StatComp Project 2: Scottish weather

Axel Eichelmann (s2030757, axeleichelmann)

Scottish Weather Data

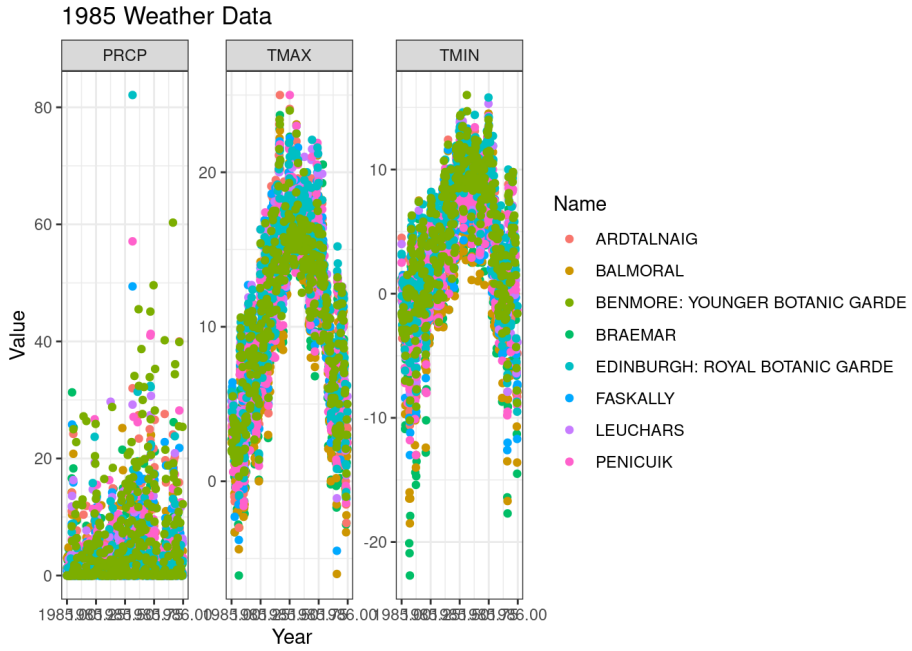
The data set `ghcnd_values` contain daily measurements of weather variables (such as precipitation, temperature, etc.) from the following 8 weather stations in Scotland

ID	Name	Latitude	Longitude	Elevation
UKE00105874	BRAEMAR	57.0058	-3.3967	339
UKE00105875	BALMORAL	57.0367	-3.2200	283
UKE00105884	ARDTALNAIG	56.5289	-4.1108	130
UKE00105885	FASKALLY	56.7181	-3.7689	94
UKE00105886	LEUCHARS	56.3767	-2.8617	10
UKE00105887	PENICUIK	55.8239	-3.2258	185
UKE00105888	EDINBURGH: ROYAL BOTANIC GARDE	55.9667	-3.2100	26
UKE00105930	BENMORE: YOUNGER BOTANIC GARDE	56.0281	-4.9858	12

from the 1^{st} of January 1960 to the 31^{st} of December 2018 in this section, we will seek to gain an understanding of how the probability of rainfall, as well as precipitation measurements vary across the year.

Seasonal Variance in Precipitation Amount

The figure below shows daily precipitation data from the year 1985 for each station. We also include the temperature data to show that the seasonal effects on precipitation are much less clear than their effects on temperature.



As we can see, there is an obvious increase in Maximum and Minimum temperature during the summer months. In contrast, there is no clear trend for the precipitation data.

In order to get a better understanding of the behaviour of rainfall throughout the year, we will carry out a Monte-Carlo premutation test on the hypothesis:

$$H_0 : \text{The rainfall distribution is the same in winter as in summer}$$
$$H_1 : \text{The winter and summer distributions have different expected values}$$

Here we will let the test statistic be $T = |\text{Winter Average Rainfall} - \text{Summer Average Rainfall}|$. Under the null hypothesis, the joint sample of Summer plus Winter rainfall would be a set of exchangeable variables, thus any random permutation of this set would have the same distribution as the original summer and winter rainfall data sets themselves. Our original data set gives the mean rainfall in summer and winter across the 8 stations, with corresponding test statistic values as:

Station ID	Avg. Summer Rainfall	Avg. Winter Rainfall	T-Value
UKE00105874	2.089790	2.878589	0.7887997
UKE00105875	1.994370	2.601891	0.6075213
UKE00105884	2.668042	4.679448	2.0114061
UKE00105885	2.130559	2.899627	0.7690686
UKE00105886	1.809706	1.931070	0.1213640
UKE00105887	2.326199	2.751264	0.4250654
UKE00105888	1.844564	1.869338	0.0247740
UKE00105930	4.995383	8.353198	3.3578143

Taking $10,000$ permutations of the Season Values for each weather station, and using $\frac{1}{J}\sum_{j=1}^{J}\delta_{(j)}$ (where $\delta_{(j)}=1$ if the permuted test statistic is greater than the one in the table above, and 0 otherwise), we get the following estimated p-values:

Station Name	Station ID	p-value
BRAEMAR	UKE00105874	0
BALMORAL	UKE00105875	0
ARDTALNAIG	UKE00105884	0
FASKALLY	UKE00105885	0
LEUCHARS	UKE00105886	0.0337
PENICUIK	UKE00105887	0
EDINBURGH: ROYAL BOTANIC GARDE	UKE00105888	0.6585
BENMORE: YOUNGER BOTANIC GARDE	UKE00105930	0

For the stations where the our estimated p-value is 0, can can calculate a corresponding 95 confidence interval for this estimate as $(0, 1-0.025^{1/N})$, where $N=10,000$ is the number of permutations we used in this test, we get it as $(0, 3.69\times10^{-4})$. Thus for any significance level above 3.69×10^{-4} (which is extremely low) there is a 95 chance our true p-value lies in the critical region. Thus, we can reject H_0, and conclude that for these stations it is almost certain that the season has an impact on the expected rainfall. As for the other two stations, we can use Jensen's inequality to find an upper bound for the standard deviation of the estimated p-value as $\sqrt{Var(\hat{p})}\leq\frac{1}{2\sqrt{N}}=0.005$, thus we know that for these stations the true p-value is likely to lie within 0.005 of our estimated p-value. In the case of Edinburgh this has little impact since after deducting this standard deviation from our estimated p-value, it still lies at 0.6535 which lies comfortably within the acceptance region, thus we cannot reject H_0. As for Leuchars, this information tells us that the true p-value is likely to lie in the interval $(0.0287, 0.0387)$. Assuming we are working at a 5 significance level, we can still reject H_0, and conclude that for this station the daily expected rainfall is different in the summer compared to the winter.

Seasonal Variance in Probability of Rainfall

In order to get a better understanding of how the probability of rainfall changes throughout the year, we will carry out another Monte-Carlo Permutation test, this time on the hypotheses

$$H_0: \text{The daily probability of rainfall is the same in winter as in the summer}$$
$$H_1: \text{The daily probability of rainfall is different in winter and in summer}$$

This time, we will use $T=|\text{winter empirical nonzero proportion}-\text{summer empirical nonzero proportion}|$ as the test statistic. The original data set gives the T-values for each station as:

Station ID	Summer Rainfall Probability	Winter Rainfall Probability	T-Value
UKE00105874	0.3065366	0.3583567	0.0518201
UKE00105875	0.2856106	0.3279466	0.0423360
UKE00105884	0.2981774	0.3463426	0.0481652
UKE00105885	0.2898350	0.3387321	0.0488971
UKE00105886	0.2545244	0.2791183	0.0245940
UKE00105887	0.2887256	0.3286598	0.0399342
UKE00105888	0.2507553	0.2815804	0.0308252
UKE00105930	0.3289275	0.3749528	0.0460253

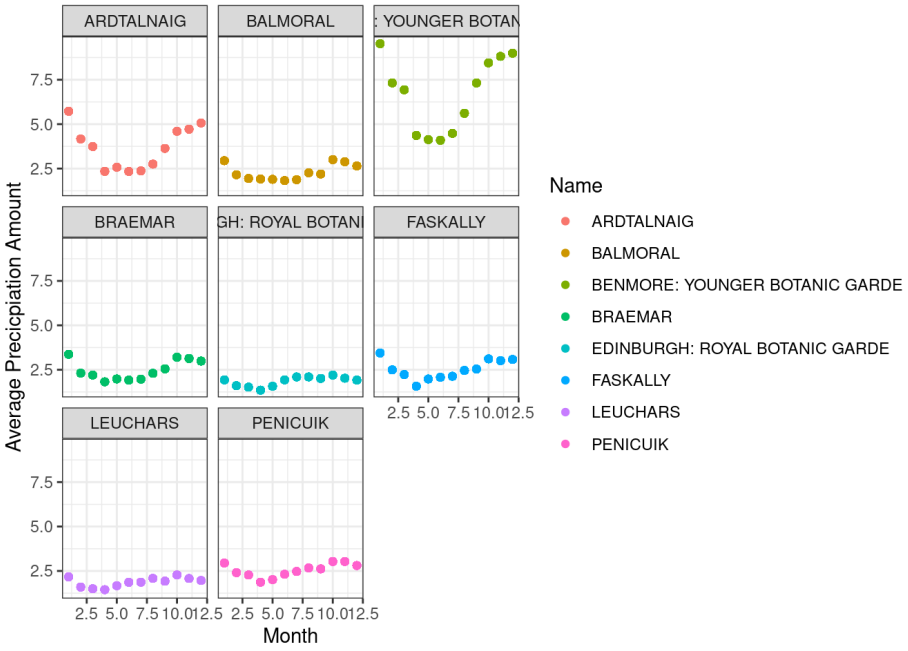
Taking $10,000$ permutations of the data for each station, and using $\frac{1}{J}\sum_{j=1}^{J}\delta_{(j)}$ where $\delta_{(j)}$ is defined as in the previous section, we get the estimated p-values for each station as:

Station Name	Station ID	p-value
BRAEMAR	UKE00105874	0
BALMORAL	UKE00105875	0
ARDTALNAIG	UKE00105884	0
FASKALLY	UKE00105885	0
LEUCHARS	UKE00105886	0
PENICUIK	UKE00105887	0
EDINBURGH: ROYAL BOTANIC GARDE	UKE00105888	0
BENMORE: YOUNGER BOTANIC GARDE	UKE00105930	0

Since each of these estimated p-values is 0, and we used $10,000$ permutations in our test, we get a confidence interval for p as $(0, 1-0.025^{1/10,000}) = (0, 3.69\times10^{-4})$. This means that we can again say that for any significance level above 3.69×10^{-4} there's a 95 chance our true p-value lies in the critical region. Thus, we can reject H_0, and conclude that at every station the daily probability of rainfall is affected by the season.

Spatial Weather Prediction

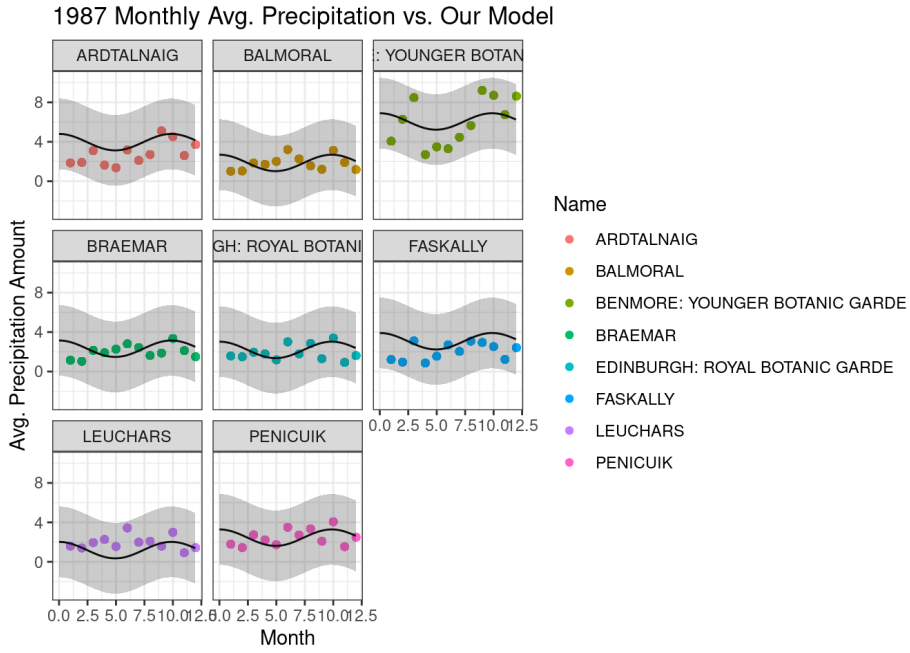
Below is a plot of the average daily precipitation in each month, averaged out across all the year in our `ghcnd_values` data set.



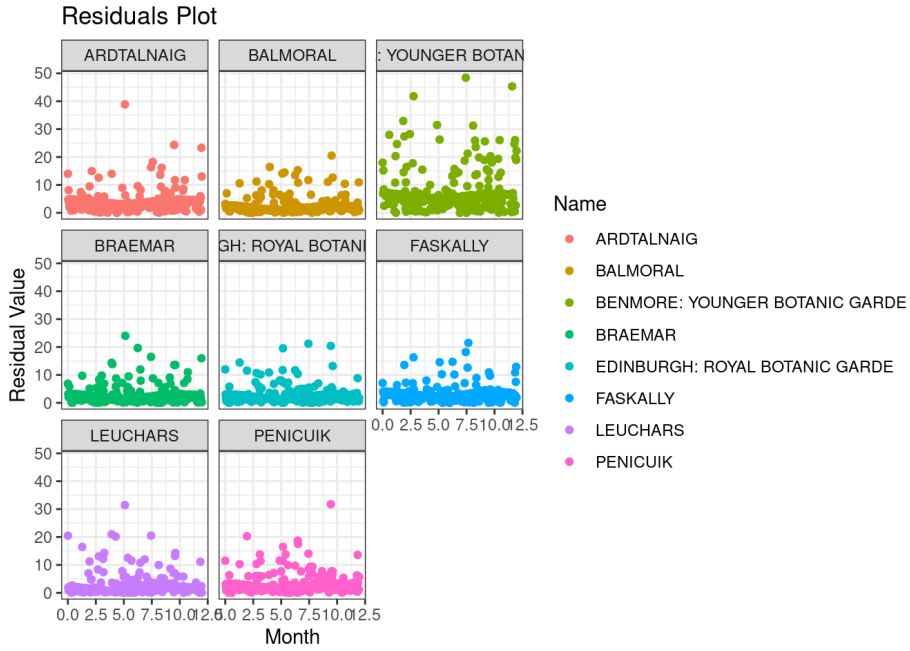
This graph shows that at each location the monthly average precipitation follows a wave-like curve with peaks in January and October meaning we can model this easily using a $\cos$ wave. We also see that for stations which are further west such as Ardtalnaig and Benmore, the amplitude of this curve is greater since there is a higher average rainfall in the winter months compared to the other stations. Now, using the `lm` function we can estimate a multiple linear regression model for the monthly average rainfall as

$$y_i = 25.84 - 2.205\theta_{i1} - 0.5493\theta_{i2} + 8.941\times10^{-4}\theta_{i3} + 0.8333\theta i4 + e_i$$

where for each $(\theta_{i1},\theta_{i2},\theta_{i3},\theta_{i4}) = (x_{i1},x_{i2},x_{i3},\cos(\frac{365x_{i4}}{300}2\pi)$ and $x_{i1},x_{i2},x_{i3},x_{i4}$ are the values of the longitude, latitude, elevation, and DecYear given in the `ghcnd_values` data frame respectively, and e_i is the random error component. Here I chose to model the seasonal variation using $\cos(\frac{365x_{i4}}{300}2\pi)$ since the peaks of this function are in January and October, thus follows the seasonal trend from the above figure fairly accurately. The figure below plots the prediciton of this model on the data subset of `ghcnd_values` from the year 1987 (chosen randomly for the sake of testing the model).



We can also plot the residuals of our model for this year.



As we can see, in this example our model work pretty well since all of the points in the data set lie within the 95% prediction interval of our model, and the Residuals plot shows that most of the fitted values lie within $10mm$ of the true precipitation value on a given day in this year.

Assesment of our Model

We now want to see whether the model is equally good at predicting the rainfall for each station, and whether the prediction accuracy is the same across the year. Carrying out a cross-validation test on each station we get scores for each station as:

Station ID	Station Name	Average Squared Error Score	Average Dawid-Sebastiani Score
UKE00105874	BRAEMAR	22.80353	17081.11
UKE00105875	BALMORAL	20.56109	16878.19
UKE00105884	ARDTALNAIG	42.77531	64036.09
UKE00105885	FASKALLY	22.96315	38031.08
UKE00105886	LEUCHARS	18.35987	21605.78
UKE00105887	PENICUIK	25.44270	18796.65
UKE00105888	EDINBURGH: ROYAL BOTANIC GARDE	17.18348	14502.72

Station ID	Station Name	Average Squared Error Score	Average Dawid-Sebastiani Score
UKE00105930	BENMORE: YOUNGER BOTANIC GARDE	106.05406	41510.11

As we can see from these results, our model is definitely better at predicting results for certain stations compared to others. In particular, it does a poor job at making predictions for Benmore and Ardtalnaig, possibly due to the fact that these are the stations whose precipitation values vary the most across the year. The lowest squared-error and Dawid-Sebastiani score here corresponds to the Edinburgh Station. This coincides with the fact that Edinburgh did not show evidence for having different precipitation variation in the different seasons (as calculated in the permutation test earlier). We can theorise from this information that our model may not account very well for seasonal variation in precipitation.

Now, we can also carry out a cross-validation test to see whether the model makes better predictions for certain months compared to others.

Month	Average Squared Error Score	Average Dawid-Sebastiani Score
1	53.71997	69197.44
2	35.70544	45975.36
3	34.00025	43531.06
4	23.24835	29518.05
5	22.33497	27947.90
6	26.03411	32652.12
7	27.67017	35645.08
8	37.03876	48516.40
9	41.76377	52339.86
10	51.36715	66177.40
11	48.89556	64140.86
12	51.79060	65563.97

As we see from these results, our model makes better predictions for the summer months than it does for the winter months as both the squared-error and Dawid-Sebastiani scores are higher in winter than in summer. I believe this is because our model may not account for the greater variation of the precipitation in the winter months, hence the actual precipitation values in these months lie further for the model's expected value than it does in the summer months. This also backs up our theory that the model does not account for seasonal variation in precipitation very well.

Code appendix

Function definitions

```
# Axel Eichelmann (s2030757, axeleichelmann)

# Place your function definitions that may be needed in analysis.R and report.Rmd
# in this file, including documentation.
# You can also include any needed library() calls here


library(shiny)
library(tidyverse)
library(dplyr)
library(StatCompLab)

data(ghcnd_stations, package = "StatCompLab")
data(ghcnd_values, package = "StatCompLab")

#Function which computes the test-stat for each station, w\ an option to randomize the season labels
test_stat <- function(randomise = FALSE) {
  PRCP_data <- mutate(ghcnd_values, Season = if_else(Month %in% c(1,2,3,10,11,12), 'Winter', 'Summer')) %>%
    filter(Element == "PRCP") %>% group_by(ID) %>% select(ID, Value, Season)
  if (randomise == TRUE) {
    for (i in ghcnd_stations$ID) {
      var <- PRCP_data$Season[PRCP_data$ID==i]
      PRCP_data$Season[PRCP_data$ID==i] <- sample(var, size = length(PRCP_data$Season[PRCP_data$ID == i]))
    }
  }
  Avg_Summer_Rf <- PRCP_data %>% summarize(Summer_rf = mean(Value[Season=='Summer']))
  Avg_Winter_Rf <- PRCP_data %>% summarize(Winter_rf = mean(Value[Season=='Winter']))
  table <- Avg_Summer_Rf %>% right_join(Avg_Winter_Rf, by = "ID")
  table <- mutate(table, t_value = abs(Summer_rf-Winter_rf))
  colnames(table) <- c("Station ID", "Avg. Summer Rf", "Avg. Winter Rf", "T-Value")
  table
}

# Create function which calculates the empirical non-zero proportion of rainfall in Winter and Summer
prcp_prob <- function(randomise = FALSE){
  PRCP_data <- mutate(ghcnd_values, Season = if_else(Month %in% c(1,2,3,10,11,12),'Winter','Summer')) %>%
    filter(Element == "PRCP") %>% group_by(ID) %>%
    mutate(Rainfall = if_else(Value == 0, 0, 1)) %>% select(ID, Rainfall, Season)
  if (randomise== TRUE){
    for (i in ghcnd_stations$ID) {
      var <- PRCP_data$Season[PRCP_data$ID==i]
      PRCP_data$Season[PRCP_data$ID==i] <- sample(var, size = length(PRCP_data$Season[PRCP_data$ID == i]))
    }
  }

  Summer_prob <- PRCP_data %>% summarise(summer_prob = sum(Rainfall[Season == 'Summer'])/length(Rainfall))
  Winter_prob <- PRCP_data %>% summarise(winter_prob = sum(Rainfall[Season == 'Winter'])/length(Rainfall))
  table <- right_join(Summer_prob, Winter_prob, by = "ID")
  table <- mutate(table, t_value = abs(summer_prob-winter_prob))
  colnames(table) <- c("Station ID", "Summer Rf Prob", "Winter Rf Prob", "T-Value")
  table
}

#version of data set with monthly average precipitation values in each year
monthly_avg_data <- ghcnd_values %>% filter(Element == "PRCP") %>% group_by(ID,Year,Month) %>%
                    mutate(monthly_avg = mean(Value)) %>% right_join(ghcnd_stations, by = "ID")

# average monthly precipitation across all years
graph_monthly_avg <- ghcnd_values %>% filter(Element == "PRCP") %>% group_by(ID, Month) %>%
                     mutate(monthly_mu = mean(Value)) %>% left_join(ghcnd_stations, by = "ID")

model <- lm(monthly_avg ~ Longitude+Latitude+Elevation + cos((DecYear-Year)*2*pi*365/300), data = monthly_avg_dat
a)


# code for the station Cross Validation
complete_ghcnd <- ghcnd_values %>% filter(Element == "PRCP") %>%
                  right_join(ghcnd_stations, by = "ID") %>% group_by(ID)

ID_set <- ghcnd_stations$ID[-1]
cv_model_data <- complete_ghcnd %>% filter(ID %in% ID_set)
CV_model <- lm(Value ~ Longitude + Latitude + cos((DecYear-Year)*2*pi*365/300), data = cv_model_data)
predict_data <- complete_ghcnd %>% filter(ID == ghcnd_stations$ID[1])
pred_values <- data.frame(predict_data, pred = predict(CV_model, predict_data),
                                        pred_sd = predict(CV_model, predict_data, se.fit = TRUE))
```

```
for (i in 2:8){
  ID_set <- ghcnd_stations$ID[-i]
  cv_model_data <- complete_ghcnd %>% filter(ID %in% ID_set)
  CV_model <- lm(Value ~ Longitude + Latitude + cos((DecYear-Year)*2*pi*365/300), data = cv_model_data)
  predict_data <- complete_ghcnd %>% filter(ID == ghcnd_stations$ID[i])
  pred_values <- rbind(pred_values, data.frame(predict_data, pred = predict(CV_model, predict_data),
                                               pred_sd = predict(CV_model, predict_data, se.fit = TRUE)))
}

# code for the month cross validation
month_ghcnd <- ghcnd_values %>% filter(Element == "PRCP") %>%
  right_join(ghcnd_stations, by = "ID") %>% group_by(Month)

month_set <- c(1,2,3,4,5,6,7,8,9,10,11,12)[-1]
cv_model_data <- month_ghcnd %>% filter(Month %in% month_set)
CV_model <- lm(Value ~ Longitude + Latitude + cos((DecYear-Year)*2*pi*365/300), data = cv_model_data)
predict_data <- month_ghcnd %>% filter(Month == 1)
pred_month_values <- data.frame(predict_data, pred = predict(CV_model, predict_data),
                          pred_sd = predict(CV_model, predict_data, se.fit = TRUE))

for (i in 2:12){
  month_set <- c(1,2,3,4,5,6,7,8,9,10,11,12)[-i]
  cv_model_data <- month_ghcnd %>% filter(Month %in% month_set)
  CV_model <- lm(Value ~ Longitude + Latitude + cos((DecYear-Year)*2*pi*365/300), data = cv_model_data)
  predict_data <- month_ghcnd %>% filter(Month == i)
  pred_month_values <- rbind(pred_month_values,
                             data.frame(predict_data, pred = predict(CV_model, predict_data),
                                        pred_sd = predict(CV_model, predict_data, se.fit = TRUE)))
}
```

Analysis code

```
# Axel Eichelmann (s2030757, axeleichelmann)
# Place analysis code that may take too long to run every time the report.Rmd
# document is run.
# Run the code in this file with
#   source("analysis.R")
# in a fresh R session to ensure consistency of the results.

# Load function definitions
source("functions.R")


# A code=readlines() code chunk in the report appendix will include the code
# in the report, without running it.
#
# You can place long-running analysis code in this file,
# and save results using
#   saveRDS(object, file = "data/object.rds")
# When the results are needed in the report.Rmd file, use
#   object <- readRDS(file = "data/object.rds")
# Make sure to use different filenames for each object, such as the object
# name itself.
#
# Remember to rerun this code to save new results when you change the code.
#
# The .gitignore file has been setup so that it ignores .rds files in the data/
# folder, so that you don't accidentally make git handle large binary data files.


#calculate the p-values of mean rainfall for each station
p <- 100
p_vals <- c(0,0,0,0,0,0,0,0)
orig_t <- as.vector(test_stat()$'T-Value')
for (i in 1:p){
  p_test_stat <- as.vector(test_stat(randomise = TRUE)$'T-Value')
  diff <- p_test_stat-orig_t
  for (i in 1:8){
    if (diff[i]>0){
      p_vals[i] = p_vals[i]+1
    }
  }
}
p_vals_mean <- p_vals/p
saveRDS(p_vals_mean, file = "data/p_vals_mean.rds")


#calculate the p-values of rainfall probability for each station
p <- 10000
p_vals_prob <- c(0,0,0,0,0,0,0,0)
orig_t <- as.vector(prcp_prob()$'T-Value')
for (i in 1:p){
  p_test_stat <- as.vector(prcp_prob(randomise = TRUE)$'T-Value')
  diff <- p_test_stat-orig_t
  for (i in 1:8){
    if (diff[i]>0){
      p_vals[i] = p_vals[i]+1
    }
  }
}
p_vals_prob <- p_vals_prob/p
saveRDS(p_vals_prob, file = "data/p_vals_prob.rds")
```