec parts or institution of an ill-conceived maintenance procedure, a logistics crisis can develop. We refer to these failure modes as systematic since they can have a fleet system-wide effects on equipment availability rates, and since they differ from chance events which may lead to isolated failures. The following list provides a set of typical systematic failure examples:

- Maintenance:
  Unexpected consequences, localized misunderstandings;
- Quality Control:
  Acceptance of out-of-specification parts;
- Operations:
  Environmental factors, optempo changes, aging;
- Design:
  Original design flaws, unexpected consequences of modifications;
- Logistics Planning:
  Contracting issues, parts allocation.

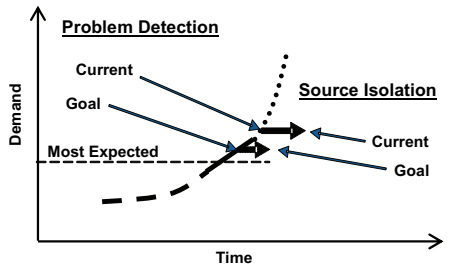

Figure 1: Demand during systematic failures.

The systematic failures usually do not occur immediately upon the change in operations, the installation of the part or the completion of the maintenance. Instead, the part and the equipment begin to create the need for early parts replacements at a later date. Human managers of the fleet typically have difficulty recognizing the initial indicators of emerging patterns of systematic failures. This is especially true with large fleets dispersed over many locations where local maintainers witness only few relevant cases. Even when parts demand data is centralized and frequently updated, the relatively limited numbers of inventory managers in central sites, each of whom manages a relatively large array of parts, are unlikely to have the luxury of focusing in on the initial evidence of unexpected demand for specific parts. Only when demand nears the safety stock level is it likely that attention will be paid. Given the lead time to fill a parts order and the likely increasing demand for replacement parts during progression of the systematic failure pattern, it is very likely that assumptions made during procurement regarding part or system reliability, service environment, supply rates, etc., will no longer be valid and the end result is that needed parts will not be available in needed quantities while it appears that excess parts were procured for other systems.. Worse yet, inventory managers must assume that parts will have to be continued to be replaced at an accelerated rates until the source of the failed assumption is isolated and the crisis mitigation plans can be made.

Figure 1 shows a graph suggesting how the problem would be perceived by inventory managers over the time of its progression. The vertical axis shows cumulative demand and horizontal axis the passage of time. The dashed section of the hypothetical curve depicts a range of demand level such that the inventory managers can expect to meet demand for parts. The demand shown by a solid line indicates where the organization is taking regular procedural steps to account for the unexpected demand by say deferring overhauls of parts. The demand indicated by a dotted line shows where continued targeted levels of operation are impossible to achieve. This is when the problem is traditionally detected. Such a supply management system can be called reactive since it waits until the dotted-line condition is reached to acknowledge that the problem has occurred, to only then possibly initiate mitigation actions. The figure also indicates that once a failure is identified, additional time is required to isolate the source of it. Once the source is isolated, a true recovery plan can be put in place.

Figure 1 also shows the goals of our research. We want to create tools to automatically recognize unexpected or anomalous failure patterns earlier than is currently possible and to reduce the time needed to isolate their source. Given that we are identifying changes in the remaining life of a set of parts, we are aiming for prognostics.

Related Work

Supply Chain Management and Prognostics. Our work combines spare parts inventory planning and data-driven equipment management. Most textbooks on these topics provide chapters on parts inventory planning [Fitzsimmons 1997]. Many vendors offer inventory planning systems with the latest focus being on integrated supply chain management systems. In general, these systems consider order quantity, reorder point and safety stock and establish these based on an estimate of demand per unit of time or usage rates. Typically, the estimates are established by looking at historical demand on inventory. Concurrently, some experimental work has been done into

data-driven prognostics for systems health management. [Schwabacher 2005] proposes an approach to detect precursors to a part's individual failure and to estimate its remaining life. The fact that models are learned from the actual operation of the equipment in question (hence the term "data-driven") differentiates this approach from hand-coded or laboratory-developed physics-based models. The end-product of data-driven prognostics is typically an embedded monitoring system which estimates the remaining life of a single unit. Recently, the attention is being paid to methods of estimation of the remaining life of a particular unit by looking at the history of performance of similar units [Bonissone 2005, Xue 2007]. Our work combines the insights of those endeavors with the field-proven algorithms of outbreak detection used in bio-surveillance in order to address a novel set of problems.

Bio-Surveillance. Early detection of bio-events is one of the key issues in maintenance of public health. Human population is vulnerable to a range of pathogens, including those causing infectious diseases and those which could enter the organism with contaminated food or water. Over the recent years, a number of research programs have been instituted to enable public health experts with the ability to collect and then to purposively and in a timely fashion analyze data which may indicate threats to public health safety. Traditional research into epidemiology has been augmented with thrusts into new analytical algorithms, data structures, information infrastructure, statistical methods of detection, etc. The original motivation stemming from the terrorist threat has been later complemented by the need to deal with naturally occurring threats, which could also be detected with the help of contemporary bio-surveillance systems. Public health and food safety are very prominent application areas of this new technology because the costs of missing a bio-event or not detecting it in its early stage may be overwhelming. According to World Health Organization, a 100 kg of inhalational anthrax released in an urban area such as Washington, D.C. may kill up to 3 million people in a very short span of time, if not mitigated rapidly after the release. According to DARPA studies, a 2-day gain in detection time of such incident would reduce the count of fatalities by a factor of six [Wagner 2001, Tsui 2005]. Naturally occurring bio-events such as outbreaks of SARS, Avian Influenza, West Nile Virus, Mad Cow Disease, Foot-and-Mouth Disease, E.coli or Salmonella, albeit not as spectacular as terror threats, also may put a huge burden on safety and economic stability of human population.

Traditionally, surveillance of public health data is based on analysis of time series of counts of individual events such as identified cases of human illness, and it is being performed in a temporal or a spatio-temporal context. Fundamental technologies range from uni-variate forecasting, to statistical process control, to clustering of events [Wagner 2006], and most of them are tailored to detect anomalous departures of the observed processes from their expected levels. More specialized approaches, many of them based on concepts of artificial intelligence, have been proposed in order to improve speed and reliability of outbreak detection, to handle specific objectives, and to learn detection models from data. Below we review a few of the methods developed at the Auton Lab.

WSARE (What's Strange About Recent Events) [Wong 2003] uses a database of past transactions such as patients visits to hospital emergency rooms as the baseline, and statistically tests whether the current data is different from the past. Records of healthcare transactions are usually annotated with multiple descriptor variables (such as patient's age, gender, home location, symptoms). WSARE in a computationally efficient manner inspects all of their combinations to detect departures of today's data from the expected in the context of specific subgroups of population. For example, it can detect that the number of children reporting today with bloody stools who live in the north of the city substantially exceeds the expectation. Traditionally, such detections were not possible in practice unless a public health official specifically suspected an outbreak that affects a particular sub-population.

If the baseline and target data are annotated geographically, a spatial scan statistic can be used. It considers regions of many different sizes, shapes, and locations and for each of them it tests a hypothesis that the data distribution is the same inside and outside of the current region. Because even millions of regions are considered, special care is being paid to assess the statistical significance of the findings, and the involved algorithms are computationally very efficient to handle voluminous data, such as daily sales of non-prescription medicines collected nationwide, in a practical amount of time [Neill and Moore 2004].

Another approach, Tip Monitor, has been successfully applied to monitor food consumer complaints for linkages between individual reports of adverse effects of food on people where the complaints may be probabilistically associated with the same underlying cause [Dubrawski 2006]. Its unique ability is to remain sensitive to signals supported by very little data – significant alerts can be raised on the basis of a very few complaints, provided that they contain significantly similar and explicable root causes.

Our current research aims at transferring experience gained from development and deployment [Sabhnani 2005] of the surveillance methods described above, to the domain of fleet health prognostics. The next section reviews our first attempts at achieving that.

Examples from the F-16 Avionics Component Supply Chain Management Experience

We hypothesize that there is an analogy between sets of humans getting an illness and sets of parts failing through a systematic issue. We argue that that the evidence of that human epidemic is analogous to the evidence of the

beginning of a systematic failure of parts. In this view, the mathematics applied to public health data bases to achieve bio-surveillance should yield prognostics of systematic failures of parts or processes when applied to maintenance and logistics data bases.

We have had the opportunity to explore this argument with actual USAF maintenance and logistics databases covering multiple years of maintenance and logistics on F-16 avionics equipment. These results have been promising. We will present the basic methods to identify unusual patterns, the tools we have to employ them in prognostics, scalability results, possibilities for controlling sensitivity, an experimental evaluation, and the ideas on the source isolation task.

Experimental setup. In the experiments documented below we hypothesized that certain types of events of interest would manifest themselves in maintenance and supply records in the form of statistically detectable patterns of departures from what is considered a normal level of activity. In particular, we will look at the time series of daily counts of maintenance and supply events of different types, recorded in the available data. A time series of daily counts of events of all types, extracted from the available maintenance data, is shown in the bottom graph in Figure 2. The horizontal axis denotes time and the vertical axis denotes daily counts of events. No doubt the changes in those counts over time could be driven by the variability in op-tempo, particular maintenance campaigns, and delivery of a long awaited set of parts or many other reasons. The top graph presents analogical plot obtained for events of a specific type. In the presented example, these events correspond to a specific type of equipment malfunction (in this case it is a failed automated diagnostic test) and to the circumstances in which the problem was discovered (in this case, in flight). In general, a query selecting events of interest may involve any subset of those and other characteristics of the records of maintenance or supply data (such as geographic location of the incident, type of the last flown mission, malfunction code, plane configuration, action taken, organization code, etc.).

Temporal Scan. The basis of our experiments has been time-based identification of unusual activity. The graphs in Figure 2 are spiky but they reveal potentially interesting longer-term temporal patterns. One way of statistical analysis of such time series is to detect and automatically report unexpected, statistically significant increases of the recorded counts of maintenance or supply actions of a certain kind. An unusual increase may indicate an emerging systematic failure pattern, and hence its automatic detection may be very useful for the purpose of pro-active management of maintenance and logistic activities. The analysis should also be robust to high frequency noise present in the daily counts data. That can be easily attained by smoothing the raw time series by considering moving sums of counts aggregated over sequences of days instead of counts recorded on individual dates. In fact, as the data at hand reveals a strong day-of-the-week effect, it makes a practical sense to consider temporal aggregation window widths to be multipliers of 7 days.

An interesting question is whether an observed escalation

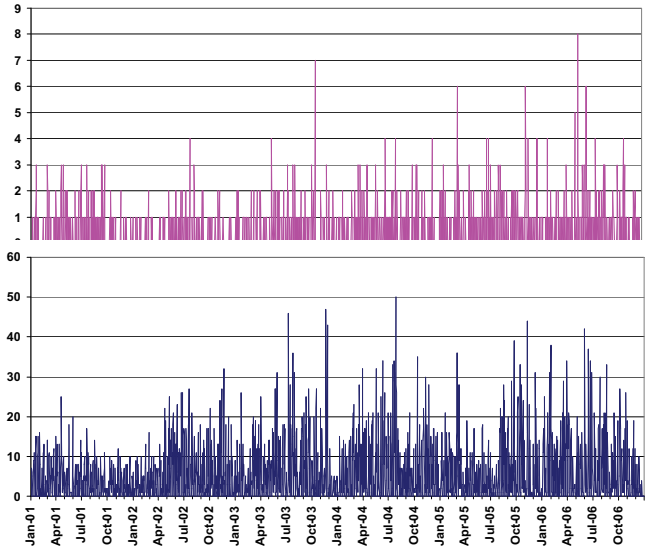

Figure 2: Time series of daily counts of maintenance activity for an avionics component of F-16 over the period from January 2001 till December 2006. Bottom: cumulative counts of all maintenance actions involving the selected avionics component. Top: subset of all counts corresponding to the events involving in this example failures of automated test procedures which have happened in flight.

in counts of a specific type of maintenance or logistic events is unique to that type, or perhaps it could be explained by the general increase in maintenance/logistics activity. We should be primarily interested in such increases that do not correlate with general activity since they may indicate emergence of the targeted systematic failures. Temporal scan [Roure 2007] is a straightforward statistical procedure which provides the needed functionality. Its algorithm can be written as follows:

For each day of analysis:

1. Establish a time window of interest ending at the current day and starting T-1 days before, inclusive.
2. Compute sum of target and baseline counts for all days inside the window and, separately, for all days outside of it.
3. Put the results in a 2-by-2 table and execute Chi-Square or Fisher's exact test of its significance.
4. Report the resulting p-value. The lower the p-value the less closely the target time series follows the baseline within the time window of interest.

For instance, setting up the analysis for the target and baseline series shown in Figure 2 using the width of the time window of T=28 days, the counts computed for June 29, 2005 yield 28 and 847 for the target series, and 95 and 11,950 for the baseline, respectively for the within time

window of analysis and outside of it. It is not hard to see that the current count of target malfunctions is disproportionally high when compared to the expectation derived from the baseline counts (if the current target counts followed the variability of the baseline, we would expect them to be near $95 \cdot (847/11{,}550) \approx 7$). This observation is confirmed by the very low p-value of $7.39 \cdot 10^{-8}$ of the Chi-square test. The relatively high count of failed automated tests in flight observed during 28 days preceding and including June 29, 2005, cannot be therefore statistically justified by the increase of general level of activity. Instead, it may indicate an on-going or developing systematic hardware, software, and support or process problem.

There usually exist other factors not represented in the data which may explain unexpected variability of the target time series. Their consideration will have to be left to human analysts who process the results of the temporal scan procedure. Accepting this limit, examples like the above have led us to consider how temporal scan can be used for prognostics.

Unguided Exhaustive Search. Maintenance and supply data is typically analyzed in a focused investigation mode in which a specific scenario of systematic failure is hypothesized and the hypothesis is then tested against the evidence available in data. This has a limitation that the analyst must know what, when and where to look for. Recent developments in data representation techniques for scalable data mining make it possible to sometimes overcome the computational limitations and offer the analysts the ability to monitor all conceivable hypotheses of failure scenarios which can possibly be tested using the available data. We have proposed, implemented and tested such an exhaustive approach, calling it the unguided search, in which the user does not have to manually select or pre-define specific queries. Its algorithm can be described as follows.

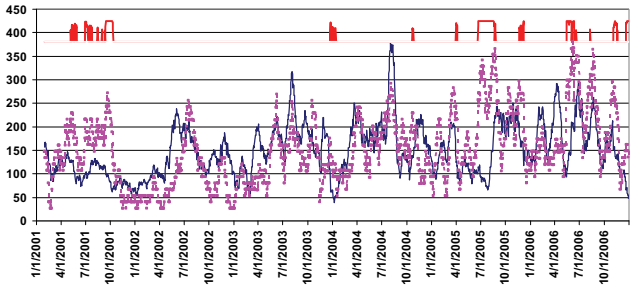

Figure 3: Results of executing temporal scan on the pair of time series of counts depicted in Figure 2. The days of significant departures of the target counts from expectations, at the level of significance α=0.05 are marked in the upper part of the graph. Solid line shows the 28-day moving sums of the baseline series while the dotted line depicts the same (but scaled up to match the volume of the baseline) for the target counts.

For each combination of descriptor variables which determines a unique query:

1. Extract daily counts of maintenance and supply events that match current query.
2. Execute the temporal scan procedure to identify days with counts significantly higher than expected based on the general activity level (assumption: the activity level can be estimated with the total daily counts of all types of maintenance/supply actions).
3. Report days associated with equal or less than critical p-values.

The graph in Figure 3 presents results of executing temporal scan against time series of counts presented in Figure 2. Markers in the upper part of the diagram indicate days for which the counts of the target query series significantly (at the level of 0.05) exceeded expectations based on the baseline counts and the target counts aggregated outside of the time window of analysis. The detection of a prolonged period of significant departure from normal during summer of 2005 would have been potentially useful according to US Air Force logistics experts. The practical application of the unguided search provides the analysts with the new ability to perform a full-scale, unconstrained surveillance of data. It is beneficial because missing important events becomes harder. However, it also raises two new challenges: (1) How to manage a potentially large number of alerts generated by the system, and (2) How to execute exhaustive search through all possible time series queries in a reasonable time. Those important issues are discussed next.

Scalability. A typical transactional data set considered in this exercise consisted of 33 demographic attributes whose values described individual maintenance actions present in data. There were over 77 thousand such actions recorded over a period of about 6 years at a daily resolution. The descriptors assumed the form of categorical variables with arities varying from 2 to hundreds. The total number of combinations of one and two unique attribute-value pairs (unique select queries of sizes 1 and 2) was close to 1 million. The total number of individual (daily) temporal scan tests for such data set exceeds 2 billion. Each such test involves a Chi-square test of independence performed on a 2-by-2 contingency table formed by the counts corresponding to the time series of interest (one of the million series) and the baseline counts, within the current temporal window of interest (one of over 2,000 windows under consideration in the complete retrospective analysis mode of the current data) and outside of it. If the users chose one of the commercial database tools, the time needed to just retrieve the time series corresponding to one of the involved queries would approach 280 milliseconds. Therefore, it would take about 3 days just to pull all the required time series from the database (non-indexed setup with processing taking place on a local server without the need for remote access), not including any processing or execution of the statistical tests.

That kind of analysis would be rendered infeasible without the efficiencies provided by the smart data structure called T-Cube [Sabhnani 2007] recently developed at the CMU Auton Lab. The T-Cube is a cached sufficient statistics, in-memory data structure designed to efficiently store answers to virtually all conceivable time series queries against sets of additive temporal data such as daily counts of transactions labeled with a set of categorical variables. Using the T-Cube, the complete set of computations, including the time necessary to retrieve and aggregate all the involved time series, compute and store the test results, load source data and build the T-Cube structure, etc., took only about 1 hour 5 minutes when executed on a dual CPU AMD Opteron 242 1,600 MHz machine (using only one CPU for computations), in the 64 bit mode, using 1MB per CPU level 2 cache and 4 GB of main memory, running under Cent OS 4 Linux operating system. With a state of the art workstation the time would be considerable lower.
T-Cube is an important enabling technology which makes the exhaustive unguided search as well as interactive ad-hoc drill-down analyses feasible. The former gives the analysts the ability to maintain high situational awareness by screening all conceivable scenarios of temporal divergences from normal, which could be checked against the available data. The latter provides the analysts with an improved data understanding as they can execute detailed follow-ups through large collections of time series in an interactive fashion. We will return to this topic shortly.

Sensitivity and Multiple Hypotheses Testing. Sensitivity is one of the key parameters determining a practical utility of the presented approach. It can be controlled with the Chi-square test's threshold of significance α. Higher sensitivity (higher α) decreases the risk of false negative detects ("misses"), but it also increases the frequency of alerts (proportion of alerts which may be false goes up). Sensitivity can be traded-off for specificity (the ability to never miss true events). This trade-off can be resolved with the Receiver Operating Characteristic (ROC) approach which conveniently allows for incorporation of the costs (if they are known) of the two types of errors into decision process so that it could be made cost-optimal [Dubrawski 2004].
At any rate, in the realistic setup considered in this paper, we need to manage a potentially huge number of supposedly statistically significant findings, which may not always be also interesting from the practical point of view. At α =0.05 we expect 5% of conducted tests (5% of 2 billion!) to appear significant, even if the NULL hypothesis held true. That is way too many alerts to deal with! One of particularly elegant and effective ways of dealing with testing multiple hypotheses in statistics is the method of False Discovery Rate [Wasserman 2004] which we currently employ. Other methods under consideration include: (1) Making sensitivity user-selectable, (2) Exploiting sequences of alerts (a sequence of consecutive days of alert is very often a consequence of a single underlying cause and the whole sequence should be treated as a single alert), (3) Constructing specific detectors tailored to monitor against particular scenarios of interest.

Discovery of Identified Historical Events: Guided Search, Timeliness of Detection. Retrospective analysis of historical data which contains identified examples of events of interest provides an excellent opportunity for objective evaluation of the developed outbreak detection algorithms. In the example discussed below, the US Air Force logistics experts provided us with one particular instance of an outbreak of systematic failures which was reportedly originally discovered in June 2006. They also provided us with the list of malfunction codes associated with the kinds of maintenance and supply events which could be linked to that particular outbreak.

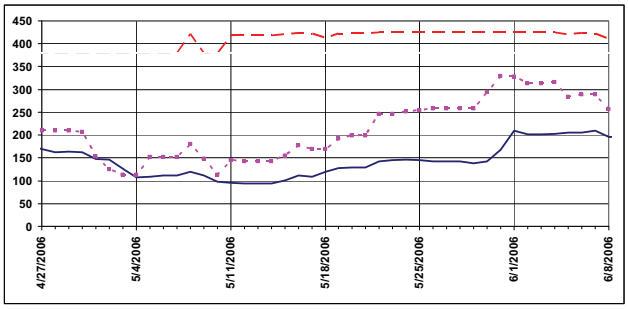

Figure 4: The particular event of interest can be reliably detected 3 weeks ahead of the estimated earliest date of its original notice, if the temporal scan detector uses 28-day time window and α=0.05. Solid line depicts the time series of baseline counts, the dotted line the target time series (scaled up to match the range of the baseline), and the dashed line in the top of the graph marks the significant alert days.

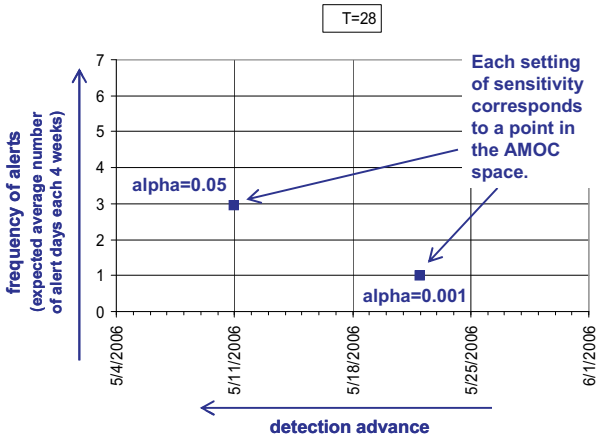

Figure 5: Activity Monitoring Operating Characteristic for the temporal scan detector evaluated on a historical event of interest at two sensitivity settings. Vertical axis denotes frequency of generated alerts (the count includes true positives as well as false alerts) measured with the average number of alert days per 4 weeks. Horizontal axis denotes time. Higher sensitivity (higher α) yields faster detection (the further left in the graph) at higher frequencies of alerts (the further up in the graph), and vice versa.

We implemented the following approach: (1) Extract daily counts of maintenance events matching the provided

indicators (simple – and therefore noisy – keyword matching of narrative fields was performed to accomplish that); (2) Attempt to detect significant increases of the corresponding counts of maintenance and supply actions in the period of time immediately preceding June 2006, which could not be explained by the increase of the general activity level.

We therefore run the temporal scan algorithm using the 28-day observation window width and significance threshold of alpha=0.05. Over the almost 6-year period of data coverage, it yield 305 days labeled as alert days. They grouped in 31 unique sequences of consecutive alert days. Assuming that the actual detection of the outbreak of interest has happened on June 1st, the temporal scan would have reliably seen the problem 3 weeks earlier (Figure 4).

Timeliness of detection can be traded off against sensitivity. Changing significance threshold to α=0.001 reduces the number of alert days to 68 (from 305), or approximately 1 per month, and the number of unique alert day sequences to 12 (from 31), or 2 per year. That comes at the cost of slower detection: now we would be able to see the problem only 10 days earlier than the assumed earliest date of the original discovery, but the hassle of having to deal with many potentially false positive alerts is being reduced. It is worth noting that similarly to ROC approach discussed earlier, AMOC can be used to optimize performance of detectors if the relative costs of dealing with potentially false detections vs. the marginal benefits of detecting events earlier can be estimated.

Explanatory Analysis. The preceding focuses on what we refer to in Figure 1 as Problem Detection. The analysts are often interested in finding out what explains an alert, part of what we refer to as Source Isolation. That is, what combination of descriptive field values contributes the most counts during the period of alert? Answering that may help in guiding the search for the underlying cause of the problem. Looking again at the graph in Figure 2 one can notice that the single day of the highest count of all types of recorded events happened in July 2004. On that day, the total of 50 maintenance events has been recorded. One may be curious to find out what types of activity contributed most to this unusually high daily count. In order to answer that question we perform another kind of an exhaustive time series scan which is being made possible by T-Cube: Peak Analysis. Peak Analysis runs a massive screening through all queries (that is, combinations of activity descriptors and their values involving all descriptive fields in data, not just the two used in the upper graph in Figure 2). It returns the descriptors of the simplest possible of all selection queries that leads to a time series which best explains the observed peak in data (note that simplicity of the query and its ability to best explain the data make contradictory requirements; the trade-off can be resolved via cross-validation). The actual result for the considered day in July 2004 identified 45 out of 50 entries and it narrowed down the scope of investigation to a relatively small subset of all organization codes, types of malfunctions and configurations of aircraft.

We observe that the ability of certain analytic techniques to point out and explain what is hidden in data in terms that are easy to interpret by humans, even though the underlying methodology is soaking in advanced computer science and statistics, is the key enabling factor in many of practical applications.

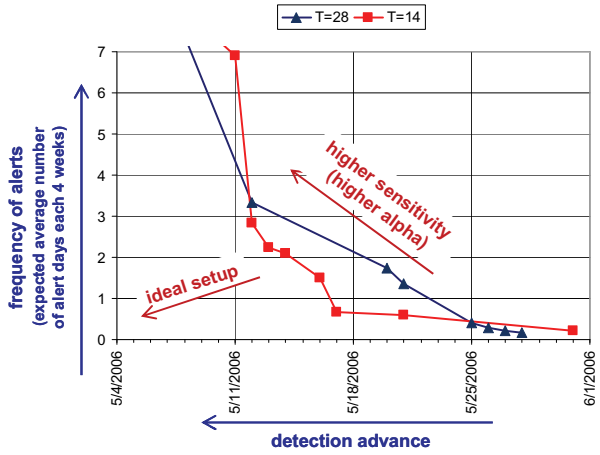

Figure 6: AMOC curves obtained for two temporal scan window widths by changing sensitivity thresholds. Shorter windows of observation lead to greater detection power (lower frequency of potentially false alerts at the same detection speeds or faster detection at the same rate of alerts) when sensitivity is not extremely low or excessively high.

Conclusion

Increasing requirements on availability and reliability of equipment challenge the existing maintenance and supply management systems. Prognostics can help. We examined the utility of a few analytic methodologies originally developed for bio-surveillance applications to support that formidable task. It is apparently possible to formulate certain fleet-level prognostics tasks in terms which are compatible with the concepts or rapid detection of outbreaks of infectious diseases. The detection algorithms can be tailored to sift through maintenance and supply records in order to spot early warnings of systematic hardware, software, and support or process failures. Being warned early about a potential problem, maintenance and supply managers can stage timely mitigation activities, while engineers and logisticians can initiate investigations into the root cause of the problem much earlier than currently possible. Their early response reduces the impact of the emerging issues on the availability and failure rates of the equipment.

The experiments have been conducted on the real maintenance and supply data for a specific subset of avionics of the F-16 fighter plane. It is important to note that the kind of data we used is being routinely collected by the United States Air Force support units, so no

investment in new infrastructure was required to make the presented work possible.

We have validated the utility of one of the applicable analytic techniques (temporal scan) in a retrospective analysis mode. The attempt to discover an identified historical event yielded, depending on the parameter settings, a 10 to 21 day earlier detection than originally attained without an automated surveillance system. We have shown how timeliness of event detection can be traded for frequency of alerts, and we hinted on how to control sensitivity of detectors to achieve cost-optimal settings of surveillance operations.

We have also shown how an efficient data representation technique such as T-Cube can enable simultaneous monitoring of a massive number of hypotheses of emergence of systematic failures. Computationally efficient detection algorithms can be therefore executed in an unguided search mode to exhaustively screen all incoming data and to report the most significant findings to human analysts for investigation. This way, the data can be analyzed automatically in vastly more configurations than possible today, which minimizes the risk of missing important signals. There is a real need for automated screening because it is clearly beyond human ability to process and internalize such a large set of potentially critical information. The undertaken statistical approach to analysis and explanation of the results brings about the ability to rank and prioritize the findings, in order to let the users pragmatically allocate their limited investigative resources.

We have presented the work in progress. More validation on a wider range of usage scenarios is required before the methodology is due for deployment. We plan to keep working on statistical methods to manage a potentially large number of significant discoveries while maintaining high rates of outbreak detectability. We will also continue research into facilitating automated explanations of the discoveries and into designing detectors tailored to detect specific types of events. The latter task can be accomplished by either constructing detectors by hand, based on experience of domain experts, or by using machine learning to automatically train them on labeled historical data. The right choice may be to combine the two approaches, depending on the accessibility of the domain experts and the availability of the sufficient amount of labeled training data. In order to tackle those questions we plan to keep drawing from the extensive experience of our team gained in applied research into detection of events of importance to public health or food safety. We believe that the bio-surveillance framework is very relevant to prognostics for fleet management.

Acknowledgement

This work has been performed under the US Government contract (AFRL Agreement Number FA8650-05-C-7264: "eLog21 Collective Mind Experiment: Early Identification of Systematic Support/Process/Hardware-based Failures").

References

Bonissone, P.B., Varma, A. 2005. Predicting the Best Units within a Fleet: Prognostic Capabilities Enabled by Peer Learning, Fuzzy Similarity, and Evolutionary Design Process. *Proc. FUZZ-IEEE 2005*, Reno, NV, 2005. 312– 318.

Dubrawski A. 2004. A Framework for Evaluating Predictive Capability of Classifiers Using Receiver Operating Characteristic (ROC) Approach: A Brief Introduction. Auton Lab Technical Report, Carnegie Mellon University, August 2004.

Dubrawski A., Elenberg, K., Moore, A., Sabhnani, M. 2006. Monitoring Food Safety by Detecting Patterns in Consumer Complaints. *Industrial Applications of Artificial Intelligence IAAI'06*, Boston, USA, July 2006.

Fitzsimmons, J.A., Fitzsimmons, M.J. 1997. *Service Management: Operations, Strategy, and Information Technology, Second Edition.* Irwin McGraw-Hill, Boston, MA, 1997.

Neill D. and Moore A. 2004. Rapid detection of significant spatial clusters, *In Proceedings of the 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 256-265.

Roure J., Dubrawski A., Schneider J. 2007. A Study into Detection of Bio-Events in Multiple Streams of Surveillance Data. In D. Zeng et al. (Eds.): *BioSurveillance 2007*, Lecture Notes in Computer Science 4506, Springer-Verlag 2007.

Sabhnani M., Moore A., Dubrawski A. 2007. Rapid Processing of Ad-hoc Queries against Large Sets of Time Series. *Advances in Disease Surveillance 2*, 2007.

Sabhnani, M., Neill, D., Moore, A., Dubrawski A., Wong, W.-K. 2005. Efficient Analytics for Effective Monitoring of Biomedical Security, *In Proceedings of the International Conference on Information and Automation*, Colombo, Sri Lanka, December 2005.

Schwabacher, M.A. 2005. A Survey of Data-Driven Prognostics. *AIAA Infotech@Aerospace Conference*,, Accessed on 08/30/07: http://ti.arc.nasa.gov/people/schwabac/AIAA-39300874.pdf Arlington, VA, 2005.

Tsui M., Espino J., Wagner M. 2005. The Timeliness, Reliability, and Cost of Real-time and Batch Chief Complaint Data. RODS Laboratory Technical Report, University of Pittsburgh, 2005.

Wagner M.M., Moore A.W., Aryel R.M., editors, 2006. *Handbook of Biosurveillance*, Academic Press 2006.

Wagner M., Tsui M., Espino J., Dato V., Sittig D., Caruana R., McGinnis L., Deerfield D., Druzdzel M., Fridsma D. 2001. The emerging science of very early detection of disease outbreaks. *Journal of Public Health Management Practice*, 7 (6):51-9, 2001.

Wasserman L. 2004. *All of Statistics: A Concise Course in Statistical Inference*, Springer-Verlag, 2004.

Wong, W.-K., Moore, A., Cooper, G., Wagner, M. 2003. What's Strange About Recent Events, *Journal of Urban Health*, 80: 66-75.

Xue, F., Bonissone, P., Varma, A., Yan, W., Eklund, N., Goebel, K. 2007. An Instance-Based Method for Remaining Useful Life Estimation for Aircraft Engines. *61ST Meeting of the Society for Machinery Failure Prevention Technology Theme—Integration of Machinery Failure Prevention Technologies into Systems Health Management*. Virginia Beach, VA, 2007.