Environmetrics Research Article

Received: 20 March 2015, Revised: 29 June 2015, Accepted: 30 June 2015, Published online in Wiley Online Library: 3 August 2015

(wileyonlinelibrary.com) DOI: 10.1002/env.2356

# Estimating trend from seasonal data: is daily, monthly or annual data best?

**Christopher S. Withers[a] and Saralees Nadarajah[b]***

**We consider a model for time series whose mean has a component that is linear in time with slope $b$ plus seasonal sinusoidal components. Suppose we have $N$ years of observations and $J$ observations per year. We answer the question: what is the loss in efficiency of estimating $b$ if the $J$ observations are grouped into weekly means, monthly means, annual means, and so on? We give answers for the cases of a single sinusoidal component and $q$ sinusoidal components. We derive tests of hypotheses, confidence intervals, and power functions for the slope. Finally, an application of the results to daily minimum and daily maximum temperature data from Auckland is presented. Copyright © 2015 John Wiley & Sons, Ltd.**

**Keywords:** efficiency; power; trend

## 1. INTRODUCTION

Aggregation of data is common practice in almost every area of the sciences; see Cholette and Chhab (1991) for a problem of aggregating weekly data into monthly values. See also Gouno *et al.* (2011). The aim of this paper is to investigate the loss of efficiency in estimation when data are aggregated.

There have been papers comparing estimates based on aggregated/non-aggregated data. Tiao (1972) compared forecasts for aggregated/non-aggregated time series data, stationary or non-stationary. Hsiao (1979) and Palm and Nijman (1982) compared estimates of a linear regression model for aggregated/non-aggregated data. But both Hsiao (1979) and Palm and Nijman (1982) assumed normality. Garrett (2003) explained why regression coefficients differ across degrees of data aggregation. More recent papers have been based on complicated models: money demand function of Japan under the low interest rate policy (Hsiao *et al*., 2005), meta-analysis of diagnostic test studies (Riley *et al*., 2008), and mixed treatment models (Saramago *et al*., 2012).

Our investigation is based on a simple and commonly used model. It does not assume normality. Furthermore, the model incorporates seasonal components, a common feature with environmental data. But we suppose that the data have no serial correlation. The serial correlation in most data sets can be removed by suitable filtering (for example, by taking maximums or minimums over windows of sufficient length), and the filtered data can be suitably modeled for possible trends or seasonality. An example is given in Section 5. Some published examples due to the authors include Nadarajah (2005) examining trends in extreme rainfall in west central Florida; Nadarajah and Shiau (2005) examining trends in extreme river flow in Pachang River, Taiwan; Withers and Nadarajah (2006) examining trends in extreme windrun in New Zealand; Feng *et al.* (2007) examining trends in extreme precipitation in China; and Withers *et al.* (2009) examining trends in extreme temperature in New Zealand.

Suppose we observe $N$ years of data with $J$ equally spaced observations per year—a total of $T = NJ$ observations. We model the data as *a linear signal + a periodic signal + noise*:

$$Y_t = a + b\left(t - \bar{t}\right) + \mathbf{c}'\mathbf{s}(t) + e_t \tag{1}$$

for $1 \leqslant t \leqslant T = NJ$, where $\bar{t} = T^{-1}\sum_{t=1}^{T} t = (T+1)/2$, $\mathbf{c}$ is a $q$ by one vector of unknown parameters, $\mathbf{s}(t)$ is a given $q$ by one function with period 1 year and mean zero:

$$\mathbf{s}(t+J) = \mathbf{s}(t),\ \sum_{t=1}^{J}\mathbf{s}(t) = \mathbf{0} \tag{2}$$

Furthermore, $\{e_t\}$ are uncorrelated and have zero mean, common variance $\sigma^2$ say. It follows by the Gauss–Markov theorem that the best linear unbiased estimator of the unknown parameters $a$, $b$ is the least squares estimator (LSE) based on all the data. Such a model is

* *Correspondence to: Saralees Nadarajah, School of Mathematics, University of Manchester, Manchester, M13 9PL, U.K. E-mail: mbbsssn2@manchester.ac.uk*

a *Industrial Research Limited, Lower Hutt, New Zealand*

b *School of Mathematics, University of Manchester, Manchester, M13 9PL, U.K.*

 

appropriate for a range of problems, for example, for analyzing a number of years of temperature or sea-level data subject to global warming, or for analyzing employment figures in a growing population allowing for seasonality.

The parameter of prime interest is the annual slope

$$B = Jb$$

A common practice is to estimate the slope using only annual means (see, for example, Donnelly *et al.* (2012)). What efficiency is lost in this way? More generally, what do we lose if we base our LSE not on the original observations but on combined data of groups of say $M$ observations? This approach reduces our number of observations from $J$ per year to $J/M$ per year.

In Section 2, we show that reducing the number of observations from $J$ per year to $J/M$ per year recovers a model of the form (1) and (2). We then answer these questions for the important case where the periodic component is a single sinusoid, that is

$$\mathbf{s}(t) = (\sin wt, \cos wt)'$$

where $w = 2\pi/J$. Of course, if we analyze only annual means, we do not have to assume any particular form for $\mathbf{s}(t)$; this is equivalent to replacing $a + \mathbf{c}'\mathbf{s}(t)$ in (1) by $d_j$ for $t = (n-1)J + j, 1 \leqslant j \leqslant J, 1 \leqslant n \leqslant N$.

In Section 3, we extend these results to the general sinusoid model, that is, $\mathbf{s}(t)$ given by

$$\mathbf{s}_{q,J}(t) = (\sin wt, \cos wt, \sin 2wt, \cos 2wt, \ldots, \sin qwt, \cos qwt)' \tag{3}$$

with $w = 2\pi/J$.

In Section 2, we show that for $\mathbf{s} = \mathbf{s}_{1,J}$ of (3), both the efficiency of $\widehat{B}$ and the LSE of the annual slope based on only annual means or totals relative to the LSE based on all $J$ observations per year never drop below 88% if there are 2 or more years of data, or 95% if there are 3 or more years of data. The only problem occurs when there is only 1 year of data; the LSE is not computable unless $J \geqslant 4$; for quarterly observations ($J = 4$), efficiency is only 48%, but for bi-monthly observations ($J = 6$), efficiency is 78%, and for monthly observations, it rises to 95%.

In Section 4, we assume independence and normality in order to calculate one-sided and two-sided confidence intervals and to derive test for the slope. Section 4 also provides power calculations. Section 5 applies some of the theoretical results to temperature data from Albert Park, Auckland, New Zealand. Section 6 notes some conclusions and future work. Section 6 also answers the question posed in the title. The proofs of all the theorems in Sections 2, 3, and 4 are given in the appendix.

Throughout, a subscript '·' denotes the mean for that subscript. For example, $Y_\cdot = T^{-1}\sum_{t=1}^{T} Y_t$.

## 2. THE LEAST SQUARES ESTIMATOR FOR COMBINED OBSERVATIONS

Suppose that the number of observations per year is written as $J = J'M$, where $J'$ and $M$ are integers. Lumping our $J$ observations per year into groups of $M$ leaves us $J'$ combined observations per year, a total of $T' = NJ' = T/M$ combined observations. Set

$$t' - 1 = [(t-1)/M]$$

for $t' = 1, \ldots, T'$ and $m = t - (t'-1)M$, where $[x]$ is the integer part of $x$. As $t$ increases from 1 to $T$, $t' = 1, \ldots 1$, ($M$ times) $2, \ldots, 2$, ($M$ times) $\ldots, T', \ldots T'$, ($M$ times) and $m = 1, \ldots, M, 1, \ldots M, \ldots, 1, \ldots M$.

Theorem 2.1 shows that reducing the observations from $J$ per year to $J' = J/M$ per year by lumping in this way recovers the same model (1).

**Theorem 2.1** *Suppose that the observations $Y_t$ satisfy the model (1) with $\{e_t\}$ uncorrelated with zero means and common variance $\sigma^2$. Lumping the observations as described, the mean of the $M$ observations $M^{-1}\sum_{t=(t'-1)M+1}^{t'M} Y_t = Y_{M,t'}$ satisfies*

$$Y_{M,t'} = a + b_1\left(t' - \overline{t'}\right) + \left(\boldsymbol{c}^*\right)'\boldsymbol{s}^*\left(t'\right) + e_{M,t'} \tag{4}$$

*where*

$$\overline{t'} = \left(T'\right)^{-1}\sum_{t'=1}^{T'} t'$$

$$b_1 = Mb,$$

$$\boldsymbol{s}^*\left(t'\right) = M^{-1}\sum_{t=(t'-1)M+1}^{t'M} \boldsymbol{s}(t) \text{ has period } J',$$

$$e_{M,t'} = M^{-1}\sum_{t=(t'-1)M+1}^{t'M} e_t$$

*where $\left\{e_{M,t'}\right\}$ are now uncorrelated with zero means and common variance $\sigma^2/M$.*



  

Consider the case $\mathbf{s}(t) = \mathbf{s}_{q,J}(t)$ of (3). By equations (1.341.1) and (1.341.3) in Gradshteyn and Ryzhik (2014),

$$\begin{aligned} M^{-1}\sum_{t=(t'-1)M+1}^{t'M}\sin(wjt) &= M^{-1}\sum_{x=0}^{M-1}\sin\left(w\left(x+\left(t'-1\right)M+1\right)j\right) \\ &= M^{-1}\sum_{x=0}^{M-1}\sin\left(wjx+w\left(t'-1\right)Mj+wj\right) \\ &= M^{-1}\csc\left(\frac{wj}{2}\right)\sin\left(\frac{wMj}{2}\right)\sin\left(wjMt'+\frac{wj(1-M)}{2}\right) \\ &= M^{-1}\csc\left(\frac{wj}{2}\right)\sin\left(\frac{w'j}{2}\right)\sin\left(w'jt'+\frac{\left(M^{-1}-1\right)w'j}{2}\right) \end{aligned}$$

and

$$\begin{aligned} M^{-1}\sum_{t=(t'-1)M+1}^{t'M}\cos(wjt) &= M^{-1}\sum_{x=0}^{M-1}\cos\left(w\left(x+\left(t'-1\right)M+1\right)j\right) \\ &= M^{-1}\sum_{x=0}^{M-1}\cos\left(wjx+w\left(t'-1\right)Mj+wj\right) \\ &= M^{-1}\csc\left(\frac{wj}{2}\right)\sin\left(\frac{wMj}{2}\right)\cos\left(wjMt'+\frac{wj(1-M)}{2}\right) \\ &= M^{-1}\csc\left(\frac{wj}{2}\right)\sin\left(\frac{w'j}{2}\right)\cos\left(w'jt'+\frac{\left(M^{-1}-1\right)w'j}{2}\right) \end{aligned}$$

where $w' = Mw = 2\pi/J'$. So, (4) holds with $\mathbf{c}^*$ and $\mathbf{s}^*(t)$ given by

$$\mathbf{c}^* = \mathbf{c} = \left(\alpha_1, \beta_1, \ldots, \alpha_q, \beta_q\right)' \tag{5}$$

and

$$\begin{aligned} \mathbf{s}^*(t) = \frac{1}{M}\Bigg(&\csc\left(\frac{w}{2}\right)\sin\left(\frac{w'}{2}\right)\sin\left(w't+\frac{\left(M^{-1}-1\right)w'}{2}\right) \\ &\csc\left(\frac{w}{2}\right)\sin\left(\frac{w'}{2}\right)\cos\left(w't+\frac{\left(M^{-1}-1\right)w'}{2}\right) \\ &\vdots \\ &\csc\left(\frac{wq}{2}\right)\sin\left(\frac{w'q}{2}\right)\sin\left(w'qt+\frac{\left(M^{-1}-1\right)w'q}{2}\right) \\ &\csc\left(\frac{wq}{2}\right)\sin\left(\frac{w'q}{2}\right)\cos\left(w'qt+\frac{\left(M^{-1}-1\right)w'q}{2}\right)\Bigg) \end{aligned}$$

respectively.

Theorem 2.1 shows that we can analyze (1) by dropping the primes. Theorem 2.2 estimates $B$ for the case $\mathbf{s}(t) = \mathbf{s}_{1,J}(t)$ of (3).

**Theorem 2.2** *Suppose $Y_t$ satisfies*

$$Y_t = a + \boldsymbol{X}_t'\,\boldsymbol{\theta} + e_t$$

*where $\boldsymbol{\theta} = (b, \alpha, \beta)'$ and $\boldsymbol{X}_t = \left(t-\bar{t}, \sin\ wt, \cos\ wt\right)'$ for $w = 2\pi/J$. Assume that the number of data points $NJ$ is not less than four, the number of unknowns of this model. For $J \neq 1$ or 2,*

$$\widehat{B}_J = J\widehat{b}_J = 12JT^{-3}\left\{\sum_{t=1}^{T}\left(t-\bar{t}+\gamma_1\sin wt-\cos wt\right)Y_t\right\}/\left(1-d_J T^{-2}\right)$$

*and*

$$\begin{aligned} var\left(\widehat{B}_J\right) &= 12J^2\ T^{-3}\left(1-d_J T^{-2}\right)^{-1}\sigma^2 \\ &= 12N^{-3}\left(1-f_J\ N^{-2}\right)^{-1}\sigma^2/J \\ &\approx 12N^{-3}\sigma^2/J \text{ for } N \text{ large and } J \text{ fixed}, \end{aligned} \tag{6}$$



  

*where* $\widehat{b}_J$ *denotes the estimator of* $b$ *and*

$$\gamma_1 = \cot(w/2) = \cot(\pi/J),\ d_J = 7 + 6\gamma_1^2,\ f_J = d_J J^{-2} = \left\{7 + 6\cot^2(\pi/J)\right\} J^{-2} \tag{7}$$

For $N = 1$ and $J \leqslant 3$, $\widehat{B}_J$ does not exist. According to Theorem 2.2, the variance of $\widehat{B}_J$ changes much faster with the number of years $\left(\sim N^{-3}\right)$ instead of the usual $N^{-1}$.

Theorems 2.3 and 2.4 show that (6) also holds for $J = 1, 2$, where now $f_J = 1$.

**Theorem 2.3** *If* $J = 2$*, then (6) holds with* $d_2 = 4,\ f_2 = 1$ *and*

$$\widehat{B}_2 = 24T^{-3}\left\{\sum_{t=1}^{T}\left(t - \bar{t} - (-1)^t/2\right)\ Y_t\right\} / \left(1 - 4T^{-2}\right)$$

*for* $N > 1$.

**Theorem 2.4** *If* $J = 1$*, then (6) holds with* $d_1 = f_1 = 1$ *and*

$$\widehat{B}_1 = S_{XX}^{-1} S_{XY} = 12\left(T^3 - T\right)^{-1} \sum_{t=1}^{T}\left(t - \bar{t}\right)\ Y_t$$

*for* $T = N > 1$.

Theorems 2.2–2.4 prove that $\text{var}\left(\widehat{B}_J\right)$ is given by (6) with $f_1 = f_2 = 1$ and $f_J$ given by (7) otherwise. By Figure 3, $f_J$ decreases from $f_3 = 1$ to $f_\infty = 6\pi^{-2} \approx 0.605$ as $J$ increases from 3 to $\infty$, so $J\,\text{var}\left(\widehat{B}_J\right)$ decreases from $12N^{-3}\left(1 - N^{-2}\right)^{-1}\sigma^2$ for $J = 1, 2$ or 3 to $12N^{-3}\left(1 - 6\pi^{-2}N^{-2}\right)^{-1}\sigma^2$ for $J = \infty$.

Now let $\widehat{B}_{J',M}$ denote the LSE for $B$ based on taking the mean of each group of $M$ successive observations, that is, based on $J' = J/M$ observations per year each with variance $\sigma^2/M$. By (6),

$$\text{var}\left(\widehat{B}_{J',M}\right) = 12N^{-3}\left(1 - d_{J'}\left(T'\right)^{-2}\right)^{-1}\sigma^2/J = 12N^{-3}\left(1 - f_{J'}N^{-2}\right)^{-1}\sigma^2/J$$

for $NJ/M \geqslant 4$. So, the efficiency of $\widehat{B}_{J',M}$ relative to $\widehat{B}_J$ is

$$\begin{aligned}\text{eff}_{J',M:J} &= \text{var}\left(\widehat{B}_J\right)/\text{var}\left(\widehat{B}_{J',M}\right)\\ &= \left(1 - d_{J'}\left(T'\right)^{-2}\right) / \left(1 - d_J T^{-2}\right)\\ &= \left(1 - f_{J'}N^{-2}\right) / \left(1 - f_J N^{-2}\right)\end{aligned} \tag{8}$$

for $NJ/M \geqslant 4$. In particular, the efficiency of $\widehat{B}_{1,J}$ that is using annual means or totals, relative to $\widehat{B}_J$ is

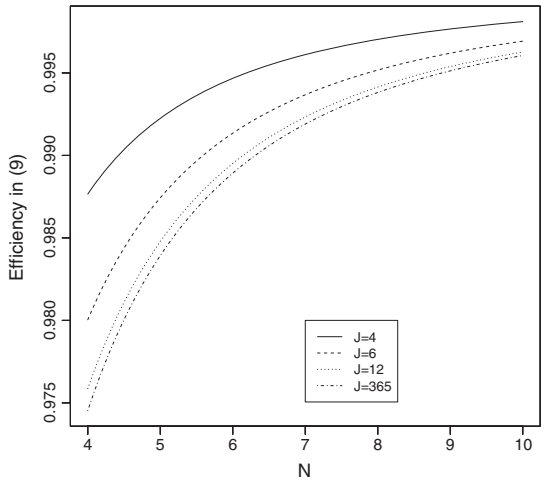


**Figure 1.** Efficiency of slope based on combining all $J$ observations per year into one annual mean or total



  

$$\text{eff}_{1,J:J} = \left(1 - J^2 T^{-2}\right) / \left(1 - d_J T^{-2}\right) = \left(1 - N^{-2}\right) / \left(1 - f_J N^{-2}\right) \tag{9}$$

for $N \geqslant 4$. By the Gauss–Markov theorem, this cannot exceed one. In fact, it equals one for $J = 1, 2, 3$. Figure 1 graphs this efficiency for selected $J$. It never drops below 88% if there is more than 1 year of data or 95% if there is more than 2 years of data.

## 3. THE EFFECT OF MORE SINUSOIDS

Suppose instead of allowing only annual frequencies in the seasonal component, we include six monthly frequencies. That is, we choose $\mathbf{s}(t) = \mathbf{s}_{2,J}(t)$ of (3).

Adapting the previous section in an obvious way, $\mathbf{S}_{XXT}$ is now 5 by 5 with extra elements $((1,4),(1,5),(2,4),(2,5),(3,4),(3,5),(4,4),(4,5),(5,5))$ given by $(-\gamma_2/(2T), 1/(2T), 0, 0, 0, 0, 1, 0, 1)$, respectively, where $\gamma_i = \cot iw/2$. This gives

$$\det(\mathbf{S}_{XXT}) = \left(1 - T^{-2} d_{J,2}\right)/192,$$

where $d_{J,2} = 13 + 6\gamma_1^2 + 6\gamma_2^2$ for $J \geqslant 5$. So, for $J \geqslant 5$ and $NJ \geqslant 6$, $\text{var}\left(\widehat{B}_J\right)$ is given by (6) with $f_J = d_{J,2} J^{-2}$. So, for $J \geqslant 5$ and $N \geqslant 6$, the efficiency of analysis on annual means or totals is given by (8) with this new $f_J$. This is graphed in Figure 2. Set

$$u_1(v) = \text{var}(\widehat{\alpha}_2) = 2^{-1} v N^{-1} \left[1 - (6 + d_J) T^{-2}\right] / \left(1 - d_{J,2} T^{-2}\right)$$
$$u_2(v) = \text{var}\left(\widehat{\beta}_2\right) = 2^{-1} v N^{-1} \left[1 - \left(6\gamma_2^2 + d_J\right) T^{-2}\right] / \left(1 - d_{J,2} T^{-2}\right)$$

If $\alpha_2 = \beta_2 = 0$, then

$$\widehat{\alpha}_2^2 / u_1(v) + \widehat{\beta}_2^2 \approx \chi_2^2 = 2\exp(1)$$

the result being exact for normal noise. Here, $\approx$ means "approximately distributed as." So, an approximately 5% level test of the hypothesis that the bi-monthly frequency is present is to accept when $\widehat{\alpha}_2^2 / u_1(\widehat{v}) + \widehat{\beta}_2^2 > -2\log_e(0.05) = 5.99\cdots$.

Theorem 3.1 estimates $B$ for the case of $q$ sinusoids.

**Theorem 3.1** *Suppose* $\mathbf{s}(t) = \mathbf{s}_{q,J}(t)$ *of (3) and set*

$$d_{J,q} = 1 + 6q + 6\sum_{j=1}^{q} \cot^2(j\pi/J)$$
$$a_1 = \left(1 - T^{-2}\right)/6,$$
$$a_i = \begin{cases} T^{-1}, & \text{for } i = 3, 5, \ldots, 2q+1 \\ -\gamma_{i/2} T^{-1}, & \text{for } i = 2, 4, \ldots, 2q \end{cases}$$

*Then,* $\text{var}\left(\widehat{B}_J\right)$ *is given by (6) with*

$$f_J = d_{J,q} J^{-2} \tag{10}$$

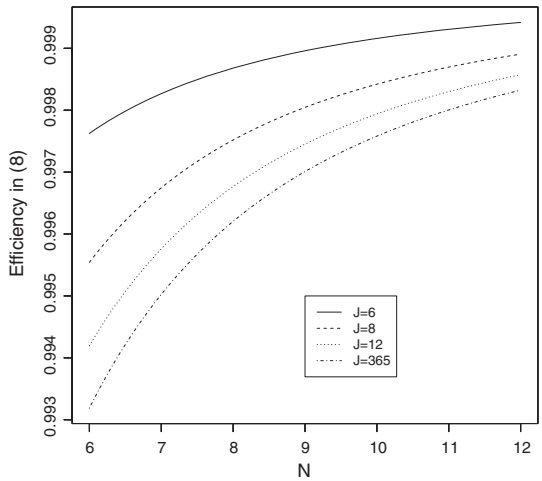


**Figure 2.** Efficiency of slope based on combining all $J$ observations per year into one annual mean or total allowing two sinusoids



  

*and the efficiency of analysis on annual means or totals is given by (8) with this new* $f_J$ *for* $J \geqslant 2q+1$ *and* $N \geqslant 2q+2$. *As for fixed* $q$, *as* $J \to \infty$, $f_J \to 6\pi^{-2}\left(\sum_{j=1}^{q} j^{-2}\right)$.

The $f_J$ in (10) is plotted in Figure 3. Note the equality $f_{J,q} = 1$ for $J = 2q+1$, confirmed numerically for $q \leqslant 10$.

## 4. CONFIDENCE INTERVALS, TESTS, AND POWER

Theorem 4.1 derives one-sided and two-sided confidence intervals and tests about $B$. It also derives the power.

**Theorem 4.1** *Suppose that our model (1) holds with periodic component* $\boldsymbol{s}(t) = \boldsymbol{s}_{1,J}(t)$ *of (3), that is*

$$\boldsymbol{s}(t) = (\sin wt, \cos wt)'$$

*with* $w = 2\pi/J$. *Suppose also that* $\{e_t\}$ *are independent and identical* $\mathcal{N}(0, \sigma^2)$. *Let*

$$\widehat{\lambda}(B) = N^{3/2} J^{1/2} \left(B - \widehat{B}\right) C_{N,J}^{-1/2} / \widehat{\sigma}$$

*where*

$$C_{N,J} = 12\left(1 - f_J N^{-2}\right)^{-1},\ \widehat{\sigma}^2 = f_J^{-1} n_J^{-1} \sum_{t=1}^{T} \left(Y_t - \widehat{Y}_t\right)^2$$

*and*

$$\widehat{Y}_t = \begin{cases} \widehat{a} + \widehat{b}\left(t - \bar{t}\right) + \widehat{\alpha} \sin wt + \widehat{\beta} \cos wt, & \text{if } J > 2 \\ \widehat{a} + \widehat{b}\left(t - \bar{t}\right) + \widehat{\beta}(-1)^t, & \text{if } J = 2 \\ \widehat{a} + \widehat{b}\left(t - \bar{t}\right), & \text{if } J = 1 \end{cases}$$

*where*

$$n_J = NJ - p \text{ for } p = \begin{cases} 4, & \text{for } J > 2 \\ 3, & \text{for } J = 2 \\ 2, & \text{for } J = 1 \end{cases}$$

*Assume that* $C_{N,J} > 0$ *and* $J + N > 3$. *Then, a two-sided* $(1-\alpha)$*-level confidence interval for* $B$ *is given by*

$$\left|\widehat{\lambda}(B)\right| < t_{n_J, 1-\alpha/2}.$$

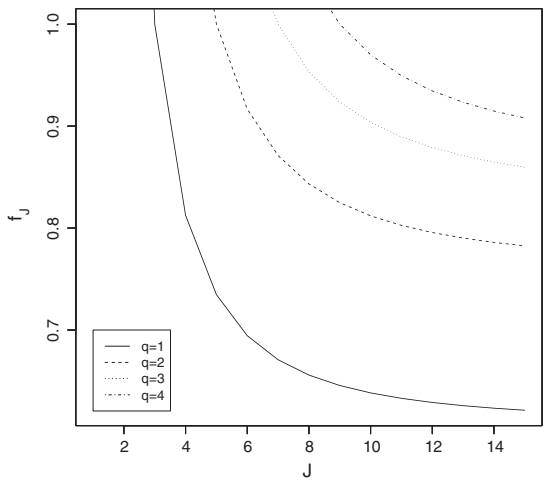


**Figure 3.** $f_J$ of (10)



  

*A one-sided* $(1-\alpha)$*-level test of* $H_0 : B = B_0$ *(a given value) versus* $H_1 : B > B_0$ *is to accept the hypothesis* $H_0$ *if and only if*

$$\widehat{\lambda}(B_0) < t_{n_J,1-\alpha}. \tag{11}$$

*Also*

$$\widehat{\lambda}(B_0) \sim t_{n_J}(\delta_J)$$

*where*

$$\begin{aligned} \delta_J &= N^{3/2} J^{1/2} (B - B_0) C_{N,J}^{-1/2} / \sigma = N^{3/2} \delta^* C_{N,J}^{-1/2} \\ \delta^* &= J^{1/2} (B - B_0) / \sigma \end{aligned} \tag{12}$$

*So, the test (11) has power*

$$P_J = P\left(t_{n_J}(\delta_J) > t_{n_J,1-\alpha}\right) \tag{13}$$

*For fixed* $J$ *and* $B - B_0$ *as* $N \to \infty$, $\delta_J \to \infty$ *under* $H_1$, *so the power* $P_J \to 1$. *For* $N > 1$ *(or* $J > 2$*),* $n_J$ *increases with* $J$. *Also* $\delta_J$ *increases with* $J$.

Figures 4–6 plot the power $P_J$ of (13) against $J$ for $\delta^*$ of (12) equal to 0.05, 0.1, and 0.5 when $\alpha = 0.05$ for a range of $N$. For $\delta^*$ equal to 1, the power is almost one for $\alpha = 0.05$. The figures illustrate that for fixed $N$, the power $P_J$ increases as $J$ increases. Its limit for $J = \infty$ is

$$\Phi(d - z) = 1 - \phi(d - z) d^{-1} \left[1 + O\left(d^{-1}\right)\right]$$

as $N \to \infty$, where $d = \delta_\infty = 12 N^{1.5} \delta^* \left(1 - 6\pi^{-2} N^{-2}\right)^{-1}$, $z = \Phi^{-1}(1-\alpha)$, $\Phi(\cdot)$ denotes the distribution function of a standard normal random variable, and $\phi(\cdot)$ denotes the density function of a standard normal random variable.

The confidence intervals, tests, and power can also be calculated for aggregated data. Let

$$\widehat{\mu}(B) = N^{3/2} \left(J'\right)^{1/2} \left(B - \widehat{B}_{J',M}\right) C_{N,J'}^{-1/2} / \widehat{\sigma}'$$

where

$$\left(\widehat{\sigma}'\right)^2 = f_{J'}^{-1} n_{J'}^{-1} \sum_{t'=1}^{NJ'} \left(Y_{M,t'} - \widehat{Y}_{M,t'}\right)^2$$

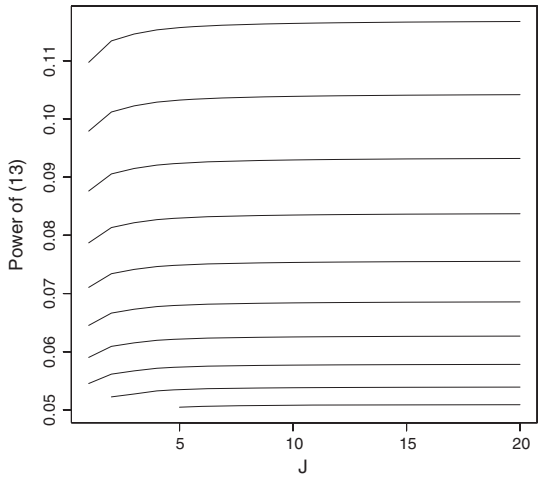


**Figure 4.** Power of one-sided 5% level test of $b_A = b_0$ when $\delta^* = 0.05$. $N = 1, 2, \ldots, 10$ correspond to the curves from the bottom to the top



  

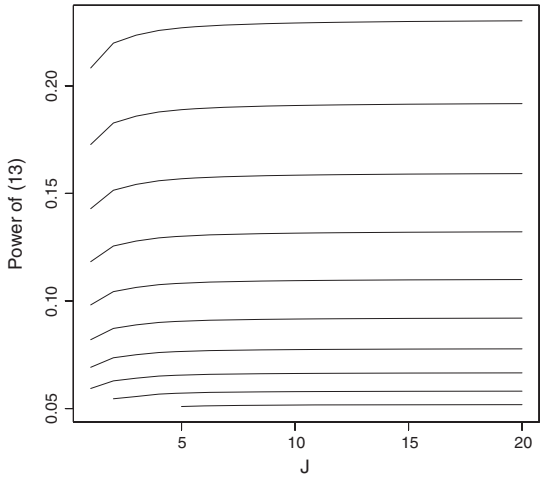


**Figure 5.** Power of one-sided 5% level test of $b_A = b_0$ when $\delta^* = 0.1$. $N = 1, 2, \ldots, 10$ correspond to the curves from the bottom to the top

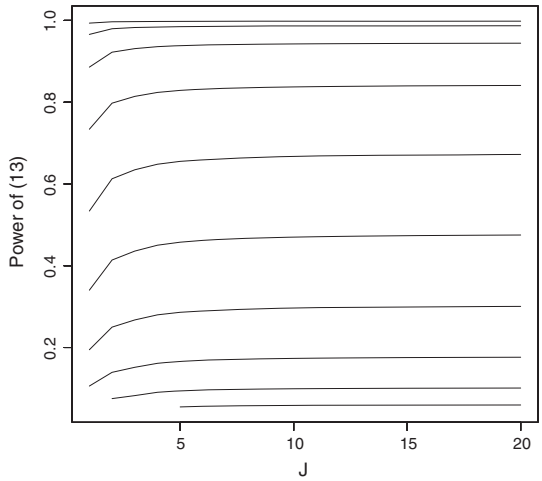


**Figure 6.** Power of one-sided 5% level test of $b_A = b_0$ when $\delta^* = 0.5$. $N = 1, 2, \ldots, 10$ correspond to the curves from the bottom to the top

and

$$\widehat{Y}_{M,t'} = \begin{cases} \widehat{a} + M\widehat{b}\left(t' - \overline{t'}\right) \\ \quad + M^{-1}\widehat{\alpha}\csc\left(\frac{w}{2}\right)\sin\left(\frac{w'}{2}\right)\sin\left(w't' + \frac{\left(M^{-1}-1\right)w'}{2}\right) \\ \quad + M^{-1}\widehat{\beta}\csc\left(\frac{w}{2}\right)\sin\left(\frac{w'}{2}\right)\cos\left(w't' + \frac{\left(M^{-1}-1\right)w'}{2}\right), & \text{if } J > 2 \\ \widehat{a} + M\widehat{b}\left(t' - \overline{t'}\right) + M^{-1}\widehat{\beta}\sum_{t=(t'-1)M+1}^{t'M}(-1)^t, & \text{if } J = 2 \\ \widehat{a} + M\widehat{b}\left(t - \overline{t}\right), & \text{if } J = 1 \end{cases}$$

If $C_{N,J'} > 0$ and $J' + N > 3$, then a two-sided $(1-\alpha)$-level confidence interval for $B$ is given by

$$|\widehat{\mu}(B)| < t_{n_{J'},1-\alpha/2}.$$

A one-sided $(1-\alpha)$-level test of $H_0 : B = B_0$ (a given value) versus $H_1 : B > B_0$ is to accept the hypothesis $H_0$ if and only if

  

$\widehat{\mu}(B_0) < t_{n_{J'},1-\alpha}$

Also

$\widehat{\mu}(B_0) \sim t_{n_{J'}}(\omega_{J'})$

where

$\omega_{J'} = M^{1/2} N^{3/2} \left(J'\right)^{1/2} (B - B_0) C_{N,J'}^{-1/2} / \sigma$

The power of the test is $P\left(t_{n_{J'}}(\omega_{J'}) > t_{n_{J'},1-\alpha}\right)$. Plots of this power versus $J'$ showed similar behavior to those exhibited in Figures 4–6.

## 5. AN EXAMPLE

Here, we apply the aforementioned results to monthly means of daily minimum and of daily maximum temperature data from Albert Park, Auckland from 1910 to 1986. The series was ended in 1989. There is missing data for 1984 and 1987. The raw data is shown in Figure 7.

The test for no serial correlation of the data sets was performed using Durbin and Watson's (1950, 1951, 1971) method. The $p$-values for minimum and maximum monthly temperature data were 0.061 and 0.055, respectively.

### 5.1. Minimum monthly temperatures

Using annual means gives $\widehat{B}_{1,12} = 0.0111$, s.e. $= 0.0026$, $\widehat{\text{var}}\left(\widehat{B}_{1,12}\right) = 0.00000676$. Using monthly data, for one sinusoid, $\widehat{B}_{12} = 0.0107$, s.e. $= 0.00255$, $\widehat{\text{var}}\left(\widehat{B}_{12}\right) = 0.0000065234$, giving $\widehat{\text{eff}}_{1:12} = 0.965$ as compared with the theoretical value $\text{eff}_{1:12} = 0.9999$. All four coefficients were highly significantly different from zero. For two sinusoids, $\widehat{B}_{12} = 0.0107$, s.e. $= 0.0025$, $\widehat{\text{var}}\left(\widehat{B}_{12}\right) = 0.00000642876$, giving $\widehat{\text{eff}}_{1,12:12} = 0.951$ as compared with the theoretical value $\text{eff}_{1,12:12} = 1.0000$. Five of the six coefficients were highly significantly different from zero, while $c_3 = \alpha_2$ of (5) was not. For three sinusoids coefficients $c_3 = \alpha_2$, $c_5 = \alpha_3$, and $c_6 = \beta_3$ of (5) are not significant, so the two sinusoid model gives the best fit. The Kolmogorov–Smirnov test gave the $p$-value 0.122 for the two sinusoids model.

### 5.2. Maximum monthly temperatures

Using annual means $\widehat{B}_{1,12} = 0.0201$, s.e. $= 0.0025$, $\widehat{\text{var}}\left(\widehat{B}_{1,12}\right) = 0.00000625$. Using monthly data, for one sinusoid, $\widehat{B}_{12} = 0.0202$, s.e. $= 0.00249$, $\widehat{\text{var}}\left(\widehat{B}_{12}\right) = 0.00000618125$, giving $\widehat{\text{eff}}_{1,12:12} = 0.989$ as compared with the theoretical value $\text{eff}_{1,12:12} = 0.9999$. All four coefficients were highly significantly different from zero. For two sinusoids, $\widehat{B}_{12} = 0.0202$, s.e. $= 0.00249$, $\widehat{\text{var}}\left(\widehat{B}_{12}\right) = 0.00000619375$, giving $\widehat{\text{eff}}_{1,12:12} = 0.991$ as compared with the theoretical value $\text{eff}_{1,12:12} = 1.0000$. All six coefficients were highly significantly different from zero. The Kolmogorov–Smirnov test gave the $p$-value 0.101 for the two sinusoids model. It is interesting to note that maximum temperature increased twice as rapidly as minimum temperature.

Figure 8 gives by month the mean residuals for one (dotted line) and two (dashed line) sinusoids. The solid line actually represents the average residuals by month for the non-sinusoidal model,

$$Y_t = a + b\left(t - \overline{t}\right) + c_j + e_t \text{ for month } j. \tag{14}$$

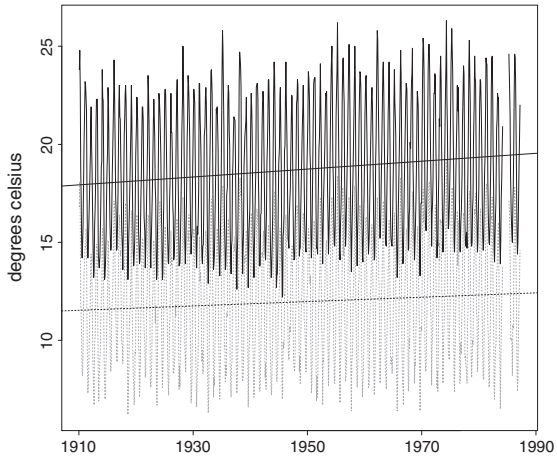


**Figure 7.** Monthly means of maximum (solid line) and minimum (dashed line) daily temperatures for Auckland with straight lines fitted to annual means



  

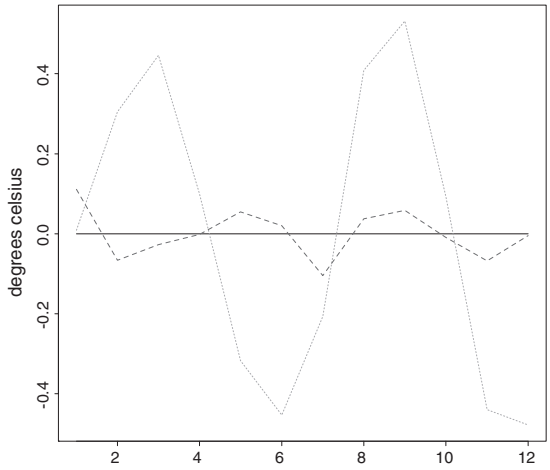


**Figure 8.** Average residuals by month for one and two sinusoids (dashed line) and for the model (14) (line) for maximum temperature

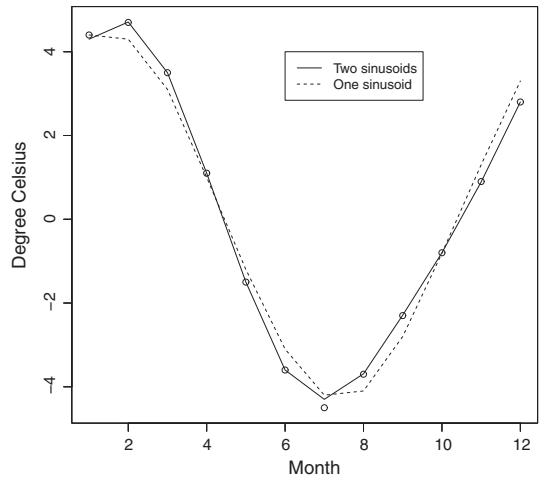


**Figure 9.** Estimated seasonal components for one and two sinusoids (line) and for the model (14) (dashed line) for maximum temperature

The mean residuals by month (i.e., the mean of all the residuals for January each year and February each year) for the fit of the non-sinusoidal model are nearly zero, although the residual sum of squares for this fit is nearly exactly as large as that for the fit with two sinusoids.

Figure 9 gives by month the estimated seasonal components for the three models. The figure shows that the seasonal part of the non-sinusoidal model is essentially the same as that for the two sinusoidal model.

## 6. CONCLUSIONS

Considering a linear model in time with slope $b$ plus an annual sinusoidal component, we have investigated the loss in efficiency in estimating $b$ when the observations are grouped into weekly means, monthly means, annual means, and so on. We have shown that the loss of efficiency in estimating $b$ using only annual means is less than 2% if we have 5 or more years of data. If only 2 years of observations are available, the loss of efficiency is still less than 12%. If only 1 year of data is available, $b$ can only be estimated if the number of observations in a year is greater or equal to the number of parameters in the model. With quarterly data for only 1 year, the estimator is 48% efficient compared with that obtained from continuous data for 1 year. With monthly data for 1 year, this rises to 95%.

To answer the question in the title, we suggest the following rough rule: use annual means if 5 or more years of data are available; and use monthly means if less than 5 years of data are available.

We have stated at several places of using annual means or annual totals instead of non-aggregated data. One should note that the efficiency of using annual means over non-aggregated data or that of using annual totals over non-aggregated data are the same.

We have assumed throughout that the noise $\{e_t\}$ are uncorrelated and have a common variance. This assumption may be simplistic. A future work is to consider some temporal dependence.

We have also considered only the loss in efficiency when lumping observations. Other issues to consider are as follows: (i) loss/gain in the bias of the estimator of $B$; (ii) loss/gain in the mean squared error of the estimator of $B$.



  

## Acknowledgements

The authors would like to thank the Editor, the Associate Editor, and the referee for careful reading and comments that greatly improved the paper.

## REFERENCES

Cholette PA, Chhab NB. 1991. Converting aggregates of weekly data into monthly values. *Applied Statistics* **40**:411–422.
Donnelly A, Broderick B, Misstear B. 2012. A novel method for defining hourly background NO2 and PM10 concentrations for use in local air quality modelling studies and comparison to existing practises. *International Journal of Sustainable Development and Planning* **7**:428–445.
Durbin J, Watson GS. 1950. Testing for serial correlation in least squares regression I. *Biometrika* **37**:409–428.
Durbin J, Watson GS. 1951. Testing for serial correlation in least squares regression II. *Biometrika* **38**:159–178.
Durbin J, Watson GS. 1971. Testing for serial correlation in least squares regression III. *Biometrika* **58**:1–19.
Feng S, Nadarajah S, Hu Q. 2007. Modeling annual extreme precipitation in China using the generalized extreme value distribution. *Journal of the Meteorological Society of Japan* **85**:599–613.
Garrett TA. 2003. Aggregated versus disaggregated data in regression analysis: implications for inference. *Economics Letters* **81**:61–65.
Gouno E, Courtrai L, Fredette M. 2011. Estimation from aggregate data. *Computational Statistics and Data Analysis* **55**:615–626.
Gradshteyn IS, Ryzhik IM. 2014. *Tables of Integrals, Series and Products, Eighth Edition*. Academic Press: New York.
Hsiao C. 1979. Linear regression using both temporally aggregated and temporally disaggregated data. *Journal of Econometrics* **10**:243–252.
Hsiao C, Shen Y, Fujiki H. 2005. Aggregate vs. disaggregate data analysis—a paradox in the estimation of a money demand function of Japan under the low interest rate policy. *Journal of Applied Econometrics* **20**:579–601.
Nadarajah S. 2005. Extremes of daily rainfall in west central Florida. *Climatic Change* **69**:325–342.
Nadarajah S, Shiau JT. 2005. Analysis of extreme flood events for the Pachang River, Taiwan. *Water Resources Management* **19**:363–374.
Palm FC, Nijman TE. 1982. Linear regression using both temporally aggregated and temporally disaggregated data. *Journal of Econometrics* **19**:333–343.
Riley RD, Dodd SR, Craig JV, Thompson JR, Williamson PR. 2008. Meta-analysis of diagnostic test studies using individual patient data and aggregate data. *Statistics in Medicine* **27**:6111–6136.
Saramago P, Sutton AJ, Cooper NJ, Manca A. 2012. Mixed treatment comparisons using aggregate and individual participant level data. *Statistics in Medicine* **31**:3516–3536.
Tiao GC. 1972. Asymptotic behaviour of temporal aggregates of time series. *Biometrika* **59**:525–531.
Withers CS, Krouse DP, Pearson CP, Nadarajah S. 2009. Modelling temperature trends in New Zealand. *Environmental Modeling and Assessment* **14**: 231–249.
Withers CS, Nadarajah S. 2006. Evidence of trend in return levels for daily windrun in New Zealand. *Journal of the Meteorological Society of Japan* **84**: 805–819.

## APPENDIX: THE PROOFS

***Proof of Theorem 2.2.*** The LSE based on $N$ years of $J$ observations per year is given by

$$\widehat{\boldsymbol{\theta}_J} = \mathbf{S}_{XX}^{-1}\,\mathbf{S}_{XY}$$

where

$$\mathbf{S}_{XX} = \sum_{t=1}^{T} \mathbf{X}_t\mathbf{X}_t' - T\mathbf{X}.\mathbf{X}.', \ \mathbf{S}_{XY} = \sum_{t=1}^{T} \mathbf{X}_t Y_t - T\mathbf{X}.Y$$

So,

$$\mathbb{E}\left[\widehat{\boldsymbol{\theta}_J}\right] = \boldsymbol{\theta} \text{ and covar}\left(\widehat{\boldsymbol{\theta}_J}\right) = \mathbf{S}_{XX}^{-1}\sigma^2$$

As $T$ increases, $\mathbf{S}_{XX}$ becomes degenerate. To avoid this, we rescale set

$$\boldsymbol{\theta}_{(T)} = \left(b_{(T)}, \alpha_{(T)}, \beta_{(T)}\right)' = T^{1/2}\left(Tb, \alpha, \beta\right)'$$
$$\mathbf{X}_{t,T}' = T^{-1/2}\left(T^{-1}\left(t-\bar{t}\right), \sin wt, \cos wt\right)$$

where $w = 2\pi/J$ and $\bar{t} = (T+1)/2$. Then $\mathbf{X}_{t,T}'\boldsymbol{\theta}_{(T)} = \mathbf{X}_t'\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}_{(T)} = \mathbf{S}_{XXT}^{-1}\mathbf{S}_{XYT}$ satisfies $\mathbb{E}\left[\widehat{\boldsymbol{\theta}_{(T)}}\right] = \boldsymbol{\theta}$ but now

$$\text{covar}\left(\widehat{\boldsymbol{\theta}}_{(T)}\right) = \mathbf{S}_{XXT}^{-1}\sigma^2$$

where

  

$$\mathbf{S}_{XXT} = \sum_{t=1}^{T} \mathbf{X}_{t,T}\, \mathbf{X}'_{t,T} - T\mathbf{X}_{\cdot T}\, \mathbf{X}'_{\cdot T}$$

retains full rank in the limit as $T \to \infty$ and

$$\mathbf{S}_{XYT} = \begin{bmatrix} T^{-3/2} \sum_{t=1}^{T} \left(t - \bar{t}\right) Y_t \\ T^{-1/2} \sum_{t=1}^{T} \sin wt Y_t \\ T^{-1/2} \sum_{t=1}^{T} \cos wt Y_t \end{bmatrix}$$

Because

$$\sum_{t=1}^{T} \left(t - \bar{t}\right) = 0$$

$$\sum_{t=1}^{T} \sin\left(wt\right) = 0$$

by equation (1.342.1) in Gradshteyn and Ryzhik (2014),

$$\sum_{t=1}^{T} \cos\left(wt\right) = 0$$

by equation (1.342.2) in Gradshteyn and Ryzhik (2014),

we have $\mathbf{X}_{\cdot T} = \mathbf{0}$. Also note that

$$\sum_{t=1}^{T} \left(t - \bar{t}\right)^2 = T\left(T^2 - 1\right)/12$$

$$\sum_{t=1}^{T} \left(t - \bar{t}\right) \sin\left(wt\right) = \sum_{t=1}^{T} t \sin\left(wt\right) = -\frac{T}{2} \cot\left(\frac{w}{2}\right)$$

by equation (1.352.1) in Gradshteyn and Ryzhik (2014),

$$\sum_{t=1}^{T} \left(t - \bar{t}\right) \cos\left(wt\right) = \sum_{t=1}^{T} t \cos\left(wt\right) = \frac{T}{2}$$

by equation (1.352.2) in Gradshteyn and Ryzhik (2014),

$$\sum_{t=1}^{T} \sin^2\left(wt\right) = \frac{T}{2}$$

by equation (1.351.1) in Gradshteyn and Ryzhik (2014),

$$\sum_{t=1}^{T} \sin\left(wt\right) \cos\left(wt\right) = 0$$

$$\sum_{t=1}^{T} \cos^2\left(wt\right) = \frac{T}{2}$$

by equation (1.351.2) in Gradshteyn and Ryzhik (2014).

  

So, for $J \neq 1$ or 2,

$$\mathbf{S}_{XXT} = \begin{bmatrix} \left(1 - T^{-2}\right)/12 & -\gamma_1 T^{-1}/2 & T^{-1}/2 \\ -\gamma_1 T^{-1}/2 & \frac{1}{2} & 0 \\ T^{-1}/2 & 0 & \frac{1}{2} \end{bmatrix}$$

where $\gamma_1 = \cot(\pi/J)$. So,

$$\det(\mathbf{S}_{XXT}) = \left[1 - \left(7 + 6\gamma_1^2\right) T^{-2}\right]/48 = \left(1 - d_J T^{-2}\right)/48$$

and

$$\mathbf{S}_{XXT}^{-1} = \frac{48}{1 - d_J T^{-2}} \begin{bmatrix} 1/4 & \gamma_1/(4T) & -1/(4T) \\ \gamma_1/(4T) & \left(1 - 7T^{-2}\right)/24 & \gamma_1/\left(4T^2\right) \\ -1/(4T) & \gamma_1/\left(4T^2\right) & \left(T^2 - 1 - 6\gamma_1^2\right)/\left(24T^2\right) \end{bmatrix}$$

The proof is complete.

***Proof of Theorem 2.3.*** We have $\sin wt = 0$ and $s(t) = \cos wt = (-1)^t$, so $\boldsymbol{\theta} = (b, \beta)'$, $\mathbf{X}_t' = \left(t - \bar{t}, (-1)^t\right)$, $\boldsymbol{\theta}_{(T)} = T^{1/2}(Tb, \beta)'$ and $\mathbf{X}_{t,T}' = T^{-1/2}\left(T^{-1}\left(t - \bar{t}\right), (-1)^t\right)$. Because

$$\sum_{t=1}^{T} \left(t - \bar{t}\right) = 0$$

$$\sum_{t=1}^{T} \left(t - \bar{t}\right)^2 = T\left(T^2 - 1\right)/12$$

$$\sum_{t=1}^{T} t\,(-1)^t = \frac{T}{2}$$

$$\sum_{t=1}^{T} (-1)^t = 0$$

we obtain

$$\mathbf{X}_{.T} = \mathbf{0}$$

$$\mathbf{S}_{XYT} = \begin{bmatrix} T^{-3/2} \sum_{t=1}^{T} \left(t - \bar{t}\right) Y_t \\ T^{-1/2} \sum_{t=1}^{T} (-1)^t Y_t \end{bmatrix}$$

$$\mathbf{S}_{XXT} = \begin{bmatrix} \left(1 - T^{-2}\right)/12 & 1/(2T) \\ 1/(2T) & 1 \end{bmatrix}$$

$$\det(\mathbf{S}_{XXT}) = \left(1 - 4T^{-2}\right)/12,$$

$$\mathbf{S}_{XXT}^{-1} = \frac{12}{1 - 4T^{-2}} \begin{bmatrix} 1 & -1/(2T) \\ -1/(2T) & \left(1 - T^{-2}\right)/12 \end{bmatrix}$$

so (6) holds. The proof is complete.

***Proof of Theorem 2.4.*** We have

$$s(t) = 0,\ \theta = b,\ X_t = t - \bar{t},\ S_{XXT}^{1,1} = S_{XXT}^{-1} = 12\left(1 - T^{-2}\right)^{-1}$$

so (6) holds. The proof is complete.



  

***Proof of Theorem 3.1.*** For $NJ \geqslant 2q+2$ and $J \geqslant 2q+1$, we obtain

$$\mathbf{S}_{XXT} = \begin{bmatrix} a_1, & a_2, & a_3, & \cdots \\ a_2, & & & \\ a_3, & & \mathbf{I}_{2q} & \\ \vdots & & & \end{bmatrix}$$

So,

$$\det(\mathbf{S}_{XXT}) = 2^{-1-2q}\left(a_1 - \sum_{j=1}^{2q+1} a_j^2\right) = \left(1 - T^{-2}d_{J,q}\right)4^{-q}/12$$

and the result follows.

***Proof of Theorem 4.1.*** Note that

$$\widehat{B}_J \sim \mathcal{N}\left(B, C_{N,J}N^{-3}\sigma^2/J\right)$$

independently of

$$n_J\widehat{\sigma}^2 \sim \sigma^2\chi^2_{n_J}$$

So,

$$\widehat{\lambda}(B) \sim t_{n_J}$$

and the results follow.



  