Emergency Medicine: Open Access
ISSN: 2165-7548

Sparks et al., Emerg Med (Los Angel) 2015, 5:3
DOI: 10.4172/2165-7548.1000254

Research Article Open Access

Syndromic Surveillance using Twitter Data

Ross Sparks[*], **Mark Cameron, Sam Woolford, Bella Robinson, Robert Power and John Colton**

The Commonwealth Scientific and Industrial Research Organisation, Australia

[*]**Corresponding author:** Ross Sparks, Team Leader of Real-Time Modelling and Monitoring, Quantitative Risk Group, Digital Productivity Flagship, Riverside Life Sciences Centre, 11 Julius Avenue, North Ryde, NSW 2113, Australia, Tel: +61 2 9325 3262; E-mail: ross.sparks@csiro.au

Received date: March 06, 2015; **Accepted date:** March 16, 2015; **Published date:** March 23, 2015

Abstract

The Federal Department of Health website in Australia on 22 October reported that in 2014 influenza A (H1N1) dominated across most jurisdictions throughout the season, however influenza A (H3N2) was predominant in New South Wales and the Australian Capital Territory, with late season increases noted in Queensland, Western Australia, the Northern Territory and Tasmania. This paper used daily counts of flu symptom tweets to describe the differences in the outbreaks across the various states and territories in Australia for the 2014 season. However, the tweet data indicates a two wave flu outbreak for Victoria that was not picked up in this report. This paper also illustrates that the flu outbreaks for flu in some states were no different to the 2013 flu season while in others it was significantly greater for a long period. The varying nature of the burden of this disease in Australia is outlined in this paper but the variable nature of the outbreaks across Australia is difficult to explain or understand. Future efforts should be devoted to understanding these differences.

Keywords: Influenza A; Syndromic surveillance; Twitter data

Introduction

This paper tests the application of statistical process control (SPC) methods to Twitter data in order to facilitate the early detection of flu outbreaks [1]. The importance of such syndromic surveillance in the early detection of disease outbreaks and other public health events was discussed in Sparks et al. [2]. Examples of these are early identification of unintentional contamination of the food or water supply where correcting the problem early avoids others being affected. Daily syndrome group counts reflect the additive effect of all co-circulating pathogens and understanding the burden of outbreaks is epidemiologically important.

Twitter, an online social media network, allows users to keep friends, family and co-workers up-to- date on their status in real time using short messages of up to 140 characters, called 'tweets', sent through computers and mobile phones. The main advantage of tweets is we capture early symptoms of flu as they present, before they are serious enough to go to emergency departments. This approach of linking Twitter data and SPC could provide a syndromic surveillance methodology for the early detection of flu events and the tweets help in tracking the burden of the disease. For example, Twitter data can provide additional contextual information not typically available from emergency department reporting.

This research considers public health conditions represented by flu. Text mining is used to determine the frequencies of occurrence of symptom keywords, or groups of words, for flu from individual tweets. For each health condition, an hourly frequency of tweets containing any of the associated symptom keywords for that health condition is established for Twitter messages sent in Australia over the period from July 1, 2013 to September 30, 2014. When there is no outbreak of a specific health condition, the frequency of counts would be expected to demonstrate the characteristic of a stable process (not necessarily at the zero level) over time. Frequency counts that are stable, and thus predictable, are defined to be in-control and therefore would be expected to be manageable with the generally available public health resources.

SPC monitoring is utilized to detect when the frequency counts of symptom keywords becomes unstable and increases beyond what would be expected for the in-control process. This could suggest an outbreak of the health condition which could result in a need for additional public health resources. Dynamic biplots are used for their capability to detect changes in multivariate data and their capability for early detection with relatively small numbers of counts. This multivariate technique will be used to track the variable start of the flu seasons in the various states of Australia. To demonstrate the efficacy of this approach, the data are split into a training period (1 July 2013 to 5 March 2014 because in 2014, like the H1N1 outbreak in 2009, the flu season started much earlier than usual) to design a surveillance methodology for the test period (6 March 2014 to the end of October 2014).

Our findings suggest that we are able to detect moderately small outbreaks within a few days and for large outbreaks within hours. Finally these results suggest that the methodology could be adapted to provide a real time surveillance approach that could be applicable for early detection and control of a variety of public health events including possible bioterrorism events. We are currently adapting the approach described in this paper to develop such a real time detector of potential syndromic events

The paper is organized in the following way. Section 2 provides detail on how the data was generated. Section 3 discusses the methodologies utilized in the analysis. Section 4 provides the results of analysis and Section 5 discusses the results. The appendices list the symptom keywords and phrases used to identify tweets potentially describing flu and present a guide to interpreting the biplots used to present our results.

Data

When people tweet that they are unwell, it is likely to be in terms of the symptoms they are experiencing. The real-time nature of Twitter and the objective of surveillance to quickly identify health risks suggest that the Twitter data should be analyzed for symptoms not conditions.

The symptom keywords used to identify flu are listed in Appendix A. Various phrases using the keyword were also included in an attempt to accommodate the various ways that the keyword might be used in a tweet. At the same time an attempt was made to eliminate any keywords or phrases that might be used to reflect something other than an individual's well-being (e.g. 'that person makes me feel sick' rather than 'I feel sick').

In order to create the frequency count data for flu, tweets originating from Australia from July 1, 2013 to October 31, 2014 were examined. These tweets have been captured by CSIRO as part of the Emergency Situation Awareness (ESA) project which collects tweets from Australia and New Zealand to identify unexpected incidents, to monitor ongoing emergency situations and provide access to an archive to explore past events. ESA is operated using a map based interactive web site and has processed nearly 2 billion Tweets since September [1]. Text mining was performed on this repository to determine the number of tweets that contained any of the symptom keywords or phrases associated with each condition during the target period. By only looking at the counts some information in the tweets was lost however, it simplified the text mining task and preserved the privacy of the Twitter user. These counts were aggregated hourly over the period from July 1, 2013 to October 31, 2014 to generate the data for each condition.

The volume of tweets is fairly stationary across the duration on the study period apart from the few times when the tweet capture system failed to operate and so we have chosen to ignore it in this paper. Technically the volume of tweets could be considered as an offset in our Poisson regression models, but we argue that the volume of tweets change because interesting topics arise which are unrelated to tweets of flu, and therefore we consider adjusting for the volume as inappropriate.

Methodology

Frequency of flu related tweet counts were analyzed using descriptive statistics to identify any underlying patterns and trends over time. Then each series was divided into two components: training data representing the frequency counts over an entire year from 30 June 2013 to 5 March 2014 and test data representing the frequency counts from 6 March 2014 to 26 October 2014 to be used to test whether the SPC methodology could detect changes in the frequency flu tweeted counts.

Separate Poisson models were fit to the training data for counts from each state or territory in Australia. We adjusted for the usual sources of variation due to the hour of the day, day of week, and date of the year (season) in order to remove broad based trends and cycles while capturing local changes in the residuals. The models utilize harmonic terms to capture the within day variation and any annual variation and indicator variables to fit the day of the week effects as shown in equation 1, where μi is the expected flu count for the ith state, dayt represents the day of the year for the tth condition count, Ij represents the specific day of the week j influence on the flu counts and the sum is from j=1 to 7. The models also allow for interaction terms.

$$\ln\ (\mu it\ )=\beta 0i+\ \beta 1idayt+\beta 2icos(2\pi\times dayt/365.25)+\beta 3isin(2\pi\times dayt/365.25)+\Sigma\ \alpha jIj \qquad 1$$

The use of SPC for the detection of process change has its origins in manufacturing processes in the 1920's and has developed typically around manufacturing applications [2]. More recently SPC has found application in many areas including healthcare process improvement. SPC typically uses data associated with a process to determine whether the process is operating in-control over time, e.g. based on the average and standard deviation of data associated with the process, and to determine when it goes out of control, e.g. the average and/or standard deviation of the process data is no longer consistent with the in-control process. Ideally, an SPC method should identify an out-of-control process as quickly as possible (implying a high degree of sensitivity) and yet not identify many false negatives (implying a high degree of specificity).

The process of people communicating through Twitter regarding the symptoms they experience suggests that the number of tweets each day is representative of tweeters with flu symptoms. Consequently, outbreaks of the condition can be associated with increases in the condition frequency counts beyond what would be expected for the stable in-control condition counts.

Figure 1 presents the number of laboratory confirmed cases of flu in the whole of Australia. The red line in the figure marks March 2014, which is significant in terms of the detection of a Twitter outbreak of people talking about flu. If Twitter can identify an outbreak well before this outbreak presents at an Emergency Department then it may be an early warning system for influenza outbreaks. In our study period, the laboratory confirmed cases first indicate an outbreak in July 2014.

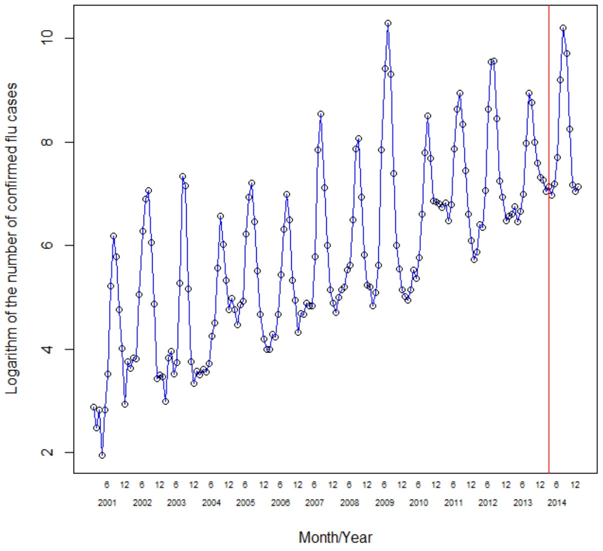

Figure 1: Number of monthly laboratory confirmed cases of influenza in Australia since 2001.

It is reasonable to assume that in various geographical regions flu outbreak will occur at different times of the flu season with some occurring simultaneously while others indicating that they are not different from previous years. In this case there may be additional power for detection of an out-of-control event by using a multivariate detection method that break down the different geographical regions. The dynamic biplot Sparks et al., [3] is demonstrated as a useful way of understanding the differences in outbreak timing across the various geographical regions. The dynamic biplot provides a multivariate approach to monitor all the flu counts from different states simultaneously. The biplot consists of two plots. The observation plots used to flag when the vectors of counts depart significantly from their forecasts by a group of recent points clustering to one side of the origin, and the variable plot which is used to describe which variables contributed to this departure. In addition, the multivariate method has the advantage of controlling the overall false detection rate for the whole of Australia.

The dynamic biplot used in this paper, as detailed in Sparks et al. [3], uses a moving window of data made up of a certain number of consecutive hourly or daily forecast errors. We start with training data that is used to define the baseline measurements associated with the in-control multivariate process and then, by moving the time window to include the next hour(s) we can detect changes in the condition counts. This paper detects and describes changes in the flu season onset around Australia by examining the standardized one day ahead forecast residuals. Multivariate tests, using Australia wide flu counts, are used to identify states or groups of states that differ from 2013's flu outbreak. The observation plot is used to identify overall shifts from the origin by flagging departures from the in-control pattern represented by the training data. The variable plot of the biplot is used to identify the nature of the outbreak and describe the characteristics of any multivariate changes from the training data. This is carried out in terms of the geographical counts in this paper. The dimensions of the variable plot represent the first two principal components of the standardized day ahead forecast errors. Univariate tests are used to identify changes in mean, variance, the mean squared error and the correlation of the individual flu counts and significant results are indicated on the variable plot by changes in the lengths and relative orientations of the vectors representing the standardized ahead forecast errors. A guide to the specific visualization cues available on the biplot are indicated in Appendix B.

Results

The Poisson models in equation 1 were fit to the training data for each geographical region of Australia and all the models had significant explanatory variables. The results for the fit of the main effects are reported in Table 1.

Syndrome	ACT	NSW	NT	QLD	SA	TAS	VIC	WA
Constant	0.787***	3.9428***	-0.97**	3.0857***	1.9441***	2.106***	3.4890***	2.4643**
Monday	0.082	0.1379***	0.086	0.1611***	0.1440*	0.0108	0.1363***	0.1217*
Tuesday	0.180	0.1736***	0.597*	0.0836	0.1587*	0.1502*	0.1757***	0.2013***
Wednesday	-0.022	0.1174***	0.383	0.0644	0.1011	0.0115	0.0988**	0.1554**
Thursday	0.053	0.0745*	0.301	0.1719***	0.0401	0.0512	0.1858***	0.1187*
Saturday	-0.757***	-0.2424***	0.058	-0.1924***	-0.3342***	-0.0832	-0.1996***	-0.0634
Sunday	-0.426**	-0.1372***	-0.058	-0.1582***	-0.1280	-0.2296**	-0.1152**	-0.1201*
time	0.0004	-0.0013***	-0.001	-0.0008***	0.000	0.0001		
cos(2π time / 365.25)	0.432***	0.2430***	0.290*	0.2407***	0.2549***	0.1902***	0.2987***	0.1755***
sin(2π time / 365.25)	-0.020	-0.1383***	-0.224	-0.2858***	-0.0765	0.0902*	-0.0940	0.0685*
*Significant at 5% level, ** significant at 1% level and *** at the 0.5%								

Table 1: The fitted Poisson regression models for total flu counts in Australia using the training data.

The models fitted to the training data should be able to predict usual behavior in the hourly counts of symptoms for each condition. If the day-ahead flu counts are predictable using the past 9 months to 15 months of flu counts, then the flu season for 2014 does not depart significantly from that of 2013. Therefore persistent high-end departures of these counts from their forecasted values are likely to flag outbreaks provided there is enough power to detect that departure.

The dynamic biplot uses as training data the Pearson residuals for the Poisson regression model for the states and territories of Australian Capital Territory (ACT), New South Wales (NSW), Northern Territory (NT), Queensland (QLD), South Australia (SA), Tasmania (TAS), Victoria (VIC) and Western Australia (WA) for 2013's winter period in July, August and September. The test data was taken from 6th March to end of October 2014. The dynamic biplot for 2013's flu season is included in Figure 2. This plot explains 50% of the variation in the 2-dimensions representation. For the 2013 flu season the outbreaks in Tasmania (TAS) and Western Australia (WA) were correlated, and the remaining states were correlated with each other (the small acute angle between the vectors in the variable plot in Figure 2 is proportional to the correlation). In addition, since WA and TAS are almost orthogonal to the other states and territories, their 2013 outbreaks are uncorrelated with these other geographical regions. This maybe because these two states are least connected to the other parts of Australia. The first dimension of the plot (which explains 30% of the variation) mostly explains the variation in all regions except TAS and WA, whereas the second dimension explains mostly TAS and WA's variation.

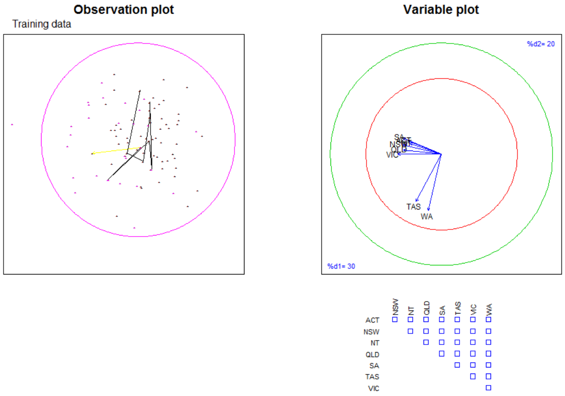

Figure 2: Biplot of the training data for flu season in 2013.

The biplot, using one-day-ahead Pearson forecast errors, indicates when and where tweet flu counts are increasing/decreasing significantly. This increase/decrease is gauged relative to the training data, i.e. relative to 2013's winter season. The 2014 flu season started very early in Australia so we started investigating the outbreak from the 6th of March 2014. Generally, we would not know this starting point when monitoring prospectively, but we did know that in 2009 the H1N1 outbreak started in early May 2009 and there were indicators that it started even earlier. The biplot for 2014 flu season was started on the 6th of March 2014 and is run to near the end of October 2014, however in future, starting on the 1st of March is recommended. The training data for this biplot is taken as the flu season of 2013 from 1 July to 30 September 2013. The ellipse constructed around the most recent points in the observation plot indicates a significant outbreak in the state and territory vector of flu tweet counts.

In Figure 3, the biplot for 17 March 2014 indicates a significant increase in flu tweet counts for Victoria (the blue thickening of the middle of the VIC vector in the variable plot flags this significant event) and a significant change in mean square forecast error in Western Australia (WA vector is colored red, largely driven by observation 7). Since the orientation of this ellipse is in the same side of the origin from vectors VIC and WA (and all other geographical regions) in the variable plot, this indicates that this significant increase in flu counts is more likely in Victoria (VIC) and Western Australia (WA). In addition, the empty boxes below the variable plot indicate non-significant changes in correlations.

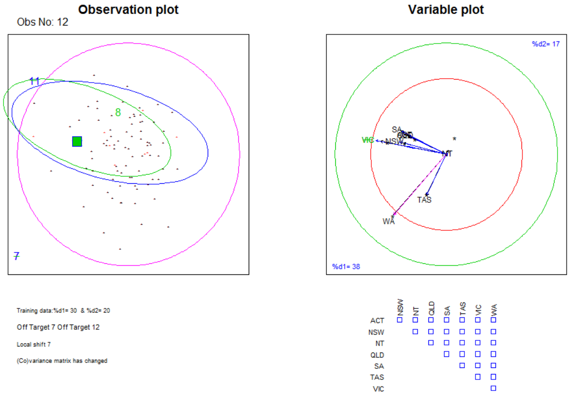

Figure 3: The biplot from 6 March to 17 March 2014 flagging the start of a flu outbreak in VIC and WA.

On the 26th of March 2014 (Obs. No. 21) significantly higher than expected tweet counts mentioning flu persisted for Victoria and Western Australia (Figure 4). The different colors and the thickening of vectors in the variable plot are used to indicate which moving average flags a significant increase in location. For WA it is the last 12 days while for VIC it is the last 20 days, however note that the ellipse for the past 12 days and past 20 days coincide. There is no evidence of a significantly larger outbreak of flu in 2014 relative to 2013 in the other states or territories.

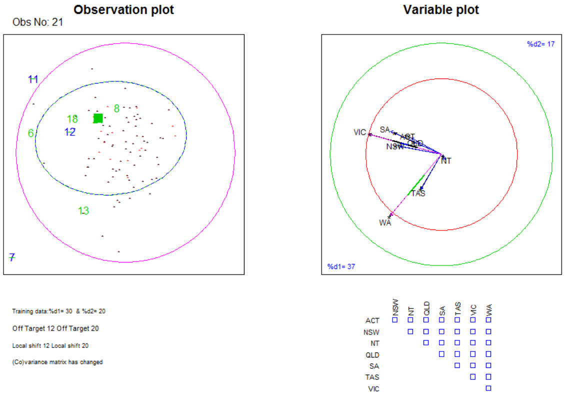

Figure 4: The biplot from 6 March to 26 March 2014 flagging the start of a flu outbreak in VIC and WA.

By the 9th of April 2014 (Figure 5), the significantly higher than expected tweet counts mentioning flu persisted for Western Australia and now is more orientated for WA. For Victoria however, this no longer flags a significant increase in flu counts, but does flag a significant change in mean square forecast error. Tasmania indicates a sudden increase in flu tweet counts but we are interested in whether this persists. Note that the coloring in the matrix below the variable plot indicates that the Northern Territory tweet flu counts has become less correlated with Queensland and South Australia's tweet flu counts (2 of its 3 neighbors).

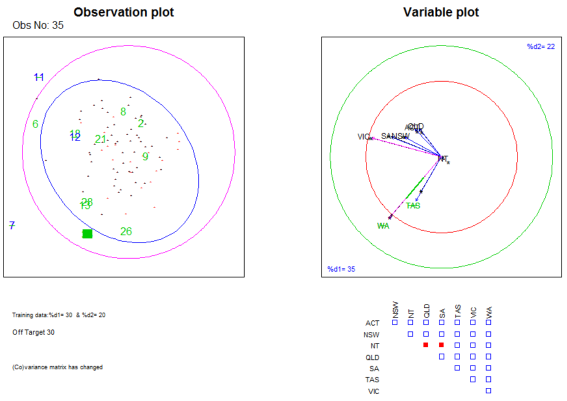

Figure 5: The biplot from 6 March to 9 April 2014 flagging the start of a flu outbreak in VIC and WA.

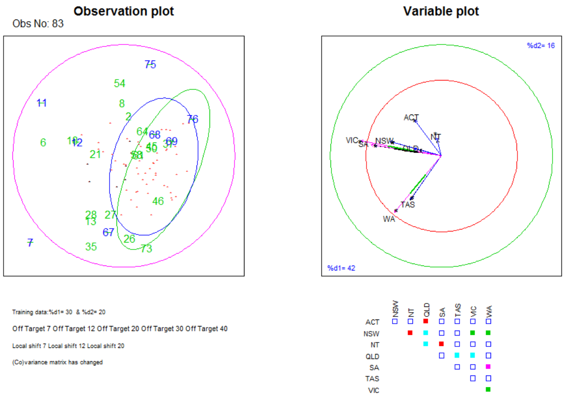

Figure 6: The biplot from 6 March to 27 May 2014.

By the 27th of May 2014 (Figure 6) there is evidence of significantly lower than expected tweet counts for all states and territories since the ellipses are located on the opposite side of the origin to all vectors in the variable plot. These lower than expected tweet flu counts are significant for VIC, WA and SA. QLD is perpendicular to the direction of vectors WA and VIC, indicating that this increase is unlikely to be occurring in Queensland. VIC and WA have become more correlated, albeit not significant at the 5% level.

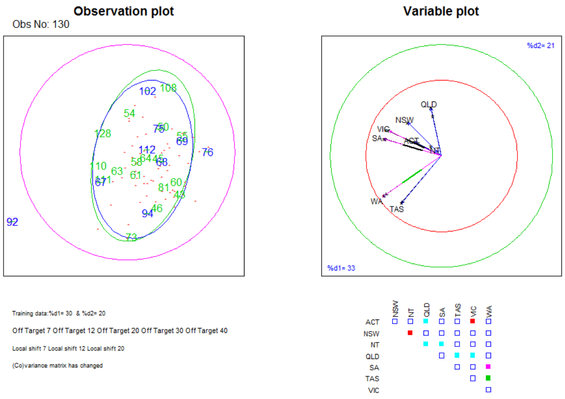

Figure 7: The biplot from 6 March to 13 July 2014 flagging.

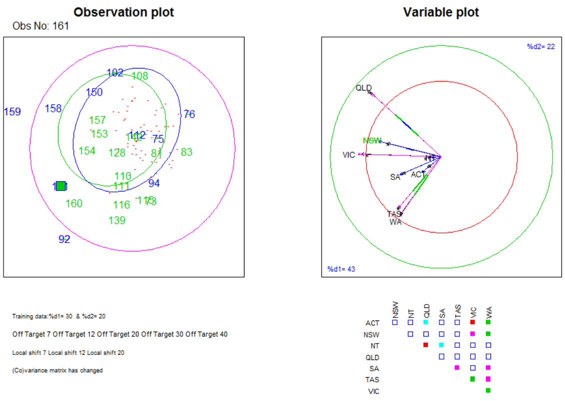

Figure 8: The biplot from 6 March to 13 August 2014 flagging.

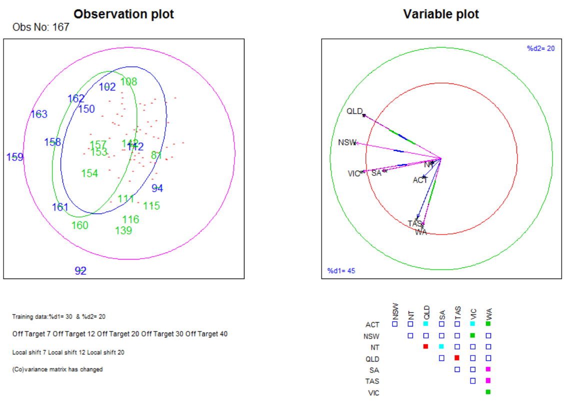

Figure 9: The biplot from 6 March to 19 August 2014.

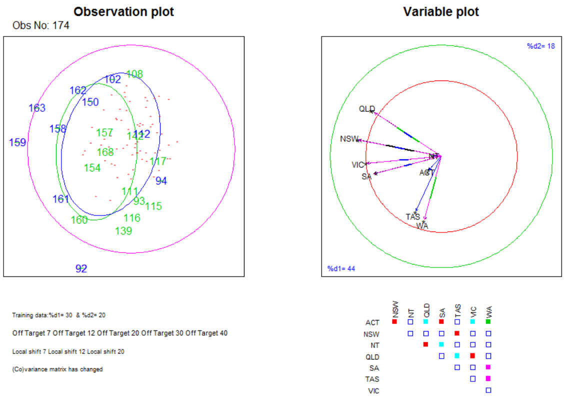

Figure 10: The biplot from 6 March to 26 August 2014.

By the 13th of July 2014 (Figure 7) there is still evidence of lower tweet flu counts when compared with the one-day ahead forecasts. This remains significantly lower than expected for VIC, WA and SA.

By the 13th of August 2014 (Figure 8) the outbreak seems to have started in QLD, TAS and WA with their tweet flu counts indicating an increase together with a potential increase in NSW, particularly test data observation numbers (obs. no.) 157 to 161. In addition, there is evidence that WA is experiencing a second wave of unexpected increased tweet flu counts, indicating that even after adjusting for higher than expected counts earlier in the season; these are now trending back into the unusual higher than expected region.

On the 19th of August 2014 (Figure 9) there is stronger evidence of increases in most states and territories, but this is significant for WA, VIC, NSW and QLD.

By the 26th of August 2014 (Figure 10) there is stronger evidence of a second wave of increased flu tweet counts. SA flags a significant increase in mean square forecast errors, but not a significant increase in counts. ACT flags a significant increase in correlation with WA and a significant reduction in its correlation with NSW and SA.

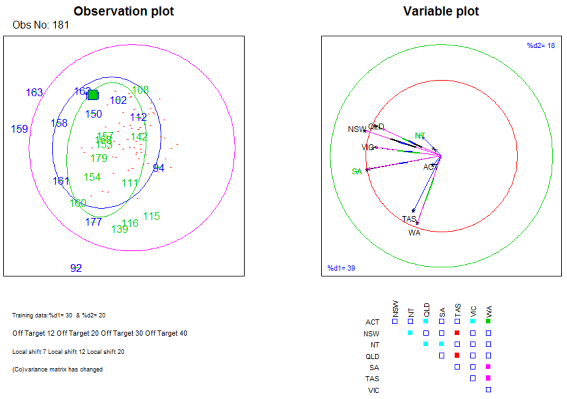

Figure 11: The biplot from 6 March to 2 September 2014.

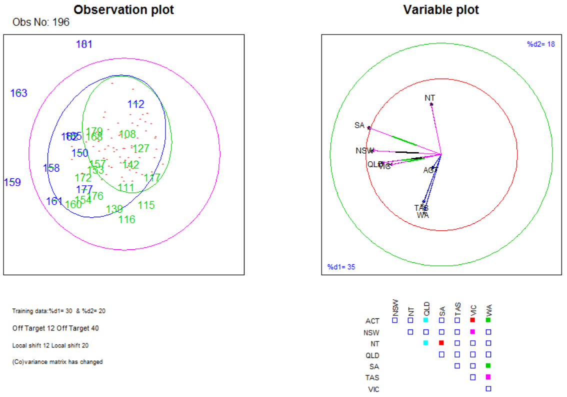

Figure 12: The biplot from 6 March to 17 September 2014.

By the 2nd of September 2014 (Figure 11), stronger evidence of increased flu tweet counts for QLD, VIC, NSW, SA and WA persists. SA now flags a significant increase in counts.

By the time we come to the 17th of September 2014 (Figure 12), there is no evidence of an unexpected increase in flu tweets for WA but there is for NSW, VIC, SA and QLD. This indicates that the unexpected increasing flu outbreak trend for WA is over but not for the other states.

By the 8th of October (Obs. No. 217) and 21st of October 2014 (Obs. No. 230) there is evidence of the flu season's impact reducing relative to what the model forecasts suggested was the expected trend. This indicates the assessment of the unusual outbreak as being over is justified. Although this may be less true of the NT, since its orientation is orthogonal to the direction of the centroid of the ellipse. There is small evidence that the outbreak in the NT may have just started (Figures 13-15).

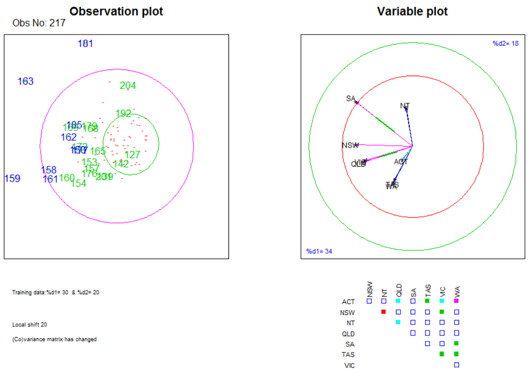

Figure 13: The biplot from 6 March to 8 October 2014.

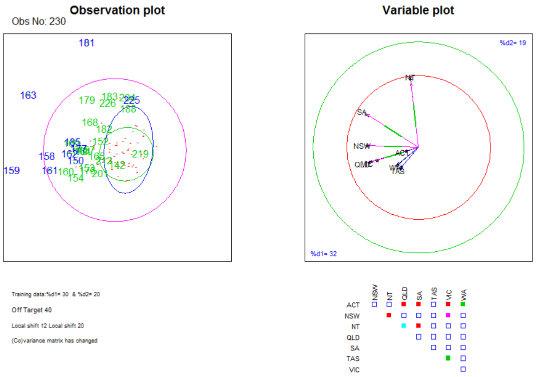

Figure 14: The biplot from 6 March to 8 October 2014.

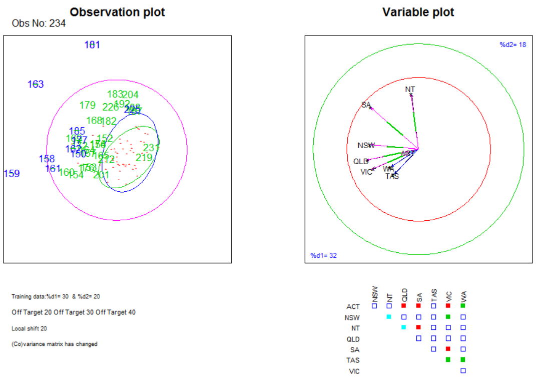

Figure 15: The biplot from 6 March to 26 October 2014.

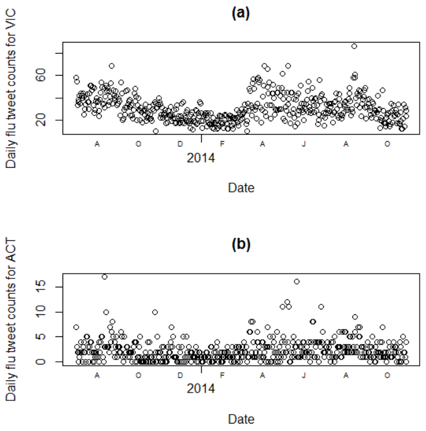

Figure 16: Times series of VIC and ACT counts of tweets mentioning flu.

Notice that the angle between the TAS and WA vectors for both the training data (2013's flu season) and the test data period (6 March to 21 October) in the variable plot was always small, indicating that Tasmania's flu tweet counts are more correlated to Western Australia's than the other states during both 2013 and 2014's flu seasons. This is interesting because these states are not near neighbours. Another interesting fact is that the flu season in 2014 lasted much longer than usual and there is evidence of a second wave of increasing flu counts in Victoria (Figure 16a). In addition, ACT never flagged an increase indicating that its 2014 flu season was not different from it 2013 flu season (Figure 16 b).

Discussion

This paper tests the use of the dynamic biplot for the early identification of possible flu outbreaks based on changes in the frequency of flu symptoms mentioned in Twitter data. This approach appears to be useful for detecting a flu outbreak in the test data (winter of 2014) based on training data from the previous year. The dynamic biplot was useful in describing the nature of the outbreak. The methodology identified very early that the 2014 flu season (March to September) was expected to be bigger than usual because it appeared very similar to trends observed for H1N1 in 2009. The large flu outbreak was confirmed with the state Health Departments of both NSW and Victoria using their reports from that period, but these were retrospective reports that could not be used to compare which data source offered the greatest benefit in terms of early detection. This should be the subject of future research. There appears to be richness and strength in the multivariate approach using the dynamic biplot. Software for the dynamic biplot is available from the lead author on request.

The approach utilized in this paper cannot be directly compared to historical data because there is no commonly agreed definition for an outbreak for the flu conditions considered in this research. Only a simulation exercise can facilitate such a comparison. However, these are our observations:

The dynamic biplot has the ability to identify flu outbreaks especially if multiple outbreaks occur simultaneously. The results demonstrate the efficacy of this methodology for use in the early detection of the flu outbreaks.

Speed of identification over emergency room reporting is facilitated because Twitter counts are available in almost real time (e.g., an hour lag time as opposed to 24 hour lag time for emergency room reporting however this paper only examines daily counts). In addition, many state health departments in Australia currently don't deliver data within 24 hours.

False alarm rates in this paper are difficult to manage. The dynamic biplot manages this by assuming normality but this is not always feasible. We believe that the methodology should be trained for each application separately. At very least the dynamic biplot offers a way of exploring the data under the assumption of normality.

The Poisson regression model used to account for seasonal variation, within day variation and day-of-week influences, offers a fairly simple model to define usual vs. expected behavior in the frequency counts. Often counts are over dispersed but this over dispersion appeared to be driven by a very small proportion of the counts that are potentially out-of-control situations. The assumption made here is that the few counts that contribute to the over-dispersion are the result of out-of-control situations and not usual behavior. Twitter also offers us the opportunity of detecting low grade flu symptoms that don't result in emergency department visits because they are not severe enough. Twitter also provides different information to emergency department presentations i.e. those that are sick enough to go and wait four or more hours (in Australia) to be treated in an emergency department. Twitter counts are likely to tell us much more about disease burden by our ability to follow the tweets of people who report certain symptoms. In addition, the younger generation that use Twitter are generally healthy enough to ride out symptoms on their own rather than present at emergency departments.

We note several limitations to the research reported here that require further study, including:

Training data that is not necessarily in-control.

Choice of keywords to use for data mining the tweets .

Selection bias is a core issue with Twitter surveillance plans.

Flu symptom keywords could appear in tweets due to their association with other health conditions that we are not attempting to detect so that our training data may still have inflated numbers for certain conditions.

Future research will be directed to link emergency department data with Twitter symptom counts (Aramaki et.al. [4]), Google flu trends (Ortiz et al. [5]), medication sales (Das et. al. [6]) and pathogen counts to integrate the value of each of these data sources in a single detection system. In addition, the timing, location, severity and disease burden also needs to be examined and assessed [7-10].

Appendix

A: Keyword Symptoms for flu

am getting the flu, feel flue, flu coming on, flu is coming on, getting flu, got flu, got grip, got grippe, got pestilence, got pneumonia, got the curse, got the grip, got the infestation, got the scourge, have flu*, have grip, have grippe, have pestilence, have pneumonia, have the affliction, have the curse, have the grip, have the infestation, have the pestilence, have the scourge, influenza.

*represents a wild card in the keyword or phrase in an attempt to capture variations of the keyword or phrase that might appear in a tweet

B: Biplot interpretations

The information that is available using the dynamic biplot is as follows:

Outlying condition count vectors are flagged by a coloured box around the numbered observation. This is found using a Hotelling's T2 statistic test to measure how far the observed condition count vector is from the sample mean condition count vector and test its level of significance.

Testing whether the sample mean condition count vector for the last 7, 12, 20, 30 and 40 observations are significantly different from the sample mean vector of the training sample is tested using the Hotelling's T2 test statistic at the 1% level of significance. An ellipse is constructed around the group of observation that contributes to this significance test (trimming out any outliers). A maximum of two ellipses are constructed and these are generally the two largest groups of significant observations. The results of these tests are recorded under the observation plot.

Under the observation plot we record the results of multivariate tests of hypotheses at the 1% level. The following tests are considered:

A location shift in the sample mean vector of n=7, 12, 20, 30 and 40 observations from the sample mean vector for the remaining observations in the the plot.

A change in the sample variance covariance matrix from that of the training sample covariance matrix.

Once a multivariate test is found to be significant then, univariate tests are carried out. These include the following:

Testing whether the mean square error (MSE) of a condition count has increased significantly. If so, then the corresponding vector in the variable plot changes colour, magenta for a significant increase in MSE and turquoise for a significant decrease in MSE.

A significant local shift in mean for individual conditions, under the assumption that there is no change in variance, is flagged by a coloured sausages being inserted on the respective vectors in the variable plot. It is used to confirm which univariate variables contributed significantly to the change in the sample mean vector from the expected vector derived from the training data.

The contribution of the most recent observation to the MSE is measured from the distance between the arrow and "*" on the respective vector in the variable plot. The larger this distance the more this variable has either shifted from target or the more variable it has become. If the most recent observations are on "one-side" of the origin then this indicates a location shift. If the observations are oscillating from one side to the next in a random way more than usual then this indicates an increase in variation in symptom counts. The contribution of the second to last observation to the MSE is given by the distance between the "*" and the "." On the variable plot – the greater this distance the larger is its contribution to the increase in MSE. If the most recent observation has significantly increased its MSE then the variable label of its respective vector changes colour and is labelled green.

If the change is an unusual increase in counts then the variable text is strikethrough using colour green.

If the change is an unusual decrease in counts then the variable text is strikethrough using colour red.

The matrix of boxes under the variable plot is included to flag the significant changes in pairwise correlations. Any change in colours of these squares linking two variables indicates either a near significant change in correlation or a significant change in correlation:

Squares are coloured red when the correlation increases significantly and dark blue when the correlations decrease significantly.

Squares are coloured magenta when the correlation increase is nearly significant (5% level of significance) and light blue when the correlation decrease is nearly significant.

The angle between the vectors in the variable plot is a rough estimate of the correlation between two variables when there is no change in location.

The direction of the location shift can be estimated from the position of the recent points in the observation plot with respect to the orientation of the vectors in the observation plot. The following interpretations are available provided the first two principal components represent a significant amount of the total variation in the full dimensional space:

Observation points that are located in a cluster away from the origin in the same direction as a vector in the variable plot flags a location change that has increased its condition counts significantly from expected.

Observation points that are located in a cluster away from the origin in the opposite direction as a vector in the variable plot flags a location change that has decreased its condition counts from expected.

Observation points that are located in a cluster away from the origin in a direction which is at right angles to the direction of a vector in the variable plot indicates this variable has not shifted location at all – all the variation has been random around the expected count.

The percentage variation explained by the first two principal components computed from the training data is always recorded below the observation plot. The percentage variation explained by the first two principal components computed from the current data is always recorded in the variable plot in the bottom left-hand corner for the major axis and in the right-hand corner for the 2nd major axis. Whenever this has changed significantly the text recording this variation changes colour from blue text to:

Turquoise when the change is a significant increase in variation explained by the two dimensions (1% level).

Magenta when the change is nearly significant at the 1% level (significant at the 5% level).

Acknowledgement

This work has been supported by the DSTO National Security Science and Technology Centre.

References

1. Power R, Robinson B, Colton J, Cameron M (2014) Emergency Situation Awareness: Twitter Case Studies. Information Systems for Crisis Response and Management in Mediterranean Countries - First International Conference, ISCRAM-med 2014, Toulouse, France, 218-231.
2. Shewhart, Walter A (1931) Economic Control of Quality of Manufactured Products. D Van Nostrand Company, New York.
3. Sparks R, Adolphson A, Phatak A (1997) Multivariate Process Monitoring Using the Dynamic Biplot. International Statistical Review, 65: 325-349.
4. Aramaki E, Maskawa S, Morita M (2011) Twitter catches the flu: detecting influenza epidemics using Twitter, In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics 1568-1576.
5. Ortiz JR, Zhou H, Shay DK, Neuzil KM, Fowlkes AL, et al. (2011) Monitoring influenza activity in the United States: a comparison of traditional surveillance systems with Google Flu Trends. PLoS One 6: e18687.
6. Das D, Metzger K, Heffernan R, Balter S, Weiss D, et al. (2005) Monitoring over-the-counter medication sales for early detection of disease outbreaks--New York City. MMWR Morb Mortal Wkly Rep 54 Suppl: 41-46.
7. Okugami C, SPARKS R, Woolford S (2014) Twitter Data Offers Opportunities for Public Health Professionals. J Health Med Informat 5: 1-3.
8. Sparks RS (2000) CUSUM charts for signalling varying locations shifts. J Qual Technol 32: 157-171.
9. Sparks RS, Keighley T, Muscatello D (2009) Improving EWMA Plans for Detecting Unusual Increases in Poisson Counts. Journal of Applied Mathematics and Decision Sciences 2009: 16.
10. Sparks R, Carter C, Graham P, Muscatello D, Churches T, et al. (2010) Understanding sources of variation in syndromic surveillance for early warning of natural or intentional disease outbreaks. IIE Transactions 42: 613-631.