Listy Biometryczne - Biometrical Letters
Vol. 39(2002), No. 2, 53-61

Forecast of the expected nonepidemic morbidity of acute diseases

Andrés M. Alonso[1], Armando Aguirre[2], Edilberto González[3]

[1]Department of Mathematics, Universidad Autónoma de Madrid, Facultad de Ciencias, CXV – Despacho 606, 28049 Madrid, Spain
[2]Unit of Experimental Clinic Research, Hospital La Candelaria de Tenerife, Spain
[3]Unit of Epidemiological Surveillance, Instituto de Medicina Tropical Pedro Kourí, Cuba

SUMMARY

In epidemiological surveillance it is important that any unusual increase of reported cases be detected as rapidly as possible. Reliable forecasts based on a suitable temporal model of an epidemiological indicator are necessary to estimate the expected nonepidemic indicator and elaborate an alert threshold. First, we present a method for identifying and replacing the abnormal values in the time series, then we apply the autoregressive integrated moving average approach to forecast the expected nonepidemic morbidity of acute respiratory infections and acute diarrhoeal diseases. Using this approach, we are able to detect the starting data of new epidemic values under routine surveillance conditions.

KEY WORDS: epidemiological surveillance; time series; alert threshold.

1. Introduction

Epidemiological surveillance is an important component for the public health system because it provides opportune information about the course of diseases and others health events that may lead to corrective actions from the health sector. Epidemiological surveillance consists of three interrelated components: continuous systematic data collection, use of analysis models and inferences from data, and rapid dissemination of findings to help the public health decision-making process (Thacker and Berkelman, 1998; Nofre, 1992). One major aspect for surveillance systems is to forecast accurately the case occurrence of health events and to detect abnormal values in case occurrence. An approach used to investigate this problem is based on time series forecasting models for specific health variables (Serfling, 1963; Choi and Thacker,

1981; Helfentein, 1986; Zaidi et al., 1989; Watier et al., 1991; Nofre, 1992), particularly the autoregressive integrated moving average models (ARIMA), developed by Box and Jenkins (1976), is used with success in different situations (Choi and Thacker, 1981; Helfentein, 1986; Zaidi et al., 1989; Watier et al., 1991).

In this paper, we use ARIMA models to forecast the weekly expected morbidity of acute respiratory infections (ARI) and acute diarrhoeal diseases (ADD), that constitute an important health problem, involving a great deal of the physician, nurses and health worker's time in the majority of the countries of our region. The data presented in this paper consist of 1985-1990 weekly reports of medical patient consultation from ambulatory facilities all over the country of ARI and ADD in children from 1 to 4 years old. This information is received each week in the Unit of Epidemiological Surveillance at the Institute of Tropical Medicine Pedro Kourí from Provincial Centers of Epidemiology and Hygiene.

2. Methods

For the ARIMA approach to time series modeling it is necessary to observe a long equally spaced series of values in a stationary mode. Usually, in infectious diseases series we observe epidemic peaks superimposed on a stationary process, which correspond to extreme high values (Watier et al., 1991). Different criteria can be used for detect and replace these abnormal values (Choi and Thacker, 1981; Watier et al., 1991; Aguirre and González, 1992) and we shall consider one of these.

For a time series $\{X_t, t = 1, ..., n\}$ observed weekly during a number of years (1985-1989 in our case) which does show trend and seasonality, we consider a point X_t as epidemic if its value is greater or equal to the upper 90% confidence limit forecast by the following model:

$$X_t = a + bt + \sum_{i=1}^{2} (c_i \sin(2\pi it/T) + d_i \cos(2\pi it/T)) + a_t, \tag{1}$$

where X_t is the number of cases in week t, T is the series periodicity, a_t are independent identically distributed random variables with mean zero and finite variance. The coefficients a, b, c_i, d_i were estimated by the least-squares method.

The epidemic points are replaced in the original series by the expected number of cases calculated from (1). Therefore we arrive to a smoothed series, which structure can be estimated by a seasonal ARIMA$(p, d, q)(P, D, Q)_{52}$ model and forecasts can be calculated for the next period of 52 weeks. A general formulation of the ARIMA$(p, d, q)(P, D, Q)_S$ model is the following:

$$\phi(B)\Phi(B^S)(1 - B^d)(1 - B^{SD})X_t = \theta(B)\Theta(B^S)e_t, \tag{2}$$

where $\phi(B) = 1-\phi_1 B-...-\phi_p B^p$ and $\Phi(B^S) = 1-\Phi_1 B^S-...-\Phi_P B^{SP}$ are the regular and seasonal autoregressive polynomials, respectively; $\theta(B) = 1 - \theta_1 B - ... - \theta_q B^q$ and $\Theta(B^S) = 1 - \Theta_1 B^S - ... - \Theta_Q B^{SQ}$ are the regular and seasonal moving average polynomials, respectively; B is the backshift operator, that is, $B^m X_t = X_{t-m}$ and $\{e_t\}$ is a series of uncorrelated random variables with mean zero and variance σ_e^2.

When a suitable model is fixed, we defined as an alert threshold the upper 95% confidence limit forecast by the seasonal ARIMA model. As a warning criterion we take the observation of two or more consecutive weeks in the case occurrence above the alert threshold. Experience with this criterion has shown that it usually indicates situations of epidemiological interest (Serfling, 1963; Aguirre and González, 1992; Aguirre and Alonso, 1993).

In the following we describe the methodology of estimation and diagnostic checking used in the next section. We use the procedure ARIMA of STATGRAF software which implement the conditional maximum likelihood estimation method, that is the parameters estimates maximize the function:

$$l(\vartheta, \sigma_e^2) = -\frac{n}{2}\ln(2\pi\sigma_e^2) - \frac{S(\vartheta)}{2\sigma_e^2}, \tag{3}$$

where ϑ denotes the model parameters ($\vartheta = (\Theta, \theta)$ in model (9) and $\vartheta = (\Theta, \phi_1, \phi_2)$ in model (10)), $S(\vartheta) = \sum_{t=SP+p+1}^{n} e_t^2(\vartheta)$ is the conditional sum of squares function and $e_t(\vartheta)$ are the residuals obtained using the value ϑ as if it were the true value of the model's parameters. A full explanation of this estimation method is presented e.g. in section 7.2.1 of (Wei, 1990).

After parameter estimation, we assess the model adequacy by checking whether the model assumptions are satisfied. Particularly, we check if the $\{e_t\}$ is a white noise process, that is $\{e_t\}$ is a series of uncorrelated random variables with mean zero and constant variance σ_e^2. Also, we check whether the errors are normally distributed since confidence intervals (8) are based on this assumption.

In order to test if the error mean is zero we use the following t-test:

$$t = \frac{\bar{\hat{e}}}{\hat{\sigma}_e}, \tag{4}$$

where $\bar{\hat{e}} = (n - SP - p)^{-1} \sum_{t=SP+p+1}^{n} \hat{e}_t$, $\hat{e}_t = e_t(\hat{\vartheta})$, denotes the estimated residuals and $\hat{\sigma}_e^2 = \frac{S(\hat{\vartheta})}{d.f.}$ with the number of degrees of freedom $d.f.$ equal to the number of terms used in the sum $S(\vartheta)$ minus the number of parameters estimated, i.e. $d.f. = (n - SP - p) - (p + q + P + Q + 1)$.

A crucial assumption in ARIMA modeling is the uncorrelation of the model's errors. Several diagnostic goodness-of-fit tests have been proposed based on the residual autocorrelations to check the joint null hypothesis $H_0 : r_1 = r_2 = ... = r_m = 0$,

where r_i is the i-th error's autocorrelation. In this paper, we use the Ljung-Box statistic (Ljung and Box, 1978), Q_{LB}, which is a modification of the portmanteau test proposed in Box and Pierce (1970):

$$Q_{LB} = n(n+2)\sum_{k=1}^{m}(n-k)^{-1}\hat{r}_k^2, \tag{5}$$

where $\hat{r}_k = \sum_{t=k+1}^{n}\hat{e}_t\hat{e}_{t-k}/\sum_{t=1}^{n}\hat{e}_t^2$ are the residual autocorrelation. The asymptotic distribution of Q_{LB} can be approximated by a χ^2 distribution with $m-(p+q+P+Q)$ degrees of freedom. Recently, Peña and Rodríguez (2002) proposed a powerful portmanteau test based on the residual correlation matrix.

The last step in the diagnostic checking stage was to test whether the errors are normally distributed. We use the Kolmogorov-Smirnov test using the Lilliefors tables, see e.g. section 4.6 in (Gibbons and Chakraborti, 1992).

Once we have estimated and checked the selected model, we calculate the forecasts h steps ahead, $X_n(h)$, as follows:

$$\begin{aligned} X_n(h) &= \Psi_1 X_n(h-1) + \ldots + \Psi_{p+P+d+D} X_n(h-p-SP-d-SD) \\ &\quad + \hat{e}_n(h) + \Xi_1\hat{e}_n(h-1) + \ldots + \Xi_{q+Q}\hat{e}_n(h-q+QS), \end{aligned} \tag{6}$$

where

$$\begin{aligned} X_n(j) = E(X_{n+j} \mid X_n, X_{n-1}, \ldots), &\quad j \geq 1 \\ X_n(j) = X_{n+j}, &\quad j \leq 0 \\ \hat{e}_n(j) = 0, &\quad j \geq 1 \\ \hat{e}_n(j) = \hat{e}_{n+j}, &\quad j \leq 1 \end{aligned}$$

and polynomials $\Psi(B)$ and $\Xi(B)$ are defined by $\Psi(B) = \phi(B)\Phi(B^S)(1-B^d)(1-B^{SD})$ and $\Xi(B) = \theta(B)\Theta(B^S)$. These forecasts have the optimal property of minimizing the expected value of:

$$\sum_{h=1}^{52}(X_{n+h} - X_n(h))^2, \tag{7}$$

where $X_n(h)$ is the forecast h steps ahead, X_{n+h} would be the observed value at time $n+h$.

If the $\{e_t\}$ are normally distributed we can calculate the $100(1-\alpha)\%$ confidence intervals for $X_n(h)$ or $100(1-\alpha)\%$ forecast limits using the following expression:

$$X_n(h) \pm z_{\alpha/2}\left(1 + \sum_{k=1}^{h-1}\varphi_k^2\right)\hat{\sigma}_e, \tag{8}$$

where φ_k are the coefficients of polynomial $\varphi(B)$ defined by relation $\Psi(B)\varphi(B) = \Xi(B)$ and $z_{\alpha/2}$ is the $\alpha/2$ percentile of the standard normal distribution.

3. Results and discussion

The typical configurations corresponding to the seasonal ARIMA$(p,d,q)(P,D,Q)_{52}$ models were obtained from the analysis of the simple and partial autocorrelation functions of the ARI and ADD series. Even though with slight differences it was observed that the series were following a pattern type (9) for ARI series and a pattern type (10) for ADD series:

$$(1-B)(1-B^{52})X_t = (1-\Theta B^{12})(1-\theta B)e_t \tag{9}$$

$$(1-\phi_1 B-\phi_2 B^2)(1-B^{52})X_t = (1-\Theta B^{12})e_t. \tag{10}$$

That is, the ARI series follows an ARIMA(0,1,1)(0;1,1)$_{52}$ model whereas the ADD series follows an ARIMA(2,0,0)(0;1,1)$_{52}$.

With the selected models we proceeded to estimate the parameters using the procedure ARIMA of STATGRAF software. The estimated parameters as well as Ljung-Box statistic (Q_{LB}), Kolmogorov-Smirnov (K-S) test of normality of residuals are given in Tables 1 and 2 for each province. Provinces were coded as follows: PR Pinar del Río, PH Provincia Habana, CH Ciudad de La Habana, MT Matanzas, VC Villa Clara, CF Cienfuegos, SS Sancti Spíritus, CA Ciego de Ávila, CM Camagüey, LT Las Tunas, HO Holguín, GM Granma, SC Santiago de Cuba, GT Guantánamo and IJ Isla de la Juventud.

Table 1. Parameter estimates and model checking test for ARI series

Series	$\theta(s.e.)$	$\Theta(s.e.)$	$Q_{LB}(p)^*$	$K-S$ test p
PR	0.67 (0.05	0.52 (0.07)	26.37 (0.09)	$p>0.9$
PH	0.70 (0.05)	0.52 (0.07)	18.27 (0.43)	$p>0.9$
CH	0.81 (0.04)	0.81 (0.04)	25.75 (0.10)	$p>0.3$
MT	0.68 (0.05)	0.50 (0.07)	17.38 (0.49)	$p>0.9$
VC	0.77 (0.05)	0.62 (0.07)	15.63 (0.61)	$p>0.9$
CF	0.70 (0.05)	0.61 (0.07)	13.33 (0.77)	$p>0.9$
SS	0.73 (0.05)	0.63 (0.06)	20.87 (0.28)	$p>0.9$
CA	0.83 (0.04)	0.61 (0.06)	23.82 (0.16)	$p>0.9$
CM	0.78 (0.05)	0.56 (0.07)	15.22 (0.66)	$p>0.9$
LT	0.89 (0.03)	0.56 (0.07)	20.39 (0.31)	$p>0.3$
HO	0.73 (0.05)	0.45 (0.07)	23.87 (0.15)	$p>0.3$
GM	0.92 (0.03)	0.58 (0.06)	15.52 (0.62)	$p>0.9$
SC	0.80 (0.04)	0.65 (0.06)	18.21 (0.44)	$p>0.9$
GT	0.94 (0.03)	0.55 (0.07)	25.92 (0.10)	$p>0.9$
IJ	0.71 (0.05)	0.49 (0.07)	20.54 (0.30)	$p>0.9$
CUBA	0.66 (0.05)	0.57 (0.07)	17.32 (0.50)	$p>0.9$

*Box-Ljung statistic based on the first 20 residual autocorrelations,
p is the p-value of Box-Ljung test

Table 2. Parameter estimates and model checking test for ADD series

Series	$\phi_1(s.e.)$	$\phi_2/\theta(s.e.)^*$	$\Theta(s.e.)$	$Q_{LB}(p)^{**}$	$K-S$ test p
PR	0.36 (0.07)	0.19 (0.07)	0.44 (0.07)	20.19 (0.26)	$p > 0.5$
PH	0.34 (0.07)	0.22 (0.07)	0.47 (0.07)	7.30 (0.97)	$p > 0.4$
CH	0.30 (0.07)	0.33 (0.07)	0.48 (0.07)	12.33 (0.77)	$p > 0.9$
MT	0.42 (0.07)	0.30 (0.07)	0.53 (0.07)	16.71 (0.47)	$p > 0.1$
VC	0.47 (0.07)	0.26 (0.07)	0.62 (0.07)	19.03 (0.32)	$p > 0.1$
CF	0.32 (0.07)	0.31 (0.07)	0.55 (0.07)	24.74 (0.10)	$p > 0.9$
SS	0.31 (0.07)	0.18 (0.07)	0.48 (0.07)	14.09 (0.66)	$p > 0.9$
CA	0.87 (0.06)	0.58 (0.09)*	0.56 (0.07)	23.78 (0.12)	$p > 0.2$
CM	0.91 (0.04)	0.55 (0.08)*	0.54 (0.07)	9.40 (0.92)	$p > 0.9$
LT	0.88 (0.05)	0.52 (0.09)*	0.61 (0.07)	19.31 (0.31)	$p > 0.3$
HO	0.88 (0.05)	0.52 (0.09)*	0.52 (0.07)	24.89 (0.09)	$p > 0.9$
GM	0.20 (0.07)	0.26 (0.07)	0.57 (0.06)	16.53 (0.48)	$p > 0.9$
SC	0.27 (0.07)	0.26 (0.07)	0.50 (0.07)	16.96 (0.45)	$p > 0.05$
GT	0.23 (0.07)	0.26 (0.07)	0.47 (0.07)	12.07 (0.79)	$p > 0.9$
IJ	0.88 (0.08)	0.68 (0.11)*	0.56 (0.07)	20.93 (0.22)	$p > 0.9$
CUBA	0.43 (0.07)	0.26 (0.07)	0.47 (0.07)	16.39 (0.49)	$p > 0.9$

*In these series indicated by (*), the parameter estimated was θ, that is, the moving average parameter

** Box-Ljung statistic based on the first 20 residual autocorrelations, p is the p-value of Box-Ljung test

After obtaining the adjusted model for each province, we calculated the 1990 forecasts and their upper 95% confidence limit. The result for CUBA series is shown in Figures 1 and 2. Also, t-test values given in Table 3 show that the residual mean is not statistically different from zero. Therefore the seasonal ARIMA type of modeling seems well adapted to describe the underlying structure of ARI and ADD cases reported in non-epidemic periods.

The alert threshold previously estimated is compared with the observed data for the year 1990, and it was effective in detecting the epidemic episodes of ARI and ADD.

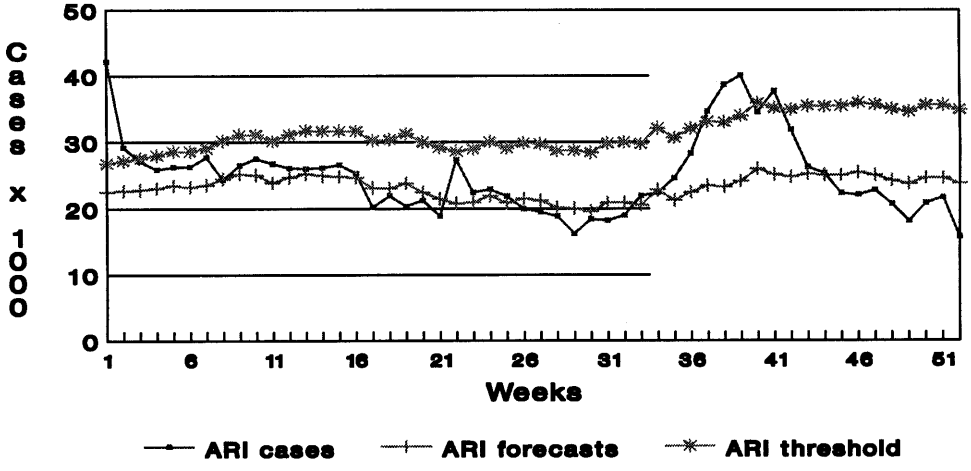

Figure 1. Estimated forecast function for ARI series. Cuba 1990.

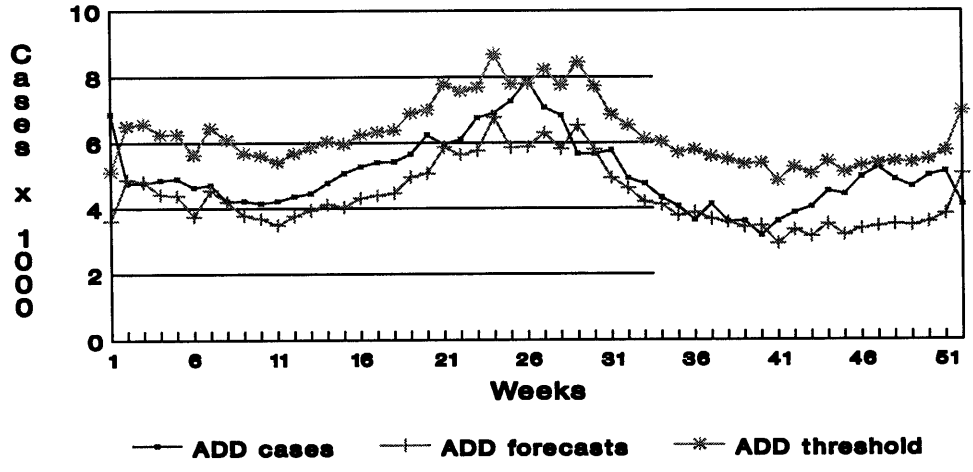

Figure 1. Estimated forecast function for ADD series. Cuba 1990.

Table 3. Mean, standard error and t-test for ARI and ADD series

	ARI ($n = 207$)			ADD ($n = 208$)		
Series	Mean e_t	σ_e	t-test p^*	Mean e_t	σ_e	t-test p^*
PR	-2.79	107.28	0.70	0.59	45.52	0.85
PH	0.61	254.08	0.97	1.00	82.01	0.85
CH	0.29	616.80	0.99	29.73	284.78	0.12
MT	-2.67	135.17	0.77	0.39	62.54	0.92
VC	0.40	182.91	0.97	-5.55	102.47	0.43
CF	1.21	117.30	0.95	4.05	43.53	0.17
SS	0.47	114.15	0.88	1.14	28.66	0.56
CA	-3.61	86.96	0.54	2.52	32.90	0.26
CM	-15.74	203.31	0.26	2.71	57.07	0.49
LT	-7.19	73.70	0.15	0.19	33.85	0.93
HO	0.28	191.98	0.98	2.67	61.94	0.53
GM	8.71	146.01	0.38	0.23	44.57	0.93
SC	-13.59	160.65	0.22	0.93	46.91	0.88
GT	-5.73	68.66	0.22	0.42	46.15	0.89
IJ	0.85	33.87	0.71	1.28	17.29	0.28
CUBA	-11.53	2153.56	0.93	56.18	759.81	0.28

*p is the p-value of the t-test

Acknowledgements

This paper was done when the two first authors stay at Instituto Pedro Kourí. We want to thanks our colleagues at the Unit of Epidemiological Surveillance, particularly to Migdonio Rodríguez and Irene Toledo for their comradeship and technical assistance.

References

Aguirre A. and González E. (1992). Forecast of Acute Respiratory Infections: Expected Nonepidemic Morbidity in Cuba. *Memorias do Instituto Oswaldo Cruz* **87**, 433-436.

Aguirre A. and Alonso A.M. (1993). Pronóstico de Situaciones Endémicas de Enfermedades Diarreicas Agudas en Cuba. *Revista Cubana de Medicina Tropical* **45**, 111-117.

Box G.E.P. and Jenkins G.M. (1976). *Time Series Analysis Forecasting and Control.* Holden-Day, San Francisco.

Box G.E.P. and Pierce D.A. (1970). Distribution of the residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association* **65**, 1509-1526.

Choi K. and Thacker S.B. (1981). An evaluation of influenza mortality surveillance 1962-1979. I. Time series forecast of expected pneumonia and influenza deaths. *American Journal of Epidemiology* **113**, 215-226.

Gibbons J.D. and Chakraborti S. (1992). *Nonparametric statistical inference.* Marcel Dekker, Inc., New York.

Helfentein U. (1986). Box-Jenkins modelling of some viral infectious diseases. *Statistics in Medicine* **5**, 37-47.

Ljung G.M. and Box G.E.P. (1978). On a measure of lack of fit in time series models. *Biometrika* **65**, 297-303.

Nobre F.F. (1992). Detecting abnormal patterns in public health surveillance data with a probability index function. *MEDINFO 92*, Elsevier Science Publishers, North Holland, 904-909.

Peña D. and Rodríguez J. (2002). A powerful portmanteau test of lack of fit for time series. *Journal of the American Statistical Association* **97**, 601-610.

Serfling R.E. (1963). Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public Health Reports* **78**, 494-506.

Thacker S.B. and Berkelman R.L. (1988). Public health surveillance in the United States. *Epidemiologic Reviews* **10**, 164-190.

Watier L.; Richarson S. and Hubert B. (1991). A time series construction of an alert threshold with application to *S.Bovimorbificans* in France. *Statistics in Medicine* **10**, 1493-1509.

Wei W.S. (1990). *Time Series Analysis. Univariate and Multivariate Methods.* Addison-Wesley Publishing Company, Redwood City.

Zaidi A.A., Schnell D.J. and Reynolds G.H. (1989). Time series analysis of syphilis surveillance data. *Statistics in Medicine* **8**, 353-362.

Received 7 May 2002; revised 15 October 2002

Prognozowanie oczekiwanej nieepidemicznej zachorowalności na choroby ostre

STRESZCZENIE

Skuteczne ostrzeganie epidemiologiczne wymaga, aby każdy wzrost przypadków chorobowych był wykryty tak szybko, jak to możliwe. Związane jest to z prognozowaniem opartym na modelu czasowym. W pracy prezentowana jest metoda służąca do identyfikowania i zastępowania nienormalnych wartości szeregu czasowego. Następnie stosuje się podejście oparte na modelu autoregresyjnym średniej ruchomej dla przewidywania liczby przypadków chorobowych. Opisana metoda pozwala na wykrycie początku epidemii w warunkach rutynowego procesu zbierania danych.

SŁOWA KLUCZOWE: ostrzeganie epidemiologiczne, szeregi czasowe, próg alarmowania