Modelling of an Epidemiological Time Series by a Threshold Autoregressive Model

Laurence Watier; Sylvia Richardson


Stable URL:
http://links.jstor.org/sici?sici=0039-0526%281995%2944%3A3%3C353%3AMOAETS%3E2.0.CO%3B2-1

*The Statistician* is currently published by Royal Statistical Society.

Your use of the JSTOR archive indicates your acceptance of JSTOR’s Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR’s Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/rss.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org/
Sat Aug 5 16:51:00 2006
Modelling of an epidemiological time series by a threshold autoregressive model

By LAURENCE WATIER† and SYLVIA RICHARDSON

Institut National de la Santé et de la Recherche Médicale, Villejuif, France

[Received March 1994. Final revision December 1994]

SUMMARY
In this paper we fit a self-exciting threshold autoregressive (SETAR) model, introduced by Tong, to a recent epidemiological time series of reported cases of Salmonella typhimurium in France. The procedure proposed by Tsay for fitting this class of model is briefly presented. The fitted ‘full’ model is compared with a simple autoregressive (AR) model. Finally, we compare the full model with the ‘restricted’ model discussed by Thanoon. Our results favour modelling by a SETAR process instead of an AR process. Thus, the time series of infections due to Salmonella typhimurium exhibits a type of non-linearity which can be accounted for by a threshold model. For parsimony and ease of interpretation of the model, the restricted SETAR model is finally preferred.

Keywords: Autoregressive models; Non-linear models; Salmonella; Self-exciting threshold autoregressive models; Time series

1. Introduction
Time series analyses carried out in epidemiology and public health research have been based essentially on a class of autoregressive moving average (ARMA) processes introduced by Box and Jenkins in 1970 (Box and Jenkins, 1970; Helfenstein, 1986). Among the characteristics of ARMA processes is the property that the expectation of current values of the process can be linearly expressed as a function of its past values, the so-called linearity property. This property, though natural, is nevertheless constraining as non-linearity can often be detected in the underlying structure of time series.

To generalize the class of ARMA models, several types of non-linear model have been proposed (Tong, 1990). We have focused our attention on piecewise linear models, i.e. models for which discontinuities arise, resulting in changes of regime. When the discontinuities are linked to an external process, this defines the class of threshold autoregressive (TAR) models. If the discontinuities result from internal changes, the relevant model class is the self-exciting threshold autoregressive (SETAR) process type. This class is particularly appealing in the context described here of modelling changes over time of indicators of infectious diseases. Indeed discontinuities of behaviour can be expected, due for example to changes in the population of bacteria.

SETAR models were introduced by Tong (1978) and have subsequently been discussed in for example Tong and Lim (1980) and Tong (1983). Motivated by the complex behaviour of solutions of non-linear discrete systems, Tong introduced a class of time series models which
could reproduce some of the features of these solutions, such as stable or unstable limit cycles. This wealth of behaviour, which has no equivalent in ARMA time series models, led Tong to propose a class of non-linear time series models where it is hypothesized that different autoregressive processes may operate. The change between the various autoregressions is governed by threshold values and a time lag. These models have since been reviewed by several researchers and compared with classical time series models often employing two well-known data sets in time series: Wolf's sunspot and the Canadian lynx data (Box and Jenkins, 1970; Campbell and Walker, 1977). We found only one example of fitting of SETAR models on epidemiological-type data (Cheng and Tong, 1992) and very few on other original series. This may be a result of the relative complexity of the identification and estimation procedures for these models which require specific programming.

The aim of this paper is to present the estimation of a SETAR model on a recent epidemiological time series where non-linearity is present. For model identification, we essentially follow the steps outlined by Tsay (1989) and present these briefly in our next section. Modelling of a time series of recorded cases of Salmonella typhimurium in France over a decade is then presented and the fit of a SETAR model is compared with that of a simple autoregressive model. Bacteria of genus Salmonella include many serotypes and the serotype typhimurium represents about 40% of Salmonella infections in France.

2. Methods

The class of SETAR processes is composed of models of the form

$$X_t = a_0^{(0)} + \sum_{i=1}^{p} a_i^{(0)} X_{t-i} + \varepsilon_t^{(0)} \quad \text{if } r_{j-1} \leq X_{t-d} < r_j$$

where \(j = 1, \ldots, k\), \(k\) is the number of regimes, with the regimes being separated by \(k - 1\) threshold values \(r_j (r_0 = -\infty; r_k = +\infty)\), \(d \in \mathbb{N}^+\) is the delay parameter \((d \leq p)\), \(\{a_i^{(0)}, a_i^{(j)}\}\), \(i = 1, \ldots, p\), \(j = 1, \ldots, k\), are the model parameters of regime \(j\) (constant and autoregressive parameters) and \(\{\varepsilon_t^{(0)}, \varepsilon_t^{(j)}\}_j = 1, \ldots, k\) are sequences of independent normal variables with zero mean and variance \(\sigma_j^2\).

These models allow for changes in the autoregressive coefficients over time, changes which are determined by comparing previous values (back shifted by a time lag equal to \(d\)) to fixed threshold values. Each different autoregression is referred to as a ‘regime’. In the definition above, the value \(p\) of the order of autoregression was set equal for all the regimes, but this need not be the case and the order of the autoregression can be different in each regime.

Before estimating a SETAR model, it is necessary to detect specific non-linear behaviour in the series analysed by using an appropriate test. Classical non-linearity tests based on maximum likelihood cannot be used to detect a SETAR alternative because the likelihood is not differentiable with respect to the unknown threshold values \(r_j\) (Tong, 1990). Several researchers have proposed methods for detecting this type of piecewise non-linearity (Tong and Lim, 1980; Keenan, 1985; Tsay, 1986; Petruccelli and Davies, 1986). We restrict ourselves to describing a test proposed by Tsay (1989) which is a combined version of the non-linearity tests of Keenan (1985), Tsay (1986) and Petruccelli and Davies (1986). It is fairly simple and widely applicable. Its asymptotic distribution under the linear model assumption is the usual \(F\)-distribution.

2.1. Test for non-linearity

To describe Tsay’s (1989) procedure, a SETAR(2; \(p, d\)) model is used as an example. It comprises two regimes and one threshold value \(r_1\). We assume the same autoregressive order

\(X_t = a_0^{(0)} + a_1^{(0)} X_{t-1} + a_2^{(0)} X_{t-2} + \varepsilon_t^{(0)} \quad \text{if } a_{j-1} \leq X_{t-d} < a_j\)
MODELLING EPIDEMIOLOGICAL TIME SERIES

\( p \) in each regime and a delay parameter equal to \( d \). This model can be written as

\[
X_t = \begin{cases} 
\alpha_1 + \sum_{i=1}^{p} a_{1i} X_{t-i} + \epsilon_{1}^{(1)} & \text{if } X_{t-d} \leq r_1, \\
\alpha_2 + \sum_{i=1}^{p} a_{2i} X_{t-i} + \epsilon_{2}^{(2)} & \text{if } X_{t-d} > r_1
\end{cases}
\]

where \( t \in \{p + 1, , N\} \), \( N \) being the number of observations, and the other parameters are defined as before.

Before defining the test statistic, some preliminary calculations are necessary.

2.1.1. **Arranged autoregression and predictive residuals.** First, observations are ranked in increasing order. Let \( n_i \) be the time index of the \( i \)th smallest observation; then model (1) can be written equivalently as

\[
X_{n_i+1} = \begin{cases} 
\alpha_1 + \sum_{v=1}^{p} a_{1v} X_{n_i+1-v} + \epsilon_{1}^{(1)} & \text{if } i \leq s, \\
\alpha_2 + \sum_{v=1}^{p} a_{2v} X_{n_i+1-v} + \epsilon_{2}^{(2)} & \text{if } i > s
\end{cases}
\]

with \( i \in \{p + 1 - d, , N - d\} \) and \( s \) satisfying \( X_{n_s} < r_1 \leq X_{n_s+1} \). In other words, \( s \) is the index of the nearest and smallest observation to the threshold value \( r_1 \). In this formulation, the subscripts of \( X \) and \( \epsilon \) contain the delay parameter because belonging to one or another regime is now characterized by the index \( i \) and not by \( X_{t-d} \).

This reorganization thus separates the observations into two non-overlapping groups, in such a way that, if the true model is indeed \( \text{SETAR}(2; p, d) \) process, the observations in each group follow the same autoregressive model. Under the null hypothesis we suppose in contrast that the model is an autoregressive process of order \( p \).

Tsay (1989) proposed to use the classical F-statistic corresponding to the regression of predictive residuals \( \tilde{\epsilon}_{n_i+1} \) (recursively estimated) of the arranged autoregression on the regressors \( (X_{n_i+1-v}, v=1, , p) \) for \( i=p+1-d, , N-d \). To calculate the statistic, we need to estimate recursively the model coefficients under the null hypothesis and a predictive residual at the following time index. Recursive estimations of parameters were computed by the algorithm proposed by Ertel and Fowlkes (1976). Tsay advised that the recursion be started with \( r_n = N/10 + p \) observations.

2.1.2. **Test statistic.** The F-statistic used is associated with the comparison with \( 0 \) of \( p + 1 \) estimated parameters in the least squares regression defined by

\[
\tilde{\epsilon}_{n_i+1} = w_0 + \sum_{v=1}^{p} w_v X_{n_i+1-v} + \tilde{\epsilon}_{n_i+1}, \quad \text{for } i \in \{m+1, , N-p\}
\]

where \( \tilde{\epsilon}_{n_i+1} \) corresponds to the standardized predictive residuals. Under the null hypothesis the standardized predictive residuals are white noise asymptotically and orthogonal to the regressors \( (X_{n_i+1-v}, v=1, , p) \). \( F \) is thus equal to

\[
F(p, d) = \frac{(\sum \hat{\epsilon}_i^2 - \sum \hat{\epsilon}_i^2)/(p+1)}{\sum \hat{\epsilon}_i^2/(N-2p-m-1)} \tag{3}
\]

and follows approximately an F-distribution with \( p + 1 \) and \( N - 2p - m - 1 \) degrees of freedom. Only the existence and not the values of the thresholds are required for building the test. When such non-linearity is detected by the test, the next delicate step in the identification of a SETAR model is to estimate the delay parameter and the threshold values.
2.2. Identification of delay parameter

Tong and Lim (1980) first fixed the model parameters (threshold values and autoregressive order) and next used the Akaike criterion to identify \( d \). Tsay proposed a different procedure, i.e. first to identify \( d \) and then the threshold values.

For an autoregressive process of fixed order \( p \), the delay value \( d_p \) is chosen from values \( (1, \ldots, p) \) as follows:

\[
d_p = \arg \max_{1 \leq \delta \leq p} \{ F(p, \delta) \}
\]

where \( F(p, \delta) \) is the statistic defined in equation (3). The notation \( \arg \max \) means that \( d_p \) is the value which maximizes \( F(p, \delta) \) and, as the delay parameter depends on \( p \), it is indexed by \( p \).

2.3. Identification of threshold values

Like Tong, Tsay proposed graphs, which are not formal tests, to identify threshold values. Two scatterplots are used for this purpose:

(a) the scatterplot of the standardized predictive residuals versus \( X_{t-d_p} \)—a non-random change will be observed at the threshold values, since the predictive residual will be biased at the threshold;

(b) the scatterplot of the t-ratios of recursive estimates of significant autoregressive coefficients versus \( X_{t-d_p} \)—the t-ratio starts to turn at the threshold value (in a simple autoregressive process, when an autoregressive coefficient is significant, the t-ratios for this coefficient gradually and smoothly converge to fixed values as recursive estimation on an increasing number of observations is carried out).

2.4. Model refinement in each regime

To compute the statistic \( F(p, d) \), we first chose a fixed value \( p \) for all the regimes. This restriction is now lifted and autoregressive orders are identified separately in each regime. However, regimes are defined by the choice of threshold values. As the choices of these parameters (autoregressive order and threshold values) are interrelated, an iterative identification procedure, based on the Schwarz information criterion (SIC), is used (Schwarz, 1978).

The SIC is calculated for all possible values of \( p \) and each potential threshold value. This criterion is defined as

\[
\text{SIC}(p) = N \log \left( \frac{\text{RSS}}{N} \right) + (p + 1) \log N
\]

where RSS is the residual sum of squares of the step-by-step fitted model, based on maximum likelihood estimates of the parameters, \( N \) is the 'effective number of observations' (dimension in each regime of the vector regressed to estimate the parameters) and \( p \) is the number of independent parameters of the model. The aim is to minimize this criterion. This minimization leads first to the choice of a value of \( p \) adapted to each regime and next to identify the threshold values. Tong and Lim (1980) have introduced this iterative procedure by using the Akaike information criterion. To penalize overparameterization more strongly, we prefer to use the SIC in a similar fashion.

In a SETAR(2; \( p \), \( d \)) process with threshold value \( r_1 \), the iterative steps are as follows. For fixed \( d \) and \( r_1 \), the SIC is first used to determine the autoregressive orders of the first and second regimes (\( p_1 \) and \( p_2 \) respectively). The autoregressive order for the first regime is defined as

\[
\hat{p}_1(r_1) = \arg \min_{1 \leq k_1 \leq p_1} \left[ N_1 \log \left( \frac{\text{RSS}_1(p_1)}{N_1} \right) + (p_1 + 1) \log N_1 \right]
\]

and similarly for \( \hat{p}_2(r_1) \).
Then, a threshold value is defined by using a criterion encompassing the two regimes:

$$\hat{r}_1 = \arg \min \left[ \text{SIC}\{\hat{r}_1(r_1)\} + \text{SIC}\{\hat{r}_2(r_1)\} \right].$$

(4)

3. Modelling of monthly notifications of cases of \textit{Salmonella typhimurium}

Bacteria of genus \textit{Salmonella} include more than 2200 serotypes which themselves are subdivided into phage types. Phage typing is a more discriminant method than serotyping. However, it is not in use routinely in France. These bacteria are, in the great majority, pathogenic to animals and humans. In France, 50 serotypes are responsible for 95\% of \textit{Salmonella} infections. These infections constitute, in addition to their consequences on health, a serious economic problem (Roberts, 1988; Hubert, 1992). Improvement of knowledge in the epidemiology of \textit{Salmonella} infections is an important task. It is thus necessary to survey the temporal evolution of many serotypes. In related work we have defined thresholds for alarm for infections due to two particular serotypes: \textit{Salmonella bovis morbillicans} and \textit{Salmonella newport} (Watier et al., 1991, 1994). These thresholds were based on forecasts derived from an estimated seasonal autoregressive integrated moving average (SARIMA) model for those series. We found that such models were appropriate for the underlying structure of several serotypes but not for some others, in particular \textit{Salmonella typhimurium} which is the focus of our present analysis.

The data used in this study come from the National Salmonella Reference Centre at the Pasteur Institute. They consist of monthly notifications of identified isolates of \textit{Salmonella typhimurium} which are available from January 1978. \textit{Salmonella} infections of the serotype \textit{typhimurium} are important as they represent about 40\% of all \textit{Salmonella} identified by the National Reference Centre. The period analysed includes 132 months of observations from January 1978 to December 1988. This sample length allows an accurate time series study.

As in most time series representing \textit{Salmonella} infections, the series analysed shows seasonality connected with the increased development of these bacteria during the summer (Fig. 1). However, the underlying structure is different from that of the serotypes previously analysed. Indeed, the residuals from the fit of similar SARIMA models are not white noise and the fit is not satisfactory. This major serotype represents in fact the aggregation of many different phage types, aggregation which could induce changes in the temporal structure. SETAR processes seemed appropriate for modelling the underlying structure of this serotype. Different regimes could represent sudden increases in prevalence of some phage types within the serotype. In this study, our goal is to achieve a better understanding of the underlying structure and the detection of occurrences (or not) of changes of regimes is the first step.

![Graph of identified cases]

Fig. 1. Comparison of observed values (———) and one-step-ahead forecasts (········) for the ‘full’ SETAR(2; 12, 7) model for \textit{Salmonella typhimurium} from 1978 to 1989
3.1. Detection of non-linearity

The values of Tsay’s statistic (see equation (3)) for autoregressive orders ranging from 1 to \( p \) are indicated in Table 1. We chose \( p = 13 \) because we are analysing a monthly seasonal time series and the partial autocorrelation function shows a significant parameter at lag 13. The recursive estimation starts with \( m = 26 \) observations. The total number of observations is 132 so the \( F \)-statistic follows a Fisher distribution with 14 and 79 degrees of freedom. This statistic is significant (\( P < 0.02 \)) for a delay of 7 and non-significant for all the other values.

3.2. Identification of delay and threshold values

3.2.1. Delay value. The delay value is defined as the value for which the statistic \( F \) is significant and maximum. This definition leads to choosing \( d = 7 \). To appreciate its stability, with respect to the autoregressive order, we calculated the same statistic with \( p \) equal to 11 and 12 (Table 2). The delay parameter is stable with the order of autoregression and henceforth we shall set its value to 7.

3.2.2. Threshold values \( r_j \). The identification of threshold value(s) is first appreciated with the help of scatterplots. To calculate \( F(p, d) \), 14 parameters were estimated: the constant and 13 autoregressive parameters. Residuals were deduced from these estimations. The first scatterplot used is the scatterplot of the standardized predictive residuals against \( X_{t-7} \) where \( X_t \) represents the time series analysed. The other scatterplots are the scatterplots of \( t \)-ratios of recursive estimates of significant autoregressive coefficients against \( X_{t-7} \). Four parameters were significant: the constant and the autoregressive parameters of orders 1, 12 and 13.

These five scatterplots indicate a threshold value around 75 and possibly another around 145 (Fig. 2). However, in this study, we only consider a SETAR process with two regimes because it is necessary to have about 50 observations per regime to obtain accurate parameter

<table>
<thead>
<tr>
<th>( d )</th>
<th>( F(13, d) )</th>
<th>( p )</th>
<th>( d )</th>
<th>( F(13, d) )</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.42</td>
<td>NS</td>
<td>8</td>
<td>0.75</td>
<td>NS</td>
</tr>
<tr>
<td>2</td>
<td>0.86</td>
<td>NS</td>
<td>9</td>
<td>0.74</td>
<td>NS</td>
</tr>
<tr>
<td>3</td>
<td>0.73</td>
<td>NS</td>
<td>10</td>
<td>0.87</td>
<td>NS</td>
</tr>
<tr>
<td>4</td>
<td>1.30</td>
<td>NS</td>
<td>11</td>
<td>0.98</td>
<td>NS</td>
</tr>
<tr>
<td>5</td>
<td>1.54</td>
<td>NS</td>
<td>12</td>
<td>0.79</td>
<td>NS</td>
</tr>
<tr>
<td>6</td>
<td>1.34</td>
<td>NS</td>
<td>13</td>
<td>0.30</td>
<td>NS</td>
</tr>
<tr>
<td>7</td>
<td>2.05</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( \dagger \) NS: \( P < 0.05 \).

\( P < 0.01 \).

\( § P < 0.001 \).

\( \dagger \) \( F_1 \) and \( F_2 \) follow a Fisher distribution with 12 and 85 degrees of freedom and 13 and 82 degrees of freedom respectively.

\( \dagger P < 0.05 \).

\( § P < 0.01 \).

\( §§ P < 0.001 \).
Fig. 2. Scatterplots for *Salmonella typhimurium*: (a) standardized predictive residuals versus $X_{t-7}$; (b) $t$-ratios of recursive estimates of the constant versus $X_{t-7}$; (c) $t$-ratios of recursive estimates of the autoregressive coefficient of order $1$ versus $X_{t-7}$; (d) $t$-ratios of recursive estimates of the autoregressive coefficient of order $12$ versus $X_{t-7}$; (e) $t$-ratios of recursive estimates of the autoregressive coefficients of order $13$ versus $X_{t-7}$.
estimation. On these scatterplots a threshold value around 75 is more clearly indicated than around 145, so we shall investigate a range of values around 75 for the threshold. Schwarz criteria were thus calculated with a wide range of order of autoregression (1–17) and threshold values between 70 and 80. In this interval there are seven different observations.

For the first regime, i.e., for values of \( X_t \) such that \( \{ X_{t-7} \} \) are less than the threshold value, the information criterion is a minimum for \( p = 12 \) and for all threshold values in the range 70–80. For the second regime, the results are less satisfying. In fact, the information criterion diminishes when \( p \) increases, whatever the threshold value. So the information criterion does not yield any information about \( p \). However, on the basis of our knowledge of the seasonality of the series, we decided to take \( p = 12 \) in the second regime also.

The criterion defined by equation (4) is now used to identify precisely the threshold value \( r_1 \). It was calculated with \( p = 12 \) in both regimes. A minimum value is obtained for a threshold value of 79 (Table 3). However, noting that the criterion was diminishing when \( r_1 \) was increasing, we extended our range of possible threshold values beyond 79. We obtained a minimum for the criterion when the threshold value was 81.

To summarize our identification step, we shall estimate a SETAR(2; 12, 7) model, i.e. a model involving two regimes; in both regimes an autoregressive process of order 12 is followed. The delay value is equal to 7. The first and the second regime comprise 56 and 64 observations respectively. Fig. 3 gives a visualization of the observations in the two regimes.

### 3.3. Parameter estimation and adequateness

The SETAR(2; 12, 7) process can be written as

\[
X_t = \begin{cases} 
    a_0^{(1)} + \sum_{i=1}^{12} a_i^{(1)} X_{t-i} + \epsilon_t^{(1)} & \text{if } X_{t-7} \leq 81, \\
    a_0^{(2)} + \sum_{i=1}^{12} a_i^{(2)} X_{t-i} + \epsilon_t^{(2)} & \text{if } X_{t-7} > 81.
\end{cases}
\]

### Table 3

<table>
<thead>
<tr>
<th>( r_1 )</th>
<th>70</th>
<th>73</th>
<th>74</th>
<th>75</th>
<th>77</th>
<th>78</th>
<th>79</th>
<th>80</th>
<th>81</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIC(( r_1 ))</td>
<td>853.49</td>
<td>853.31</td>
<td>852.64</td>
<td>852.25</td>
<td>833.85</td>
<td>833.33</td>
<td>831.78</td>
<td>829.35</td>
<td>830.89</td>
</tr>
</tbody>
</table>

Fig. 3. Monthly number of identified cases of *Salmonella typhimurium* from 1978 to 1989: ······, first regime; −−−−−−, second regime
In references concerned with this class of models, after identifying the delay and threshold values, most researchers estimate in each regime an autoregressive process of order \( p \) including all autoregressive parameters and retain this ‘full’ model even if some autoregressive parameters are non-significant (Tong and Lim, 1980; Tsay, 1989). Recently, Thanoon (1990) proposed a ‘restricted’ SETAR model including only the significant autoregressive parameters. Estimating such models on classical series, he showed that the restricted model fits better than the full model for the lynx data. We have thus estimated a full model, but for parsimony we have compared it with a restricted model.

Estimations of autoregressive parameters for the full model are indicated in Table 4. The autoregressive parameter of order 1 is the same for the two regimes. For the other parameters, there are some differences, notably with the seasonal parameters of order 11 and 12. We note that the standard errors of these two parameters are important. Overall many autoregressive parameters are low and statistically non-significant. For comparison, we have also estimated a full autoregressive process of order 12 on the series. It is interesting that the estimated values of the AR(12) parameters are situated either between the estimated values in each regime or near the values of one of the regimes.

To compare the fit of the SETAR and AR(12) models, we have regrouped the residuals obtained in both regimes. For the full SETAR model, the residuals are independent (Box–Ljung statistic, 27.84; 24 degrees of freedom; \( P = 0.24 \)) but their normality is rejected (Shapiro–Wilk statistic, 0.957; \( P = 0.004 \)). Nevertheless, the null hypothesis of normality is no longer rejected \( (P = 0.18) \) if we remove only one value corresponding to a bad fit of the model. For the AR(12) process, independence and normality of residuals are rejected \( (P = 0.04 \) and \( P = 0.02 \) respectively). The sum of squares of residuals for the SETAR process is clearly lower than for the AR process (66 502 against 80 769). This indicates a better fit for the SETAR process (see Fig. 1).

These remarks are in favour of modelling our series by a SETAR process. Thus, time series of infections due to *Salmonella typhimurium* show non-linearity which could be taken into account by a threshold model. For an observation \( X_t \), the model change is identified by using a delay of seven time lags, a delay value equal to approximately half the periodicity.

**TABLE 4**

Estimation of the autoregressive parameter for a full SETAR(2; 12, 7) and an AR(12) process

<table>
<thead>
<tr>
<th>( i )</th>
<th>Results for ( \hat{\alpha}<em>i^{(1)} ) ( (\hat{\sigma}</em>{\hat{\alpha}_i}) )</th>
<th>Results for ( \hat{\alpha}<em>i^{(2)} ) ( (\hat{\sigma}</em>{\hat{\alpha}_i}) )</th>
<th>Results for ( \hat{\alpha}<em>i ) ( (\hat{\sigma}</em>{\hat{\alpha}_i}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.396 (0.138)</td>
<td>0.396 (0.145)</td>
<td>0.386 (0.086)</td>
</tr>
<tr>
<td>2</td>
<td>0.188 (0.154)</td>
<td>0.265 (0.150)</td>
<td>0.130 (0.094)</td>
</tr>
<tr>
<td>3</td>
<td>0.136 (0.178)</td>
<td>0.040 (0.117)</td>
<td>0.053 (0.095)</td>
</tr>
<tr>
<td>4</td>
<td>0.320 (0.183)</td>
<td>0.105 (0.090)</td>
<td>0.196 (0.096)</td>
</tr>
<tr>
<td>5</td>
<td>0.087 (0.241)</td>
<td>0.047 (0.081)</td>
<td>0.012 (0.098)</td>
</tr>
<tr>
<td>6</td>
<td>0.170 (0.329)</td>
<td>0.086 (0.076)</td>
<td>0.117 (0.096)</td>
</tr>
<tr>
<td>7</td>
<td>0.229 (0.431)</td>
<td>0.068 (0.077)</td>
<td>0.121 (0.097)</td>
</tr>
<tr>
<td>8</td>
<td>0.121 (0.299)</td>
<td>0.020 (0.074)</td>
<td>0.000 (0.098)</td>
</tr>
<tr>
<td>9</td>
<td>0.277 (0.259)</td>
<td>0.012 (0.077)</td>
<td>0.011 (0.097)</td>
</tr>
<tr>
<td>10</td>
<td>0.036 (0.211)</td>
<td>0.060 (0.080)</td>
<td>0.015 (0.097)</td>
</tr>
<tr>
<td>11</td>
<td>0.126 (0.188)</td>
<td>0.258 (0.089)</td>
<td>0.111 (0.096)</td>
</tr>
<tr>
<td>12</td>
<td>0.371 (0.148)</td>
<td>0.204 (0.128)</td>
<td>0.404 (0.089)</td>
</tr>
</tbody>
</table>

\( \dagger P < 0.001. \)
\( \ddagger P < 0.001. \)
\( \S P = 0.08. \)
\( \S S P < 0.05. \)
A second restricted model including only significant autoregressive parameters and a constant in both regimes was also estimated. It can be written as

\[
X_t = \begin{cases} 
74.84 + 0.378X_{t-1} - 0.367X_{t-4} + 0.341X_{t-12} + \epsilon_t^{(1)} & \text{if } X_{t-7} \leq 81, \\
(20.274) & (0.097) & (0.096) & (0.087) \\
-0.392 + 0.658X_{t-1} + 0.314X_{t-11} + \epsilon_t^{(2)} & \text{if } X_{t-7} > 81. \\
(5.216) & (0.068) & (0.037)
\end{cases}
\] (5)

where \(\hat{\delta}_t^{(1)} = 32.53, \hat{\delta}_t^{(2)} = 14.66\) and standard errors are indicated in parentheses. The standard errors have clearly diminished in comparison with the full model. In both regimes, a short-term dependence and a seasonality are apparent. However, this dependence is expressed differently in each regime.

For completeness, this restricted model is compared with an autoregressive process which includes only significant autoregressive parameters. This latter model can be written as

\[
X_t = 26.373 + 0.438X_{t-1} - 0.176X_{t-4} + 0.504X_{t-12} + \epsilon_t
\] (6)

where \(\hat{\delta}_t = 26.38\) and standard errors are indicated in parentheses.

For both restricted models, the residuals are Gaussian white noise. However, the sum of squares for the SETAR model is smaller (74 058 against 85 268).

A comparison between the full and the restricted SETAR model shows a Schwarz criterion minimum for the restricted model (779 against 829) but, because of the number of important parameters for the full model (24 against 7), its sum of squares is smaller (66 502 against 74 058).

4. Discussion

To detect non-linearity in relation to piecewise processes, we have used Tsay's approach in preference to other tests which have been proposed (Petruccelli and Davies, 1986). Tsay's test is easy to compute and a simulation study has shown that its power was satisfactory. Thus, non-linearity was detected for the time series of \textit{Salmonella typhimurium} infections.

As we have shown, the steps necessary to identify a SETAR model are relatively complex, in that there is no procedure allowing the \textit{simultaneous identification} of the delay parameter, the number of regimes, the threshold values and the autoregressive order. Tong (1990) has proposed graphical tools for this identification step. In our case, the delay parameter was easy to identify with Tsay's procedure and the number of regimes was limited because of the total number of observations. In fact, it would be hazardous to try to estimate a model with much fewer than 50 observations in each regime. To identify threshold values and autoregressive order, we first used some graphs proposed by Tsay and secondly the Schwarz criterion, in the same way that the Akaike criterion was used by Tong and Lim (1980). For parsimony, the Schwarz criterion was preferred to the Akaike criterion.

In our epidemiological application, a SETAR model gives satisfactory results. For parsimony and ease of interpretation of the model, a restricted SETAR model is preferred to a full model. Moreover the fit of the restricted SETAR model is better than the fit of a simple autoregressive model. SETAR models are interesting since they give more information than classical models about the underlying structure of the time series of interest. The discontinuity that was identified may correspond to competition between phage types resulting in a sudden prevalence of some phage types within the serotype. Time changes in the occurrence of particular phage types have been observed in some \textit{Salmonella} (Lancet, 1988). In particular the increase of \textit{Salmonella enteritidis} observed in European and American countries is due to the emergence of phage 4 (Cowden \textit{et al.}, 1989; Rampling \textit{et al.}, 1989).
In our case, it is interesting to see that the fitted SETAR model given in equation (1) admits a limit cycle of period 13 close to the observed annual periodicity of the data. We also note that the global linear model (equation (2)) is not very dissimilar from the linear model for the lower regime. The lower regime corresponds overall to data observed in warm months whereas the upper regime reflects mostly the structure for the cold months. This is due first to the finding of a delay equal to approximately half the periodicity, periodicity which is linked to a seasonal phenomenon, and secondly to the threshold value which splits the data roughly in half. The observed differences between the two regimes could indicate that different phage types are prevalent in the two (warm and cold) periods. As the globally linear model and the linear model in the lower regime are similar, an alternative explanation would be an increase in prevalence of some phage types in the warm period coupled with substantial background noise in the cold period. This is in accordance with a previous study on Salmonella enteritidis infections where the notion of warm and cold periods has been introduced (Watier et al., 1994). Indeed we know that eggs are a major source of Salmonella typhimurium as well as Salmonella enteritidis infections. In Watier et al. (1994), support was found for the hypothesis that, in France, non-commerically distributed free-range eggs from smallholdings could be leading to an increase in Salmonella enteritidis. Moreover climatic conditions between November and April lead to an interruption of laying of eggs by non-industrial free-range hens. However, we have no information to confirm that hypothesis with respect to Salmonella typhimurium.

This study points to the fact that serotyping, as opposed to phage typing, gives only a broad classification which limits the understanding of the underlying epidemiological process.

Acknowledgements

The authors wish to thank P. A. D. Grimont and P. Bouvet of the Pasteur Institute of Paris for providing the data. They also wish to thank B. Hubert of the National Public Health Network of Saint Maurice for helpful discussions. This paper has benefited from constructive comments made by H. Tong.

References


