BCB744: Biostatistics R Exam

Smit, A. J.

2025-05-30

About the Exam

The Biostatistics Exam will start at 8:30 on 30 May, 2025 and you have until 8:30 on 31 May, 2025 to complete it. This exam may be conducted anywhere in the world, and it will contribute 70% of the final assessment marks for the Biostatistics component of the BCB744 module.

Assessment Criteria

Your responses will be evaluated based on the following criteria:

1. **Technical Accuracy (50%)**
   - Correct application of data analyses and statistical methods. Statistical tests should address the hypotheses within what was taught in BCB744. The assessor recognises that students will not have access to the level of statistical knowledge and experience a research statistician might have. For example, in places, linear mixed models might be more suited to the questions, but students only were taught the basics about ANOVA and simple linear model (relatively simple designs). When non-parametric alternatives are required, I'll assign marks to the any statement that suggests the correct test to use, but not actually marks the execution of those tests.
   - Use of appropriate R packages, functions, and syntax (code style and liberal commenting)
   - Appropriate choice and justification of techniques, including due consideration for the assumptions of the methods used
   - Accurate calculations and results interpretation, down to small details such as how many decimal places to use
2. **Depth of Analysis (20%)**
   - Comprehensive exploration of the problem
   - Insightful interpretation of results
   - Consideration of shortfalls in the analysis (due to data limitations, assumptions, etc.), and suggestions for improvement
   - Application of out-of-the-box thinking to the problem
3. **Clarity and Communication (20%)**
   - Logical organisation of ideas, including clear section headings and subheadings
   - Clear and concise explanations at each stage of the analysis
   - Effective use of publication quality visualisations where appropriate, including all necessary annotations
   - Communication of results in a way that is appropriate for a scientific audience (e.g. a journal article)
4. **Critical Thinking Shown in Final Conclusion/Synopsis (10%)**
   - Discussion of the findings in the context of the problem (add ecological context, etc., as you deem necessary)
   - Identification of limitations
   - Discussion of assumptions
   - Consideration of broader implications

General Notes to Assessor (applies to all tasks):

- Heavily penalise untidy formatting, good document structure according to logical headings and heading hierarchies, excessive output of long, unnecessary data printouts (other than the obvious

and required used of `head()`, `tail()`, `glimpse()`, and `summary()`) that serve no purpose [-15%].
- Answers where the code gives error messages (it fails to run to provide the required output) get 0 for that question.
- Where students write long-form text feedback/answers within code blocks, where it should have been more appropriately placed within the markdown text between code blocks and presented in full sentences, should be penalised [-10% for each question where this occurs].
- Text answers written as bullet points and which lacks detailed explanatory power gets penalised [-10%].
- Untidy presentation and formatting that fails to resemble my model answers (below) penalised [-15%].

The marks indicated for each task reflect the relative weight and expected depth of your response. Focus on demonstrating both technical proficiency and conceptual understanding in your answers.

Please refer to the Assessment Policy for more information on the test format and rules.

Instructions

This is the open book assessment.

You must address all tasks in the allocated time of 24-hr. Please submit your answers in a neatly formatted .html document (produced from a Quarto document in RStudio) and submit it to the iKamva platform.

Clearly structure the document according to the task numbers, i.e., use appropriately hierarchical headings, subheadings, and sub-subheadings to structure your document logically.

Naming convention: `BCB744_Biostatistics_Prac_Exam_YourSurname.html`

Background

These data represent the aerial cover of kelp canopy in South Africa, as measured by Landsat satellites, for the period 1984 to 2024 at a quarterly interval. The intention is to understand the spatio-temporal patterns in kelp canopy cover and to explore how these patterns may be related to coastal sections and biogeographical provinces.

You are provided with two datasets at the Google Drive link emailed to you:

1. A table of 58 coastal sections (`58_sections.csv`) that partitions the South African coastline into approximately 50 km intervals. Each section is defined by a single coordinate point (latitude, longitude) representing the boundary of the section.
2. A table of the biogeographical provinces (`bioregions.csv`) that the 58 coastal sections fall within. There is one row for each of the 58 sections. For this exercise, the biogeographical classification by Professor John Bolton is of interest.
3. A netCDF file (`kelpCanopyFromLandsat_SouthAfrica_v04.nc`) of kelp sampling locations and aerial cover data – these are presented as various variables at grid points across time.

Task 1: Initial Processing

- **[Task Weight: 10%]**
- **[Components (1) and (2) marked on a 0–100 scale, then scaled to equal proportions of the Task Weight of 10%]**

You are provided with a NetCDF file that contains satellite-derived measurements of kelp canopy area across the South African coastline from 1984 to 2024, sampled quarterly. Each observation corresponds to a grid cell at a specific time point.

1. Read the kelp canopy area, time, location (latitude/longitude), and satellite pass data from the NetCDF file. Once unpacked, it contains over 5 million rows. Your processing workflow will include:

- extracting data from the netCDF file where `area` and `passes` are variables defined over 3D space (`longitude`, `latitude`, and `time`); and
- using functions such as `tidync::hyper_tibble()` or `ncdf4::ncvar_get()` to read these values.

Answer

Note to assessor: Students may have used any of a number of NetCDF targeted packages, such as **tidync**, **stars**, or **terra**. Below I use **ncdf4**.

```
library(tidyverse)
library(ncdf4)
library(geosphere)
library(mgcv)

nc_path <- "../data/Kelpwatch/"
nc_file <- paste0(nc_path, "kelpCanopyFromLandsat_SouthAfrica_v04.nc")
nc <- nc_open(nc_file)

area <- ncvar_get(nc, "area")
time <- ncvar_get(nc, "time")
year <- ncvar_get(nc, "year")
quarter <- ncvar_get(nc, "quarter")
latitude <- ncvar_get(nc, "latitude")
longitude <- ncvar_get(nc, "longitude")
passes <- ncvar_get(nc, "passes")

nc_close(nc)
```

POSIX timestamps rather than raw seconds:

```
time <- as.POSIXct(time, origin = "1970-01-01", tz = "UTC")
```

Create index vectors:

```
time_idx <- seq_along(time) # 1…ntime
loc_idx <- seq_along(latitude) # 1…nloc
```

2. Restructure the data into a data.table or data.frame:

- the data should have six columns: `longitude`, `latitude`, `year`, `quarter`, `area`, and `passes`;
- each row should correspond to a unique pixel in space-time (i.e., one location at one time point); and
- note that the `time` variable in the netCDF file is in numeric format (e.g., days since origin, where `origin = "1970-01-01"`), so you'll have to convert it to POSIX timestamps using appropriate tools (e.g., `as.POSIXct()`).

If you are unable to read the NetCDF file, you may request access to a processed version of this file (in long CSV format) from me, but you'll be penalised by 10% if you do so.

Answer

Note to assessor: Find some evidence for the successful execution of the above instructions such as the presence of the required dataframe columns, correct conversion of the time variable, and so on.

Cartesian join of (`time_idx` × `loc_idx`), in an order that matches how `as.vector()` will flatten a $nloc \times ntime$ matrix. Then, add the flattened area and the six columns and select the required variables and make a long:

```
# Create Cartesian product of indices
long_df <- crossing(
  time_idx = time_idx,
  loc_idx = loc_idx
) |>
  mutate(
    area = as.vector(area),
    time = time[time_idx],
    year = year[time_idx],
    quarter = quarter[time_idx],
    latitude = latitude[loc_idx],
    longitude = longitude[loc_idx],
    passes = passes[loc_idx]
  ) |>
  select(area, time, year, quarter, latitude, longitude, passes)

summary(long_df)
```

```
      area             time                          year         quarter
```

```
 Min.   :  0.0     Min.   :1984-04-01 00:00:00   Min.   :1984   Min.   :1.000
 1st Qu.:  0.0     1st Qu.:1996-01-01 00:00:00   1st Qu.:1996   1st Qu.:1.000
 Median :  0.0     Median :2005-07-01 00:00:00   Median :2005   Median :2.000
 Mean   :190.3     Mean   :2005-05-24 09:13:06   Mean   :2005   Mean   :2.477
 3rd Qu.:324.0     3rd Qu.:2015-01-01 00:00:00   3rd Qu.:2015   3rd Qu.:3.000
 Max.   :900.0     Max.   :2024-04-01 00:00:00   Max.   :2024   Max.   :4.000
 NA's   :1061652
    latitude        longitude         passes
 Min.   :-34.83   Min.   :14.83   Min.   :0.000
 1st Qu.:-34.66   1st Qu.:18.12   1st Qu.:1.000
 Median :-34.35   Median :18.48   Median :1.000
 Mean   :-33.42   Mean   :18.55   Mean   :1.399
 3rd Qu.:-32.98   3rd Qu.:19.40   3rd Qu.:2.000
 Max.   :-25.44   Max.   :19.97   Max.   :4.000
```

Task 2: Exploratory Data Analysis

- **[Task Weight: 10%]**
- **[Tasks 2.1, 2.2, and 2.3, each marked on a 0–100 scale, then scaled to equal proportions of the Task Weight of 10%]**

2.1 Weighted Mean Time Series

1. For each `year` and `quarter` combination:

- compute the weighted mean of the kelp canopy `area` across all locations, using the number of satellite `passes` as weights;
- exclude observations where `passes = 0` or `area` is `NA`; and
- plot the resulting time series of weighted mean kelp area, using i) `quarters` on the x-axis, and ii) a continuous `time` index from 1984–2024.

Answer

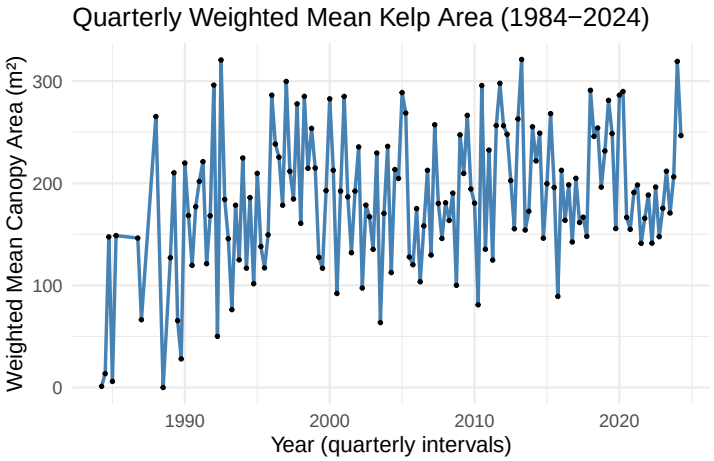
Note to assessor: If the student produced the figure exactly as I have it below, this question can get full marks. Else, assess the analysis workflow and assign marks in accordance with the portions of the script that are correctly executed.

```r
# Filter out missing values and zero passes at the start
quarter_means <- long_df |>
  filter(!is.na(area), !is.na(passes)) |>
  group_by(year, quarter) |>
  summarise(
    weighted_area = weighted.mean(area, w = passes, na.rm = TRUE),
    .groups = "drop"
  ) |>
  filter(!is.na(weighted_area))

summary(quarter_means)
```

```
      year         quarter      weighted_area
 Min.   :1984   Min.   :1.000   Min.   :  0.0
 1st Qu.:1996   1st Qu.:1.000   1st Qu.:145.8
 Median :2005   Median :2.000   Median :186.0
 Mean   :2005   Mean   :2.477   Mean   :185.1
 3rd Qu.:2014   3rd Qu.:3.000   3rd Qu.:232.1
 Max.   :2024   Max.   :4.000   Max.   :321.2
```

```r
library(tidyverse)

# Filter out invalid observations
quarter_means <- long_df |>
  filter(!is.na(area), !is.na(passes), passes > 0) |>
  group_by(year, quarter) |>
  summarise(
    weighted_area = weighted.mean(area, w = passes, na.rm = TRUE),
    .groups = "drop"
  ) |>
  filter(!is.na(weighted_area)) |>
  mutate(
    quarter = as.integer(quarter),
```

```
    time_index = year + (quarter - 1) / 4
  )

# Plot quarterly weighted mean time series
ggplot(quarter_means, aes(x = time_index, y = weighted_area)) +
  geom_line(color = "steelblue", size = 0.8) +
  geom_point(color = "black", size = 0.6) +
  labs(
    title = "Quarterly Weighted Mean Kelp Area (1984-2024)",
    x = "Year (quarterly intervals)",
    y = "Weighted Mean Canopy Area (m²)"
  ) +
  theme_minimal()
```



2. Compute the weighted mean `area` at each unique (`longitude`, `latitude`) pixel across `time`. Then:

- select a random sample of 100 pixels;
- for each sampled pixel, extract the full time series of weighted mean area;
- plot all 100 time series in a single panel (overlayed), using semi-transparent lines; and
- label axes appropriately.

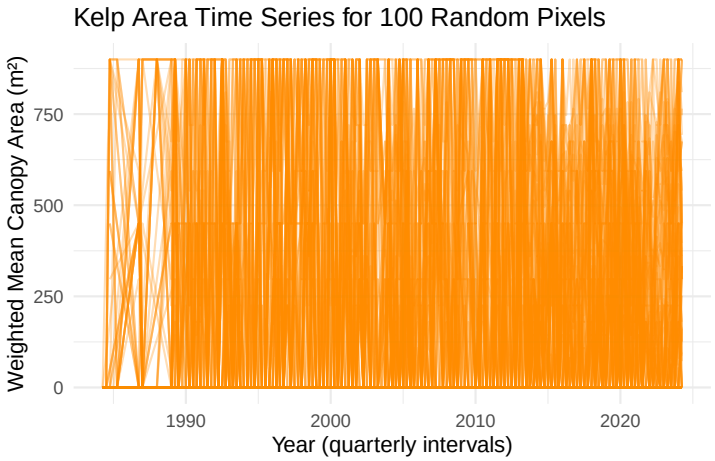
Answer

Note to assessor: Again, if they produced the figure exactly as below, this question can get full marks. Else, assess the analysis workflow and assign marks in accordance with the portions of the script that are correctly executed.

```
pixel_means <- long_df |>
  filter(!is.na(area), !is.na(passes), passes > 0) |>
  group_by(year, quarter, longitude, latitude) |>
  summarise(
    weighted_area = weighted.mean(area, w = passes, na.rm = TRUE),
    .groups = "drop"
  )

summary(pixel_means)
```

```
      year         quarter        longitude        latitude
 Min.   :1984   Min.   :1.000   Min.   :15.08   Min.   :-34.83
 1st Qu.:1997   1st Qu.:2.000   1st Qu.:18.33   1st Qu.:-34.67
 Median :2006   Median :3.000   Median :18.84   Median :-34.36
 Mean   :2006   Mean   :2.503   Mean   :18.64   Mean   :-33.62
 3rd Qu.:2016   3rd Qu.:3.000   3rd Qu.:19.41   3rd Qu.:-33.95
 Max.   :2024   Max.   :4.000   Max.   :19.97   Max.   :-26.48
 weighted_area
 Min.   :  0.0
 1st Qu.:  0.0
 Median :  0.0
 Mean   :196.1
 3rd Qu.:360.0
 Max.   :900.0
```

```r
# Step 1: Calculate weighted mean per pixel over time
pixel_means <- long_df |>
  filter(!is.na(area), !is.na(passes), passes > 0) |>
  group_by(year, quarter, longitude, latitude) |>
  summarise(
    weighted_area = weighted.mean(area, w = passes, na.rm = TRUE),
    .groups = "drop"
  ) |>
  mutate(time_index = year + (quarter - 1) / 4)

# Step 2: Identify 100 unique (lon, lat) pairs
set.seed(42)
sample_pixels <- pixel_means |>
  distinct(longitude, latitude) |>
  sample_n(100)

# Step 3: Filter time series for the sampled pixels
pixel_timeseries_sample <- pixel_means |>
  inner_join(sample_pixels, by = c("longitude", "latitude"))

# Step 4: Plot all 100 pixel time series
ggplot(pixel_timeseries_sample,
       aes(x = time_index, y = weighted_area,
           group = interaction(longitude, latitude))) +
  geom_line(alpha = 0.3, color = "darkorange") +
  labs(
    title = "Kelp Area Time Series for 100 Random Pixels",
    x = "Year (quarterly intervals)",
    y = "Weighted Mean Canopy Area (m²)"
  ) +
  theme_minimal()
```




```r
# This figure is uninterpretable...
```

2.2 Summary Statistics

1. Using the weighted data prepared for each `year` and `quarter` combination (prepared in 2.1.1), compute and report summary statistics for the levels of temporal aggregation:
   - by `year`;
   - by `quarter`;
   - by `year/quarter` combination;

- include: weighted mean, median, standard deviation, interquartile range, skewness, and kurtosis; and
- comment on the appropriateness of each statistic for these data, and justify your choices in light of the data distribution.

Answer

Note to assessor: Since this question can be objectively assessed relative to the expectations (specific reporting of correct summary stats per each level of aggregation as I have them below), marks can simply

be assigned based on the correct reporting of the summary statistics.

```
library(e1071)  # for skewness and kurtosis

# Yearly aggregation
year_stats <- quarter_means |>
  group_by(year) |>
  summarise(
    mean = mean(weighted_area, na.rm = TRUE),
    median = median(weighted_area, na.rm = TRUE),
    sd = sd(weighted_area, na.rm = TRUE),
    iqr = IQR(weighted_area, na.rm = TRUE),
    skew = skewness(weighted_area, na.rm = TRUE),
    kurt = kurtosis(weighted_area, na.rm = TRUE)
  )
year_stats
```

```
# A tibble: 41 x 7
        year  mean median    sd   iqr     skew   kurt
   <int[1d]> <dbl>  <dbl> <dbl> <dbl>    <dbl>  <dbl>
 1      1984  54.1   13.6  81.2  73.2   0.375   -2.33
 2      1985  77.4   77.4 101.   71.4   0       -2.75
 3      1986 146.   146.   NA     0   NaN      NaN
 4      1987  66.3   66.3  NA     0   NaN      NaN
 5      1988 133.   133.  188.  133.    0       -2.75
 6      1989 108.    96.3  79.7  91.8   0.250   -2.04
 7      1990 171.   173.   41.1  31.5  -0.0806  -1.89
 8      1991 178.   185.   43.8  50.2  -0.272   -2.03
 9      1992 213.   240.  124.  152.   -0.329   -2.05
10      1993 131.   135.   42.8  41.1  -0.191   -1.95
# i 31 more rows
```

```
# Quarterly aggregation (across all years)
quarter_stats <- quarter_means |>
  group_by(quarter) |>
  summarise(
    mean = mean(weighted_area, na.rm = TRUE),
    median = median(weighted_area, na.rm = TRUE),
    sd = sd(weighted_area, na.rm = TRUE),
    iqr = IQR(weighted_area, na.rm = TRUE),
    skew = skewness(weighted_area, na.rm = TRUE),
    kurt = kurtosis(weighted_area, na.rm = TRUE)
  )
quarter_stats
```

```
# A tibble: 4 x 7
  quarter  mean median    sd   iqr   skew    kurt
    <int> <dbl>  <dbl> <dbl> <dbl>  <dbl>   <dbl>
1       1  216.   220.  66.6  83.4 -0.898  0.922
2       2  185.   204.  74.2 114.  -0.361 -0.598
3       3  172.   179.  71.2  85.6 -0.326 -0.0222
4       4  166.   166.  49.7  46.2  0.177  1.32
```

```
# Year-quarter combination (already in quarter_means)
year_quarter_stats <- quarter_means |>
  mutate(year_quarter = paste0(year, " Q", quarter)) |>
  summarise(
    mean = mean(weighted_area, na.rm = TRUE),
    median = median(weighted_area, na.rm = TRUE),
    sd = sd(weighted_area, na.rm = TRUE),
    iqr = IQR(weighted_area, na.rm = TRUE),
    skew = skewness(weighted_area, na.rm = TRUE),
    kurt = kurtosis(weighted_area, na.rm = TRUE)
  )
year_quarter_stats
```

```
# A tibble: 1 x 6
   mean median    sd   iqr   skew     kurt
  <dbl>  <dbl> <dbl> <dbl>  <dbl>    <dbl>
1  185.   186.  68.3  86.2 -0.323 -0.00284
```

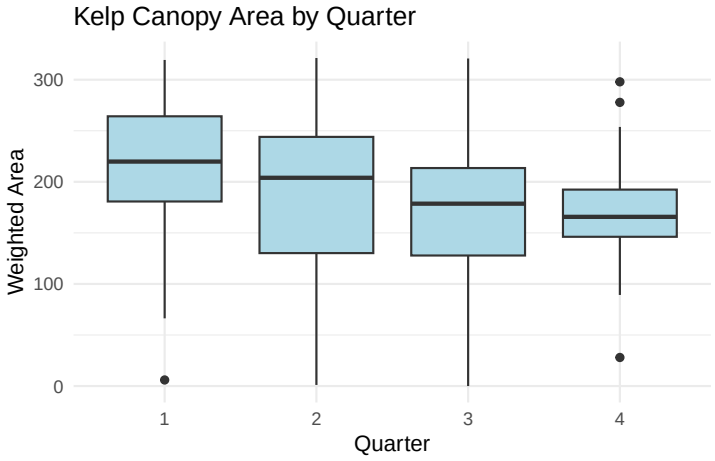
2. Create visualisations (e.g. boxplots, violin plots, histograms) to support your interpretations.

Answer

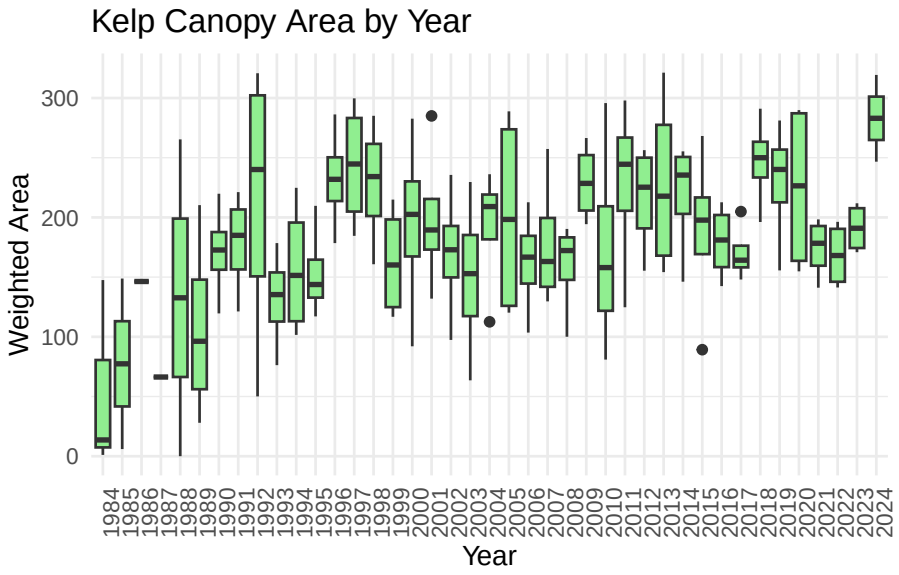
Note to assessor: These or similar figures are required, showing the seasonal effects, the annual effect,

a histogram of the frequency of the weighted areas, and something similar to the density distribution (or a frequency histogram) that shows distribution within a quarter. I'd also accept variations of these plots, as long as they speak to the nature of the data being assessed. They need to capture the feature of the summary statistics that I have above.

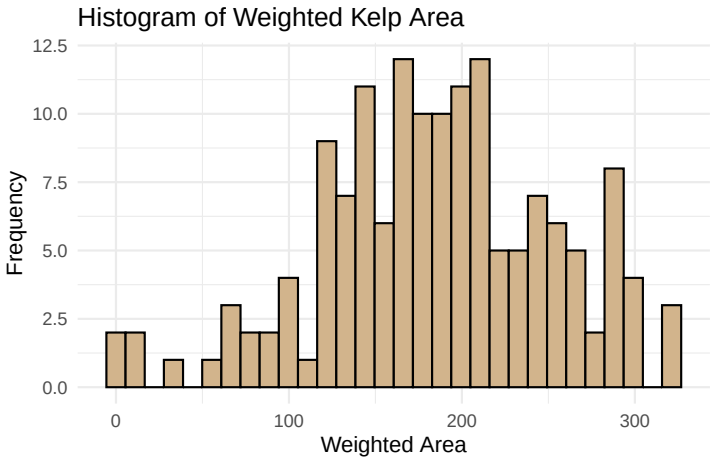
```
# Boxplot by quarter
ggplot(quarter_means, aes(x = factor(quarter), y = weighted_area)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Kelp Canopy Area by Quarter",
       x = "Quarter", y = "Weighted Area") +
  theme_minimal()
```



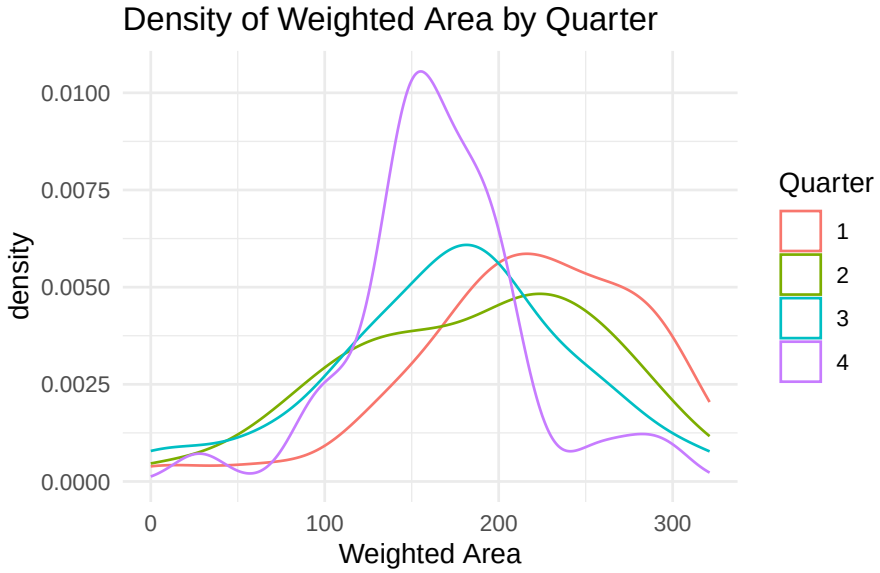

```
# Boxplot by year
ggplot(quarter_means, aes(x = factor(year), y = weighted_area)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Kelp Canopy Area by Year",
       x = "Year", y = "Weighted Area") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```




```
# Histogram of all weighted areas
ggplot(quarter_means, aes(x = weighted_area)) +
  geom_histogram(bins = 30, fill = "tan", color = "black") +
  labs(title = "Histogram of Weighted Kelp Area",
       x = "Weighted Area", y = "Frequency") +
  theme_minimal()
```




```r
# Density plot by quarter
ggplot(quarter_means, aes(x = weighted_area, color = factor(quarter))) +
  geom_density() +
  labs(title = "Density of Weighted Area by Quarter",
       x = "Weighted Area", color = "Quarter") +
  theme_minimal()
```



3. Based on these, discuss any discernible temporal trends (e.g. decadal increases/decreases) and seasonal patterns (quarterly effects).

Answer

Note to assessor: Some variation of this:

Temporal variation in kelp canopy area, as observed in the quarterly weighted means from 1984 to 2024, exhibits a broadly increasing trend until the early-2000s, followed by more irregular fluctuations at a higher weighted canopy area than at the start of the time series. The yearly summary statistics confirm this, and show rising means and medians through the late 1990s into the early 2000s, peaking during the 2010s. Although values for some early years (e.g., 1986, 1987) suffer from missing data (e.g., NA for SD and skewness), the central decades present coherent patterns.

The skewness values tend to hover close to zero, with mild left-skew in later years, suggesting distributions that are not heavily distorted but do include more frequent lower outliers in certain quarters. The kurtosis estimates, consistently below 0 across many years, indicate relatively flat distributions with light tails, again pointing to the presence of moderate, widespread values and a paucity of extreme outliers in the tails.

The quarterly summaries reveal a seasonal signal: Q1 (January–March) is associated with the highest

mean kelp area (216 km²), while Q4 (October–December) shows the lowest (166 km²). This likely reflects seasonal growth dynamics, perhaps linked to photoperiod, wave exposure, or nutrient regimes (factors known to modulate canopy) forming macroalgae. The density plot by quarter reinforces this picture: Q1's density curve is skewed leftward, indicative of a modal concentration around high values, whereas Q4 shows more compact distributions with lower peaks.

The boxplots by year reinforce this temporal variation. Inter-annual variation is relatively low during some periods (e.g., 1990s) but expands considerably in the 2010s and 2020s, suggesting a greater degree of spatial heterogeneity or more frequent episodes of extreme coverage. Notably, the increase in spread does not always correspond with a decline in median or mean area, implying that while certain coastal sectors may experience loss, others persist or expand.

The histogram of weighted area suggests a right-skewed global distribution when pooled over time, driven by many instances of low or zero canopy and fewer but notable high-coverage events.

So, these patterns suggest:

- A long-term trend toward increased kelp canopy extent over the study period, potentially stabilising or fragmenting in recent years;
- Clear intra-annual (seasonal) variation, with Q1 consistently supporting greater canopy development than later quarters;
- Increasing variability over time, which might reflect changing climate regimes, hydrodynamic forcing, or anthropogenic disturbance gradients.

These findings provide a strong descriptive foundation for the inferential modelling that follows.

2.3 Observation Density Map

Create a map plotting each observed pixel location (defined by `longitude` × `latitude`):

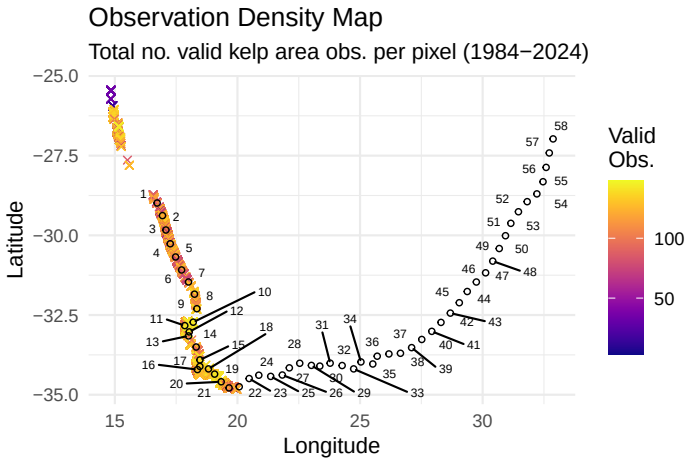
- colour each pixel by the total number of valid observations (i.e., non-NA values of `area`) across all `time` points;
- overlay the 58 coastal sections as reference points or lines, numbered from west (1) to east (58); and
- use an appropriate geographic projection and include a legend.

Answer

Note to assessor: I'm not concerned too much about having a coastline as the data clearly show the coastal profile, but, of course, having the coastline as well is great too (maybe worth a mark or three extra). I'd also be okay if the figure focuses only on sections 1-22 where kelp is present. Again, this is a fairly objective question, so marks can be assigned based on the correct figure (full marks for all the labels, etc., correctly applied as per standards of scientific publications).

```
library(ggrepel)  # for labelling points
library(viridis)  # for color scale

# Load the long_df and coastal sections
# Assume long_df has already been created from earlier tasks
# Load the 58 sections file (replace with actual path if needed)
sections_df <- read.csv("../data/Kelpwatch/58_sections.csv")

# 1. Count valid observations per pixel
obs_counts <- long_df |>
  filter(!is.na(area)) |>
  group_by(longitude, latitude) |>
  summarise(
    n_obs = n(),
    .groups = "drop"
  )

# 2. Prepare coastal sections data
sections <- sections_df |>
  mutate(section_id = row_number())

# 3. Plot
ggplot() +
  geom_point(
    data = obs_counts, shape = 4,
    aes(x = longitude, y = latitude, colour = n_obs)
```

```
  ) +
scale_colour_viridis(name = "Valid\nObs.", option = "plasma") +
geom_point(
  data = sections,
  color = "black", size = 1.1, shape = 1,
  aes(x = Longitude, y = Latitude)
  ) +
geom_text_repel(
  data = sections,
  size = 2,
  color = "black",
  max.overlaps = 58,
  aes(x = Longitude, y = Latitude, label = section_id)
) +
coord_fixed(1.3) +
labs(
  title = "Observation Density Map",
  subtitle = "Total no. valid kelp area obs. per pixel (1984-2024)",
  x = "Longitude",
  y = "Latitude"
) +
theme_minimal()
```



Task 3: Inferential Statistics (Part 1)

- **[Task Weight: 20%]**
- **[Components (1), (2), (3), and (4) each marked on a 0–100 scale, then scaled to equal proportions of the Task Weight of 20%]**

You are now asked to formally test whether the weighted mean kelp canopy area has changed over time, and whether it shows evidence of seasonal variation.

You should:

1. Formulate and clearly state the null and alternative hypotheses for each of the following:
   - a temporal effect (i.e., whether kelp canopy area has changed across the study period); and
   - a seasonal effect (i.e., whether kelp canopy area differs between quarters).

Answer

Note to assessor: The null and alternative hypotheses can be clearly and strictly stated as I have them. This provides strong and an unambiguous basis for the inferential tests that follow, and marks should be assigned accordingly (questions 1-4).

Temporal effect (across years):

- H0: There is no effect of year on kelp canopy area.
- H1: There is a significant effect of year on kelp canopy area.

Seasonal effect (across quarters):

- H0: There is no effect of quarter on kelp canopy area.
- H1: There is a significant effect of quarter on kelp canopy area.

2. Choose and implement a statistical model appropriate to this task.

You may also consider:

- whether to model individual observations or to aggregate the data across spatial pixels; and
- how to treat missing or zero-valued observations.

The model you choose should reflect your understanding of the data structure and the nature of the questions being asked.

Answer

Note to assessor: Given the relatively small number of observations (n = 164 quarters over 41 years), and the fact that these are weighted mean values (not raw spatially disaggregated pixels), a linear model is appropriate. We'll use `lm()` here with both year and quarter as additive predictors, but I'll also accept a linear model to test for the `year` effect and an ANOVA to test the `quarter` effect. However, I prefer the `lm()` as it is more inclusive, comprehensive, and efficient. I did not specifically ask for an interaction effect, so no extra marks for assessing the interaction term.

```
# Ensure correct types
quarter_means <- quarter_means |>
  mutate(
    year = as.integer(year),
    quarter = factor(quarter) # categorical, not numeric
  )

# Fit additive model: weighted area ~ year + quarter
lm_additive <- lm(weighted_area ~ year + quarter, data = quarter_means)

# Summary of the model
summary(lm_additive)
```

```
Call:
lm(formula = weighted_area ~ year + quarter, data = quarter_means)

Residuals:
     Min       1Q   Median       3Q      Max
-170.137  -42.569    4.579   40.684  174.330

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3751.4614   921.5608  -4.071 7.65e-05 ***
year            1.9786     0.4596   4.305 3.05e-05 ***
quarter2      -31.0695    14.2591  -2.179 0.030941 *
quarter3      -43.5227    14.3554  -3.032 0.002877 **
quarter4      -49.3584    14.3555  -3.438 0.000763 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 62.55 on 146 degrees of freedom
Multiple R-squared:  0.1838,    Adjusted R-squared:  0.1614
F-statistic: 8.219 on 4 and 146 DF,  p-value: 5.25e-06
```

```
# I'd also accept simply two models that only look at year and season
# as the main effects
```

3. Justify your modelling approach, including:
    - why you chose that particular method (rather than alternatives);
    - the assumptions involved; and
    - how those assumptions might be violated in this dataset.

Answer

Note to assessor: A linear model was selected to test for additive effects of time (`year`, numeric) and seasonality (`quarter`, categorical) on the weighted mean kelp canopy area. This is justified by the structure of the quarter_means dataset, which is already aggregated by time and does not include repeated spatial

measurements that would justify a hierarchical or mixed model. The model assumes independence of residuals, homoscedasticity, and approximate normality of errors—conditions that are examined below.

A similar justification would be needed should the choice have been a `lm()` for `year` and an `aov()` for `quarter`.

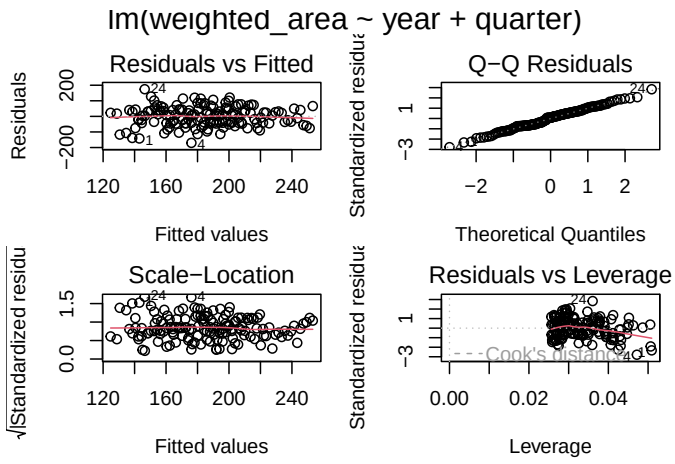
It is plausible the the normality or heteroscedasity assumptions are violated as a result of the slightly right-skewed data seen in an earlier analysis. We will test these in the assumption tests which follow.

4. Present and interpret your results as you would in a scientific paper.

Answer

Note to assessor: The diagnostics might not be part of the Results, but it should be reported somewhere to convince the examiner that these were checked. It needs to be mentioned in the Results (below) what the findings where to justify the test you ultimately selected.

```
# Default diagnostic plots
# Set up compact margins and multi-panel layout
par(
  mfrow = c(2, 2),     # 2 rows, 2 columns
  mar = c(4, 4, 2, 1), # inner margins: bottom, left, top, right
  oma = c(1, 1, 1, 1)  # outer margins
)

# Plot standard lm diagnostics
plot(lm_additive)
```



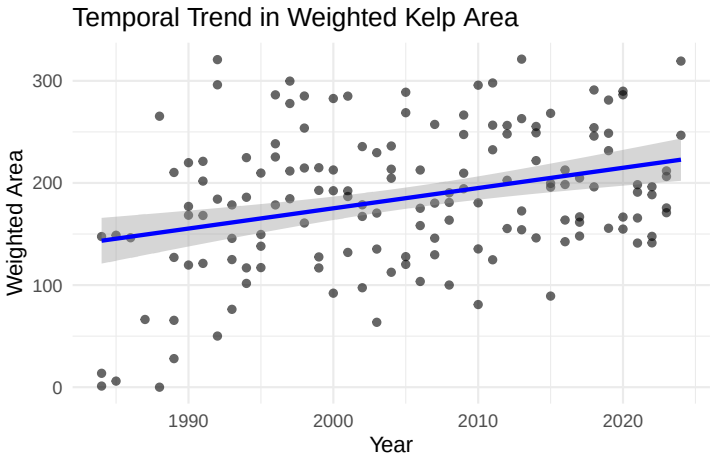

```
# Reset to default layout afterwards
par(mfrow = c(1, 1))
```

These two figure will need to be reported with the Results, and explained:

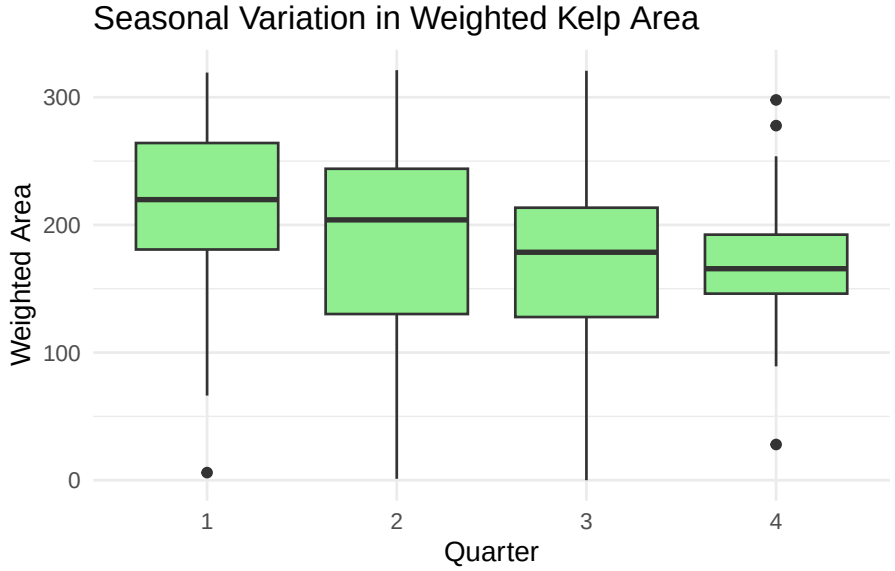
```
anova(lm_additive)
```

```
Analysis of Variance Table

Response: weighted_area
           Df Sum Sq Mean Sq F value    Pr(>F)
year        1  72915   72915 18.6353 2.905e-05 ***
quarter     3  55724   18575  4.7472  0.003446 **
Residuals 146 571260    3913
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Effect of year
ggplot(quarter_means, aes(x = year, y = weighted_area)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  labs(title = "Temporal Trend in Weighted Kelp Area", x = "Year", y = "Weighted Area") +
  theme_minimal()
```




```r
# Effect of quarter
ggplot(quarter_means, aes(x = quarter, y = weighted_area)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Seasonal Variation in Weighted Kelp Area", x = "Quarter", y = "Weighted Area") +
  theme_minimal()
```



Results The linear model examining the additive effects of year (numeric) and quarter (categorical) on quarterly weighted mean kelp canopy area reveals two prominent patterns: a clear temporal trend and a seasonal modulation of kelp cover (Figure x).

Regarding the temporal trend, the coefficient for year is positive and statistically significant ($= 1.98$, $p < 0.001$), indicating a gradual increase in kelp canopy area over the 41-year period. On average, the weighted canopy area increased by approximately 2 m^2 per year. This suggests a persistent long-term expansion of kelp cover.

All three quarter indicators (Q2–Q4, with Q1 as reference) are significantly negative, confirming that Q1 supports the highest canopy cover (Figure y). The estimated differences range from -31 m^2 (Q2) to -49 m^2 (Q4), all statistically significant at $p < 0.05$. This seasonal signal aligns with ecological expectations regarding seasonal effects.

The model explains approximately 18% of the total variance ($R^2 = 0.18$), which, while slight, is reasonable for ecological time series aggregated at the quarterly scale. Residual plots suggest approximate normality, with slight heteroscedasticity at the extremes. The Q–Q plot exhibits minor departures in the tails, but no worrying outliers or leverage points compromise the fit. These diagnostics support the model's suitability for inference, though future work might benefit from a more flexible framework (e.g., GAMs) to account for potential nonlinearities.

These findings point to a persistent long-term increase in kelp canopy cover along the South African coast, potentially reflecting broader shifts in oceanographic or climatic regimes, such as changes in upwelling intensity, nutrient flux, or wave climate. The consistent seasonal signature, with maximal cover in Q1, supports the hypothesis of photoperiod- or temperature-driven phenology. Notably, the lower cover in Q4 might coincide with stormier periods or post-reproductive senescence. Taken together, the temporal and seasonal dynamics provide a robust baseline for assessing spatial heterogeneity and long-term ecological change in subsequent tasks.

Task 4: Assigning Kelp Observations to Coastal Sections

- **[Task Weight: 20%]**
- **[Tasks 4.1 and 4.2 each marked on a 0–100 scale, then scaled in the proportion 0.7 and 0.3 of the Task Weight of 20%]**

Using the data prepared above, your task now is to spatially classify each kelp canopy observation by assigning it to two types of geographic units.

4.1 Assignment to Coastal Sections

You are provided with a table of 58 coastal sections, each defined by a single geographic coordinate (`Latitude` and `Longitude`). These points mark successive ~50 km intervals along the South African coastline, numbered from west (1) to east (58).

Assign each kelp canopy observation to the nearest coastal section based on geographic proximity:

- use a geodesic (great-circle) distance metric to compute proximity between kelp sampling points and section coordinates (assume all coordinates are in WGS84);
- add a new column to your kelp dataset called `section_id`, indicating the row number (1–58) of the nearest section; and
- you may use any R packages or methods you like, but your code should be efficient and well-commented.

Answer

Note to assessor: Check for objective evidence that a dataframe has been created and that it has the correct columns, i.e., `section_id` or `section`, `longitude`, `latitude`, `year`, `quarter`, and `area` or `weighted_area`. Calculations must show evidence of using the Haversine distance metric.

I use only the relevant sections:

```
sections_df <- sections_df |>
  slice(1:22) # no Ecklonia further east of section 22
```

Prepare a script for sectioning the kelp data:

```
processed_file_path <- "../data/Kelpwatch/pixel_means.RData"

if (file.exists(processed_file_path)) {
  message("Loading pre-processed pixel_means from: ", processed_file_path)
  load(processed_file_path)
  message("Successfully loaded pixel_means with sections.")
} else {
  message("Processing and assigning sections to pixel_means...")
}

# Check if the processed file exists and load it
if (file.exists(processed_file_path)) {
  message("Loading pre-processed pixel_means from: ", processed_file_path)
  load(processed_file_path)
  message("Successfully loaded pixel_means with sections.")
} else {
  message("Processing and assigning sections to pixel_means...")

  # Create section assignments
  section_coords_matrix <- as.matrix(sections_df[, c("Longitude", "Latitude")])

  find_closest_section_idx <- function(kelp_lon, kelp_lat, all_section_coords) {
    if (is.na(kelp_lon) || is.na(kelp_lat)) {
      return(NA_integer_)
```

```
    }
    current_kelp_coord <- c(kelp_lon, kelp_lat)
    distances <- distHaversine(p1 = current_kelp_coord, p2 = all_section_coords)
    return(which.min(distances))
  }

  # Apply the function to each kelp data point
  assigned_section_indices <- mapply(find_closest_section_idx,
    kelp_lon = pixel_means$longitude,
    kelp_lat = pixel_means$latitude,
    MoreArgs = list(
      all_section_coords = section_coords_matrix
    ),
    SIMPLIFY = TRUE
  )

  # Add section column to pixel_means
  # pixel_means2[, section := assigned_section_indices]
  pixel_means$section <- assigned_section_indices

  # Ensure the data directory exists
  if (!dir.exists(dirname(processed_file_path))) {
    dir.create(dirname(processed_file_path), recursive = TRUE)
  }

  # Save the processed data
  save(pixel_means, file = processed_file_path)
  message("Processing complete. Data saved to: ", processed_file_path)
}

head(pixel_means)
```

```
# A tibble: 6 x 7
       year   quarter longitude  latitude weighted_area time_index section
  <int[1d]> <int[1d]> <dbl[1d]> <dbl[1d]>         <dbl>  <dbl[1d]>   <int>
1      1984         2      15.1     -26.6             0      1984.       1
2      1984         2      15.1     -26.6             0      1984.       1
3      1984         2      15.1     -26.6             0      1984.       1
4      1984         2      15.1     -26.6             0      1984.       1
5      1984         2      15.1     -26.6             0      1984.       1
6      1984         2      15.1     -26.6             0      1984.       1
```

4.2 Assignment to Biogeographical Provinces

You are also provided with a table that maps each coastal section (1–58) to a biogeographical province, based on a classification by Professor John Bolton.

- Using your previous assignment of each kelp observation to a `section_id`, add a second column called `bioregion_id` that indicates which biogeographical province the observation falls within.
- Your final kelp dataset should contain the following key columns (alongside the original data):
  - `longitude`, `latitude`
  - `year`, `quarter`, `area`, `passes`
  - `section_id` (integer 1–58)
  - `bioregion_id` (character or factor)
- Include your full, annotated R code that performs both spatial assignments into your resultant .html document. Your method should be reproducible, and your code should be easy to follow. Print the `head()` and `tail()` of your final dataset, and include a `summary()` of the data.

Answer

Note to assessor: Check for objective evidence that a dataframe has been created and that it has the correct columns, i.e., `section_id` or `section`, `longitude`, `latitude`, `year`, `quarter`, `area` or `weighted_area`, and also `bioregion` or `bolton` or some unique identifier for the bioregion.

I used a simple merge to assign bioregions to the kelp data, but there are other ways to do it. Again, we want objective evidence that the data have been assigned correctly (e.g. the `head()` and `tail()` of the data, or a `summary()`).

```r
bioreg <- read.csv("../data/Kelpwatch/bioregions.csv") |>
  mutate(section = row_number())  # Make rownames explicit as 'section'

# Merge into pixel_means (assumed already loaded in environment)
pixel_means <- pixel_means |>
  left_join(bioreg, by = "section") |>
  select(-spal.prov, -spal.ecoreg, -lombard)

# The question asks for passes to be present in the data, but
# I'm okay with omitting it as it is included in the weighted
# area calculation

head(pixel_means)
```

```
# A tibble: 6 x 7
      year   quarter longitude  latitude weighted_area section bolton
  <int[1d]> <int[1d]> <dbl[1d]> <dbl[1d]>         <dbl>   <int> <chr>
1      1984         2      15.1     -26.6             0       1 BMP
2      1984         2      15.1     -26.6             0       1 BMP
3      1984         2      15.1     -26.6             0       1 BMP
4      1984         2      15.1     -26.6             0       1 BMP
5      1984         2      15.1     -26.6             0       1 BMP
6      1984         2      15.1     -26.6             0       1 BMP
```

```r
tail(pixel_means)
```

```
# A tibble: 6 x 7
      year   quarter longitude  latitude weighted_area section bolton
  <int[1d]> <int[1d]> <dbl[1d]> <dbl[1d]>         <dbl>   <int> <chr>
1      2024         2      20.0     -34.8             0      22 AMP
2      2024         2      20.0     -34.8             0      22 AMP
3      2024         2      20.0     -34.8             0      22 AMP
4      2024         2      20.0     -34.8             0      22 AMP
5      2024         2      20.0     -34.8             0      22 AMP
6      2024         2      20.0     -34.8             0      22 AMP
```

```r
summary(pixel_means)
```

```
      year         quarter        longitude        latitude
 Min.   :1984   Min.   :1.000   Min.   :15.08   Min.   :-34.83
 1st Qu.:1997   1st Qu.:2.000   1st Qu.:18.33   1st Qu.:-34.67
 Median :2006   Median :3.000   Median :18.84   Median :-34.36
 Mean   :2006   Mean   :2.503   Mean   :18.64   Mean   :-33.62
 3rd Qu.:2016   3rd Qu.:3.000   3rd Qu.:19.41   3rd Qu.:-33.95
 Max.   :2024   Max.   :4.000   Max.   :19.97   Max.   :-26.48
 weighted_area      section         bolton
 Min.   :  0.0   Min.   : 1.00   Length:5446570
 1st Qu.:  0.0   1st Qu.:15.00   Class :character
 Median :  0.0   Median :19.00   Mode  :character
 Mean   :196.1   Mean   :15.82
 3rd Qu.:360.0   3rd Qu.:20.00
 Max.   :900.0   Max.   :22.00
```

Task 5: Inferential Statistics (Part 2)

- [**Task Weight: 30%**]
- [**Tasks 5.1, 5.2, 5.3, 5.4, and 5.5 each marked on a 0–100 scale, then scaled to equal proportions of the Task Weight of 30%**]

You are now asked to evaluate a series of research questions concerning the spatial and temporal structure of kelp canopy area. These questions are to be answered using the kelp dataset that has already been processed to include both `section_id` and `bioregion_id`. Use the weighted kelp canopy area (area, weighted by passes) as your response variable throughout – you should have already prepared this dataset in Task 2.

You may use ANOVAs and/or linear models. In each case you must clearly state your hypotheses, justify your choice of model, and interpret your findings both statistically and ecologically.

5.1 Spatial Differences Between Coastal Sections

Question: Is there a statistically significant difference in mean kelp canopy area between coastal sections?

Answer

Note to Assessor (applies to Questions 5.1–5.5): It is understood that students are not expected to possess the depth of statistical training characteristic of professional statisticians. While some questions may, in principle, warrant more advanced treatments (e.g. linear mixed-effects models), the scope of instruction in this module was limited to foundational techniques—namely, simple linear models and ANOVA for relatively straightforward designs. **If students applied LMEs or other more complicated models without demonstrating a clear understanding for why they did so, give them zero percent for that answer.** Where assumptions of parametric tests are violated, students should identify suitable non-parametric alternatives. Marks will be awarded for correctly identifying such alternatives, even if these tests are not implemented in code.

Hypotheses:

- Null (H): There is no difference in mean kelp canopy area across coastal sections.
- Alternative (H): Mean kelp canopy area differs among at least one pair of coastal sections.

Justification: A one-way ANOVA is appropriate for testing whether the means of a continuous response variable (kelp area) differ across the levels of a single categorical factor (coastal section).

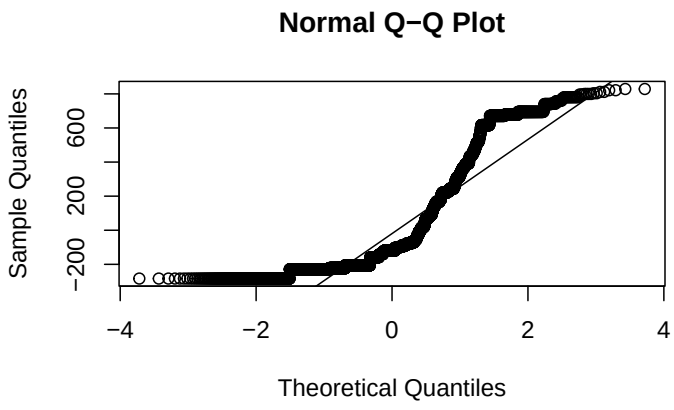
```
library(lmtest)
library(car)

# Set section and bioreg as a factor as this is needed for the tests
# that will assess the section and bioreg effects
pixel_means <- pixel_means |>
  mutate(section = as.factor(section),
         bolton = as.factor(bolton))

model_5.1 <- aov(weighted_area ~ section, data = pixel_means)
summary(model_5.1)
```

```
                Df    Sum Sq   Mean Sq F value Pr(>F)
section         18 2.020e+10 1.122e+09   13252 <2e-16 ***
Residuals   5446551 4.612e+11 8.467e+04
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Also test assumptions...
# Residuals
resid_5.1 <- sample(residuals(model_5.1), 5000)

# Normality
shapiro.test(resid_5.1)  # Subsample for tractability
```

```
    Shapiro-Wilk normality test

data:  resid_5.1
W = 0.80962, p-value < 2.2e-16
```

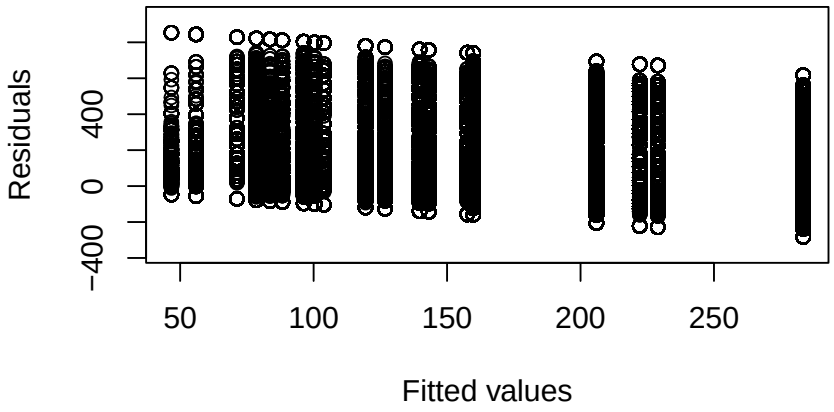
```
qqnorm(resid_5.1); qqline(resid_5.1)
```



```r
# Homogeneity of variances
leveneTest(weighted_area ~ section, data = pixel_means)
```

```
Levene's Test for Homogeneity of Variance (center = median)
           Df F value    Pr(>F)
group      18   14154 < 2.2e-16 ***
      5446551
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Residuals vs fitted
plot(model_5.1, which = 1)
```



One could do a TukeyHSD test, but the question did not explicitly ask for this. Add two marks if correctly given and executed.

Results Mean kelp canopy area varied significantly across the 19 coastal sections assessed ($F18, 5446551 = 13{,}252$, $p < 0.001$; one-way ANOVA). Residual diagnostics indicated substantial departures from normality (Shapiro–Wilk $W = 0.81$, $p < 0.001$) and non-constant variance across groups (Levene's test: $F18, 5446551 = 14{,}154$, $p < 0.001$), though the extremely large sample size renders the ANOVA robust to these violations. The fitted model accounted for a substantial proportion of the spatial variance in kelp canopy area. Residual plots suggested heteroscedasticity consistent with spatial heterogeneity in canopy distributions.

To provide a fully defensible hypothesis test (considering both assumptions were violated), do a Kruskal–Wallis rank sum test. This is the standard non-parametric alternative to one-way ANOVA when comparing more than two independent groups (here, coastal sections). It tests whether the distributions of kelp canopy area differ across sections without assuming normality or equal variances.

5.2 Spatial Differences Between Biogeographical Provinces

Question: Is there a statistically significant difference in mean kelp canopy area between biogeographical provinces?

Answer

Hypotheses:

- Null (H): Mean kelp canopy area is not different amongst bioregions.
- Alternative (H): Mean kelp canopy area differs between bioregions.

Justification: Here, bioregion is treated as a categorical predictor, suitable for a one-way ANOVA to test inter-bioregional differences. The weighted kelp area is the continuous response variable. Assumptions of normality and heteroscedasticity hold.

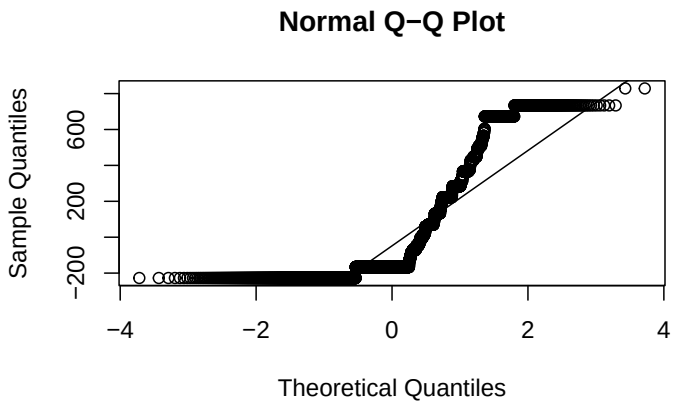
```r
model_5.2 <- aov(weighted_area ~ bolton, data = pixel_means)
summary(model_5.2)
```

```
                Df    Sum Sq   Mean Sq F value Pr(>F)
bolton           2 6.897e+09 3.448e+09   39585 <2e-16 ***
Residuals  5446567 4.745e+11 8.711e+04
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Also test assumptions...
resid_5.2 <- sample(residuals(model_5.2), 5000)

shapiro.test(resid_5.2)
```

```
    Shapiro-Wilk normality test

data:  resid_5.2
W = 0.74748, p-value < 2.2e-16
```

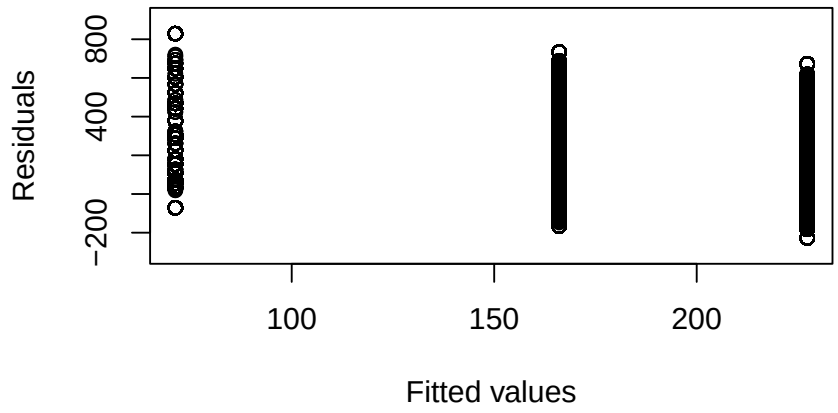
```r
qqnorm(resid_5.2); qqline(resid_5.2)
```




```r
leveneTest(weighted_area ~ bolton, data = pixel_means)
```

```
Levene's Test for Homogeneity of Variance (center = median)
           Df F value    Pr(>F)
group       2   39585 < 2.2e-16 ***
      5446567
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
plot(model_5.2, which = 1)
```



Results A one-way ANOVA revealed significant differences in mean kelp canopy area between the three biogeographical provinces (BMP, SWP, AMP) ($F2, 5446567 = 39{,}585$, $p < 0.001$). Residual analysis again revealed non-normal error distribution (Shapiro–Wilk $W = 0.75$, $p < 0.001$) and unequal variances among provinces (Levene's test: $F2, 5446567 = 39{,}585$, $p < 0.001$). Despite these assumption violations, the effect was large and consistent with substantial regional differentiation in canopy extent. Residuals from the model were symmetrically distributed but with longer tails in high-variance provinces.

As before, a Kruskal–Wallis rank sum test is recommended. Again suitable, since the bioregion variable has three independent levels. This test evaluates whether the central tendency of canopy area differs among provinces without assuming parametric conditions.

5.3 Interaction Between Section and Province

Question: Is there an interaction between coastal section and biogeographical province in explaining variation in kelp canopy area?

Answer

Hypotheses:

- H (Main effect – Section): There are no differences in mean kelp canopy area across coastal sections.
- H (Main effect – Province): There are no differences in mean kelp canopy area across biogeographical provinces.
- H (Interaction): The effect of coastal section on kelp canopy area is the same across all provinces (i.e., no interaction).
- H (Alternative): At least one main effect is non-zero, or there is a significant interaction (i.e., the influence of section depends on the province).

A two-way ANOVA with interaction is appropriate here because:

- Both predictors, i.e., `section` (a factor) and province (`bolton`, also a factor), are categorical.
- The response variable (`weighted_area`) is continuous and assumed to meet the assumptions of normality and homogeneity of variance.
- The interaction term allows us to assess whether the spatial granularity of section varies in importance across broader-scale bioregions.

This model allows us to partition variation in canopy area into hierarchical (or cross-cutting) spatial components.

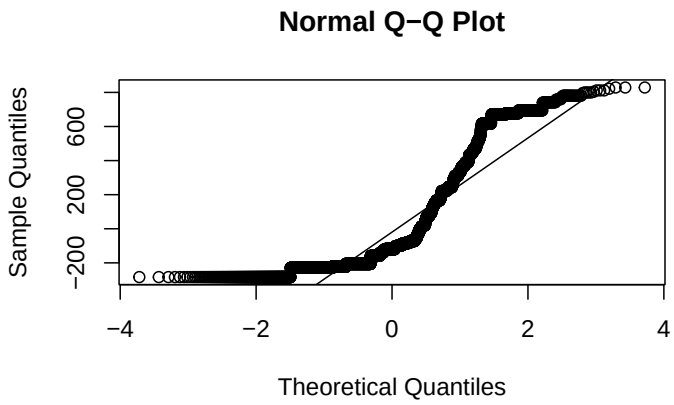
```
# Two-way ANOVA with interaction
model_5.3 <- aov(weighted_area ~ section * bolton, data = pixel_means)

# Summary of the model
summary(model_5.3)
```

```
                Df    Sum Sq   Mean Sq F value Pr(>F)
section         18 2.020e+10 1.122e+09   13252 <2e-16 ***
Residuals   5446551 4.612e+11 8.467e+04
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Also test assumptions...
resid_5.3 <- sample(residuals(model_5.3), 5000)

shapiro.test(resid_5.3)
```

```

    Shapiro-Wilk normality test

data:  resid_5.3
W = 0.81015, p-value < 2.2e-16
```

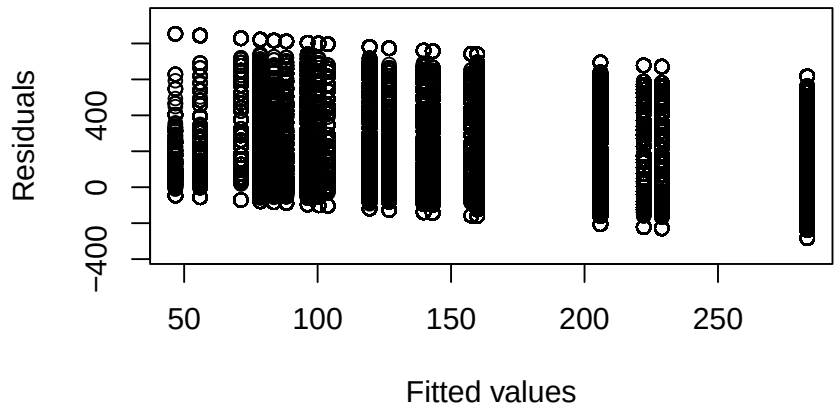
```
qqnorm(resid_5.3); qqline(resid_5.3)
```




```
car::leveneTest(model_5.3)
```

```
Levene's Test for Homogeneity of Variance (center = median)
           Df F value    Pr(>F)
group      18   14154 < 2.2e-16 ***
      5446551
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(model_5.3, which = 1)
```



In this case, interaction terms cannot be estimated, because there's no shared replication across bioregions within sections. The design is not factorial. The solution is to fix mixed models with random and fixed

effects.

Results A two-way ANOVA incorporating both section and biogeographical province, including their interaction, confirmed strong spatial structuring of kelp canopy area ($F_{18,5446551} = 13{,}252$, $p < 0.001$ for the section main effect). However, the interaction term was non-estimable due to the confounded structure of the design — each section belonged uniquely to a single province, precluding factorial replication. As such, a fully crossed design could not be evaluated within a fixed-effects ANOVA framework. Model residuals displayed similar deviations from normality and homogeneity as in previous analyses, but were centered and broadly symmetric.

A non-parametric two-way tests for interaction effects do not exist. Because of the nested design, a fully non-parametric two-way ANOVA is not viable. As an alternative approach, perform separate Kruskal–Wallis tests within each province, or apply a permutation-based two-way ANOVA that accommodates unbalanced and non-normal designs.

5.4 Linear Trend Over Time by Province

Question: Is there a linear trend in kelp canopy area over time, and does the direction or strength of this trend differ between biogeographical provinces?

Answer

Hypotheses:

- H (Main effect of time): There is no linear trend in kelp canopy area over time.
- H (Time × Province interaction): The effect (slope) of year is the same across all provinces.
- H (Alternative): Kelp canopy area shows a linear trend over time, and the rate or direction of change differs by province.

This is a typical ANCOVA design because:

- The continuous predictor `year` captures temporal trends.
- The categorical predictor `bolton` accounts for regional differences.
- The interaction term `year` × `bolton` assesses whether temporal slopes differ among provinces.

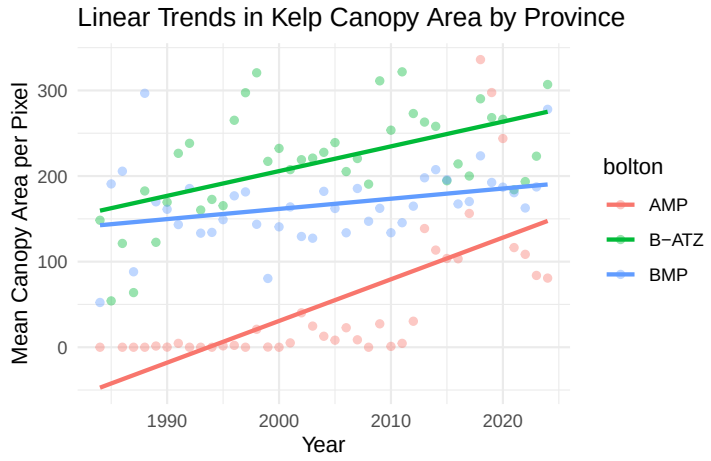
ANCOVA is ideal for testing whether regression lines have the same slope across groups; here, whether kelp decline/growth rates differ by province.

```
# Aggregate by year and province
province_annual <- pixel_means |>
  group_by(year, bolton) |>
  summarise(mean_area = mean(weighted_area, na.rm = TRUE), .groups = "drop")

# Fit the ANCOVA model
model_5.4 <- lm(mean_area ~ year * bolton, data = province_annual)

# Summary of model
# summary(model_5.4)  # Dont print... too long!

# A Type III ANOVA, as in the car::Anova() function with type = 3, tests the
# significance of each effect after accounting for all other effects in the
# model, including interactions. It asks: "What is the unique contribution
# of this factor, given that all others (including interactions) are already
# in the model?"
Anova(model_5.4, type = 3)
```

```
Anova Table (Type III tests)

Response: mean_area
             Sum Sq  Df F value    Pr(>F)
(Intercept) 125927   1 45.2807 6.830e-10 ***
year        127296   1 45.7731 5.702e-10 ***
bolton       38795   2  6.9749  0.001379 **
year:bolton  37570   2  6.7548  0.001679 **
Residuals   322599 116
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ggplot(province_annual,
       aes(x = year, y = mean_area, color = bolton)) +
```

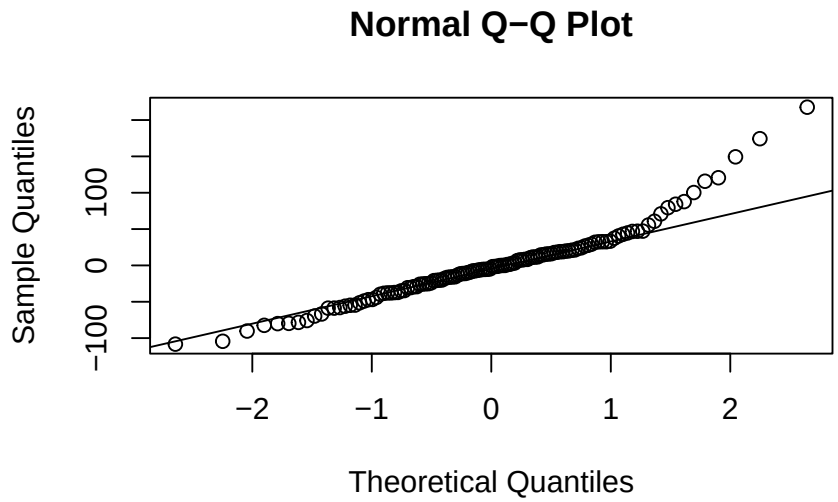
```
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", se = FALSE) +
  theme_minimal() +
  labs(
    title = "Linear Trends in Kelp Canopy Area by Province",
    y = "Mean Canopy Area per Pixel",
    x = "Year"
  )
```



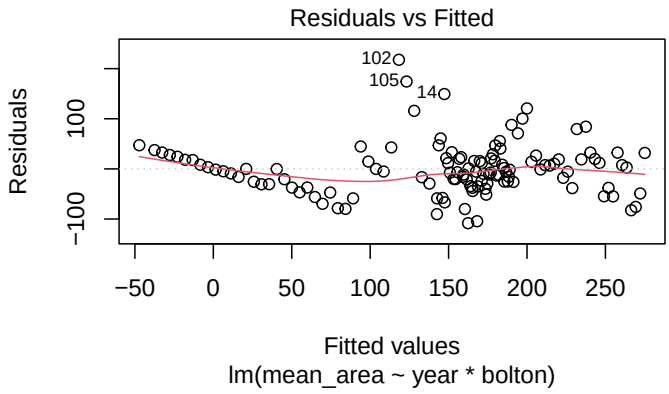

```
# Also test assumptions...
resid_5.4 <- residuals(model_5.4)

# Normality
shapiro.test(resid_5.4)
```

```
    Shapiro-Wilk normality test

data:  resid_5.4
W = 0.93452, p-value = 1.599e-05
```

```
qqnorm(resid_5.4); qqline(resid_5.4)
```




```
# Homoscedasticity
plot(model_5.4, which = 1)
```




```
bptest(model_5.4)  # Breusch-Pagan test or any other suitable test
```

```
    studentized Breusch-Pagan test

data:  model_5.4
BP = 16.228, df = 5, p-value = 0.006223
```

```
# Multicollinearity
# Don't penalise students for not testing multicollinearity, but
# add a bonus two marks if included
vif(model_5.4)
```

```
                    GVIF Df GVIF^(1/(2*Df))
year        3.138942e+00  1        1.771706
bolton      8.468999e+08  2      170.591747
year:bolton 8.466877e+08  2      170.581063
```

```
# Independence of residuals
dwtest(model_5.4)  # Durbin-Watson test
```

```
    Durbin-Watson test

data:  model_5.4
DW = 1.7605, p-value = 0.1114
alternative hypothesis: true autocorrelation is greater than 0
```

Results The ANCOVA model indicated a significant linear increase in mean kelp canopy area over time across all provinces ($F1, 116 = 45.8$, $p < 0.001$), with significant differences in temporal trends among provinces (year × province interaction: $F2, 116 = 6.75$, $p = 0.0017$). This suggests that both the direction and magnitude of temporal trends in canopy area are modulated by biogeographical province. Residuals exhibited slight right-skew (Shapiro–Wilk $W = 0.93$, $p < 0.001$), and heteroscedasticity was supported by a significant Breusch–Pagan test (BP = 16.23, df = 5, $p = 0.006$). No evidence of residual autocorrelation was detected (Durbin–Watson DW = 1.76, $p = 0.11$). Variance inflation factors indicated high collinearity between province and interaction terms, reflecting the nested structure of the provinces over time.

Pragmatic alternative to account for non-normality, heteroscedasticity, and collinearity: perform separate Spearman's rank correlation tests between year and mean area within each province to assess monotonic trends, then compare slope magnitudes informally or using non-parametric trend tests like Mann–Kendall.

5.5 Seasonal Variation Across Provinces

Question: Does the seasonal pattern in kelp canopy area differ between provinces?

Answer

Hypotheses:

- H (Main effect of quarter): Mean kelp canopy area does not vary across seasons (quarters).
- H (Quarter × Province interaction): The pattern of seasonal variation is the same in all provinces — i.e., there is no interaction.

- H_A (Alternative): There are seasonal differences in kelp canopy area, and the shape or magnitude of the seasonal cycle differs by province.

A two-way ANOVA is appropriate here, with:

- `quarter` as a categorical predictor representing seasonal variation.
- `bolton` as a categorical grouping variable for province.
- An interaction term (`quarter` × `bolton`) to test whether seasonal cycles differ by region.

This model assesses both the amplitude and phase-shift of seasonal dynamics across spatial domains.

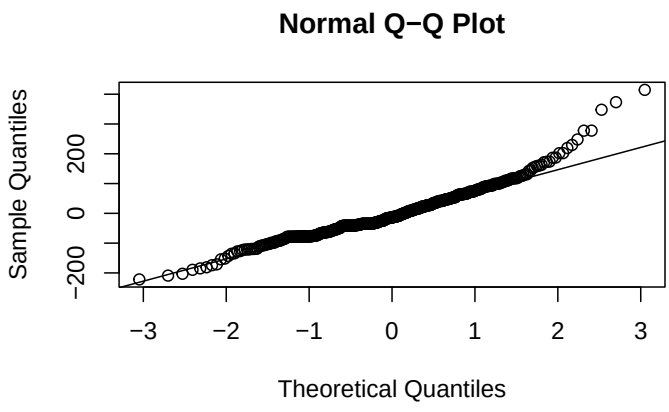
```
# Ensure quarter is treated as a categorical variable
pixel_means <- pixel_means |>
  mutate(quarter = as.factor(quarter))

# Group by year, quarter, and province to retain replication across years
province_seasonal_replicated <- pixel_means |>
  group_by(year, quarter, bolton) |>
  summarise(mean_area = mean(weighted_area, na.rm = TRUE), .groups = "drop")

# Fit two-way ANOVA with interaction
model_5.5 <- aov(mean_area ~ quarter * bolton, data = province_seasonal_replicated)

# Model summary
summary(model_5.5)
```

```
               Df  Sum Sq Mean Sq F value Pr(>F)
quarter         3   66835   22278   3.061 0.0281 *
bolton          2 1944753  972377 133.590 <2e-16 ***
quarter:bolton  6  103175   17196   2.362 0.0295 *
Residuals     423 3078948    7279
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Extract residuals
resid_5.5 <- residuals(model_5.5)

# Test for normality of residuals
shapiro.test(resid_5.5)
```

```
    Shapiro-Wilk normality test

data:  resid_5.5
W = 0.95331, p-value = 1.7e-10
```

```
qqnorm(resid_5.5); qqline(resid_5.5)
```

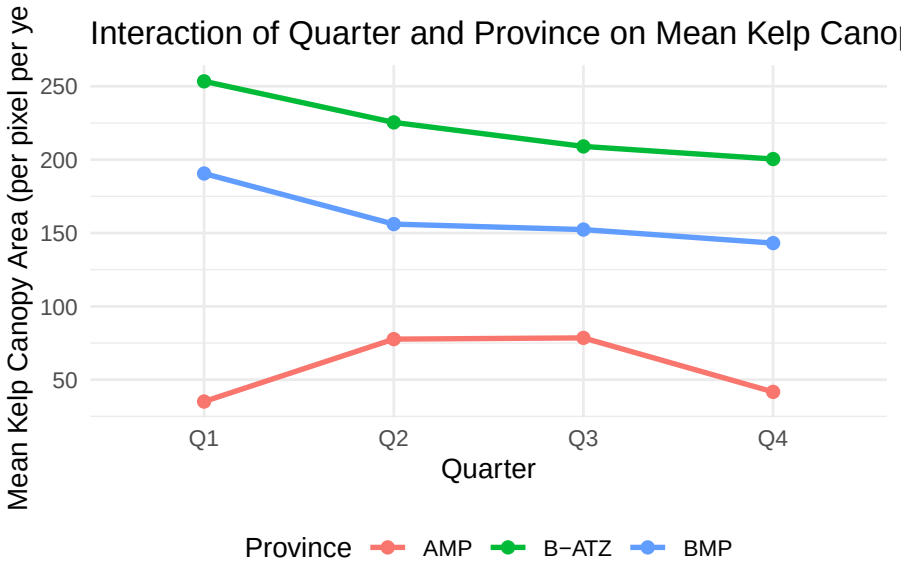



```
# Test for homogeneity of variances across quarter × province combinations
province_seasonal_replicated$group <- interaction(province_seasonal_replicated$quarter,
                                                  province_seasonal_replicated$bolton)
leveneTest(mean_area ~ group, data = province_seasonal_replicated)
```

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value   Pr(>F)
group  11  2.3599 0.007804 **
```

```
      423
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Interaction plot: mean kelp canopy area by quarter and province
ggplot(province_seasonal_replicated,
                            aes(x = quarter,
                                y = mean_area,
                                group = bolton,
                                color = bolton)) +
  stat_summary(fun = mean, geom = "line", size = 1) +
  stat_summary(fun = mean, geom = "point", size = 2) +
  theme_minimal(base_size = 11) +
  scale_x_discrete(labels = c("Q1", "Q2", "Q3", "Q4")) +
  labs(
    title = "Interaction of Quarter and Province on Mean Kelp Canopy Area",
    x = "Quarter",
    y = "Mean Kelp Canopy Area (per pixel per year)",
    color = "Province"
  ) +
  theme(legend.position = "bottom")
```



Results Seasonal variation in kelp canopy area differed significantly among provinces (two-way ANOVA: quarter × province interaction: $F6, 423 = 2.36$, p = 0.030). Both main effects were significant, with variation among quarters ($F3, 423 = 3.06$, $p = 0.028$) and among provinces ($F2, 423 = 133.6$, $p < 0.001$). Residuals exhibited mild non-normality (Shapiro–Wilk $W = 0.95$, $p < 0.001$), and homoscedasticity was violated across the interaction groups (Levene's test: $F11, 423 = 2.36$, $p = 0.008$). Nevertheless, interaction plots indicated province-specific seasonal cycles in canopy area that were consistent across years.

There is no easy non-parametric alternative.

General Instructions for Task 5 (above)

For each sub-question, above, consider:

- formally state the null and alternative hypotheses;
- justify your choice of model;
- justify your choice of predictors;
- justify your decision to aggregate or not aggregate the data at various levels;
- discuss the assumptions involved and any violations you detect;
- present the relevant model outputs and statistical tests;
- include visualisations where appropriate (e.g. interaction plots, trend lines, diagnostic plots);
- justify your choice of visualisation; and
- present the results in a clear and concise manner, including tables and figures where appropriate, in a manner that would be appropriate for a scientific audience (e.g. a journal article).

You are not required to use the same modelling approach for all five sub-questions, though consistency across related questions is encouraged.

Task 6: Write-up

- **[Task Weight: 10%]**

Write a short report (maximum 2 pages of text) that synthesises your findings across Tasks 2 through 5. This report should be written in the style of the Discussion section of a scientific paper, intended for an ecological audience.

Your goal is to interpret the major patterns and relationships you have identified, and to comment meaningfully on their ecological significance. Your write-up should include:

- Temporal Trends and Seasonality.
- Spatial Structure and Biogeography.
- Interaction Effects and Spatial–Temporal Coupling.
- Limitations and Assumptions.
- Ecological Interpretation.

Format and tone:

- Aim for clarity and economy of expression.
- Don't generate any new tables and figures. The tables and figures from Tasks 2 through 5 should be sufficient.
- Write in complete paragraphs. Avoid bulleted summaries.
- Add references to the tables and figures from Tasks 2 through 5 as needed.
- Cite any additional references you use.

Answer

Note to Assessor: Penalise all text that was clearly AI generated by subtracting 50% off the mark.

Discussion

Our analysis of long-term kelp canopy dynamics revealed strong spatio-temporal structuring in both the mean extent and variability of kelp forests across southern Africa. While the limitations of simple linear and ANOVA-based models prevent exhaustive ecological inference, the results consistently support our observations that kelp canopy cover is not uniformly variable in space or time, but instead shaped by regionally distinct seasonal regimes and long-term trends that interact with the underlying biogeographic patterns.

At the broadest temporal scale, the annual means (Task 2) show evidence for a long-term linear increase in kelp canopy area across the study period, although the strength and direction of this trend varied substantially by biogeographical province. The Benguela Marine Province (BMP) exhibited a relatively strong positive trend, while the Agulhas and Benguela-Agulhas Transition Zone (AMP and B-ATZ) showed weaker or inconsistent trajectories. These patterns likely reflect contrasting oceanographic influences, specifically the differing degrees of upwelling intensity and sea surface temperature variability across provinces. When the annual trends were disaggregated seasonally (Task 3), the signal became more complex. Seasonal means indicated a marked peak in kelp canopy area during austral summer and early autumn (quarters 1 and 2). This seasonal cycle was most pronounced in BMP, less so in B-ATZ, and relatively flat in AMP, suggesting that regional exposure to storm-driven disturbance, solar irradiance, or herbivore activity may modulate canopy recovery and loss at intra-annual scales.

Spatial patterns in kelp canopy extent, assessed both at the bioregional level (Task 4.1) and at the finer spatial resolution of coastal sections (Task 5.1), reinforce the view that biogeographic context is a dominant driver of canopy variability. A large proportion of the overall variance in canopy area was attributable to differences among sections, but these differences were nested within broader provincial contrasts. The one-way and two-way ANOVAs demonstrated strong main effects of both spatial variables, although the interaction between section and province could not be statistically partitioned due to their hierarchical structure. Nonetheless, the magnitude of spatial variation suggests that local environmental factors (such as wave exposure, topographic complexity, and nutrient availability) likely modulate kelp dynamics on sub-provincial scales, even as broader provincial regimes exert overarching influence.

Where temporal and spatial effects intersect, the coupling of seasonal cycles with regional identity becomes ecologically meaningful. Task 5.5 illustrated this clearly: although all provinces experience some degree of seasonal fluctuation, the shape and amplitude of these cycles differed significantly. The interaction effect between quarter and province was significant despite the relatively simple model structure, indicating that the timing of canopy peaks and troughs is not synchronised coast-wide. This asynchrony complicates large-scale generalisations about kelp seasonality and underscores the importance of considering regional context in ecological forecasting. That the BMP, for instance, displays both the strongest seasonal cycle and the most pronounced long-term increase suggests a possible reinforcing interaction between natural seasonal pulses and long-term drivers such as ocean cooling (prevalent in the kelp's distributional range) or changing storm regimes.

Several methodological limitations constrain the scope of these interpretations. First, violations of model assumptions (non-normality, heteroscedasticity, and the lack of full factorial design in spatial terms) limit the interpretive power of parametric tests. While non-parametric alternatives were noted where appropriate (e.g. Kruskal–Wallis, etc.), their implementation was outside the scope of this coursework. Secondly, while the data offer high temporal and spatial resolution, they lack covariates such as sea surface temperature, nutrient levels, or herbivore abundance that would allow causal inference. Finally, the use of mean canopy area as the primary response variable conceals potentially important variation in canopy fragmentation, persistence, or patch turnover.

Despite these limitations, the ecological interpretation remains convincing: kelp canopy extent is increasing overall, but in a regionally specific manner that reflects underlying biogeographic structure. Seasonal variation is evident and substantial, but not temporally synchronised across the coastline. These findings align with previous studies that describe the interaction between physical forcing and biological response in kelp-dominated ecosystems (Dayton 1985; Wernberg et al. 2016). Future work should explore whether these spatial–temporal patterns correspond to known gradients in oceanographic regime or are indicative of broader shifts in coastal ecosystem functioning under climate change.

References

Dayton, P. K. (1985). Ecology of kelp communities. Annual Review of Ecology and Systematics, 16, 215–245.

Wernberg, T., et al. (2016). Climate-driven regime shift of a temperate marine ecosystem. Science, 353(6295), 169–172.