Statistical Examination of Temperature Shifts: Sydney Airport 1973 vs. 2003

Tianyi Wang

7 Mar 2025

1 Introduction

In today's era of heightened environmental awareness, understanding urban temperature trends is crucial for policymakers, researchers, and the general public. This report presents a comprehensive statistical analysis of weekly temperature data in Sydney, comparing measurements from 1973 and 2003. Our analysis employs robust statistical techniques to examine these patterns, aiming to contribute valuable insights into how climatic shifts may influence urban life, airport operations and infrastructure planning over time.

2 Data

This dataset contains 52 rows and 3 columns.

- Week: An integer from 1 to 52, representing the week number of the year.
- Temp1973: The lowest temperature is 18.5°C and the highest is 39.7°C. The median temperature is 27.7°C , with an average of 27.04°C.
- Temp2003: The lowest temperature is slightly higher at 19.1°C, and the highest is 41.1°C. The median temperature is 26.1°C, with an average of 27.41°C.

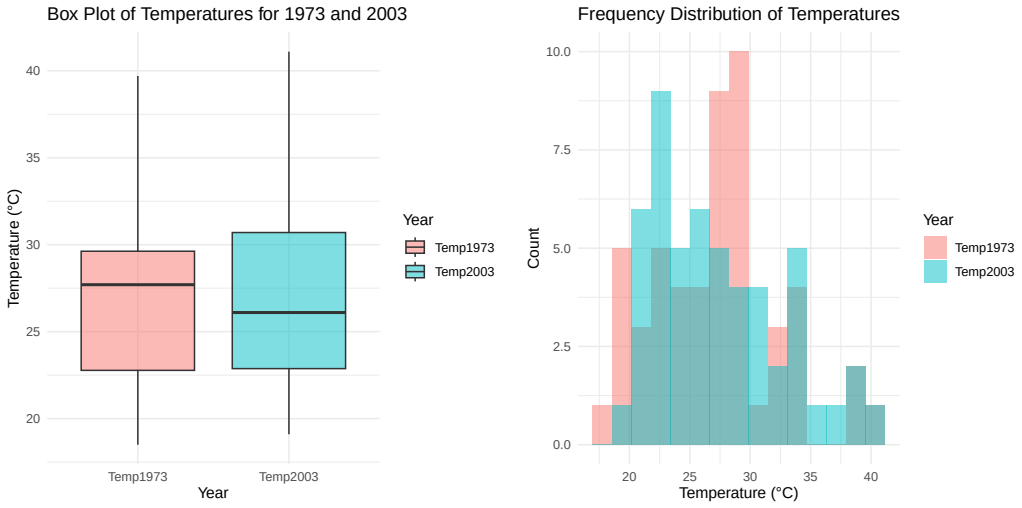

Figure 1: Box Plot of Temperatures for 1973 and 2003 and Frequency Distribution of Temperatures

Compare the temperatures in two years: The box plot shows that the typical (median) temperature in 1973 is a bit higher than in 2003. However, the overall averages are very close, meaning both years experienced similar general temperatures.

The histogram reveals how often temperatures occurred within certain ranges. Both years have most temperatures clustering in the 23°C to 30 °C. While the general shape is similar, the 2003 data seems to extend a bit more towards the higher temperatures, meaning that extremely warm days were a little more common in 2003 compared to 1973.

3 Hypothesis formulation and testing procedure

3.1 Hypothesis Formulation and Assumptions

We consider the paired temperature observations from 1973 and 2003. Let

$$D_i = \text{Temp1973}_i - \text{Temp2003}_i$$

represent the difference in temperature for the ith paired observation. The Wilcoxon signed rank test is employed to assess whether there is a systematic (median) difference in these paired measurements.

The null hypothesis is:

$$H_0 : \text{Median}(D) = 0,$$

which implies that there is no location shift in the temperature distribution between 1973 and 2003.

In contrast, the alternative hypothesis is:

$$H_1 : \text{Median}(D) \neq 0,$$

indicating that a significant shift in the paired temperature values exists.

Statistical Assumptions and Caveats:

- **Paired Observations:** Each observation in 1973 is paired with a corresponding observation in 2003.
- **Independence:** The pairs are assumed to be independent of one another.
- **Measurement Scale:** The differences are measured on a continuous scale, allowing for non-parametric ranking.
- **Symmetry of Differences:** The test assumes that the distribution of the differences is symmetric about the median. Violations of this assumption may affect the test's validity.
- **Handling of Ties and Zero Differences:** Tied values and zero differences are handled by the test, though an excessive number of ties can reduce its power.

3.2 Hypothesis Testing Procedure and Conclusion

We employed the Wilcoxon signed rank test to evaluate the null hypothesis. In R, the test was executed as follows:

```
wilcox_result <- wilcox.test(data$Temp1973, data$Temp2003, paired = TRUE, alternative = "
print(wilcox_result)
```

The output was:

```
Wilcoxon signed rank test with continuity correction

data:  data$Temp1973 and data$Temp2003
V = 633, p-value = 0.6132
alternative hypothesis: true location shift is not equal to 0
```

3.3 Conclusion

Since the computed p-value (0.6132) exceeds 0.1, we fail to reject the null hypothesis. This indicates that there is no statistically significant difference between the paired temperature observations in 1973 and 2003, suggesting that no notable shift in the central tendency or overall distribution has occurred over the 30-year period, subject to the assumptions outlined above.

4 Smooth the data

4.1 Kernel Smoothing Using the Nadaraya–Watson Estimator

We apply the Nadaraya–Watson kernel smoother to uncover the underlying temperature trends in the 1973 and 2003 datasets.

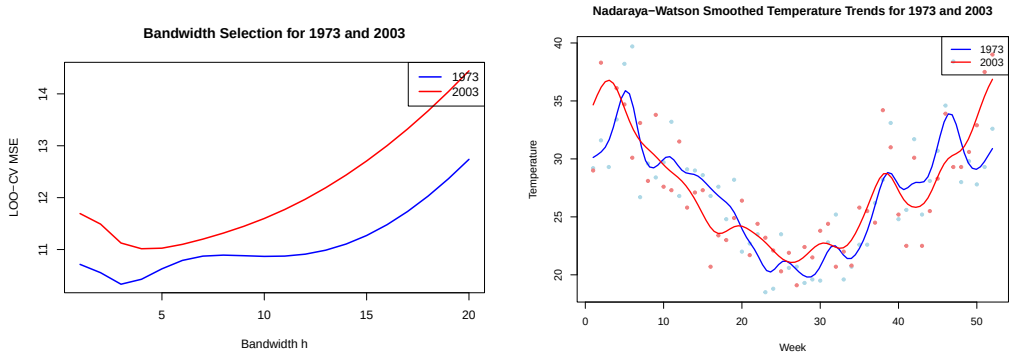

Figure 2: LOOCV for the bandwith of the Nadaraya-Watson smoother on data and Nadaraya-Watson Smoothed Temperature Trends for 1973 and 2003

Justification of Choices

Kernel Choice: We selected the Gaussian kernel because its smooth, exponentially decaying weights ensure that all observations contribute to the estimate, with closer points having a higher influence. This property is especially advantageous for modeling temperature data, where gradual changes are expected. The smooth weighting also prevents abrupt jumps in the smoothed curve, which aligns well with the continuous nature of temperature variations.

Bandwidth Selection: We use MSE to determine bandwidth. In LOO-CV, each data point is sequentially omitted, and the model is fitted using the remaining data. The prediction error for the omitted point is then computed. The average of these errors (the MSE) serves as an evaluation metric. Selecting the bandwidth that minimizes the MSE implies that the chosen h offers the best predictive performance, balancing bias

and variance effectively.The bandwidth h is critical in controlling the smoothness of the resulting curve. As illustrated in Figure 2, varying h directly influences the trade-off between bias and variance:

- A smaller h results in a curve that closely follows the data, which might lead to overfitting by capturing noise.
- A larger h produces an oversmoothed curve that may obscure significant local trends.

In our analysis, we chose $h = 3$ for the 1973 dataset and $h = 4$ for the 2003 dataset. These values provide a balanced compromise by capturing the overall trend while minimizing the impact of random fluctuations.

4.2 Smoothing Spline Smoother

The graph displays weekly temperature trends for 1973 and 2003 using a smoothing spline. This method fits a smooth line through noisy data, reducing random fluctuations while keeping the main seasonal patterns clear. It makes it easier to see the overall trend in temperatures for both years.

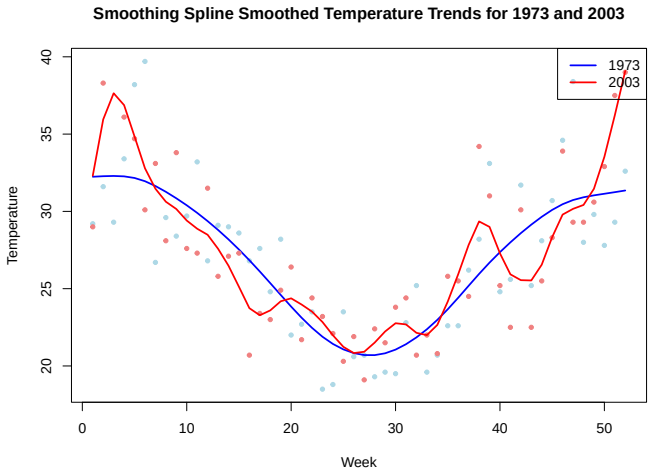

Figure 3: Smoothing Spline Smoothed Temperature Trends for 1973 and 2003

4.3 Comparison of the Smoothers

Nadaraya–Watson Smoother:

- Provides a balanced fit that captures the overall seasonal trends in both 1973 and 2003.
- Avoids overfitting by smoothing out noise effectively, resulting in a more consistent curve.

Smoothing Spline Smoother:

- For 1973, the smoothing spline appears to overfit the data, producing an overly wiggly curve that captures too much noise.
- For 2003, the curve is not smooth enough, potentially missing key trend features.

Overall Comparison:

- The Nadaraya-Watson method works better in these examples, as it maintains a more appropriate balance between bias and variance across both datasets.

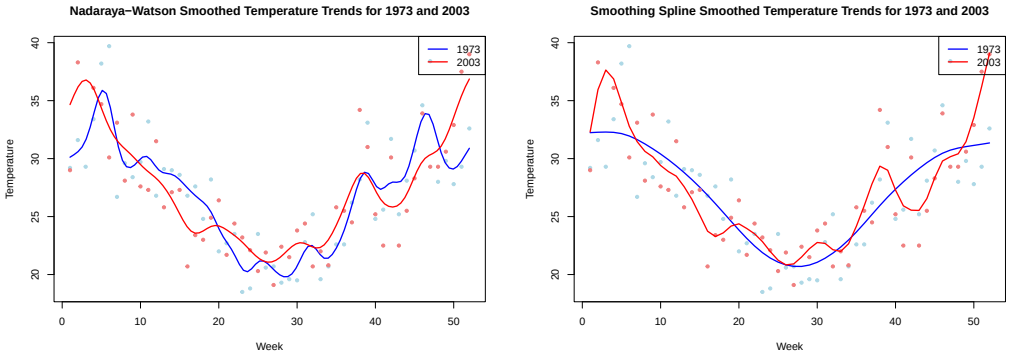

Figure 4: Comparison of two smoother

R code is atached below

project3

1073354

2025-03-03

```
#data
# Load necessary libraries
library(tidyr)
library(ggplot2)
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Load the dataset
data <- read.csv("~/Downloads/sydney-temperatures.csv", stringsAsFactors = TRUE)
attach(data)

# Summary statistics
summary(data)
```

```
##       Week          Temp1973        Temp2003
##  Min.   : 1.00   Min.   :18.50   Min.   :19.10
##  1st Qu.:13.75   1st Qu.:22.77   1st Qu.:22.88
##  Median :26.50   Median :27.70   Median :26.10
##  Mean   :26.50   Mean   :27.04   Mean   :27.41
##  3rd Qu.:39.25   3rd Qu.:29.62   3rd Qu.:30.70
##  Max.   :52.00   Max.   :39.70   Max.   :41.10
```

```
# Convert the data to long format
data_long <- data %>%
  pivot_longer(cols = c(Temp1973, Temp2003), names_to = "Year", values_to = "Temperature")

# Create a frequency distribution plot (histogram) for the temperatures
frequency_plot <- ggplot(data_long, aes(x = Temperature, fill = Year)) +
  geom_histogram(alpha = 0.5, position = "identity", bins = 15) +
  labs(title = "Frequency Distribution of Temperatures",
       x = "Temperature (°C)",
       y = "Count",
       fill = "Year") +
  theme_minimal()
#save graph
ggsave("Frequency Distribution of Temperatures.pdf", height = 5, width = 5)

# Create a box plot comparing the two years
box_plot <- ggplot(data_long, aes(x = Year, y = Temperature, fill = Year)) +
  geom_boxplot(alpha = 0.5) +
  labs(title = "Box Plot of Temperatures for 1973 and 2003",
       x = "Year",
       y = "Temperature (°C)") +
  theme_minimal()

# Save the plot
ggsave("Box Plot of Temperatures for 1973 and 2003.pdf", height = 5, width = 5)
```

```
# Use Wilcoxon signed rank test: paired replicates
wilcox_result <- wilcox.test(data$Temp1973, data$Temp2003, paired = TRUE, alternative = "two.sided")
print(wilcox_result)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  data$Temp1973 and data$Temp2003
## V = 633, p-value = 0.6132
## alternative hypothesis: true location shift is not equal to 0
```

```
# ----- Nadaraya-Watson Smoother -----
# Optimized Bandwidth Selection using the entire dataset and ksmooth-based LOO-CV

# Assume 'data' is a data frame with columns Temp1973 and Temp2003.
# Use the entire dataset:
x <- 1:nrow(data)
y1973 <- data$Temp1973
y2003 <- data$Temp2003

# Define a range of candidate bandwidths
bandwidths <- seq(1, 20, by = 1)

# Function to compute the leave-one-out CV MSE using ksmooth for a given bandwidth h
loo_cv_ksmooth <- function(x, y, h) {
  n <- length(x)
  cv_errors <- sapply(1:n, function(i) {
    # Leave out the i-th observation
    x_train <- x[-i]
    y_train <- y[-i]
    # Predict the left-out point using ksmooth with the given bandwidth
    y_pred <- ksmooth(x_train, y_train, kernel = "normal", bandwidth = h, x.points = x[i])$y
    (y[i] - y_pred)^2
  })
  mean(cv_errors)
}

# Compute LOO-CV MSE for each candidate bandwidth for both 1973 and 2003 data
mse_1973 <- sapply(bandwidths, function(h) loo_cv_ksmooth(x, y1973, h))
mse_2003 <- sapply(bandwidths, function(h) loo_cv_ksmooth(x, y2003, h))

# Determine the optimal bandwidth (minimizing LOO-CV MSE) for each year
optimal_h_1973 <- bandwidths[which.min(mse_1973)]
optimal_h_2003 <- bandwidths[which.min(mse_2003)]

cat("Optimal bandwidth for 1973 (minimizing LOO-CV MSE):", optimal_h_1973, "\n")
```

```
## Optimal bandwidth for 1973 (minimizing LOO-CV MSE): 3
```

```
cat("Optimal bandwidth for 2003 (minimizing LOO-CV MSE):", optimal_h_2003, "\n")
```

```
## Optimal bandwidth for 2003 (minimizing LOO-CV MSE): 4
```

```
# Plot the LOO-CV MSE curves for both years and save the plot as a PDF file
pdf("Bandwidth_Selection_Combined.pdf", height = 5, width = 7)
plot(bandwidths, mse_1973, type = "l", lwd = 2, col = "blue",
     xlab = "Bandwidth h", ylab = "LOO-CV MSE",
     main = "Bandwidth Selection for 1973 and 2003",
     ylim = range(c(mse_1973, mse_2003)))
lines(bandwidths, mse_2003, lwd = 2, col = "red")
legend("topright", legend = c("1973", "2003"), col = c("blue", "red"), lwd = 2)
dev.off()
```

```
## quartz_off_screen
##                 2
```

```
# Apply the Nadaraya-Watson kernel smoother using the Gaussian kernel.
# The bandwidth is set to 3 and 4
smoothed_1973 <- ksmooth(x, data$Temp1973, kernel = "normal", bandwidth = 3)
smoothed_2003 <- ksmooth(x, data$Temp2003, kernel = "normal", bandwidth = 4)

# Save the Nadaraya-Watson smoothed plot to a PDF file.
pdf("NadarayaWatson_Temperature_Trends.pdf", width = 8, height = 6)
plot(x, data$Temp1973, col = "lightblue", pch = 20,
     main = "Nadaraya-Watson Smoothed Temperature Trends for 1973 and 2003",
     xlab = "Week", ylab = "Temperature")
lines(smoothed_1973$x, smoothed_1973$y, col = "blue", lwd = 2)
points(x, data$Temp2003, col = "lightcoral", pch = 20)
lines(smoothed_2003$x, smoothed_2003$y, col = "red", lwd = 2)
legend("topright", legend = c("1973", "2003"), col = c("blue", "red"), lwd = 2)
dev.off()
```

```
## quartz_off_screen
##                 2
```

```
# ----- Smoothing Spline Smoother----
# Fit smoothing splines for both temperature series using smooth.spline

# Fit smoothing spline for the 1973 temperature series
spline_1973 <- smooth.spline(x, data$Temp1973)
# Fit smoothing spline for the 2003 temperature series
spline_2003 <- smooth.spline(x, data$Temp2003)

# Predict the smoothed values for each observation
smoothed_spline_1973 <- predict(spline_1973, x)$y
smoothed_spline_2003 <- predict(spline_2003, x)$y

# Save the smoothed plot to a PDF file
pdf("SmoothingSpline_Temperature_Trends.pdf", width = 8, height = 6)
plot(x, data$Temp1973, col = "lightblue", pch = 20,
     main = "Smoothing Spline Smoothed Temperature Trends for 1973 and 2003",
     xlab = "Week", ylab = "Temperature")
lines(x, smoothed_spline_1973, col = "blue", lwd = 2)
points(x, data$Temp2003, col = "lightcoral", pch = 20)
lines(x, smoothed_spline_2003, col = "red", lwd = 2)
legend("topright", legend = c("1973", "2003"), col = c("blue", "red"), lwd = 2)
dev.off()
```

```
## quartz_off_screen
##                 2
```