Early Prediction of Seasonal Influenza using School Absenteeism

by

Anu Stanley

A Thesis

presented to

The University of Guelph

In partial fulfilment of requirements

for the degree of

Master of Science

in

Mathematics and Statistics

Guelph, Ontario, Canada

© Anu Stanley, December, 2014

ABSTRACT

EARLY PREDICTION OF SEASONAL INFLUENZA USING SCHOOL ABSENTEEISM

Anu Stanley
University of Guelph, 2014

Advisors: Dr. Rob Deardon
Dr. Zeny Feng

Syndromic surveillance uses non-traditional health-related data to detect regularly occurring or emerging infectious disease outbreaks. A school absenteeism surveillance system was implemented by Wellington-Dufferin-Guelph Public Health (WDGPH) since February-2008 using an arbitrary 10% absenteeism threshold. The primary focus of this thesis is to refine the current methods to allow early detection of seasonal influenza outbreaks in the community. Surveillance systems were developed linking real outbreaks, defined by aggregated hospital data within the WDG area, to the school absenteeism data. We used the moving average (MA), exponentially weighted moving average (EWMA) and logistic regression (LR) to compute a unique baseline for each school on a given day and compared its false alarm rate (FAR) and accumulated days delay (ADD) to that of a steady baseline currently used by the WDGPH. This study concludes that the current methods of WDGPH appear insufficient in comparison to the surveillance systems implemented in this thesis.

ACKNOWLEDGMENTS

I wish to sincerely thank my advisors Dr. Rob Deardon and Dr. Zeny Feng, for their constant guidance and support throughout the study.

I am also thankful to the Wellington-Dufferin-Guelph Public Health for providing the data sets used in this study. I am especially thankful to Dr. Lise Trotz-Williams for the constant help in understanding the various aspects of the surveillance process.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Syndromic surveillance is a relatively new, but rapidly developing practice that uses non-traditional health-related data to detect regularly occurring or emerging infectious disease outbreaks [1]. These data types do not rely on a confirmed diagnosis but include behaviours that occur following the onset of disease symptoms, such as absences from work or school, over-the-counter drug sales and calls to health telephone help lines to name a few [1].

Syndromic surveillance has been recognized for its sensitivity or ability to detect actual infectious disease outbreaks earlier than the traditional methods of clinical or laboratory diagnoses [1;2]. However, these systems can generate frequent false alarms since an increase in sensitivity usually comes at the cost of decreased specificity; specificity being the ability to recognize when aberrations are not related to specific disease events [1;2]. Therefore, the perceived usefulness and challenge of a syndromic surveillance system lies not only in making early predictions, but also in finding an appropriate balance between true-positive and false-positive alerts [1;2].

School absenteeism surveillance is of particular importance to public health since children tend to be more susceptible to acute respiratory virus infections [1;2] and schools play an important role in the spread of influenza during an epidemic [1;3]. The school absenteeism surveillance program at Wellington-Dufferin-Guelph Public Health (WDGPH) has collected daily counts of students absent from schools in the WDG area since 2008. The threshold used for following up suspected infectious disease prevalence increases in schools has to-date been based on the arbitrary level of 10% absenteeism traditionally used by public health units and provincial agencies in Ontario. However, baseline levels of absenteeism vary from school to school. For some schools, this threshold leads to poor sensitivity, while in others it can lead to an unacceptable number of 'false-positive' alerts. In a study conducted in Quebec, the daily influenza related absenteeism threshold of 10% was found to only be able to raise the earliest alert at two weeks before the peak of the pandemic hospitalization, leaving insufficient room for intervention [4].

The primary focus of this thesis is to refine the methods used for analyzing WDG school absenteeism data in order to provide timely alerts and control the number of false-positive and false-negative absenteeism alerts. The successful completion of this project will improve WDGPH's ability to quickly identify expected outbreaks of seasonal influenza in the population. This will enable resources and possible control measures to be put into place more efficiently and potentially limit the course of the epidemic. In doing so, we explore different decisions pertaining to which schools to monitor, how to aggregate the data in time, choosing outbreak detection methods and how to evaluate their performances.

1.1 Data Aggregation

Decisions regarding data aggregation are important in reducing variability and when adjusting for time or resource constraints. These decisions included determining which schools to include in our analysis and how to group them in time and space. In the past, school absenteeism surveillance studies have been conducted with less than 10 schools [3] to over 100 schools [4-7]. Performing separate analysis on each school may allow greater sensitivity, however it could use more resources in investigating the resulting multiple alarms. In the above studies, most analyses were conducted at the individual school level [3-6] or aggregated at a district level [7]. However, a few studies grouped the schools by age group [5;7]. Aggregation with respect to time is either done by weekly [4] or (predominantly) daily [3;5-7]. Implications of the different levels of aggregation are considered and we attempt to discuss their relevance for the effectiveness of early outbreak detection.

1.2 Surveillance Systems

The current arbitrary threshold of 10% absenteeism used by WDGPH generates too many signals in schools that have an elevated average level of absenteeism or high noise, and too few signals in schools with low mean absenteeism. Consequently, a unique threshold can be determined for each school using individual school absenteeism statistics. However, this might still generate alarms too frequently as it does not account for the random fluctuations generated from, for example, field trip days, seasonal effects or day of the week effects. These challenges call for a procedure that can smooth out the day-to-day variation while remaining sensitive to important changes.

Statistical process control (SPC) methods from the field of manufacturing have often been applied in public health surveillance to detect changes in influenza activity over time [8-10]. Methods with memory such as the moving average (MA) process [10], especially the exponentially weighted moving average (EWMA) have a long history of application to problems in public health surveillance [8-11]. Regression methods are also widely used in outbreak detection and commonly a harmonic logistic regression [12] with sine and cosine functions are used to account for the seasonality of influenza [8]. These surveillance systems will be considered in this thesis.

1.3 Evaluation Metrics

To evaluate the performance of the surveillance systems, it is customary to use a combination of numerical indices; most commonly used are binary classification tests such as sensitivity and specificity. These indices are usually combined in the form of receiver operating characteristic (ROC) curve or activity monitoring operating characteristic (AMOC) curve to obtain a single performance measure which can be compared meaningfully across different surveillance systems [8]. However, these tests assume that data are independent and identically distributed and thus cannot be straightforwardly applied in the context of early outbreak detection since surveillance data are frequently auto-correlated. A few metrics exist that consider the auto-correlation but they assume that the outbreak is already occurring [8], whereas the focus here is to raise a signal before the outbreak for a timely public health intervention. Therefore, given the lack of appropriate evaluation methods, we derive a set of evaluation metrics that will measure the timeliness of alarms and the false alarm rate.

1.4 Overview of the Thesis

In this thesis, the various surveillance systems for outbreak detection will be compared and contrasted. In Chapter 2, we describe the data sets used and the preliminary analysis done to determine how the data should be divided for analysis. Chapter 3 outlines the three surveillance systems that will be applied: moving average (MA), exponentially weighted moving average (EWMA) and logistic regression (LR). We also detail the evaluation metrics that will compare the performance of each of the systems. All statistical analyses are implemented in R [13] and the results are presented in Chapter 4. In the last chapter, we summarize the findings of this thesis, and provide suggestions for future work relating to school absenteeism surveillance.

Chapter 2

Data Description

In this chapter, we describe the two types of routinely collected data used in this thesis and the preliminary analysis performed to inform data aggregation decisions.

2.1 School Data

Daily counts of absenteeism for up to fifty elementary and high schools in the WDG area has been collected by WDGPH for the period of February 20th, 2008 to April 30th, 2014. The schools were identified in the data only via anonymized ID numbers. These data contained information on numbers and percentage of students absent per day and the type of school (i.e. elementary or high school). Entries containing obvious typographical error (absenteeism $>$ 100%), and other extreme points (absenteeism $>$ 50%) were removed from this dataset before proceeding to the preliminary analysis.

2.2 Influenza Data

The second type of data obtained from WDGPH consisted of aggregated daily counts of laboratory-confirmed influenza in the WDG area for the same time period. These data indicated when WDGPH had first been informed of the case (i.e. when a lab report documenting that an individual had tested positive for influenza had been received).

2.3 Preliminary Data Analysis

As Figure 1 illustrates, absenteeism data were not available from the start of the influenza season for the 2007-2008 academic year. For this reason the 2007-2008 data were excluded from our analysis. Another problem can be seen in the 2009-2010 academic year where the first influenza case appeared well before the start of the school year. This was also the year with the pHINI outbreak and therefore did not reflect a standard "seasonal influenza" season. Therefore, for the purpose of this evaluation, the 2009-2010 academic year was also excluded and our analyses were based on the remaining five years.

The daily mean absenteeism level and standard deviation (SD) over the five-year period for all elementary schools was 5.51% (SD 3.18%), and 8.61% (SD 5.62%) for high school students. There was a statistical significant difference in the percentage of absenteeism between elementary school students and high school students with a p-value of less than 2.2e-16 under both Wilcoxon-Mann-Whitney and Kolmogrov-Smirnov tests. Therefore, separate analyses were

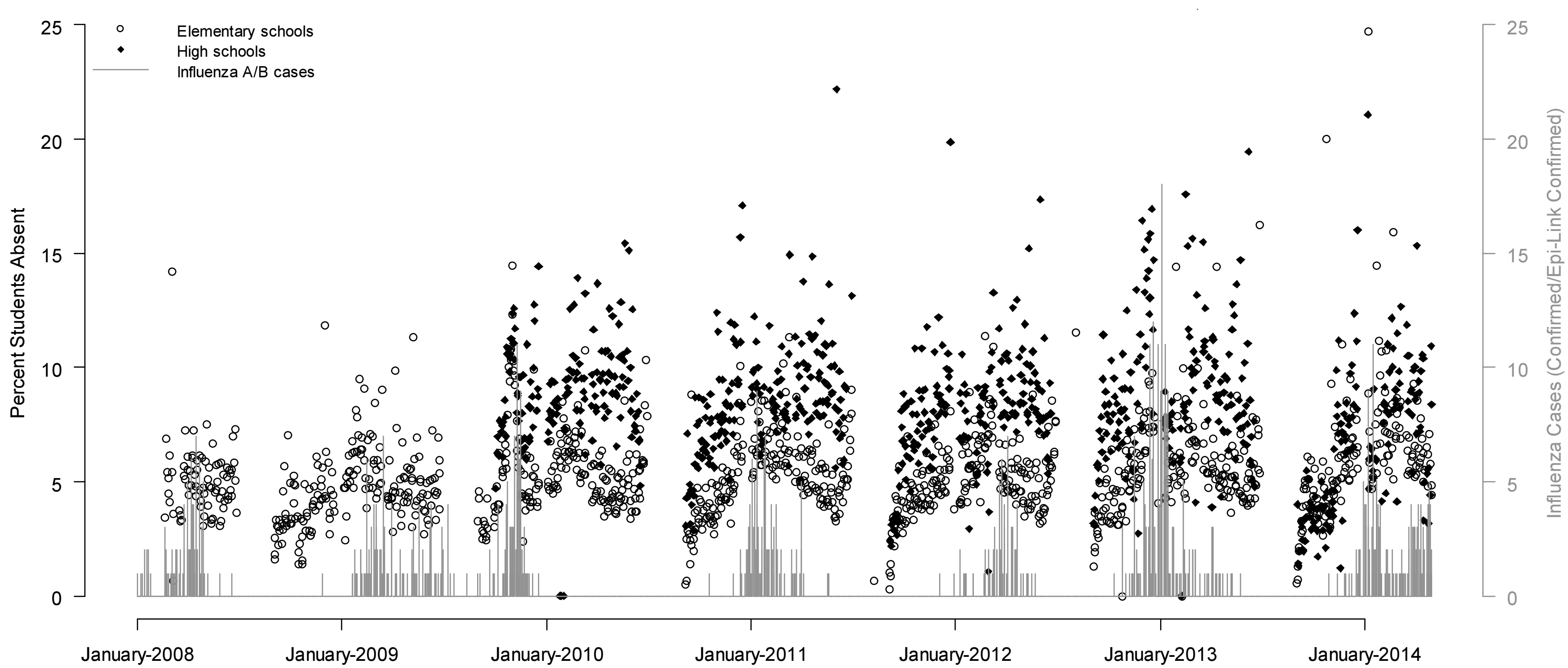


Figure 1: Average daily percent absent among elementary school and high school students in the WDG school system from 2008-2014 academic years. The grey lines indicate the numbers of confirmed cases of flu reported to public health on the given day.

performed for elementary and high schools to account for the age-specific differences.

In the school absenteeism data there were, of course, missing values present due to there being no observations during weekends, winter breaks and summer vacations. The simplest approach to this problem is to delete the missing days from the analysis. Another way to handle missing information (e.g. caused by the weekends) is to aggregate data by week for analysis. This thesis also explores the advantages and disadvantages of aggregating the data weekly instead of daily.

The school absenteeism data showed inconsistency in the number of days reported to WDGPH. Some schools only reported absenteeism for fifty days in the five-year period, whereas other schools reported absenteeism for more than 800 days over the same period. These inconsistencies would be expected to continue into the foreseeable future and so we evaluate the merit in using all schools in the WDG area against using just a selection of schools that report consistently, when trying to predict influenza epidemics. Therefore, the elementary and high school datasets were further divided: average absenteeism from all schools and average absenteeism of the three "best" schools, chosen based on their consistency in reporting and noise level. The average absenteeism was calculated both per day and per week. Analysis was also performed on individual schools and the results for one of each school type are reported. Table 1 displays the names and description of the different subsets of the data, along with the absenteeism mean and standard deviation of each dataset.

Table 1: Summary of the different school data under consideration, along with their mean and standard deviation (percent absenteeism).

Data Type	**Description**	**Mean(SD)**	
		Daily	**Weekly**
Elementary			
ES-89	Daily /Weekly absenteeism from one elementary school (anonymized id#89) with most consistent reporting.	5.31% (2.13%)	5.44% (1.80%)
ES-3avg	Daily / Weekly average of three most informative elementary schools	6.47% (2.57%)	6.45% (2.00%)
ES-allavg	Daily/Weekly average of all reporting elementary schools	5.34% (2.04%)	5.35% (1.52%)
High school			
HS-4	Daily /Weekly absenteeism from one high school (anonymized ID # 4) with most consistent reporting.	5.68% (1.93%)	5.64% (1.22%)
HS-3avg	Daily / Weekly average of three most informative high schools	7.86% (3.22%)	7.82% (2.36%)
HS-allavg	Daily/Weekly average of all reporting high schools	8.13% (3.15%)	8.11% (2.39%)
Average of E+H ESHS-avg	Daily/Weekly average of all reporting schools	5.67% (2.11%)	5.67% (1.52%)

* ES stands for Elementary School, HS stands for High School

Chapter 3

Methods

In this chapter, we describe the different surveillance systems under evaluation, and the metrics used to evaluate their performances.

For simplicity, let X_t be the percentage of absenteeism and Y_t be the counts of influenza cases, at time t ($t = 1, \dots, T_i$). Here, t could be the t^{th} observed day or t^{th} observed week depending on the aggregation method. T_i is the total number of observation times in the i^{th} year, such that i=1, 2, 3, 4 or 5 for years 2009, 2011, 2012, 2013 and 2014, respectively.

3.1 Moving Average

The moving average uses averaging over time to smooth out day-to-day variation. It predicts whether the current observation is signal-worthy by comparing it to the average of the values in the last w time steps, where w is the size of the "window" of data used [14]. Suppose that $x_1, x_2, \dots, x_t$ are individual observations that have been collected. The moving average of span w at time t is defined as:

$$MA_t = \frac{x_t + x_{t-1} + \cdots + x_{t-w+1}}{w}$$

The mean (MA_t) and standard deviation ($\hat{\sigma}_t$) of observations within this span w is calculated and used to generate an upper control limit (UCL_t) value at time t where

$$\hat{\sigma}_t = \sqrt{\frac{1}{w-1}\sum_{j=1}^{w}(X_{t-j+1} - MA_t)^2}$$

$$UCL_t = MA_t + k\hat{\sigma}_t$$

where k usually takes values between 1 and 10.

Traditionally, if the next observation X_{t+1} is greater than the UCL_t, then an alarm is raised [14]. Another scenario is also explored where an alarm is raised if the moving average of the next observation MA_{t+1} is greater than the UCL_t [15]. The aim is to determine whether this is more robust than using single day-to-day changes.

The degree of smoothing is directly proportional to the window size, which, in turn directly affects the sensitivity of the signal detection. If the window is wide, it will be less sensitive to smaller shifts. However, if the window is too narrow, there is a risk of having too many false alarms due to small noise-based shifts. Thus, the user-defined constant k determines the sensitivity of the surveillance system: lower values of k will result in reduced sensitivity, as alarms will be triggered with less of a deviation from the mean process [14]. Different combinations of window size ($2 \leq w \leq 25$) and k ($1 \leq k \leq 10$) are considered for each data set to determine optimal parameters.

3.2 Exponentially Weighted Moving Average

The exponentially weighted moving average (EWMA) is a development of the moving average algorithm in which successively lower weights are assigned to observations further back in time, controlled by a smoothing constant [14].

Let z_t be the EWMA statistic for the current data point X_t and z_t is defined as an exponentially weighted sum of past values as follows:

$$z_t = \lambda X_t + (1 - \lambda) z_{t-1}$$

where λ is the user-defined smoothing constant, $0 < \lambda \leq 1$. If λ is large, z_t depends more on current and recent observations. If λ is small, z_t depends on more distant observations. For example, if $\lambda = 0.2$, then the weight assigned to the current observation is 0.2, and the preceding observations are given a weight of 0.16, 0.128, 0.1024, and so forth. However, if $\lambda = 0.85$, then the weight assigned to the current observation is 0.85, and the preceding observations are given a weight of 0.1275, 0.019125, 0.002869, and so forth. Note that when $\lambda = 1$ no smoothing of data occurs [15].

Assuming that the observations x_t are independent random variables with variance σ^2, then the variance of z_t denoted as $\sigma_{z_t}^2$ is given by

$$\sigma_{z_t}^2 = \sigma^2 \left(\frac{\lambda}{2 - \lambda}\right) \left[1 - (1 - \lambda)^{2t}\right]$$

Using this variance, an upper control limit can be established as

$$UCL = \mu_0 + L\sigma_{z_t}$$

Here, μ_0 and σ are the target value of mean and standard deviation respectively [15]. This estimate of μ_0 is also used as the starting value z_0. Note that L is the desired width of the control limits and functions in the same manner as the parameter k in the moving average.

In this analysis, EWMA is applied on each year individually, while using the log transformed mean of the remaining years as the μ_0 and a value of 1 as the σ. The various options for lambda ($0.05 \leq \lambda \leq 1$) and L ($0.1 \leq L \leq 10$) are considered. The EWMA analyses are implemented in R using the qcc package.

3.3 Logistic Regression

Regression methods are widely used in syndromic surveillance [8]. Linear regression, however, assumes normality of the residuals, but since the data is collected for a small region, the average number of cases/day during an outbreak season is < 2, therefore eliminating the possibility of assuming a normal distribution [16]. Thus, generalized linear models are better suited here and thus Poisson regression is typically used. However, since the focus is more on predicting whether there is presence/absence of influenza in the community, rather than the expected count of influenza cases, here logistic (or binomial) regression is used.

Logistic regression allows you to generate models that predict a dichotomous (binary) dependent variable from one or more predictor variables. For the case at hand, the binary dependent variable is 1 if there is at least one case on a given day, and 0 if otherwise. The canonical link function (logit), which is the log odds of the outcome, is modelled as a linear function of the predictor variables (x_i):

$$\text{logit}(\rho) = \log\left(\frac{\rho}{1-\rho}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_i x_i$$

and

$$\rho = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_i x_i)}}$$

where ρ denotes the probability of the case occurrence. The intercept and coefficients, $\underline{\beta} = (\beta_1, \dots, \beta_i)$, are estimated while minimizing the error between actual data and the predicted value in some way [17]. In our case, the predictor variable x_1 is the absenteeism value for a given day.

Since seasonal influenza is not usually considered to be a life-threatening disease, only those with severe symptoms are likely to seek medical attention, while milder symptoms prompt self-medication [18;19]. As a result, influenza may be circulating within the community at a given point, but it may take some time before it is first diagnosed and reported. Therefore, the lagged cross-correlation between the absenteeism and flu cases is taken into account, and the various lagged values of absenteeism are modelled by,

$$\text{logit}(\rho) = \log\left(\frac{\rho}{1-\rho}\right) = \beta_0 + \beta_1 x_{t-l}$$

Here, x_{t-l} denotes the value of absenteeism observed at time $t-l$ with lag $l \in \{1,2,\dots\}$. Hence, for larger values of l, absenteeism data from further into the past are used. This is the first model under consideration.

Our second model aims to capture the seasonality of influenza, along with its relationship with absenteeism. Since influenza follows a cyclic trend, trigonometric sine and cosine functions are added to the model to capture the underlying sinusoidal behavior of seasonal influenza. This "harmonic logistic regression" is given by

$$\text{logit}(\rho) = \log\left(\frac{\rho}{1-\rho}\right) = \beta_0 + \beta_1 x_{t-l} + \beta_2 \sin\left(\frac{2\pi t}{T^*}\right) + \beta_3 \cos\left(\frac{2\pi t}{T^*}\right)$$

where, T^* equals 365 for daily aggregated data and 52 for weekly aggregated data.

Both models are tested using cross-validaton to estimate how accurately the predictive model would perform in practice. Since this is time-series data, a modified leave-one-out-cross-validation (LOOCV) is performed by using a single school year as the validation data and the remaining years as the training data. This is repeated until each school year has been used as the validation data. Predicting an outbreak is based on different probability thresholds $(0.1 \leq \Theta \leq 1)$ for the varying absenteeism lags $(0 \leq l \leq 30)$. That is, if ρ is estimated to be greater than the chosen threshold a prediction of a influenza epidemic is made (i.e. an alarm is raised).

3.4 Evaluation Metrics

Our aim is to develop surveillance systems that use school absenteeism rates to help predict an influenza epidemic before hospital cases are reported. This should allow public health professionals to investigate and instigate early interventions and, in doing so, help retard the course of the epidemic, hence lowering causality rates. According to the epidemiologist at WDGPH, an alarm should ideally be raised one to two weeks before the first hospital diagnosed case

Report date of the first hospital diagnosed case within each academic year is used as the reference date for all metrics. Obviously, there is no merit in using school absenteeism to detect influenza if an alert is not raised before this reference date. Therefore, only alerts produced before the first case will be used in the evaluation metrics. Of these, the alerts produced within two weeks before the first case are considered true alarms and any alerts before this two-week period are considered to be false alarms (See Figure 2).

The following metric evaluates the false alarm rate (FAR) of a surveillance system:

$$FAR = \begin{cases} \dfrac{n_f}{n_f + c}, & \text{if } n_f + c > 0 \quad \text{(i.e. alarms are produced)} \\ 1, & \text{if } n_f + c = 0 \quad \text{(i.e. no alarms produced)} \end{cases}$$

where, n_f is the total number of false alarms produced (i.e. alarms produced 14 days prior to the first case), and c indicates whether there were any true alarms, equalling 0 in the case of no true alarms and 1 if a true alarm was raised.

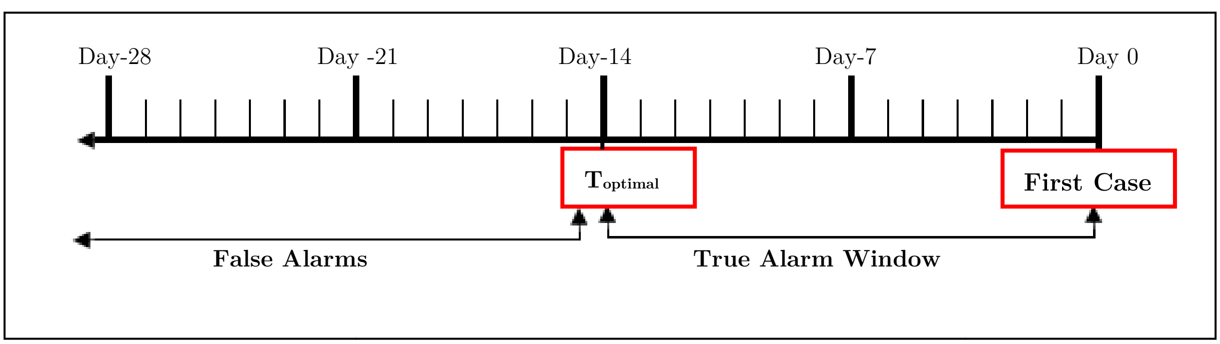


Figure 2: Differentiating between true and false alarms.

Therefore, the FAR lies between 0 and 1, and a value of 0 indicates only true alarms, whereas a value of 1 indicates only false alarms. Further corrections are made to processes that failed to produce any alarm by assigning a value of 1 to the FAR.

An ideal predictive surveillance system will not only have an appropriate balance between true and false alarms, but it should also be able to give predictions in a timely manner. For simplicity, let τ be the difference between time t and the reference date, such that $\tau \in \{\tau_{max}, \dots, -2, -1, 0\}$, where τ_{max} is the difference between the first school day and the first case, and therefore differs each year. In order to measure the timeliness of the true alarms, the "accumulated days delay" (ADD) of the true alarm, χ_τ, is used:

$$\chi_\tau = \begin{cases} \tau_{obs} - \tau_{optimal}, & \text{if } c = 1 \\ \mid \tau_{max} \mid, & \text{if } c = 0 \end{cases}$$

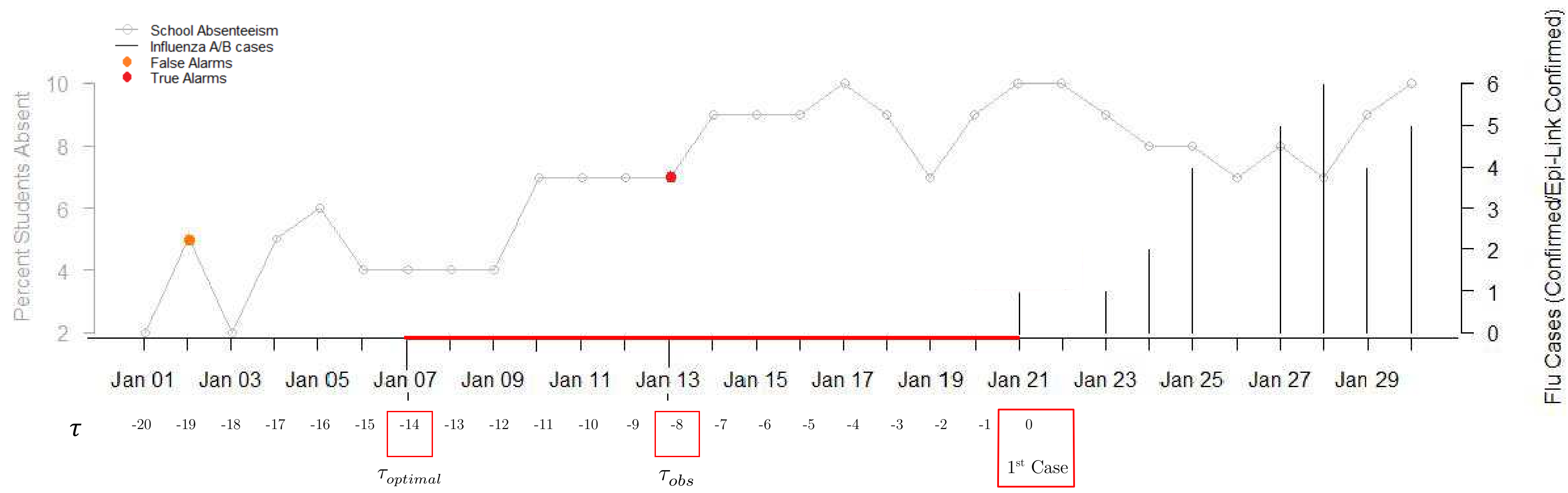
where, τ_{obs} is the day of the first true alarm, and $\tau_{optimal}$ is when our optimal alarm should have been raised. In this case, the optimal alarm should be raised fourteen days before the first hospital case recorded, and thus $\tau_{optimal} = -14$. The idea behind this is to penalize the processes that produce alarms too late, leaving little or no room for public health intervention. Therefore, if a true alarm is produced, χ_τ can take values between 0 (indicating the earliest alarm two weeks before the first case) and 14 (indicating that the alarm was received on the same date as the first case). To account for surveillance systems that fail to produce any true alarms, such systems can be assigned a large arbitrary value [8;14]; in this instance $|\tau_{max}|$ equals the total number of days before the first case. This approximates to a surveillance system which relies only on reports of a hospital confirmed case.

The ADD and FAR are not comparable between the daily- and weekly-aggregated analyses data due to the nature of the observations. For weekly data we are predicting if there is a outbreak at the end of each week, and therefore by default, the FAR rates will be lower and the ADD will be higher than for daily data.

Therefore, an optimal model should have a low FAR and small ADD. In the event that two surveillance systems produce the same FAR and ADD values, another evaluation metric, precision measure (PM) is proposed. PM measures the relative timeliness of the true alarms given by a predictive model and can take values between 0 and 1. We define the PM as

$$PM = \begin{cases} \dfrac{(\chi_\tau^2 + 0.01\ )}{(\chi_1^2 + \cdots\ + \chi_F^2) + (\chi_\tau^2 + 0.01)}, & \text{if } c = 1 \\ 0, & \text{if } c = 0 \end{cases}$$

where, $\chi_1\ , \dots\ , \chi_F$ are the timeliness of the false alarms, measured in the same manner as the ADD. A high PM indicates that either the number of false alarms, F, is small, or that χ_F is small implying that the day of the false alarm is very close to $-\ \tau_{optimal}$. So, a surveillance system with a higher PM should be more precise than a system with a lower PM. A value of 0 indicates a system either produces no alarms or produces all false alarms. Figure 3 illustrates how the above metrics are calculated for a surveillance system that produced one false alarm and one true alarm.



$$\tau_{obs} = -8$$
$$n_f = 1$$
$$c = 1$$
$$\boldsymbol{FAR} = \frac{1}{1+1} = 0.5$$
$$\boldsymbol{ADD} = \tau_{obs} - \tau_{optimal} = -8 - (-14) = 6$$

$$\chi_F = -19 - (-14) = -5$$

$$\boldsymbol{PM} = \frac{(6^2 + 0.01\ )}{(-5^2) + (6^2 + 0.01)} = 0.590$$

Figure 3: Sample evaluation metric calculation for a surveillance system that produced one true alarm and one false alarm.

The above metrics are applied to all surveillance systems evaluated reporting the average false alarm rates and alarm delays over all years under consideration (see Section 2.3). All surveillance systems and their respective parameters are summarized in Table 2. The Precision Measure will only be reported if needed (i.e. when two processes have the same FAR and ADD).

Table 2: Summary of the different surveillance systems and their parameters under evaluation

Method	**Subtype / Model**	**Alarm Criteria**	**Parameters**
Moving Average	MA-mean	$MA_{t+1} > UCL_t$	Window size $(2 \leq w \leq 25)$ k $(1 \leq k \leq 10)$
	MA-x	$X_{t+1} > UCL_t$	
Exponentially Weighted Moving Average		$z_t > UCL$	Lambda $(0.05 \leq \lambda \leq 1)$ L $(0.1 \leq L \leq 10)$
Logistic Regression	Absent	$\rho >$ threshold (Θ)	Probability threshold $(0.1 \leq \Theta \leq 1)$ Absenteeism lags $(0 \leq l \leq 30)$
	Absent+cos+sin		

Chapter 4

Results

4.1 Preliminary Results

Irregular patterns were observed in the moving average and logistic regression systems as the window sizes and lags increased, respectively. Further investigation showed that this was due to the inconsistency in the number of observations before the first case appeared per year. For example, in some years there were only 43 days between the first school day and the first influenza case. Therefore, when a 20 day lagged logistic regression was fit, there were only 23 days to evaluate, of which 14 days will be in the true alarm window, leaving only 9 days in the false alarm period. As a result this produced a low false alarm rate by default. The same pattern was observed in the moving average process, which depended on the number of school days (rather than all days) before the first case. We could tackle this problem by running separate analysis over each year, but since we want a surveillance system that would be consistent over all years, it was decided to omit the higher moving average windows and lags from our results. Also, in past studies the maximum cross-correlation coefficient of school absenteeism on influenza was found between a lag of 2 days [7] to 1 week [20], and therefore, large lags may be biologically less relevant.

4.2 Main Results

The tables in Appendix 1 and 2 summarize the average FAR and ADD under the different combination of parameters for all surveillance systems and data sets, for daily and weekly data, respectively. The results from the EWMA revealed that the FAR reduced as the λ increased towards 1. This is surprising since in practice $0.05 \leq \lambda \leq 0.25$ are found to work well [15]. See Discussion for a detailed explanation of what this implies.

The parameters that produced lowest FAR and its corresponding ADD under the different surveillance systems (MA, EWMA and LR) for the daily aggregation are shown in Table 3. Likewise, Table 4 shows the parameters that produces the lowest FAR for weekly data.

The moving average systems performed moderately well for the daily data but poorly for the weekly set. Moving average (mean) produced two good sets of daily results (ES-89 and ES-3avg) with low false alarm rates but these were counteracted by high ADD values from not being able to detect outbreaks occurring in all the years. The moving average (x) system produced the best daily result for individual high school (anonymized ID #4). Figure 4 displays this result for MA-x ($w = 5$, $k = 2.0$) which produced the lowest daily FAR of 0.250 and average alarm delay of 6.25 days. The EWMA performed poorly overall with low false alarm rates counteracted by the high ADD under many scenarios. However, as mentioned earlier, when λ approached 1 better results(lower FAR) were observed, especially in the case of individual high school (anonymized ID #4) where it produced exceptional results for both the daily and weekly data.

Table 3: Daily data results. Lowest False Alarm Rate (FAR) and corresponding Accumulated Days Delay (ADD) for the different surveillance systems.

Surveillance System		**Data**	**Parameters** (w/λ/ l)	(k/L/Θ)	**FAR**	**ADD**
Moving Average	MA-mean	**ES-89**	w **= 3**	**k = 1.5**	**0.375**	**19**
		ES-3avg	w **= 3**	**k = 3.5**	**0.375**	**28.25**
		ES-allavg	w = 3	k = 2.0	0.500	6.6
		HS-4	w **= 2**	**k = 1.5**	**0.427**	**3.25**
		HS-3avg	w = 3	k = 2.0	0.667	27.5 (PM=0.253)
		HS-allavg	w = 4	k = 1.0	0.667	27.5 (PM=0.251)
		ESHS-avg	w = 3	k = 2.5	0.567	24
	MA-x	ES-89	w = 3	k = 5.0	0.500	19
		ES-3avg	w = 9	k = 2.5	0.438	13.5
		ES-allavg	w = 4	k = 5.0	0.400	22.6
		HS-4	w **= 5**	**k = 2.0**	**0.250**	**6.25**
		HS-3avg	w = 5	k = 3.5	0.500	42.25
		HS-allavg	w = 3	k = 5.0	0.667	27
		ESHS-avg	w = 5	k = 4.0	0.500	29
Exponentially Weighted Moving Average		ES-89	λ = 0.30	L = 0.2	0.642	11.25
		ES-3avg	λ = 0.20 - 1	L = 1.0-3.4	0.500	23.50
		ES-allavg	λ = 0.95 - 1	L = 0.7-0.8	0.583	4.6
		HS-4	λ **= 1**	L = **1-1.1**	**0.250**	**14.25**
		HS-3avg	λ = 0.05	L = 0.2-0.5	0.500	29.5
		HS-allavg	λ = 0.05	L = 0.1-0.2	0.500	33
		ESHS-avg	λ = 0.55	L = 0.1	0.521	21.8
Logistic Regression	Absent	**ES-89**	l **= 11**	**Θ = 0.150**	**0.450**	**1.75**
		ES-3avg	l **= 11**	**Θ = 0.175**	**0.354**	**8.75**
		ES-allavg	l = 10	Θ = 0.175	0.500	9.0
		HS-4	l = 7	Θ = 0.300	0.462	4.75
		HS-3avg	l = 14	Θ = 0.300	0.477	19.5
		HS-allavg	l = 14	Θ = 0.300	0.481	19.25
		ESHS-avg	l = 0	Θ = 0.300	0.482	6
	Abs+cos+sin	ES-89	l = 0	Θ = 0.250	0.500	22.5
		ES-3avg	l = 3	Θ = 0.125	0.500	24
		ES-allavg	l **= 0**	**Θ = 0.175**	**0.500**	**20.2**
		HS-4	l = 7	Θ = 0.275	0.750	36.5*
		HS-3avg	l = 0 and 7	Θ = 0.250	0.750	36.5*
		HS-allavg	l = 0 and 7	Θ = 0.250	0.750	36.5*
		ESHS-avg	l = 3	Θ = 0.100	0.533	20

* These systems produced the same results (PM = 0.25)

Table 4: Weekly data results. Lowest False Alarm Rate (FAR) and corresponding Accumulated Days Delay (ADD) for the different surveillance systems.

Surveillance System		**Data**	**Parameters** (w/λ/ l)	(k/L/Θ)	**FAR**	**ADD**
Moving Average	MA-mean	ES-89	w = 3	k = 0.5	0.604	2
		ES-3avg	**w = 3**	**k = 0.5**	**0.436**	**0.25**
		ES-allavg	w = 3	k = 0.5	0.500	0.4
		HS-4	w = 2	k = 2.5	0.500	5.25
		HS-3avg	w = 3	k = 0.5	0.625	6.75
		HS-allavg	w = 3	k = 0.5	0.563	3.75
		ESHS-avg	w = 3	k = 0.5	0.500	0.2
	MA-x	ES-89	w = 3	k = 0.5	0.492	0.5
		ES-3avg	w = 3	k = 1.0	0.450	0.25
		ES-allavg	**w = 3**	**k = 1.0**	**0.400**	**0.2**
		HS-4	w = 3	k = 0.5	0.458	0.5
		HS-3avg	w = 3	k = 0.5	0.458	0.25 (PM=0.252)
		HS-allavg	w = 3	k = 0.5	0.458	0.25 (PM=0.252)
		ESHS-avg	w = 3	k = 2.5	0.500	3
Exponentially Weighted Moving Average		ES-89	λ = 0.05-0.45	L = 1.4-2.5	0.750	6.75
		ES-3avg	λ = 0.95-1	L = 0.7	0.500	3.50
		ES-allavg	λ = 0.30-1	L = 0.1-2.1	0.800	8.20
		HS-4	**λ = 1**	**L = 0.1**	**0.125**	**0.75**
		HS-3avg	λ = 0.2-0.4	L = 0.1-0.8	0.500	5.5
		HS-allavg	λ = 0.25-0.9	L = 0.1-0.8	0.500	5.75
		ESHS-avg	λ = 0.8 -1	L = 0.1	0.600	5.6
Logistic Regression	Absent	**ES-89**	**l = 3**	**Θ = 0.400**	**0.325**	**1**
		ES-3avg	**l = 3**	**Θ = 0.350**	**0.313**	**1**
		ES-allavg	l = 0	Θ = 0.350	0.425	0.2
		HS-4	**l = 3**	**Θ = 0.500**	**0.250**	**1.5**
		HS-3avg	l = 3	Θ = 0.400	0.458	3.75 (PM=0.500)
		HS-allavg	l = 3	Θ = 0.400	0.458	3.75 (PM=0.500)
		ESHS-avg	**l = 1**	**Θ = 0.325**	**0.271**	**0.8**
	Abs+cos+sin	ES-89	l = 0	Θ = 0.700	0.500	3.5
		ES-3avg	l = 2	Θ = 0.125	0.583	2
		ES-allavg	l = 2	Θ = 0.100	0.505	1.6
		HS-4	**l = 1**	**Θ = 0.225**	**0.450**	**2.25**
		HS-3avg	l = 2	Θ = 0.100	0.464	2
		HS-allavg	l = 2	Θ = 0.100	0.464	1.75
		ESHS-avg	l = 3	Θ = 0.100	0.471	1.6

Figure 5 displays the best weekly results on HS-4 using EWMA ($\lambda = 1$, L= 0.1), producing a FAR of 0.125 and average alarm delay of 0.75 weeks. It can be seen that logistic regression with only absenteeism data produced satisfactory results with both daily and weekly datasets. However, the addition of the harmonic, sine and cosine, variables led to models that were not able to detect outbreaks that occurred earlier in the year (2010-2011 and 2012-2013 academic years), therefore resulting in high ADD values.

These results can be compared to the performance of the traditional 10% threshold method (Table 5), which was only able to produce satisfactory results on the individual elementary school (anonymized id #89), with a FAR of 0.75 and average alarm delay of 45.25 days (or a 7 week delay in the weekly data). All surveillance systems applied in the study, even including the relatively poorly performing harmonic logistic regression, performed better than the traditional 10% baseline method in producing timely alerts with low false alarm rate.

Table 5: Daily and Weekly results for 10% absenteeism threshold method.

Method	**Data**	**Daily**		**Weekly**	
		FAR	**ADD**	**FAR**	**ADD**
10% Absenteeism Threshold	ES-89	0.75	45.25	0.75	7
	ES-3avg	0.75	45.25	1	10
	ES-allavg	0.8	53.6	1	10.6
	HS-4	1	59.25	1	10
	HS-3avg	0.8125	30.5	1	10
	HS-allavg	0.8125	30.5	1	10
	ESHS-avg	0.8	53.6	1	10.6

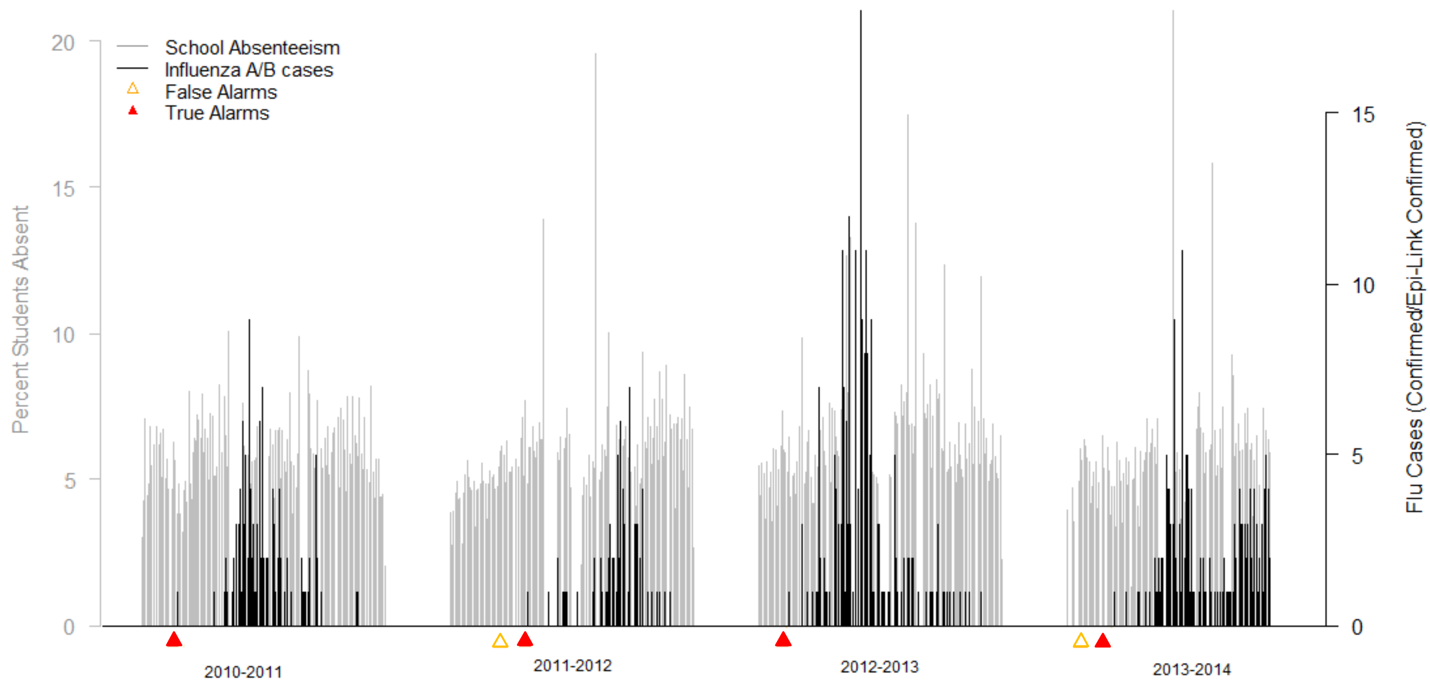


Figure 4: Best Daily Result was found on the dataset containing absenteeism from a single high school (anonymized ID #4). The surveillance system here is Moving Average(x) (w =5, k = 2.0), producing a FAR of 0.25 and an average accumulated delay of 6.25 days.

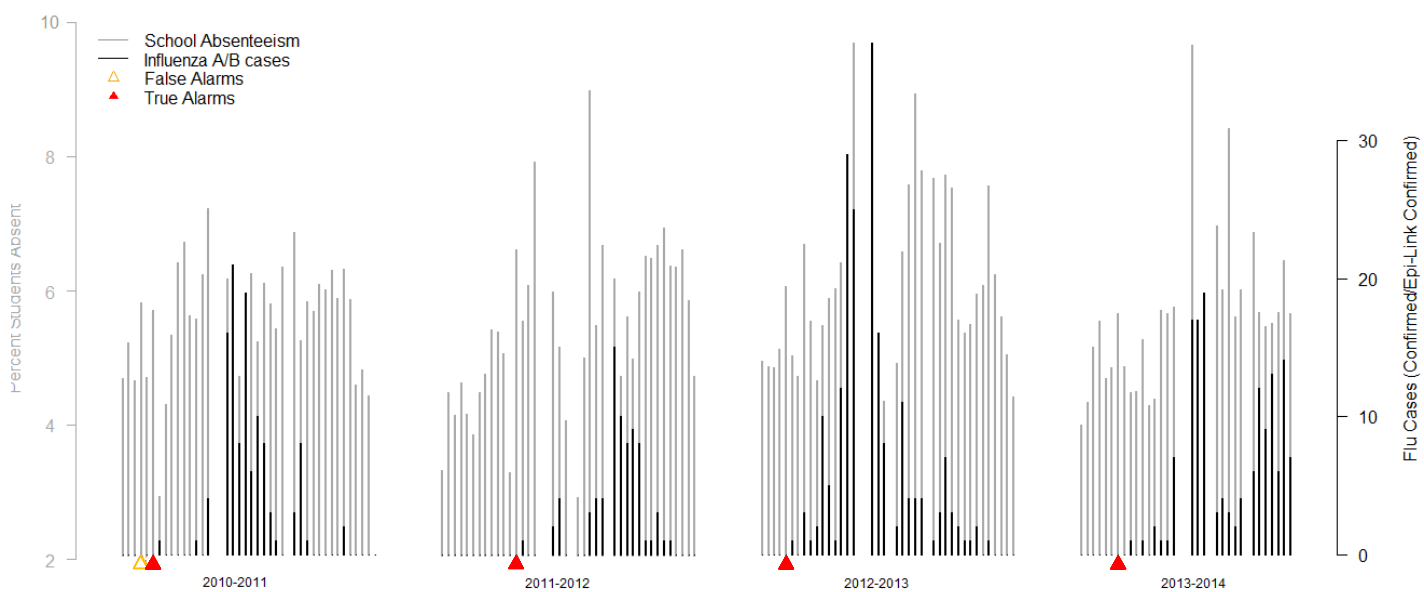


Figure 5: Best Weekly Result was found on the dataset containing absenteeism from a single high school (anonymized ID # 4). The surveillance system here is EWMA (λ = 1, L= 0.1), producing a FAR of 0.125 and average accumulated delay of 0.75 weeks.

Chapter 5

Discussion

Our results suggest that the proposed surveillance systems are superior to the traditional methods used for the analysis of school absenteeism in early detection of influenza. The traditional methods, at best, produce a false alarm rate of 0.75 and an average delay of 45.25 days, whereas the surveillance systems implemented here were able to reduce the FAR to a more acceptable level of 0.250 and a corresponding average accumulated delay of 6.25 days. Of the surveillance systems considered, the simpler MA and EWMA outperformed the more complicated systems such as harmonic logistic regression. This suggests that placing undue importance on expected seasonal patterns may be counterproductive, as it hinders the system's ability to detect outbreaks that occur earlier in the year, even when there is significant school absenteeism.

Interestingly, some of the best results produced by the EWMA system occurred when the λ equalled 1 or approached closer to 1 (i.e. more weight on recent observation and little or no weight on past observations), suggesting that the absenteeism increase associated with influenza is not a small sustained shift but rather a large or sudden shift. This study also suggests that monitoring a single school or a few consistently reporting schools may provide better results than monitoring a large number. This would appear to be a better use of resources since there are so many schools with inconsistent reporting patterns. It was also hypothesized by WDGPH that monitoring high schools would produce a high number of false alarms due to the high noise level. Our results have shown that reliable results can be generated from high schools as well. However, a clear conclusion cannot be reached in deciding whether elementary or high schools would be better suited for monitoring. We would suggest that maybe monitoring both types would be sensible, allowing us to take account of age-specific differences.

We hypothesized that monitoring at a weekly level would reduce the false alarm rate, but monitoring at a daily level would improve timeliness. Our results suggest that monitoring the schools on a weekly basis produces a relatively low level of false alarm rate, while still producing alerts early enough for intervention. Therefore the weekly monitoring appears to be a viable option. Unfortunately, we cannot come to a definitive conclusion as to whether weekly monitoring should be chosen over the daily, as our FAR and ADD are not comparable between the two data aggregate types.

Despite the performance of our surveillance systems, there are several limitations. Firstly, only 4 to 5 years worth of data satisfied our criteria for analysis. Although the surveillance systems were applied in a cross-validated setting, to minimize dependencies between years and to mimic a real time monitoring setting, inferential conclusions from such a small dataset are still limited.

Secondly, the missing values resulting from school holidays and weekends were not accounted for. They could, for example, have been imputed, which may have impacted the performance of the surveillance systems. It should be noted that the logistic regression only used a single day's absenteeism from the past (x_{t-l}). It might be felt that looking at all days going back to $t-l$ would be sensible. Unfortunately, this could not be done without imputation because the multiple lags cause multiple missing values, and therefore our sample size would have decreased substantially due to listwise deletion. Imputation methods that respect the prospective nature of the analysis and maintain the integrity of the actual data would hopefully maximize the use of all the available data, and would be an interesting avenue for further work.[1]

Thirdly, no single surveillance system (MA, EWMA or LR) produced consistently strong results across all data aggregate types. One reason for this could be because some data aggregate types have too much important information smoothed out (e.g ESHSavg). However, this could also be due to the other aforementioned limitations of the surveillance systems considered, and it might be that other systems may perhaps perform more consistently. Since this study already involved large amounts of computation, resulting from the consideration of different parameter combinations for each system, we did not evaluate a number of such other methods that have been proposed for syndromic surveillance [8]. Also due to the anonymized nature of our data, we were not able

[1] Deleting the missing values can also allow us to fit more lagged variables. We performed this analysis on all data sets, but the results were not a noticeable improvement, and therefore not reported in this thesis.

to take advantage of any spatial scan statistic methods [8]. These may have better suited the study, as the WDG district spans a large area of approximately 4140 sq. km. Disparities in population densities are evident as 98% of the area is rural and 54% of the population live in these rural areas [21;22]. This raises questions such as, if a rural school produced the alert, should the alarm be raised across the whole district? This remains an area for future research.

Lastly, all the years analysed in this thesis had the first case appear between October to late November. As such, we have not explored how a late epidemic occurring after the winter break would affect the surveillance systems. The school winter break would produce missing values for two consecutive weeks, greatly affecting the ability to detect an outbreak. Even the best imputation methods would likely produce highly imprecise estimates, which puts into question whether school absenteeism surveillance would be a reliable system for such years.

Since the surveillance systems were not applied to emerging diseases (i.e. pHINI), our results also cannot be generalized to outbreaks outside of the seasonal levels. An area of future research could measure the properties of emerging outbreaks and determine how to remain sensitive to them while allowing for good prediction of seasonal outbreaks. Although we found harmonic logistic regression did not perform well, other methods have been proposed to account for the year-by-year variation in the onset of influenza. For example, Fanshaw *et al.* [23] suggested a harmonic regression model where the coefficients of sine and cosine can change over time according to a random walk model.

Future research could consider incorporating such covariates into the logistic regression model to improve its predictive performance.

Bibliography

[1] Provincial Infectious Diseases Advisory Committee, "Syndromic Surveillance Discussion Paper,"2012.

[2] K. M. Neuzil, C. Hohlbein, and Y. Zhu, "Illness among schoolchildren during influenza season: effect on school absenteeism, parental absenteeism from work, and secondary illness in families," *Arch. Pediatr. Adolesc. Med.*, vol. 156, no. 10, pp. 986-991, Oct.2002.

[3] W. P. Schmidt, R. Pebody, and P. Mangtani, "School absence data for influenza surveillance: a pilot study in the United Kingdom," *Euro. Surveill*, vol. 15, no. 3 Jan.2010.

[4] C. A. K. Mogto, G. De Serres, M. D. Fradet, G. Lebel, S. Toutant, R. Gilca, M. Ouakki, N. Z. Janjua, and D. M. Skowronski, "School Absenteeism As an Adjunct Surveillance Indicator: Experience during the Second Wave of the 2009 H1N1 Pandemic in Quebec, Canada,", 7 ed 2012.

[5] M. Besculides, R. Heffernan, F. Mostashari, and D. Weiss, "Evaluation of school absenteeism data for early outbreak detection, New York City," *BMC Public Health*, vol. 5, no. 1, p. 105, 2005.

[6] P. Mann, E. O'Connell, G. Zhang, A. Llau, E. Rico, and F. C. Leguen, "Alert system to detect possible school-based outbreaks of influenza-like illness," *Emerg. Infect. Dis.*, vol. 17, no. 2, pp. 262-264, Feb.2011.

[7] J. R. Egger, A. G. Hoen, J. S. Brownstein, D. L. Buckeridge, D. R. Olson, and K. J. Konty, "Usefulness of school absenteeism data for predicting influenza outbreaks, United States," *Emerg. Infect. Dis.*, vol. 18, no. 8, pp. 1375-1377, Aug.2012.

[8] S. Unkel, C. P. Farrington, P. H. Garthwaite, C. Robertson, and N. Andrews, "Statistical methods for the prospective detection of infectious disease outbreaks: a review," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 175, no. 1, pp. 49-82, Jan.2012.

[9] C. Sonesson and D. Bock, "A review and discussion of prospective statistical surveillance in public health," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 166, no. 1, pp. 5-21, Feb.2003.

[10] K. L. Tsui, W. Chiu, P. Gierlich, D. Goldsman, X. Liu, and T. Maschek, "A Review of Healthcare, Public Health, and Syndromic Surveillance," *Quality Engineering*, vol. 20, no. 4, pp. 435-450, Sept.2008.

[11] S. H. Steiner, K. Grant, M. Coory, and H. A. Kelly, "Detecting the start of an influenza outbreak using exponentially weighted moving average charts," *BMC. Med. Inform. Decis. Mak.*, vol. 10, p. 37, 2010.

[12] A. M. Stolwijk, H. Straatman, and G. A. Zielhuis, "Studying seasonality by using sine and cosine functions in regression analysis," *J. Epidemiol. Community Health*, vol. 53, no. 4, pp. 235-238, Apr.1999.

[13] R Development Core Team, "A language and environment for statistical computing.," Vienna, Austria: 2005.

[14] M. M. Wagner, A. W. Moore, and R. M. Aryel, *Handbook of Biosurveillance* Elsevier Science, 2011.

[15] D. C. Montgomery, *Introduction to Statistical Quality Control* Wiley, 2008.

[16] P. A. Rogerson and I. Yamada, "Approaches to syndromic surveillance when data consist of small regional counts," *MMWR Morb. Mortal. Wkly. Rep.*, vol. 53 Suppl, pp. 79-85, Sept.2004.

[17] K. W. Yih, A. Abrams, R. Danila, K. Green, K. Kleinman, M. Kulldorff, B. Miller, J. Nordin, and R. Platt, "Ambulatory-care diagnoses as potential indicators of outbreaks of gastrointestinal illness--Minnesota," *MMWR Morb. Mortal. Wkly. Rep.*, vol. 54 Suppl, pp. 157-162, Aug.2005.

[18] T. Kass-Hout and X. Zhang, *Biosurveillance: Methods and Case Studies* Taylor & Francis, 2010.

[19] W. K. Yih, S. Deshpande, C. Fuller, D. Heisey-Grove, J. Hsu, B. A. Kruskal, M. Kulldorff, M. Leach, J. Nordin, J. Patton-Levine, E. Puga, E. Sherwood, I. Shui, and R. Platt, "Evaluating real-time syndromic surveillance signals from ambulatory care data in four states," *Public Health Rep.*, vol. 125, no. 1, pp. 111-120, Jan.2010.

[20] C. K. Cheng, B. J. Cowling, E. H. Lau, L. M. Ho, G. M. Leung, and D. K. Ip, "Electronic school absenteeism monitoring and influenza surveillance, Hong Kong," *Emerg. Infect. Dis.*, vol. 18, no. 5, pp. 885-887, May2012.

[21] Statistics Canada, "Census," 2011.

[22] Wellington-Duffein-Guelph Public Health, "Infectious Disease in Wellington-Dufferin-Guelph,"2013.

[23] T. R. Fanshawe, P. J. Diggle, S. Rushton, R. Sanderson, P. W. W. Lurz, S. V. Glinianaia, M. S. Pearce, L. Parker, M. Charlton, and T. Pless-Mulloli, "Modelling spatio-temporal variation in exposure to particulate matter: a two-stage approach," *Environmetrics*, vol. 19, no. 6, pp. 549-566, 2008.

APPENDIX A. Daily Results

Table A.1: Moving Average(x). FAR and ADD for alarms raised under different moving window sizes(w) and k.

MA-x	***k***																			
	Average False Alarm Rate										**Average Accumulated Delay**									
W	**0.5**	**1**	**1.5**	**2**	**2.5**	**3**	**3.5**	**4**	**4.5**	**5**	**0.5**	**1**	**1.5**	**2**	**2.5**	**3**	**3.5**	**4**	**4.5**	**5**
2	0.894	0.864	0.838	0.798	0.757	0.740	0.734	0.815	0.825	0.825	1.20	1.59	2.03	2.81	3.70	11.52	12.48	18.24	20.11	25.53
3	0.891	0.850	0.814	0.764	0.806	0.779	0.731	0.715	0.714	0.661	2.06	2.48	3.70	4.31	12.84	20.66	20.66	20.95	23.24	23.24
4	0.885	0.835	0.750	0.696	0.709	0.716	0.706	0.651	0.664	0.632	1.94	3.49	7.41	11.25	20.75	22.14	28.96	29.21	33.96	33.96
5	0.877	0.813	0.754	0.704	0.690	0.768	0.727	0.732	0.764	0.871	2.66	3.66	7.69	16.69	26.04	37.84	40.07	42.07	46.32	53.64
6	0.870	0.816	0.755	0.737	0.724	0.749	0.763	0.843	0.843	0.871	3.24	3.66	7.43	17.69	29.15	36.79	46.24	51.35	51.35	53.64
7	0.867	0.808	0.750	0.680	0.711	0.762	0.802	0.831	0.843	0.843	3.38	3.63	10.20	15.05	26.21	38.47	46.64	48.24	51.35	51.35
8	0.862	0.785	0.772	0.701	0.759	0.799	0.831	0.843	0.843	0.843	3.38	3.74	12.77	16.68	29.77	39.59	48.24	51.35	51.35	51.35
9	0.856	0.771	0.711	0.654	0.717	0.800	0.814	0.814	0.871	0.871	3.07	3.77	13.77	18.09	31.07	46.61	51.35	51.35	53.64	53.64
10	0.844	0.725	0.709	0.751	0.730	0.814	0.814	0.814	0.843	0.871	2.93	3.91	17.42	24.84	38.11	51.35	51.35	51.35	52.49	53.64

Table A.2: Moving Average (mean). FAR and ADD for alarms raised under different moving window sizes *(w)* and *k*.

MA-mean										**k**										
	Average False Alarm Rate										**Average Accumulated Delay**									
W	**0.5**	**1**	**1.5**	**2**	**2.5**	**3**	**3.5**	**4**	**4.5**	**5**	**0.5**	**1**	**1.5**	**2**	**2.5**	**3**	**3.5**	**4**	**4.5**	**5**
2	0.880	0.794	0.760	0.791	0.804	0.786	0.785	0.793	0.784	0.795	1.67	1.81	8.98	17.04	20.36	26.05	33.73	36.59	36.59	38.01
3	0.853	0.711	0.690	0.669	0.692	0.764	0.707	0.757	0.761	0.761	6.16	11.88	23.24	27.61	37.04	45.07	45.07	48.36	50.36	50.36
4	0.790	0.769	0.825	0.907	0.907	0.907	0.907	0.907	0.943	0.971	13.49	34.14	50.35	55.64	55.64	55.64	55.64	55.64	57.64	59.24
5	0.847	0.814	0.871	0.943	0.971	1.000	1.000	1.000	1.000	1.000	27.94	48.89	53.64	57.64	59.24	60.84	60.84	60.84	60.84	60.84
6	0.802	0.871	0.871	0.943	0.971	1.000	1.000	1.000	1.000	1.000	26.16	53.64	53.64	57.64	59.24	60.84	60.84	60.84	60.84	60.84
7	0.805	0.871	0.943	0.971	1.000	1.000	1.000	1.000	1.000	1.000	38.47	53.64	57.64	59.24	60.84	60.84	60.84	60.84	60.84	60.84
8	0.800	0.907	0.971	1.000	1.000	1.000	1.000	1.000	1.000	1.000	40.89	55.64	59.24	60.84	60.84	60.84	60.84	60.84	60.84	60.84
9	0.871	0.871	0.943	1.000	1.000	1.000	1.000	1.000	1.000	1.000	53.64	53.64	57.64	60.84	60.84	60.84	60.84	60.84	60.84	60.84
10	0.871	0.871	0.971	1.000	1.000	1.000	1.000	1.000	1.000	1.000	53.64	53.64	59.24	60.84	60.84	60.84	60.84	60.84	60.84	60.84

Table A.3: Exponentially Weighted Moving Average. FAR and ADD for alarms raised under different smoothing parameters (λ) and L deviations.

EWMA	L																			
	Average False Alarm Rate										**Average Accumulated Delay**									
λ	**0-1**	**1-2**	**2-3**	**3-4**	**4-5**	**5-6**	**6-7**	**7-8**	**8-9**	**9-10**	**0-1**	**1-2**	**2-3**	**3-4**	**4-5**	**5-6**	**6-7**	**7-8**	**8-9**	**9-10**
0.05	0.772	0.853	0.909	0.957	0.971	0.964	0.964	0.964	0.964	0.964	46.93	49.69	53.60	55.63	57.93	58.70	58.84	58.84	58.84	58.84
0.1	0.825	0.843	0.833	0.840	0.915	0.923	0.964	0.964	0.964	0.964	43.40	46.40	47.58	49.36	53.62	55.24	58.84	58.84	58.84	58.84
0.15	0.797	0.833	0.851	0.838	0.870	0.886	0.925	0.954	0.964	0.964	39.00	43.94	47.05	48.18	51.20	52.77	56.55	58.24	58.84	58.84
0.2	0.764	0.836	0.846	0.835	0.865	0.877	0.893	0.926	0.943	0.964	34.49	42.68	45.31	47.29	50.43	52.20	54.66	56.68	57.64	58.84
0.25	0.740	0.823	0.861	0.833	0.847	0.875	0.893	0.900	0.920	0.938	30.99	40.66	45.29	46.50	48.97	52.31	54.66	55.24	56.36	57.38
0.3	0.739	0.801	0.859	0.841	0.836	0.859	0.893	0.900	0.900	0.921	28.71	38.49	44.69	46.58	48.16	51.58	54.66	55.24	55.24	56.40
0.35	0.750	0.788	0.848	0.846	0.832	0.841	0.888	0.900	0.900	0.900	27.27	37.24	43.80	46.75	47.95	50.78	54.47	55.24	55.24	55.24
0.4	0.770	0.772	0.837	0.850	0.833	0.837	0.871	0.900	0.900	0.900	26.62	35.75	43.08	46.81	48.03	50.78	53.64	55.24	55.24	55.24
0.45	0.778	0.769	0.826	0.853	0.834	0.836	0.871	0.883	0.900	0.900	25.54	34.99	42.13	46.89	48.06	50.78	53.64	54.28	55.24	55.24
0.5	0.781	0.769	0.816	0.856	0.832	0.839	0.871	0.871	0.894	0.900	24.69	34.52	41.41	46.96	48.10	51.06	53.64	53.64	54.92	55.24
0.55	0.785	0.766	0.819	0.854	0.831	0.843	0.871	0.871	0.883	0.900	23.60	33.75	41.45	46.73	48.10	51.35	53.64	53.64	54.28	55.24
0.6	0.785	0.765	0.817	0.852	0.830	0.843	0.871	0.871	0.871	0.897	21.83	33.24	41.39	46.48	48.14	51.35	53.64	53.64	53.64	55.09
0.65	0.789	0.766	0.817	0.853	0.830	0.843	0.871	0.871	0.871	0.887	20.44	32.92	41.09	46.52	48.14	51.35	53.64	53.64	53.64	54.51
0.7	0.784	0.765	0.815	0.849	0.827	0.848	0.871	0.871	0.871	0.879	19.01	32.63	40.80	46.21	48.14	51.64	53.64	53.64	53.64	54.07
0.75	0.786	0.767	0.808	0.846	0.827	0.851	0.871	0.871	0.871	0.871	17.96	32.51	40.17	46.18	48.14	51.64	53.64	53.64	53.64	53.64
0.8	0.785	0.766	0.803	0.841	0.831	0.856	0.871	0.871	0.871	0.871	16.62	32.10	39.65	45.77	48.42	51.92	53.64	53.64	53.64	53.64
0.85	0.784	0.759	0.804	0.838	0.831	0.858	0.871	0.871	0.871	0.871	16.16	31.25	39.39	45.77	48.46	51.92	53.64	53.64	53.64	53.64
0.9	0.777	0.753	0.801	0.835	0.833	0.861	0.871	0.871	0.871	0.871	14.97	30.61	38.99	45.65	48.46	52.21	53.64	53.64	53.64	53.64
0.95	0.768	0.748	0.799	0.830	0.838	0.861	0.871	0.871	0.871	0.871	13.84	30.14	38.44	45.34	48.74	52.21	53.64	53.64	53.64	53.64
1	0.757	0.740	0.799	0.827	0.840	0.865	0.871	0.871	0.871	0.871	12.70	29.61	37.97	45.21	48.74	52.49	53.64	53.64	53.64	53.64

Table A.4: Logistic Regression- Absenteeism only FAR and ADD for alarms raised under different lags and thresholds

Absent	Threshold																	
	Average False Alarm Rate									**Average Accumulated Delay**								
Lag	**0.1-0.2**	**0.2-0.3**	**0.3-0.4**	**0.4-0.5**	**0.5-0.6**	**0.6-0.7**	**0.7-0.8**	**0.8-0.9**	**0.9-1**	**0.1-0.2**	**0.2-0.3**	**0.3-0.4**	**0.4-0.5**	**0.5-0.6**	**0.6-0.7**	**0.7-0.8**	**0.8-0.9**	**0.9-1**
1	0.949	0.867	0.771	0.796	0.871	0.871	0.871	0.871	0.956	0.03	6.49	24.50	40.08	53.55	53.64	53.64	53.64	58.36
2	0.951	0.817	0.821	0.871	0.898	0.936	0.963	0.971	1.000	0.51	14.85	44.60	53.76	55.24	57.30	58.77	59.26	60.84
3	0.861	0.795	0.871	0.907	0.914	0.936	0.936	0.971	1.000	5.10	37.36	53.89	55.82	56.21	57.36	57.36	59.29	60.84
4	0.798	0.766	0.828	0.871	0.871	0.871	0.871	0.871	0.957	10.00	34.56	49.27	54.02	54.02	54.02	54.02	54.02	58.56
5	0.806	0.769	0.832	0.871	0.871	0.871	0.871	0.900	0.977	9.62	34.51	50.20	54.15	54.15	54.15	54.15	55.64	59.65
6	0.871	0.842	0.991	1.000	1.000	1.000	1.000	1.000	1.000	5.80	40.94	60.10	60.84	60.84	60.84	60.84	60.84	60.84
7	0.940	0.753	0.919	1.000	1.000	1.000	1.000	1.000	1.000	0.29	15.37	53.34	60.84	60.84	60.84	60.84	60.84	60.84
8	0.935	0.790	0.685	0.879	1.000	1.000	1.000	1.000	1.000	0.21	6.67	27.20	49.70	60.84	60.84	60.84	60.84	60.84
9	0.937	0.737	0.947	1.000	1.000	1.000	1.000	1.000	1.000	0.29	16.41	55.41	60.84	60.84	60.84	60.84	60.84	60.84
10	0.775	0.849	1.000	1.000	1.000	1.000	1.000	1.000	1.000	6.25	45.88	60.84	60.84	60.84	60.84	60.84	60.84	60.84
11	0.694	0.715	0.921	1.000	1.000	1.000	1.000	1.000	1.000	10.75	36.30	53.96	60.84	60.84	60.84	60.84	60.84	60.84
12	0.705	0.707	0.965	1.000	1.000	1.000	1.000	1.000	1.000	9.47	35.68	58.51	60.84	60.84	60.84	60.84	60.84	60.84
13	0.811	0.863	1.000	1.000	1.000	1.000	1.000	1.000	1.000	7.29	44.44	60.84	60.84	60.84	60.84	60.84	60.84	60.84
14	0.863	0.704	0.896	0.996	1.000	1.000	1.000	1.000	1.000	2.81	21.81	50.26	60.05	60.84	60.84	60.84	60.84	60.84

Table A.5: Logistic Regression- Absenteeism + Cos + Sin FAR and ADD for alarms raised under different lags and thresholds

Absent +cos+sin	**Threshold**																	
	Average False Alarm Rate									**Average Accumulated Delay**								
Lag	**0.1-0.2**	**0.2-0.3**	**0.3-0.4**	**0.4-0.5**	**0.5-0.6**	**0.6-0.7**	**0.7-0.8**	**0.8-0.9**	**0.9-1**	**0.1-0.2**	**0.2-0.3**	**0.3-0.4**	**0.4-0.5**	**0.5-0.6**	**0.6-0.7**	**0.7-0.8**	**0.8-0.9**	**0.9-1**
1	0.810	0.737	0.750	0.938	0.973	1.000	1.000	1.000	1.000	28.33	35.65	41.95	56.75	59.34	60.84	60.84	60.84	60.84
2	0.911	0.843	0.938	1.000	1.000	1.000	1.000	1.000	1.000	35.01	43.47	55.68	60.84	60.84	60.84	60.84	60.84	60.84
3	0.826	0.920	1.000	1.000	1.000	1.000	1.000	1.000	1.000	38.16	54.37	60.84	60.84	60.84	60.84	60.84	60.84	60.84
4	0.691	0.750	0.863	0.880	0.945	0.971	0.971	0.971	0.994	32.26	44.26	53.33	54.49	57.90	59.32	59.32	59.32	60.53
5	0.738	0.725	0.900	0.909	0.936	0.971	0.979	1.000	1.000	31.48	42.03	55.64	56.10	57.49	59.35	59.72	60.84	60.84
6	0.863	0.875	0.991	1.000	1.000	1.000	1.000	1.000	1.000	37.38	50.51	60.09	60.84	60.84	60.84	60.84	60.84	60.84
7	0.922	0.796	0.875	1.000	1.000	1.000	1.000	1.000	1.000	34.65	41.01	50.88	60.84	60.84	60.84	60.84	60.84	60.84
8	0.918	0.814	0.814	0.973	1.000	1.000	1.000	1.000	1.000	34.25	38.71	45.44	58.77	60.84	60.84	60.84	60.84	60.84
9	0.900	0.812	0.929	0.973	1.000	1.000	1.000	1.000	1.000	34.91	42.88	54.87	58.64	60.84	60.84	60.84	60.84	60.84
10	0.848	0.920	1.000	1.000	1.000	1.000	1.000	1.000	1.000	38.05	54.39	60.84	60.84	60.84	60.84	60.84	60.84	60.84
11	0.817	0.832	0.991	1.000	1.000	1.000	1.000	1.000	1.000	36.67	47.31	60.15	60.84	60.84	60.84	60.84	60.84	60.84
12	0.845	0.879	1.000	1.000	1.000	1.000	1.000	1.000	1.000	36.61	50.89	60.84	60.84	60.84	60.84	60.84	60.84	60.84
13	0.883	0.857	0.964	1.000	1.000	1.000	1.000	1.000	1.000	37.24	49.09	58.05	60.84	60.84	60.84	60.84	60.84	60.84
14	0.906	0.818	0.929	1.000	1.000	1.000	1.000	1.000	1.000	34.31	41.50	55.15	60.84	60.84	60.84	60.84	60.84	60.84

APPENDIX B. Weekly Results

Table B.1: Moving Average(x). FAR and ADD for alarms raised under different moving window sizes*(w)* and *k*.

MA-x										*k*										
	Average False Alarm Rate										Average Accumulated Delay									
W	**0.5**	**1**	**1.5**	**2**	**2.5**	**3**	**3.5**	**4**	**4.5**	**5**	**0.5**	**1**	**1.5**	**2**	**2.5**	**3**	**3.5**	**4**	**4.5**	**5**
2	0.662	0.645	0.750	0.745	0.766	0.780	0.799	0.782	0.814	0.819	0.14	0.54	2.34	2.66	5.29	6.13	6.87	6.87	7.09	7.43
3	0.519	0.482	0.617	0.748	0.772	0.743	0.771	0.807	0.836	0.836	0.18	0.25	2.55	4.26	5.86	6.33	7.51	8.01	8.41	8.41

Table B.2: Moving Average (mean). FAR and ADD for alarms raised under different moving window sizes*(w)* and *k*.

MA-mean										*k*										
	Average False Alarm Rate										Average Accumulated Delay									
W	**0.5**	**1**	**1.5**	**2**	**2.5**	**3**	**3.5**	**4**	**4.5**	**5**	**0.5**	**1**	**1.5**	**2**	**2.5**	**3**	**3.5**	**4**	**4.5**	**5**
2	0.664	0.653	0.788	0.814	0.808	0.790	0.814	0.829	0.875	0.911	0.14	0.70	4.54	7.09	7.09	7.09	7.43	7.83	8.64	9.10
3	0.526	0.551	0.762	0.807	0.900	0.900	0.964	0.964	0.964	0.964	0.28	2.91	6.61	8.05	9.22	9.22	9.74	9.74	9.74	9.74

Table B.3: Exponentially Weighted Moving Average. FAR and ADD for alarms raised under different smoothing parameters (λ) and L deviations.

EWMA	L																			
	Average False Alarm Rate										Average Accumulated Delay									
λ	**0-1**	**1-2**	**2-3**	**3-4**	**4-5**	**5-6**	**6-7**	**7-8**	**8-9**	**9-10**	**0-1**	**1-2**	**2-3**	**3-4**	**4-5**	**5-6**	**6-7**	**7-8**	**8-9**	**9-10**
0.05	0.989	0.970	0.964	0.964	0.964	0.968	1.000	1.000	1.000	1.000	9.71	9.71	9.74	9.74	9.74	9.80	10.17	10.17	10.17	10.17
0.1	0.982	0.977	0.964	0.964	0.964	0.964	0.979	1.000	1.000	1.000	9.68	9.71	9.74	9.74	9.74	9.74	9.93	10.17	10.17	10.17
0.15	0.954	0.982	0.964	0.964	0.964	0.964	0.964	1.000	1.000	1.000	9.42	9.71	9.73	9.74	9.74	9.74	9.74	10.17	10.17	10.17
0.2	0.894	0.981	0.964	0.964	0.964	0.964	0.964	0.982	1.000	1.000	8.75	9.66	9.73	9.74	9.74	9.74	9.74	9.96	10.17	10.17
0.25	0.855	0.970	0.964	0.964	0.964	0.964	0.964	0.964	1.000	1.000	8.36	9.49	9.72	9.74	9.74	9.74	9.74	9.74	10.17	10.17
0.3	0.833	0.946	0.967	0.964	0.964	0.964	0.964	0.964	0.986	1.000	8.11	9.23	9.72	9.74	9.74	9.74	9.74	9.74	10.00	10.17
0.35	0.824	0.925	0.968	0.964	0.964	0.964	0.964	0.964	0.979	1.000	7.96	9.00	9.72	9.74	9.74	9.74	9.74	9.74	9.91	10.17
0.4	0.819	0.908	0.970	0.964	0.964	0.964	0.964	0.964	0.971	1.000	7.80	8.80	9.73	9.74	9.74	9.74	9.74	9.74	9.83	10.17
0.45	0.811	0.890	0.972	0.964	0.964	0.964	0.964	0.964	0.964	1.000	7.65	8.61	9.73	9.74	9.74	9.74	9.74	9.74	9.74	10.17
0.5	0.807	0.881	0.973	0.964	0.964	0.964	0.964	0.964	0.964	0.996	7.57	8.49	9.73	9.74	9.74	9.74	9.74	9.74	9.74	10.13
0.55	0.808	0.871	0.967	0.964	0.964	0.964	0.964	0.964	0.964	0.993	7.56	8.38	9.64	9.74	9.74	9.74	9.74	9.74	9.74	10.09
0.6	0.806	0.867	0.962	0.964	0.964	0.964	0.964	0.964	0.964	0.993	7.47	8.30	9.60	9.74	9.74	9.74	9.74	9.74	9.74	10.09
0.65	0.800	0.871	0.959	0.964	0.964	0.964	0.964	0.964	0.964	0.989	7.35	8.32	9.56	9.74	9.74	9.74	9.74	9.74	9.74	10.04
0.7	0.793	0.868	0.959	0.964	0.964	0.964	0.964	0.964	0.964	0.989	7.19	8.29	9.57	9.74	9.74	9.74	9.74	9.74	9.74	10.04
0.75	0.793	0.865	0.955	0.964	0.964	0.964	0.964	0.964	0.964	0.989	7.06	8.25	9.53	9.74	9.74	9.74	9.74	9.74	9.74	10.04
0.8	0.784	0.869	0.952	0.964	0.964	0.964	0.964	0.964	0.964	0.989	6.86	8.28	9.48	9.74	9.74	9.74	9.74	9.74	9.74	10.04
0.85	0.781	0.862	0.952	0.964	0.964	0.964	0.964	0.964	0.964	0.989	6.69	8.21	9.48	9.74	9.74	9.74	9.74	9.74	9.74	10.04
0.9	0.773	0.865	0.952	0.964	0.964	0.964	0.964	0.964	0.964	0.993	6.57	8.22	9.48	9.74	9.74	9.74	9.74	9.74	9.74	10.09
0.95	0.770	0.865	0.947	0.964	0.964	0.964	0.964	0.964	0.964	0.993	6.43	8.23	9.45	9.74	9.74	9.74	9.74	9.74	9.74	10.09
1	0.749	0.863	0.947	0.964	0.964	0.964	0.964	0.964	0.964	0.996	6.21	8.19	9.45	9.74	9.74	9.74	9.74	9.74	9.74	10.13

Table B.4: Logistic Regression- Absenteeism only. FAR and ADD for alarms raised under different lags and thresholds

Absent	**Threshold**																	
	Average False Alarm Rate									**Average Accumulated Delay**								
Lag	**0.1-0.2**	**0.2-0.3**	**0.3-0.4**	**0.4-0.5**	**0.5-0.6**	**0.6-0.7**	**0.7-0.8**	**0.8-0.9**	**0.9-1**	**0.1-0.2**	**0.2-0.3**	**0.3-0.4**	**0.4-0.5**	**0.5-0.6**	**0.6-0.7**	**0.7-0.8**	**0.8-0.9**	**0.9-1**
0	0.813	0.773	0.685	0.681	0.675	0.777	0.875	0.943	0.971	0.46	0.98	1.25	2.78	4.58	6.51	8.36	9.49	9.83
1	0.720	0.622	0.585	0.583	0.583	0.730	0.879	0.936	0.971	0.21	1.07	1.69	3.04	4.48	6.64	8.63	9.46	9.86
2	0.655	0.604	0.557	0.633	0.745	0.911	0.982	1.000	1.000	0.06	0.92	2.27	4.01	6.04	8.55	9.83	10.17	10.17
3	0.495	0.485	0.506	0.594	0.755	0.915	0.964	0.982	1.000	0.20	1.65	2.55	4.16	6.72	8.99	9.71	9.94	10.17

Table B.5: Logistic Regression- Absenteeism + Cos + Sin. FAR and ADD for alarms raised under different lags and thresholds

Abs+cos+sin	**Threshold**																	
	Average False Alarm Rate									**Average Accumulated Delay**								
Lag	**0.1-0.2**	**0.2-0.3**	**0.3-0.4**	**0.4-0.5**	**0.5-0.6**	**0.6-0.7**	**0.7-0.8**	**0.8-0.9**	**0.9-1**	**0.1-0.2**	**0.2-0.3**	**0.3-0.4**	**0.4-0.5**	**0.5-0.6**	**0.6-0.7**	**0.7-0.8**	**0.8-0.9**	**0.9-1**
0	0.702	0.830	0.838	0.850	0.826	0.746	0.764	0.839	1.000	3.53	5.40	5.79	6.29	6.46	6.53	7.07	8.18	10.17
1	0.670	0.741	0.793	0.855	0.821	0.764	0.782	0.866	0.993	3.46	4.62	5.38	6.40	6.49	6.57	7.38	8.50	10.09
2	0.636	0.812	0.867	0.879	0.861	0.818	0.806	0.875	0.986	2.82	5.05	5.97	6.58	6.89	6.94	7.42	8.62	10.00
3	0.660	0.822	0.889	0.865	0.821	0.810	0.804	0.866	0.993	3.37	5.15	6.31	6.41	6.54	6.97	7.50	8.42	10.08