# Spatial and Temporal Algorithm Evaluation for Detecting Over-The-Counter Thermometer Sale Increases during 2009 H1N1 Pandemic

**Jialan Que[1,2,3,*], Fu-Chiang Tsui[1,2]**

**[1]RODS Laboratory, Department of Biomedical Informatics, University of Pittsburgh**
**[2]Intelligent Systems Program, University of Pittsburgh**
**[3]Division of Biomedical Informatics, University of California, San Diego**

## *Abstract*

***Background***

***Spatial outbreak detection algorithms using routinely collected healthcare data have been developed since the late 90s to identify and locate disease outbreaks. However, current well-received spatial algorithms assume only one outbreak cluster present at the same point of time which may not be valid during a pandemic when several clusters of geographic areas concurrently occur. Based on a retrospective evaluation on time-series and spatial algorithms, this paper suggests that time series analysis in detection of pandemics is still a desirable process, which may achieve more sensitive performance with better timeliness.***

***Methods***

***In this paper, we first prove in theory that two existing spatial models, the likelihood ratio and the Bayesian spatial scan statistics, are not useful if multiple clusters occur at the same point of time in different geographic regions. Then we conduct a comparison between a spatial algorithm, the Bayesian Spatial Scan Statistic (BSS), and a time series algorithm, the wavelet anomaly detector (WAD), on the performance of detecting the increase of the over-the-counter (OTC) medicine sales during 2009 H1N1 pandemic.***

***Results***

***The experiments demonstrated that the Bayesian spatial algorithm responded to the increase of thermometer sales about 3 days later than the time series algorithm.***

***Conclusion***

***Time-series algorithms demonstrated an advantage for early outbreak detection, especially when multiple clusters occur at the same time in different geographic regions. Given spatial-temporal algorithms for outbreak detection are widely used, this paper suggests that epidemiologists or public health officials would benefit by applying time series algorithms as a complement to spatial algorithms for public health surveillance.***

***Keywords: biosurveillance, disease outbreak detection, time series***

## Introduction

Over the past decades, infectious disease outbreaks routinely devastated the world's urban and suburban population. The release of anthrax in 2001, the Severe Acute Respiratory Syndrome (SARS) outbreaks in 2002, and recent H1N1 swine flu outbreaks in 2009 are examples [1,2,3]. Lessons learned from those outbreaks include development of disease specific vaccines and effective outbreak detection algorithms that can be employed in biosurveillance systems [2,4,5].

## Background

Currently, principal approaches for outbreak detection include temporal analysis, spatial analysis and tempo-spatial analysis. Temporal analysis using time series algorithms is a conventional approach due to its simplicity when compared with spatial algorithms, which require additional geographical information. One of the algorithms, the *wavelet anomaly detector* (WAD) algorithm [4] developed by the University of Pittsburgh, is used in the Realtime Outbreak and Disease Surveillance (RODS) system. It computes a score for each unit (*e.g.*, ZIP code) area based on how many standard deviations that the number of the cases from that area in the most recent day is elevated from the expected, and then reports all the areas with higher-than-threshold scores. Some other algorithms applying temporal analysis have also been explored in [5,6,7].

In addition to time series algorithms, researchers have developed spatial and tempo-spatial algorithms to take into account geographical information in the belief that the additional spatial information may lower false alarm rates and better localize outbreaks. There are two approaches for spatial and tempo-spatial algorithms: the frequentist approach and the Bayesian approach. A representative algorithm using frequentist approach is the spatial scan statistic (KSS) developed by Kulldorff et. al. in 1997 [8], which scans the region of interest for clusters using circular windows of various sizes. Each scanning window may cover a number of ZIP code areas and is considered as a cluster candidate. This frequentist approach uses a likelihood-ratio test, a statistical test used to compare the fit of two models: the null hypothesis (no outbreaks in a region) model and the alternative hypothesis (an outbreak in a region) model, to find a window (cluster) with maximum likelihood ratio. The derivatives of KSS include the elliptic spatial scan statistic and the flexible spatial scan statistic (FSS), which are derived by relaxing the constraint of a circular cluster shape (window) [9,10].

A representative spatial algorithm using the Bayesian approach is the Bayesian spatial scan statistic (BSS) developed by Neill et. al. [11]. It computes the posterior probability of the alternative hypothesis $H_1$, $P(H_1(S)|Data)$, in a region S. It creates a conjugate Gamma-Poisson model to compute the posterior probability of having an outbreak in region S. BSS employs a rectangular scanning window (aligning with $x$ and $y$ axes) to search for clusters over a $m \times m$ grid covering the whole region of interest. Each window can comprise one or more grid cells and is considered as a potential cluster. The algorithm identifies the outbreak cluster with the highest posterior probability of having an outbreak. Another spatial clustering algorithm recently developed by the authors, the rank-based spatial clustering (RSC), employs a different searching scheme which has been demonstrated to improve computational complexity [12].

## Problem and objective

Several previous studies have demonstrated that the aforementioned spatial algorithms are able to identify and localize outbreaks and they are preferable in some ways over time series algorithms when applied to different data sources [8,10,11,13,14,15]. Nonetheless, those studies, regardless of whether they examined frequentist or Bayesian models, share the same assumption: only one outbreak cluster at a same point in time exists in the entire study region. The fact is that such an assumption may not hold during an outbreak or a pandemic, when disease activities can be found in multiple geographically separate places across a large region.

In this paper, we studied the applicability of spatial algorithms based on the single outbreak cluster assumption for detecting H1N1 pandemics. The methods section provides 1) a theoretical derivation of the deteriorated performance of the spatial algorithms when the single cluster assumption is no longer true, and 2) an evaluation scheme using real-world OTC data collected from Texas State during April 2009 (the starting period of the H1N1 pandemic) to compare the performance of a spatial algorithm and a time series algorithm. The hypothesis in this paper is that current existing spatial disease detection algorithms cannot detect outbreaks earlier than the time series algorithm if their underlying statistic models have the one-cluster assumption.

## Methods

### Spatial algorithms for outbreak detection

The common statistical models for spatial detection algorithms include the likelihood ratio (used in KSS and FSS) and the Bayesian posterior probability (used in BSS and RSC). Both models presume that there is only one cluster of an outbreak at a time possible within the whole study region [8,11,12]. However, this assumption would be violated in circumstances where a disease simultaneously spread from multiple separate geographic areas. In the following, we prove that both models will be defective given the assumption of one outbreak cluster is violated. Table 1 lists the main symbols used in this paper and their respective meanings.

**Table 1**: List of symbols.

| Symbol | Meaning |
|---|---|
| $S$ | a region of interest |
| $C_G$ | the total observed counts in the entire study region $G$ |
| $C_S$ | the summed observed counts in the areas within cluster $S$ |
| $C_{G-S}$ | the summed observed counts in the areas outside cluster $S$ |
| $B_G$ | the total expected counts in the entire study region $G$ |
| $B_S$ | the summed expected counts in the areas within cluster $S$ |
| $B_{G-S}$ | the summed expected counts in the areas outside cluster $S$ |
| $c_i$ | the observed counts in area $i$ |
| $b_i$ | the expected counts in area $i$ |
| $q_1$ | $q_1 = C_S/B_S$, infection rate within cluster $S$ |
| $q_1'$ | $q_1' = C_{G-S}/B_{G-S}$, infection rate outside cluster $S$ |
| $q_0$ | $q_0 = C_G/B_G$, infection rate within the entire region $G$ |

### 1) Likelihood Ratio Model

The frequentist approach in spatial scan statistics models uses likelihood ratio ($LR(S)$) and p-value to determine an outbreak region. Equation (1) is the likelihood ratio ($LR(S)$) of having an outbreak in a region $S$ against having no outbreak in the entire study region [8]. If $LR(S)$ is greater than 1 and its p-value is the most significant in the randomization test, it is likely that the cluster $S$ is the one having an outbreak.

$$\boldsymbol{LR(S) = \frac{L_1(S)}{L_0} = \frac{\left(\frac{C_S}{B_S}\right)^{C_S}\left(\frac{C_{G-S}}{B_{G-S}}\right)^{C_{G-S}}}{\left(\frac{C_G}{B_G}\right)^{C_G}} = \frac{{q_1}^{C_S}{q_1'}^{C_{G-S}}}{{q_0}^{C_G}}} \quad (1)$$

Under the one outbreak cluster assumption, region $S$, with more observed counts due to the outbreak and no outbreaks outside of $S$, is likely to have $LR(S) \gg 1$ given ${q_1}' \approx q_0$ and $q_1 \gg q_0$. However, when we consider a scenario where multiple ($K$) outbreak clusters are occurring concurrently in several geographically separate regions, we define $q_k = C_{S_k}/B_{S_k}$ as the infection rate for each cluster region $S_k$ and $k = 1, \cdots K$, and we assume $q_k \approx q_1$ for $k = 2, \cdots K$. The likelihood ratio for region $S$ becomes close to 1 as the cluster number $K$ becomes greater (*e.g.*, $K \to \infty$) and the outbreak regions become to cover the entire study region, $\sum_{k=1}^{K} S_k \to G$, as shown in Eq. (2):

$$\begin{aligned}\lim_{K\to\infty} \boldsymbol{LR(S)} &= \lim_{K\to\infty} \frac{{q_1}^{C_{S_1}}\left(\frac{\sum_{k=2}^{K} q_k B_{S_k} + q_1' B_{G-\sum_{k=1}^{K} S_k}}{\sum_{k=2}^{K} B_{S_k} + B_{G-\sum_{k=1}^{K} S_k}}\right)^{C_{G-S_1}}}{\left(\frac{\sum_{k=1}^{K} q_k B_{S_k} + q_1' B_{G-\sum_{k=1}^{K} S_k}}{\sum_{k=1}^{K} B_{S_k} + B_{G-\sum_{k=1}^{K} S_k}}\right)^{C_G}} \\ &= \lim_{K\to\infty} \frac{{q_1}^{C_{S_1}}\left(\frac{\sum_{k=2}^{K} q_k B_{S_k}}{\sum_{k=2}^{K} B_{S_k}}\right)^{C_{G-S_1}}}{\left(\frac{\sum_{k=1}^{K} q_k B_{S_k}}{\sum_{k=1}^{K} B_{S_k}}\right)^{C_G}} \\ &= \frac{{q_1}^{C_{S_1}}\left(\frac{\sum_{k=2}^{K} q_1 B_{S_k}}{\sum_{k=2}^{K} B_{S_k}}\right)^{C_{G-S_1}}}{{q_1}^{C_G}} = \frac{{q_1}^{C_{S_1}}{q_1}^{C_{G-S_1}}}{{q_1}^{C_G}} = \boldsymbol{1}\end{aligned} \quad (2)$$

Equation (2) shows that given any cluster region $S_k$ with $q_k \approx q_1$ and a large $K$, the likelihood ratio of having an outbreak approaches 1, which means there is no difference between the non-hypothesis and the alternative hypothesis even though both $q_k$ and $q_0$ are much higher than 1 (*e.g.*, their observed values are much higher than baseline values); consequently, this leads to the incorrect result of no found outbreak instead of multiple probable outbreak clusters.

### 2) Bayesian Gamma-Poisson Model

The Bayesian approach in spatial scan statistics uses the posterior probability of a region having an outbreak to determine an outbreak region, as shown in Eq. (3), which requires three variables: $P(D|H_1(S))$, $P(H_1(S))$ and $P(D)$. $P(D|H_1(S))$ in Eq. (4) is the likelihood of the alternative hypothesis $P(H_1(S))$ (i.e., having an outbreak in region $S$) based on a Gamma-Poisson model.

$P(H_1(S))$ is the prior probability of having an outbreak in any region $S$, while $P(H_0)$ represents the prior of having no outbreaks. In our study, we estimated $P\left(H_1(S)\right) = P_1/M$ by assuming a uniform distribution among all possible $M = |\mathrm{Z}|$ regions having an outbreak and used $P_1$ to denote the prior probability of having an outbreak cluster in the entire study region. Given $P(H_0) + \sum_{S\in \mathrm{Z}} P(H_1(S)) = 1$ where $P(H_0) = 1 - P_1$, the probability of data in Eq. (5), $P(D)$, is computed as the sum of two components: one is the sum of the likelihood of all possible clusters multiplied by their priors $P\left(H_1(S)\right)$, and the other is the sum of the likelihood of the non-hypothesis multiplied by the prior $P(H_0)$.

$$P(H_1(S)|D) = \frac{P\left(D\middle|H_1(S)\right)P(H_1(S))}{P(D)} \tag{3}$$

$$P\left(D\middle|H_1(S)\right) = \varphi \cdot \frac{{\beta_S}^{\alpha_S}\Gamma(\alpha_S + C_S)}{(\beta_S + B_S)^{\alpha_S + C_S}\Gamma(\alpha_S)} \cdot \frac{{\beta_{G-S}}^{\alpha_{G-S}}\Gamma(\alpha_{G-S} + C_{G-S})}{(\beta_{G-S} + B_{G-S})^{\alpha_{G-S} + C_{G-S}}\Gamma(\alpha_{G-S})} \tag{4}$$

$$P(D) = P(D|H_0)P(H_0) + \sum_{S\in \mathrm{Z}} P\left(D\middle|H_1(S)\right)P(H_1(S)), \tag{5}$$

where $\varphi$ is a constant factor.

We will focus on the computation of the likelihood $P(D|H_1(S))$ since $\frac{P(H_1(S))}{P(D)}$ is a constant across any cluster in Eq. (3). Consider a specific case where there are $K$ ($K \gg 1$) outbreak regions and the summed time series for each region follow a Poisson distribution with the same conjugate Gamma priors (*i.e.*, $q_1 \sim \Gamma(\alpha, \beta)$) and have expected values and observed values which are close to each other (*e.g.*, $C_{S_i} \approx C_{S_j}$ and $B_{S_i} \approx B_{S_j}$, where $i, j = 1, \cdots, K$), respectively. The likelihoods of all the candidate clusters will thus result in the values, $P\left(D\middle|H_1(S_1)\right)$, $P\left(D\middle|H_1(S_2)\right)$, $\cdots$, $P\left(D\middle|H_1(S_K)\right)$, where $P\left(D\middle|H_1(S_i)\right) \approx P\left(D\middle|H_1\left(S_j\right)\right)$ for $i, j = 1, \cdots, K$ which we denote $P\left(D\middle|H_1(\cdot)\right)$ for any $P\left(D\middle|H_1(S_i)\right)$. The posterior probability of each one of the $K$ regions in Eq. (3) can be derived in Eq. (6) when $K$ becomes large, where $Q$ is the summation of the likelihood of all the possible regions other than the $K$ clusters multiplied by the priors.

$$\begin{aligned}\lim_{K\to\infty} P(H_1(S)|D) &= \lim_{K\to\infty} \frac{P(D|H_1(\cdot)) \times \frac{P_1}{N}}{K \times P\left(D\middle|H_1(\cdot)\right) \times \frac{P_1}{N} + Q + P(D|H_0) \times (1 - P_1)} \\ &= \lim_{K\to\infty} \frac{P(D|H_1(\cdot)) \times \frac{P_1}{N}}{K \times P\left(D\middle|H_1(\cdot)\right) \times \frac{P_1}{N}} = \lim_{K\to\infty} \frac{1}{K} = 0\end{aligned} \tag{6}$$

If $K$ is large, such as during a pandemic period, the posterior probability $1/K$ for each examined region would approach zero, which indicates false negatives for outbreaks. Although it is an extreme example, we are addressing the issue of lowered posterior probabilities of outbreak regions due to the violation of one-outbreak-region assumption. Thus, the Bayesian Gamma-Poisson model can be challenged as well.

**An Example: 2009 H1N1 Flu Pandemic**

Beginning in the state of Veracruz, Mexico, the 2009 H1N1 flu outbreak spreads quickly and globally. In the U.S., within about 3 weeks, the H1N1 virus became widespread in 8 states and

infected about 5,000 people [16]. Compared with the previous year, this outbreak provided much stronger signals captured in the CDC seasonal ILI trend (Fig. 1), thus it was selected as a real-life example to test our hypothesis.

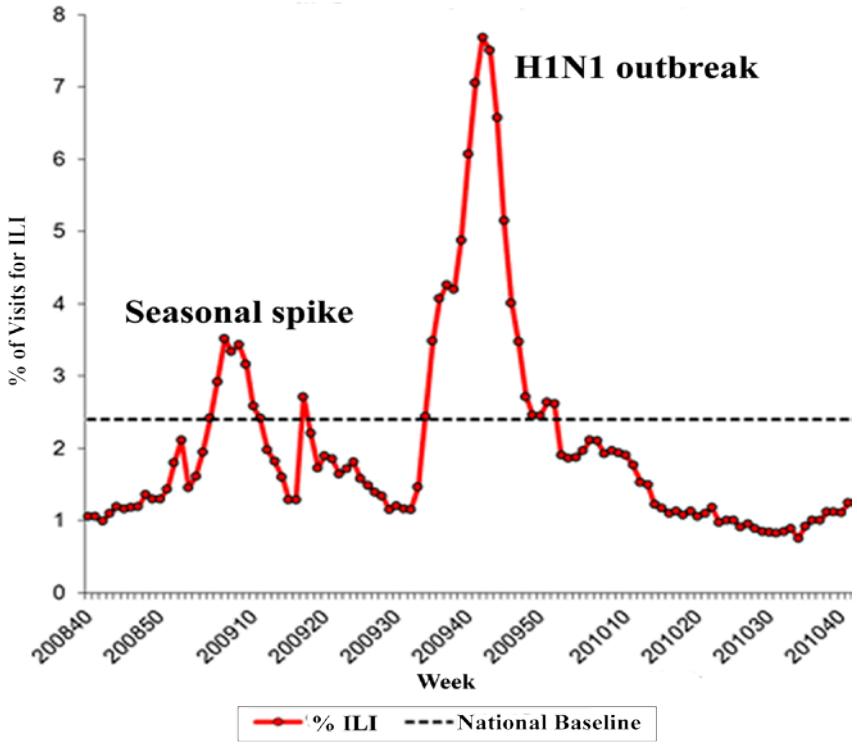



**Figure 1:** Percentage of visits for Influenza-Like Illness (ILI) reported by the U.S. outpatient ILI surveillance network (ILINet), weekly national summary, September, 2008 – April, 2009 (http://www.cdc.gov/flu/weekly)

We applied BSS (a spatial algorithm) and WAD (a time series algorithm) to detect the significant increase of thermometer sales which may indicate the onset of the H1N1 pandemic that occurred in the state of Texas at the end of April, 2009. The reasons for choosing these two algorithms for this study are 1) BSS is a spatial algorithm preferred over the frequentist algorithms and has been tested and applied in the applications of prospective disease surveillance [11,13,17]; and 2) WAD is a well established time series algorithm which has been evaluated in multiple studies and found favorably compared with other time series algorithms [7,18]. The goal for these two algorithms in this study is to find emerging clusters of geographic areas having significantly increased thermometer sales which may indicate increased H1N1 infections in the early stage of the pandemic.

### 1) Study Dataset

In this study, we used routinely collected over-the-counter (OTC) sales data as the data source for detection of flu outbreaks. Because previous studies from literature have demonstrated that OTC data such as cold remedies and diarrhea remedies sales can serve as good indicators for

outbreak detection and are more timely than physician diagnosis, these data were used as an influenza outbreak indicator [19,20,21,22,23,24]. We obtained the dataset from the National Retail Data Monitor (NRDM) system [25] developed by the RODS laboratory, which has been collecting OTC sales data from 30,000+ retail stores across the country on a daily basis since 2003. The NRDM classifies each retail product sale into one of twenty three categories, taking into account both purpose of the treatment and consumer age group, such as anti-fever adult or cold relief pediatric.

We chose to study the data from Texas for the following reasons: 1) Texas is one of the two states (the other is California) where H1N1 pandemic was confirmed and identified in the early stages, 2) Texas shares the longest border line with Mexico, where H1N1 was first identified, and 3) Texas is the 2$^{nd}$ largest state and it also has the second largest population in the U.S..

Among the 23 OTC categories NRDM provides, we chose the thermometer sales category as our indicator of the flu outbreak for three reasons. First, the RODS disease surveillance system signaled enormous spikes in the time series data of over-the-counter (OTC) thermometer sales in Texas at the end of April corresponding to the chronology of the 2009 H1N1 pandemic (Fig. 2). Second, we found a strong correlation (correlation coefficient is 0.91 with 95% confidence interval [0.89,0.92]) between patients with constitutional syndrome visiting emergency departments (EDs) and OTC thermometer sales in Pennsylvania in the past flu seasons as shown in Fig. 3. Finally, Villamarin et al. also demonstrated high correlation (0.89) between actual and predicted ED visits using thermometer sale data [26].

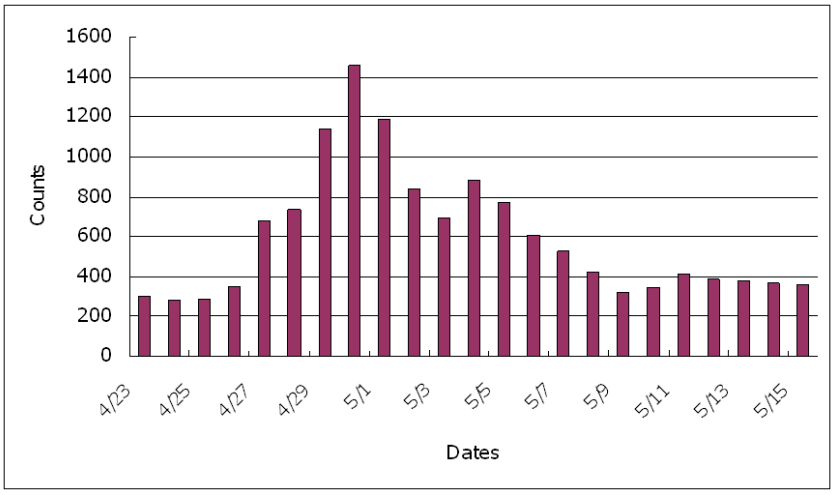



**Figure 2:** Total thermometer sales in Texas collected from NRDM between 4/23/09 and 5/15/09.

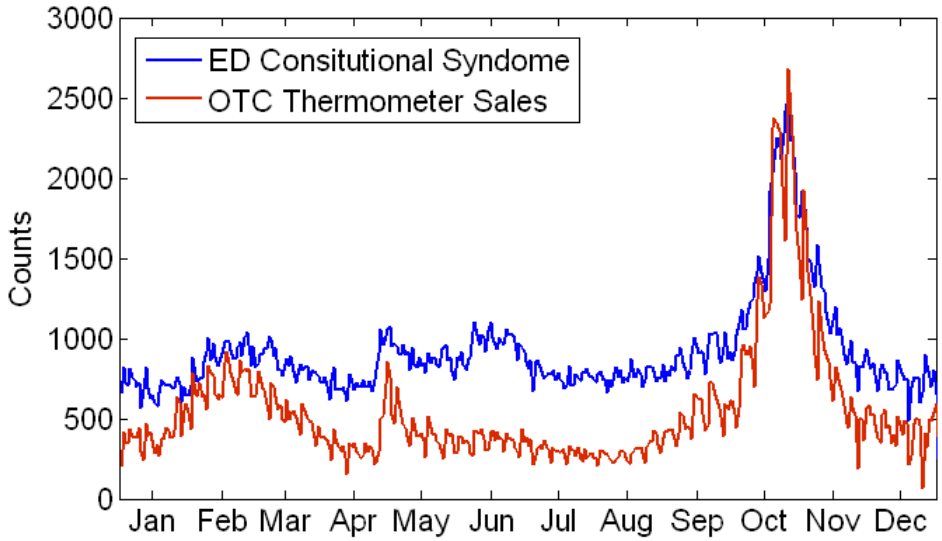



**Figure 3:** Time series of patient ED visits with constitutional syndrome and counts for OTC thermometer sales in the state of Pennsylvania in 2009.

Our experimental dataset covers purchases made between March 1, 2007 and May 31, 2009. The thermometer sales data included records from 1,413 pharmacy stores in 581 ZIP code areas. The data prior to April 1, 2009 were used to train algorithms and estimate false alarm rates while data from April 1, 2009 to May 31, 2009 was used to evaluate the algorithms (evaluation period).

Although we conducted a retrospective study, a prospective analysis was mimicked by incrementally adding each day's data to the algorithms as during the evaluation period. For each day, the algorithms use previous 730 days to predict the current day's sales and estimate the values of prior parameters (*e.g.*, $\alpha$'s and $\beta$'s in Gamma distribution used in the Bayesian spatial detection algorithm). In order to define the alert threshold, $P_t$, used by the detection algorithms, we applied the analysis to each of the 365 days between April 1, 2008 and March 31, 2009 and recorded the highest score for each day. This allows us to calculate the false alarm rate as $i/365$ if $p_t$ is the $i$-th greatest in the set of 365 recorded scores. The assumption here is that during the 3-year period before April 1, 2009, there were no H1N1 pandemics. But note that the ignored other strains of flu that did occur in this period will result in an underestimation of algorithm performance.

### 2) Date of Outbreak Onset

Figure 4 shows the accumulation of confirmed H1N1 cases in Texas after April 23, 2009 (when CDC started counting cases) as posted on CDC's official website [16]. Within about 3 weeks, the H1N1 virus spread quickly and infected more than 500 people in Texas.

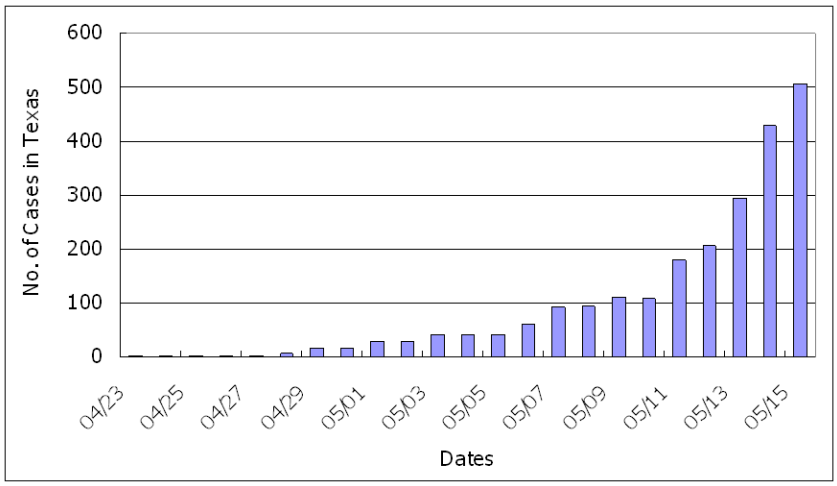



**Figure 4**: Cumulative number of confirmed H1N1 cases in Texas reported to CDC between 4/23/09 and 5/15/09

In this study, the date of April 24, 2009 is considered to be an indicator of the onset of the H1N1 flu pandemic hitting Texas and is used to evaluate the detection algorithms. On this day, the World Health Organization (WHO) issued the first Disease Outbreak Notice on the H1N1 flu pandemic, confirming the infection of a number of people in Mexico and the United States [27]. On the same day, the US Centers for Disease Control and Prevention (CDC) announced that 7 of the 14 Mexican samples contained the same virus strain as that found in California and Texas and suggested that the containment in the USA was "not very likely" [28]. Another appropriate date can be considered as the pandemic onset is the date when the first case was confirmed in Texas. Since we are comparing the relative timeliness between different detection algorithms, the absolute difference between each algorithm and the real onset date will not affect our findings.

### 3) Algorithm Evaluations

We evaluated the *Bayesian spatial scan statistic* (BSS) and the *Wavelet anomaly detection* (WAD) algorithms using data from the real OTC thermometer sales data from the 581 ZIP code areas in Texas where reporting stores were located.

We applied WAD to these 581 ZIP code areas and chose the threshold based on the false alarm rate computed from the training data set.

BSS was applied to the same data sets. We laid a 24x24 grid (576 grid cells) over a Texas state map that included the 581 ZIP code centroids. All of the resulting rectangles, of varying sizes and covering different locations, were examined. The baseline (expected count) for each ZIP code was estimated using a wavelet transform from the previous two years of data (as used in WAD). By assuming a uniform prior distribution, this approach computes the posterior probability of having an outbreak using the Gamma-Possion model we described earlier in the section. The cluster with the highest posterior probability is reported if it exceeds a threshold, which is estimated based on a preset false alarm rate described in the subsection of Study Dataset.

We examined the detection timeliness of both the WAD (*Wavelet anomaly detector*) and the BSS (*Bayesian spatial scan statistic*) algorithms, respectively. We chose two false alarm rates: one false alarm per two months and one false alarm per month, allowing us to evaluate the two algorithms from a practical perspective as resources are limited to public health officials. The false alarm rate reflects how often an alarm is triggered by chance, assuming that analyses are repeated on a regular basis with a periodicity equal to the specified time interval length, e.g., on a daily basis.

## Results

The results of WAD are illustrated in Fig. 5 and Fig. 6. Fig. 5 shows the number of significantly elevated ZIP code areas in Texas between the April 23 and May 15 in the evaluation period. In addition, Fig. 6 shows the geographical distribution of the ZIP code areas with significantly elevated thermometers sales from April 25, 2009 to April 30, 2009, using the same color scheme to represent the two significance levels. As shown in both figures, the number of significantly elevated ZIP code areas stayed low (less than 20 ZIP codes) before April 27. From our past experience, the spikes shown in these areas probably resulted from an imperfect data collection process or from some other stochastic reasons (*e.g.*, non-continuous data reporting from some stores) since the spikes are distributed randomly and only last for a day (see Fig. 6(a), 6(b)). However, starting from April 27, 66 ZIP code areas simultaneously signaled alarms. Furthermore, the thermometer sales stayed significantly high for several days, until around May 7, and the number of significantly elevated ZIP codes exceeded 100 on April 29 and April 30 (in Fig. 6(c-f)). More specifically, the same ZIP codes in the counties Nueces, Travis, Bexar, Collin, Dallas and later El Paso, Bowie, Tarrant and Cameron, repeatedly reported significantly elevated thermometers sales within those days. These results suggest that WAD was able to detect the significant increase on April 27, 2009, by identifying 66 ZIP code areas in about 20 counties showing a significantly elevated amount of thermometer purchases from those drug stores under surveillance.

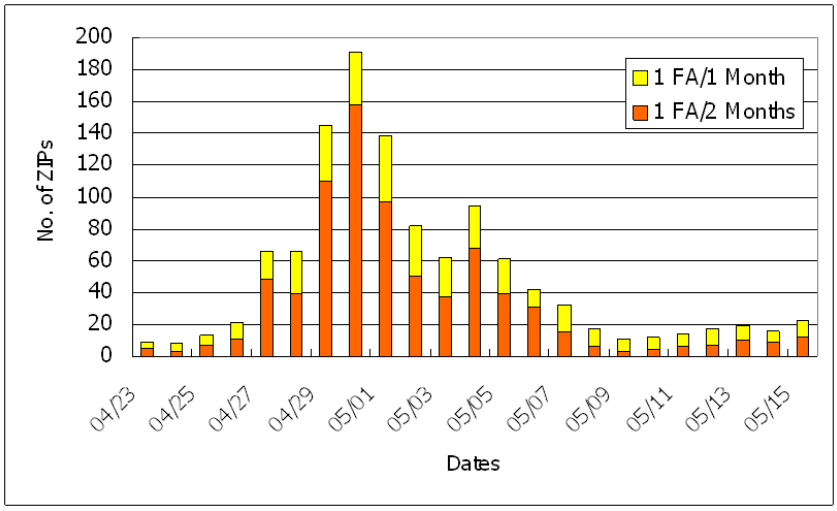


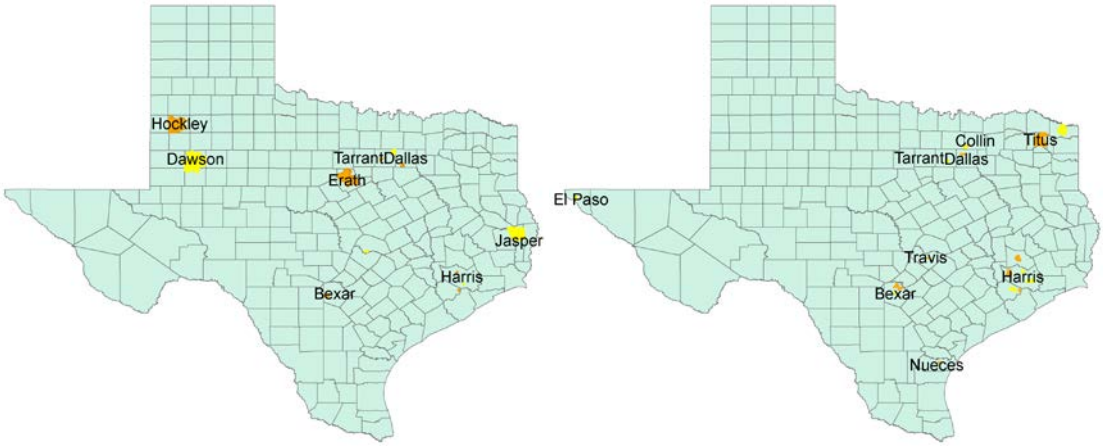

**Figure 5:** The number of significantly elevated ZIP code areas in Texas analyzed by Wavelet transform between 4/23/09 and 5/15/09. The orange bars represent the numbers of elevated ZIP codes with scores equal to or larger than the threshold score, which allows only one false alarm

(FA) per two months by chance. The yellow bars represent the numbers of the ZIP codes having a significant increase of cluster posterior probabilities corresponding to FA between once per two months and once per month.

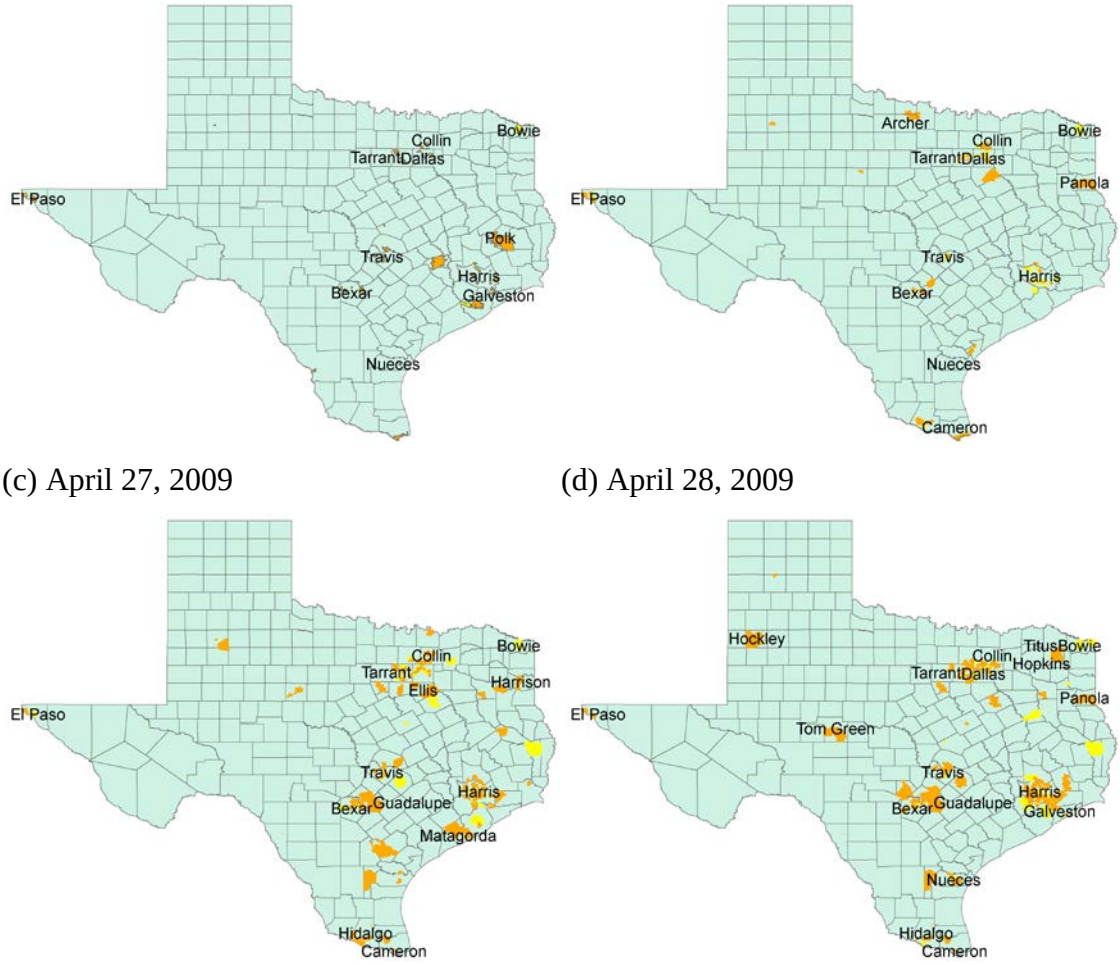
BSS, on the other hand, showed different results from WAD when analyzing the same set of data. The first day BSS fired an alarm was April 30, 2009 which was 3 days later than WAD did. It reported a cluster including 409 out of total 581 ZIP codes with the posterior probabilities above the threshold corresponding to one false alarm per month (Fig. 7(a)). The cluster was located in the eastern central part of Texas. On the next day, May 1, 2009, BSS fired the 2nd alarm with a cluster comprising 272 ZIP codes with significance of one false alarm per two months (Fig. 7(b)). Fig. 7 shows that the cluster had become more localized and moved to the southeast. However, during the first 6 days of the real outbreak, from April 24, 2009 to April 29, 2009, BSS did not identify any significant cluster. It is also worth noting that after May 1, although the counts of thermometers sales were still high, BSS did not fire any alarms at all for the rest of days within the evaluation period (except for May 5, when a 3 ZIP code cluster was found but was believed to be a false alarm).

## Discussion

In the example described above, we found that WAD was able to detect the significant increase of thermometer sales, which may indicate the onset of H1N1 pandemic, 3 days earlier than BSS did. The slow timeliness of BSS in this study compared with WAD can be attributed to the violation of the basic assumption that only one outbreak cluster can occur at the same point in time in different geographic regions. Therefore, in order to rapidly detect a pandemic, such as the 2009 H1N1 flu outbreak, which takes place in multiple distant places in a sudden and simultaneous way, our results show that the spatial models with the one-cluster assumption are not preferred.




(a) April 25, 2009 (b) April 26, 2009



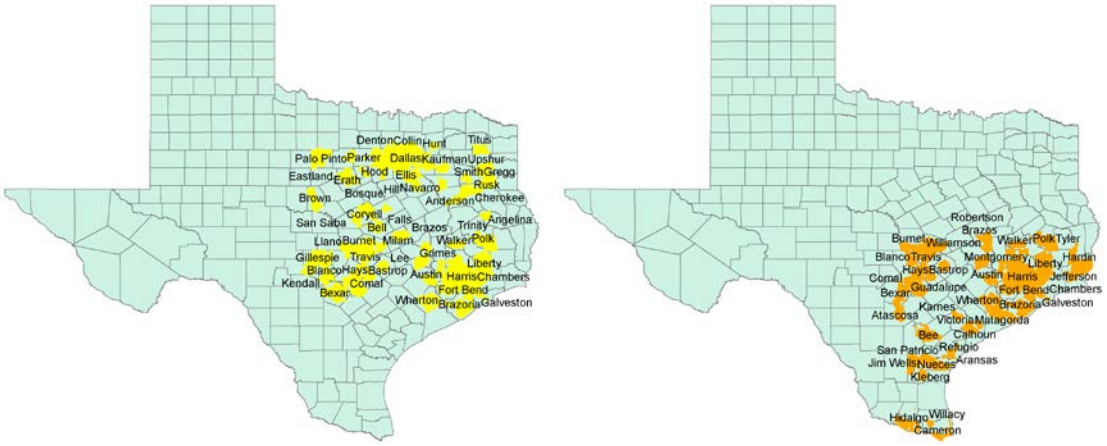

(c) April 27, 2009

(d) April 28, 2009

(e) April 29, 2009

(f) April 30, 2009

**Figure 6:** The counties in Texas with ZIP codes having significant elevated counts on the sales of thermometers. Orange areas represent the ZIP code areas with significance level of 1 false alarm per 2 months and yellow areas represent the ZIP code areas with significance level of 1 false alarm per 1 month.




(a) April 30, 2009 (b) May 1, 2009

**Figure 7:** Two significant clusters detected by BSS on April 30, 2009 and May 1, 2009. The ZIP code areas in lime are the areas covered by NRDM; the ZIP code areas in yellow are the ones in the cluster with significance level of 1 false alarm per month detected by BSS and the areas in orange are the ZIP code areas in the cluster with significance level of 1 false alarm per 2 months.

However, although the time series algorithm, WAD, was able to respond 3 days earlier than the Bayesian spatial algorithm, BSS, which suggests the time series algorithms are more sensitive to pandemic detection, it also has a few drawbacks. Time series algorithms are less robust on noisy data (*e.g.*, caused by imperfect data collection process, etc.) since they signal alarms whenever the deviation of the observed counts from the expected counts in an area exceeds a threshold. Thus, a secondary analysis is recommended to be performed. A secondary analysis may include 1) considering the output for the previous days as well to determine if the newly alarmed areas are correlated with the previous ones; 2) studying if there is a data quality problem that may cause an abnormal increase of sales; and 3) waiting for the outcome from the next time period to make better decision.

It is also important to address how strong the OTC data used in this study have the signal of real H1N1 pandemic. In the study dataset, the number of thermometer sales is much higher than the number of confirmed cases each day at the end of April. This implies that a big proportion of the elevated thermometer purchases possibly were not due to H1N1 infection but other reasons. Although the literature has shown that OTC data can be used to predict number of cases in Emergency Departments [26], research on the relationship between OTC sales and medical treatment seeking is still needed especially by the people from the public health or social behavioral fields, such as the study on the relations between hospitality and behaviors during outbreaks [29,30].

## Limitations

Both the experimental data and the applied algorithms limited this study. Our experimental design was restricted to analyzing only one type of data (i.e., the OTC thermometer sales) due to data availability. Some public health organizations or biosurveillance systems, however, may have more than one possible data source (e.g., emergency department patient visits, etc.) available, which to some extent may contain better signals of H1N1 pandemic. Also it would be ideal to analyze multiple states of data and to use ED data to create a better gold standard in terms of deciding outbreak onset period. We hope in the future to use ISDS distribute project data for a larger scale study. Furthermore, the comparison in this study was only performed between one spatial algorithm and one time series algorithm. Evaluation using other algorithms (e.g., moving average, the spatial scan statistic, etc.) would be helpful to support our findings. In addition, this study only used one known H1N1 pandemic to test the performance of the two algorithms; thus confidence interval for timeliness is not available.

## Conclusion

We have conducted a study on the detection of significantly increased thermometer sales which may indicate 2009 H1N1 pandemic in Texas by applying a time series algorithm and a spatial algorithm. Although the spatial algorithm was more robust (fewer false alarms) and informative (suggesting geographical distribution of outbreaks) than the pure time series algorithm, our results suggest that the time series analysis is still desirable in detection of pandemics as it may achieve a more sensitive performance with better timeliness. The use of time series algorithms, therefore, is still necessary for rapid outbreak detection, especially in scenarios where the single-cluster assumption does not hold. Instead of replacing the time series algorithms, we suggest epidemiologists or biomedical informaticians apply time series algorithms as a complement to current spatial algorithms for public health surveillance purposes.

## Acknowledgement

The project was funded by CDC grants P01 HK000086 and 1U38 HK000063-01, and PA Department of Health grant SAP #40000012020.

## Corresponding author

Jialan Que
Department of Biomedical Informatics
University of Pittsburgh, PA
Email: jiq4@pitt.edu

## References


1. Amerithrax or AnthraxiInvestigation. , U.S. Federal Bureau of Investigation.
2. Hung LS. The SARS epidemic in Hong Kong: what lessons have we learned? The Royal Society of Medicine. 2003; 96(8): 374–378.
3. WHO. Influenza-like illness in the United States and Mexico. [Online].; 2009. Available from: http://www.who.int/csr/don/2009_04_24/en/index.html.
4. Rongkavilit C. Importance of immunization: a serious disease is just a plane ride away. Consultant for Pediatricians. 2010; 9(10).
5. Hulth A, Andrews N, Ethelberg S, et al. Practical usage of computer-supported outbreak detection in five European countries. Eurosurveillance. 2010; 15(36).
6. Zhang J, Tsui FC, Wagner MM. Detection of outbreaks from time series data using wavelet transform. AMIA Annu Symp Proc; 2003. p. 410-4.
7. Wagner MM, Moore WA, Ron AM. Handbook of biosurveillance. Burlington: Elsevier Academic Press; 2006.
8. Hunter JS. The exponentially weighted moving average. Journal of Quality Technology. 1986; 18: 155-62.
9. Buckeridge D, Burkon H, Campbell M, et al. Algorithms for rapid outbreak detection: a research synthesis. J Biomed Inform. 2005 Apr; 38(2): 99-113.
10. Kulldorff M. A spatial scan statistic. Commun Stat Theory Methods. 1997; 26(6): 1481-96.
11. Takahashi K, Kulldorff M, Tango T, et al. A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. Int J Health Geogr. 2008 Apr; 7(14).
12. Kulldorff M, Huang L, Pickle L, et al. An elliptic spatial scan statistic. Stat Med. 2006 Nov; 25(22): 3929-43.
13. Neill DB, Moore AW, Cooper GF. A Bayesian spatial scan statistic. Advances in Neural Information Processing Systems. 2005; 18: 1003-10.
14. Que J, Tsui FC. Rank-based spatial clustering: an algorithm for rapid outbreak detection. J Am Med Inform Assoc. 2011 May; 18(3): 218-24.
15. Neill DB. An empirical comparison of spatial scan statistics for outbreak detection. International Journal of Health Geographics. 2009; 8(20).
16. Que J, Tsui FC. A multi-level spatial clustering algorithm for detection of disease outbreaks. AMIA Annu Symp Proc; 2008. p. 611-5.
17. Que J, Tsui FC. A z-score based multi-level spatial clustering algorithm for the detection of disease outbreaks. Lecture Notes in Computer Science, BioSecure; 2008. p. 108-18.
18. CDC. H1N1 Flu (Swine Flu): Past Situation Updates. [Online]. [cited 2010. Available from: http://www.cdc.gov/h1n1flu/updates.
19. Jiang X, Cooper GF. A recursive algorithm for spatial cluster detectior. AMIA Annu Symp; 2007. p. 369-73.
20. Siegrist D, Pavlin J. BioALIRT biosurveillance testbed evaluation. Syndromic surveillance. ; 2004; New York. p. 152-8.
21. Hogan WR, Tsui FC, Ivanov O, et al. Detection of Pediatric Respiratory and diarrheal outbreaks from sales of over-the-counter electrolyte products. J Am Med Inform Assoc. 2003

Nov-Dec; 10(6): 555-62.

22. Magruder SF. Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease. Johns Hopkins APL technical digest. 2003; 24(4): 349-53.
23. Welliver RC, Cherry JD, Boyer KM, et al. Sales of Nonprescription Cold Remedies: A Unique Method of Influenza Surveillance. Pediatric Research. .
24. Das D, Metzger K, Heffernan R, et al. Monitoring over-the-counter medication sales for early detection of disease outbreaks --- New York City. MMWR Morb Mortal Wkly Rep. 2005; 54(Suppl): 41-46.
25. Hogan WR, Wagner MM. Sales of over-the-counter healthcare products. In Wagner MM, Moore AW, Aryel RM. Handbook of Biosurveillance. Amsterdam; Boston: Academic Press; 2006.
26. Daily L, Watkins RE, Plant AJ. Timeliness of data sources used for influenza surveillance. J Am Med Inform Assoc. 2007; 14(5): 626-31.
27. Wagner MM, Tsui FC, Espino JU, et al. National retail data monitor for public health surveillance. Morbidity & Mortality Weekly Report. 2004 Sep; 53(Suppl): 40-2.
28. Villamarin R, Cooper GF, Tsui FC, et al. Estimating the incidence of influenza cases that present to emergency departments. Emerging Health Threats Journal. 2011; 4: s57.
29. CDC. CDC Briefing on Public Health Investigation of Human Cases of Swine Influenza. [Online].; 2009. Available from: http://www.cdc.gov/media/transcripts/2009/t090424.htm.
30. Isaacs D. Lessons from the swine flu: pandemic, panic and/or pandemonium? J Paediatr Child Health. 2010; 46(11): 623-6.
31. Wong LP, Sam IC. Behavioral responses to the influenza A(H1N1) outbreak in Malaysia. J Behav Med. 2011; 34(1): 23-31.

doi: 10.5210/ojphi.v4i1.3915
Cite this item as: Que, J., & Tsui, F. 2012 May 17. Spatial and Temporal Algorithm Evaluation for Detecting Over-The-Counter Thermometer Sale Increases during 2009 H1N1 Pandemic. *Online Journal of Public Health Informatics* [Online] 4(1):e1.