Paper 197-31

BIOSURVEILLANCE AND OUTBREAK DETECTION USING THE ARIMA AND LOGISTIC PROCEDURES

Ernest S. Shtatland, Ken Kleinman, Emily M. Cain
Harvard Medical School, Harvard Pilgrim Health Care, Boston, MA

ABSTRACT

The main objective of this paper is to show potential usefulness of the combination of autoregressive integrated moving average (ARIMA) models and logistic regression with automatic model selection (see our work presented at SUGI'28 and SUGI'29.) Time-series analysis with ARIMA provides only one perspective of the information in the surveillance data (i.e. the number of patients as a function of time). The information about the geographical location of the patients provides a second perspective. We would like to combine both perspectives.

INTRODUCTION

A number of different approaches have been developed for syndromic surveillance, with systems monitoring over-the-counter drug sales (Goldenberg et al. (2002)), web-based physician reports (Shannon et al. (2002)), consumer health hotline telephone calls, ambulatory care visit records (Lazarus et al. (2002), and Kleinman et al. (2004)) and visit data from emergency department (ED) information systems (Reis, Pagano, and Mandl (2003), Reis and Mandl (2003)). It is shown in Reis and Mandl (2003) and Reis, Pagano, and Mandl (2003) that ARIMA and its simpler cousins, ARMA, AR, or MA are some of the fundamental methods for time series forecasting. This work was first developed by Box and Jenkins (1976). Below we will use terms ARIMA and ARMA as interchangeable, although ARIMA is a more general class of models capable of describing non-stationary behaviors. According to Kulldorff et al. (2005) with reference to Reis and Mandl (2003), most analytical methods in use for the early detection of disease outbreaks are purely temporal in nature. For ED data, the best-fitting ARIMA-type model is applied to the residuals after fitting the trend and seasonality. Due to the timeliness and data availability considerations, the majority of systems to date have focused on monitoring visit data from ED. In Reis, Pagano and Mandl (2003) and Reis and Mandl (2003), ED data are typically modeled by using low-order ARIMA models such as ARMA(1,2), ARMA(2,1), and ARMA(1,1). In Earnest et al. (2005) the following models are used: AR(1), ARMA(1,1), ARMA(1,2), ARMA(1,3), ARMA(1,4), ARMA(0,1) (MA(1)), ARMA(0,2) (MA(2)) and ARMA(0,3) (MA(3)). It is well known (Box and Jenkins (1976)) that there is duality between autoregressive and moving average processes that can be formulated as follows:

1) The conditions for stationarity of an AR process mirror the conditions for invertibility of an MA process;

2) An AR process can be expressed as an infinite MA process and vice versa;

3) An AR process of order p has a PACF (partial autocorrelation function), which cuts off at lag p and a MA process of order q has an ACF (autocorrelation function), which cuts off at lag q.

We prefer to use either purely autoregressive models or a combination of AR and MA components (which are not purely MA processes). As a result we will use the most fundamental low-order models such as AR(1), ARMA(1,1) and AR(2). They are the simplest and most frequently used ARMA models. These models can be found in the following sources: Lai (2005), where the author applies the models AR(1) and ARMA(1,1) in monitoring the SARS epidemic in mainland China; Earnest et al. (2005) where the authors use AR(1) and ARMA(1,1) and also ARMA(1,3) to monitor a SARS outbreak in Singapore; Reis and Mandl (2003), where ARMA(1,1) is found to work best for respiratory-related ED volume), and Rizzo et al. (2005).

ARMA MODELING IN DISEASE SURVEILLANCE: ADVANTAGES AND DISADVANTAGES

Thus, as can be seen from the Introduction, typical models for temporal data are ARIMA. According to Reis and Mandl (2003), one of the primary benefits of ARIMA models is their ability to correct for local trends in the data – what has happened on the previous day is incorporated into the forecast of what will happen today. This works well, for example, during a particularly severe flu season, where prolonged periods of high visit rates are adjusted to by the ARIMA model, thus preventing the alarm from being triggered every day throughout the flu season.

However, if the ARIMA model "adjusts" to an actual outbreak instead of detecting it, a slowly spreading outbreak or attack might be missed because of this correction. This correction is most likely to affect detection of outbreaks occurring over several days, rather than those that occur suddenly. It is therefore also important to rely on the non-ARIMA or non-classical ARIMA models for outbreak detection.

ARMA models in Reis and Mandl (2003) and Reis, Pagano, and Mandl (2003) require large historic records of patient visits in order to begin surveillance. This is a substantial disadvantage. As can be seen from Moore et al. (2002), in some cases long historical data are not available and not necessary. Also combining both historical and recent trends is quite realistic. Another disadvantage of ARMA is that the corresponding detector is not sensitive to the slow growth.

According to Rizzo et al. (2005), outbreaks that evolve over a matter of days, for example, can often be detected with ARMA models that generate single-day predictions based on historical data. More gradually developing outbreaks are generally easier to detect by using such techniques as CuSum (Hawkins and Olwell (1998)).

In this paper, we propose an approach to ARIMA modeling that allows potentially to detect changes in variance and even in correlation structure of the process. The correlation structure is determined by the ARIMA coefficients.

EXAMPLES OF ARMA MODELING IN DISEASE SURVEILLANCE

There is a consensus that ARMA models seem better suited to describe historical visit rates and account for temporal dependency. It is shown in Reis and Mandl (2003) that of all the ARMA models tested, ARMA(2,1) works best for overall ED volume with the error of 9.37%. For respiratory-related ED visits the best model is ARMA(1,1) with the error of 27.54%. In Earnest et al. (2005) it is shown that using ARMA models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore, ARMA(1,3) has been found to be optimal. Its mean absolute percentage error (MAPE) for the training and validation sets is 5.7% and 8.6% respectively. At the same time, simpler models ARMA(1,0) and ARMA (1,1) demonstrate performance very close to optimal (see MAPE for the training and validation sets):

ARMA(1,0)	6.0%	9.0%
ARMA(1,1)	6.2%	9.2% .

Guided by the parsimony principle we can prefer either ARMA(1,0) or ARMA(1,1) to the optimal but *more complex model* ARMA(1,3). Also in Lai (2005) the author discusses using models AR(1) and ARMA(1,1) in monitoring the SARS epidemic in Mainland China.

AR(1)

$y_t = \phi_1 y_{t-1} + w_t$; *(AR(1) equation)*
where w_t is a white noise;
$-1 < \phi_1 < 1$; *(interval of stationarity)*
$\sigma_y^2 = \sigma_w^2 / (1 - \phi_1^2)$; *(Variance)*
$R^2 = 1 - \sigma_w^2 / \sigma_y^2$; *(R-Square for AR(1))*

If the coefficient ϕ_1 converges to the ends of the interval $-1 < \phi_1 < 1$, then R^2 converges to 1.

AR(2)

$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + w_t$; *(AR(2) equation)*
where w_t is a white noise;
$\phi_2 + \phi_1 < 1$; *(triangular of stationarity)*
$\phi_2 - \phi_1 < 1$;
$-1 < \phi_2 < 1$;

$\sigma_y^2 = \sigma_w^2 (1 - \phi_2 / (1 + \phi_2) (1 / (1 - \phi_1 - \phi_2)(1 - \phi_2 + \phi_1)$; *(Variance)*

$R^2 = 1 - \sigma_w^2 / \sigma_y^2 =$ *(R-Square for AR(2))*
$1 - (1+ \phi_2 / 1- \phi_2)(1 - \phi_1 - \phi_2) (1-\phi_2 + \phi_1);$
If the coefficients (ϕ_1, ϕ_2) converge to the boundary of the triangular of stationarity, then R^2 converges to 1.

ARMA(1,1)

$y_t - \phi_1 y_{t-1} = w_t - \theta_1 w_{t-1};$ *(ARMA(1,1) equation)*
where w_t is a white noise;
$-1 < \phi_1 < 1;$
$-1 < \theta_1 < 1;$

$\sigma_y^2 = \sigma_w^2 (1+ \theta_1^2 -2 \phi_1\theta_1 / (1- \theta_1^2);$ *(Variance)*

$R^2 = 1 - \sigma_w^2 / \sigma_y^2 =$ *(R-Square for ARMA(1,1))*
$1 - (1 -\phi_1^2) / (1+\theta_1^2 - 2\phi_1\theta_1)$
If the coefficients (ϕ_1,θ_1) converge to the boundary of the rectangular of stationarity and invertibility, then R^2 converges to 1.

R-SQUARE FOR AR(1), AR(2) AND ARMA(1,1)

To the best of our knowledge, formulas for R-Square for AR(1), AR(2) and ARMA(1,1) appeared originally in SAS / ETS® User's Guide (1999). However, characteristics similar to R-Square were described in Shtatland and Sandri (1990) and Shtatland (1993) (the entropy as a measure of self-organization of a process). Also, formulas for the variances on which R-Square are based were described in Box and Jenkins (1976), pp. 58 – 77.

Note that in Reis, Pagano and Mandl (2003) it is proposed to use the variance of residuals as the measure of goodness-of fit. We prefer to use R-Square and Adjusted R-Square (see below) as more superior than the variance of residuals.

Note also that ARMA models, for example ARMA(1,1), with roots close to the unit circle are capable of representing long-range dependence over a wide range. However, generally speaking ARMA are short-memory models. Also, all three models of our choice: AR(1), AR(2) and ARMA(1,1), can model approximately periodical phenomena.

SELECTING A MODEL

Our approach is to use the data for 30 – 60 days for model development, see for comparison Earnest et al. (2005). We have to choose between AR(1), AR(2) or ARMA(1,1) based either on the *adjusted* R-Square

$1 - (n - 1) / (n - k) (1 - R^2)$ (n is the sample size and k is the number of parameters),

or Akaike Information Criterion (AIC), or Schwarz Bayesian Criterion. Note that all criteria mentioned above: R-Square, Adjusted R-Square, AIC, and some others can be used as measures of *practical significance* as opposed to *statistical significance*.

From one 30 – 60 days period to another, we can observe the behavior of the R-Square, adjusted R-Square, AIC, and also parameter points: ϕ_1 for AR(1), (ϕ_1 , ϕ_2) for AR(2) and (ϕ_1,θ_1) for ARMA(1,1). Our approach can most likely detect gradual changes in the variance and some other characteristics that can been obtained within CUSUM (Rizzo et al. (2005)). Clustering in the parameter space would allow us to define conditions on dynamic patterns in the time domain. Such clustering would be impossible within the classic ARMA approach.

COMBINING ARMA TECHNIQUES AND REGRESSION METHODS: LOGISTIC, POISSON REGRESSION, GLMM, ETC.

We suggest the following workflow for analysis of surveillance data:

1) In analyzing spatio-temporal data, work with the temporal series *first*, and then with the spatial series (see for example, Ozonoff et al. (2003));
2) Build an ARMA model for each individual ZIP code based on hospital ED data instead of one gross time series (see Kleinman et al. (2004b) and Platt et al. (2003));
3) Select *just a few* ARMA models with the *highest* level of adjusted R-Square or AIC. Nowadays, the following approach is much more common: for each ZIP code to perform a separate statistical significance test, obtaining its *p*-value and finally report all ZIP codes that are significant at some level α. Here we have a multiple hypothesis testing problem with potentially large numbers of false alarms. To avoid this problem, we propose to work with R-Square, Adjusted R-Square or AIC as measures of *practical significance* as oppose to *statistical significance*);
4) Take ambulatory care visit data with selected ZIP codes that correspond to the chosen ARMA models;
5) In doing so we reduce the extremely large size of spatio-temporal surveillance data sets that typically preclude more complex models;
6) In Kleinman et al. (2004a), the authors recommend that individual-level covariates be used whenever possible, because of their predictive value. Practically, however, their inclusion can make fitting models computationally difficult. And in Kleinman (2005), it is assumed that no covariates, such as age and gender, will be used.
7) After reducing the data described above, we arrive at a smaller surveillance data set and we can possibly use individual covariates with automatic method selection. Thus we have a two-stage monitoring system that can combine the strengths of ARIMA and LOGISTIC / POISSON /GLMM methods bases on ambulatory care visits.

CONCLUSIONS:

In syndromic surveillance systems, all available information, temporal and spatial, should be used. Temporal information analysis, followed by spatial analysis should be performed. For temporal analysis we propose to use the most parsimonious and at the same time the most fundamental ARMA models: AR(1), AR(2) and ARMA(1,1) for each individual ZIP code. We propose to calculate R-Square , Adjusted R-Square or AIC and select a number of models (ZIP codes) with the highest level of mentioned above measures of practical significance as opposed to the approach based on statistical significance. Following this, we propose to proceed to the spatial step with logistic, Poisson regressions or GLMM with the option of using their own built-in R-Square measures (Brownstein, Kleinman, and Mandl (2005).

REFERENCES:

Box , G. E. P. and Jenkins G. M. (1976). Time series analysis: forecasting and control. San Francisco, CA: Holden Day.

Brownstein, J. S., Kleinman K. P., and Mandl, K. D. (2005). Identifying pediatric age group for influenza vaccination using a real-time regional surveillance system. American Journal of Epidemiology, v. 162, pp. 686 – 693.

Earnest, A., Chen, M. I., Ng, D., and Sin, L. Y. (2005). Using ARIMA models to predict and monitor the numbers of beds occupied during a SARS outbreak in tertiary hospital in Singapore. BMC Health Services Research, 5:36.

Goldenberg, A., Shmuel, G., Caruana, R. A., and Fienberg, S. E. (2002). Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. Proc. Natl. Acad. Sci. USA, 99(8), 5237-5240.

Hawkins, M. D. and Olwell, D. H. (1998). Cumulative Sum Charts and Charting for Quality Improvement. Berlin / Heidelberg: SpringerVerlag.

Kleinman, K., Lazarus, R., and Platt, R. (2004a). A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with application to biological terrorism. American Journal of Epidemiology, v.159, number 3, pp. 217 – 224

Kleinman, K., Lazarus, R., and Platt, R. (2004b). Respond to “Surveilling Surveillance”. American Journal of Epidemiology, v.159, number 3, p. 228.

Kleinman, K. (2005). Generalized linear models and generalized linear mixed models for small-area surveillance. In Spatial and Syndromic Surveillance for Public Health. Edited by A. B. Lawson and K. Kleinman, John Wiley & Sons, Ch. 5.

Kulldorff, M., Heffernan, R, Hartman, J., Assunçao, R. and Mostashari, F. (2005). A space time permutation scan statistic for disease outbreak detection. PloS Medicine, v.**2**, # 3.

Lazarus, R., Kleinman, K., Dashevsky, I., Adams, C., Kludt, P., DeMaria, A. Jr., and Platt, R. (2002). Use of automated ambulatory-care encounter records for detection of acute illness clusters, including potential bioterrorism events. Emerg. Infect. Dis, 8(8), 753-760.

Lai, D.(2005). Monitoring the SARS epidemic in China: time series analysis. Journal of Data Science, 3, pp. 279 – 293.

Ozonoff , A., Forsberg, L., Bonetti, M., and Pagano., M. (2004). Bivariate method for spatio-temporal syndromic surveillance. MMWR, vol.53, Supplement, pp. 61 – 66.

Platt, R., Bocchino, C., Caldwell, et al. (2003). Syndromic surveillance using minimum transfer of identifiable data: the example of the National Bioterrorism Syndromic Surveillance Demonstration Program. Journal of Urban Health: Bulletin of the New York Academy of Medicine Vol. 80, No. 2, Supplement 1.

Reis, B. Y., Pagano, M., and Mandl, K. D. (2003). Using temporal context to improve biosurveillance. Proceedings National Academy of Sciences, USA, 100(4), 1961-1965.

Rizzo, S. L., Grigoryan, V. V., Wallstrom, G. L., Hogan, W. R., Wagner M. M. (2005). The use of case studies for the evaluation of surveillance systems. RODS Technical Report Center for Biomedical Informatics, University of Pittsburgh.

Shannon, M., Burstein J., Mandl K.,and Fleisher, G. (2002). Usage of a web-based decision support tool for bioterrorism detection. Amer. J. Emerg. Med., **20**(4), 384-385.

Shtatland, E. S. and Sandri, G. v. H. (1990).Statistical analysis of chaotic systems. Bulletin of the American Physical Society, v. 35, No 10, pp. 2228-2229.

Shtatland, E. S. (1993). ARMA modeling in geophysics. Earth, Atmospheric, and Planetary Sciences, Earth Resources Laboratory report.

Shtatland, E. S., Barton, M. B., and Cain. E. M. (2001). The perils of stepwise logistic regression and how to escape them using information criteria and the Output Delivery System, SUGI '26 Proceeding, Paper 222-26, Cary, NC: SAS Institute, Inc.

Shtatland, E. S., Kleinman, K., and Cain, E. M. (2003). Stepwise methods in using SAS® PROC LOGISTIC and SAS® ENTERPRISE MINER for prediction. SUGI '28 Proceeding, Paper 258-28, Cary, NC: SAS Institute, Inc.

Shtatland, E. S., Kleinman, K., and Cain, E. M. (2004). A new strategy of model building in PROC LOGISTIC with automatic variable selection, validation, shrinkage and model averaging. SUGI '29 Proceeding, Paper 191-29, Cary, NC: SAS Institute, Inc.

CONTACT INFORMATION:

Ernest S. Shtatland
Department of Ambulatory Care and Prevention
Harvard Pilgrim Health Care & Harvard Medical School
133 Brookline Avenue, 6th floor
Boston, MA 02115
tel: (617) 509-9936
email: ernest_shtatland@hphc.org