

Time series analysis as a framework for the characterization of waterborne disease outbreaks

ELENA N. NAUMOVA & ROBERT D. MORRIS

*Department of Family Medicine and Community Health, Tufts University School of Medicine,
136 Harrison Avenue, Boston, Massachusetts 02111, USA*

e-mail: enaumova@opal.tufts.edu

Abstract The time series approach is a powerful tool that has the potential to provide valuable insight into epidemic and endemic waterborne disease. Detecting waterborne disease requires a quantitative framework for the definition of an outbreak as well as analytical methods for the evaluation of an outbreak. This paper provides a model for characterizing a disease outbreak using time series methods, evaluates this model using a simulated disease outbreak, and demonstrates the method developed to analyse the relationship between time series representing water quality and time series representing disease.

INTRODUCTION

Current approaches to characterizing the incidence of waterborne disease rely heavily on the detection of water related outbreaks. One of the major difficulties in identification of waterborne disease outbreaks is the lack of a clear and consistent definition for an outbreak. One common epidemiology text defines an outbreak as an increase in the occurrence of disease in members of a defined population that is “clearly in excess of the number of cases usually or normally found in that population” (Friedman, 1993). Applying this definition in any meaningful way requires a precise definition of what is meant by “clearly in excess” and what is the usual or normal number of cases. It also requires that we identify the defined population and specify a time frame in which an outbreak might occur.

By considering the incidence of disease in the context of time series analysis, we can provide a quantitative framework for the definition of an outbreak as well as analytical methods for the evaluation of an outbreak. Time series analysis is well equipped to identify aberrations from baseline rates of disease. Furthermore, methods have been developed to analyse the relationship between time series representing water quality and time series representing disease. This paper provides a model for characterizing a disease outbreak using time series methods and evaluates this model using a simulated disease outbreak.

To describe a formal framework, three major assumptions were made:

- (a) Waterborne infections are manifested by increased disease rates in a community following increased pathogen contamination in community water sources.
- (b) An outbreak of infectious disease is an increase of disease rate above some endemic level localized in time and space. In defining an outbreak we consider the following outbreak characteristics to be essential:
 - localization in space implies a link with a water source shared by a particular community;

- localization in time means that an outbreak has an onset, magnitude and duration consistent with the incubation period for primary and secondary spread; and
 - the incubation period is the time from exposure to the onset of symptoms.
- (c) The link between drinking water contamination and disease in the community is characterized by three factors:
- the rate of infectious disease in the community;
 - the level of drinking water contamination; and
 - the time between the occurrence of contamination and the incidence of disease in a community.

In other words, a waterborne outbreak investigation attempts to assess an association between response and exposure in terms of the time sequence of events and the dose-response relationship. In the context of such a relationship it is essential to describe the probabilistic properties for model components and their measures.

Characterization of health outcomes

Disease incidence in a community can be estimated by a variety of indicators, such as laboratory confirmed cases, emergency room visits, medication use or seroprevalence. The marker of infection or health outcome measured as an observable rate of infection, Y , in a community, consists of two components: the baseline endemic rate, Y_e , and the rate of infection in a community associated with an outbreak, Y_o . An endemic rate, Y_e , is observable only in a time period with no outbreak and is assumed to be constant for a given community. The rate of infection associated with an outbreak, Y_o , is related to the incubation period. Thus, the distribution of an observable rate of infection in a community can be represented as a mixture: $Y = Y_e + Y_o$.

In a time series context, the endemic rate is measured as the number of events or counts collected in regularly spaced intervals over some period T , following a Poisson distribution with parameter λ : $Y_e \sim \text{Poisson}(\lambda)$. This time series is a stationary stochastic process having a constant mean, λ , which is equal to the variance. Assuming that this process is stable for a given community, the serial autocorrelation for such a process would be zero. The component reflecting the rate associated with an outbreak, Y_o , can be estimated using the equation: $Y_o = Y - Y_e$.

Outbreak description

Let us define the presence of an infectious disease outbreak as an increase in the daily counts of our disease indicator above some pre-specified level. Such an outbreak should have a detectable time of onset, magnitude (size) and duration consistent with the incubation period of the responsible pathogen.

In the time series context, the rate associated with outbreak would be proportional to a temporal curve of the incubation period so that $Y_o = \alpha Y_e$, where α is a coefficient of proportionality. The time series describing the incubation period can be characterized as a set of values describing the proportion of the population infected or the number of failures on a series of days. In the context of our time series model, α equals zero before an outbreak, rises to a peak in accordance with a specific incubation

period and then decays to zero. The sum of (αY_e) over the period of the increase can be considered the magnitude of the outbreak and can be estimated as the sum of daily counts over the time period representing the outbreak duration.

The timing, duration and magnitude of the outbreak depend on the pathogen, the magnitude of the exposure, the route of transmission and the sensitivity of the exposed population. For example, the length of the incubation period might be age-dependent. For waterborne infection it ranges from a few hours for Salmonella and Norwalk virus to 1–4 weeks for *Giardia*. For *Cryptosporidium* the mean incubation period is 7–8 days.

Outbreak simulation

Although the distribution of incubation periods for individuals in a population may be very complex, we limited our study to a temporal pattern with a steady increase before a single peak, followed by a slow decay. This kind of curve might be obtained using a Poisson distribution. The distribution of the magnitude of epidemic for large populations, under certain conditions, was shown to be close to a Poisson form based on an approximate formula for the total size of epidemic.

Let D denote a random variable reflecting the length of incubation period for an individual (time from day of exposure until day of onset of symptoms, or, in other words, a failure time). Let D_1, D_2, \dots, D_M be independent and identically distributed discrete failure times taking values in $M = (1, 2, \dots)$. The probability distribution of D_i can be specified by probability function, $p_d, d \in M$. The probability of disease related to an exposure event (i.e. the occurrence of a sharp increase in contaminant level), becomes non-zero immediately following the event, rises to a maximum at the mean incubation period for the pathogen and then decays. Hence, the duration of an outbreak in the general case, or $\max\{D_i\}$, would be the period from the time of exposure until the onset of symptoms for the last case of disease related to the event.

The parameter α , is the relative proportion of infected persons on a given day. Disease occurrence might follow a binomial distribution, so it is reasonable to use a Poisson distribution for defining the probability of a given number of cases for each day, $\text{prob}\{D = d\} \sim \text{Poisson}(s)$, where s is a mean day of incubation period.

DESCRIPTION OF EXPOSURE

The level of individual exposure is a function of the quantity and the quality of the water ingested (Haas *et al.*, 1993). In general, water consumption can be approximated as a log-normal distribution (Rosebury & Burmaster, 1992) but the amount of water ingested is a function of age (Shimokura *et al.*, 1998) and behavioural characteristics. Over the time period likely to be involved in an outbreak, water consumption can be assumed to be stable in a fixed age group. On the other hand, the quality of the water ingested is an important time-varying component of individual exposure. In a time series analysis context, this means that the individual exposure to a pathogen can be reasonably assumed to be proportional to drinking water quality. We assume that water quality (specific pathogen amount) can change over time and some specific water quality characteristics can serve as a surrogate for pathogen contamination. Turbidity

level in drinking water, for example, is correlated with high microbial contamination. Temporal characteristics of drinking water turbidity depend on many factors related to the characteristics of the water source and treatment plant.

Exposure simulation

Let X denote a positive real-valued random variable, representing the daily turbidity levels, recorded in nephelometric turbidity units (NTU). For simplicity, we are using a uniform distribution to simulate the time series. Zero or a lowest detectable level can be taken as a lower boundary for the distribution. The highest observable value of turbidity or regulation standard can be used for the upper boundary. It is important to note that the distribution may be sensitive to measurement precision, i.e. a smaller number of significant digits would tend to distort the distribution from uniform to polynomial. The distortion in the turbidity distribution would occur in the presence of a single spike. For treated water turbidity, a single spike would be a sudden burst of very large amplitude with low background.

MODELLING THE RELATIONSHIP BETWEEN EXPOSURE AND OUTCOME

To demonstrate the use of time series analysis we conducted two models: model 1, for endemic process and model 2, for outbreak. Model 1 reflects the endemic process over 500 days with the outcome modelled as a Poisson process (with a mean of 3 cases per day) and exposure modelled as a uniform variable distributed in the range of (0, 0.5). Model 2 reflects the endemic process with a simulated outbreak; an outcome was modelled as a mixture of a Poisson process with a mean of three cases per day, and a simulated outbreak with a mean incubation period of eight days. The temporal curve of the incubation period was simulated as a Poisson distribution with $s = 8$ days as the mean incubation period for a population of 80 infected people. The outbreak started at day 200 and continued for 15 days. Exposure was modelled as a uniform variable distributed in a range of (0, 0.5) with the simulated outbreak at day 200. The spike in water quality was modelled by adding five fixed values (0.25, 0.5, 1, 0.5, 0.25) to the simulated time series starting at day 198 for five consecutive days. Figure 1 shows the time series for the simulated outcome presented as a number of cases of waterborne infection and simulated exposure, as a characteristic of water quality.

To evaluate the relationship between exposure and outcome we used an autocorrelation function (ACF) or cross-correlation function in a multivariate case. The autocorrelation and autocovariance functions describe the serial dependence structure of a time series. The autocovariance function is estimated by summing the lagged products and dividing by the length of the series. For the ACF, all covariances are further divided by the geometric mean of the corresponding variances. We computed the estimates of the cross-correlation for a specified number of lags using standard statistical software.

In the case of the endemic process, the ACF was near zero for all lagged time points. For the simulated outbreak, the steady rise and extended decline in daily counts associated with an outbreak were subject to a serial autocorrelation. The ACF was high for five consecutive days and peaked at day eight, corresponding to the simulated temporal curve of

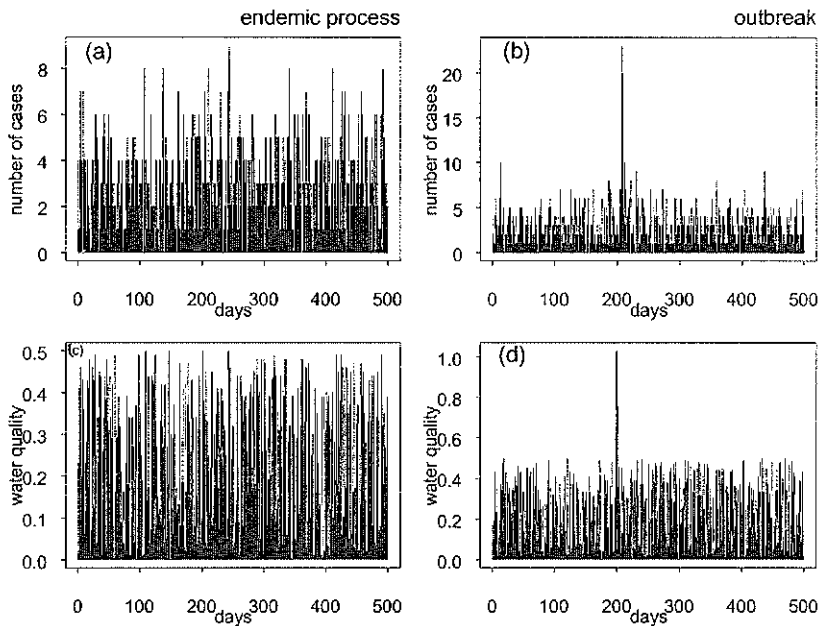


Fig. 1 Simulated time series for a given number of cases, (a) and (b), and water quality (c) and (d), for an endemic process, (a) and (c), and for an outbreak (b) and (d).

the incubation period. Using Box-Jenkins techniques for examining the time series of outcomes by testing the ACF, we found a classical moving average behaviour when the ACF is non zero for lags of the first four days, and zero thereafter (Shumway, 1988).

To describe the lagged relationship between outcome and exposure in more detail we applied a newly developed model for constructing the temporal exposure response surfaces (TERS). This TERS model reflects both the underlying tendency for the exposure to cause disease outbreak and the detection of the lags, which has a most significant influence on disease rate. The creation of lagged predicting surfaces for time series was based on a generalized additive model (Hastie & Tibshirani, 1990) and is described in detail elsewhere (Naumova & Morris, 1995). In Fig. 2 the TERS are shown: for the simulated endemic process and the simulated outbreak. The TERS for the endemic process is flat, reflecting stability over time and no relationship between water quality and disease rate. The TERS for the outbreak has an increase in disease rate associated with an increase in turbidity at a lag of 7 to 9 days. These surfaces provide more precise estimates of the hazard of drinking water contamination and can give some insight into the possible identity of the pathogen (Morris *et al.*, 1996, 1998).

CONCLUSIONS

If one accepts our initial assumptions about the nature of waterborne disease outbreaks, then time series analysis is the most appropriate framework for their quantitative evaluation. Time series methods provide a set of powerful tools that have the potential to provide valuable insight into epidemic and endemic waterborne disease.

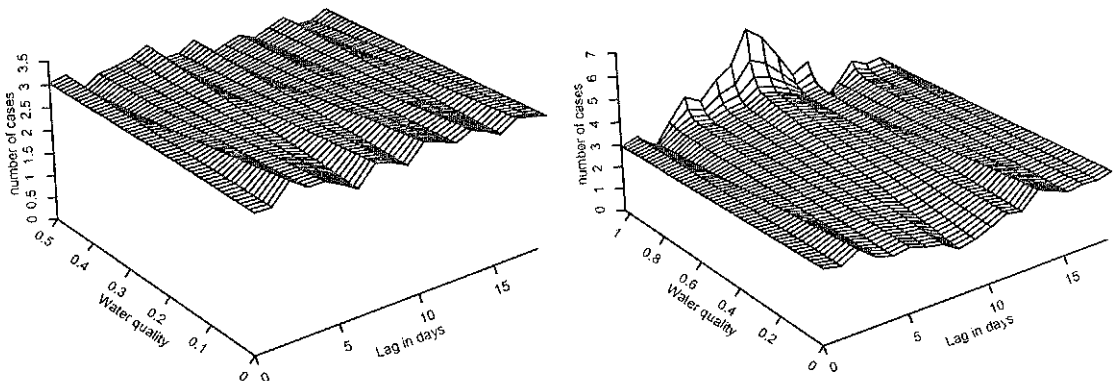


Fig. 2 The temporal exposure response surface describes the relationship between water quality (x-axis) and number of cases of infectious disease (z-axis) over a time period of 20 days for an endemic process (left) and a simulated outbreak (right).

Detecting waterborne disease requires the analysis of some form of surveillance data that describes the incidence of disease in a community over time, in other words, a time series. Determining that an outbreak has occurred requires detection of an aberration in this time series. The definition of an outbreak should, therefore, be based on the characterization of events in a time series.

We recommend that an outbreak be defined based on the distribution of the time series data. To describe the distribution of daily counts associated with an outbreak, we can use a frequency of frequencies distribution which ignores the temporal sequence. The presence of an outbreak means departure from the initial Poisson process leading to overdispersion relative to the Poisson model. The degree of departure reflects the magnitude and duration of an outbreak.

Once we have identified an outbreak, the TERS model provides a tool that has been specifically designed to evaluate water borne disease outbreaks. The results of this model can help to confirm that an outbreak has occurred and can give some insight into the possible identity of the pathogen.

REFERENCES

- Friedman, G. D. (1993) *Primer of Epidemiology*. McGraw Hill, New York.
- Haas, C. N., Rose, J. B., Gerba, C. & Regli, S. (1993) Risk assessment of virus in drinking water. *Risk Analysis* 13(5), 545–552.
- Hastie, T. J. & Tibshirani, R. J. (1990) *Generalized Additive Models*. Chapman and Hall, London, UK.
- Morris, R. D., Naumova, E. N., Levin, R. & Munasinghe, R. L. (1996) Temporal variation in drinking water turbidity and physician diagnosed gastrointestinal infections in Milwaukee. *Am. J. Public Health* 86, 237–239.
- Morris, R. D., Naumova, E. N. & Griffiths, J. K. (1998) Did Milwaukee experience waterborne cryptosporidiosis before the large documented outbreak in 1993? *Epidemiology* 9(3), 264–270.
- Naumova, E. N. & Morris, R. D. (1995) NAARX and GAM predicted surfaces in studying association between drinking water turbidity and gastrointestinal infections. In: *Proc. Joint Statistical Meetings: Epidemiology Section*, 141–146.
- Rosebury, A. M. & Burmaster, D. E. (1992) Lognormal distribution for water intake by children and adults. *Risk Analysis* 12, 99.
- Shimokura, G. H., Savitz, D. A. & Symanski, E. (1998) Assessment of water use for estimating exposure to tap water contaminants. *Environ. Health Perspect.* 106, 55–59.
- Shumway, R. H. (1988) *Applied Statistical Analysis*. Prentice Hall, Englewood Cliffs, New Jersey, USA.