Paper 363-2008

# Another Look at Low-Order Autoregressive Models in Early Detection of Epidemic Outbreaks and Explosive Behaviors in Economic and Financial Time Series

Ernest S. Shtatland, eStatConsulting, Stoneham, MA
Timur Shtatland, eStatConsulting, Stoneham, MA

## ABSTRACT

In our SUGI 2006 presentation, we suggested using low-order autoregressive models, AR(1) and AR(2), in biosurveillance and outbreak detection (PROC ARIMA, SAS/ETS®). Our suggestion was based on empirical data. In the NESUG 2007 paper, we proposed strong theoretical grounds for this. Here we provide further development of our approach. Based on a classic susceptible-infectious-recovered (SIR) model, we arrive at AR(1) models of epidemics where we need to estimate only one parameter, the first-order autoregressive coefficient. Its least squares estimate has a very simple epidemiological meaning. In the vast majority of applications, AR and ARMA are used as purely empirical, stationary models, with no specific substance matter meaning for coefficients. The value of our first-order autoregressive coefficient less than one corresponds to a stationary, no-epidemic regime. If the parameter is greater than one, we have an explosive case (an outbreak of epidemic). When the coefficient is equal to one, we have a unit root case. Based on the observed data in a chosen time window, least squares estimates and confidence intervals allow us to decide which case is more appropriate. The question of bias correction of our estimates is also discussed. After purely temporal analysis, we can proceed to the spatial step with logistic or Poisson regressions as in our SUGI 2006 paper. The approach described above can also be used in describing explosive behaviors of economic and financial time series (e.g., stock market bubbles).

The intended audience: SAS users of all levels who work with SAS/STAT® and SAS/ETS®.

## INTRODUCTION

Early recognition of either bioterrorist attacks or emerging epidemics is of greatest importance nowadays. With such highly infectious and rapidly spreading diseases as SARS, influenza, measles or smallpox it is a challenge to build adequate models for early detection of these exponentially developing processes. According to Mohtashemi, Kleinman and Yih (2007), temporal anomaly detection is a key component of real-time syndromic surveillance. Although space-time detection methods are sometimes more powerful at anomaly detection than purely temporal ones, spatial data are typically scarce. In addition, many health-care institutions routinely collect data in the form of *time series*. That is why effective purely temporal surveillance techniques are still needed. Autoregressive moving average (ARMA) processes are among the most popular and frequently used temporal models. The literature on using ARMA models in biosurveillance is abundant. Here we mention only the most important for us sources: Reis and Mandl (2003), Reis, Pagano and Mandl (2003), Earnest et al. (2005), and Lai (2005). Also see Shtatland, Kleinman and Cain (2006) and Shtatland (2007) for numerous additional references. In spite of their popularity, AR and ARMA models in biosurveillance share a common disadvantage: they are used as purely empirical models, with no specific substance matter meaning for coefficients. Our suggestions in Shtatland, Kleinman and Cain (2006) on using low-order ARMA processes are based solely on the parsimony principle. In Shtatland (2007**)**, we proposed strong theoretical grounds for this. In this presentation, we provide further development of our approach. In doing so, we are guided by the pioneering article by Mohtashemi et al. (2006) in which the workhorse model in epidemiology, a classic susceptible-infectious-recovered (SIR) model was successfully used for early detection of respiratory infection outbreaks. In the course of linearization of *nonlinear* SIR difference equations, which is justified by the requirements of early detection of the initial phase of the epidemic, we arrive at a *linear* AR(1) model of epidemics that inherits first principles of SIR. Here, we need to estimate only one parameter, the first-order autoregressive coefficient. This parameter and its least squares estimate have a very simple epidemiological meaning. It is interesting that in our approach detection thresholds are defined naturally in terms of this coefficient. In all papers cited above, thresholds are defined more or less arbitrarily. Our approach can also be used in describing explosive behaviors of economic and financial time series (e.g., stock market bubbles and higher frequency phenomena such as micro-bubbles). See for example Phillips, Wu and Yu (2007), Shive (2006), Shiller (2005), Abreu and Brunnermeier (2003), and Brunnermeier and Pedersen (2002).

## SIR MODELS IN EARLY DETECTION

A SIR model can be described by the following first-order nonlinear system of difference equations (we use the notations of Mohtashemi et al. (2006)):

$$\begin{aligned} S_{n+1} &= S_n - \beta S_n I_n, \\ I_{n+1} &= I_n + \beta S_n I_n - \delta I_n, \qquad (1) \\ R_{n+1} &= R_n + \delta I_n \end{aligned}$$

where $S_n$, $I_n$ and $R_n$, represent the respective numbers of susceptible, infected and recovered individuals correspondingly on day $n$; $\beta$ is the infection transmission rate and $\delta$ is the average rate of recovery from infection; $1/\delta$ can be considered the mean duration of the infectivity (in days). Below, we will use another transmission parameter, alternative to $\beta$. Note that Mohtashemi et al. (2006) use only the first two equations in (1) with the motivation that there is not enough time for temporary removal of the recovered population to be of significance to the dynamics. Note also that in the context of their article (the emergency department of a large, academic pediatric hospital), variables $S_n$, $R_n$ cannot be observed or measured systematically. Only $I_n$ can be obtained indirectly, through some calculations, by using the observable daily number of patients $v_j$ presenting to the emergency department on day $j$. Mohtashemi et al. (2006) propose the following approximate formula for $I_n$:

$$I_n = \Sigma^{n}_{j=n-d+1}\, v_j \qquad (2)$$

where $d$ is an average number of days of infectivity per patient, i.e. $d = 1/\delta$. Thus Mohtashemi et al. (2006) assume that the overall number of infected on day $n$ can be approximated by the sum of the number of visits to the emergency department during the past $d$ days. This is a reasonable approximation if parameter $d$ is adequately chosen or estimated. Also Mohtashemi et al. (2006) suppose that the mean duration of infectivity $d = 1/\delta = 7$, which is a clinically realistic assumption for influenza. And even though the assumption $d = 7$ is rather approximate (according to Wearing, Rohani and Keeling (2005), influenza lasts for 3 to 5 days), it is very convenient because it allows to compensate for the day of the week effect in $v_j$ variability. According to Mohtashemi et al. (2006), the goodness of such approximation can potentially impact the accuracy of early detection results and is open to discussion. Summarizing, we can see an urgent need in a closed equation for $I_n$ alone. Fortunately, such an equation can be obtained from the first two equations in (1) by eliminating unobservable variable $S_n$ (see equation (2) in Mohtashemi et al. (2006)):

$$I_{n+2} = I^2_{n+1} / I_n - \beta I_{n+1} (I_{n+1} - (1 - \delta) I_n) \qquad (3)$$

Equation (3) is a strongly nonlinear second-order difference equation, which is not easy to interpret. In (3) we have only one parameter to estimate, $\beta$. Mohtashemi et al. (2006) use this equation for estimating $\beta$ through least-squares regression for each time window of a chosen length. The authors proceed from the premise that the transmission rate $\beta$ is seldom unchanging and may be considered approximately constant only over short time periods. For each day in a year, they choose a time window of length $T$ (from the present day to the past), form $T$ - 2 equations and estimate $\beta$ by using least-squares regression. In their paper, the results for $T$ = 7 are reported. Again, this choice of $T$ seems reasonable since it compensates for the day of the week variation in data. In addition, it is made in the spirit of early detection requirements. At the same time the choice $T$ = 7 results in a very high everyday variability of the $\beta$ estimates, comparable to the variability of the time series $I_n$ itself. It looks unusual since $\beta$ is the only fundamental parameter of our model. We can conjecture that the time window length of 14 would provide smoother estimates of $\beta$ and in some cases serve the needs of early detection of outbreaks at the same level as $T$ = 7. Though, sometimes it is necessary to work with $T$ < 7 also. Unfortunately, there is no explicit formula for the least-squares estimates of $\beta$ and it is not easy to investigate their properties, in particular sensitivity to the time-window length $L$. This disadvantage will be overcome within our autoregression modeling approach. After getting the estimate of $\beta$ for each current day, Mohtashemi et al. (2006) compare it with the mean infection rate for the same day of year in their historical data (they have 7 training years). If the current estimate exceeds the threshold based on the training data (in terms of the mean and standard deviation of $\beta$), their detection algorithm generates an alarm. Mohtashemi et al. (2006) test the performance of their detection system by using simulated outbreaks.

## AUTOREGRESSIVE MODELS IN EARLY DETECTION

In trying to build autoregressive models for early detection of epidemic outbreaks, we use SIR model (1) and equation (3). The 2nd order of equation (3) suggests that we can limit ourselves to AR(2) models. Thus, regrouping (3) we arrive at the equation of the form

$$I_{n+2} = a_1 I_{n+1} + a_2 I_n \qquad (4)$$

where

$$a_1 = I_{n+1} / I_n - \beta(I_{n+1} - I_n) \qquad (5)$$

$$a_2 = - \beta \delta I_{n+1} \qquad (6)$$

Equation (4) is a *nonlinear* difference equation with time- and state-dependent coefficients. This class of equations has been described, for example, in Kato and Ozaki (2002). Also, (4) can be considered an equation with *slowly varying* coefficients. Indeed, in the context of Mohtashemi et al. (2006) we can think that $\beta \approx 2/10000$, $\delta \approx 1/7$, $I_{n+1} \approx 350$, and $(I_{n+1} - I_n) \approx 20$ or less at the early phase of epidemics. Thus, $a_2 \approx 0.01$, this coefficient can be neglected and our 2nd order difference equation is reduced to the 1st order equation. Further, it can be shown that term $\beta(I_{n+1} - I_n)$ in (5) is usually less than 0.004 and can also be neglected. Thus, $I_{n+1} / I_n$ can be considered the main part of parameter $a_1$, and consequently this ratio $I_{n+1} / I_n$ should be a good estimate of parameter $a_1$ ( see Hurwicz (1950), Zielenski (1999) and Luger (2005)). Below we will return to the ratio-based estimates. Of course, both of the dropped terms can produce an effect on the course of epidemics during much longer time than the starting period of the epidemic in which we are interested. As shown, equation (4) can be reduced to the following 1st order linear equation (we change the time index: $n + 2 \rightarrow n + 1$):

$$I_{n+1} \approx a_1 I_n .$$

This equation can be considered an indication of an exponential growth. Also, at the early phase of an outbreak, we can assume $S_n \approx S_0 \approx N$ (where $S_0$ is the initial number of susceptible at the beginning of the epidemic (if it ever happens) and $N$ is the total number of people in our *closed* population: $N = S + I + R$, with no birth, no death). With this assumption we arrive at the equation

$$I_{n+1} \approx (1 + \beta N - \delta )I_n \qquad (7)$$

Now we introduce an *alternative transmission rate* $\beta^* = \beta N$. Both $\beta$ and $\beta^*$ are popular in the epidemiological community. For more information about various forms of transmission rates see, for example, McCallum, Barlow and Hone (2001). According to Begon et al. (2002), transmission rates $\beta$ and $\beta^*$ discussed here are likely to remain benchmarks against which actual transmission dynamics are judged, and to remain key elements in most mathematical models of transmission. Note that $\beta$ and $\beta^*$ have different dimensions. With the alternative infection transmission rate, our difference equation takes a very simple form

$$I_{n+1} \approx (1 + \beta^* - \delta)I_n \qquad (8)$$

Finally, note that we have performed a number of approximations in developing *linear* equation (8) from strongly *nonlinear* equation (3). Each of these approximations results in some small error, and it is not easy to take those errors into account individually and altogether. In addition, we have some stochastic variation in $\beta$ and $\delta$. Combining all these sources of uncertainty in one, we arrive at our final AR(1) model

$$I_{n+1} = (1 + \beta^* - \delta)I_n + w_n \qquad (9)$$

where $w_n$ is an additive white Gaussian noise. At this moment, we are not interested in specifics of this noise. Note that we can derive the equations (7), (8) and (9) not from (4), but directly from the second equation in (1):

$$I_{n+1} = I_n + \beta S_n I_n - \delta I_n$$

By using the same assumption $S_n \approx S_0 \approx N$ as before, we immediately get linear models (7), (8) and (9). This is a result of linearization: our originally *nonlinear* system becomes (to a very good approximation) *linear* because the assumption $S \approx N$ turns our nonlinear transmission term $\beta SI$ into a linear one, $\beta^* I$. The sequence of approximations (4) - (9) discussed above can be symbolically expressed as SIR → AR(2) → AR(1). We adopt model (9) as our basic model for pre-epidemic and emerging epidemic processes. Note that almost all ARMA / ARIMA models for syndromic surveillance for respiratory diseases (influenza, SARS) use a 1st order autoregressive component (see Shtatland (2007)). That is why using AR(1) to describe a pre-epidemic process does not seem unusual. However, we are not interested in the most adequate description of a quiet, stationary period before the epidemic. Our intention is early detection of an outbreak, the take-off of the epidemic, and our very simple model (9) seems to be adequate for this purpose. When $\beta^* - \delta < 0$, i.e. the rate of recovery is greater than the transmission rate, we have a no-epidemic, stationary regime. When $\beta^* - \delta > 0$, there is an exponential growth, outbreak of epidemic. If $\beta^* - \delta = 0$, we have the so-called unit root case. Thus, there is a natural threshold in terms of $\beta^* - \delta$ to discriminate between these regimes. It is very important because of the lack of such a natural threshold in all previous works on time series modeling for biosurveillance. See, for example, Reis, Pagano and Mandl (2003) and Mohtashemi et al. (2006). Finally note that our AR(1) model (9) is based on the SIR model from which it inherits first principles.

## STATISTICAL INFERENCE ABOUT THE AR(1) MODEL

### ESTIMATING AR(1) PARAMETERS

Statistical inference about model (9) (confidence intervals, hypotheses testing, etc.) is based on estimates of the sole autoregressive parameter $a_1 = 1 + \beta^* - \delta$. For simplicity, we drop subscript index 1 here and to the rest of the paper. Thus, equation (9) can be written as $I_{n+1} = aI_n + w_n$. It is safe to assume that $a$ is *always* greater than 0 in our context. There exists a well-developed theory of estimating the AR(1) parameter. It includes ordinary least-squares, Yule-Walker, Burg, and various modified least-squares estimators (see, for example, Provost and Sanjel (2005) and references therein). The most widely used estimator is the ordinary least-squares (OLS) method that provides the following estimate of parameter *a* based on the time series $I_1$ , $I_2$ , …, $I_T$:

$$\hat{a}(T)= \sum^{T}_{n=2} I_n I_{n-1} \,/\, \sum^{T-1}_{n=1} I^2_n \qquad (10)$$

*T* can be considered a baseline of historical data used for estimating parameters of the model and making decisions. Here and further in the paper, a typical value of *T* is 7, though larger values (e.g., *T* = 14) and especially smaller values (from 2 to 6) can be useful. Note that for general ARMA processes, a *nonlinear* iterative least-squares procedure must be used for estimating ARMA parameters. It is well known that this procedure does not always converge successfully for a given set of data, particularly if the starting values of parameters are far from the resulting least-squares estimates (SAS/ETS® User's Guide (1993), pp 140-141). In case of AR(1), we have a very simple, explicit formula (10), which can be easily used and interpreted. The properties of estimate (10) are well known. It is consistent, i.e. $\hat{a}(T) \to a$ (the real value of the autoregressive parameter) as $T \to \infty$. At the same time $\hat{a}(T)$ is a biased estimator. The bias $b_T(a)$ depends on the real value of the parameter to be estimated and the time window *T*. According to Le Breton and Pham (1989), if $0 < a < 1$ (a stationary case), then

$$b_T(a) \approx - 2a / T \qquad (11)$$

and if $a > 1$ (an explosive, outbreak case), then

$$b_T(a) \approx - C\, T^{½} / a^T \qquad (12)$$

where *C* is a known constant which depends on *a*. In Le Breton and Pham (1989), there is a more complicated formula for $b_T(a)$ in the unit root case ($a = 1$) that is similar to (11) with a different constant. This case is not as important to us as the two previous ones. In practice we are never in a position of knowing the true mean of the process of interest. If the mean is unknown, we have to use a model with intercept. In this case the bias becomes even larger than (11), it is given by following formula

$$b_T(a) \approx - (1 + 3a ) / T \qquad (11')$$

(see Kendall (1954), Sawa (1978), Orcutt and Winokur (1969)). It is interesting that the bias $b_T(a)$ *is always negative* which means that the ordinary least-squares estimator *always underestimates a*. The bias can be rather substantial when the unknown parameter *a* is in the vicinity of 1. For example, according to (12) an outbreak value *a* = 1.1 might be estimated as 0.96 (the bias equals 0.14 or 13% of the real value) with *T* = 7. Let us remind that *T* = 7 is used in Reis, Pagano and Mandl (2003) and Mohtashemi et al. (2006). The example above shows how easy it is to misspecify the real regime if we do not take the bias into account**.** It is important to remember that formulas (11), (11') and (12) are asymptotic and ideally are meant to be used for large enough *T*. So using them for *T* = 7 is rather problematic. Instead of using formulas (11), (11') and (12) for bias correction we can also use a modified least-squares estimator defined by formula (10')

$$\hat{a}(T)= \sum^{T}_{n=2} I_n I_{n-1} \,/\, \sum^{T-1}_{n=2} I^2_n \qquad (10')$$

which differs from (10) only in the denominator (now summation is performed from *n* = 2 to *n* = *T* - 1, rather than from *n* = 1 to *n* = *T* - 1). The idea behind this correction is very simple: dropping a positive term in the denominator results in overall estimate increase. According to Provost and Sanjel (2005), this simple correction can be surprisingly effective**,** including cases with small values of *T*.

Another interesting and more general approach to unbiased estimating of parameter $a = 1 + \beta^* - \delta$ in (9), was initiated by Hurwicz (1950) and developed further by Zielinski (1999) and

Luger (2005). In particular, Hurwicz (1950) observed that every ratio $I_{n+1} / I_n$, $n$ = 1, 2, …, $T$-1, is a *median-unbiased* estimator of $a$ (though inefficient in statistical terms, since these ratios have a Cauchy distribution). Hurwicz (1950) hypothesized that the *median* of the ratios mentioned above would be a more efficient estimate of $a$ and perhaps an unbiased one. This hypothesis has been proved by Zielinski (1999) who has showed that the Hurwicz estimator

$$\hat{a}^{Hur}(T)=median(I_2 / I_1, I_3 / I_2, \ldots, I_T / I_{T-1}) \qquad (13)$$

is median-unbiased, robust against any deviation from Gaussian distribution, including heavy tails as well as contamination with outliers. Moreover, it has been proved that innovations $w_n$ in (9) are not necessarily identically distributed. Luger (2005) has showed that the results mentioned above remain true under more general distributional assumptions, without assuming statistical independence. Thus, assumptions basic for the least squares theory which are very difficult to verify in practice or to prove theoretically, are not necessary for the median-unbiased estimator approach. The fact that ratio $I_{n+1} / I_n$ is the *main part* of coefficient $a$ in (5) (see above) and properties of this ratio as an unbiased estimator of $a$ will allow us to use $I_{n+1} / I_n$ as the most *local* estimate which is based only on the *two consecutive* terms of time series. This locality will be important for us in the light of early detection requirements (epidemics and explosive behaviors in finance). Mohtashemi et al. (2006) propose the following classification of timeliness of detection (with 7-day time window and simulating outbreaks lasting 7 days): “early detection” is defined as detection during the first 3 days of outbreak, “intermediate detection” represents detection during the 4th and 5th days of outbreak, and “late detection” corresponds to detection in the last 2 days of outbreak. Using the SIR model with detection time window $T$ = 7 and simulated outbreaks, Mohtashemi et al. (2006) report cumulative sensitivity of 10% on the very 1st day of the outbreak, 30% - on the 2nd day, 50% - on the 3rd day, 63% - on the 4th day, 75% - on the 5th day, and 78% and 87.5% - on the 6th and 7th days correspondingly. Thus, according to Mohtashemi et al. (2006) their SIR model combined with Least-Squares estimation provides 50% sensitivity at the early stage of the epidemics (first 3 days) and 75% sensitivity at the end of the intermediate detection stage. Of course, real outbreaks may last well beyond 7 and even 14 days. Mohtashemi et al. (2006) and Reis, Pagano and Mandl (2003) focused on the first few days because useful detection systems should be able to recognize outbreaks within that time frame. Combining this information with theoretical results regarding formulas (10) – (13), the simulation results from Marriott and Pope (1954), Orcutt and Winokur (1969), Sawa (1978), Andrews (1993), and Provost and Sanjel (2005), and the outcomes of our simulations, we can conclude:

● As our AR(1) model is a linear approximation (and a good one) of the SIR model in Mohtashemi et al. (2006), we can expect a comparable performance for such short-term dynamics as flu-like epidemics (with seven-day simulated outbreaks). To achieve such a performance it is recommended to use a combination of estimates (10), (10’) and (13) with bias-corrections (11), (13) and (11’) not only for $T$ = 7 but also for shorter time windows including $T$ = 2. According to Reis, Pagano and Mandl (2003), time window $T$ should be *at most* 7 days, *i. e.* the authors assume the possibility of T < 7. Also, to have a non-zero sensitivity on the 1st day, we must use the ratio $I_{n+1} / I_n$ among other statistics. Note that for $T$ = 2, both OLS estimator (10) and Hurwicz estimator (13) are equal to the ratio $I_{n+1} / I_n$. This is the only case when estimators (10) and (13) are equal and both represent the *instant change* on the recent day comparatively to the previous one. This is why the ratio $I_{n+1} / I_n$ should be an important component of our statistical toolbox.

● We have compared the OLS estimate (10) with $\hat{a}^{Hur}$ estimate (13) using simulations. Our simulations results with 10,000 simulations for $T$ = 4, 5, 6, 7 and 14, and $a$ = 0.7, 0.8, 0.9, 1.0, 1.1, 1.2 and 1.47 (the last value is taken as a parameter of exponential growth for simulating outbreaks in Mohtashemi et al. (2006)) show that $\hat{a}^{Hur}(T)$ is much less biased and much less variable across the values of $T$ than its OLS counterpart (10) for all combinations of $T$ and $a$. Bias of all our estimates decreases with the increase in $a$ for $a$ > 1. When we average our OLS estimates across simulations, using median rather than mean, both bias and variability dramatically decrease. Also, the results depend on the fact whether we apply some induction period with discarding some terms (in our case 50 terms) of simulated time series. We find that discarding can significantly improve both bias and variability. See also Anderson (1979).

● Actually, statistics (10), (10'), (11), (11'), (12) and (13) are developed for steady-state regimes in which parameter *a* is unchanged during baseline T, and we have either a stationary regime or an exponential growth. In early detection, at some moments we are dealing with a *mix* of these regimes, so strictly speaking our statistics are not applicable. In particular, statistic $\hat{a}^{Hur}(T)$ (with $T$ = 7) being a robust and stable estimate is ideal for either stationary, pre-epidemic state or purely epidemic regime, but not for the periods including both pre-epidemic and epidemic days. For example, with $T$ = 7 the estimate $\hat{a}^{Hur}(T)$ will likely not react to the first days of the epidemic (from the 1st to the 3rd day), so we can miss the early detection at all. At the same time, OLS estimate can potentially react to the very 1st day, though probably in a rather weak form. To have this reaction stronger we have to use smaller time window, including $T$ = 3 and even $T$ = 2. As noted above, $I_{n+1} / I_n$ is the only estimate capable of detecting change-point at the very first day.

● Using higher-order moment estimates (Huzii (1981)) and a partial correlation approach (Anderson (1990)) is also a possibility.

Finally, when estimating parameter $a = 1 + \beta^* - \delta$ we actually estimate $\beta^* - \delta$, the difference between the transmission rate and recovery rate. This difference is a *threshold parameter*, closely related to another famous threshold parameter, $R_0 = \beta^* / \delta$, which is known in epidemiology as the basic reproductive ratio. Parameter $\beta^* - \delta$ can be called "an engine of the epidemic". Also note that here we use $\beta^*$ as the transmission rate parameter, and not $\beta$. Another interesting observation is that the type of behavior (a no-epidemic vs. an outbreak regime) is determined not by $\beta^*$ and $\delta$ *separately*, but by their difference $\beta^* - \delta$. In our approach we do not need to estimate or make some assumptions about $\beta^*$ and $\delta$ separately (for example, assume that $\delta \approx 1/7$ and then estimate $\beta^*$). We need to estimate only the difference $\beta^* - \delta$.

## CONFIDENCE INTERVALS AND HYPOTHESES TESTING FOR AR(1) MODELS

As in the previous section, the theory below is developed only for stable regimes in which parameter *a* is unchanged. Thus strictly speaking, it cannot be very useful in early detection situation, when the first days of the time window *T* are non-epidemic and then the epidemic begins. Also, we consider here only OLS estimates though generalizations are possible to other classes of estimates, including median-unbiased ones (in the spirit of Andrews (1993) and Luger (2005)). We will return to confidence intervals for median-unbiased estimators elsewhere.

Even with bias correction discussed above, the point estimate of the form (10) cannot serve as the sole basis to distinguish the case $a < 1$ (stationarity) from $a > 1$ (outbreak). The point estimate can be considered rather as the "best guess" for an unknown parameter. To choose between these possibilities, we have to use either confidence intervals, or hypotheses testing, or both. Although a statistician will recognize the typical hypothesis-testing problem in biosurveillance (with Type I and Type II errors), we prefer to use an equivalent confidence-interval language following Tukey (1991) who argues strongly that confidence intervals are more informative and more honest than p-values. Confidence interval results are based on the following statements:

● If $a < 1$, then the probability distribution of

$$(\hat{a}(T) - a)\, T^{½} / (1 - \hat{a}^2(T))^{½}$$

is close to the Gaussian distribution $N(0,1)$ when the time window $T$ is *large enough* (see Giraitis and Phillips (2006)).

● If $a > 1$, then the probability distribution of

$$(\hat{a}(T) - a)\, \hat{a}^T(T) / (\hat{a}^2(T) - 1)$$

is close to the standard Cauchy distribution $C$ when the time window $T$ is *large enough* (see Phillips, Wu and Yu (2007)).

As a result, we have the following two families of the approximate two-sided confidence intervals for parameter *a*:

● If *a* < 1 (a stationary case), the 100(1 - *α*)% confidence interval for *a* is given by the formula

$$(\hat{a}(T) - ((1 - \hat{a}^2(T))^{½} / T^{½})\, N_\alpha ,\ \ \hat{a}(T) + ((1 - \hat{a}^2(T))^{½} / T^{½})\, N_\alpha ) \qquad (14)$$

where $N_\alpha$ is the two-tailed α percentile critical value of the standard Gaussian distribution. For 90, 95 and 99 percent confidence intervals, these critical values are as follows: $N_{0.10}$ = 1.645, $N_{0.05}$ = 1.96, $N_{0.01}$ = 2.576. Note that for small sample sizes (e.g., *T* = 7), using confidence limits based on the Student's *t* distribution rather than Gaussian one, is more appropriate.

● If *a* > 1 (an explosive case), the 100(1 - *α*)% confidence interval for *a* is given by the formula

$$(\hat{a}(T) - ((\hat{a}^2(T) - 1)) / \hat{a}^T(T))C_\alpha ,\ \ \hat{a}(T) + ((\hat{a}^2(T) - 1)) / \hat{a}^T(T))C_\alpha ) \qquad (15)$$

where $C_\alpha$ is the two-tailed α percentile critical value of the standard Cauchy distribution. For 90, 95 and 99 percent confidence intervals, these critical values are as follows: $C_{0.10}$ = 6.315, $C_{0.05}$ = 12.7, $C_{0.01}$ = 63.657 (see Phillips, Wu and Yu (2007)). Gaussian critical values are much smaller than the corresponding Cauchy ones, which is not surprising since the Cauchy distribution has heavier tails.

It is worth noting that we are confronted here with a rather unconventional hypothesis testing / confidence interval situation:

● Both the null hypothesis of stationarity $H_0$ (0 < *a* < 1) and the alternative hypothesis of outbreak $H_1$ (*a* > 1) are composite, which makes hypothesis testing more difficult. The problem can be reduced to testing $H_0$: *a* = 1 versus $H_1$: *a* > 1, though we are not very interested in a borderline case of *a* = 1 (the unit root or random walk case) since it is too special for our practical needs and at the same time very difficult analytically.

● Test statistics and their approximate distributions in (13) and (14) are *different* (which is unusual).

● Confidence intervals in (14) are often too wide to be practical for the time window *T* = 7 adopted in Reis, Pagano and Mandl (2003) and Mohtashemi et al. (2006). For example, if *â(T)* = 1.25 (with bias correction already taken into account and *δ* ≈ 1/7) which corresponds to a high value of the basic reproductive ratio: $R_0$ ≈ 2.75, then the 90% two-sided confidence interval based on the Cauchy distribution is of the form: (1.25 - 0.36, 1.25 + 0.36). This interval contains the borderline value *a* = 1 and therefore we are in the "Grey Zone" and fail to detect a rather strong signal as a take-off of the epidemic. Now let us consider an example with *â(T)* = 1.5 (with bias correction already taken into account and *δ* = 1/7) which corresponds to *β** ≈ 0.643 and $R_0$ ≈ 4.5. It is an important example, since *â(T)* = 1.5 is very close to the parameter of exponential growth for simulated outbreaks **(***λ* ≈ 1.47) in Mohtashemi et al. (2006). Using 90 percent confidence interval (14) with $C_{0.10}$ = 6.315, we obtain the interval (1.5 - 0.46, 1.5 + 0.46) which does *not* contain *a* = 1 and thus we have to reject stationarity or a no-outbreak regime. But this rejection (with 10% significance level) is almost borderline. To improve the situation we can build a one-sided confidence interval, instead of the typically recommended two-sided one. As a result, we arrive at the 95 % confidence interval (1.5 - 0.46, + ∞) that does not cover *a* = 1. Of course, elevating the significance level from α= 10% to α= 5% makes us psychologically more ready to reject stationarity in favor of an outbreak. But what we did can be considered a kind of statistical trick to get a more or less tolerable (in terms of the level of confidence) confidence interval. We can add to this that early detection in syndromic surveillance is inherently *multiple testing* problem since estimating, testing and making decisions are performed daily. Thus, significance levels used in testing procedures cannot be taken at face value. Moreover, they could be absolutely misleading. Note that most probably our early detection problem with *T* = 7 (or even less) should not be considered in terms of *statistical significance* only, *practical significance* can be even more important.

## APPLICATIONS TO ECONOMIC AND FINANCIAL TIME SERIES: SIR, AR(1) MODELS, STOCK MARKET BUBBLES AND MICRO-BUBBLES

It can be shown that the results discussed above are applicable to the problems of testing explosive behaviors in economics and finance, in particular in date-stamping the origination of economic bubbles (see, for example, Phillips, Wu and Yu (2007), Phillips and Magdalinos (2007) and references therein). In economic and financial applications, only the time-scale can be different: time is usually measured in *months*, rather than *days*. But the autoregressive model AR(1) of the form (9), the least-squares estimate (10) and confidence intervals (13) and (14) are the same. However, a substantial difference between both applications is the following: in our epidemiological application, we are interested in early detection only; at the same time in economics, we are equally interested in the origination and collapses of economic bubbles. It is worth noting that the authors in Phillips, Wu and Yu (2007) use also one-sided tests, more exactly, they apply right-sided unit root tests against explosiveness to have more discriminatory power in detecting explosive processes.

Note that the paper by Phillips, Wu and Yu (2007) being a fundamental research article on modeling bubble formation, uses AR(1) as an ad-hoc, purely empirical model, without any explanation of why AR models, in particular AR(1), should be used. In this connection, it is interesting to mention the paper by Shive (2006) in which the author used a SIR model as an epidemic model of investor behavior that results in building a stock market bubble. Shive (2006) reports periods of significant epidemic behavior in buying and selling that correspond to the dramatic price movements in some stocks in Finland from 12/1994 to 1/2004. The author refers to the book of Yale economics professor Robert Shiller (2005) in which he defines a speculative bubble “as a situation in which news of price increases spurs investor enthusiasm, which spreads by *psychological contagion* from person to person”. We can add to this that Shiller is world famous by predicting the year 2000 stock market crash. Also, he is one of the first econometricians to describe a new bubble in the making – the housing bubble.

Thus, we can see that psychological and physiological contagion is a common and natural justification for using SIR models and their linear approximations, AR(1) models both in epidemiology and finance. The real strength of the approach with using SIR and then AR(1) models in financial applications is the straightforward, almost clinical way in which it explains how and why financial bubbles happen as they do. Still, we can conclude that using autoregressive models in describing explosive phenomena seems to be in an emerging state.

See also Abreu and Brunnermeier (2003) for a comprehensive paper on bubbles and crashes, Brunnermeier and Pedersen (2002) for an interesting work on predatory trading, and Brunnermeier and Nagel (2004) for a very influential paper on hedge funds and technology bubbles. In Brunnermeier and Pedersen (2002), the authors focus more on higher frequency phenomena, including mini- or micro-bubbles, etc, rather than bubbles.

## WORKFLOW FOR EARLY DETECTING EXPLOSIVE BEHAVIORS

We suggest the following workflow for early detecting explosive behaviors:

1) Each day, the ordinary least-square (OLS) estimate (10) and median-unbiased estimate (13) of the only autoregression parameter *a* are iteratively calculated, based on the past *T* values of the time series $I_n$ (the present date value is included). The changes in the time index are obvious.

↓

2) Bias correction is performed by using either formulas (11), (11’) and (12) or a modified least-squares estimator given by (10’). Comparing OLS and median-unbiased estimates (13) can be useful here.

↓

3) The confidence intervals are built either of the form (14) if the OLS estimate *â(T)* is smaller than one or of the form (15) if this estimate exceeds one.

↓

4) If the confidence interval (14) does not contain the borderline *a* = 1, the decision "**No Alarm**" can be made.

↓

5) If the confidence interval (15) does not contain the borderline *a* = 1, the decision "**Alarm**" can be made.

↓

6) Steps 4 and 5 are formulated it terms of statistical significance. Taking into account the small length of the time window (at most *T* = 14, typically *T* = 7 or even less) and as a result large width of confidence intervals, most likely the final decisions "**Alarm**" or "**No Alarm**" will be made very seldom based on confidence intervals only.

↓

7) If the interval (14) or (15) covers *a* = 1, we are in the "Grey Zone" and can apply some of the following empirical rules (reasonable, though more or less arbitrary):

    a) If the most recent ratio $I_{n+1} / I_n$ is greater than 1, then "**Alarm**" . This rule maximizes the likelihood of the anomaly detection on the very first day, though at a prize of increase in the number of false alarms

    b) If two most recent ratios $I_{n+1} / I_n$ in a row are greater than 1, then "**Alarm**". This rule facilitate the outbreak detection during the first two days

    c) If three days in a row we have *â(T)* > 1, then "**Alarm**", etc.

In general, using *multiple time windows* (*T* = 2, 3, 4, 5, 6, 7 and 14) can be strongly advised and not only for the Grey Zone. The estimates obtained for time window *T* = 14 will be smoother and Gaussian / Cauchy distribution approximations more accurate. At the same time, estimates and confidence intervals with time window T = 14 would "feel" the start of epidemics even later than with T = 7. Comparing results for T = 7 and T = 2, 3, 4, 5, 6 and 14 might be helpful.

## SUMMARY AND CONCLUSIONS

In this paper, we propose a simple but powerful model for early detection of epidemic outbreaks. First, based on a SIR model, a workhorse of epidemiology, we show that in the autoregressive approach we can limit ourselves with the models of the second or even first order. Then, by using methods of linearization, we conclude that the most appropriate model is AR(1). The autoregression coefficient of this model has a very simple epidemiological meaning and can be used as a natural threshold parameter in epidemic detection. To the best of our knowledge, this paper and Shtatland (2007) are the first publications that report this threshold parameter for early detection in biosurveillance applications. We can consider our AR(1) model as a paradigm-model for early detection of epidemic outbreaks. Note that AR(1) is already widely used as a paradigm-model in the unit root theory. The proposed detection method is based on a combination of the least-squares estimator of the autoregressive coefficient with approximate two-sided and one-sided confidence intervals (built by using the Gaussian and Cauchy limit distributions) and median-unbiased estimators. Note that typically ARMA processes are considered synonyms to the stationary models, and the explosive case is ignored as a nuisance. Here AR(1) model with autoregressive coefficient greater than one plays a *major role*. Our theoretical results are comparable with the results of Mohtashemi et al. (2006), which are based on the real data from the emergency department of a large urban hospital, with simulated outbreaks. However, there is an important difference between our approach and that of Mohtashemi et al. (2006), which is worth mentioning. Mohtashemi et al. (2006) heavily use historical data (more than 4389 days), and we do not (14 days or less). Our approach can be preferred when historical data are either not available at all or are unrepresentative of the current behavior for a variety of reasons: changes in treatments, coding, and reporting practices, appearance of new influenza strains, etc. (See more about these extreme situations with historical data in Burkom, Murphy and Shmueli (2007)). In this respect, the two approaches can be considered complementary. Also, our approach can be used in economic and financial time series and any application where we are interested in detection of abrupt switches from the normal, stationary regime to the exponential growth. Models of the form (9) seem well suited to capturing the essential features of epidemic, economic and financial processes that can undergo mildly explosive behavior (including stock market bubbles and higher frequency phenomena such as micro-bubbles, etc.) The purely temporal modeling results discussed above are interesting by themselves because they provide the key components of real-time syndromic surveillance. Following Shtatland, Kleinman and Cain (2006), we can add that such results could also be used in a more general, space-time detection schemes with possible application of logistic and Poisson regressions.

## REFERENCES

Abreu, D. and Brunnermeier, M. (2003) Bubbles and crashes, *Econometrica*, 71, 173-204.

Anderson, O. D. (1979). On warming-up time series simulations generated by Box-Jenkins models. *Journal of the Operational Research Society*, 30(6), 587-589.

Anderson, O. D. (1990). On the partial autocorrelations for an explosive process. *Communications in Statistics - Theory and Methods*, 19(9), 3505-3526.

Andrews, D. W. K. (1993). Exactly median-unbiased estimation of first order autoregressive / unit root models. *Econometrica*, 61(1), 139-165.

Begon, M., Bennett, M., Bowers, R. G., French N. P., Hazel, S. M., and Turner, J. (2002). A clarification of transmission terms in host-microparasite models: numbers, densities and areas. *Epidemiology and Infection*, 129, 147-153.

Brunnermeier, M. K. and Nagel, S. (2004). Hedge funds and the technology bubble. *Journal of Finance*, 59, 2013-2040.

Brunnermeier, M. K. and Pedersen, L. H. (2002). Predatory trading. Working paper, Princeton University and NYU.

Burkom, H. S., Murphy, S. P. and Shmueli, G. (2007). Automated time series forecasting for biosurveillance. *Statistics in Medicine* (in press).

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.

Giraitis, L. and Phillips, P. C. B. (2006). Uniform limit theory for stationary autoregression. *Journal of Time Series Analysis*, 27(1), 51-60.

Huzii, M. (1981). Estimation of coefficients of an autoregressive process by using a higher order moment. *Journal of Time Series Analysis*, 2(2), 87-93.

Hurwicz, L. (1950). Least-squares bias in time series. In *Statistical Inference in Dynamic Economic Models* (ed. T. C. Koopmans). New York, NY: Wiley.

Kato, H. and Ozaki, T. (2002). Adding data process feedback to the nonlinear autoregressive model. *Signal Processing*, 82, 1189-1204.

Kendall, M. G. (1954). Note on bias in the estimation of autocorrelation. Biometrika, 41, 403-404.

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

Le Breton, A. and Pham, D. T. (1989). On the bias of the least squares estimator for the first order autoregressive process. *Ann. Instit. Statist. Math.* 41(3), 555-563.

Luger, R. (2005). Median-unbiased estimation and exact inference methods for first-order autoregressive models with conditional heteroscedasticity of unknown form. *Journal of Time Series Analysis*, 27(1), 119-128.

Marriott, F. H. C. and Pope, J.A . (1954). Bias in the estimation of the autocorrelation. Biometrika, 41, 390-402.

Orcutt, G. H. and Winokur, H. S. (1969). First order autoregression: inference, estimation, and prediction. *Econometrica*, 37, 1-14.

Phillips, P. C. B., Wu Y. and Yu J. (2007). Explosive behavior and the NASDAQ bubble in the 1990s: When did irrational exuberance escalate asset values? *Cowles Foundation for Research in Economics*, Yale University, Working Paper.

Phillips, P. C. B. and Magdalinos, T. (2007). Limit theory for moderate deviations from a unit root. *Journal of Economics*, 136, 115-130.

Phillips, P. C. B. and Han, C. (2006). Gaussian inference in AR(1) time series with or without a Unit Root. *Cowles Foundation Discussion Papers*, #1546, Yale University.

Provost, S. B. and Sanjel, D. (2005). Inference about the first-order autoregressive coefficient. *Communications in Statistics – Theory and Methods*, 34, 1183-1201.

Reis, B. Y. and Mandl, K. D. (2003). Time series modeling for syndromic surveillance. *BMC Medical Informatics and Decision Making*, 3:2.

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.

SAS Institute Inc. (1993), *SAS/ETS® User's Guide, Version 6, Second Edition*, Cary, NC: SAS Institute Inc.

Sawa, T. (1978). The exact moments of the least squares estimators for the autoregressive model, *Journal of Econometrics*, 8, 159-172.

Shiller, R. J. (2005), *Irrational Exuberance*, *Second Edition,* Princeton University Press.

Shive, S. (2006). An epidemic model of investor behavior. Working Paper, *Ross Business School, University of Michigan*.

Shtatland, E. S., Kleinman, K. and Cain, E. M. (2006). Biosurveillance and outbreak detection using the ARIMA and LOGISTIC procedures. *SUGI'31 Proceedings*, Paper 197-31, Cary, NC: SAS Institute, Inc.

Shtatland, E. S. (2007). Low-order autoregressive models in early detection of epidemic outbreaks and explosive behaviors in economic and financial time series. *NESUG'20 Proceedings*, Paper PO19, NorthEast SAS® Users Group, Inc.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, 6(1), 100-116.

Wearing, H. J., Rohani, P. and Keeling, M. J., (2005). Appropriate models for the management of infectious diseases. *PLoS Medicine*, 2:7, 621-627.

Zielinski, R. (1999). A median-unbiased estimator of the AR(1) coefficient. *Journal of Time Series Analysis*, 20(4), 477-481.

**CONTACT INFORMATION**

Ernest S. Shtatland
eStatConsulting
200 Park Terrace Drive, Suite 234
Stoneham, MA 02180
tel: (781) 662-9578
email: eshtatland@yahoo.com