
Seasonality Assessment for Biosurveillance Systems

Elena N. Naumova and Ian B. MacNeill

Tufts University School of Medicine, Boston, MA, USA
University of Western Ontario, London, ON, Canada

Abstract: Biosurveillance systems for infectious diseases typically deal with nonlinear time series. This nonlinearity is due to the non-Gaussian and non-stationary nature of an outcome process. Infectious diseases (ID), waterborne and foodborne enteric infections in particular, are typically characterized by a sequence of sudden outbreaks, which are often followed by long low endemic levels. Multiple outbreaks occurring within a relatively short time interval form a seasonal pattern typical for a specific pathogen in a given population. Seasonal variability in the probability of exposure combined with a partial immunity to a pathogen adds to the complexity of seasonal patterns. Although seasonal variation is a well-known phenomenon in the epidemiology of enteric infections, simple analytical tools for examination, evaluation, and comparison of seasonal patterns are limited. This obstacle also limits analysis of factors associated with seasonal variations. The objectives of this paper are to outline the notion of seasonality, to define characteristics of seasonality, and to demonstrate tools for assessing seasonal patterns and the effects of environmental factors on such patterns. To demonstrate these techniques, we conducted a comparative study of seasonality in *Salmonella* cases as reported by the state surveillance system in relation to seasonality in ambient temperature, and found that the incidence in *Salmonella* infection peaked two weeks after a peak in temperature. The results suggest that ambient temperature can be a potential predictor of *Salmonella* infections at a seasonal scale.

Keywords and phrases: Seasonality, δ -method, ambient temperature, *Salmonella* infection, biosurveillance

28.1 Introduction

We define “disease seasonality” as systematic periodic fluctuations within the course of a year that can be characterized by the magnitude, timing, and duration of a seasonal increase. Variations in seasonal characteristics in temporal, spatial, or demographic contexts provide important clues to factors influencing disease occurrence. We consider stability in seasonality, expressed by some measure of variation in the above-mentioned characteristics of a seasonal pattern, as an indicator of synchronization in disease incidence by environmental and/or social processes. Meteorological factors, and ambient temperature in particular, appear to be critically linked to seasonal patterns of disease. Recent studies indicate that meteorological disturbances may influence the emergence and proliferation of water- or foodborne pathogens. It is quite plausible that seasonal fluctuation in ambient temperature might affect the timing and intensity of infectious outbreaks. Therefore, we examined seasonal patterns in both infections and temperature time series and then compared characteristics of seasonality.

28.1.1 Conceptual framework for seasonality assessment

This synchronization in disease incidence and environmental factors can be viewed as a special case when multiple time series exhibit common periodicities [MacNeill (1977)]. The conceptual format for measuring the temporal relation between seasonal patterns in environmental temperatures and disease incidence is shown in Figure 28.1. Considering two characteristics of seasonality: the magnitude and the timing of a seasonal peak, we define a set of measures. The measures related to timing are (1) the position of the maximum point on the seasonal curve of exposure (i.e., temperature) or disease incidence, (2) the position of the minimum point on the seasonal curve of exposure or disease incidence, and (3) the lag, which is the difference between time of exposure maximum and time of disease incidence maximum. The magnitude related measures are (1) maximum value on the seasonal curve of exposure or incidence of disease, (2) minimum value on the seasonal curve of exposure or incidence of disease, (3) the amplitude, which is the difference between maximum and minimum values on the seasonal curve for exposure or incidence of disease, and (4) the relative intensity, which is the ratio of maximum value and minimum value on the seasonal curve. Thus, the task of measuring the temporal relation between seasonal patterns is translated to the problem of estimating the lag and associations among measures of timing and intensity.

This concept is easy to express via Model 1 as follows.

$$Y(t) = \gamma \cos(2\pi\omega t + \psi) + e(t), \quad (28.1)$$

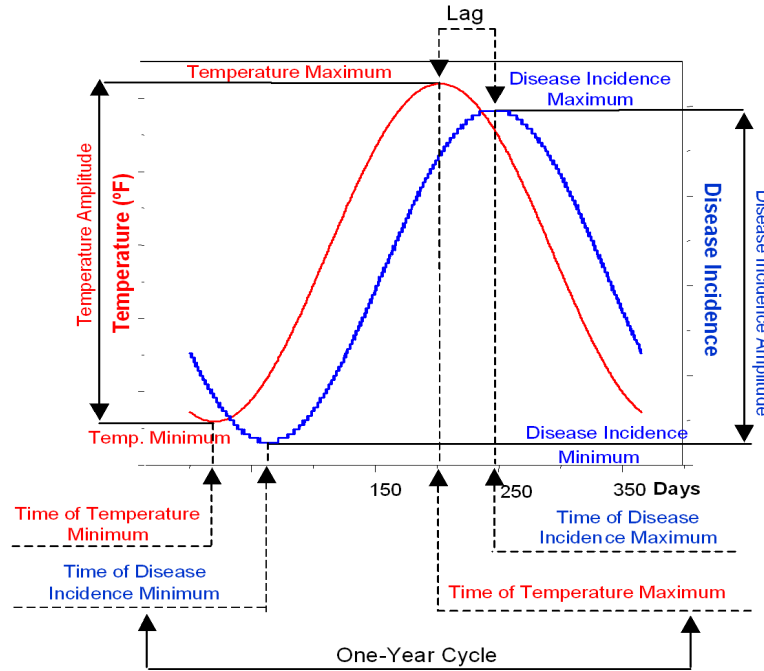


Figure 28.1: Characteristics of seasonality: Graphical depiction and definition for daily time series of exposure (ambient temperature) and outcome (disease incidence) variables

where $Y(t)$ is a time series, the periodic component has a frequency of ω , an amplitude of γ , and a phase angle of ψ , and $\{e(t), t = 1, 2, \dots, n\}$ is an i.i.d. sequence of random variables with $E[e(t)] = 0$ and $\text{Var}[e(t)] = \sigma^2$. From a user standpoint, this model offers the highly desirable property of being easy to interpret. The model describes a seasonal curve by a cosine function with symmetric rise and fall over a period of a full year. The locations of two points at which this seasonal curve peaks and has the lowest value can be determined using a shift, or phase angle parameter, ψ . This parameter reflects the timing of the peak relative to the origin. For convenience, an origin can be set at the beginning of a calendar year, January 1. So, if $\psi = 0$, there is no shift of a peak relative to the origin. If $\psi = \pi$, the peak shifts to the summer, that is, to the 182nd day. If $\pi < \psi < 2\pi$, there is a shift toward fall; or if $\psi < \pi$, there is a shift toward spring. The parameter can be used for seasonality comparison and can be expressed in days. The amplitude of fluctuations between two extreme points is controlled via a parameter γ ; if $\gamma = 0$, there is no seasonal increase.

This Model 1 is equivalent to Model 2:

$$Y(t) = \beta_1 \sin(2\pi\omega t) + \beta_2 \cos(2\pi\omega t) + e(t), \tag{28.2}$$

which is more convenient to fit by least squares, a procedure available in much commercial statistical software. Below we demonstrate an approach that allows us to combine the ease of fitting Model 2 and the simplicity and elegance of interpretation of Model 1, by using the δ -method.

28.2 δ -Method in Application to a Seasonality Model

This methodology, whose origin is remote, enables one to obtain a workable approximation to the mean, variances, and covariances of a function of random variables whose means and variances are either known or for which there exist consistent estimators.

28.2.1 Single-variable case

Let X be a random variable with $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$. Also let X_1, X_2, \dots, X_n , be a sequence of i.i.d. random variables each with the same distribution as X . If $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$, then $E(\bar{X}_n) = \mu$ and $\sigma_{\bar{X}_n}^2 = \sigma^2/n$. Hence, $\bar{X}_n \rightarrow \mu$ as $n \rightarrow \infty$, where convergence is in each of: probability, a.e., and mean square. Now let $f(\cdot)$ be a function of one variable that may be expanded in Taylor's series; that is,

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{(x - x_0)^2}{2!}f''(x_0) + \dots$$

In consequence of the above,

$$Y_n = f(\bar{X}_n) = f(\mu) + (\bar{X}_n - \mu)f'(\mu) + O\left(\frac{1}{n}\right).$$

Because $\bar{X}_n \rightarrow \mu$, $f(\bar{X}_n) \rightarrow f(\mu)$, and $f'(\bar{X}_n) \rightarrow f'(\mu)$,

$$\begin{aligned} f(\bar{X}_n) - f(\mu) &= (\bar{X}_n - \mu) \left[f'(\bar{X}_n) - \{f'(\bar{X}_n) - f'(\mu)\} \right] + O\left(\frac{1}{n}\right) \\ &= (\bar{X}_n - \mu) f'(\bar{X}_n) - (\bar{X}_n - \mu) \{f'(\bar{X}_n) - f'(\mu)\} + O\left(\frac{1}{n}\right) \\ &= (\bar{X}_n - \mu) f'(\bar{X}_n) + O\left(\frac{1}{n}\right). \end{aligned}$$

Therefore, $E[Y_n - f(\mu)]^2 = E[\bar{X}_n - \mu]^2 \{f'(\bar{X}_n)\}^2 + O\left(\frac{1}{n^{1/2}}\right)$. That is

$$\sigma_{Y_n}^2 = \sigma_{\bar{X}_n}^2 \{f'(\bar{X}_n)\}^2 + O\left(\frac{1}{n^{1/2}}\right),$$

where σ^2 may be estimated consistently by $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (X_i - \hat{X}_n)^2$. Thus, for large samples: $E[Y_n] \cong f'(\bar{X}_n)$ and $\sigma_{Y_n}^2 \cong \{f'(\bar{X}_n)\}^2 (\hat{\sigma}_{X_n}^2/n)$.

28.2.2 Two-variables case

Let $f(\cdot, \cdot)$ be a function of two random variables X and Y , whose means (μ_x and μ_y), variances (σ_x^2 and σ_y^2), and covariance (σ_{xy}) are known. Consider the first few terms in the Taylor's series expansion of $f(\cdot, \cdot)$:

$$\begin{aligned} f(X, Y) &= f(\mu_x, \mu_y) + (X - \mu_x) \left. \frac{\partial f}{\partial X} \right|_{\mu_x, \mu_y} + (Y - \mu_y) \left. \frac{\partial f}{\partial Y} \right|_{\mu_x, \mu_y} \\ &\quad + \frac{1}{2}(X - \mu_x)^2 \left. \frac{\partial^2 f}{\partial X^2} \right|_{\mu_x, \mu_y} + \frac{1}{2}(Y - \mu_y)^2 \left. \frac{\partial^2 f}{\partial Y^2} \right|_{\mu_x, \mu_y} \\ &\quad + (X - \mu_x)(Y - \mu_y) \left. \frac{\partial^2 f}{\partial X \partial Y} \right|_{\mu_x, \mu_y} + \dots \end{aligned}$$

Then, by neglecting higher-order terms in the expansion, one can obtain a large sample approximation to $E[f(X, Y)]$ as follows.

$$E[f(X, Y)] \cong f(\mu_x, \mu_y) + \frac{\sigma_x^2}{2} \left. \frac{\partial^2 f}{\partial X^2} \right|_{\mu_x, \mu_y} + \frac{\sigma_y^2}{2} \left. \frac{\partial^2 f}{\partial Y^2} \right|_{\mu_x, \mu_y} + \sigma_{xy} \left. \frac{\partial^2 f}{\partial X \partial Y} \right|_{\mu_x, \mu_y}.$$

Similarly, a large sample approximation to the variance of $f(X, Y)$ can be obtained as follows.

$$\text{Var}[f(X, Y)] \cong \sigma_x^2 \left(\left. \frac{\partial f}{\partial X} \right|_{\mu_x, \mu_y} \right)^2 + \sigma_y^2 \left(\left. \frac{\partial f}{\partial Y} \right|_{\mu_x, \mu_y} \right)^2 + 2\sigma_{xy} \left(\left. \frac{\partial f}{\partial X} \right|_{\mu_x, \mu_y} \left. \frac{\partial f}{\partial Y} \right|_{\mu_x, \mu_y} \right).$$

Now we consider large sample results. Let X_{1n} and X_{2n} be two sequences of random variables with $\text{Var}(X_{1n}) = n^{-1}\sigma_{11}$, $\text{Var}(X_{2n}) = n^{-1}\sigma_{22}$, and $\text{Cov}(X_{1n}, X_{2n}) = n^{-1}\sigma_{12}$. The estimators for the parameters are denoted by $\hat{\mu}_{jn}$ and $\hat{\sigma}_{jkn}$, $j, k = 1, 2$. These estimators are assumed to be consistent. Then

$$\begin{aligned} Y_n &= f(\hat{\mu}_{1n}, \hat{\mu}_{2n}) + (\hat{\mu}_{1n} - \mu_1) \left. \frac{\partial f}{\partial \mu_1} \right|_{\hat{\mu}_{1n}, \hat{\mu}_{2n}} + (\hat{\mu}_{2n} - \mu_2) \left. \frac{\partial f}{\partial \mu_2} \right|_{\hat{\mu}_{1n}, \hat{\mu}_{2n}} \\ &\quad + \frac{1}{2}(\hat{\mu}_{1n} - \mu_1)^2 \left. \frac{\partial^2 f}{\partial \mu_1^2} \right|_{\hat{\mu}_{1n}, \hat{\mu}_{2n}} + \frac{1}{2}(\hat{\mu}_{2n} - \mu_2)^2 \left. \frac{\partial^2 f}{\partial \mu_2^2} \right|_{\hat{\mu}_{1n}, \hat{\mu}_{2n}} \\ &\quad + (\hat{\mu}_{1n} - \mu_1)(\hat{\mu}_{2n} - \mu_2) \left. \frac{\partial^2 f}{\partial \mu_1 \partial \mu_2} \right|_{\hat{\mu}_{1n}, \hat{\mu}_{2n}} + \dots \end{aligned}$$

Because $E(\hat{\mu}_{jn} - \mu_j) \rightarrow 0$, $j = 1, 2$, and $\hat{\sigma}_{jkn} = (1/n)\hat{\sigma}_{jk}$ $j, k = 1, 2$, $E[Y_n] \cong f(\hat{\mu}_{1n}, \hat{\mu}_{2n})$, then

$$\begin{aligned} \sigma_{Y_n}^2 &\cong \hat{\sigma}_{11n} \left(\frac{\partial f}{\partial \mu_1} \Big|_{\hat{\mu}_{1n}, \hat{\mu}_{2n}} \right)^2 + \hat{\sigma}_{22n} \left(\frac{\partial f}{\partial \mu_2} \Big|_{\hat{\mu}_{1n}, \hat{\mu}_{2n}} \right)^2 \\ &\quad + 2\hat{\sigma}_{12n} \left(\frac{\partial f}{\partial \mu_1} \Big|_{\hat{\mu}_{1n}, \hat{\mu}_{2n}} \right) \left(\frac{\partial f}{\partial \mu_2} \Big|_{\hat{\mu}_{1n}, \hat{\mu}_{2n}} \right). \end{aligned}$$

More generally, it can be shown that

$$\begin{aligned} E[f(X_1, X_2, \dots, X_k)] \\ \cong f(\mu_{x_1}, \mu_{x_2}, \dots, \mu_{x_k}) + \frac{1}{2} \sum_{i=1}^k \sum_{j=1}^k \sigma_{x_i x_j} \frac{\partial^2 f}{\partial X_i \partial X_j} \Big|_{\mu_{x_1}, \mu_{x_2}, \dots, \mu_{x_k}} \end{aligned}$$

and

$$\text{Var}[f(X_1, X_2, \dots, X_k)] \cong \sum_{i=1}^k \sum_{j=1}^k \sigma_{x_i x_j} \left(\frac{\partial f}{\partial X_i} \frac{\partial f}{\partial X_j} \Big|_{\mu_{x_1}, \mu_{x_2}, \dots, \mu_{x_k}} \right).$$

28.2.3 Application to a seasonality model

Now by using the δ -method we demonstrate how we can relate parameters of two models: Model 1 and Model 2. Consider a time series $\{Y(t), t = 1, 2, \dots, n\}$ and the traditional model for seasonality, Model 2:

$$Y(t) = \beta_1 \sin(2\pi\omega t) + \beta_2 \cos(2\pi\omega t) + e(t),$$

where $\{e(t), t = 1, 2, \dots, n\}$ is an i.i.d. sequence of random variables with $E[e(t)] = 0$ and $\text{Var}[e(t)] = \sigma^2$. Note that $\cos(2\pi\omega t + \psi) = \cos(2\pi\omega t) \cos(\psi) - \sin(2\pi\omega t) \sin(\psi)$.

If $\beta_1 = -\gamma \sin \psi$ and $\beta_2 = \gamma \cos \psi$, then $\gamma \cos(2\pi\omega t + \psi) = \beta_1 \sin(2\pi\omega t) + \beta_2 \cos(2\pi\omega t)$.

Also, $\beta_1/\beta_2 = -\sin(\psi)/\cos(\psi) = -\tan(\psi)$ and $\gamma^2 = \beta_1^2 + \beta_2^2$. Therefore, $\gamma = a(\beta_1^2 + \beta_2^2)^{1/2}$, where $a = -1$ when $\beta_2 < 0$ and $a = 1$ otherwise; and $\psi = -\arctan(\beta_1/\beta_2)$ with $\frac{\pi}{2} < \psi < \frac{3\pi}{2}$. It may be noted that a change of sign of gamma results in a phase shift of $\pm\pi$. Also,

$$\begin{aligned} \frac{\partial \gamma}{\partial \beta_1} &= a\beta_1/(\beta_1^2 + \beta_2^2)^{1/2}, & \frac{\partial \gamma}{\partial \beta_2} &= a\beta_2/(\beta_1^2 + \beta_2^2)^{1/2}, \\ \frac{\partial \psi}{\partial \beta_1} &= -\beta_2/(\beta_1^2 + \beta_2^2), & \frac{\partial \psi}{\partial \beta_2} &= \beta_1/(\beta_1^2 + \beta_2^2). \end{aligned}$$

To fit the Model 2 by OLS, we let $Y' = \{Y(1), Y(2), \dots, Y(n)\}$ be the vector of observations; $e' = \{e(1), e(2), \dots, e(n)\}$ be the vector of noise variables; $\beta' =$

$\{\beta_1, \beta_2\}$ be the vector of parameters; and X be the design matrix where $X_{k1} = \sin(\omega k)$ and $X_{k2} = \cos(\omega k)$, $k = 1, 2, \dots, n$. Then for $Y = X\beta + e$, the least squares estimators are $\hat{\beta} = (X'X)^{-1}X'Y$, and the variance-covariance matrix is $\text{Cov}(\hat{\beta}) = \sigma^2(X'X)^{-1}$. If $\hat{Y} = X\hat{\beta}$, then we denote the vector of residuals by $\hat{e} = Y - \hat{Y}$. The unknown variance may be consistently estimated by $\hat{\sigma}^2 = n^{-1}e'e$. Because OLS estimators of the parameters of Model 2 are consistent, we have consistent estimators for β_1 , β_2 , and σ^2 .

Models 1 and 2 are equivalent, therefore we can fit Model 2 to obtain the estimates for the amplitude and phase parameters by applying the δ -method. Thus, we have $E[\hat{\beta}'] = E[(\hat{\beta}_1, \hat{\beta}_2)] \rightarrow (\beta_1, \beta_2) = \beta'$ and

$$\text{Cov}(\hat{\beta}) = \sigma^2(X'X)^{-1} = \begin{pmatrix} \hat{\sigma}_{\beta_1}^2 & \hat{\sigma}_{\beta_1\beta_2} \\ \hat{\sigma}_{\beta_1\beta_2} & \hat{\sigma}_{\beta_2}^2 \end{pmatrix} \approx \sigma^2(X'X)^{-1}.$$

For the amplitude $\gamma = f(\beta_1, \beta_2) = (\beta_1^2 + \beta_2^2)^{1/2}$, the estimates are

$$\hat{\gamma} = f(\hat{\beta}_1, \hat{\beta}_2) = (\hat{\beta}_1^2 + \hat{\beta}_2^2)^{1/2}$$

and

$$\text{Var}(\hat{\gamma}) = (\hat{\sigma}_{\beta_1}^2 \hat{\beta}_1^2 + \hat{\sigma}_{\beta_2}^2 \hat{\beta}_2^2 + 2\hat{\sigma}_{\beta_1\beta_2} \hat{\beta}_1 \hat{\beta}_2) / (\hat{\beta}_1^2 + \hat{\beta}_2^2).$$

The phase angle estimate is $\hat{\psi} = -\arctan(\hat{\beta}_1/\hat{\beta}_2)$ and corresponding variance estimate is

$$\text{Var}(\hat{\psi}) = (\hat{\sigma}_{\beta_1}^2 \hat{\beta}_2^2 + \hat{\sigma}_{\beta_2}^2 \hat{\beta}_1^2 - 2\hat{\sigma}_{\beta_1\beta_2} \hat{\beta}_1 \hat{\beta}_2) / (\hat{\beta}_1^2 + \hat{\beta}_2^2)^2.$$

28.2.4 Potential model extension

Model 2 can be extended to a more general Model 3 for seasonality as

$$Y(t) = \beta_0 + \beta_1 \sin(2\pi\omega t) + \beta_2 \cos(2\pi\omega t) + \beta_3 \sin(4\pi\omega t) + \beta_4 \cos(4\pi\omega t) + e(t), \quad (28.3)$$

where ω is the frequency of the seasonal component and 2ω is its first harmonic. The vector form of the model uses the following notation.

$$Y' = \{Y(1), Y(2), \dots, Y(n)\};$$

$$e' = \{e(1), e(2), \dots, e(n)\};$$

$$\beta' = \{\beta_0, \beta_1, \beta_2, \beta_3, \beta_4\};$$

and the $(n \times 5)$ design matrix X has as k th row

$$\{1, \sin(2\pi\omega t), \cos(2\pi\omega t), \sin(4\pi\omega t), \cos(4\pi\omega t)\}.$$

Model 3 can now be rewritten as described above with an alternative Model 4 as

$$Y(t) = \beta_0 + \gamma_1 \cos(2\pi\omega t + \psi_1) + \gamma_2 \cos(4\pi\omega t + \psi_2) + e(t). \quad (28.4)$$

The relations between the parameters of Models 3 and 4 are as follows: $\gamma_1^2 = \beta_1^2 + \beta_2^2$; $\psi_1 = -\arctan(\beta_1/\beta_2)$; $\gamma_2^2 = \beta_3^2 + \beta_4^2$; $\psi_2 = -\arctan(\beta_3/\beta_4)$ and $\beta_1 = -\gamma_1 \sin(\psi_1)$; $\beta_2 = \gamma_1 \cos(\psi_1)$; $\beta_3 = -\gamma_2 \sin(\psi_2)$; $\beta_4 = \gamma_2 \cos(\psi_2)$. The estimation for β_0 is the same for each of Models 3 and 4. The partial derivatives are those given in Models 1 and 2, but it should be noted that many of the derivatives are zero; that is,

$$\begin{aligned} \frac{\partial \gamma_1}{\partial \beta_0} = \frac{\partial \gamma_1}{\partial \beta_3} = \frac{\partial \gamma_1}{\partial \beta_4} = \frac{\partial \gamma_2}{\partial \beta_0} = \frac{\partial \gamma_2}{\partial \beta_3} = \frac{\partial \gamma_2}{\partial \beta_4} = 0, \\ \frac{\partial \psi_1}{\partial \beta_0} = \frac{\partial \psi_1}{\partial \beta_3} = \frac{\partial \psi_1}{\partial \beta_4} = \frac{\partial \psi_2}{\partial \beta_0} = \frac{\partial \psi_2}{\partial \beta_3} = \frac{\partial \psi_2}{\partial \beta_4} = 0, \end{aligned}$$

and

$$\frac{\partial \beta_0}{\partial \beta_i} = 0, \quad i = 1, \dots, 4; \quad \text{also} \quad \frac{\partial \beta_0}{\partial \beta_0} = 1.$$

This simplifies the computation of the standard error estimates for $(\beta_0, \gamma_1, \psi_1, \gamma_2, \psi_2)$. Thus, we have the following estimates for Model 4 based on the fit for Model 3:

$$\begin{aligned} \hat{\beta}_0 &= \hat{\beta}_0; & \hat{\gamma}_1 &= (\hat{\beta}_1^2 + \hat{\beta}_2^2)^{1/2}; & \hat{\gamma}_2 &= (\hat{\beta}_3^2 + \hat{\beta}_4^2)^{1/2}; \\ \hat{\psi}_1 &= -\arctan(\hat{\beta}_1/\hat{\beta}_2); & \hat{\psi}_2 &= -\arctan(\hat{\beta}_3/\hat{\beta}_4). \end{aligned}$$

To obtain the estimates of the standard errors of the parameters $\Omega = (\beta_0, \gamma_1, \psi_1, \gamma_2, \psi_2)$, we denote the matrix of partial derivatives of the Model 4 parameters with respect to Model 3 by $\left. \frac{\partial \Omega}{\partial \beta} \right|_{\beta}$. Then

$$\text{Cov}(\hat{\Omega}) \cong \left(\left. \frac{\partial \Omega}{\partial \beta} \right|_{\beta} \right)' (\hat{\sigma}^2 (X'X)^{-1}) \left(\left. \frac{\partial \Omega}{\partial \beta} \right|_{\beta} \right).$$

28.2.5 Additional considerations

Here we intend to apply the proposed method for assessing seasonality in infectious diseases that typically have one annual peak and are measured by counting cases occurring over prespecified time periods (e.g., days, weeks, months). Therefore, two main aspects of the method of implementation should be discussed: the underlying distribution of the case counting process and the orthogonality of the design matrix.

It is plausible to assume a Poisson process for a rare event such as a case of infection in a large closed population that satisfies a requirement of non-negativity in a time series of counts. Suppose the mean-value function for a Poisson process follows Model 2 or 3. Then the process will have as its t th component a Poisson variate with parameter $\lambda(t)$. Unless $\lambda(t) = \lambda$ for all t , the process will not have constant variance. In a case of nonconstant variance, the OLS parameters will be biased. To reduce this bias, we will use an iterative weighted least squares approach implemented via standard statistical software for a generalized Poisson regression.

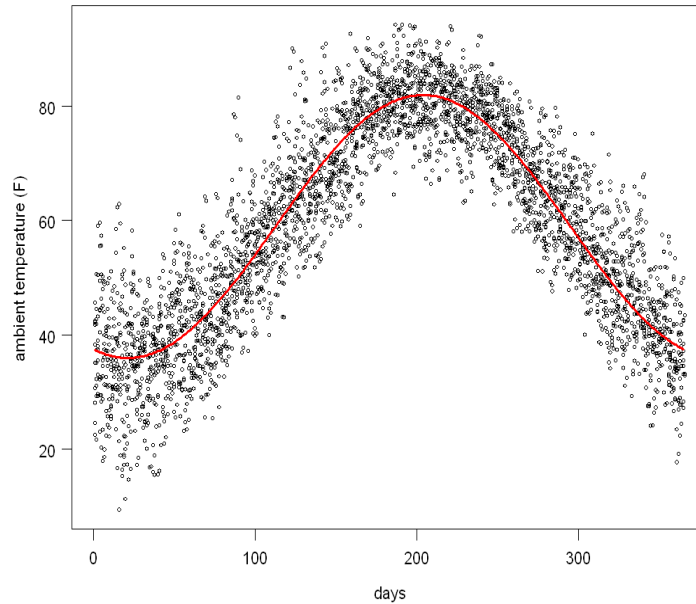


Figure 28.2: Seasonal curve for ambient temperature in temperate climate of Massachusetts, USA. Solid line is the fitted mean-value function

When fitting a trigonometric polynomial to a set of data, it is helpful if the columns of the design matrix X are orthogonal. Suppose that a time period consists of n equal subintervals of length $1/n$, and the data are collected at the end of each time subinterval. For a one-year study period, a year is a time period of one unit in length divided into 365 subunits, that is, days, each of $1/365$ th of the unit. For the proposed Model 1, the frequency ω equals 1, meaning that in one full cycle per unit of time (year) the first harmonic has twice this frequency, or there are two full cycles per year. Then the design matrix X will have columns that are orthogonal and $X'X$ will be a diagonal matrix.

28.3 Application to Temperature and Infection Incidence Analysis

To examine the relations between seasonal patterns in ambient temperature and disease incidence, we study time series of daily mean temperature and counts of *Salmonella* that have been established over the last decade in Massachusetts, USA. Ten years of temperature daily observations and *Salmonella* counts, superimposed for ease of seasonality visualization are shown in Figures 28.2 and 28.3. Figure 28.4 demonstrates daily counts of *Salmonella* with respect to the corresponding temperature values. As we can see, there are apparent increases in *Salmonella* cases in warm summer months, as well as respective increases in variability in these time periods. A daily rate of *Salmonella* is 0.78 cases per 1,000,000 population.

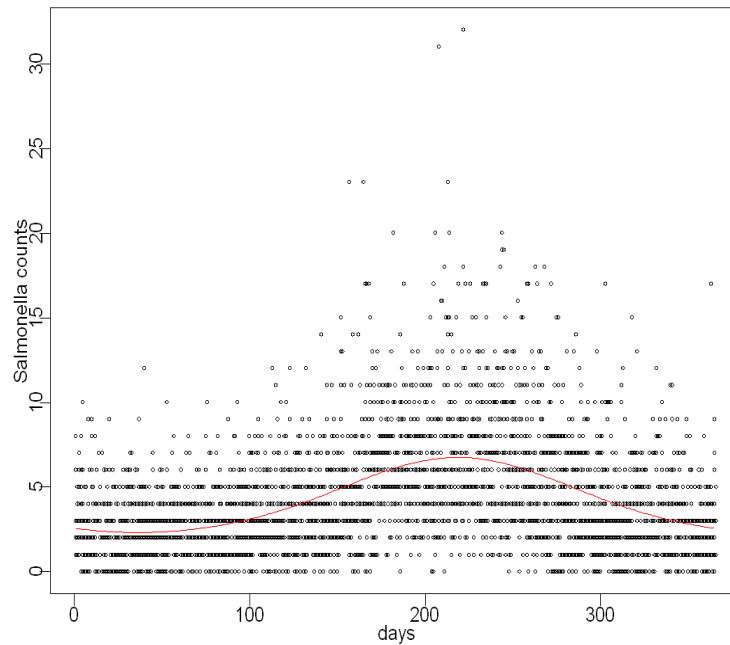


Figure 28.3: Seasonal curve for salmonella cases in Massachusetts, USA. Solid line is the fitted mean-value function

Our first step was to describe the seasonal pattern in temperature and infections over the last decade. We used a generalized linear model (GLM) with a Gaussian distribution for the outcome when the variable of interest is ambient

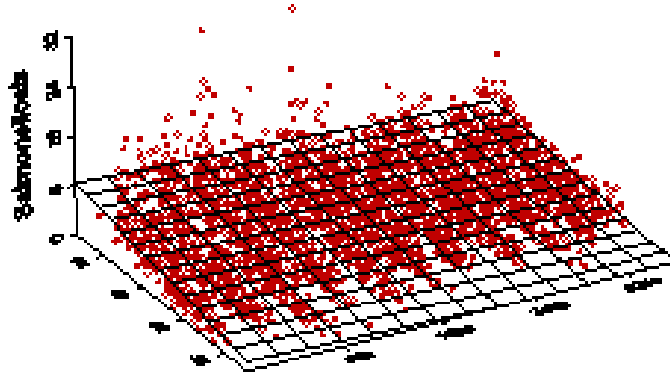


Figure 28.4: Temporal pattern in daily *Salmonella* cases (Z -axis) with respect to ambient temperature values in $^{\circ}\text{C}$ (Y -axis) over time (X -axis)

temperature,

$$Y(t) = \beta_0 + \beta_1 \sin(2\pi\omega t) + \beta_2 \cos(2\pi\omega t) + e(t), \quad (28.5)$$

and a Poisson distribution if the studied outcome is daily disease counts,

$$\log(E[Y(t)]) = \beta_0 + \beta_1 \sin(2\pi\omega t) + \beta_2 \cos(2\pi\omega t) + e(t). \quad (28.6)$$

In both cases, β_0 is an intercept that estimates a baseline of a seasonal pattern. With t as time, expressed in days for a time series of length N ($t = 1, 2, \dots, N$, where N is the number of days in a time series), we set $\omega = 1/365$ to properly express the annual cycle. The $\exp\{\beta_0\}$ for the Poisson regression reflects a mean daily disease count over a study period. We estimated the mean-value function using Models 5 and 6, as well as using the estimates of the amplitude and phase angle, and obtained the exact same plot (Figures 28.2 and 28.4).

Now, using the estimates of the amplitude and the phase angle, the proposed characteristics of seasonality can be expressed as follows.

1. The average maximum value on the seasonal curve of exposure, $\max\{Y(t)\} = \beta_0 + \gamma$, or incidence of disease, $\max\{Y(t)\} = \exp\{\beta_0 + \gamma\}$;
2. The average minimum value on the seasonal curve of exposure, $\min\{Y(t)\} = \beta_0 - \gamma$, or incidence of disease, $\min\{Y(t)\} = \exp\{\beta_0 - \gamma\}$;

3. The average intensity, the difference between maximum and minimum values on the seasonal curve for exposure, $I = 2\gamma$, or incidence of disease, $I = \exp\{\beta_0 + \gamma\} - \exp\{\beta_0 - \gamma\}$;
4. The average relative intensity, the ratio of maximum value and minimum value on the seasonal curve, for exposure, $I_R = (\beta_0^2 - \gamma^2)/(\beta_0 - \gamma)^2$, or incidence of disease, $I_R = \exp\{2\gamma\}$;
5. The average peak timing (in days), a position of the maximum point on the seasonal curve of exposure or disease incidence, $P = 365(1 - \psi/\pi)/2$;
6. The average lag, $P_E - P_D$, the difference between peak timing of exposure, P_E , and peak timing of disease incidence, P_D .

The results of fitting Model 2, as well as the estimated amplitude and phase angle parameters are shown in Table 28.1. We used S-Plus glm-function to fit the models. S-Plus codes for estimation of seasonality parameters are available on request. Suggested models demonstrate that a seasonal component explained 83% of variability in daily temperature and 23% in counts of *Salmonella* infections. The *Salmonella* infections peaked two weeks after a peak in temperature.

Table 28.1: Characteristics of seasonal curves for ambient temperature and *Salmonella* cases

	Temperature Parameters Value (Std.error)	Disease Parameters Value (Std.error)
Intercept— β_0	58.971 (0.1213)	1.377 (0.0086)
$\sin(2*\pi*time/365)$ — β_1	-9.187 (0.1716)	-0.3111 (0.0117)
$\cos(2*\pi*time/365)$ — β_2	-21.093 (0.1715)	-0.4331 (0.0118)
Null variance (df = 3652)	1163591	8967
Residual variance (df = 3650)	196222	6871
% variance explained	83%	23%
Amplitude— γ	22.9698	0.5412
Phase angle— ψ	-0.4071	-0.6489
Relative intensity— I_R	2.2760	2.9519
Peak timing— P	206.1	220.2

Next, we hypothesized that temporality in ambient temperature will determine, in part, the timing and magnitude of peaks and we explored associations between seasonal characteristics in disease and temperature. Specifically, we asked the question, “Do the timing and/or intensity of a seasonal peak in ambient temperature predict the timing and/or intensity of the seasonal peak for

an enteric infection?” In order to answer this question, we examined seasonal characteristics in ambient temperature and *Salmonella* infections in the manner described above for each year separately and then examined the synchronization of seasonal patterns in temperature and *Salmonella* counts. A few interesting observations are: moderate association exists between relative intensities for temperature and *Salmonella* cases ($\rho = 0.648$), and negative correlation exists between average minimum values for temperature and average maximum values for *Salmonella* infections ($\rho = -0.806$). These results suggest that ambient temperature can be a potential predictor of *Salmonella* infections at a seasonal scale.

28.4 Conclusion

An ability to provide estimates for seasonality characteristics as a set of parameters was the main objective in developing the presented models. The presented set of analytical tools allows for comprehensive, systematic, and detailed examination of a seasonal pattern in daily time series of continuous and discrete outcomes. The application indicates the promise of these techniques to produce sensible and intuitively appearing functional relationships. The suggested conceptual structure permits the description of seasonal patterns and their comparison. In fitting a GLM with a cosine function for a seasonal curve we assumed that a pattern described by a cosine curve has a symmetric rise and fall, and a cosine curve with a period of a full year has a point at which it peaks and a point with the lowest value. We demonstrated an approach, in which we combine the ease of fitting one model with the simplicity and elegance of interpretation of another one, by using the δ -method. We also demonstrated that the proposed technique could be extended to a more general case, for example, when two seasonal peaks can be identified. Clearly, further experience in using these techniques and some theoretical work are required. It is important to compare the performance of models with well-documented statistical techniques for seasonality evaluation, to expand visual presentation of modeling results, and to provide step-by-step instructions for implementing these statistical procedures in practical settings for public health professionals. The presented models and parameter estimation procedures allow for a straightforward interpretation, are easy to perform using commercial statistical software, and are valuable tools for investigating seasonal patterns in biosurveillance.

One methodological aspect of this exercise deserves special comment. The vast majority of epidemiological studies that have examined the seasonality of diseases used crude quarterly or monthly aggregate data which prevent a fully detailed, accurate, or comprehensive analysis of a seasonal pattern and may

even be misleading [da Silva Lopes (1999)]. Examination of weekly rates substantially improves the evaluation of seasonal curves when compared to monthly data, but a systematic approach to the issue of week standardization has often been lacking. The use of daily time series enabled us to detect significant differences in the seasonal peaks of infections, which would have been lost in a study that used monthly cumulative information. The effective use of the presented methods requires data collected over a long period with sufficient frequency. An efficient surveillance system has similar requirements. The vast majority of continuously monitored surveillance systems collect data on a daily basis and focus on the use of daily time series.

Acknowledgements. We wish to thank Drs. Jeffrey Griffiths and Andrey Egorov for their thoughtful suggestions, and Drs. Alfred DeMaria and Bela Matyas and the Massachusetts Department of Public Health for providing us with surveillance data. We would also like to thank the USA EPA, the National Institute of Allergy and Infectious Diseases, and the National Institute of Environmental Health Sciences that provided funding through the Subcontract Agreement, AI43415, and ES013171 grants, respectively.

References

1. MacNeill, I. B. (1977). A test of whether several time series share common periodicities, *Biometrika*, **64**, 495–508.
2. da Silva Lopes, A. C. B. (1999). Spurious deterministic seasonality and autocorrelation corrections with quarterly data: Further Monte Carlo results, *Empir Econ.*, **24**, 341–359.