PH 724 Data Analysis & Scientific Communication Incorporating Use of Generative AI

Part 13: Introduction to Practice Project 4 (Time series, Chapters 6–7 of [Textbook])

Zhiyang Zhou (zhou67@uwm.edu, zhiyanggeezhou.github.io)

2026/11/22 00:16:34

Scientific background

Many scientific measurements are collected **repeatedly over time**.

- Environmental conditions such as temperature and air pollution change from time to time.
- Public-health outcomes such as respiratory infections may vary across seasons.
- Observations close together in time may be more similar than observations far apart.

Classical time-series examples, such as daily temperature, measure a continuous quantity. Public-health surveillance often records **how many events occur** in each time interval.

Scientific background: Central England temperature

Chapters 6–7 study the **Central England daily mean temperature record**.

- Temperatures show a pronounced **annual seasonal cycle**.
- Weather also fluctuates over **days and weeks**.
- A long record can reveal **slower changes over years and decades**.

The challenge is to distinguish systematic patterns from temporally correlated fluctuations and short-term noise.

Scientific question

How does temperature vary over time, and what can we learn from the different time scales of variation?

More specifically:

1. What is the typical **seasonal pattern** within a year?
2. How strongly are temperatures on nearby days **associated**?
3. Is there evidence of **long-term variation or trend**?
4. Which **frequencies** (annual, slower, or faster) explain the variation?
5. ...

Study design: an observational study

This is **not a randomized experiment**.

- **Source:** Hadley Centre Central England Temperature (HadCET)
- **Observation:** One daily mean temperature for each calendar day

- **Study period:** Jan. 1, 1772-–current

Final data

https://www.metoffice.gov.uk/hadobs/hadcet/index.html

```r
library(ggplot2)

# Download data
url <- paste0(
  "https://www.metoffice.gov.uk/hadobs/hadcet/",
  "data/meantemp_daily_totals.txt"
)
cet <- read.table(
  url,
  header = TRUE,
  stringsAsFactors = FALSE
)

# Convert date and temperature
names(cet) <- c("date", "temp")
cet$date <- as.Date(cet$date)
cet$temp <- as.numeric(cet$temp)

# Replace missing values
cet$temp[cet$temp == -99.9] <- NA

# Restrict to textbook period
cet <- subset(
  cet,
  date >= as.Date("1772-01-01") &
  date <= as.Date("2019-12-31")
)

# Plot
ggplot(cet, aes(x = date, y = temp)) +
  geom_line(color = "steelblue", linewidth = 0.2) +
  labs(
    title = "Central England Daily Mean Temperature (1772-2019)",
    x = "Calendar time",
    y = "Daily mean temperature (°C)"
  ) +
  theme_bw()
```

Assumptions

Time-series methods must account for **temporal dependence**:

$$\mathrm{Cov}(Y_t, Y_{t+h}) \neq 0, \quad h \in \mathbb{Z}^+$$

- $\mathbb{Z}^+$: the set of all the positive integers
- For a *stationary* process, the statistical behavior does not change with calendar time; in particular, the above covariance depends on the **lag** h, rather than the date t.

Statistical models used by [Textbook]: time-domain analysis (Chapter 6)

Decompose the observed time series into a mean, a temporally correlated process, and independent noise:

$$Y_t = m(t) + Z_t + \varepsilon_t,$$

- $m(t)$: mean process
- Z_t: zero-mean, a temporally correlated process
- ε_t: independent noise

For $m(t)$, [Textbook] considers the harmonic regression:

$$m(t) = \beta_0 + a_1 \cos(2\pi t/\tau) + b_1 \sin(2\pi t/\tau) + a_2 \cos(4\pi t/\tau) + b_2 \sin(4\pi t/\tau)$$

- $\tau \approx 365.2425$ days
- First harmonic: broad annual cycle
- Second harmonic: additional flexibility in the seasonal shape

One may assume the temporally correlated component follows a stationary Gaussian process with a covariance function $C_\theta(h)$:

$$Z_t \sim \mathrm{GP}(0, C(h)),$$

where $C(h) = \mathrm{Cov}(Z_t, Z_{t+h})$ merely relies on the time lag h.

Gaussian autoregressive moving-average (ARMA) models are special cases of stationary Gaussian processes

Consider the Gaussian MA(1) model:

$$Z_t = u_t + \theta u_{t-1}, \qquad u_t \overset{iid}{\sim} N(0, \sigma_u^2).$$

It is equivalent to $Z_t \sim \mathrm{GP}(0, C(h))$ such that

$$C(h) = \begin{cases} (1+\theta^2)\sigma_u^2, & h = 0, \\ \theta\sigma_u^2, & |h| = 1, \\ 0, & |h| \geq 2. \end{cases}$$

Another familiar example of a stationary Gaussian process is Gaussian AR(1):

$$Z_t = \theta Z_{t-1} + u_t, \qquad u_t \overset{iid}{\sim} N(0, \sigma_u^2),$$

where $|\theta| < 1$. It is equivalent to $Z_t \sim \mathrm{GP}(0, C(h))$ such that

$$C(h) = \sigma_Z^2 \theta^{|h|}$$

with $\sigma_Z^2 = \sigma_u^2/(1-\theta^2)$.

Potential statistical models: frequency-domain analysis (Chapter 7)

Again, decompose the observed time series into

$$Y_t = m(t) + Z_t + \varepsilon_t,$$

where

- $m(t)$: mean process, modeled using harmonic regression.
- Z_t: zero-mean, stationary Gaussian process.
- ε_t: independent Gaussian noise, independent of Z_t.

In time-domain analysis, temporal dependence is characterized by the **spectral density** rather than the covariance function $C(h)$. The spectral density describes how variation is distributed across frequencies.

The covariance function $C(h)$ and its spectral density $f_Z(\omega)$ are Fourier-transform pairs:

- For given $C(h)$, the Fourier transform yields

$$f_Z(\omega) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} C(h) e^{-i\omega h};$$

  - $\omega \in [-\pi, \pi]$
- If $f_Z(\omega)$ is known, the inverse Fourier transform gives:

$$C(h) = \int_{-\pi}^{\pi} f_Z(\omega) e^{i\omega h} \, \mathrm{d}\omega.$$

Key idea: Time-domain and frequency-domain models describe the same temporal dependence from different perspectives.

Gaussian ARMA models also have spectral representations

Consider the stationary Gaussian MA(1) model:

$$Z_t = u_t + \theta u_{t-1}, \qquad u_t \overset{iid}{\sim} N(0, \sigma_u^2).$$

Its spectral density is

$$f_Z(\omega) = \frac{\sigma_u^2}{2\pi}(1 + \theta^2 + 2\theta \cos \omega)$$

for $\omega \in [-\pi, \pi]$.

Another familiar example is stationary Gaussian AR(1):

$$Z_t = \theta Z_{t-1} + u_t, \qquad |\theta| < 1.$$

Its spectral density is

$$f_Z(\omega) = \frac{\sigma_u^2}{2\pi(1 + \theta^2 - 2\theta \cos \omega)}$$

Interpretation:

- The covariance function describes correlation at different time lags.
- The spectral density describes variation at different frequencies.

Parametric spectral model used by [Textbook]

Recall the decomposition

$$Y_t = m(t) + Z_t + \varepsilon_t.$$

[Textbook] defines the discrete Fourier transform:

$$\widetilde{Y}(\omega) = \sum_{t=1}^{n} Y_t \exp(i\omega t) = \widetilde{m}(\omega) + \widetilde{Z}(\omega) + \widetilde{\varepsilon}(\omega),$$

- ω takes values at n equally spaced Fourier frequencies in $(-\pi, \pi]$.

- $\widetilde{m}(\omega) = \sum_{t=1}^{n} m(t) \exp(i\omega t)$
- $\widetilde{Z}(\omega) = \sum_{t=1}^{n} Z_t \exp(i\omega t)$
- $\widetilde{\varepsilon}(\omega) = \sum_{t=1}^{n} \varepsilon_t \exp(i\omega t)$

Thus, the discrete Fourier transform of the observed series is the sum of the discrete Fourier transforms of its three components.

The spectral power at frequency ω is

$$P(\omega) = |\widetilde{Y}(\omega)|^2.$$

Greater spectral power at a particular frequency means that fluctuations at that frequency contribute more to the variation in the time series.

Taking expectation of the spectral power,

$$\mathrm{E}\{P(\omega)\} = |\widetilde{m}(\omega)|^2 + \mathrm{E}|\widetilde{Z}(\omega)|^2 + \mathrm{E}|\widetilde{\varepsilon}(\omega)|^2.$$

Key idea: Independent components contribute additively to the expected spectral power.

For the Central England temperature data, [Textbook] specifies a three-component model for the expected spectral power:

$$\mathrm{E}|P(\omega)|^2 \approx \underbrace{\left|\sum_{t=1}^{n} m(t)e^{i\omega t}\right|^2}_{\text{Mean component}} + \underbrace{n\sigma_1^2 e^{-\sqrt{|\lambda\omega|}}}_{\text{Correlated component}} + \underbrace{n\sigma_0^2}_{\text{Independent noise}}.$$

- σ_0^2: variance of independent noise.
- σ_1^2: scale of the correlated component's spectral power.
- λ: temporal-scale parameter controlling spectral decay.

Time-domain vs. frequency-domain interpretations

Both models describe the same temperature variation from different perspectives.

Component	Time domain	Frequency domain
Mean $m(t)$	Temperature follows a predictable annual pattern, with warmer summers and colder winters.	Spectral power is concentrated near annual and semiannual frequencies.
Correlated process Z_t	Temperatures on nearby days tend to be similar. Unusually warm or cold conditions may persist for several days.	Spectral power is greater at low frequencies (smaller $\|\omega\|$) and decreases as $\|\omega\|$ increases.
Independent noise ε_t	Random day-to-day deviations that are independent across days.	Contributes the same expected spectral power at all frequencies, regardless of ω.

Key takeaway:

Time-domain models focus on how observations change and relate to one another over time, making them especially useful for predicting values at specific times.

Frequency-domain models focus on how variation is distributed across frequencies, making them especially useful for identifying strong repeating patterns (e.g., seasonal cycles) and understanding slow versus rapid fluctuations.