Early Detection of Important Animal Health Events

J. L. Andrews[*], I. E. Diaz Bobadilla[†], Y. Huang[‡], C. Kitchen[‡], K. Malenfant[‡], P. D. Moloney[§], M. A. Steane[*], S. Subedi[*][¶], R. Xu[*], X. Zhang[‖], P. D. McNicholas[*], and J. Stockie[**]

PIMS-MITCAS Industrial Problem Solving Workshop,
University of Calgary, May 2009.

1 Introduction

The Alberta Agriculture and Rural Development (AARD) is designing and implementing real time animal health surveillance systems to support Alberta's livestock and food production industries. A surveillance network "AB Vet Surveillance Network" has been in place since 2005 which is continuously collecting and analysing data on AB livestock and poultry for the early detection of and rapid response to important public health events.

Information that was submitted to "AB Vet Surveillance Network" from 2005–2009 was available on a variety of variables including:

- Location (county and type);
- Type of operation (eg dairy, feed lot, etc.);
- Number of animals affected;
- Number of animals on the farm;
- Number of animals tested;
- Syndrome; and
- Diagnosis.

However, no incidence of major animal health events was reported.

The objective of the study was to optimize existing statistical methodologies for early detection of important animal health events. So, we proposed to tackle the problem using both univariate and multivariate approaches. Univariate approaches like Poisson Du Jour, recursive CUSUM, and ARIMA were performed. The problem was also addressed using multivariate ARMA and ARIMA. Some alternative approaches were also suggested. One desired property of the method was a low false discovery rate.

[*] Department of Mathematics & Statistics, University of Guelph.

[†] Department of Mathematics & Statistics, University of Regina.

[‡] Department of Mathematics & Statistics, University of Calgary.

[§] School of Mathematical & Geospatial Sciences, RMIT Melbourne.

[¶] Corresponding author. E-mail ssubedi@uoguelph.ca.

[‖] Department of Statistics, University of British Columbia.

[**] Department of Mathematics, Simon Fraser University.

2 Cleaning the Data

As is often the case with real data, a bit of preprocessing was needed for the AARD data. Most significant issues could be traced back to data entry errors, or at least suspected data entry errors. For instance, one entry noted that over ten-thousand cattle, 100% of the farm, had suffered heart failure. When trying to detect an outbreak in a population, entry errors of this nature need to be minimized.

We also detected some interesting quirks inside the data set. An example would be that all county codes used all five-characters regardless of the actual code length. In other words, attempting to look at county "MD60" would require calling upon "MD60␣" (where '␣' is a blank space).

Perhaps the most important issue to the task at hand, detecting outbreaks, is that there is a considerable lag between veterinary visits and data entry. At times there are lags of over a year — though this may be entry error as well. Recognizing that entry lag of as little as a few days could be significantly detrimental to early detection of an outbreak, this is a problem that needs to be addressed.

3 Spatiotemporal Analysis

When the problem was first proposed there was a discussion on using spatiotemporal analysis. In order implement this approach we narrowed our focus to six key indicators, namely: (x, y) coordinates; operation; diagnosis; frequency; and time. This created a six-dimensional problem so in order to simplify this we looked at individual or combinations of operations and diagnosis leaving us with a four dimensional problem to look at. A spatiotemporal analysis is a relating of both space and time and we wanted the space represented on a two-dimensional plane and time to be represented as a third dimension so our goal was to imbed the frequency of instances into this three-dimensional framework by assigning each coordinate in our three-dimensional view a number representing the frequency. Our goal was then to count the frequency contained within a sphere of constant radius for every location and time combination of the data. If the frequency breaks a yet to be determined threshold this could be a potential indicator of an outbreak.

The main problem encountered in this approach is the data did not include (x, y) coordinates and only specified location by county. We could not assign a numbering system to the counties since they are not of consistent size or relative position so we created a grid and assigned an (x, y) location based on the center of each county. If, however, we consider a sphere centered around a smaller county that barely touches a larger county we would find the frequency count to be unnecessarily inflated. To remedy this we considered assigning coordinates to the boundaries of the counties and include fractional frequencies in the total frequency count. This would then introduce error due to the fact that a larger county would most likely not have consistent population throughout. The difficulty here could be avoided if the data contained specific locations, such as GPS coordinates, of each farm. Unfortunately the effectiveness of this method could not be measured since the existing data was collected in such a way that we could not determine if there exists a threshold that could be used as an indicator.

Given we already had coordinates defined for every county used in the data we created an interactive graph that animated across time the number of instances at each location for fixed operations and diagnoses. This could not be used as a signal processing tool but shows the history of outbreaks over time.

4 'Poisson du Jour'

Another option was to attack the problem from a univariate stand point, where each symptom, disease, county and enterprise type has some validity. All symptoms and diagnoses are to be considered, but without any correlation matrix to show possible links between diagnoses it may be better to handle each independently. As we are interested in alerting authorities that an outbreak of some disease has started, it makes sense to create a system that generates a flag when the system is out of control. There are many ways a process may be out of control and we examine some of them.

Let us assume that the number of new cases reported of diagnosis type d, on a given day t, in a given area a, follows a Poisson distribution; that is $X_{a,d,t} \sim \text{Poisson}(\lambda_{a,d,t})$. We could then calculate the probability $P(X_{a,d,t} \geq x_{a,d,t})$, where $x_{a,d,t}$ is the actual observed number of new cases reported of diagnosis type d on day t in region a. Then, depending on the level of significance one is willing to assign — that is, on the probability of Type I error that one is willing to accept — any highly unusual or suspect results can be highlighted. A problem arises in the estimation of the parameter $\lambda_{a,d,t}$ (the true mean number of new cases). From the time series results it is clear that there is some seasonal component to the number of expected new cases. To account for this we shall estimate the mean for the current day by taking the average value for previous years around the same date, or

$$\lambda_{a,d,t} = \frac{1}{y(2w+1)} \sum_{j=1}^{y} \sum_{i=-w}^{w} x_{a,d,t+i-j},$$

where w is the width of the interval that we want to generate the estimate from (days before and after the current date) and y is the number of years of observed data you wish to use to form the estimate. This estimate is used to generate upper control limits, either using a Poisson method

$$UCL_{\text{Poisson}_{a,d,t}} = P(X_{a,d,t} \geq UCL_{Poisson_{a,d,t}}),$$

or using a normal approximation to the Poisson, as used in c-charts,

$$UCL_{\text{Approx}_{a,d,t}} = \lambda_{a,d,t} + 3\sqrt{\lambda_{a,d,t}}.$$

To demonstrate these approaches we have taken the data for the Lethbridge from the start of 2008 until May 25, 2009 and checked the number of times "BRD Complex" was given as the diagnosis as displayed in Figure 1. You can see that, given a level of significance of 0.001, three days would have triggered a flag using the approximate method, that is less than 0.6% days, whilst the Poisson method has generated no flags. This is partially due to the fact that the normal approximation to the Poisson distribution is not appropriate for small values of $\lambda_{a,d,t}$.

Alternatively, monitoring the recent number of days over the mean could show whether the system is out of control, possibly due to an outbreak of disease. Consider the case where, over the last n days, we wish to see if the number of days over the mean, $U_{a,d,t}$, is higher than expected. The associated probability could be approximated using a binomial distribution, $U_{a,d,t} \sim \text{binomial}(n, p_{a,d,t})$, given

$$p_{a,d,t} = P(X_{a,d,t} > \lambda_{a,d,t}).$$

An example of how this works, using the same diagnosis and county as before, is in Figure 2. You can clearly see that on three separate occasions over the roughly 18 month run a flag would have been triggered.

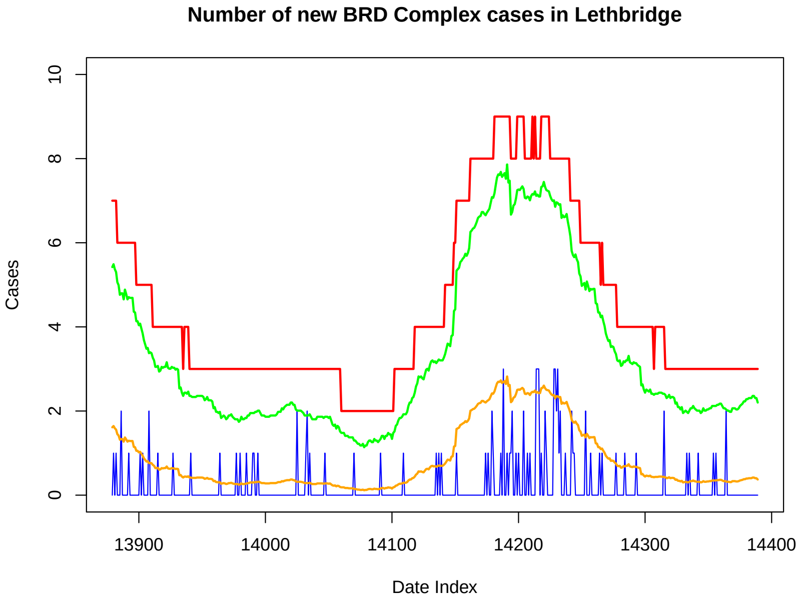

Figure 1: The number of new diagnosis of BRD Complex in blue, and the critical values using the Poisson method (red) and approximate method (green) from 1/1/2008 to 25/5/2009 in Lethridge county. The mean value, $\lambda_{a,d,t}$ is shown in orange.

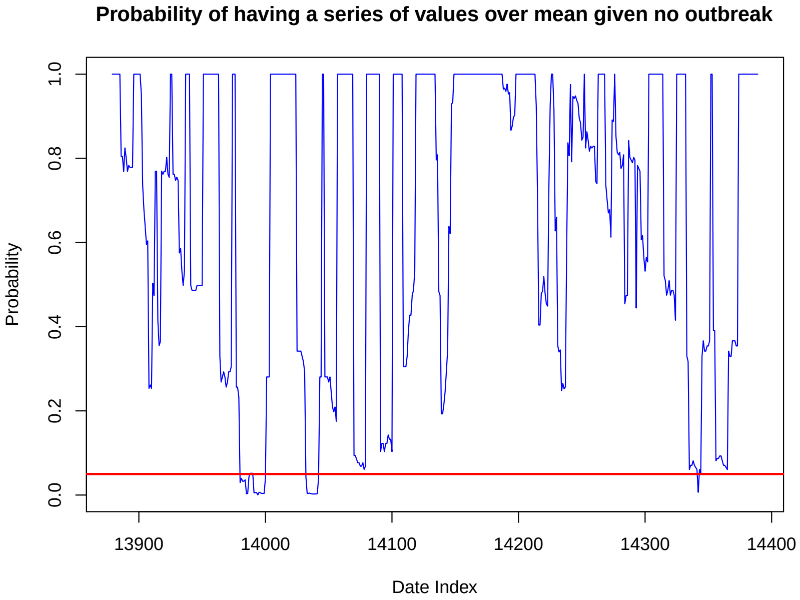

Figure 2: The probability that over ten days the number of days over the mean is given in blue, with the red line at the $\alpha = 0.05$ for days from 1/1/2008 to 25/5/2009.

These examples, while showing the process's functionality, may not be replicated for all counties and diagnoses. Where the number of cases is small, the number of false positives will increase. The more data that can go into calculating the estimates, the better the estimates should be. Hence, the longer the data is collected, the less variation in the estimates. To illustrate this point an analysis of Alberta as a whole for “Undetermined” diagnoses was carried out. The increased number of cases will hopefully improve the estimates for the mean number of new cases reported on a day. To justify this aggregation it is noted that with cattle being regularly transferred between properties across Alberta, spatial differences between herds does not prevent infections across Alberta.

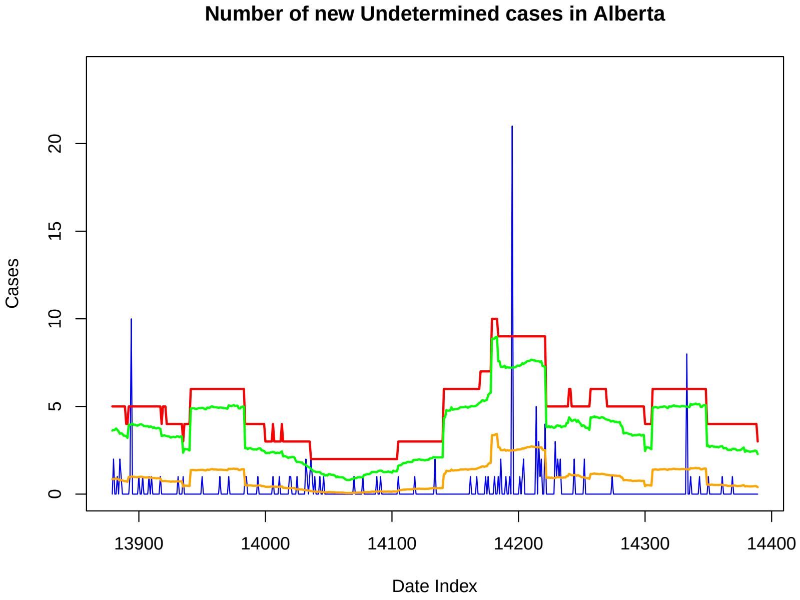

Figure 3: The number of new undetermined diagnoses in blue, and the critical values using the Poisson method (red) and approximate method (green) from 1/1/2008 to 25/5/2009 in across Alberta. The mean value, $\lambda_{a,d,t}$ is shown in orange.

From Figures 3 and 4 it can be seen that there are only three events that would arouse suspicion. Two of the points of interest in Figure 3 are points that are double the upper control limit, and would warrant further inspection. Whilst in Figure 4 the sections below the critical line relate to times where the $\lambda_{a,d,t}$ is very close to zero, highlighting an issue with this method when the expected number of new cases is very small.

In terms of validation, there is still the issue of whether these approaches are sensitive enough to flag an actual outbreak. Using them in combination, so that you could receive a flag on on day for more than one reason would be advisable. Other options for flagging unusual behavior in reporting numbers could be to look at: the number of consecutive days of non-decreasing reports; and the number of days above a bound higher than the mean. Any of these methods may be about to use fewer days to monitor if the process is under control. However, non-reporting days will need to be taken into account, as while no new cases may be reported, that does not mean that new cases have developed that are unreported.

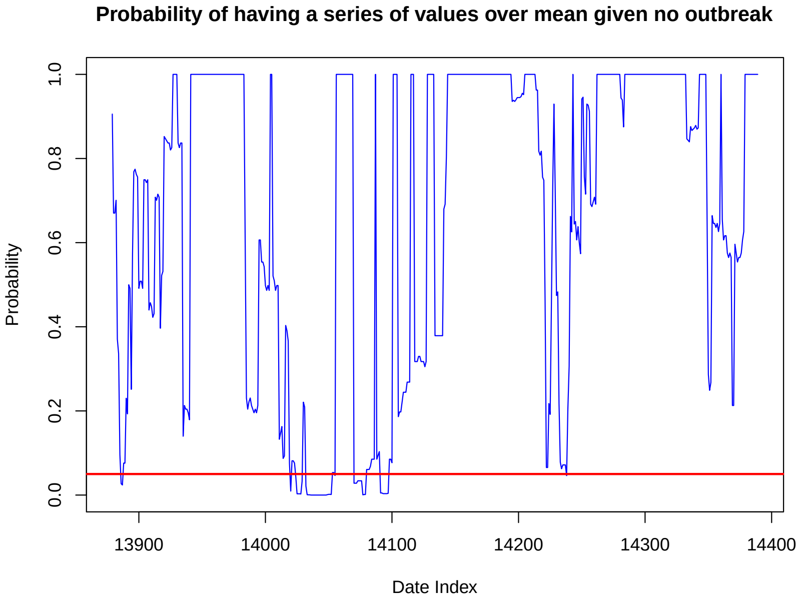

Figure 4: The probability that over ten days the number of days over the mean is given in blue, with the red line at the $\alpha = 0.05$ for days from 1/1/2008 to 25/5/2009.

5 Recursive CUSUM

In the absence of covariates and Poisson model parameters for both normal and outbreak situations, recursive CUSUM based on the cumulative sum of the recursive residuals was used. The basic underlying principle for detection of an outbreak is that if the cumulative sum of the residuals is outside the threshold, an outbreak is detected. The detector is written in the form:

$$S_0 = 0, \quad S_n = \max\left\{0,\ S_{n-1} + \log\left(\frac{f_{\theta_1}(y_n)}{f_{\theta_0}(y_n)}\right)\right\},\ n \geq 1.$$

An outbreak is detected at time $N = \min\{n : S_n \geq c\}$, where c is the user specified threshold. The algorithm for finding the recursive CUSUM was implemented in R (R Development Core Team, 2009) package `strucchange`.

In case of a potential outbreak, the increase in the number of occurrences of the disease will be observed not only on a single day but for a few days. A sudden high incidence rate for disease on a particular day might be a potential outlier. In Figure 5, a very high peak was observed at around day 100 which fell back to the regular incidence rate the very next day suggesting that it was a potential outlier. The CUSUM model did detect an increase however it was within the threshold suggesting that the CUSUM model seems able to detect potential outbreaks but is less sensitive to outliers resulting fewer false positives.

On the other hand, traits like reproduction possess a seasonal rhythm. For example, a seasonal calving has been practiced in most of the farms to better suit the climate as well as for better labour management. As a result, more incidences of reproductive syndromes are observed during the calving seasons than the rest of the year. However, as seen in Figure 6, increases due to seasonal trends were detected as potential outbreaks using the CUSUM model.

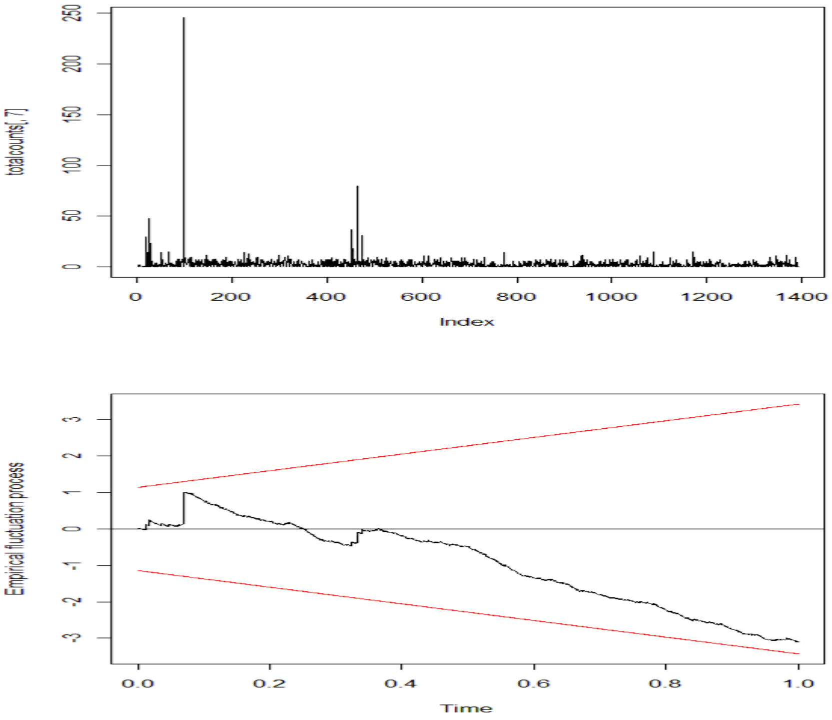

Figure 5: Observed total number of occurrence of neurological/recumbent syndromes at a particular day is shown at the top and recursive CUSUM of the residuals for neurological/recumbent syndromes in black, and along with threshold at 5% significance level denoted by red is shown at the bottom.

6 ARMA/ARIMA

6.1 Introduction

Yet another approach to these data is to fit time series models. A time series X_t is a set of data where X_t=$\{X_1, X_2, \ldots, X_t\}$ are realisations of X at times $1, 2, \ldots, t$. The data X_t is stationary if and only if, for any positive integer m and times $t_i : 1, \ldots, t_m$, the joint probability distribution of $\{X_1, X_2, \ldots, X_t\}$ is the same as $\{X_{1+s}, X_{2+s}, \ldots, X_{t+s}\}$ for all t and s; that is, the distribution of the time series is the same for all data sets of t terms in length.

A white noise process $X_t = \epsilon_t$ comes from a distribution with a mean of zero and a constant variance σ^2_ϵ and is stationary; the ϵ_t are also mutually independent.

ARMA model consists of two parts: AR (Autoregressive) and MA (Moving Average). In autoregressive model, each X_t is defined as a linear combination of its predecessors such that

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \epsilon_t,$$

where ϵ_t is a iid random disturbance/error with mean 0 and variance σ^2_ϵ and come from a white noise function.

In moving average model, each X_t is defined as a linear combination of its white noise processes and

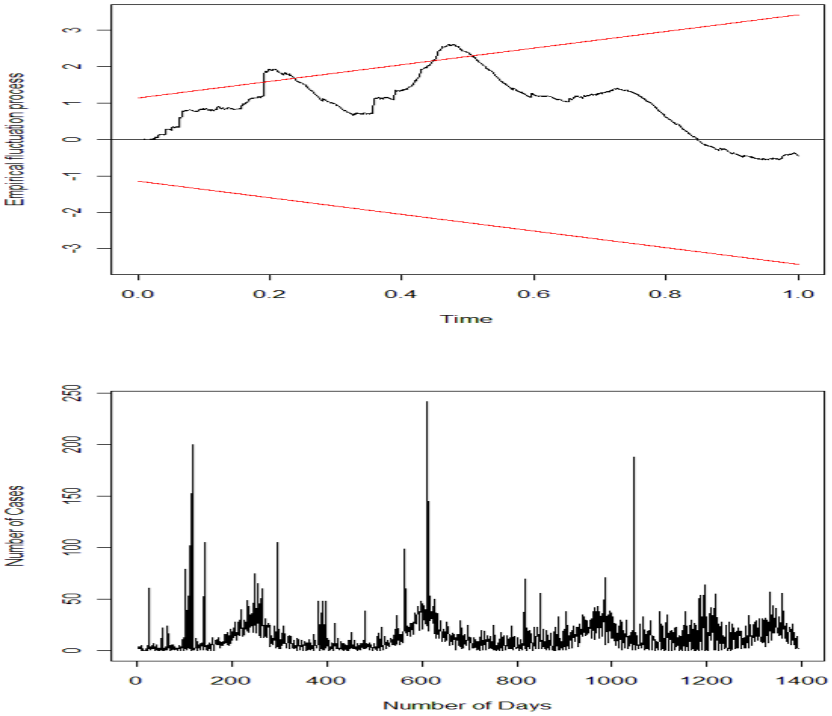

Figure 6: Observed total number of occurrences of Reproductive/Obstetrics syndromes at a particular day is shown at the top and recursive CUSUM of the residuals for reproductive/obstetrics syndromes in black, and along with threshold at 5% significance level denoted by red is shown at the bottom.

is given by:

$$X_t = \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \ldots + \theta_q\epsilon_{t-q}$$

where, ϵ_t's are iid with mean 0 and variance σ_ϵ^2 and come from a white noise function. The general MA(q) model stationary since the ϵ's are defined as stationary, and this model is a linear combination of ϵ's.

The autoregressive moving average model, or ARMA(p, q), combines both the AR and MA. For a time series X_t, given the white noise ϵ_t, the model is written in the form

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + \epsilon_t + \theta_1\epsilon_{t-1} + \ldots + \theta_q\epsilon_{t-q},$$

where, p is order of the autoregressive process and q is the order of the moving average process. When the time series of data X_t is not stationary, the autoregressive integrated moving average model, or ARIMA(p, d, q), which is an extension of the ARMA model, is used where model is build on the number of differences between two time points to achieve stationary. In ARIMA(p, d, q), similar to p and q are orders of AR and MA processes respectively, and d is the number of differences taken to achieve stationary.

6.2 Univariate ARMA

6.2.1 The Model

Given a time series of data X_t, the univariate ARMA method allows us to uncover the hidden patterns in the data but also predicts future values in this series. The model consists of two parts, an autoregressive (AR) part and a moving average (MA) part and it is given by

$$X_t = \alpha + \varepsilon_t + \sum_{i=1}^{p} \phi_{t-i} X_{t-i} + \sum_{i=1}^{q} \theta_{t-i} \varepsilon_{t-i},$$

where $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ with σ^2 is the variance and:

- α is a constant (intercept);
- p is order of the autoregressive process and q is the order of the moving average process;
- ϕ_i is the parameter of the autoregressive process; and
- θ_i is the parameter of the moving average process.

This model is denoted as ARMA(p, q).

6.2.2 The Approach

The goal is to develop an approach to optimize the sensitivity detection. Since the data is associated with time, ARMA would be applied to detect a possible outbreak.

- Calculate the time as the difference between the visited date and the earliest visited date.
- Calculate the affected rate which is the number of affected animals over the number of animals on that affected farm.
- Apply the square root transformation to the affected rate.
- Apply the univariate ARMA model.

In our data, there is more than one observation in some days. However, since there is only one observation for each single time in the univariate ARMA model, we have to get one and only one response for each day. In order to consider more information, we calculate the average affected rate for each day as our single observation for each time unit in the ARMA model. Then Figure 7 suggests to apply ARMA(7,0). A time series plot of the historical average affected rates as well as the predictions and its confidence intervals for the next 60 days is shown in Figure 8. The upper bound of confidence interval is about 5% in this case. Once the average of affected rate in the next 60 days is above this upper bound, an alarm is sounded. This approach has to update as long as you have more historical data, and the predictions and the upper bound might change depending on the update data.

We applied the same approach using the submission date instead of the visited date and the results were similar. There are possible data entry errors found in the data. For example, there are cases with 100% affected rate and only one animal on farm. Obviously, this is going to influence the average affected rate per day which will definitely cause false positives. As well, this approach is not able to take enough information into account; for example, the effect of counties and different combination of syndromes. We can apply the same approach to each selected combination which includes only one county, one operation, one syndrome and clinical diagnosis, but there will be more than 50,000 such combinations in this data. Therefore, it is not a good idea to use a univariate ARMA model to detect the outbreak of this data.

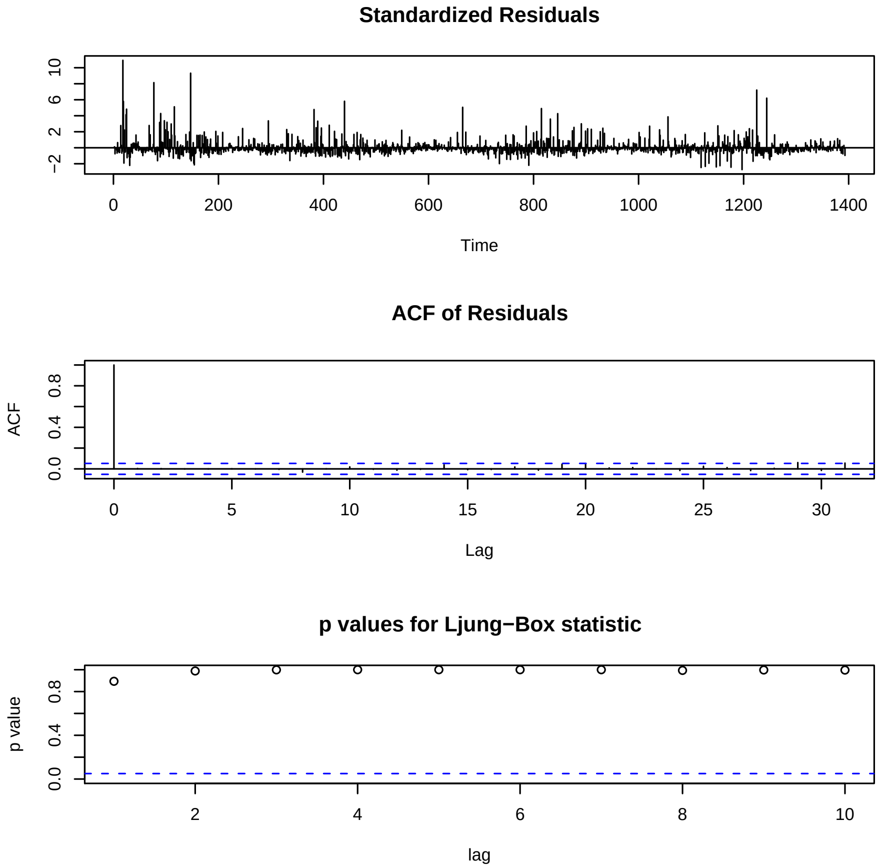

Figure 7: Diagnosis plots for ARMA(7,0).

6.3 Multivariate ARMA

6.3.1 The Model

Sometimes the occurrence of one syndrome might have an influence on the other syndrome. Taking into account that the syndromes are not independent of each other, a multivariate ARMA approach was proposed. Given a time series of data X_t, the multivariate ARMA(p,q) model is given by:

$$X_{jt} = \alpha_j + \varepsilon_{jt} + \sum_{i=1}^{p} \phi_{j,t-i} X_{j,t-i} + \sum_{i=1}^{q} \theta_{j,t-i} \varepsilon_{j,t-i}$$

where $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ with σ^2 is the covariance matrix for syndrome j, α is a vector of constants (intercepts), p and q are orders of the autoregressive and moving average processes respectively,and ϕ_j and $theta_j$ are the vectors of AR and MA parameters for syndrome j respectively.

6.3.2 The Approach

If using ARMA model:

- Calculate the standard deviation (sd) of the syndrome of interest over the past year.
- Today's predicted interval = Today's predicted value ± 2sd.

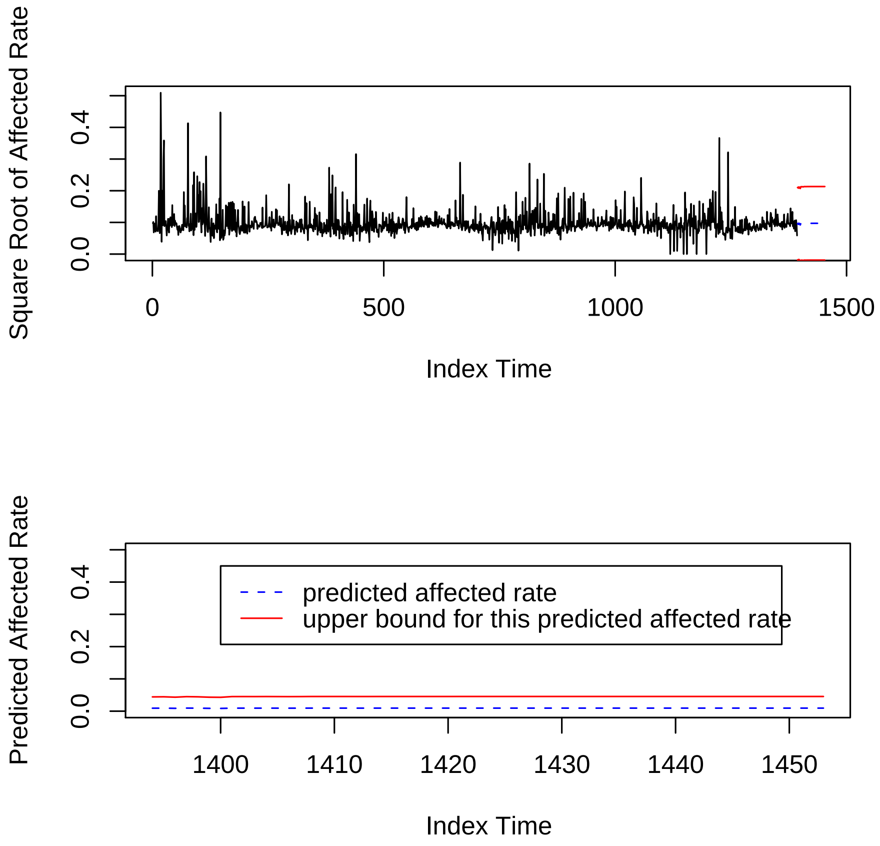

Figure 8: Prediction and confidence interval.

- Does today's actual value fall outside this interval? ⇒ Yes? ⇒ Investigate!

If using ARIMA model:

- Today's predicted value = Yesterday's actual value + Today's predicted difference.
- Today's predicted interval = Today's predicted value ± 2sd.
- Does today's actual value fall outside this interval?⇒ Yes? ⇒ Investigate!

In order to allow room for some flexibility and optimize the method individually for a particular syndrome, the time period over which the sd is calculated can be changed by the user. Also, the number of sd's used to calculate the prediction interval can be changed depending on the user's desired test sensitivity. The multivariate ARMA and ARIMA methods were implemented in the R package `dse`.

In our case, we created a thirteen-variable model estimated using the number of cases per day per syndrome as the response variable. For the multivariate ARMA, the resultant model was ARMA(6,0) aka AR(6) and for the multivariate ARIMA, the resultant model with first differences was ARIMA(6,1,0).

In Figure 9, the black line indicates the observed number of diseased animals at different time frames and the red line represents number of diseased animals at different time frames using the fitted multivariate ARMA model. The model seems to capture the trend very well.

Similarly, in Figure 10, the black line indicating the difference in observed number of diseased animals at two consecutive times and the red line representing the difference in diseased animals at two consecutive

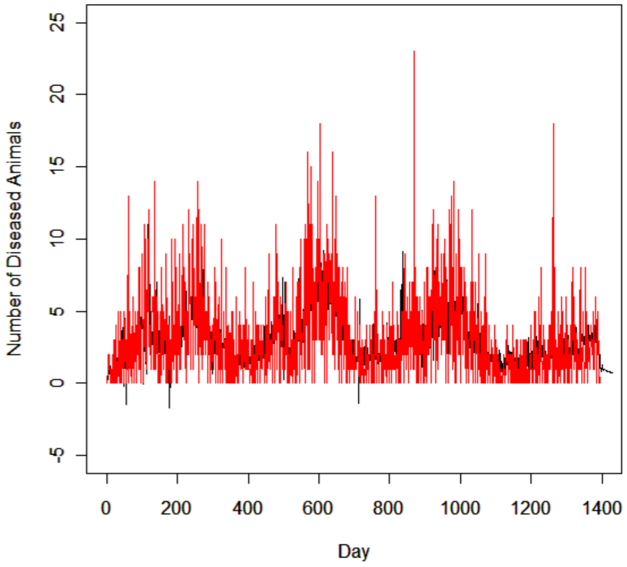

Figure 9: ARMA: neurological/recumbent syndromes.

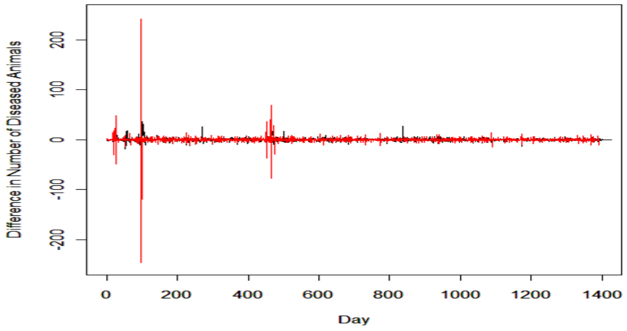

Figure 10: ARIMA: neurological/recumbent syndromes.

time frames using ARIMA overlapped quite a bit. In the presence of an increase in number of diseased animals, the method seemed to detect it.

Also in Figure 11, the black line and the red line indicated the observed number of diseased animals and the number of diseased animals using ARMA respectively. As mentioned in Section 5, the reproductive/obstetrics syndrome has a seasonal trend which was very well captured by the ARMA model. The increase in the occurrence of the disease as a result of seasonal trend was not flagged as an outlier reducing the cost associated with the false positives.

Also, Figure 12 shows that the seasonality associated with the reproductive/obstetrics syndrome was removed using the ARIMA model. Thus, overall, the multivariate ARMA and ARIMA model seemed to

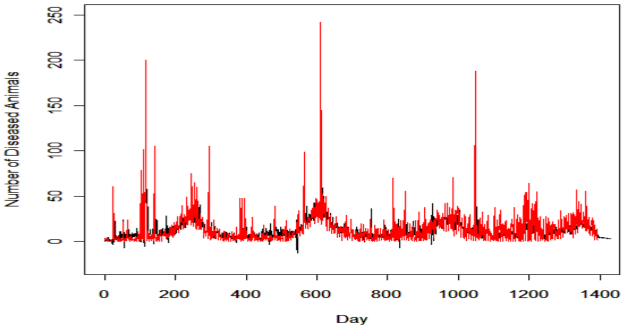

Figure 11: ARMA: Reproductive/Obstetrics Syndromes

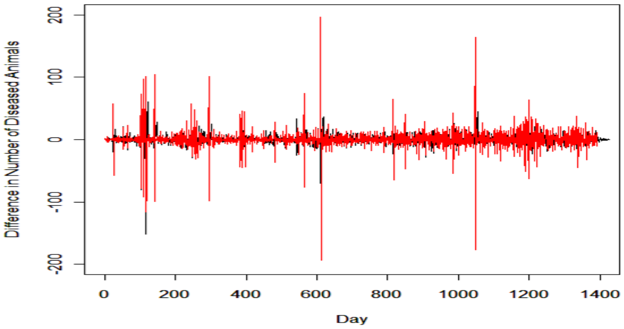

Figure 12: ARIMA: Reproductive/Obstetrics Syndromes

capture the seasonality very well. Theoretically, the detection of outliers can be done by calculating the prediction interval. However, due to the lack of incidence of an outbreak in the data set, the effectiveness of the model proposed could not be tested.

7 Other Potential Approaches

7.1 Other Options

We suggest three alternate approaches based on

- Dynamic specificity
- Fearnheads model
- Inhomogeneous Poisson model

None have been tested yet on the given disease data. Instead, the methods were illustrated with toy data. It is assumed that a real disease outbreak is associated with an increase in reported disease counts which must last for several days and the infectious diseases may exhibit some spatial clustering.

So, the general approach is:

- Report an outbreak when counts show significant increase over a short period of time.
- Define rules for detecting events which dynamically controls the specificity and sensitivity.
- Rules are updated daily using new data.

More counts were reported on weekdays than on weekends, so the Sunday data was deleted. Some seasonal effects were eliminated using a fitted Fourier spline (constrained B-spline is a better alternative). So it was assumed that after removing these patterns, any change in pattern will be related to an outbreak.

7.2 Dynamic Specificity

An outbreak was reported if there were more than K days in a 7-day window having more than C events. The specificity was given by:

$$\text{Specificity} = 1 - P(\sum_{i=1}^{7} (X_i > C) > K),$$

where X_i is the count of outbreak-free days in a 7-day window around day i. And the sensitivity was given by:

$$\text{Sensitivity} = P(\sum_{i=1}^{7} (Y_i > C) > K),$$

where Y_i is the count of days with outbreaks in a 7-day window around day i.

Specificity and sensitivity can be updated dynamically every day and the threshold C and K can be selected to achieve the best specificity and sensitivity. For each time series we can obtain a decision with known specificity and sensitivity. There are multiple time series in the data (e.g., 68 counties). which can be colored each county on a map to denote specificity and sensitivity (an idea borrowed from naive disease mapping).

7.3 Fearnhead's Model

Fearnhead (2006) presents an exact and efficient Bayesian inference for multiple changepoint problems where the changepoints correspond to outbreaks. The probability of changepoints is updated dynamically.

7.4 Inhomogeneous Poisson Model

A similar problem was solved in IPSW 2004 by testing an inhomogeneous Poisson model. To detect hacker attacks, multiple ports are monitored for frequency of visit. Similarly, multiple syndromes can be monitored for counts of disease.

8 Recommendations

There were notable differences in the results of analysis based on report dates as opposed to those based on actual dates on which the farm was visited. It was observed that sometimes the lag time between visit and submission was several days. For a communicable disease like flu which requires immediate response, a lag time of more than 10 days might rapidly increase the spread of the disease. In order to have an effective early warning system (≤ 3 days), the lag between visitation and submission needs to be reduced dramatically. Also, at the time of data submission, users could be asked to indicate whether or not the disease is communicable.

When an error is discovered at the AARD end (e.g. one case had 12,000 affected animals at one farm), it could be corrected in the data file to maintain the integrity of the data. Further consideration should be given to treating the data in a more natural way, instead of flagging it if the percentage infected is $>$ 5%.

As mentioned earlier, some traits have a seasonal pattern. Seasonality plays a significant role in occurrence of a disease and should be taken into account, but not for every diagnosis. Taking into account the seasonality will reduce the potential false positives and hence will be more time efficient.

From our study, we suggest that some combination of univariate and multivariate approaches will give much better results. In particular, multivariate ARIMA and Poisson approaches show definite promise.

References

Fearnhead, P. (2006). Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and Computing 16*, 203–213.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.