Summary of Biosurveillance-relevant technologies

Andrew Moore

School of Computer Science
Carnegie Mellon University
awm@cs.cmu.edu

Greg Cooper **Rich Tsui** **Mike Wagner**

Center for Biomedical Informatics
University of Pittsburgh
{gfc,tsui,mmw}@cbmi.upmc.edu

This short report, compiled upon request from Dave Siegrist and Ted Senator, surveys the spectrum of technologies that can help with Biosurveillance. We indicate which we have chosen, so far, to use in our development of analysis methods and our reasons.

1 Time-weighted averaging

This is directly applicable to a scalar signal (such as "number of respiratory cases today". This method, more commonly used in computational finance, simply compares the count during the current time period with the weighted average of the counts of recent days. Exponential weighting is typically used, where the half-life is known as the "time window" parameter. This time-window parameter is typically chosen by hand. We prefer the Serfling and Univariate HMM methods described below.

2 Serfling method

This method (Serfling, 1963) is a cyclic regression model, and is the standard CDC algorithm for flu detection. It is, again, applicable to scalar signals. It assumes that the signal follows a sinusoid with a period of one year, and thus finds the four parameters a, bc and d in

$$y_t = a + bt + ccos(t/365.25 + d) + \epsilon_t \quad (1)$$

where the parameters are chosen to minimize the sum of squares of residuals $\sum_t \epsilon_t^2$. It is an easy matter of regression analysis to determine, on any date, whether

y_t is sufficiently large that the probability of such a large y_t in the absence of an epidemic (the p-value) is so low that an alarm should be signaled. Rich Tsui has investigated this method extensively (Tsui, Wagner, Dato, & Chang, 2001) and it is used within the current RODS system.

3 ARIMA model

This (along with its simpler cousin, the ARMA model) is the staple method (Hamilton, 1994) of understanding time series data, though it is primarily used for prediction instead of signaling alarms. It asks the question "what is the signal likely to be in the current time period based on recent time periods". The "AR" in "ARIMA" refers to Auto-regression: the idea that we'll try to predict y_t as a linear function of $y_{t-1}, y_{t-2}, \ldots y_{t-p}$ for some time-window p. The parameters of this linear function can be learned by linear regression over large amounts of historical data. The "IMA" part of ARIMA comes from the correction that we should take into account that $y_{t-1}, y_{t-2}, \ldots y_{t-p}$ are themselves noisy values. It can be turned into an alarm-signaller in a similar manner to the Serfling method: do a statistical test to see how surprised we are by the recent value of our signal compared with the mean and variance of our prediction. Rich Tsui has implemented this within the RODS system.

4 Combining ARIMA with other factors

ARIMA will not, by itself, model seasonal effects in the data in the same way that Serfling would. It is not hard to get the best of both worlds by adding a sinusoidal term to the regression. But we can go further than that. For example we could add a day-of-week-specific term, or other terms to compensate for effects that would otherwise distort the predicted signal. (Remember, distorting the predicted signal could cause false positives or negatives if the basis of sounding an alarm is comparison between the predicted signal and the actual signal). Based on the DARPA Biosurveillance Kickoff meeting, I believe this combination ARMA/regression approach is being used by members of Howard Burkom's team at JHU/APL. We think it is an excellent approach, but we do not plan to duplicate effort within the DARPA program by reproducing it (though we will likely trade ideas and code with Howard's team).

5 Univariate HMM

This approach uses Hidden Markov Models (Rabiner, 1989; Moore, 2001b) with discrete hidden state and a continuous output signal. The hidden state is whether or not there is an epidemic of some disease (e.g. influenza) in progress. There can be more than two possible states if we allow more than two levels of epidemic, or if we allow epidemics of different diseases. The parameters of the HMM can be learned from data and, as before, the probability distribution of today's signal can be predicted and compared with today's actual value. A great merit of this approach is that it can easily allow non-Gaussian noise, and even multimodal distributions for the signal. Rich Tsui has implemented and tested this method extensively within RODS and favors it as a practical detector.

6 Kalman filter used as an adaptive ARIMA model

Loosely speaking, this is similar to the HMM approach, but in which the hidden state is a continuous variable. Although Kalman filters (Hamilton, 1994) are primarily used as state estimators for control systems it is quite easy to use them as underlying state estimators for a signal such as "rate of recent respiratory problems". And again, it is then easy in principle (though fiddly in practice) to learn the parameters from historical data, and signal alarms when today's observation falls outside the model's expectations. Rich Tsui has implemented this but prefers the HMM approach.

7 Recursive-least-square (RLS) adaptive filter

This is another AR regression method, implemented in an efficient way using some nifty linear algebra tricks. Apart from computational speed, which is not important for scalar problems with merely tens of thousands of datapoints (which is what we are dealing with with hourly signals over a period of years), it is hard to distinguish the advantages and disadvantages of this from ARIMA-type methods.

8 Support vector machine (SVM)

This is a non-linear AR method in which today's signal is predicted as a non-linear function of recent days' signals. SVM (Burges, 1998; Moore, 2001c) could have been replaced by any of dozens of machine learning algorithms that can predict real-valued outputs based on inputs. Rich Tsui has implemented this method, but we find no particular advantage compared with other methods. (Note that traditionally SVMs are used for classification instead of regression, so additional pain is needed to apply them to this problem).

9 Artificial neural networks (ANN)

Neural networks can be used in an AR-like fashion, or by numerous other methods such as recurrent connections, to model time series with nonlinear dynamics. They require careful attention and tweaking. Rich Tsui has implemented this approach but we do not make much use of it.

10 CuSUM method

This econometric method (Bos & Fetherston, 1992) tests to see whether there has been a recent change in the behavior of a time series. Unlike most of the above methods it can detect changes in variance instead of merely changes in value. It will, however, signal a false alarm if changes occur that other methods such as HMM or ARIMA would not react to, such as a steady rise in the mean. Rich Tsui has implemented it but does not find it very useful.

11 Randomization tests

Many of the above methods can precisely quantify, for every test, what is the level of anxiety (formally, what is the p-value...the chance that we'd have seen a signal as strong as this if the process was proceeding as normal). But in some of the more powerful tests below, a correct p-value is hard to find because of the effect of *multiple testing*. If I do 100 tests instead of 1 test, and I signal an alarm if any one test reveals an anomaly, then my chance of false positives will probably go up considerably—up to 100-fold depending on the degree of independence

between then tests. This is very serious: if a surveillance system were installed at all of 100 military bases, for example, there would probably be one or two false alarms every day, and any central monitoring agency would find it hard to believe subsequent alarms. The extent of my danger of false alarms depends entirely on how strongly correlated my multiple tests are. A conservative way of dealing with this is the Bonferroni correction. But that can be so conservative that it allows an unacceptable chance of a false negative. An alternative is to simulate 1000 samples of the data from the recent picture and perform the multiple tests on each of those samples. Then we can ask the question: on what fraction of simulated runs did the surprise level of the loudest alarm beat the surprise level of the loudest alarm on the real data? That fraction is a very good approximation for the true p-value of today's data. This technique, known as *randomization-based testing* (Efron & Tibshirani, 1993) is employed extensively in our own systems, and is also used in spatial scan statistics and WSARE described below.

12 Spatial Scan Statistics

This method (Kulldorf, 1997) searches for geographical overdensities of disease cases. It looks for small geographical regions (usually circular) in which the fraction of the population with some disease is significantly higher than in general. By using sensible statistics and randomization tests, great care is taken to ensure

- We don't get fooled by multiple testing (randomization is the most common defense against that).
- We don't get fooled by general population overdensities.

A more sophisticated kind of scan statistic also looks for overdensities in time: it searches the joint space of local geography and recent time windows to find out whether there's been a recent upswing in disease levels. JHU/APL, under the guidance of Dr. Howard Burkom, has been extensively investigating this approach. We (Pitt/CMU) also plan to use this approach and, projecting out from preliminary discussions with Dr. Burkom in Feb 2002, will probably do this in collaboration with Dr. Burkom instead of as a separate endeavor.

13 Bayesian Networks

Given a multidimensional database with many attributes, Bayesian networks (Cooper & Herskovits, 1992; Cooper, 1995, 1999; Pearl, 1996; Moore, 2001a) are a popular and well-principled way of concisely representing the interrelationships between the attributes. They give a succinct model of the joint probability density function from which the attributes are drawn. Common extensions include Dynamic Bayesian Networks, which additionally model the relationship between the attributes in this time period to those in the previous time period. We make extensive use of Bayesian networks in our approaches to Biosurveillance, including their use within WSARE and PANDA (described below).

14 Contingency Table Analysis

This tool is the primary workhorse of traditional epidemiology. Table 1 shows an example 2-by-2 table analysing the number of thirty-somethings who have needed treatment today compared with a previous time-period. Contingency table analysis asks the question: if there was really no relationship between thirty-somethings and the time they choose treatment, what is the chance we'd have seen a distribution that skewed as much as, or more than, this? This question *could* be answered by an extensive randomization test, but happily there is an exact analytical way to compute this chance (which is a p-value). The method is known as the Fisher Exact Test (Good, 2000). Running Fisher's Exact Test on Table 1 yields a score of 0.00005058, which indicates that the count C_{today} for cases matching the rule Age_Decile $= 3$ are significantly different from the count C_{other}....there's only a 5-in-100,000 chance we'd have seen a distribution this extreme under the null hypothesis.

Table 1: A sample 2x2 Contingency Table

	C_{today}	C_{other}
$Age\_Decile = 3$	48	45
$Age\_Decile \neq 3$	86	220

We (and almost everyone else) use contingency table analysis and Fisher's

exact test in many subcomponents of our detection system.

15 Scalar Outlier detection

Outlier detection is simple. From historical data we learn a probability density function of the signal we are monitoring. Then on each time period we signal a warning if the current signal is in an area with very low density. This is a simple approach often used in Statistical Quality Control (SQC) but it is unlikely to be useful to us compared with the earlier ARIMA or Scalar HMM methods.

16 Anomaly detection

In exactly the same general approach as outlier detection it is possible to model the joint distribution of a set of variables (see, for example (Eskin, 2000)). For example, we could learn the joint distribution over all features of a case admitted to an ED. Then, for each case, we can compute the likelihood of that particular case and compare it to the likelihood of other cases. We can then signal an anomaly when we see extremely strange cases. This method has been used very successfully for detecting strange galaxies for astrophysicists, and for detecting errors in consumer marketing databases. It has also been used for network intrusion detection, where the extent of its success is debatable. It is possible that it might be useful for biosurveillance but it is not a high priority for us because we anticipate a greater likelihood of attacks being discovered by strange commonalities between multiple cases (see WSARE below, or Spatial statistics above), or by cumulative evidence from multiple cases (see PANDA below)

17 Change-point statistics

There is a branch of statistics (Carlstein, 1988) that considers a time series of (usually scalar) signals in which each is generated independently from some fixed but unknown distribution F until a certain unknown time T after which they may instead be generated independently from some fixed but unknown distribution G. This technology is good at finding out if such a change ever occurred and when it occurred as accurately as possible with very few assumptions about the

nature of the distribution. We believe this could be relevant technology, though the assumptions it is built upon are still idealized compared with our scenario.

18 FDR tests

The False Discovery Rate statistic is a relatively new approach to compensating for testing multiple hypotheses. It is

- ...much more computationally efficient than randomization, though sometimes not as accurate.
- ...somewhat computationally slower than the traditional Bonferroni approach, but much more accurate.

The False Discover Rate (FDR) method (Benjamini & Hochberg, 1995; Miller, Genovese, Nichol, Wasserman, Connolly, Reichart, Hopkins, Schneider, & Moore, 2001) guarantees that the fraction of the number of false positives over the number of tests in which the null hypothesis was rejected will be no greater than α for a user-specified choice of α. We use FDR instead of randomization in cases where randomization would be impractical. One example is "outer-loop" cases of multiple testing, such as when we report on historical data which of our earlier dates had significant events.

19 Kd-trees

These are classical geometric data structures (Friedman, Bentley, & Finkel, 1977; Bentley, 1980; Omohundro, 1987; Moore, 2001d) for retrieving nearby points efficiently, and finding all individual points within a certain distance from a query quickly. These are not of much use as statistical techniques in themselves, but they are very useful for accelerating spatio-temporal queries in which we need to quickly retrieve a set of cases from some region at some time. There are also disk-resident versions of the same structures, which have attractive paging properties. These include R-trees (Guttman, 1984).

An extension of kd-trees are *Multi-resolution kd-trees.* Kd-trees have traditionally been used for the efficient retrieval of data points nearest a query point. By adding sufficient statistics about all the data points (Priebe, 1994; Deng &

Moore, 1995) below each node in the tree it is possible to speed up other operations, such as locally weighted regression (Moore, Schneider, & Deng, 1997) and mixture model based clustering and density estimation algorithms (Moore, 1999). At each node during the search, the algorithm considers whether it 1) can ignore all the points below that node because they are irrelevant to the current query, 2) can estimate the effects of all the points below that node without visiting them individually, or 3) must recurse further down the tree. The result is that most queries need only visit a small number of nodes in the tree.

We have not yet used kd-tree technology in the biosurveillance work, but we intend to do so in two ways.

- We will try a mixture of a Gaussian and Background distribution, as described in the kickoff meeting, to identify compact regions with increased prodrome levels.
- We plan to collaborate with Dr. Burkom's group to use kd-trees to scale up spatial scan statistic computations.

20 All-dimensions trees

Many data mining algorithms are built upon making a huge number of queries to the data of the form: how many records are there with attribute x has value y and attribute z has value a, etc. When the number of records in the data set becomes large, the cost of answering these queries can dominate the computation involved. Algorithms that rely on this include rule learning, decision trees, Bayes net learning, and frequent sets. An all-dimension tree (Moore & Lee, 1998; Anderson & Moore, 1998; Komarek & Moore, 2000) builds a cache that efficiently stores these answers in a form that allows the queries to be answered without going back to the data. The result is that after the cache is built, the queries are answered in computation time independent of the number of records in the data set. We will use this to scale up WSARE to allow it to run on surveillances that are receiving thousands up to millions of new observations a day.

21 WSARE

WSARE (Wong, Moore, Cooper, & Wagner, 2002) is a generic detection system built specifically for quick integration into data sources of the kind being investigated in the DARPA Biosurveillance program. It looks for answers to the question "What's Strange About Recent Events" in which no individual record might look anomalous yet there has been a significant change in some aspect of the multi-dimensional interrelationship between all the features being monitored. WSARE is designed to do a different and complementary task to the traditional detection systems which focus on one attribute of the data, typically a daily aggregate count such as the number of respiratory cases in a certain day. The traditional systems will be much more sensitive to the specific diseases they are trying to detect, but may be less sensitive to an insidious disease that only affects a specific group while not causing enough of a perturbation in the daily aggregate count of the monitored symptom. For example, suppose that a disease causes elderly males from a certain neighborhood to be sick but the number of cases caused by this disease is not sufficient to skew the daily aggregate counts above an alert threshold.

WSARE approaches this problem using a rule-based anomaly detection approach. This system searches for irregularities in the data using rules. An example of such a rule would be "Prodrome = Respiratory AND Home Zipcode = 84102". This rule indicates that WSARE is determining if the number of cases with respiratory problems in the 84102 zipcode area are unusual or not. On each day, WSARE looks at the current 24 hours worth of data and compares it against the past week's worth. WSARE considers all possible rules and selects the most statistically significant rule for the current 24 hours, using a randomization test to guard against multiple-hypothesis testing errors. This rule is printed as the final result, along with a message stating if the anomalous pattern, which corresponds to the rule, should be interpreted as significant.

The current version of WSARE under development uses a Bayesian Network to model what distribution of cases we should expect to see to day based on known facts (e.g. day of week, time of year, recent FLUSTAR reports for the city) so that it can more accurately determine deviations from the predicted background. The Bayes Net is learned from historical data.

WSARE is different from the above methods primarily because it inspects the interrelationships of very many attributes simultaneously instead of monitoring one individual signal. It is complementary to PANDA, below, because PANDA begins with a many-attribute investigation of how surprising each case is and then

very carefully analyses the degree of surprise about the recent pattern of surprising cases: PANDA has the potential to spot a problem based on a small handful of very specific cases whereas WSARE screens for problems detectable by a large scale change over a sizable sub-population of cases.

22 PANDA

We are in the process of developing the PANDA system (Patient-based ANomaly Detection and Assessment). PANDA will contain probabilistic causal models of patient diseases, including diseases that are likely due to bioterrorism. For each patient, PANDA will dynamically construct a probabilistic causal model (Cooper, 1999; Cooper & Yoo, 1999; Cooper, 2000) using a Bayesian network representation. Such a patient-specific causal model includes variables that represent risk factors (e.g., infectious disease exposures of various types), disease states, and patient symptoms. Some of the variables in each patient-specific network are linked (via arcs) to population variables, such as a variable that represents spatio-temporal information about the release of an infectious agent. Also, if the disease state of a patient P is potentially influenced by another patient Q (e.g., if Q has a contagious disease and P was exposed to Q), then the causal model for P would have arcs into it from the model for Q.

The inter-linked patient-specific causal models will form a large causal Bayesian network that represents the entire population being modeled. PANDA will use such a network to infer the spatio-temporal probability distribution of disease for the population as a whole, as well as the probability distribution of disease for each patient in the population. None of the other approaches in this document (or elsewhere) construct a detailed spatio-temporal probabilistic causal model of the population and use that model to infer the disease status of (1) the population and (2) each member of the population.

The PANDA approach is powerful and general. There are two major challenges that we will face in implementing it: (1) assessing the causal models, and (2) achieving computational tractability. The former will require focusing on the most important causal relationships, rather than trying to be causally complete. The latter will probably require approximation algorithms for inference, including simulation algorithms.

23 FLUMOD (Spatial HMM)

In addition to analysing temporal sequences of univariate signals over a region with HMMs as described above, we can also analyse patterns and dynamics in space and time, by defining an HMM in which the hidden state corresponds to the cross-product of disease states in many neighborhoods. We will use this to further refine our models of the background distribution—what we "should" be seeing on each day in the absence of any new public health threat. This project is under way as of February 2002.

References

Anderson, B., & Moore, A. W. (1998). AD-trees for fasting counting and rule learning. In *KDD98 Conference*.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B, 57*, 289–300.

Bentley, J. L. (1980). Multidimensional Divide and Conquer. *Communications of the ACM*, *23*(4), 214—229.

Bos, T., & Fetherston, T. A. (1992). Market Model Nonstationarity in the Korean Stock Market. In *Pacific-Basin Capital Markets Research, Vol. 3*, pp. 287–301. Elsevier Science Publishers B. V. (North-Holland), Amsterdam.

Burges, C. (1998). A tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, *2*(2), 955–974.

Carlstein, E. (1988). Nonparametric Change-point Estimation. *The Annals of Statistics*, *16*(1), 188–197.

Cooper, G. F. (1995). A method for learning belief networks that contain hidden variables. *Journal of Intelligent Information Systems*, *4*, 1–18.

Cooper, G. F. (1999). An overview of the representation and discovery of causal relationships using Bayesian networks. In C. Glymour and G. F. Cooper (Ed.), *Computation, Causation, and Discovery. Menlo Park, CA*. AAAI Press and MIT Press.

Cooper, G. F. (2000). A Bayesian method for causal modeling and discovery under selection. In *Proceedings of the Confernce on Uncertainty in Artificial Intelligence*, pp. 98–106. Morgan Kaufmann .

Cooper, G. F., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data.. *Machine Learning , 9*, 309–347.

Cooper, G. F., & Yoo, C. (1999). Causal discovery from a mixture of experimental and observational data. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pp. 116–125. Morgan Kaufmann.

Deng, K., & Moore, A. W. (1995). Multiresolution instance-based learning. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, pp. 1233–1239 San Francisco. Morgan Kaufmann.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman and Hall.

Eskin, E. (2000). Anomaly detection over noisy data using learned probability distributions. In *Proceedings of the 2000 International Conference on Machine Learning (ICML-2000)* Palo Alto, CA.

Friedman, J. H., Bentley, J. L., & Finkel, R. A. (1977). An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software, 3*(3), 209–226.

Good, P. (2000). *Permutation Tests - A Practical Guide to Resampling Methods for Testing Hypotheses* (2nd edition). Springer-Verlag, New York.

Guttman, A. (1984). R-trees: A dynamic index structure for spatial searching. In *Proceedings of the Third ACM SIGACT-SIGMOD Symposium on Principles of Database Systems*. Assn for Computing Machinery.

Hamilton, J. (1994). *Time Series Analysis*. Princeton University Press.

Komarek, P., & Moore, A. W. (2000). A Dynamic Adaptation of AD-trees for Efficient Machine Learning on Large Data Sets. In Langley, P. (Ed.), *Proceedings of the 17th International Conference on Machine Learning*, pp. 495–502.

Kulldorf, M. (1997). A spatial scan statistic. *Communications in Statistics—Theory and Methods, 26,* 1481–1496.

Miller, C. J., Genovese, C., Nichol, R. C., Wasserman, L., Connolly, A., Reichart, D., Hopkins, A., Schneider, J., & Moore, A. (2001). Controlling the false discovery rate in astrophysical data analysis. submitted to AJ.

Moore, A. W. (1999). Very fast mixture-model-based clustering using multiresolution kd-trees. In Kearns, M., & Cohn, D. (Eds.), *Advances in Neural Information Processing Systems 10*, pp. 543–549 San Francisco. Morgan Kaufmann.

Moore, A. W. (2001a). A Powerpoint tutorial on Bayes Nets. Available from http://www.cs.cmu.edu/∼awm/781/timetable.html.

Moore, A. W. (2001b). A Powerpoint tutorial on Hidden Markov Models. Available from http://www.cs.cmu.edu/∼awm/781/timetable.html.

Moore, A. W. (2001c). A Powerpoint tutorial on Support Vector Machines. Available from http://www.cs.cmu.edu/∼awm/781/timetable.html.

Moore, A. W. (2001d). A tutorial on kd-trees. Available from http://www.cs.cmu.edu/∼awm/papers.html.

Moore, A. W., Schneider, J., & Deng, K. (1997). Efficient locally weighted polynomial regression predictions. In D. Fisher (Ed.), *Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 196–204 San Francisco. Morgan Kaufmann.

Moore, A. W., & Lee, M. S. (1998). Cached Sufficient Statistics for Efficient Machine Learning with Large Datasets. *Journal of Artificial Intelligence Research, 8*.

Omohundro, S. M. (1987). Efficient Algorithms with Neural Network Behaviour. *Journal of Complex Systems, 1*(2), 273–347.

Pearl, J. (1996). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Priebe, C. (1994). Adaptive Mixtures. *Journal of the American Statistical Association, 89,* 796–806.

Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. IEEE, 77*(2), 257–285.

Serfling, R. E. (1963). Methods for Current Statistical Analysis of Excess Pneumonia-Influenza Deaths. *Publich Health Reports, 78*, 494–506.

Tsui, F. C. R., Wagner, M., Dato, V., & Chang, H. C. (2001). Value of ICD-9–Coded Chief Complaints for Detection of Epidemics. In *Symposium of Journal of American Medical Informatics Association*.

Wong, W., Moore, A. W., Cooper, G., & Wagner, M. (2002). Rule-based Anomaly Pattern Detection for Detecting Disease Outbreaks. Tech. rep. CMU-CS-02-106, Carnegie Mellon University, School of Computer Science.