MAT215: Project 5

Zaigham Abbas Randhawa

Wednesday, 05/04/2017

Data

The data you will be working with is a part of a project that examines the issue of storm water runoff in Lancaster, PA. http://forestsforwatersheds.org/reduce-stormwater/ describes stormwater runoff as "rainfall that flows over the ground surface. It is created when rain falls on roads, driveways, parking lots, rooftops and other paved surfaces that do not allow water to soak into the ground. Stormwater runoff is the number one cause of stream impairment in urban areas. Where rain falls on paved surfaces, a much greater amount of runoff is generated compared to runoff from the same storm falling over a forested area. These large volumes of water are swiftly carried to our local streams, lakes, wetlands and rivers and can cause flooding and erosion, and wash away important habitat for critters that live in the stream."

We would like to know if there is a difference between different tree species in their ability to remove water out of the ground. The data contains the following variables:

- *month*, *day*, *hour* - designate a time point at which the data point was collected
- *date*, *dateF* - designate the date of data collection
- *sunrise*, *sunset* - designate the sunrise and sunset time for the day on which data was collected
- *location* - designates the park in Lancaster where the data was collected
- *units* - units of measurement
- *plant* - plant species
- *temperature* - hourly temperature
- *rh* - relative humidity
- *vpd* - vapor pressure deficit
- *rain* - amount of rainfall
- *srad* - solar radiation
- *y* - response: the amount of sap flowing up (positive values - transpiration) or down (negative values - foliar water uptake) the tree

For your analysis, keep only the data with positive values of y. Also, keep only the data with the units of L/hr. You should have $n = 45228$ observations.

For all your analyses tranform your response y by calculating $log(y)$ and use $log(y)$ as your response.

Questions

Question 1

To get a feel for what our data looks like, we should make some plots. We will look at our data over time. To create the time variable (i.e. convert dateF, which is character variable, to a numeric date R would understand) use the following R command

```
library(mosaic)
library(corrplot)
library(ggplot2)

mydata<-read.csv(file="lancTranspiration2016-01-11.csv", header=TRUE)
mydata$datum <- strptime(mydata$dateF, format = "%Y-%m-%d %H:%M:%S")
newdata=subset(mydata,y>=0&units=="L/hr",select=c(datum,y,plant,location,temper
                                                  ,rh,vpd,rain,srad))
```

a. Make a separate plot for each plant in your data. Your x-axis should be the date (i.e. the new *datum* variable) and your y-axis should by y, the sap flow.

This data is very noisy. To smoothen out the plot to see the data patterns, add a moving average line to each plot. See below the R code to compute a moving average with *movingAverage* function. You have to decide what value for n looks good. n too small will not smoothen out your curve a lot, while an n too big will tend to smoothen out too much.

```
plot(newdata$datum[newdata$plant=="AmberMaple"],
newdata$y[newdata$plant=="AmberMaple"],las=1,col="black",type="l",
ylab="Sap Movement Rates (L/hr)",xlab="Time (months)")
movAvg <- movingAverage(newdata$y[newdata$plant=="AmberMaple"], n = 170,
                        centered = TRUE)
lines(newdata$datum[newdata$plant=="AmberMaple"], movAvg, col = "blue", lwd = 2)
```

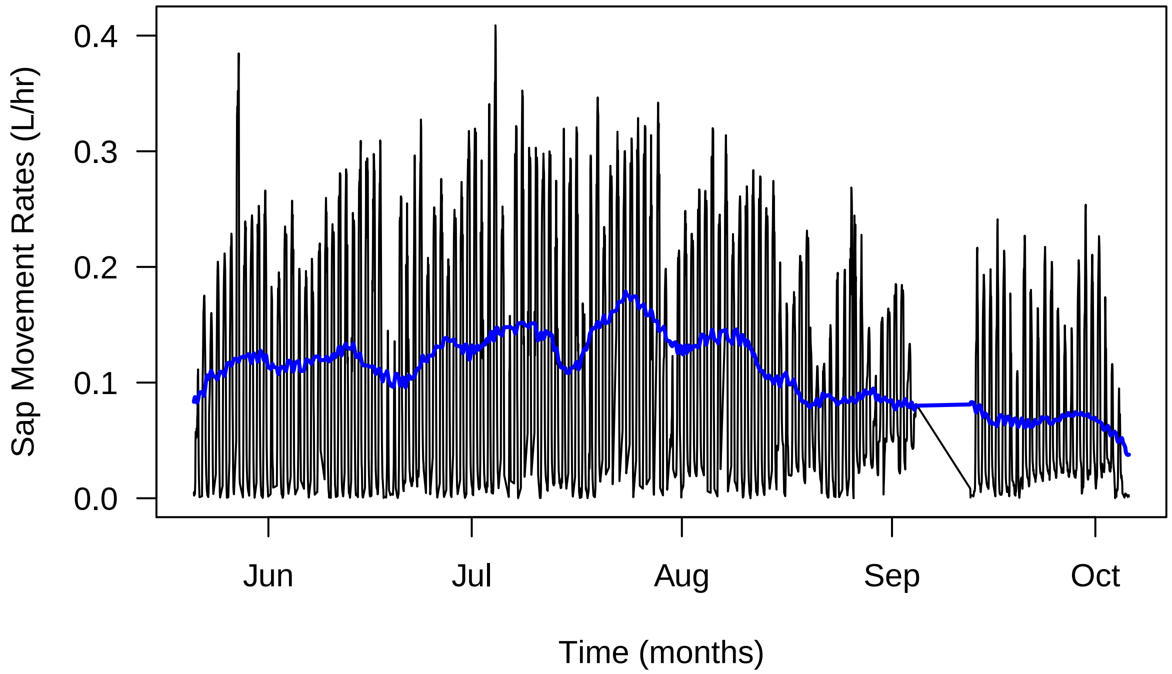


Figure 1a. Sap Flow rates over the year for plants of type Amber Maple

```
plot(newdata$datum[newdata$plant=="BlackAsh2"],
newdata$y[newdata$plant=="BlackAsh2"],las=1,col="black",type="l",
ylab="Sap Movement Rates (L/hr)",xlab="Time (months)")
movAvg <- movingAverage(newdata$y[newdata$plant=="BlackAsh2"], n = 170,
                        centered = TRUE)
lines(newdata$datum[newdata$plant=="BlackAsh2"], movAvg, col = "blue", lwd = 2)
```

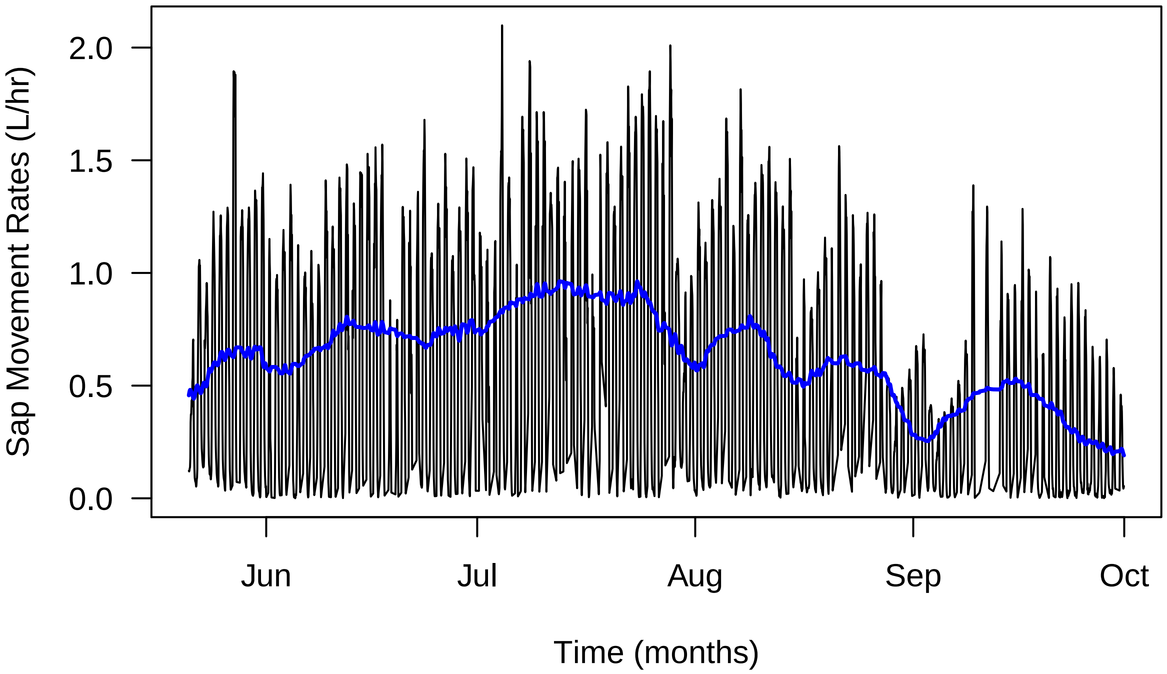


Figure 1b. Sap Flow rates over the year for plants of type BLack Ash 2

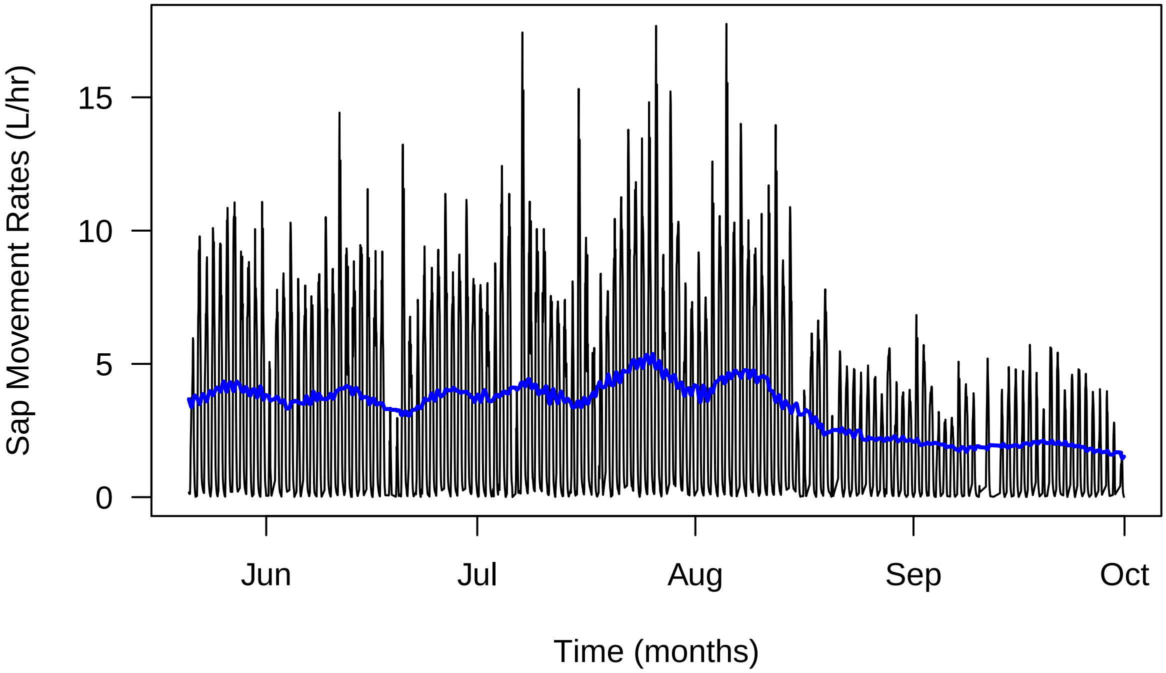


Figure 1c. Sap Flow rates over the year for plants of type NM 1

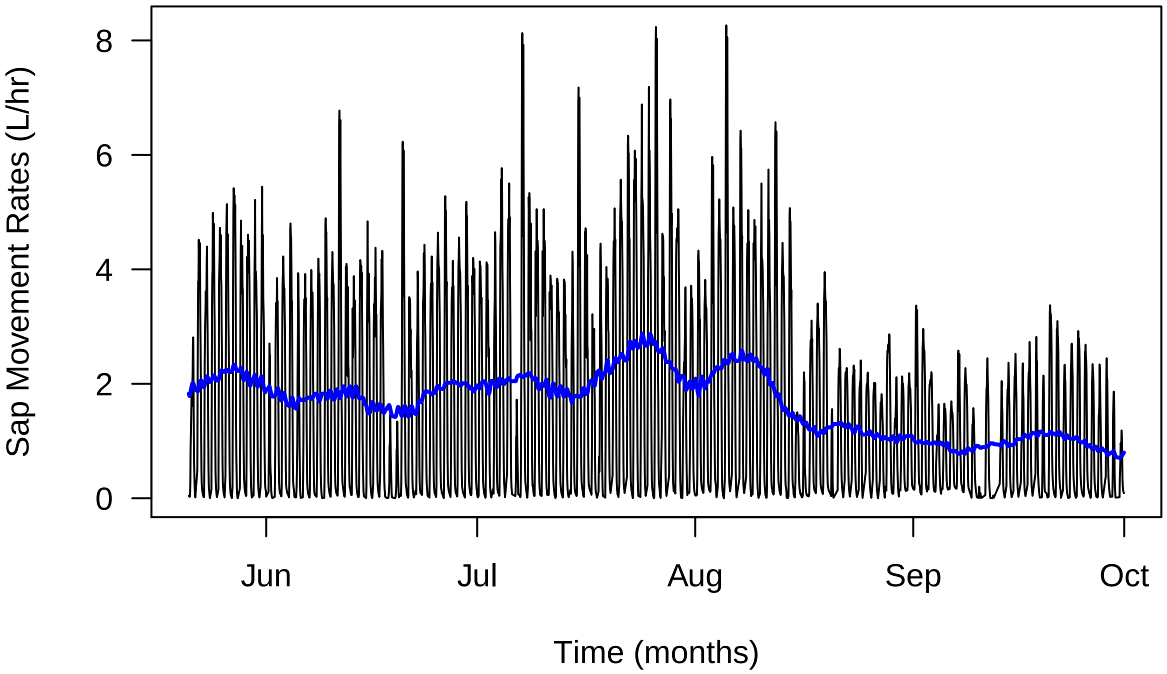


Figure 1d. Sap Flow rates over the year for plants of type NM 2

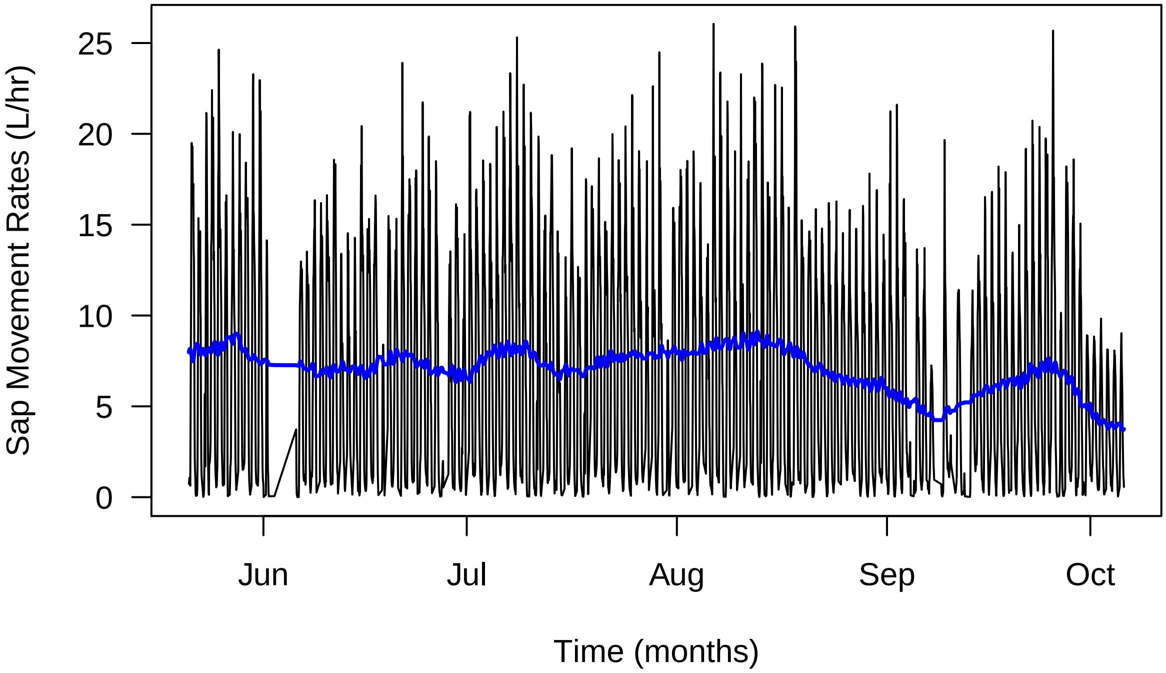


Figure 1e. Sap Flow rates over the year for plants of type Sycamore 1

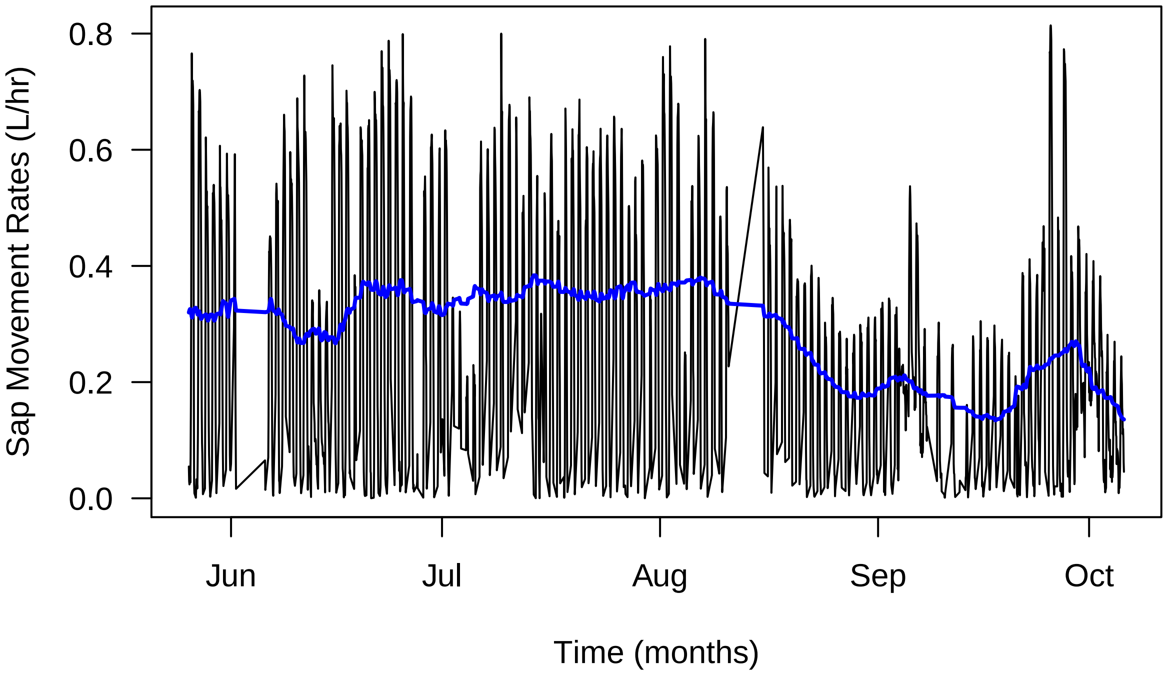


Figure 1f. Sap Flow rates over the year for plants of type Red Maple 1

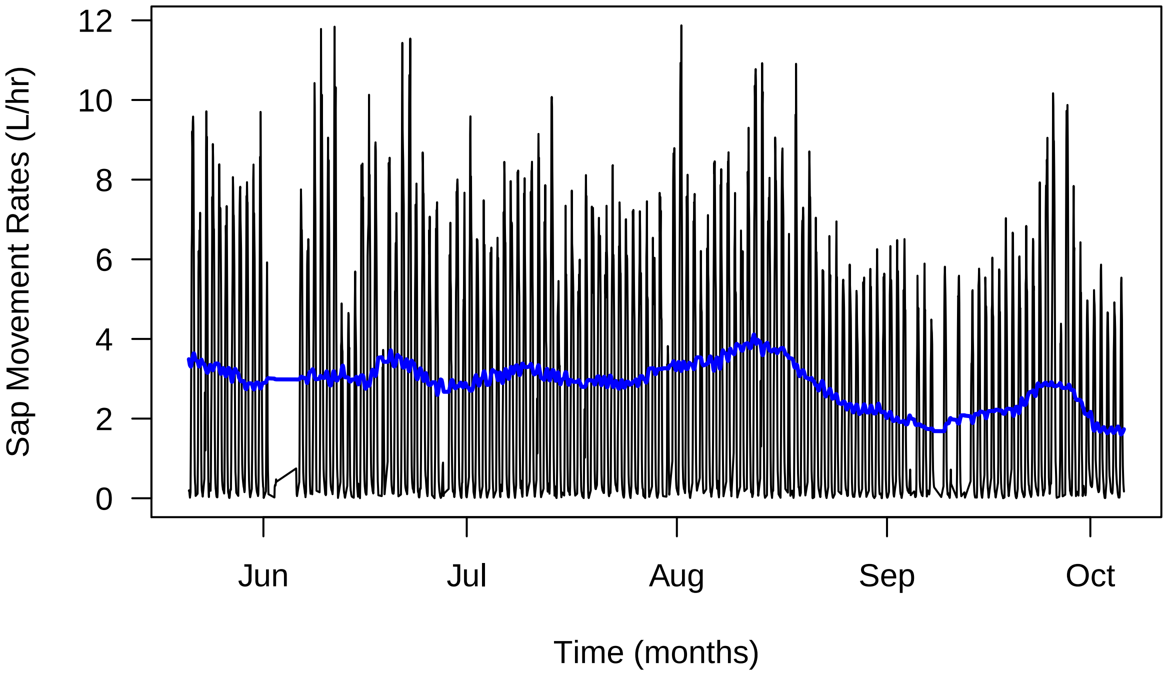


Figure 1g. Sap Flow rates over the year for plants of type Red Maple 2

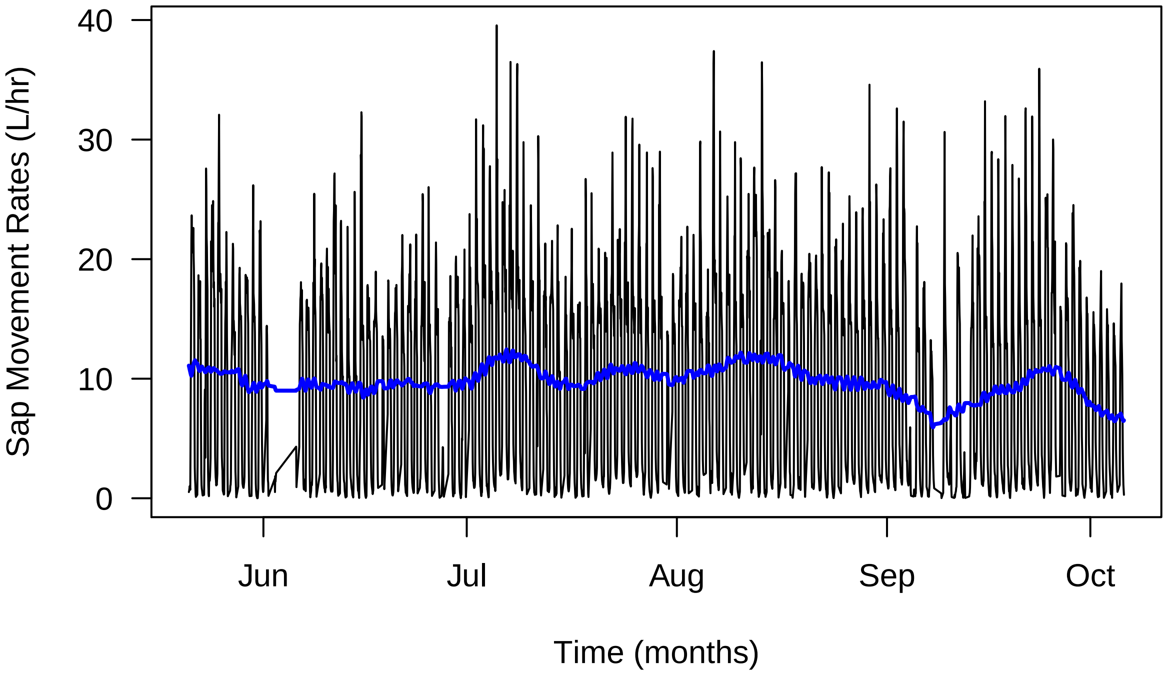


Figure 1h. Sap Flow rates over the year for plants of type Sycamore 2

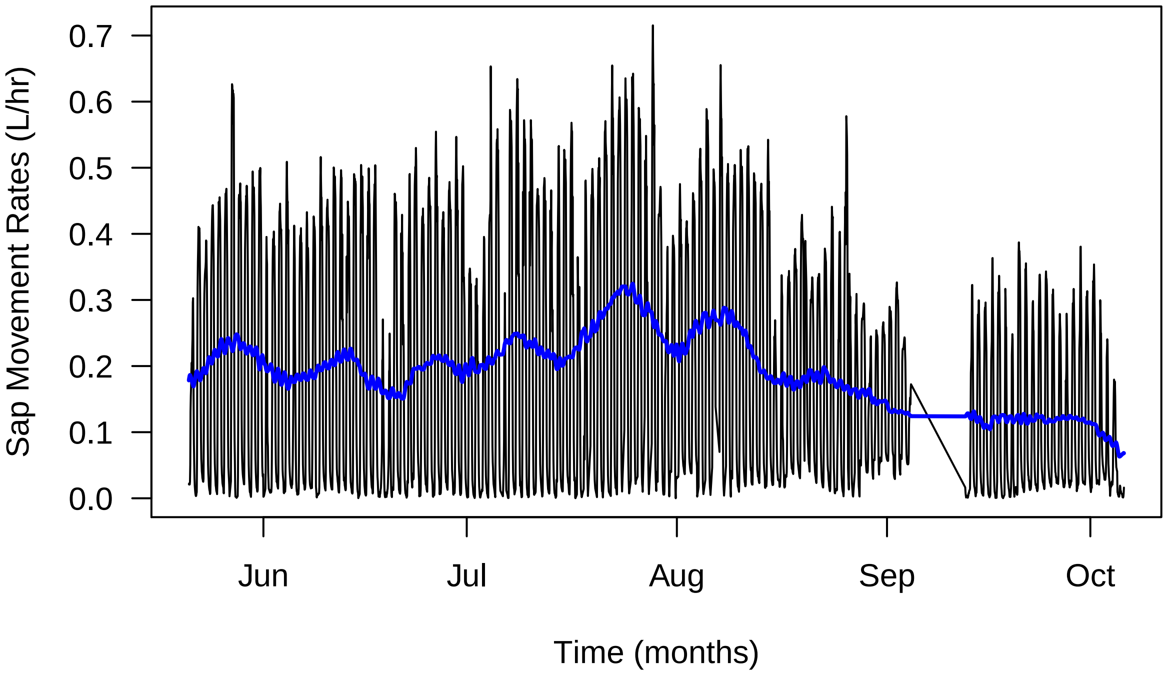


Figure 1i. Sap Flow rates over the year for plants of type Green Vase

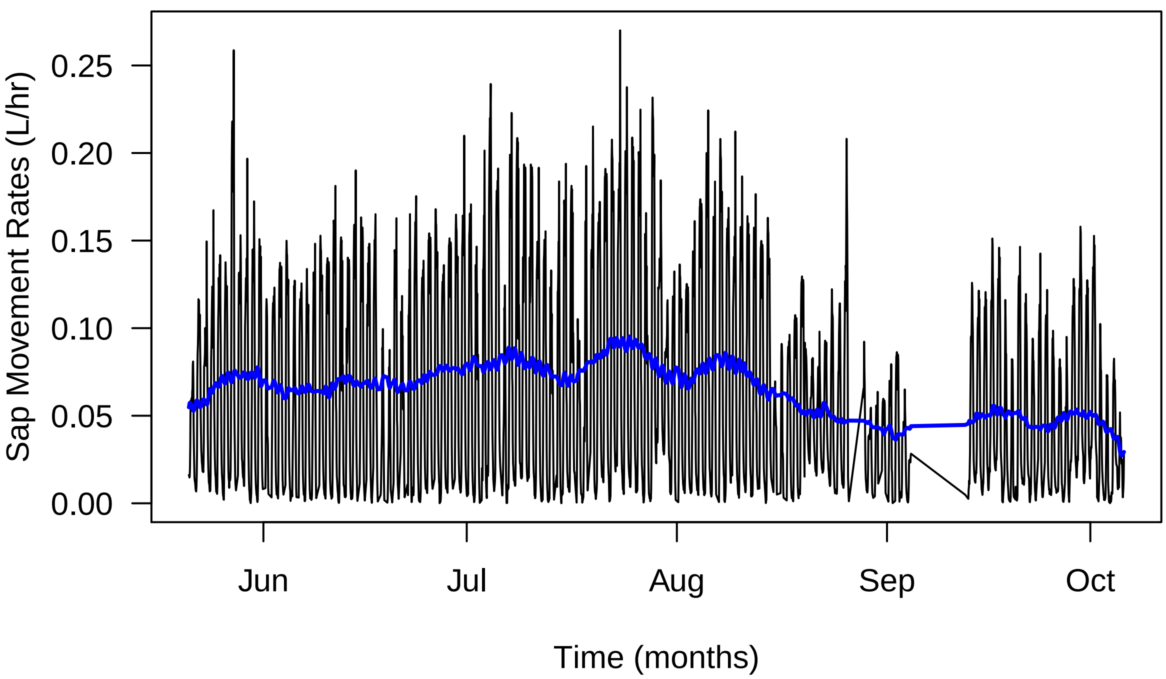


Figure 1j. Sap Flow rates over the year for plants of type Red Maple

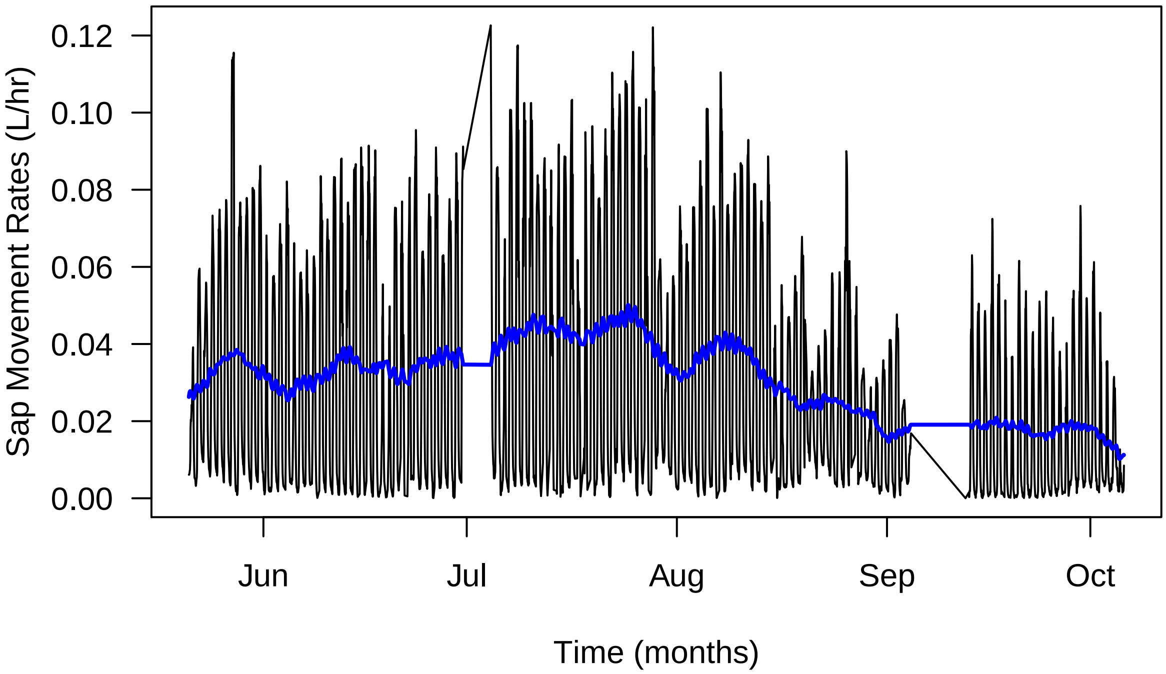


Figure 1k. Sap Flow rates over the year for plants of type Gold Ginko

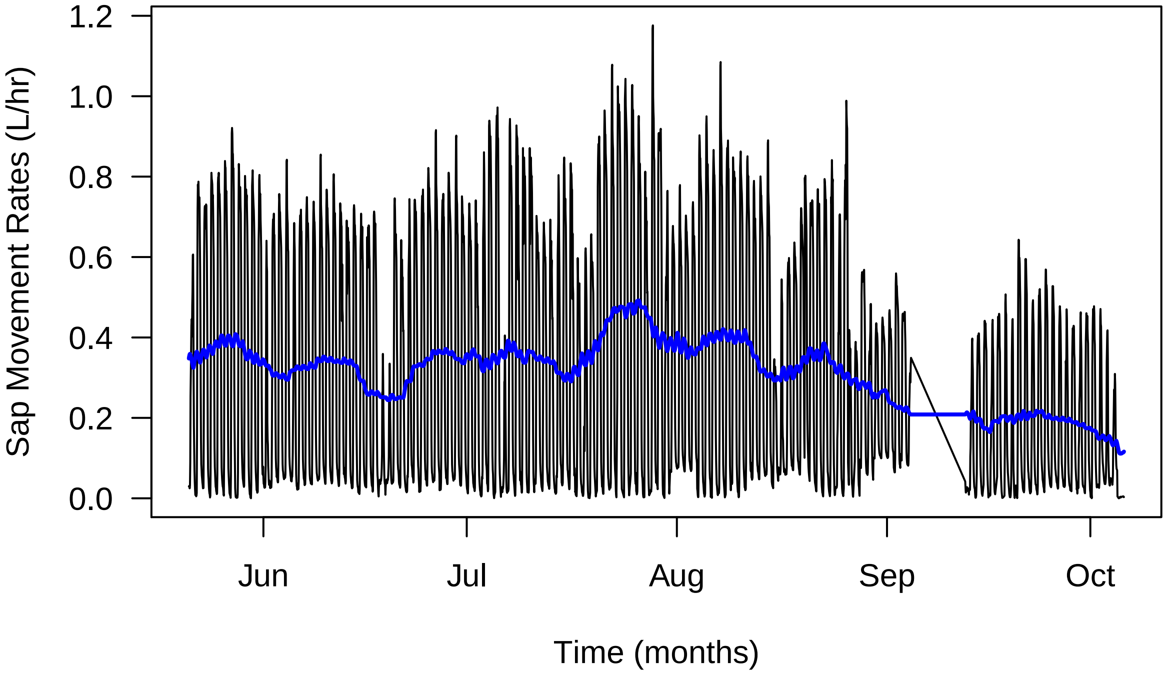


Figure 11. Sap Flow rates over the year for plants of type Paperback Maple

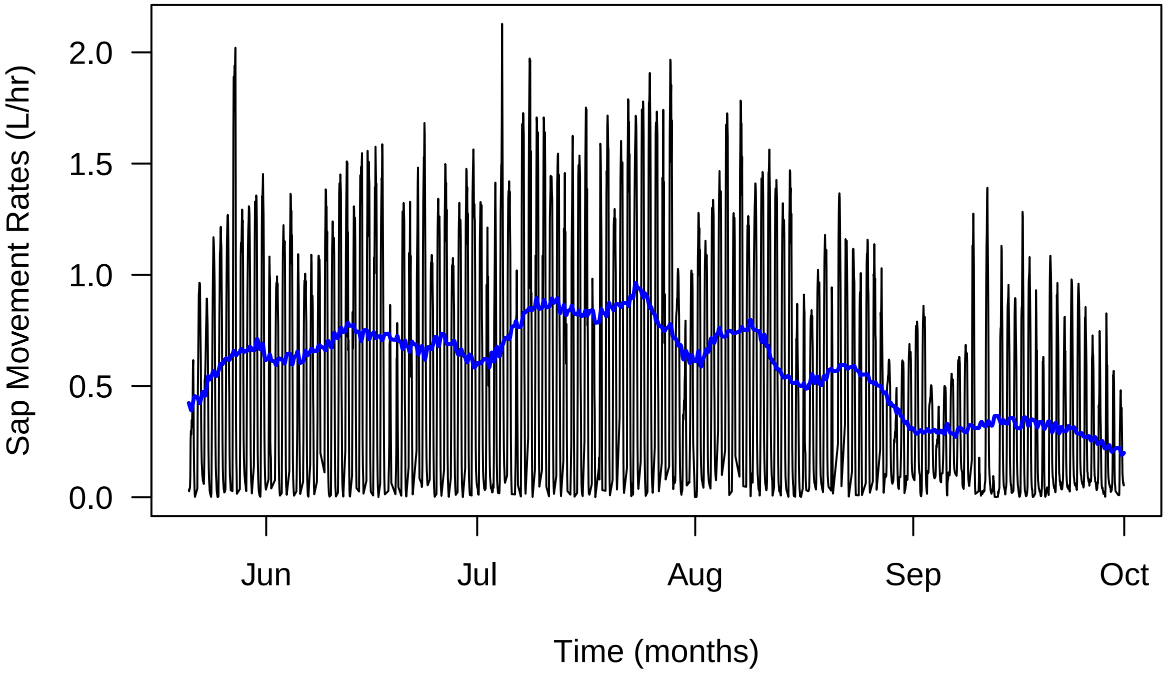


Figure 1m. Sap Flow rates over the year for plants of type Black Ash 1

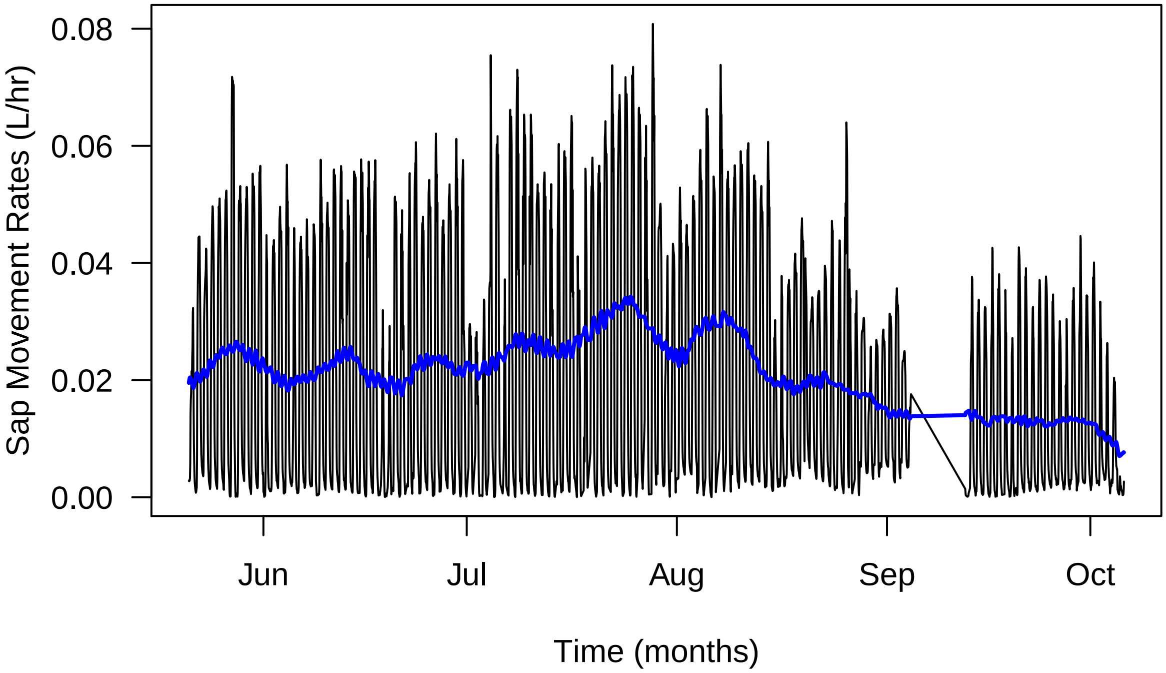


Figure 1n. Sap flow rates over the year for plants of type Parrottia

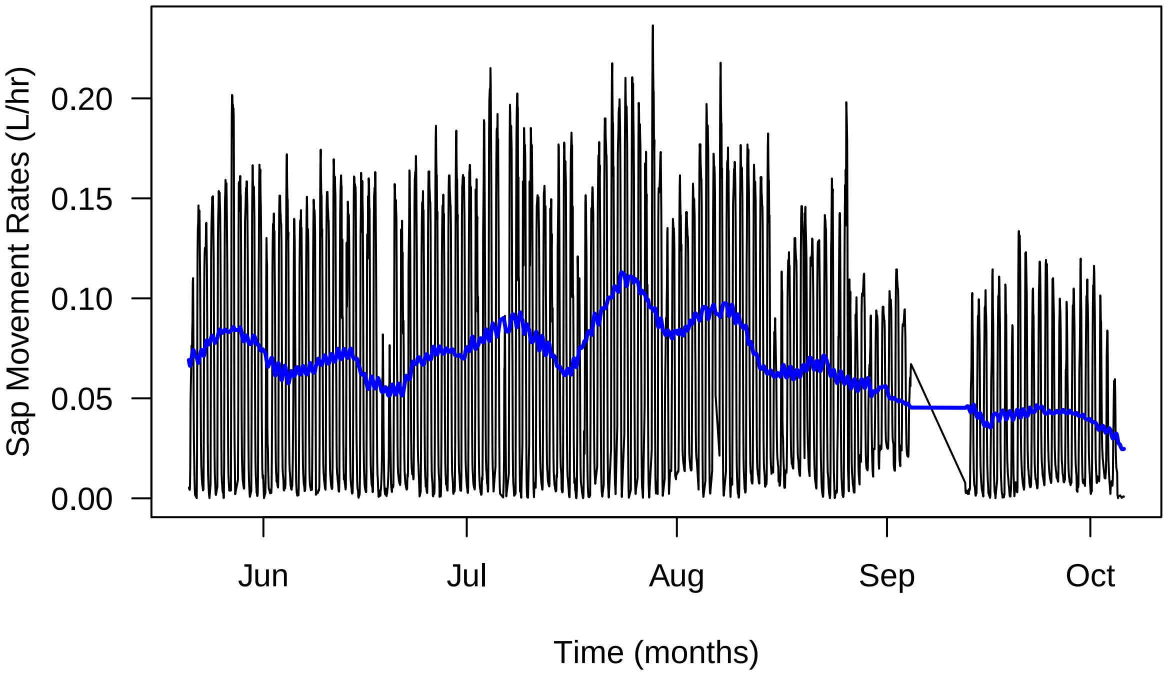


Figure 1o. Sap flow rates over the year for plants of type Sugar Maple

Do you see any clear patterns in your plots? Do data patterns differ for different plants? How about different locations? Is there any time effect? (10 pts)

For most of the plants (except Sycamore and Red Maple varieties) the sap flow rate reached their shortly before the month of august. The sap flow rates also decreased sharply (e.g. Amber Maple, NM varieties) or became constant (e.g. Green Vase or Gold Ginko) immidiately after september. For the plants which go through the former change, the decrease is followed by an increase (e.g. Sycamore and Black Ash Varieties).

All of the plants at the buchanan park (which include Amber Maple, Green Vase, Gold Ginko, Paperback Maple, Sugar Maple, Red Maple, and Parrottia) in their distributions show a single straight line with a negative slope slightly after september.

If plants are of the same type and are located at the same location then they have almost identical sap flow rate distributions. Examples justifying this can be found when the distribtuions for Black Ash 1, Red Maple 1, NM 1, and Sycamore 1 are compared with those for Black Ash 2, Red Maple 2, NM 2, and Sycamore 2 respectively. If the location changes, then the distribution also changes. Evidence for this can be found when the distributions for Red Maple 1 or 2 are compared with Red Maple's-Red Maple is located at buchanan while Red Maple 1 and 2 are located at the reservoir.

Question 2

a. We would like to investigate whether there is a difference between different parks considered. Construct a table that contains the averages and 95% CI's for each level of variable *location*. (5 pts)

Location	Y (L/hr)	95% CI
Longspark	1.610	(1.541, 1.679)
Buchanan	0.116	(0.067, 0.165)
Reservoir	5.482	(5.412, 5.551)

Table 1. Sap flow rates at various locations along with a 95 percent confidence interval

b. Construct a plot that contains the boxplots for each level of *location*. Alternatively, you can construct a plot that contains the averages and 95% CIs for each park. Does there seem to be a difference between the different parks? (5 pts)

```r
p <- ggplot(newdata, aes(x=location, y=y))
p + geom_boxplot() + scale_y_log10()+labs(x="Location",
y="Sap Movement Rate (L/hr)")+coord_flip(ylim = c(1e-6,1e2))
```

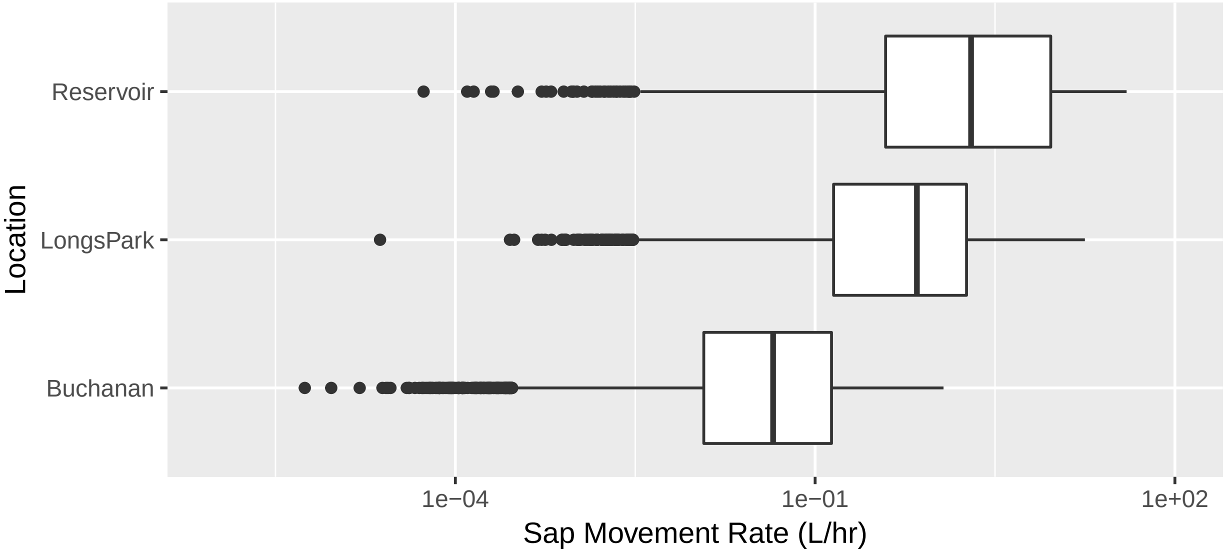


Figure 2. Sap flow rates of plants at different locations

Please note that the scale here is logarithmic, and its interpretation is quiet different from an un-logarithmic scale. The positioning of a line on the scale matters a lot in addition to its length: If we have 2 lines of equal magnitude but one of them is appears up a bit earlier than the other on the scale, then the former one has a smaller range than the latter.

So coming back to our data, there is deinitely a difference between the distributions of the sap movement rates for different locations. Buchanan's range, and minimum, maximum are the smallest.

These values for LongsPark are somewhat larger. However, the reservoir has the highest magnitudes for these statistics.

c. Fit an appropriate model to investigate your claim from part (a). Write down your hypotheses and make your conclusions in terms of the problem at the level of significance of 5%. (5 pts)

```
Location_Model<- lm(log(newdata$y) ~ newdata$location)
anova(Location_Model)
```

```
## Analysis of Variance Table
##
## Response: log(newdata$y)
##                      Df Sum Sq Mean Sq F value    Pr(>F)
## newdata$location      2 107975   53987   16379 < 2.2e-16 ***
## Residuals         40081 132114       3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

H_0 states that a plant's location has no affect on its sap flow rates. Its H_A is that a plant's location has an affect on the sap flow rates.

The p-value in the anova report for the 'plant' variable is below 0.05, therefore we can reject the null hypothesis and accept the alternative one. Hence it can be concluded that a plant's particular location does significantly affect its sap flow rates.

Question 3

Repeat Question 2 - label the parts a, b, and c in your report - but replace variable *location* with variable *plant*, i.e. we would like to see if there is a difference between plants. (15 pts)

```
model_mean=lm(newdata$y~newdata$plant-1)
```

a.

Plant	Y (L/hr)	95% CI
AmberMaple	0.056	(-0.204, 0.317)
BlackAsh1	2.433	(2.177, 2.689)
BlackAsh2	2.158	(1.881, 2.434)
GoldGinko	0.019	(-0.231, 0.268)
GreenVase	0.099	(-0.144, 0.342)
NM1	11.423	(11.167, 11.679)
NM2	8.135	(7.883, 8.386)
PaperbackMaple	0.163	(-0.0763, 0.402)
Parrottia	0.014	(-0.229, 0.258)
RedMaple	0.036	(-0.215, 0.287)
RedMaple1	1.341	(1.031, 1.652)
RedMaple2	15.175	(14.919, 15.431)
SugarMaple	0.037	(-0.207, 0.282)
Sycamore1	11.768	(11.518, 12.018)

Plant	Y (L/hr)	95% CI
Sycamore2	15.137	(14.888, 15.386)

Table 2. Sap flow rates for various plants along with a 95 percent confidence interval

b.

```
p <- ggplot(newdata, aes(x=plant, y=y))
p + geom_boxplot() + scale_y_log10()+labs(x="Plant Specie",
y="Sap Movement Rate (L/hr)")+coord_flip(ylim = c(1e-6,1e2))
```

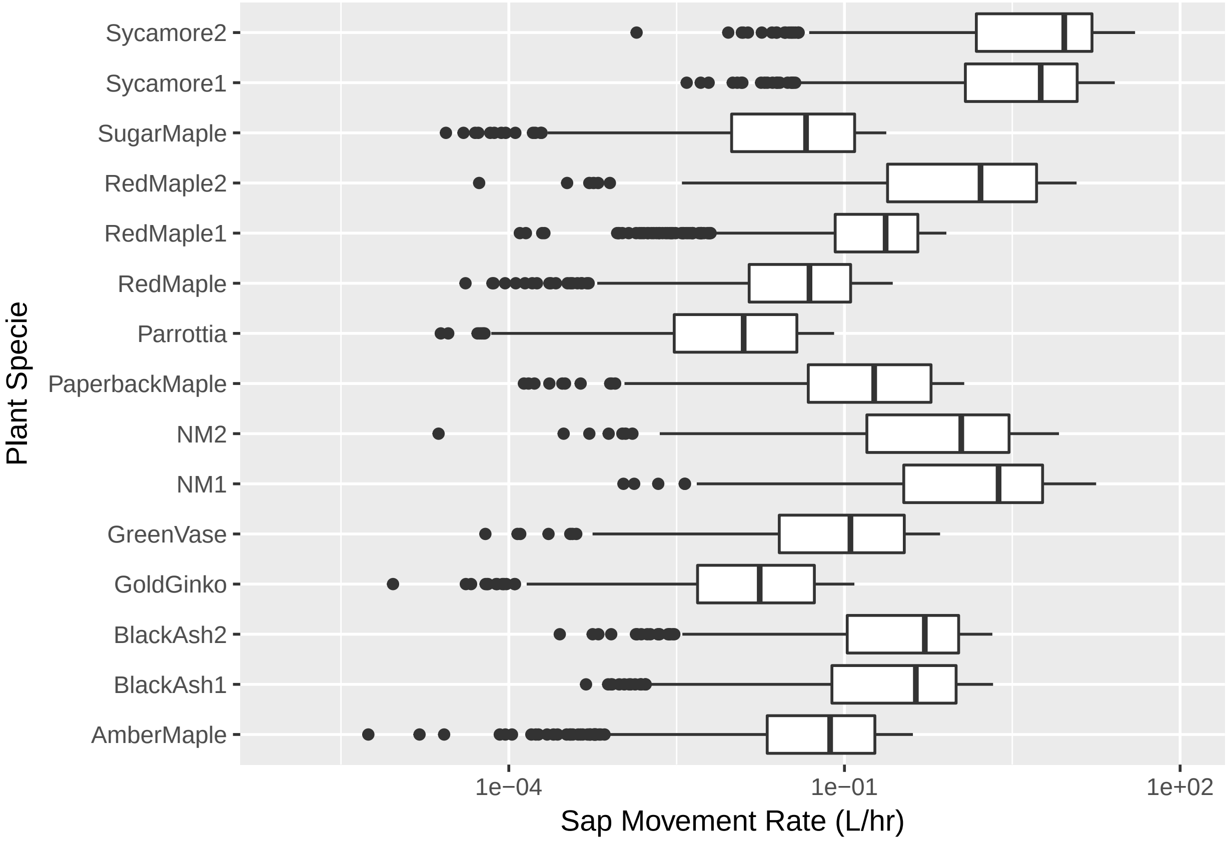


Figure 3. Sap flow rates for different plant species

Each specie has a distinct distribution for its sap movement rate, i.e. none of them are identical. Some of these distribution (e.g. Red Maple 2) are pretty wide while others are quite narrow (e.g. Gold Ginko).

c.

```
Plant_Model<- lm(log(newdata$y) ~ newdata$plant)
anova(Location_Model)
```

```
## Analysis of Variance Table
```

```
##
## Response: log(newdata$y)
##                      Df Sum Sq Mean Sq F value    Pr(>F)
## newdata$location      2 107975   53987   16379 < 2.2e-16 ***
## Residuals         40081 132114       3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

H_0 is that a plant's specie has no affect on its sap flow rates. A plausible H_A could be that a plant's specie does have an affect on its sap flow rates.

The p-value in the anova report for the 'plant' variable is below 0.05, therefore we can reject the null hypothesis and accept the alternative one. Hence it can be concluded that a plant's particular specie does significantly affect its sap flow rates.

Question 4

We also have environmental variables that could influence sap flow so we would like to consider them for analysis.

For parts (a) and (b) keep only the data for plant $BlackAsh1$ since, for the same day, all plants will have the same values for all environmental variables. Therefore, keeping the data for all plants would artificially inflate our sample size.

a. To investigate any potential issues of colinearity, create a scatterplot matrix between $temper$, rh, vpd, $rain$, and $srad$ using R's function $pairs$. Scatterplot matrix contains all scatterplots between the 4 environmental variables. Are there any variables that seem related so that some of them can be eliminated from further consideration? (5 pts)

```r
BlackAsh1_data=subset(newdata,plant=="BlackAsh1",select=c(temper,rh,vpd,rain,srad))
pairs(BlackAsh1_data)
```

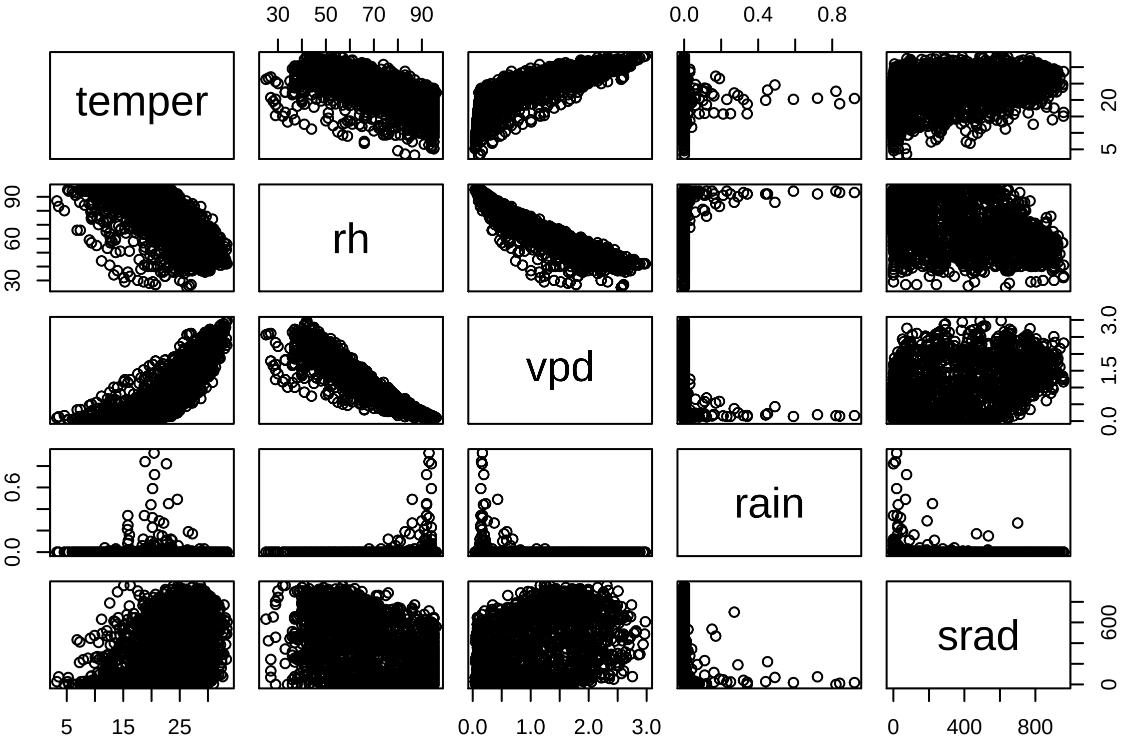
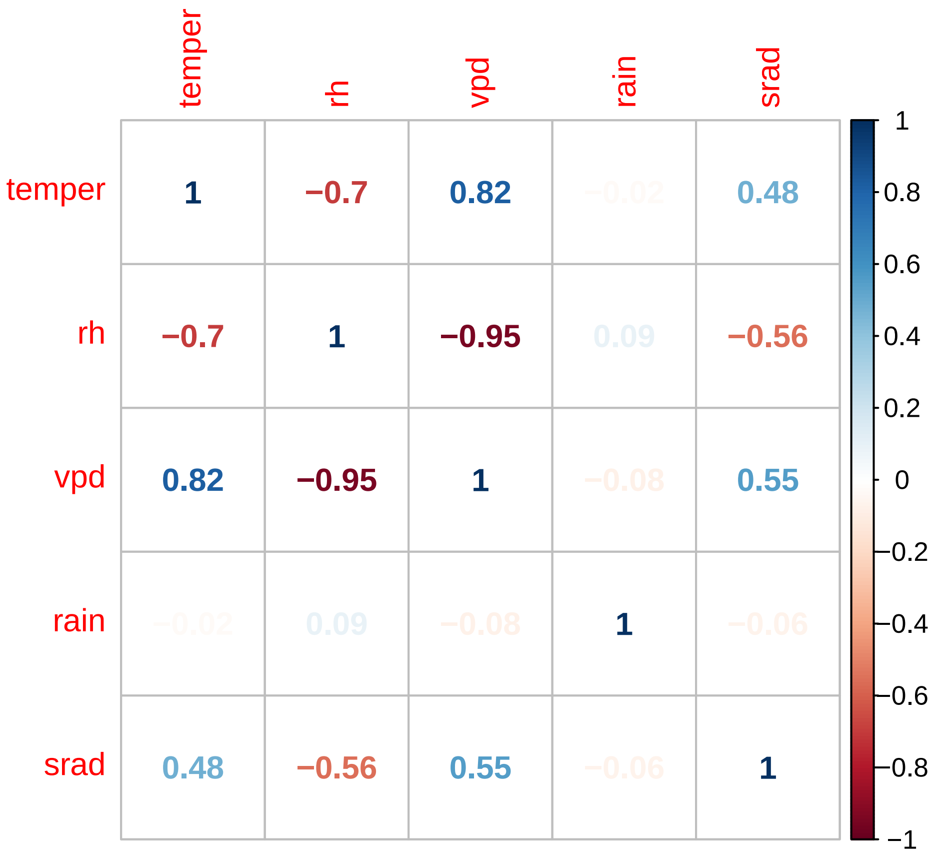


Figure 4. Scatter plot matrix for different enviromental variables

Vapor pressure deficit(vpd) definitely has some significant colinearity with relative humidity(rh) and temperature(temper). Relative humidity and temperature also seem to be related to some extent.

b. Calculate all correlations between *temper*, *rh*, *vpd*, *rain*, and *srad* (google "correlation matrix in r" to see how to do this). Construct a 5 by 5 table that contains these correlations rounded to two decimal places. Your table's rows and columns should be the variable names or some reasonable shortcuts like *SR* for solar radiation, etc. Identify variables such that $|r| < 0.55$ that should be considered in further analyses. (5 pts)

```
M <- cor(BlackAsh1_data,use="complete.obs")
corrplot(M, method = "number")
```



	temper	rh	vpd	rain	srad
temper	1	−0.7	0.82	−0.02	0.48
rh	−0.7	1	−0.95	0.09	−0.56
vpd	0.82	−0.95	1	−0.08	0.55
rain	−0.02	0.09	−0.08	1	−0.06
srad	0.48	−0.56	0.55	−0.06	1

Table 3. Correlation matrix for different enviromental variables

In the table, rh and vpd in association with other variables give $|r|$ values that are greater than or equal to 0.55. Therefore, these variables will not be considered for further analysis. Thus the variables that will be accounted for are temper, rain, and srad only.

c. For analysis, fit a model that includes *location*, *plant* and the variables you chose from part (b). Your model should contain *location*, *plant*, and all possible two-way interactions between the variables from part (b). To obtain the simplest possible model, we will perform model selection procedure called *backward selection*. Here are the steps to carry out backward selection:
   1. Fit a model containing all model terms.
   2. Identify a term with the highest p-value.
   3. Re-fit the model without this term.
   4. Keep repeating steps 2 and 3 until all p-values are less than 0.05.

Write down your final model obtained through backward selection. (10 pts)

```
data_new=subset(newdata,select=c(y,plant,location,temper,rain,srad))
ln_y<-log(data_new$y)
fit_stuff <- lm(ln_y ~ location+plant+temper+rain+srad+
                  temper:rain+temper:srad+rain:srad, data=data_new)
```

All of the terms in this model have a p-value of less than 0.05. Therefore no futher improvement is possible.

Therefore the final model is:

$$
\begin{gathered}
\ln(\text{Sap Flow Rate}) = -5.917 + 2.439 I_{\text{Longspark}}(X_{\text{L}}) + 4.548 I_{\text{Reservoir}}(X_{\text{L}}) \\
- 0.951 I_{\text{BlackAsh1}}(X_{\text{S}}) - 0.912 I_{\text{BlackAsh2}}(X_{\text{S}}) - 1.185 I_{\text{GoldGinko}}(X_{\text{S}}) \\
+ 0.588 I_{\text{GreenVase}}(X_{\text{S}}) + 0.689 I_{\text{NM1}}(X_{\text{S}}) + 1.165 I_{\text{PaperbackMaple}}(X_{\text{S}}) \\
- 1.605 I_{\text{Parroitta}}(X_{\text{S}}) - 0.342 I_{\text{RedMaple}}(X_{\text{S}}) - 3.676 I_{\text{RedMaple1}}(X_{\text{S}}) \\
- 1.575 I_{\text{RedMaple2}}(X_{\text{S}}) - 0.433 I_{\text{SugarMaple}}(X_{\text{S}}) - 0.303 I_{\text{Sycamore1}}(X_{\text{S}}) \\
+ 0.110(\text{Temperature}) + 6.541(\text{Rain}) + 0.007(\text{Solar radiation}) \\
-0.379(\text{Temperature:Rain}) - 0.0002(\text{Temperature:Solar Radiation}) \\
+0.008(\text{Rain:Solar Radiation})
\end{gathered}
$$

where X_{L} is the location and X_{S} is the specie of of a particular plant